ACADÉMIE DE MONTPELLIER

UNIVERSITÉ MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

THÈSE

présentée à l'Université Montpellier II – Sciences et Techniques du Languedoc
pour obtenir le diplôme de **DOCTORAT**

Spécialité: Mathématiques (Probabilités)
Formation Doctorale: Biostatistique

UNE THÉORIE ASYMPTOTIQUE DES ALGORITHMES GÉNÉTIQUES

par

Raphaël CERF

Soutenue le 24 mars 1994 devant le Jury composé de

MM.	AZENCOTT Robert,	Professeur,	ENS de Cachan,	Rapporteur
	BEN AROUS Gérard,	Professeur,	Université Paris XI,	Examinateur
	BERLINET Alain,	Professeur,	Université Montpellier II,	Directeur de Thèse
	ESCOUFIER Yves,	Professeur,	Université Montpellier II,	Président
	YCART Bernard,	Professeur,	Université Grenoble I,	Examinateur
	BOLTHAUSEN Erwin,	Professeur,	Université de Zürich,	Rapporteur

Remerciements

Voici enfin venu le temps de conclure ce travail. Bien souvent, cette page de remerciements reste formelle. En ce qui me concerne, je voudrais ici parler de façon moins impersonnelle de tout ce qui, directement ou non, a contribué à la genèse de ce travail. Qu'on veuille donc me pardonner ces longueurs, dont la lecture restera bien sûr facultative.

Mes premières pensées vont à ma famille –à ma grand–mère, à mon frère Max– à mes parents Edith et Jean[†] qui m'ont soutenu et encouragé tout au long de mes études. Leur capacité d'écoute et de compréhension envers mes histoires de mathématicien et la sagesse de leurs conseils furent maintes fois déterminantes pour mon orientation. La maison de Romans a toujours été un havre de paix et de tranquillité où, magnifiquement accueilli, j'étais heureux de revenir travailler.

C'est Pierre–Jean Hormière, mon Professeur de math–sup, qui m'a insufflé la passion des mathématiques. Et dans mon travail d'aujourd'hui, je ressens encore l'influence de son enseignement.

En 1988, j'accédai à la rue d'Ulm: le directeur d'alors, Georges Poitou, fut pour beaucoup dans cette décision. Les quatre années que j'ai passées dans cet eden universitaire m'ont marqué à jamais et m'ont fait découvrir l'humanisme et la tolérance intellectuelle. Là se sont forgées de solides amitiés et j'ai eu la chance de côtoyer des individus hors normes. Quelques noms me viennent à l'esprit:
TMOY, Jean–François Mallordy, Jean–Marc Daul, Osnola, Julien Cassaigne, FX–Dehon, Taar, Titou, Nathanaël Enriquez, Gaël Giraud, Antoine Balan, Lafforgue & Cie, Fage, petit Patrick, Chief, Vmaillot, Franck Plouraboué, Laurent Dumas, Pascal Boyer, TVTT, Jean–Louis Pretini, Olivier Compagnon, Barthélémy Fayes, Julien Mechler, David Harari, Glockenspiel, Stéphane Grognet, Pollardux, Jean–David Fermanian, Stéphane Guillermou, Kirone Mallick, Jean–Marc Bottazzi ... Que ceux que j'oublie me pardonnent.

L'histoire de cette thèse est peu banale: mais je crois que pour réaliser un travail de cette sorte, il faut s'échapper de la routine universitaire. Au cours de l'année 1990, j'entendis pour la première fois parler d'algorithmes génétiques. Je souhaitais alors faire de l'informatique et j'étais engagé dans un milieu de biologistes et d'informaticiens. Là, Jean–Arcady Meyer se faisait l'apôtre de ces techniques, désignées sous le terme générique de "systèmes adaptatifs". La démarche intellectuelle qui avait abouti à la conception de ces algorithmes me séduisit d'emblée. Cependant, je continuai sur ma voie et fis un stage d'analyse d'images sous la direction de Pierre Vincens dans le groupe de BioInformatique de Philippe Tarroux et Jean–Arcady Meyer. A l'issue de ce stage, je décidai de me consacrer à nouveau aux mathématiques.

Toujours hanté par ces mystérieux algorithmes génétiques, je me lançai dans l'étude de la théorie des probabilités. Mes premiers pas dans ce domaine furent guidés par Omer Adelman, qui me fit découvrir l'extraordinaire richesse de la marche symétrique simple.

[†] En l'occurrence, il ne s'agit pas du mathématicien!

Après l'agrégation, je partis au service militaire, toujours à l'affût d'une piste qui permettrait d'attaquer efficacement ces algorithmes. La littérature et les prépublications sur le sujet s'accumulaient autour de moi sans qu'apparaisse l'ébauche d'une théorie solide: essentiellement des simulations et de petits calculs combinatoires partiels.

Au mois de mars 1992, Jean–Pierre Aubin, avec beaucoup de gentillesse, m'invita à assister à son cours sur la théorie de la Viabilité aux Houches. Là non plus je ne trouvai pas d'outil adapté à l'étude des algorithmes génétiques, mais je fis la connaissance d'un mathématicien italien, Carlo Mariconda, avec qui je devais faire mes premières armes en recherche mathématique. A partir d'un joli problème suggéré par Arrigo Cellina, nous avons écrit ensemble trois articles sur le principe boum–boum en théorie du contrôle.

J'explorais aussi la littérature de la génétique mathématique: sans résultats tangibles. Et de mes multiples discussions avec les probabilistes ne sortaient que des indications assez vagues. Mais j'étais plus que jamais décidé à en faire mon sujet de thèse.
Pourquoi rester à Paris puisqu'aucun mathématicien ne s'intéressait à ce sujet?

En juin 1992, j'écrivis trois lettres, dans les Universités de Lyon, Grenoble et Montpellier pour chercher un directeur. Yves Escoufier m'accueillit avec chaleur et enthousiasme à Montpellier et je commençai à travailler sous la direction d'Alain Berlinet.
Jean–Pierre Labesse, directeur des études du DMI à l'ENS, fut très intrigué, voire inquiété par ma démarche.
–“Ce n'est pas dans les habitudes de la maison d'empêcher les gens de faire ce qu'ils veulent, me dit–il, cependant votre truc me semble louche. Vous allez au moins en discuter encore un peu avant de vous décider.”
Il me fit rencontrer Jean–François Le Gall et Gérard Ben Arous. Ce dernier me conseilla d'aller voir Marie Duflo. Et un beau jour, début octobre 1992, je débarquai dans le bureau de Marie Duflo à Orsay. Gérard Ben Arous était aussi là.
–“Je veux quand même savoir de quoi il retourne.”
J'expose donc une nouvelle fois le problème.
–“Vous tenez là un très beau sujet” s'exclama Gérard Ben Arous. “Ce qu'il vous faut maintenant, c'est une puissante machinerie mathématique pour faire marcher tout ça. Regardez le livre de Marie Duflo: si tel ou tel objet de votre algorithme est une martingale, ça vous donnera des résultats de convergence. Et sinon, c'est Freidlin et Wentzell.”
–“Freidlin et Wentzell? Qu'est–ce que c'est?”
–“Voilà: vous avez un paysage dynamique et des attracteurs avec leurs bassins d'attraction, et vous étudiez la chaîne des visites ... Oh et puis non. Je ne peux pas vous expliquer.
Il faut que vous lisiez le chapitre 6 du Livre. Au début, vous allez atterrir là–dedans et vous serez noyé. Puis vous comprendrez. Allez donc aussi voir le travail de Catoni, et apprenez à faire converger le recuit simulé, juste pour le plaisir. Après vous pourrez retourner à votre algorithme.”
De tout ceci, exprimé avec un enthousiasme communicatif et dans un style percutant, je retins aussi quelques phrases chocs.
“On n'a pas besoin de connaître explicitement la matrice des transitions d'une chaîne de Markov: des estimations asymptotiques des coefficients suffisent.”

Je me plongeai donc dans la théorie de Freidlin–Wentzell et il me fallut trois mois d'efforts pour absorber ce fameux chapitre 6. Et je compris que je tenais enfin là le bon outil pour décortiquer les mécanismes cachés des algorithmes génétiques. La voie était ouverte, et pendant plus d'une année, je travaillais sans relâche avec une intensité nouvelle.

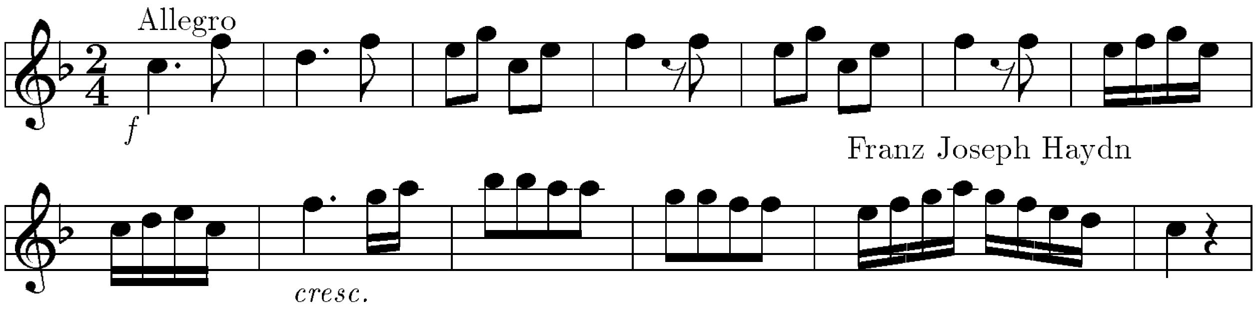
Accomplir ce travail sous la direction d'Alain Berlinet fut pour moi un plaisir. J'ai admiré son honnêteté intellectuelle exemplaire ainsi que son ouverture d'esprit. Constamment il m'a incité à nouer des contacts et à ne pas rester replié sur moi–même. Ses encouragements et ses conseils avisés m'ont été une aide précieuse. Je suis également reconnaissant à Yves Escoufier de m'avoir accueilli dans son équipe et de m'avoir fait bénéficier d'un cadre de travail privilégié. Tous deux resteront pour moi des modèles de l'Universitaire.

Avec le recul, je vois que les outils que j'ai utilisés ont été développés dans l'équipe de Robert Azencott. Mais mon isolement a eu pour effet de décupler mon énergie au travail, et aujourd'hui, je suis heureux d'avoir acquis mon autonomie intellectuelle.

Il n'est rien de plus précieux que de recevoir le bon conseil au bon moment. Et il est impossible de savoir quelle est la personne qui détient la lumière. C'est pourquoi il ne faut jamais relâcher sa vigilance.

Voilà. Je suis très reconnaissant envers les Professeurs Robert Azencott et Erwin Bolthausen qui ont accepté la tâche ingrate de rapporteur, et je remercie les Professeurs Gérard Ben Arous, Robert Azencott, Alain Berlinet, Yves Escoufier et Bernard Ycart qui m'honorent en participant au jury de cette thèse.

"Les théories sont des filets: seul celui qui lance, pêchera." Novalis



Patience, patience,
Patience dans l'azur!
Chaque atome de silence
Est la chance d'un fruit mûr!

Paul Valéry

Résumé

Les algorithmes génétiques sont des procédures d'exploration inspirées des mécanismes génétiques qui guident l'évolution naturelle: mutation, crossing–over, sélection. Ils sont aujourd'hui massivement utilisés pour de nombreux problèmes d'optimisation, allant du problème du voyageur de commerce à la conception d'architectures de réseaux. Cependant les résultats théoriques de convergence manquent cruellement pour ces algorithmes.

La théorie de Freidlin–Wentzell étudie les perturbations aléatoires de systèmes dynamiques. Nous construisons plusieurs modèles d'algorithmes génétiques en perturbant aléatoirement des processus rudimentaires. La dynamique asymptotique des processus ainsi obtenus est analysée à l'aide des outils puissants développés par Freidlin et Wentzell.

Dans un premier article, un modèle markovien de l'algorithme génétique simple de Holland est construit en perturbant aléatoirement un mécanisme de sélection très simple: les mutations et les crossing–over sont considérés comme des perturbations aléatoires qui s'évanouissent. La convergence vers les maxima globaux de la fonction d'adaptation devient possible lorsque la taille de la population dépasse une valeur critique qui dépend fortement du problème d'optimisation. Le crossing–over n'est pas du tout fondamental pour entraîner cette convergence: le point crucial est la délicate interaction asymptotique entre les mutations et la force de sélection.

Dans un second article, nous utilisons les concepts introduits par Catoni et ensuite généralisés par Trouvé pour sonder plus profondément la dynamique de l'algorithme de mutation–sélection lorsque la taille de la population devient très grande. Le résultat clef repose dans la structure des trajectoires de populations qui joignent deux populations uniformes: un petit groupe d'individus se sacrifie afin de créer un chemin idéal qui est ensuite emprunté par tous les autres individus. Les différentes quantités associées à l'algorithme sont alors des fonctions affines de la taille de la population. La hiérarchie des cycles sur l'ensemble des populations uniformes se stabilise. En outre, si le noyau de mutation est symétrique, la distribution limite est la distribution uniforme sur l'ensemble des maxima globaux de la fonction d'adaptation.

Dans un troisième article, un nouvel algorithme génétique est présenté. Il est construit en perturbant aléatoirement un schéma à deux opérateurs, crossing–over et sélection. Trois conditions de nature biologique sont imposées sur le crossing–over. Un nouveau mécanisme de sélection est utilisé, qui possède l'avantage décisif de préserver la diversité des individus dans la population. Les attracteurs du processus non perturbé sont des sous–ensembles particuliers de populations d'équi–adaptation munis d'une structure riche. Les perturbations aléatoires évanescentes sont de deux sortes: des mutations et un relâchement de la force de sélection. Lorsque la taille de la population dépasse une valeur critique, leur délicate interaction asymptotique entraîne la convergence de la population vers l'attracteur idéal dont les populations contiennent tous les maxima globaux de la fonction d'adaptation.

Mots clefs Freidlin–Wentzell, algorithmes génétiques, optimisation stochastique

Plan de la thèse

Cette thèse se compose de trois articles en anglais:

Asymptotic convergence of genetic algorithms,
The dynamics of mutation-selection algorithms with large population sizes,
A new genetic algorithm.

Le premier chapitre est une synthèse en français des résultats essentiels.
Les trois chapitres suivants correspondent aux trois articles.
L'annexe A est une étude bibliographique qui avait été réalisée au début de ce travail afin de déterminer les directions de recherche possibles sur la théorie des algorithmes génétiques.
Les annexes B et C sont des notes qui résument le premier et le troisième article, soumises aux "Comptes Rendus de l'Académie des Sciences".

Table des matières

Main notations

$\lfloor s \rfloor$	the unique integer such that $\lfloor s \rfloor \le s < \lfloor s \rfloor + 1$	$46, 99, 144$
$\lvert X \rvert, \operatorname{card} X$	the cardinality of the set X	$45, 99, 144$
E	a finite set	$46, 99, 144$
i, j, e	points of E	$46, 99, 144$
f	a positive function defined on E (the fitness function)	$46, 99, 144$
f_λ	the set $\{\, i \in E : f(i) = \lambda \,\}$	$65, 145$
f_λ^-	the set $\{\, i \in E : f(i) < \lambda \,\}$	$65, 145$
f_λ^+	the set $\{\, i \in E : f(i) > \lambda \,\}$	$65, 145$
f^*	the set of the global maxima of f	$46, 99, 144$
m	a population size	$46, 99, 144$
E^m	the set of m–uples of elements of E	$46, 99, 144$
x, y, z, u, v	elements of E^m	$46, 99, 144$
$\widehat{f}(x)$	the maximum of the set $\{\, f(x_1), \cdots, f(x_m) \,\}$	$46, 99, 144$
$[x]$	the set $\{\, x_1, \cdots, x_m \,\}$	$46, 99, 144$
$\widehat{x}$	the points of $[x]$ which realize the value $\widehat{f}(x)$	$46, 99, 144$
(i)	the m–uple whose all components are equal to i	$46, 99$
$x(i)$	the number of occurrences of i in x	$46, 99$
S	the set of equi–fitness populations	46
U	the set of uniform populations (A in the first paper)	99
$\mathfrak{S}_r, \mathfrak{S}(r)$	the group of permutations over $\{1 \cdots m\}$	$46, 145$
a, b, c	three positive real numbers	$55, 100, 151$
$\delta(i, j)$	the Kronecker symbol (the identity matrix indexed by E)	$46, 99, 144$
α	a communication kernel on E (the mutation kernel)	$50, 100, 151$
β	a communication kernel on $E \times E$ (the crossover kernel)	$50, 146$
δ	the minimum of $\{\, f(i) - f(j) : i, j \in E,\ f(i) \ne f(j) \,\}$	$52, 99, 144$
Δ	the maximum of $\{\, f(i) - f(j) : i, j \in E \,\}$	$68, 99, 144$
δ^*	the minimum of $\{\, f(f^*) - f(i) : i \in E \setminus f^* \,\}$	136
ρ	the minimum $\min(a, b/2, c\delta)$	70
Δ°, Δ^*	quantities related to f and the communication kernel α	68
$\Delta^{\circledast}$	the maximum $\max(\Delta^\circ, \Delta^*)$	68
R	minimal number of transitions to join 2 points through α	$68, 103, 164$
p, q	paths in E^m and $\mathcal{P}(E)$	$104, 156$
$\lvert p \rvert$	the length of the path p	$104, 156$
$m(p)$	the dimension of the vectors in the path p	105
$d(x, y)$	the Hamming distance between x and y	$54, 101, 155$
g	a graph (usually on E or E^m)	
g_λ	the restriction of g to the level λ	$65, 166$

$G(x)$	the set of x–graphs	$61, 102, 157$
$G_X(Y)$	the set of Y–graphs over $X \cup Y$	$65, 166$
q_1	a communication kernel on E^m	$102, 155$
V_1	the one–step communication cost	$102, 104, 155$
V	the minimal communication cost	$58, 105, 157$
$\widetilde{V}$	the cost constrained outside the attractors	58
V_H	the cost constrained outside the set H	107
$\overline{V}$	the cost for paths corresponding to the whole process	$104, 156$
W_1, W_H, W	the virtual energies associated with V_1, V_H, V	$102, 105, 107$
A_1, A_H, A	the communication altitude associated with V_1, V_H, V	$106, 107, 157$
W_1^*, W^*	the set of the minima of the virtual energy	$102, 105, 157$
$W_X(Y)$	the minimal cost of a graph in $G_X(Y)$	$65, 166$
$G_X^*(Y)$	the graphs realizing the value $G_X(Y)$	$65, 166$
$[V], [\overline{V}]$	the communication costs induced on $\mathcal{P}(E)$ by V and $\overline{V}$	105
π	a cycle	
H_e	the height of exit	134
H_m	the height of mixing	134
H_1	the critical height	$103, 136, 173$
H_e^*	the height of exit of the attraction basin of W^*	$136, 173$
$l, l(n)$	the parameter controlling the random perturbations	$47, 100, 151$
X_n^∞	the unperturbed process	$47, 101, 146$
X_n^l	the perturbed process	$47, 100, 151$
$U_n^l, V_n^l, Z_n^l, Y_n^l$	intermediate populations in the transition mechanism	$47, 100, 151$
K	an attractor	148
$\mathcal{K}$	the set of all attractors	148
$[K]$	the set of all points present in the attractor K	145
$\widehat{f}(K)$	the maximum of f over the set $[K]$	145
$\widehat{K}$	the points of $[K]$ realizing the value $\widehat{f}(K)$	145
Λ	the maximum cardinal of a set $f_\lambda, \lambda \in \mathbb{R}^+$	145
$\mathcal{K}_\lambda$	the attractors included in f_λ	148
$\mathcal{K}_\lambda^-, \mathcal{K}_\lambda^+$	the attractors included in f_λ^-, f_λ^+	148
x^λ	the points of $[x]$ belonging to f_λ	145
λ^x	the sequence of values appearing in $f[x]$	145
$\tau(k, h)$	a triangular array of integers	147
$\underset{\infty}{\precsim}, \prec$	two relations over $\mathcal{K}$	$150, 171$
$\mathcal{E}$	a subset of E in one–to–one correspondence with $\mathcal{K}$	149
K^*	the ideal attractor which satisfies $[K^*] = f^*$	148
m^*	the critical population size	$72, 103, 170$
α_{opt}	the optimal convergence exponent	$137, 174$

Chapitre Premier

UNE THÉORIE ASYMPTOTIQUE DES ALGORITHMES GÉNÉTIQUES

(SYNTHÈSE)

UNE THÉORIE ASYMPTOTIQUE DES ALGORITHMES GÉNÉTIQUES (SYNTHÈSE)

Raphaël CERF

Université de Montpellier

Janvier 1994

Ce texte est une synthèse simplifiée des trois articles sur les algorithmes génétiques

Asymptotic convergence of genetic algorithms,

The dynamics of mutation-selection algorithms with large population sizes,

A new genetic algorithm.

Seuls les résultats essentiels sont cités et toutes les preuves ainsi que les détails techniques sont omis. Nous présentons tout d'abord dans une première partie le fonctionnement de l'algorithme génétique simple classique, tel qu'il apparaît actuellement dans la littérature. Nous donnons ensuite dans une seconde partie un bref aperçu de la théorie de Freidlin–Wentzell des perturbations aléatoires de systèmes dynamiques: tous nos modèles sont analysés grâce aux idées fondamentales de cette théorie. Les trois sections suivantes sont des résumés des trois articles. Nous suggérons pour finir quelques directions de travail.

1. L'algorithme génétique simple classique

Soit f une fonction définie sur l'espace $\{0,1\}^N$ à valeurs dans $\mathbb{R}_+^*$. Le problème est de localiser l'ensemble f^* des maxima globaux de f, ou, à défaut, de trouver le plus rapidement et le plus efficacement possible des points sous–optimaux (l'entier N est grand et une recherche exhaustive est exclue).

Les algorithmes d'évolution sont des procédures de recherche qui s'inspirent des mécanismes de la sélection naturelle et des phénomènes génétiques [7,8]. Le principe consiste à faire évoluer un ensemble de points de l'espace $\{0,1\}^N$ comme s'il s'agissait d'une population naturelle d'individus ou de chromosomes, la fonction f mesurant l'adaptation d'un individu à l'environnement. Nous décrivons maintenant le fonctionnement de l'algorithme génétique simple classique, qui est l'algorithme de base le plus répandu dans la littérature et la pratique des simulations [7]. De nombreuses variantes et extensions sont possibles.

Il s'agit donc d'un algorithme stochastique et itératif qui opère sur des ensembles de points, et qui est bâti à l'aide de trois opérateurs: mutation, crossing–over, sélection. Soit $m \in \mathbb{N}$ la taille (fixe) de la population. Notons X_n la population à l'instant n: c'est un vecteur $X_n = (X_n^1, \cdots, X_n^m)$ dont les m composantes sont des mots de longueur N sur l'alphabet $\{0,1\}$. Le passage de X_n à X_{n+1} se décompose en trois étapes:

$$X_n \quad \xrightarrow{\text{mutation}} \quad Y_n \quad \xrightarrow{\text{crossing–over}} \quad Z_n \quad \xrightarrow{\text{sélection}} \quad X_{n+1}.$$

Décrivons en détail chacune de ces étapes.

$X_n \longrightarrow Y_n$: **mutation.** Un paramètre p entre 0 et 1 est initialement choisi: c'est la probabilité de mutation. Pour chaque lettre de chaque chromosome $X_n^1, \cdots, X_n^m$, une variable de Bernoulli de paramètre p est tirée indépendamment et, suivant le résultat, la lettre associée est changée ou non, i.e. avec probabilité p elle est remplacée par l'autre lettre $(0 \to 1, 1 \to 0)$ et avec probabilité $1-p$ elle n'est pas modifiée. En général, le paramètre p est choisi "petit". La nouvelle population obtenue après application de l'opérateur de mutation est notée Y_n.

$Y_n \longrightarrow Z_n$: **crossing–over.** Ici encore, un paramètre q entre 0 et 1 est fixé initialement: c'est la probabilité de crossing–over. Pour construire la population Z_n, $m/2$ couples sont formés à partir de la population Y_n (par exemple en appariant les individus consécutifs de Y_n : $(Y_n^1, Y_n^2), (Y_n^3, Y_n^4), \cdots$ ou bien en choisissant au hasard et uniformément des individus dans Y_n). Pour chaque couple, une variable de Bernoulli de paramètre q est tirée pour décider si le crossing–over a lieu. Si c'est le cas, un site de coupure est tiré au hasard, uniformément entre 1 et $N-1$, et les segments finaux des deux chromosomes sont échangés.

$$\begin{array}{r|l} 000\text{———}00011 & 01101\text{———}001 \\ 100\text{———}01110 & 00110\text{———}111 \end{array} \quad \longrightarrow \quad \begin{array}{r|l} 000\text{———}00011 & 00110\text{———}111 \\ 100\text{———}01110 & 01101\text{———}001 \end{array}$$

site de coupure

Une nouvelle paire d'individus est ainsi obtenue (identique à l'ancienne s'il n'y a pas eu de crossing–over) et est stockée dans la population Z_n. En général, le paramètre q est choisi "grand".

Remarquons que les opérateurs de mutation et de crossing–over ne font pas intervenir la fonction f: ce sont des opérateurs stochastiques d'exploration. C'est le troisième et dernier opérateur, la sélection, qui guide la population vers les valeurs élevées de la fonction f.

$Z_n \longrightarrow X_{n+1}$**: sélection.** Les m individus de la population X_{n+1} sont sélectionnés indépendamment à l'aide d'une distribution de probabilité sur la population Z_n qui favorise les individus de Z_n les mieux adaptés. Le choix le plus fréquent est l'unique distribution telle que la probabilité d'un individu soit proportionnelle à son adaptation, i.e.

$$\forall h \in \{1 \cdots m\} \qquad P\left(Z_n^h\right) \;=\; \frac{f(Z_n^h)}{f(Z_n^1) + \cdots + f(Z_n^m)}.$$

Mathématiquement, la suite $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov d'espace d'états $\left(\{0,1\}^N\right)^m$. La loi de X_n est déterminée de manière unique par la donnée de la loi de X_0 (en général, la loi uniforme sur $\left(\{0,1\}^N\right)^m$) et du mécanisme de transition précédemment décrit.
Ce mécanisme, au demeurant fort compliqué (il est scindé en trois étapes non triviales), possède néammoins quelques propriétés essentielles.
- Il est homogène (i.e. indépendant du temps n).
- Il est irréductible; étant données deux populations arbitraires x et y, la probabilité partant de x d'obtenir y en un nombre fini de transitions est non nulle:

$$\forall x, y \in \left(\{0,1\}^N\right)^m \quad \exists r \in \mathbb{N} \quad P\left(X_{n+r} = y / X_n = x\right) > 0.$$

Autrement dit, le mécanisme de transition permet d'explorer tout l'espace des populations.
- Il est apériodique, i.e.

$$\forall x \in \left(\{0,1\}^N\right)^m \qquad \mathrm{pgcd}\,\{\, r : P\left(X_{n+r} = x / X_n = x\right) > 0 \,\} = 1.$$

Cette dernière condition, plus technique que les précédentes, n'est pas fondamentale.
Ouvrons à ce point n'importe quel ouvrage de théorie élémentaire des chaînes de Markov. Très rapidement apparaît le théorème suivant:

Théorème. *Une chaîne de Markov homogène irréductible apériodique à espace d'états fini est ergodique et possède une unique mesure de probabilité stationnaire ou invariante.*

Ainsi, pour toute population de départ x, la limite

$$\lim_{n\to\infty} P\left(X_n = y / X_0 = x\right) = \mu(y)$$

existe et est indépendante de x. La mesure de probabilité μ correspond en quelque sorte à l'état d'équilibre du processus $(X_n)_{n\in\mathbb{N}}$ et la loi de X_n converge vers μ. De plus cette mesure μ charge positivement chaque population. Le mécanisme de transition donne cependant l'impression que la mesure μ est plutôt concentrée sur les populations qui contiennent des points où la fonction f est élevée, vu que la sélection favorise ces points.

L'algorithme décrit ici n'est donc pas à proprement parler un algorithme d'optimisation, mais plutôt une promenade au hasard dans l'espace des populations qui est attirée vers les hautes valeurs de la fonction d'adaptation f. Beaucoup plus intéressant serait de construire un processus $(X_n)_{n\in\mathbb{N}}$ dont la loi se concentre progressivement sur les populations ne contenant que des maxima globaux, i.e.

$$\forall x \in \left(\{0,1\}^N\right)^m \qquad \lim_{n\to\infty} P\left(\{X_n^1, \cdots X_n^m\} \subset f^* / X_0 = x\right) = 1.$$

Remarquons enfin que le processus (X_n) est contrôlé par de nombreux paramètres: les probabilités de mutation et de crossing–over p et q, la taille de la population m. En outre, de nombreux choix, pas du tout innocents, doivent être faits pour le mettre en œuvre: par exemple, les mécanismes de mutation et de crossing–over, et aussi la distribution de probabilité utilisée pour la sélection. Tout ceci suscite naturellement de nombreuses questions. Jusqu'à présent, seules des considérations heuristiques et des indications expérimentales guidaient les décisions des utilisateurs d'algorithmes génétiques. Mais les problèmes fondamentaux ne sont pas résolus. Comment faut–il contrôler un tel algorithme? Vers quoi est-il possible de le faire converger?

Essayons tout d'abord d'avoir une idée des paramètres à l'œuvre dans la nature.
Voici quelques phrases tirées de [9, p31] au sujet du taux de mutation.
"La détermination du taux de mutation spontanée est fondamentale pour l'étude de l'évolution, cependant ce paramètre crucial n'a pas été mesuré avec une précision suffisante au niveau moléculaire: le taux de mutation est généralement très faible et pour obtenir une estimation précise, il est nécessaire d'étudier un grand nombre de gènes. Les généticiens classiques ont déterminé que le taux de mutation qui change des caractères phénotypiques ou provoque des effets mortels est de l'ordre de 10^{-5} par locus par génération dans les organismes supérieurs comme l'homme, la drosophile et le maïs. Cependant, on en sait très peu sur la nature de ces mutations au niveau de l'ADN. En outre, la plupart de ces mutations sont délétères et contribuent peu aux changements dans l'évolution des organismes. Récemment, il y a eu de nombreuses tentatives de mesurer le taux de mutation au niveau moléculaire. En étudiant 7 loci d'enzyme et 3.111.598 générations de gènes (=nombre de gènes examinés $\times$ nombre de générations), Voelker et al. [13] ont estimé que les taux des changements de mobilité électrophorétique et des mutations nulles étaient de 1.28×10^{-6} et 3.86×10^{-6} par locus et par génération, respectivement."
Quant au crossing–over, la situation est plus délicate: il ne semble pas qu'il y ait une quantité biologique directement comparable avec les paramètres de l'algorithme génétique. Cependant, il apparaît que plusieurs crossing–over se produisent d'une génération à l'autre.

2. La théorie de Freidlin–Wentzell

Le but de cette section est de donner un bref aperçu de la théorie de Freidlin–Wentzell des petites perturbations aléatoires de systèmes dynamiques [6]. Il s'agit d'une présentation très simplifiée, voire simplificatrice, dépourvue de tout détail technique.

Considérons un système dynamique dans $\mathbb{R}^d$:

$$dx_t \;=\; b(x_t)\,dt. \tag{1}$$

Sous des hypothèses de régularité minimales sur la fonction b, il existe une unique trajectoire $x(t)$ satisfaisant l'équation (1) et une condition initiale $x_0 = x \in \mathbb{R}^d$.
Que peut–il arriver à cette trajectoire?
Elle peut par exemple converger vers un point d'équilibre:

$$\lim_{t\to\infty} x_t = x_\infty.$$

Plus généralement, elle peut se rapprocher indéfiniment d'un ensemble limite. Définissons l'ensemble ω–limite associé au point de départ x comme l'ensemble des valeurs d'adhérence (les points dont tout voisinage est visité une infinité de fois) de la trajectoire x_t issue de x. Un tel ensemble est compact et sera désormais désigné par le terme “attracteur”.
Nous faisons l'hypothèse suivante:

Hypothèse. *L'espace $\mathbb{R}^d$ contient un nombre fini r d'attracteurs $K_1, \cdots, K_r$.*

A chaque attracteur est associé un bassin d'attraction: c'est l'ensemble des points de départ pour lesquels le système converge vers l'attracteur en question. L'espace $\mathbb{R}^d$ est ainsi partitionné en r bassins d'attractions. Nous distinguons deux catégories d'attracteurs: les attracteurs instables (il existe des points infiniment proches de l'attracteur qui appartiennent à un autre bassin d'attraction) et les attracteurs stables (le bassin d'attraction est un voisinage de l'attracteur).

La théorie de Freidlin–Wentzell étudie ce qui se passe lorsque le système précédent est soumis à de petites perturbations browniennes. L'intensité des perturbations aléatoires est contrôlée par un paramètre positif ϵ (ϵ est petit). L'équation déterministe

$$dx_t \;=\; b(x_t)\,dt, \qquad x_0 = x \tag{2}$$

se transforme en une équation différentielle stochastique

$$dX_t^\epsilon \;=\; b(X_t^\epsilon)\,dt \,+\, \epsilon\,dw_t, \qquad X_0^\epsilon = x \tag{3}$$

où $(w_t)_{t\in\mathbb{R}_+}$ est le mouvement brownien standard dans $\mathbb{R}^d$.
L'objet obtenu $(X_t^\epsilon)_{t\in\mathbb{R}_+}$ est un processus stochastique guidé par le mouvement brownien $(w_t)_{t\in\mathbb{R}_+}$ et dépendant de ϵ. Le paramètre ϵ est petit et, avec grande probabilité, les trajectoires de (X_t^ϵ) sont proches des trajectoires du système déterministe (2).

Cependant, la situation change radicalement. En effet, le mouvement brownien possède la propriété fondamentale suivante: étant donnés un ouvert borné arbitraire de $\mathbb{R}^d$ et un point x de cet ouvert, le mouvement brownien issu de x sort de l'ouvert au bout d'un temps fini avec probabilité 1. Il s'ensuit que le processus (X_t^ϵ) s'échappe avec probabilité 1 de n'importe quel bassin d'attraction, et en fait, il les visite tous.
De plus, le processus (X_t^ϵ) est ergodique et admet une unique mesure de probabilité invariante, i.e. la limite

$$\lim_{t \to \infty} P(X_t^\epsilon \in A / X_0^\epsilon = x) \;=\; \mu^\epsilon(A)$$

existe pour tout ensemble borélien A de $\mathbb{R}^d$ et est indépendante du point de départ x.
La quantité $\mu^\epsilon(A)$ peut s'interpréter comme la probabilité de présence du processus dans A lorsque le système a atteint son état d'équilibre.
Cette mesure μ^ϵ charge positivement tous les ouverts de $\mathbb{R}^d$. Toutefois, lorsque ϵ devient petit et que les perturbations aléatoires disparaissent, le processus passe le plus clair de son temps au voisinage des attracteurs et μ^ϵ se concentre sur les attracteurs: pour tout voisinage $\mathcal{V}(K_1 \cup \cdots \cup K_r)$ de l'ensemble des attracteurs,

$$\lim_{\epsilon \to 0} \mu^\epsilon\big(\mathcal{V}(K_1 \cup \cdots \cup K_r)\big) \;=\; 1$$

et pour tout ensemble borélien A disjoint de $K_1, \cdots, K_r$,

$$\lim_{\epsilon \to 0} \mu^\epsilon(A) \;=\; 0.$$

Plusieurs questions surgissent naturellement.
Est-ce que la limite $\lim_{\epsilon \to 0} \mu^\epsilon$ existe?
Est-ce que la mesure μ^ϵ se concentre de manière identique sur tous les attracteurs ou bien existe–t–il des attracteurs privilégiés et d'autres pour lesquels

$$\lim_{\epsilon \to 0} \mu^\epsilon(\mathcal{V}(K_h)) \;=\; 0 \;?$$

Quelles sont les vitesses de convergence?

A partir du processus $(X_t^\epsilon)_{t \in \mathbb{R}_+}$ à temps continu et espace d'états continu $\mathbb{R}^d$, Freidlin et Wentzell construisent un processus $(Z_n^\epsilon)_{n \in \mathbb{N}}$ à temps discret et espace d'états fini $\{1 \cdots r\}$, défini ainsi: Z_n^ϵ est le numéro (symbolique) du nième attracteur visité par le processus (X_t^ϵ). Si le point de départ est "proche" de l'attracteur K_h, alors $Z_0^\epsilon = h$. Le processus passe d'abord un certain temps dans le bassin d'attraction de K_h, puis le mouvement brownien cause un changement de bassin d'attraction: le processus est attiré par K_s et $Z_1^\epsilon = s$. Nous passons sous silence tous les détails techniques relatifs à la construction de Z_n^ϵ: il faut définir précisément ce que signifie visiter un attracteur, sortir et rentrer dans un bassin d'attraction. De la nature markovienne des perturbations, il découle que Z_n^ϵ est une chaîne de Markov. Cette chaîne est irréductible et a un espace d'états fini $\{1 \cdots r\}$. Elle possède une unique mesure de probabilité invariante, soit ν^ϵ.
L'idée essentielle sur laquelle repose l'étude de Freidlin et Wentzell peut s'énoncer ainsi:

Principe. *L'étude asymptotique lorsque ϵ tend vers 0 de la dynamique du processus (X_t^ϵ) est "équivalente" à l'étude asymptotique lorsque ϵ tend vers 0 de la dynamique de la chaîne symbolique (Z_n^ϵ) des visites successives des attracteurs.*

En particulier, l'étude du comportement asymptotique de la mesure μ^ϵ se ramène à l'étude du comportement asymptotique de la mesure ν^ϵ. Plus précisément, il existe une formule qui exprime μ^ϵ en fonction de ν^ϵ et des excursions du processus en dehors des attracteurs: soit A un borélien de $\mathbb{R}^d$, $\mathcal{V}$ un voisinage de l'ensemble des attracteurs, $\partial\mathcal{V}$ sa frontière, et $\tau_{\mathcal{V}}$ le temps d'atteinte de $\mathcal{V}$. Alors

$$\mu^\epsilon(A) \ = \ \int_{\partial\mathcal{V}} \nu^\epsilon(dy) E_y \left[\int_0^{\tau_{\mathcal{V}}} 1_A(X_t^\epsilon)\, dt \right]$$

(où E_y dénote l'espérance pour le processus partant de y).
L'étape suivante consiste à étudier la chaîne (Z_n^ϵ), et tout d'abord son mécanisme de transition. Soient K_i et K_j deux attracteurs. La probabilité de transition $P\,(Z_n^\epsilon = j / Z_n^\epsilon = i)$ peut s'écrire comme une intégrale sur un espace fonctionnel. Grossièrement,

$$P\,(Z_n^\epsilon = j / Z_n^\epsilon = i) \ = \ \int_{\{\ \phi : \phi_0 \in K_i,\, \phi_T \in K_j\ \}} dP\,(X_\cdot^\epsilon \text{ suive } \phi \text{ sur } [0,T])\,. \tag{4}$$

L'ensemble d'intégration est l'ensemble des fonctions continues définies sur un intervalle $[0,T]$ (T est laissé variable) à valeurs dans $\mathbb{R}^d$ et qui connectent les attracteurs K_i et K_j. L'intégrand $dP\,(X_t^\epsilon$ suive $\phi_t)$ est la probabilité élémentaire que la trajectoire de (X_t^ϵ) soit confinée dans un petit tube autour de la fonction ϕ.
Fixons maintenant une fonction $\phi : [0,T] \longmapsto \mathbb{R}^d$ et considérons la quantité

$$P\,(X_t^\epsilon \text{ suive } \phi_t \text{ sur } [0,T])\,.$$

Si la fonction ϕ n'est pas une solution du système déterministe (1), cette quantité tend vers 0 avec ϵ. De la nature profonde des perturbations browniennes, il découle que

$$P\,(X_t^\epsilon \text{ suive } \phi_t) \underset{\ln}{\sim} \exp -\frac{V(\phi)}{\epsilon^2} \qquad \text{lorsque } \epsilon \to 0$$

où $\sim_{\ln}$ signifie logarithmiquement équivalent (les logarithmes des deux termes sont des quantités équivalentes) et $V(\phi)$ est une constante associée à la fonction ϕ, qui caractérise la vitesse de convergence vers 0. Plus précisément,

$$V(\phi) \ = \ \int_0^T |\dot{\phi}_t - b(\phi_t)|^2\, dt.$$

Ainsi $V(\phi)$ est nulle si et seulement si ϕ est une solution du système déterministe (1) et est d'autant plus grande que ϕ est "loin" des trajectoires du système déterministe (1). La quantité $V(\phi)$, appelée fonction d'action, mesure le coût pour le système perturbé pour suivre la fonction ϕ.
L'intégrale précédente (4) est donc une intégrale de termes qui tendent vers 0 à des vitesses $\exp -\frac{V(\phi)}{\epsilon^2}$. A ce point, effectuons quelques rappels sur la méthode de Laplace.
Considérons tout d'abord une somme finie d'exponentielles

$$a_1 \exp -\frac{\alpha_1}{\epsilon^2} + \cdots + a_n \exp -\frac{\alpha_n}{\epsilon^2}$$

où les α_i et les a_i sont des coefficients positifs.
Lorsque ϵ tend vers 0, le terme prépondérant est le terme correspondant au plus petit des coefficients α_i, soit $\min \alpha_i$, et la somme est logarithmiquement équivalente à $\exp -\frac{\min \alpha_i}{\epsilon^2}$.
Considérons ensuite le cas d'une intégrale sur la droite réelle

$$\int_{\mathbb{R}} g(x) \exp -\frac{f(x)}{\epsilon^2}\, dx$$

où f et g sont deux fonctions positives. Là encore, seuls les minima de f et son comportement près des minima sont importants. L'intégrale est équivalente à l'intégrale restreinte à un voisinage de ces minima et est logarithmiquement équivalente à $\exp -\frac{\min f}{\epsilon^2}$.
La méthode de Laplace classique donne le développement asymptotique complet de ce type d'intégrales en fonction du seul comportement local de f au voisinage de ses minima.
Dans notre situation (4), nous avons affaire à une intégrale sur une espace fonctionnel de dimension infinie, mais la méthode de Laplace s'applique encore: les termes intégrés se comportent comme $\exp -\frac{V(\phi)}{\epsilon^2}$, et le terme prépondérant est celui pour lequel $V(\phi)$ est minimal. Notons donc

$$V(i,j) = \inf \left\{ V(\phi) : \phi \text{ continue } [0,T] \mapsto \mathbb{R}^d, \phi_0 \in K_i, \phi_T \in K_j \right\}.$$

Nous avons alors

$$P\left(Z_n^\epsilon = j / Z_n^\epsilon = i\right) \underset{\ln}{\sim} \exp -\frac{V(i,j)}{\epsilon^2}.$$

Le coût $V(i,j)$ correspond ainsi à un problème du calcul des variations et mesure la difficulté de passage de l'attracteur K_i à l'attracteur K_j. Un tel passage s'effectuera avec grande probabilité en suivant une fonction ϕ dont le coût $V(\phi)$ est proche de $V(i,j)$.

Nous sommes maintenant confrontés au problème suivant:
Etant données une chaîne de Markov d'espace d'états fini et des estimations de ses probabilités de transition, est–il possible d'en déduire des estimations de sa mesure invariante?
Ici encore, Freidlin et Wentzell ont trouvé les outils adaptés pour traiter cette question. En fait, il existe une formule explicite qui exprime la mesure invariante de la chaîne comme une fonction rationnelle des coefficients de la matrice de transition.

Définition. *(i–graphe)*
Soit i un élément de $\{1\cdots r\}$. Un i–graphe sur $\{1\cdots r\}$ est un graphe tel que, pour tout point j distinct de i, le graphe contient une unique flèche issue de j, et il existe un chemin dans le graphe qui mène de j à i. Le graphe ne contient pas de flèche issue de i.
L'ensemble des i–graphes sur $\{1\cdots r\}$ est noté $G(i)$.

Un i–graphe est donc un graphe sans cycle formé de chemins qui aboutissent en i.
Supposons donnée une matrice de transition irréductible $(p_{ij}, 1\leq i,j\leq r)$ sur $\{1\cdots r\}$.
A chaque graphe g sur $\{1\cdots r\}$, nous associons le réel

$$\pi(g) \;=\; \prod_{(\alpha\to\beta)\in g} p_{\alpha\beta}.$$

Théorème. *(représentation de la mesure invariante)*
La mesure de probabilité invariante ν de la chaîne de Markov d'espace d'états $\{1\cdots r\}$ et de matrice de transition $(p_{ij}, 1\leq i,j\leq r)$ est donnée par

$$\text{(5)}\qquad \forall i\in\{1\cdots r\}\qquad \nu(i) \;=\; \frac{\displaystyle\sum_{g\in G(i)}\pi(g)}{\displaystyle\sum_{j\in\{1\cdots r\}}\;\sum_{g\in G(j)}\pi(g)}.$$

Notons $p^\epsilon_{ij} = P\left(Z^\epsilon_n = j / Z^\epsilon_n = i\right)$. La formule précédente donne l'expression de la mesure invariante ν^ϵ de la chaîne (Z^ϵ_n) en fonction des p^ϵ_{ij}. Nous savons aussi que $p^\epsilon_{ij}\sim_{\ln}\exp -\frac{V(i,j)}{\epsilon^2}$. Le numérateur et le dénominateur de la fraction (5) sont des sommes de produits de termes qui se comportent comme $\exp -\frac{Cte}{\epsilon^2}$. Nous sommes donc une nouvelle fois dans une situation typique de la méthode de Laplace: dans chaque somme, le terme prépondérant est celui qui est associé à la constante la plus petite.
Pour tout graphe g sur $\{1\cdots r\}$, nous avons

$$\pi^\epsilon(g) \;\underset{\ln}{\sim}\; \exp -\frac{\sum_{(\alpha\to\beta)\in g} V(\alpha,\beta)}{\epsilon^2}.$$

Pour le numérateur, la constante minimale est donc

$$W(i) \;=\; \min_{g\in G(i)} \sum_{(\alpha\to\beta)\in g} V(\alpha,\beta)$$

et pour le dénominateur, c'est

$$W(W^*) \;=\; \min\big\{\, W(i) : i\in\{1\cdots r\}\,\big\}.$$

La fonction W est l'énergie virtuelle et W^* est l'ensemble des minima globaux de W.

Finalement,

$$\forall i \in \{1 \cdots r\} \qquad \nu^\epsilon(i) \underset{\ln}{\sim} \exp - \frac{W(i) - W(W^*)}{\epsilon^2}$$

et la question du comportement asymptotique de ν^ϵ (et donc aussi de μ^ϵ) est résolue:
la mesure ν^ϵ se concentre sur les attracteurs dont l'indice est dans W^* et décroît vers 0 à des vitesses $\exp - \frac{\mathrm{Cte}}{\epsilon^2}$ pour les autres attracteurs.
Ainsi, pour tout voisinage $\mathcal{V}(\cup_{i \in W^*} K_i)$ des attracteurs de W^*,

$$\lim_{\epsilon \to 0} \lim_{t \to \infty} P\left(X_t^\epsilon \in \mathcal{V}(\cup_{i \in W^*} K_i) \,/\, X_0^\epsilon = x\right) \;=\; 1.$$

Il existe donc un sous–ensemble W^* d'attracteurs sur lequel se concentre la mesure invariante du processus. Notons que ce sous–ensemble dépend de tout le paysage dynamique du système déterministe. La vue isolée d'un attracteur ne permet pas de conclure quant à son appartenance à W^*. Il existe cependant des résultats généraux sur W^*: par exemple, l'ensemble W^* ne contient que des attracteurs stables, et les attracteurs instables peuvent en quelque sorte être évacués pour l'étude de la dynamique asymptotique. En effet, les petites perturbations browniennes permettent au processus de s'échapper très facilement des attracteurs instables.

La dynamique asymptotique du processus peut être décrite de manière très précise. Essayons d'en donner une image sommaire. Supposons que le processus soit initialement dans le bassin d'attraction de K_1. Il quitte K_1 au bout d'un temps fini. Parmi toutes les trajectoires de sortie, il en existe une plus "probable" que les autres, qui l'amène vers un nouvel attracteur; par exemple K_2. De même, de K_2, il tombe sur K_3. L'ensemble des attracteurs étant par hypothèse fini, le processus finit par revisiter un attracteur. Par exemple, de K_3, il retourne vers K_1. Nous obtenons alors un cycle d'ordre 1

$$K_1 \to K_2 \to K_3 \to K_1$$

sur lequel le processus tourne longtemps, très longtemps. Englobons maintenant ces trois attracteurs dans une boîte. Comme toujours, les perturbations browniennes finissent par pousser le processus hors de cette boîte, et ici encore, il existe une trajectoire de sortie canonique qui fait tomber le processus dans un nouveau bassin d'attraction, ou plus généralement, dans un autre cycle d'ordre 1.
Les cycles d'ordre 1 sont aussi en nombre fini, et le processus finit par revisiter un cycle d'ordre 1: un cycle d'ordre 2 est alors formé, dans lequel le processus reste piégé très très longtemps. En continuant de la sorte, il est possible de construire toute une hiérarchie de cycles qui épuise l'ensemble des attracteurs et fournit une image très précise de la dynamique asymptotique du processus. A chaque transition entre cycles est associée une constante qui caractérise la difficulté de la transition.

Enfin, lorsque ϵ décroît avec le temps, i.e. $\epsilon = \epsilon(t)$ est une fonction de t qui tend en décroissant vers 0, nous obtenons un processus de Markov inhomogène (le mécanisme de transition dépend du temps); nous supprimons l'exposant $\epsilon(t)$ dans les notations.
Si $\epsilon(t)$ décroît très lentement, de sorte qu'à chaque instant la loi de X_t soit proche de l'état d'équilibre associé au niveau de perturbation $\epsilon(t)$, la situation ne change pas fondamentalement. La loi limite correspond à la limite de la suite des lois d'équilibre.
Si au contraire $\epsilon(t)$ décroît très rapidement, le processus risque de rester piégé dans certains sous–ensembles d'attracteurs: plus précisément, dans la hiérarchie des cycles, certaines transitions ne pourront être effectuées qu'un nombre fini de fois, alors que d'autres, plus "faciles", seront réalisées une infinité de fois avec probabilité 1. La loi limite dépend alors fortement du point de départ.
La hiérarchie des cycles permet ainsi de décrire les dynamiques possibles de (X_t) en fonction de la vitesse de décroissance de $\epsilon(t)$.

3. Convergence asymptotique des algorithmes génétiques

Armé de la puissante machinerie mathématique de Freidlin–Wentzell, je suis retourné à l'étude des algorithmes génétiques et j'ai tenté de discerner les mécanismes profonds à l'œuvre dans ces algorithmes.
Existe-t-il un processus de fond qui guide la dynamique de l'algorithme génétique, et celui–ci peut–il être construit en rajoutant à ce processus de petites perturbations aléatoires?
Afin de bien mettre en lumière les éléments cruciaux qui gouvernent la dynamique de l'algorithme, nous proposons dans le premier article un modèle abstrait et général que nous analysons à l'aide de la théorie de Freidlin–Wentzell.

Soit E un ensemble fini et f une fonction définie sur E à valeurs dans $\mathbb{R}_+^*$.
Nous cherchons à localiser l'ensemble f^* des maxima globaux de f.
Nous modélisons l'algorithme génétique par une chaîne de Markov d'espace d'états E^m, où m désigne la taille de la population. Introduisons à ce point quelques notations.
Si $x = (x_1, \cdots, x_m)$ est un élément de E^m et i un point de E, nous notons

$$\begin{gathered}
\widehat{f}(x) = \widehat{f}(x_1, \cdots, x_m) = \max\{\, f(x_k) : 1 \leq k \leq m\,\}, \\
\widehat{x} = \{\, x_k : 1 \leq k \leq m,\, f(x_k) = \widehat{f}(x)\,\}, \\
x(i) = \operatorname{card}\{\, k : 1 \leq k \leq m,\, x_k = i\,\}, \\
[x] = \{\, x_k : 1 \leq k \leq m\,\},
\end{gathered}$$

et $(i) = (i, \cdots, i)$ est la population uniforme dont toutes les composantes sont égales à i.
De manière générale, les lettres x, y, z, u, v désignent des populations, i.e. des éléments de E^m, et les lettres i, j des points de E.

Voici maintenant le processus de fond $(X_n^\infty)_{n\in\mathbb{N}}$ que nous allons perturber aléatoirement pour reconstruire l'algorithme génétique; il modélise la force fondamentale qui guide la dynamique de l'algorithme. L'exposant ∞ exprime qu'il s'agit du processus limite, lorsque les perturbations ont disparu. Ce processus (X_n^∞) est une chaîne de Markov sur E^m dont le mécanisme de transition peut être décrit ainsi: les m composantes de X_{n+1}^∞ sont choisies indépendamment avec la loi uniforme sur l'ensemble $\widehat{X}_n^\infty$. Les individus dont l'adaptation (mesurée par f) n'est pas maximale à l'instant n sont donc détruits à l'instant $n+1$, et les individus d'adaptation maximale ont des chances de survie égales. Avec une formule, ceci donne

$$P\left(X_{n+1}^\infty = z / X_n^\infty = y\right) \;=\; \prod_{k=1}^{m} \frac{1_{\widehat{y}}(z_k) y(z_k)}{\operatorname{card}\widehat{y}}.$$

Cette chaîne est instantanément piégée dans l'ensemble des populations d'équi–adaptation

$$S \;=\; \{\, x \in E^m : f(x_1) = \cdots = f(x_m)\,\}$$

puis, au bout d'un certain temps, elle est absorbée par une population uniforme:

$$\forall x \in E^m \qquad P\left(\exists i \in \widehat{x} \quad \exists N \quad \forall n \geq N \quad X_n^\infty = (i) / X_0^\infty = x\right) \;=\; 1.$$

En effet, le mécanisme de sélection ne préserve pas la diversité et certains individus disparaissent par hasard. Lorsque la population est devenue uniforme, il ne se passe plus rien. Les populations uniformes sont donc exactement les états absorbants de la chaîne (X_n^∞). Les populations d'équi–adaptation sont les attracteurs de la chaîne (X_n^∞) et les populations uniformes sont les attracteurs stables.
La chaîne (X_n^∞) est ensuite perturbée aléatoirement. L'intensité des perturbations est contrôlée par un paramètre l et nous obtenons une suite de chaînes de Markov (X_n^l).
Le mécanisme de transition comporte trois étapes:

$$X_n^l \xrightarrow{\text{mutation}} U_n^l \xrightarrow{\text{crossing–over}} V_n^l \xrightarrow{\text{sélection}} X_{n+1}^l.$$

Décrivons maintenant chacune de ces étapes.

$X_n^l \longrightarrow U_n^l$**: mutation.** Les mutations sont de petites perturbations aléatoires indépendantes des individus de la population X_n^l. La probabilité pour que le point i de E se transforme par mutation en j est donnée par $p_l(i,j)$ où p_l est un noyau markovien sur E, i.e. p_l est à valeurs dans $[0,1]$ et

$$\forall i \in E \qquad \sum_{j \in E} p_l(i,j) \;=\; 1.$$

Ainsi

$$P\left(U_n^l = u / X_n^l = x\right) \;=\; p_l(x_1, u_1) \cdots p_l(x_m, u_m)$$

et nous imposons la convergence de $p_l(i,j)$ vers la matrice identité

(6) $$\forall i,j \in E \qquad \lim_{l \to \infty} p_l(i,j) \;=\; \delta(i,j)$$

(où δ est le symbole de Kronecker)
de sorte que les mutations disparaissent lorsque l tend vers l'infini.

$U_n^l \longrightarrow V_n^l$**: crossing–over.** Les crossing–over sont modélisés comme des petites perturbations aléatoires des couples formés par les individus consécutifs de la population U_n^l. Ainsi

$$P\left(V_n^l = v / U_n^l = u\right) \;=\; q_l((u_1,u_2),(v_1,v_2))\, q_l((u_3,u_4),(v_3,v_4)) \cdots$$

où q_l est un noyau markovien sur $E \times E$, et nous imposons

$$\text{(7)} \qquad \forall i_1,i_2,j_1,j_2 \in E \qquad \lim_{l\to\infty} q_l((i_1,i_2),(j_1,j_2)) \;=\; \delta(i_1,j_1)\,\delta(i_2,j_2)$$

de sorte que les crossing–over disparaissent lorsque l tend vers l'infini.

$V_n^l \longrightarrow X_{n+1}^l$**: sélection.** Les individus de la population X_{n+1}^l sont sélectionnés indépendamment à l'aide d'une distribution de probabilité sur la population V_n^l définie par une fonction de sélection.

Définition. *Une fonction de sélection d'ordre m est une fonction F définie sur l'ensemble $\{1\cdots m\}\times(\mathbb{R}_+^*)^m$ à valeurs dans $[0,1]$ telle que, pour tout m–uplet $(f_1,\cdots,f_m)$ de $(\mathbb{R}_+^*)^m$:*

a) $F(1,f_1,\cdots,f_m)+F(2,f_1,\cdots,f_m)+\cdots+F(m,f_1,\cdots,f_m) \;=\; 1,$

b) $\forall \sigma \in \mathfrak{S}_m \quad \forall k \in \{1\cdots m\} \qquad F(\sigma(k), f_{\sigma(1)},\cdots,f_{\sigma(m)}) = F(k,f_1,\cdots,f_m),$

c) $f_1 \geq f_2 \geq \cdots \geq f_m \Rightarrow F(1,f_1,\cdots,f_m) \geq F(2,f_1,\cdots,f_m) \geq \cdots \geq F(m,f_1,\cdots,f_m).$

La condition *a*) dit que $F(\cdot,f_1,\cdots,f_m)$ est une probabilité sur l'ensemble $\{1\cdots m\}$.
La condition *b*) dit que cette probabilité est indépendante de l'indexation de $f_1,\cdots,f_m$.
La condition *c*) dit que cette probabilité favorise les valeurs élevées.
Soit donc F_l une fonction de sélection (dépendant de l). Alors

$$P\left(X_{n+1}^l = x / V_n^l = v\right) \;=\; \prod_{r=1}^{m} \sum_{k: v_k = x_r} F_l\big(k, f(v_1),\cdots,f(v_m)\big).$$

Le mécanisme de sélection de la chaîne (X_n^∞) correspond à la fonction de sélection

$$F_\infty(k,f(x_1),\cdots,f(x_m)) \;=\; \frac{1_{\widehat{x}}(x_k)}{\operatorname{card}\widehat{x}}$$

(i.e. c'est la loi uniforme sur $\widehat{x}$).
Nous imposons la convergence de la suite F_l vers F_∞:

$$\text{(8)} \qquad \forall x \quad \forall k \qquad \lim_{l\to\infty} F_l(k,f(x_1),\cdots,f(x_m)) \;=\; F_\infty(k,f(x_1),\cdots,f(x_m)).$$

Les conditions (6), (7) et (8) entraînent la convergence, lorsque l tend vers l'infini, du mécanisme de transition de la chaîne (X_n^l) vers le mécanisme de transition de la chaîne (X_n^∞):

$$\forall y, z \in E^m \qquad \lim_{l\to\infty} P\left(X_{n+1}^l = z / X_n^l = y\right) \;=\; P\left(X_{n+1}^\infty = z / X_n^\infty = y\right)$$

et en ce sens, la chaîne (X_n^l) est une perturbation de la chaîne (X_n^∞).

Afin d'analyser la dynamique asymptotique de (X_n^l) lorsque l tend vers l'infini, nous avons besoin de davantage d'information sur le mode de convergence des probabilités de transition. Nous faisons les hypothèses de développement asymptotique suivantes:
Pour le noyau p_l: il existe un réel positif a et un noyau irréductible α sur E, i.e.

$$\forall i, j \in E \quad \exists i_0, i_1, \cdots, i_r \quad i_0 = i, \quad i_r = j, \qquad \prod_{0\le k\le r-1} \alpha(i_k, i_{k+1}) > 0$$

et p_l admet le développement

$$\forall i, j \in E \quad \forall s > 0 \qquad p_l(i,j) = \begin{cases} \alpha(i,j)\, l^{-a} + o(l^{-s}) & \text{si } i \neq j \\ 1 - \alpha(i,j)\, l^{-a} + o(l^{-s}) & \text{si } i = j \end{cases}$$

Pour le noyau q_l: il existe un réel positif b et un noyau β sur $E \times E$ tels que

$$\forall (i_1, j_1) \in E \times E \quad \forall (i_2, j_2) \in E \times E \quad \forall s > 0$$
$$q_l\left((i_1,j_1),(i_2,j_2)\right) = \begin{cases} \beta\left((i_1,j_1),(i_2,j_2)\right)\, l^{-b} + o(l^{-s}) & \text{si } (i_1,j_1) \neq (i_2,j_2) \\ 1 - \beta\left((i_1,j_1),(i_2,j_2)\right)\, l^{-b} + o(l^{-s}) & \text{si } (i_1,j_1) = (i_2,j_2) \end{cases}$$

Pour la fonction de sélection F_l, nous choisissons

$$F_l(k, f_1, \cdots, f_m) \;=\; \frac{\exp(c f_k \ln l)}{\sum_{r=1}^m \exp(c f_r \ln l)}$$

où c est un réel positif.
Le point crucial est de bien doser les vitesses de convergence dans (6), (7) et (8), afin que chaque opérateur intervienne dans la dynamique asymptotique. Plus précisément, il faut que les vitesses de convergence soient logarithmiquement du même ordre (ce qui autorise l'ajout des paramètres de contrôle a, b, c). L'hypothèse d'irréductibilité sur le noyau α est essentielle: elle assure que les petites perturbations des individus permettent asymptotiquement de visiter tout l'espace. Lorsque l croît vers l'infini, les perturbations s'évanouissent et avec grande probabilité, la chaîne (X_n^l) se comporte comme le ferait (X_n^∞): elle est tout d'abord attirée par l'ensemble S des populations d'équi–adaptation.
Nous utilisons la théorie de Freidlin et Wentzell pour analyser la dynamique asymptotique

de la chaîne (X_n^l) lorsque les perturbations disparaissent. Soit T_n l'instant de la nième visite dans l'ensemble S. Posons $Z_n^l = X_{T_n}^l$. La chaîne (Z_n^l) est la chaîne de Markov induite par la chaîne (X_n^l) sur l'ensemble S. L'étude de cette chaîne des visites successives des attracteurs est "équivalente" à l'étude de tout le processus. Les attracteurs stables sont les populations uniformes. Nous commençons par estimer les probabilités de transition de la chaîne (Z_n^l). Soit, pour deux populations y et z, $\mathcal{D}(y,z)$ l'ensemble des chemins $\mathcal{P}$ dans E^m

$$x^0 \to u^0 \to v^0 \to x^1 \to u^1 \to \cdots \to x^{r-1} \to u^{r-1} \to v^{r-1} \to x^r$$

qui correspondent aux trajectoires possibles de tout le processus

$$X_n^l \to U_n^l \to V_n^l \to X_{n+1}^l \to U_{n+1}^l \to \cdots \to X_{n+r-1}^l \to U_{n+r-1}^l \to V_{n+r-1}^l \to X_{n+r}^l$$

et qui joignent y et z: $x^0 = y, x^r = z$ (la longueur r est laissée variable), et soit $\widetilde{\mathcal{D}}(y,z)$ l'ensemble des chemins de $\mathcal{D}(y,z)$ qui restent confinés hors de S, i.e. la population x^k n'est pas dans S pour $k \in \{1 \cdots r-1\}$.
Le coût d'un chemin $\mathcal{P}$ de $\mathcal{D}(y,z)$ est défini par

$$V(\mathcal{P}) = \sum_{k=0}^{r-1} ad(x^k, u^k) + b\bar{d}(u^k, v^k) + c\Big(m\widehat{f}(v^k) - \sum_{h=1}^{m} f(x_h^{k+1})\Big)$$

où

$$d(x,u) = \operatorname{card}\{\, k : 1 \le k \le m, x_k \neq u_k \,\},$$
$$\bar{d}(u,v) = \operatorname{card}\{\, k : 1 \le k \le m/2, \quad (u_{2k-1}, u_{2k}) \neq (v_{2k-1}, v_{2k}) \,\}$$

et x_h^{k+1} est la hième composante du vecteur x^{k+1}. La quantité $V(\mathcal{P})$ mesure la difficulté pour le processus perturbé pour suivre le chemin $\mathcal{P}$ et est l'analogue de la fonction d'action

$$V(\phi) = \int_0^T |\dot{\phi}_t - b(\phi_t)|^2 \, dt.$$

Ici, $d(x^k, u^k)$ est le nombre de mutations nécessaires pour passer de x^k à u^k, $\bar{d}(u^k, v^k)$ est le nombre de crossing–over nécessaires pour passer de u^k à v^k et $\sum_{h=1}^m \widehat{f}(v^k) - f(x_h^{k+1})$ est le coût dû à la sélection dans v^k d'individus d'adaptation non maximale (cette somme s'annule si et seulement si $[x^{k+1}] \subset \widehat{v}^k$). Nous définissons alors deux fonctions de coût de communication V et $\widetilde{V}$ sur $E^m \times E^m$ par

$$V(y,z) = \inf\{\, V(\mathcal{P}) : \mathcal{P} \in D(y,z) \,\}, \qquad \widetilde{V}(y,z) = \inf\{\, V(\mathcal{P}) : \mathcal{P} \in \widetilde{D}(y,z) \,\}.$$

Nous obtenons les estimations, pour y et z dans S,

$$P\left(Z_{n+1}^l = z / Z_n^l = y\right) \underset{l\to\infty}{\sim} \widetilde{C}(y,z)\, l^{-\widetilde{V}(y,z)}$$

où $\widetilde{C}(y,z)$ est un réel positif. De ceci, nous déduisons à l'aide de la formule de représentation de Freidlin–Wentzell des estimations de la mesure invariante ν^l de la chaîne (Z_n^l). La quantité cruciale est l'énergie virtuelle W, définie pour x dans S par

$$W(x) = \min_{g \in G(x)} \sum_{(y \to z) \in g} \widetilde{V}(y,z) = \min_{g \in G(x)} \sum_{(y \to z) \in g} V(y,z)$$

où $G(x)$ est l'ensemble des x–graphes sur S.
Les populations non uniformes de S étant instables, il est suffisant (d'après un résultat de Freidlin et Wentzell) d'étudier W en se restreignant aux populations uniformes. Nous confondons de plus la population uniforme (i) et le point i de E. Nous étudions donc W comme fonction de E dans $\mathbb{R}^+$, avec

$$W(i) = \min_{g \in G(i)} \sum_{(i_1 \to i_2) \in g} V(i_1, i_2)$$

où nous nous restreignons aux i–graphes sur l'ensemble E.
L'ensemble W^* des minima de W

$$W^* \;=\; \{\, i \in E : W(i) = \min_{j \in E} W(j) \,\}$$

est aussi considéré comme une partie de E. C'est sur cet ensemble que se concentre la suite des mesures stationnaires des chaînes (X_n^l) lorsque l tend vers l'infini:

$$\forall x \in E^m \qquad \lim_{l \to \infty} \quad \lim_{n \to \infty} \quad P\left(X_n^l \in W^* / X_0^l = x\right) = 1.$$

Le résultat fondamental que nous démontrons s'énonce ainsi:

Théorème. *(existence d'une taille de population critique)*
Supposons fixés l'espace E, la fonction d'adaptation f, les noyaux de mutation α et de crossing–over β, et les constantes positives a, b, c.
Il existe une valeur critique m^* qui dépend de tous ces éléments telle que, lorsque la taille m de la population de l'algorithme est plus grande que m^*, l'ensemble W^* des minima de l'énergie virtuelle W est inclus dans l'ensemble f^* des maxima globaux de la fonction f.

Esquissons brièvement le schéma de la preuve.
Soit i un point de E n'appartenant pas à f^*. Soit j un point de E tel que $f(j) > f(i)$ et qu'il existe une séquence de mutations $e_1 \to \cdots \to e_r$ qui transforme i en j en restant en–dessous de $f(j)$:

$$\forall k \in \{1 \cdots r\} \qquad f(e_k) \le f(j).$$

Un tel point existe forcément: le noyau de mutation est irréductible et finit par créer à partir de i des points d'adaptation supérieure à $f(i)$.

Le coût $V(i,j)$ de la transition entre les populations uniformes (i) et (j) est alors une fonction bornée de m. Plus précisément, cette transition s'opère ainsi:
Toute la population est initialement concentrée en i. Un individu part tout seul et suit la séquence de mutations $e_1 \to \cdots \to e_r$. Dès que cet individu atteint un niveau d'adaptation supérieur à $f(i)$ (ce qui arrive fatalement, vu que $f(e_r) = f(j) > f(i)$), tout le reste de la population, qui attendait sagement en i, se précipite massivement sur le point découvert par notre explorateur. La trajectoire continue de la sorte jusqu'en (j). Le coût de ce chemin correspond aux seuls coûts (mutations et anti–sélections) occasionnés par l'individu explorateur, vu que tous les autres individus ne font que suivre le processus de fond (X_n^∞). Ce coût est donc indépendant de la taille m de la population. Le résultat est encore valide si l'inégalité est large: $j \neq i, f(j) \geq f(i)$.
Considérons maintenant deux points i, j de E tels que $f(i) > f(j)$. Nous avons au contraire

$$\liminf_{m\to\infty} \frac{V(i,j)}{m} \; > \; 0$$

d'où en particulier $\lim_{m\to\infty} V(i,j) = \infty$. En effet, le passage de la population uniforme (i) à la population uniforme (j) nécessite de réaliser un événement de coût positif sur chaque individu (mutation ou anti–sélection). La tendance naturelle des individus est de rester au point i: d'eux–mêmes, ils ne sont pas attirés vers les niveaux d'adaptation inférieurs.

Soit maintenant i un point de f^*.
Pour tout point j distinct de i, il existe un chemin $e_1 \to \cdots \to e_r$ joignant j à i (i.e. $e_1 = j, e_r = i$) tel que $f(e_1) \leq \cdots \leq f(e_r)$ et pour tout k dans $\{1 \cdots r-1\}$, le point e_{k+1} peut être créé à partir du point e_k par une séquence de mutations qui reste en–dessous du niveau d'adaptation $f(e_{k+1})$. Il s'ensuit que $V(e_1, e_2) + \cdots + V(e_{r-1}, e_r)$ est une fonction bornée de m. En réunissant de tels chemins, il est possible de construire un i–graphe dont toutes les transitions ont des coûts bornés, d'où

$$(9) \qquad \sup_{m\in\mathbb{N}^*} \min_{e\in f^*} W(e) \; \leq \; \sup_{m\in\mathbb{N}^*} W(i) \; < \; \infty.$$

Soit j un point de E n'appartenant pas à f^*.
Tout j–graphe contient nécessairement une transition $e_1 \to e_2$ correspondant à une décroissance d'adaptation: $f(e_1) > f(e_2)$. Il suffit par exemple de considérer le chemin qui mène d'un point de f^* à j. Le coût de cette transition croît linéairement avec m, de sorte que $\lim_{m\to\infty} W(j) = \infty$ et ainsi

$$(10) \qquad \lim_{m\to\infty} \min_{j\in E\setminus f^*} W(j) \; = \; \infty.$$

Les résultats (9) et (10) entraînent que, pour m suffisamment grand, les minima W^* de W sont inclus dans f^*.

Nous donnons plusieurs conditions suffisantes pour avoir $W^* \subset f^*$, ainsi que des majorations de la taille critique m^*. Ces conditions sont naturellement fortement liées au problème d'optimisation. Nous les affaiblissons progressivement afin de les rendre plus "lisibles". La borne la plus grossière de m^* est

$$m^* \;\leq\; \frac{aR + c(R-1)\Delta}{\min(a, b/2, c\delta)}$$

où $\delta = \min\{\, |f(i) - f(j)| : i, j \in E,\ f(i) \neq f(j)\,\}$, $\Delta = \max\{\, |f(i) - f(j)| : i, j \in E\,\}$, et R est le nombre minimal de transitions permettant de joindre deux points arbitraires de E par une séquence de mutations, i.e. c'est le plus petit entier tel que

$$\forall i, j \in E \quad \exists\, r \leq R \quad \exists e_1, \cdots, e_{r+1} \in E \quad \text{tels que}$$
$$e_1 = i,\ e_{r+1} = j, \quad \forall k \in \{1, \cdots, r\} \quad \alpha(e_k, e_{k+1}) > 0.$$

La difficulté du problème est liée au coût minimal nécessaire pour, partant d'un point quelconque, arriver à créer un point d'adaptation supérieure: ce coût correspond au trajet de l'individu "explorateur". Notons enfin que notre résultat est obtenu sans faire appel à l'opérateur de crossing–over, qui n'est donc pas fondamental pour assurer la convergence vers f^*: l'interaction asymptotique entre les mutations et la force de sélection suffit.

L'énoncé de notre théorème appelle naturellement la question suivante: est–il vraiment nécessaire de dépasser la taille critique m^* pour avoir $W^* \subset f^*$?
Il est clair que l'algorithme ne peut pas fonctionner avec un seul individu: l'opérateur de sélection ne fait plus rien et la fonction f n'intervient plus dans le mécanisme de transition. Mais peut–être deux individus suffisent–ils pour résoudre tous les problèmes?
La réponse est non: nous allons maintenant exhiber des situations où la taille critique m^* est arbitrairement grande.

Considérons l'espace $E = \{\, 0, 1, \cdots, N\,\}$ et soit f une fonction de E dans $\mathbb{R}_+^*$ telle que

$$f(1) = f(2) = \cdots = f(N-1), \quad f(0) = f(1) + \gamma, \quad f(N) = f(N-1) + \Gamma,$$

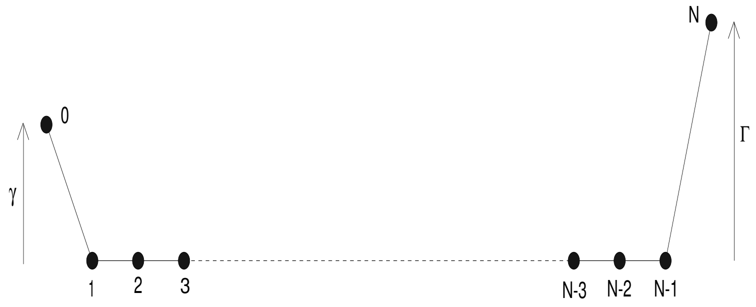
où $0 < \gamma < \Gamma$. Soit α un noyau markovien de mutation tel que

$$\forall i \in \{1 \cdots N-1\} \qquad \alpha(i, i-1) + \alpha(i, i+1) + \alpha(i, i) = 1, \quad \alpha(i, i-1) > 0, \quad \alpha(i, i+1) > 0,$$

(les mutations possibles à partir de i sont $i \to i-1$ et $i \to i+1$), et pour les points 0 et N,

$$\alpha(0,1) + \alpha(0,0) = 1, \quad \alpha(0,1) > 0, \quad \alpha(N, N-1) + \alpha(N, N) = 1, \quad \alpha(N, N-1) > 0.$$

D'où le schéma



Nous avons
$$\forall i \in \{1 \cdots N-1\} \qquad V(i,i-1) = V(i,i+1) = a,$$
(ces transitions nécessitent juste la mutation d'un individu: tout le reste de la population vient alors sur le nouveau point)
$$V(0,1) = \min\big(am, a + cm\gamma\big), \quad V(N,N-1) = \min\big(am, a+cm\Gamma\big).$$
(le passage de 0 en 1, par exemple, s'effectue soit par une mutation générale, de coût am, soit par une anti–sélection générale après qu'un individu ait muté en 1, de coût $a + cm\gamma$). Enfin
$$V(0,N) = \min\big(am + a(N-1), a + cm\gamma + a(N-1), aN + c(N-1)\gamma\big).$$
Il y a deux trajectoires candidates pour passer de 0 en N: soit toute la population passe de 0 en 1 (coût $V(0,1)$) puis de 1 en N (coût $a(N-1)$), soit un explorateur part tout seul de 0 et s'en va jusqu'en N, tandis que les autres individus attendent en 0 (coût $aN + c(N-1)\gamma$): lorsque l'explorateur arrive en N, toute la population se précipite en N vu que $\Gamma > \gamma$.
De même
$$V(N,0) = \min\big(am + a(N-1), a + cm\Gamma + a(N-1), aN + c(N-1)\Gamma + cm(\Gamma-\gamma)\big).$$
Notons toutefois une différence cruciale pour le passage de N vers 0: avec le second type de trajectoire, lorsque l'explorateur arrive en 0, il est nécessaire de faire une anti–sélection générale de coût $cm(\Gamma - \gamma)$ pour amener tout le monde en 0.
Nous vérifions en particulier sur cet exemple que
$$\lim_{m\to\infty} V(0,N) \;=\; aN + c(N-1)\gamma \;<\; \infty$$
et
$$\lim_{m\to\infty} \frac{V(N,0)}{m} \;=\; \min\big(a, c(\Gamma-\gamma)\big) \;>\; 0.$$

Evaluons maintenant l'énergie virtuelle W.
Pour i dans $\{1 \cdots N-1\}$, le i–graphe de coût minimal est

$$0 \to 1 \cdots \to i-1 \to i \leftarrow i+1 \leftarrow \cdots \leftarrow N-1 \leftarrow N$$

et nous avons

$$W(i) \;=\; a(N-2) \;+\; V(0,1) \;+\; V(N,N-1).$$

Pour 0, il y a deux 0–graphes candidats:

$$N \to N-1 \to \cdots \to 1 \to 0 \quad \text{et} \quad N-1 \to \cdots \to 1 \to 0 \leftarrow N$$

d'où

$$\begin{aligned} W(0) \;&=\; a(N-1) \;+\; \min\big(V(N,N-1), V(N,0)\big) \\ &=\; a(N-1) \;+\; \min\big(am, a+cm\Gamma, aN + c(N-1)\Gamma + cm(\Gamma-\gamma)\big). \end{aligned}$$

De même, nous avons

$$W(N) \;=\; a(N-1) \;+\; \min\big(am, a+cm\gamma, aN + c(N-1)\gamma\big).$$

Comme $N > 1$ et que $V(0,1) > a$, $V(N,N-1) > a$, pour tout i dans $\{1 \cdots N-1\}$, nous avons $W(0) < W(i)$, $W(N) < W(i)$, et donc $W^* \subset \{0, N\}$. Supposons que

$$am \leq a + cm\gamma \quad \text{et} \quad am \leq aN + c(N-1)\gamma.$$

Pour réaliser ces conditions, il suffit par exemple d'avoir $a \leq c\gamma$ et $m \leq N$.
Dans ce cas,

$$W(0) = W(N) = a(N-1) + am$$

et l'ensemble $W^* = \{0, N\}$ n'est pas inclus dans $f^* = \{N\}$.
Finalement, lorsque $a \leq c\gamma$, nous avons $m^* > N$.
Ce phénomène peut s'expliquer ainsi: dans les conditions ci–dessus, les départs des bassins d'attraction de 0 et de N s'effectuent par une mutation générale, de coût am, qui amène toute la population en 1 et en $N-1$. Les hauteurs γ et Γ n'interviennent donc pas et la dynamique ne permet pas de discriminer les points 0 et N, si bien que la loi limite se concentre sur les deux points $\{0, N\}$. Si le noyau de mutation α n'est pas symétrique, il est en outre possible de construire des exemples où $f^* \cap W^* = \emptyset$ lorsque $m < m^*$.

Dès lors que $m \geq m^*$, nous savons que la suite des mesures stationnaires se concentre sur f^* lorsque l tend vers l'infini. Afin d'obtenir un algorithme réalisable en pratique et implémentable sur ordinateur, nous faisons croître l avec n et nous obtenons une chaîne de Markov inhomogène $X_n^{l(n)}$ dont le mécanisme de transition dépend de n. L'exposant $l(n)$

est désormais supprimé dans les notations. Le problème est d'adapter soigneusement la vitesse de croissance de $l(n)$ de sorte que la loi de X_n converge vers la limite de la suite des mesures stationnaires des chaînes (X_n^l). Ce sera le cas si $l(n)$ croît très lentement, car alors la loi de X_n sera à chaque instant proche de la loi d'équilibre associée au niveau de perturbation $l(n)$. Si au contraire $l(n)$ croît très vite, X_n risque de rester piégée dans certains sous–ensembles de populations, car l'intensité des perturbations ne permet alors plus de s'échapper de tous les bassins d'attraction. L'idéal est d'ajuster $l(n)$ de sorte que X_n s'échappe des bassins d'attraction de tous les points qui ne sont pas des maxima globaux et soit piégé dans le bassin d'attraction des points de f^*: la population X_n contient alors définitivement des points de f^* après un nombre fini de transitions.
Nous imposons une condition de nature technique sur la suite $l(n)$ afin de faciliter l'étude. La plupart de nos conditions portent sur l'exposant de convergence de $l(n)$, i.e. l'unique réel λ tel que la série

$$l(1)^{-\theta} + \cdots + l(n)^{-\theta} + \cdots$$

converge pour $\theta > \lambda$ et diverge pour $\theta < \lambda$.

Théorème. *(condition nécessaire pour coloniser f^*)*
Pour que la chaîne $Z_n = X_{T_n}$ des visites des attracteurs soit piégée dans f^* après un nombre fini de transitions, i.e. pour avoir

$$\forall x \in E^m \qquad P\left(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset f^*/X_0 = x\right) = 1,$$

il est nécessaire que l'exposant de convergence λ de la suite $l(n)$ appartienne à un intervalle critique $]\phi, \psi[$ (où ϕ et ψ sont deux constantes caractéristiques du problème d'optimisation), i.e. il faut que

$$\sum_{n\geq 0} l(n)^{-\phi} = \infty \qquad et \qquad \sum_{n\geq 0} l(n)^{-\psi} < \infty.$$

L'intervalle $]\phi, \psi[$ est non vide pour m assez grand et il existe alors toujours des suites de la forme n^θ qui conviennent.

Théorème. *(condition suffisante pour coloniser f^*)*
Il existe deux constantes θ_1, θ_2 telles que, si l'exposant de convergence λ de la suite $l(n)$ est dans l'intervalle $]\theta_1, \theta_2[$, i.e. si

$$\sum_{n\geq 0} l(n)^{-\theta_1} = \infty \qquad et \qquad \sum_{n\geq 0} l(n)^{-\theta_2} < \infty$$

alors, avec probabilité 1, après un nombre fini de transitions, la chaîne (X_{T_n}) est piégée dans f^* et la population (X_n) contient toujours des individus appartenant à f^*:

$$\forall x \in E^m \qquad P\Big(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset f^*,\, \widehat{X}_n \subset f^*/X_0 = x\Big) = 1.$$

Pour terminer, nous montrons comment notre modèle général s'adapte au cas de l'espace $E = \{0, 1\}^N$ et nous retraduisons dans ce cadre les principaux résultats.

4. La dynamique des algorithmes de mutation–sélection avec de grandes tailles de population

Un résultat important du premier article est que, contrairement à une opinion très répandue dans la littérature des algorithmes génétiques, l'opérateur de crossing–over n'est pas du tout indispensable pour assurer la convergence de l'algorithme vers les maxima globaux de la fonction d'adaptation.

Dans le second article, nous reprenons le modèle étudié dans le premier article mais nous supprimons le crossing–over. Nous étudions donc une chaîne de Markov (X_n^l) construite à l'aide des opérateurs de mutation et de sélection:

$$X_n^l \quad \xrightarrow{\text{mutation}} \quad Y_n^l \quad \xrightarrow{\text{sélection}} \quad X_{n+1}^l.$$

$X_n^l \longrightarrow Y_n^l$**: mutation.** Soit α un noyau markovien irréductible sur E. Posons

$$\alpha_l(i,j) = \begin{cases} \alpha(i,j)\, l^{-a} & \text{si } i \neq j \\ 1 - \sum_{e \neq i} \alpha(i,e)\, l^{-a} & \text{si } i = j \end{cases}$$

La matrice des transitions de X_n^l à Y_n^l est définie par

$$\forall x, y \in E^m \qquad P\left(Y_n^l = y / X_n^l = x\right) = \alpha_l(x_1, y_1) \cdots \alpha_l(x_m, y_m).$$

$Y_n^l \longrightarrow X_{n+1}^l$**: sélection.** Nous utilisons le même mécanisme de sélection que dans le modèle précédent:

$$P\left(X_{n+1}^l = x / Y_n^l = y\right) = \prod_{r=1}^{m} \frac{y(x_r) \exp(cf(x_r) \ln l)}{\sum_{k=1}^m \exp(cf(y_k) \ln l)}$$

et le processus (X_n^l) est toujours une perturbation du processus (X_n^∞), au sens que

$$\forall y, z \in E^m \qquad \lim_{l \to \infty} P\left(X_{n+1}^l = z / X_n^l = y\right) = P\left(X_{n+1}^\infty = z / X_n^\infty = y\right).$$

Nous effectuons le développement des probabilités de transition lorsque l tend vers l'infini:

$$P\left(X_{n+1}^l = z / X_n^l = y\right) \underset{l \to \infty}{\sim} q_1(y,z) \exp\left(-V_1(y,z) \ln l\right)$$

où $q_1(y,z)$ est une constante positive ou nulle et $V_1(y,z)$ est la fonction de coût de communication définie par

$$V_1(y,z) = \min \Big\{ ad(y,x) + c \sum_{k=1}^{m} (\widehat{f}(x) - f(z_k)) : \\ x \text{ tel que } P\left(X_{n+1}^l = z, Y_n^l = x / X_n^l = y\right) > 0 \Big\}.$$

Nous sommes alors dans le cadre du recuit simulé généralisé étudié par Trouvé: les probabilités de transition du processus (X_n^l) forment une famille de noyaux markoviens sur E^m indexée par l qui est admissible pour le noyau de communication q_1 et la fonction de coût V_1 [10, Définition 3.1]. L'analogue de la température du recuit est $1/\ln l$. Tous les outils développés par Catoni et Trouvé [1,2,10,11,12] sont alors à notre disposition.
L'énergie virtuelle W_1 associée au coût de communication V_1 est

$$\forall x \in E^m \qquad W_1(x) = \min \left\{ \sum_{(x_1 \to x_2) \in g} V_1(x_1, x_2) : g \in G(x) \right\}$$

où $G(x)$ désigne l'ensemble des x–graphes sur E^m.
Soit W_1^* l'ensemble des minima globaux de W_1. La probabilité se concentre sur W_1^* lorsque l tend vers l'infini:

Proposition. *(Freidlin et Wentzell)*

$$\forall x \in E^m \qquad \lim_{l \to \infty} \quad \lim_{n \to \infty} \quad P\left(X_n^l \in W_1^* / X_0^l = x\right) \;=\; 1.$$

Nous confondons la population uniforme (i) avec le point i de E. L'ensemble W_1^* est inclus dans l'ensemble des attracteurs stables du processus (X_n^∞), i.e. les populations uniformes, et peut donc être considéré comme une partie de E. Le résultat essentiel du premier article se formule ainsi:

Théorème. *(existence d'une taille de population critique)*
Il existe une taille critique de population m^* qui dépend de l'espace E, de la fonction f, du noyau de mutation α et des paramètres a, c, telle que

$$m \geq m^* \qquad \Longrightarrow \qquad W_1^* \subset f^*.$$

La quantité fondamentale pour décrire la dynamique de l'algorithme est l'altitude de communication A_1 introduite par Trouvé et définie par

$$A_1(x,y) = \inf \left\{ \max_{1 \leq k < r} W_1(p_k) + V_1(p_k, p_{k+1}) : p_1 = x, p_r = y \right\}.$$

Dans le cas du recuit simulé séquentiel classique, où $V(i,j) = \big(U(j) - U(i)\big)^+$, l'altitude de communication est la hauteur d'énergie minimale à franchir pour joindre x et y.
L'altitude de communication permet de construire la hiérarchie des cycles et de définir les hauteurs de sortie et de mélange (H_e, H_m) de sous–ensembles quelconques.
Nous commençons par remplacer la fonction V_1 par la fonction V définie par

$$V(x,y) \;=\; \min \left\{ V_1(p_1, p_2) + \cdots + V_1(p_{r-1}, p_r) : p_1 = x, p_r = y \right\}$$

et nous prouvons un résultat valide dans le cadre général du recuit simulé généralisé: l'énergie virtuelle W et l'altitude de communication A associées à V coïncident avec W_1 et A_1.

Notre objectif est d'étudier la dynamique de l'algorithme lorsque m devient grand: une difficulté majeure est que la taille de l'espace d'états E^m croît géométriquement avec m. Toute l'étude repose sur le résultat suivant (également valide dans le cadre du recuit simulé généralisé).

Théorème. *Soit H une partie de E^m telle que*

$$\forall x \in E^m \quad \exists y \in H \quad V(x,y) = 0.$$

Définissons un coût de communication V_H sur $H \times H$ par

$$V_H(x,y) \;=\; \inf \left\{\, V_1(p_1,p_2) + \cdots + V_1(p_{r-1},p_r) : p_1 = x, p_r = y, p_k \notin H, 1 < k < r \,\right\}$$

et soient A_H et W_H l'altitude de communication et l'énergie virtuelle sur H associées à V_H. Alors $W = W_H$ et $A = A_H$.

Nous appliquons ce résultat avec pour H l'ensemble U des populations uniformes; cet ensemble U est en bijection avec E et son cardinal ne varie pas avec m.
Nous montrons ensuite qu'il est équivalent de construire toute la hiérarchie des cycles sur E^m, puis de la projeter sur U (en ne conservant dans chaque cycle que les populations uniformes), ou de construire directement la hiérarchie des cycles sur U à l'aide de A_U. En quelque sorte, les opérations "construction des cycles" et "projection sur U" commutent. Nous nous ramenons donc à l'étude de la restriction de la dynamique à l'ensemble des attracteurs stables U. Nous examinons ensuite les trajectoires de coût minimal entre deux populations uniformes. Lorsque la taille de la population m est grande, leur structure peut être décrite ainsi: un petit groupe d'individus se "sacrifie" pour ouvrir un chemin idéal de coût minimal qui est ensuite massivement suivi par le reste de la population. Quantitativement, ce phénomème se traduit par le

Théorème. *(linéarité du coût de communication)*
Soient i,j deux points de E. Pour m assez grand, $V(i,j)$ est une fonction affine de m:

$$\exists M \in \mathbb{N} \quad \forall m \geq M \qquad V(i,j) \;=\; \theta(i,j) \;+\; m\,\Omega(i,j).$$

Le coefficient $\theta(i,j)$ correspond au coût des trajectoires du petit groupe d'individus et $\Omega(i,j)$ est le coût du chemin idéal. Nous donnons différentes propriétés des coefficients $\theta(i,j)$ et $\Omega(i,j)$. Nous décrivons le comportement de Ω dans deux cas extrêmes, lorsque a/c est très petit ou très grand devant les variations de la fonction f.
Toutes les quantités décrivant l'algorithme sont définies à travers V comme le minimum ou le maximum d'un ensemble fini de sommes construites avec V et sont donc aussi des fonctions affines de m lorsque m est grand: il en est ainsi de l'énergie virtuelle et de l'altitude de communication.

Toutes ces fonctions affines, qui sont en nombre fini, possèdent un nombre fini de points d'intersection. Soit M^* un entier assez grand au–delà duquel il n'y a plus de changements de position relative entre deux fonctions affines. La hiérarchie des cycles sur U est construite à l'aide de classes d'équivalence de relations de comparaison liées à l'altitude de communication. Ces relations sont figées lorsque m est supérieur à M^* si bien que la hiérarchie des cycles se stabilise et ne dépend plus de m dès que m dépasse M^*.
Nous étudions la structure de cette hiérarchie limite. Nous avons $W^* \subset f^*$. Nous montrons que la fonction f est constante sur les cycles qui ne rencontrent pas W^*. Lorsque le noyau de mutation est en outre symétrique, nous avons $W^* = f^*$ et la loi limite est la loi uniforme sur l'ensemble f^*.
Nous donnons pour finir différents résultats de convergence.

Théorème. *(hauteur critique H_1, Trouvé* [10, Théorème 3.23]*)*
Supposons α symétrique et m assez grand pour que $W^* = f^*$. Il existe une valeur critique H_1, appelée hauteur critique, qui est une fonction bornée de la taille m, telle que, pour toute suite croissante $l(n)$ tendant vers l'infini, nous avons l'équivalence

$$\sup_{x \in E^m} \quad P\left(X_n \notin f^*/X_0 = x\right) \underset{n\to\infty}{\longrightarrow} 0 \qquad \Longleftrightarrow \qquad \sum_{n=0}^{\infty} l(n)^{-H_1} = \infty.$$

Au contraire de H_1, la hauteur de sortie H_e^* du bassin d'attraction de f^* croît linéairement avec m. Ainsi, en adaptant correctement la suite $l(n)$, il est possible d'obtenir un algorithme qui contient définitivement des points de f^* après un nombre fini de transitions.

Théorème. *Soit $H_e^* = H_e(\{\, x \in E^m : [x] \cap f^* \neq \emptyset \,\})$.*
Supposons m assez grand pour avoir $W^* = f^*$ et $H_1 < H_e^*$. Pour toute suite croissante $l(n)$ tendant vers l'infini, nous avons l'équivalence

$$\begin{aligned} \forall x \in E^m \qquad & P\left(\exists N \quad \forall n \geq N \quad [X_n] \cap f^* \neq \emptyset / X_0 = x\right) = 1 \\ \Longleftrightarrow \qquad & \sum_{n=0}^{\infty} l(n)^{-H_1} = \infty, \qquad \sum_{n=0}^{\infty} l(n)^{-H_e^*} < \infty. \end{aligned}$$

Nous étudions l'exposant de convergence optimale α_{opt} défini par Azencott, Catoni et Trouvé qui caractérise la vitesse de convergence optimale de l'algorithme en temps fini.

Théorème. *Il existe deux constantes positives R_1 et R_2 telles que pour tout n*

$$\frac{R_1}{n^{\alpha_{\text{opt}}}} \quad \leq \quad \inf_{0 \leq l(1) \leq \cdots \leq l(n)} \quad \max_{x \in E^m} \quad P\left(X_n \notin f^*/X_0 = x\right) \quad \leq \quad \frac{R_2}{n^{\alpha_{\text{opt}}}}.$$

De plus l'exposant α_{opt} est une fonction affine strictement croissante de m.

La vitesse de convergence croît donc linéairement avec la taille de la population. Ce dernier résultat montre que l'algorithme de mutation–sélection est intrinsèquement parallèle.

5. Un nouvel algorithme génétique

Nous présentons dans le troisième article un nouvel algorithme génétique qui trouve simultanément tous les maxima globaux de la fonction d'adaptation.

Le cadre de travail est inchangé: nous considérons une fonction f définie sur un espace fini E, à valeurs dans $\mathbb{R}_+^*$, dont nous cherchons à déterminer l'ensemble f^* des maxima globaux. Nous introduisons les notations suivantes: si K est une partie de E^m,

$$\begin{gathered}
[K] = \{\, x_k : 1 \leq k \leq m,\ x = (x_1, \cdots, x_m) \in K \,\}, \\
\widehat{f}(K) = \max \{\, f(x_k) : 1 \leq k \leq m,\ x = (x_1, \cdots, x_m) \in K \,\}, \\
\widehat{K} = \{\, x_k : 1 \leq k \leq m,\ x = (x_1, \cdots, x_m) \in K,\ f(x_k) = \widehat{f}(K) \,\}.
\end{gathered}$$

En particulier, si x est un élément de E^m, nous notons $[x] = [\{x\}], \widehat{x} = \widehat{\{x\}}$. Si F est une partie de E, $f(F)$ (ou fF) dénote l'ensemble $\{\, f(i) : i \in F \,\}$. Pour un réel positif λ, nous posons $f_\lambda = f^{-1}(\{\lambda\})$ i.e. f_λ est l'ensemble des éléments de E de niveau d'adaptation λ. Pour x élément de E^m, nous notons $x^\lambda = [x] \cap f_\lambda$ et nous numérotons les éléments de l'ensemble x^λ:

$$x^\lambda = \{\, x_1^\lambda, \cdots, x_{|x^\lambda|}^\lambda \,\}$$

en suivant leur ordre d'apparition dans la suite $(x_k, 1 \leq k \leq m)$. Lorsque $\lambda = \widehat{f}(x)$, nous utilisons la notation $\widehat{x} = \{\, \widehat{x}_1, \cdots, \widehat{x}_{|\widehat{x}|} \,\}$. Enfin, pour un élément x de E^m, nous dénotons par $\lambda_1^x, \cdots, \lambda_{|f[x]|}^x$ les $|f[x]|$ éléments de l'ensemble $f[x]$, où à nouveau l'indexation respecte l'ordre d'apparition des éléments de $f[x]$ dans la suite $(f(x_k), 1 \leq k \leq m)$.

Nous décrivons maintenant le processus de fond (X_n^∞) qui guide l'algorithme.
Ce processus est toujours une chaîne de Markov d'espace d'états E^m.
Le mécanisme de transition comporte deux étapes:

$$X_n^\infty \xrightarrow{\text{crossing–over}} Z_n^\infty \xrightarrow{\text{sélection}} X_{n+1}^\infty.$$

$X_n^\infty \longrightarrow Z_n^\infty$**: crossing–over.** L'opérateur de crossing–over est un opérateur stochastique qui agit sur les couples formés par les individus consécutifs de la population X_n^∞. Soit β un noyau markovien sur l'espace des couples $E \times E$. La quantité $\beta((i_1, j_1), (i_2, j_2))$ est la probabilité d'obtenir la paire d'individus (i_2, j_2) en faisant un crossing–over sur le couple (i_1, j_1). Nous imposons trois conditions d'origine biologique sur le noyau β.

- Lorsque deux individus identiques s'accouplent, les descendants sont identiques

$$\forall i \in E \qquad \beta((i,i),(i,i)) = 1.$$

- La probabilité que rien ne se passe pendant le crossing–over est non–nulle

$$\forall i, j \in E \qquad \beta((i,j),(i,j)) > 0.$$

- La population est asexuée et les parents jouent des rôles symétriques

$$\forall i_1, j_1, i_2, j_2 \in E \qquad \beta((i_1, j_1), (i_2, j_2)) = \beta((j_1, i_1), (j_2, i_2)).$$

Remarquons que la première condition est tout à fait essentielle pour la bonne marche de l'algorithme. Les probabilités de transition de X_n^∞ à Z_n^∞ sont

$$P\left(Z_n^\infty = z / X_n^\infty = x\right) = \beta\left((x_1, x_2), (z_1, z_2)\right)\, \beta\left((x_3, x_4), (z_3, z_4)\right) \cdots$$

Soulignons ici une différence fondamentale avec le précédent modèle: le crossing–over est maintenant intégré au processus non perturbé.

$Z_n^\infty \longrightarrow X_{n+1}^\infty$**: sélection.** Nous proposons ici une version améliorée du mécanisme de sélection précédemment utilisé: son avantage décisif est de préserver la diversité des individus dans la population.
Nous définissons un tableau triangulaire d'entiers $\tau(k,h),\ 0 \leq k \leq h+1,\ 1 \leq h \leq m$, par

$$\begin{array}{lll} \forall h \in \{1 \cdots m\} & & \tau(0,h) = 1, \quad \tau(h+1,h) = m, \\ \forall k, h \quad 1 \leq k \leq h \leq m & & \tau(k,h) = 2k \left\lfloor \dfrac{m}{2(h+1)} \right\rfloor + 1. \end{array}$$

Supposons que $Z_n^\infty = z$. Pour construire la population $X_{n+1}^\infty = x$, nous commençons par choisir aléatoirement avec la loi uniforme une permutation σ de $\{1 \cdots |\widehat{z}|\}$ (i.e. $\sigma \in \mathfrak{S}(|\widehat{z}|)$). A l'aide de la subdivision

$$1 = \tau(0, |\widehat{z}|) \leq \tau(1, |\widehat{z}|) \leq \cdots \leq \tau(|\widehat{z}|, |\widehat{z}|) \leq \tau(|\widehat{z}|+1, |\widehat{z}|) = m$$

de $\{1 \cdots m\}$, nous divisons l'ensemble d'indexation $\{1 \cdots m\}$ en $|\widehat{z}|+1$ parties.
Les composantes de la r–ième partie, pour $1 \leq r \leq |\widehat{z}|$, prennent alors la valeur $\widehat{z}_{\sigma(r)}$, i.e. pour les indices k tels que $\tau(r-1, |\widehat{z}|) \leq k < \tau(r, |\widehat{z}|)$ nous posons $x_k = \widehat{z}_{\sigma(r)}$.
Les composantes de la $(|\widehat{z}|+1)$–ième partie (dont les indices k vérifient $\tau(|\widehat{z}|, |\widehat{z}|) \leq k \leq m$) sont choisies indépendamment et uniformément sur l'ensemble $\widehat{z}$.
Remarquons que les $|\widehat{z}|$ premières parties sont de cardinal pair, de sorte que le crossing–over ne peut agir sur une paire d'individus appartenant à deux parties distinctes. Comme de plus chaque partie ne contient qu'un seul type d'individus, la première condition imposée sur le crossing–over entraîne que celui–ci n'a aucun effet sur les $|\widehat{z}|$ premières parties. L'intérêt essentiel de la $(|\widehat{z}|+1)$–ème partie est de permettre à des individus différents de s'accoupler sans restriction.
Le processus (X_n^∞) est instantanément piégé dans des sous–ensembles particuliers de E^m que nous appelons les attracteurs.

Définition–Proposition. (caractérisation des attracteurs)
Les attracteurs du processus (X_n^∞) sont les ensembles de populations K tels que

$i)$ $[K] = \widehat{K}$

$ii)$ une population $x = (x_1, \cdots, x_m)$ de E^m appartient à K si et seulement si

- $\forall r \in \{1 \cdots |\widehat{K}|\} \quad \forall k, h \qquad \tau(r-1, |\widehat{K}|) \leq k, h < \tau(r, |\widehat{K}|) \quad \Longrightarrow \quad x_k = x_h,$
- $\{\, x_k : \ 1 \leq k < \tau(|\widehat{K}|, |\widehat{K}|) \,\} = [K],$
- $\{\, x_k : \tau(|\widehat{K}|, |\widehat{K}|) \leq k \leq m \,\} \subset [K].$

L'ensemble de tous les attracteurs est dénoté par $\mathcal{K}$ (ainsi $\mathcal{K} \in \mathcal{P}(\mathcal{P}(E^m))$).

Un attracteur K est stable si et seulement si le crossing–over ne peut créer des individus d'adaptation supérieure à $\widehat{f}(K)$ à partir de deux individus quelconques de $[K]$. La structure de l'ensemble $\mathcal{K}$ des attracteurs est beaucoup plus riche que dans le précédent modèle. Tous les points présents dans les populations d'un attracteur fixé ont même adaptation. Notons $\mathcal{E}$ l'ensemble des parties de E formées de points de même adaptation et de cardinal inférieur à m:

$$\mathcal{E} = \{\, F \subset E : |F| \leq m,\ \widehat{F} = F \,\}.$$

L'application qui à un attracteur K de $\mathcal{K}$ associe l'ensemble $[K]$ de $\mathcal{E}$ est une bijection. Lorsque la taille m de la population est suffisamment grande, il existe un unique attracteur idéal K^* associé à l'ensemble des maxima globaux de f^*: en particulier, les populations de K^* contiennent tous les points de f^*.
Le processus (X_n^∞) est ensuite perturbé aléatoirement par deux mécanismes distincts. Le premier agit directement sur la population et modélise la mutation. Le second consiste à relâcher la sélection des individus. Comme précédemment, l'intensité des perturbations est contrôlée par un paramètre l, et celles–ci s'évanouissent lorsque l tend vers l'infini.
Le mécanisme de transition de la chaîne perturbée X_n^l est décomposé en trois étapes:

$$X_n^l \xrightarrow{\text{mutation}} Y_n^l \xrightarrow{\text{crossing–over}} Z_n^l \xrightarrow{\text{sélection}} X_{n+1}^l.$$

Nous décrivons maintenant en détail chaque étape.

$X_n^l \longrightarrow Y_n^l$: **mutation.** Les mutations sont de petites perturbations aléatoires des individus de la population X_n^l. Soit α un noyau markovien irréductible sur l'espace E. Soit α_l le noyau irréductible sur E défini par

$$\alpha_l(i,j) = \begin{cases} \alpha(i,j)\, l^{-a} & \text{si } i \neq j \\ 1 - \sum_{e \neq i} \alpha(i,e)\, l^{-a} & \text{si } i = j \end{cases}$$

où a est un réel positif.
Les probabilités de transition de X_n^l à Y_n^l sont données par

$$P\left(Y_n^l = y / X_n^l = x\right) = \alpha_l(x_1, y_1) \cdots \alpha_l(x_m, y_m).$$

$Y_n^l \longrightarrow Z_n^l$**: crossing–over.** Le crossing–over n'est pas perturbé et cette étape est exactement la même que le passage de X_n^∞ à Z_n^∞.

$Z_n^l \longrightarrow X_{n+1}^l$**: sélection.** Le mécanisme de sélection de X_n^∞ est perturbé en relâchant aléatoirement la force de sélection: les individus d'adaptation non maximale ont une probabilité non nulle de survie.
Résumons brièvement et sans formule ce mécanisme.
Supposons que $Z_n^l = z$ et que nous voulons construire le vecteur $X_{n+1}^l = x$.
Nous traversons d'abord l'ensemble $f[z]$. A chaque élément λ de cet ensemble, nous associons une suite $\psi_1, \cdots, \psi_{n_\lambda}$ obtenue en réordonnant aléatoirement (tous les ordres étant également probables) l'ensemble $f^{-1}(\{\lambda\}) \cap f[z]$.
La population x est ensuite construite ainsi: pour chaque composante $x_k, 1 \leq k \leq m$, nous tirons une valeur λ avec une loi de probabilité sur l'ensemble $f[z]$ qui favorise les valeurs élevées. A cette valeur λ, nous avions précédemment associé une suite $\psi_1, \cdots, \psi_{n_\lambda}$. Nous divisons l'ensemble d'indices $\{1 \cdots m\}$ en $n_\lambda + 1$ parties. Si l'indice k considéré appartient à la partie numéro r, où $1 \leq r \leq n_\lambda$, nous prenons $x_k = \psi_r$. Si k appartient à la $(n_\lambda + 1)$–ième partie, nous choisissons x_k au hasard et uniformément sur l'ensemble $\{\psi_1, \cdots, \psi_{n_\lambda}\}$.
Décrivons maintenant précisément ce mécanisme.
Pour chaque indice h dans $\{1 \cdots |f[z]|\}$, nous sélectionnons aléatoirement et indépendamment avec la loi uniforme une permutation σ^h de $\mathfrak{S}(|z^{\lambda_h}|) = \mathfrak{S}(|[z] \cap f_{\lambda_h}|)$.
La loi de chaque composante $x_k\,(1 \leq k \leq m)$ de x est définie ainsi:
Nous sélectionnons aléatoirement une valeur λ dans l'ensemble $f[z]$ avec la distribution de probabilité

$$\forall h \in \{1 \cdots |f[z]|\} \qquad P(\lambda = \lambda_h) = \frac{\exp(c\lambda_h \ln l)}{\sum_{r=1}^{|f[z]|} \exp(c\lambda_r \ln l)}$$

où c est un réel positif.
La valeur de x_k est alors choisie dans l'ensemble z^λ selon la valeur de l'indice k:

- Si $\tau(|z^\lambda|, |z^\lambda|) \leq k \leq \tau(|z^\lambda| + 1, |z^\lambda|) = m$, alors x_k est choisi au hasard avec la loi uniforme sur $z^\lambda = [z] \cap f_\lambda$ (k est dans la dernière partie de l'ensemble des indices).
- S'il existe un indice r dans $\{1 \cdots |z^\lambda|\}$ tel que $\tau(r-1, |z^\lambda|) \leq k < \tau(r, |z^\lambda|)$, alors nous prenons $x_k = z^\lambda_{\sigma^h(r)}$, où h est l'unique entier dans $\{1 \cdots |z^\lambda|\}$ tel que $\lambda = \lambda_h$ (l'indice k se trouve dans la r–ième partie de l'ensemble des indices, avec $1 \leq r \leq |z^\lambda|$).

Nous avons

$$\forall y, z \in E^m \qquad \lim_{l \to \infty} P\left(X_{n+1}^l = z / X_n^l = y\right) = P\left(X_{n+1}^\infty = z / X_n^\infty = y\right)$$

et le processus (X_n^l) est une perturbation du processus (X_n^∞).
Nous effectuons le développement des probabilités de transition de la chaîne (X_n^l) lorsque l tend vers l'infini et nous obtenons

$$P\left(X_{n+1}^l = v / X_n^l = u\right) \underset{l \to \infty}{\sim} q_1(u, v) \exp\left(-V_1(u, v) \ln l\right)$$

où q_1 est un noyau positif ou nul sur E et V_1 est un coût de communication sur E^m.
Nous sommes donc toujours dans le cadre du recuit simulé généralisé étudié par Trouvé et nous menons l'étude en utilisant conjointement les techniques des deux premiers articles.

Théorème. *(existence d'une taille de population critique)*
Il existe une taille critique de population m^* qui dépend de la fonction f, des noyaux de mutation α et de crossing–over β, et des paramètres a et c, telle que $W^* = K^*$ lorsque m est supérieur à m^*, d'où

$$\forall m \geq m^* \quad \forall x \in E^m \qquad \lim_{l\to\infty} \quad \lim_{n\to\infty} \quad P\left([X_n^l] = f^* / X_0^l = x\right) = 1.$$

De plus, la loi limite est la loi uniforme sur l'attracteur idéal K^*.

Nous donnons différentes conditions (qui dépendent fortement du problème d'optimisation) qui entraînent $W^* = K^*$, ainsi que des majorations de m^*.
Pour finir, nous appliquons les résultats de Catoni et Trouvé pour traiter le cas inhomogène où $l = l(n)$ est une fonction croissante de n.

Théorème. *(hauteur critique)*
Supposons que m est assez grand de sorte que $W^* = K^*$. Il existe une valeur positive H_1 telle que, pour toute suite croissante $l(n)$ qui tend vers l'infini, nous avons l'équivalence

$$\sup_{x\in E^m} \quad P\left(X_n \notin K^* / X_0 = x\right) \underset{n\to\infty}{\longrightarrow} 0 \qquad \Longleftrightarrow \qquad \sum_{n=0}^{\infty} l(n)^{-H_1} = \infty.$$

La hauteur critique H_1 est une fonction bornée de m.

Théorème. *Soit $H_e^* = H_e(\{\, x \in E^m : f^* \subset [x] \,\})$.*
Supposons m assez grand de sorte que $W^* = \{K^*\}$ et $H_1 < H_e^*$.
Pour toute suite croissante $l(n)$ tendant vers l'infini, nous avons l'équivalence

$$\begin{aligned} &\forall x \in E^m \qquad P\left(\exists N \quad \forall n \geq N \quad f^* \subset [X_n] / X_0 = x\right) = 1 \\ &\Longleftrightarrow \qquad \sum_{n=0}^{\infty} l(n)^{-H_1} = \infty, \qquad \sum_{n=0}^{\infty} l(n)^{-H_e^*} < \infty. \end{aligned}$$

Proposition. *L'exposant de convergence optimale est encadré par deux fonctions affines strictement croissantes de m.*

Ce résultat montre que l'algorithme est intrinsèquement parallèle; il utilise essentiellement de petites opérations locales indépendantes.
Enfin, lorsque m est suffisamment grand, les cycles sur l'ensemble des attracteurs qui ne contiennent pas l'attracteur idéal K^* (i.e. les mauvais cycles qui ralentissent la convergence vers K^*) sont réduits à un seul attracteur.

6. Perspectives

Le cadre théorique proposé dans ce travail a déjà permis d'obtenir des résultats substantiels de convergence. De nombreuses directions de recherche sont maintenant possibles.

Tout d'abord, il serait souhaitable d'effectuer une étude plus fine des quantités cruciales qui apparaissent dans ces premiers modèles. Existe–t–il une description "directe" (en fonction du paysage d'énergie) de la taille critique de population m^* et de l'intervalle critique dans lequel doit se trouver l'exposant de convergence de la suite $l(n)$? Comment le coût de communication V varie–t–il en fonction des différents paramètres (m, a, c)?

Une conclusion essentielle est que l'implémentation efficace d'un algorithme génétique dépend très fortement du problème d'optimisation. Supposons ensuite que la fonction d'adaptation ne soit pas quelconque mais qu'elle soit choisie dans une classe de fonctions particulières définies sur un espace fixe. Quels sont alors les mécanismes de mutation et de crossing–over et les paramètres (m, a, c) les plus efficaces pour cette classe de fonctions?

Ce premier travail ne concerne que le plus simple des algorithmes génétiques. De nombreuses variantes et extensions construites à l'aide d'opérateurs génétiques plus sophistiqués peuvent être analysés par les mêmes techniques. Plus généralement, il s'agirait d'évaluer l'impact et l'intérêt de chaque phénomène génétique (redondance, diploïdie, sexe, dominance, parties non codantes, ...) pour le processus d'optimisation.

Enfin, différentes versions continues de ces modèles discrets sont envisageables; il s'agit là d'un axe de travail important qui nécessitera de développer de nouveaux outils.

Références

1. O. Catoni, *Large deviations for annealing*, PhD Thesis, University Paris XI (1990).
2. O. Catoni, *Rough large deviations estimates for simulated annealing. Application to exponential schedules*, The Annals of Probability **20** (1992), 1109–1146.
3. R. Cerf, *Asymptotic convergence of genetic algorithms*, Preprint, submitted (1993).
4. R. Cerf, *The dynamics of mutation-selection algorithms with large population sizes*, Preprint, submitted (1993).
5. R. Cerf, *A new genetic algorithm*, Preprint, submitted (1993).
6. M.I. Freidlin–A.D. Wentzell, *Random perturbations of dynamical systems*, Springer–Verlag, New York, 1984.
7. D. Goldberg, *Genetic algorithms in search, optimization and machine learning*, Addison–Wesley, 1989.
8. J.H. Holland, *Adaptation in natural and artificial systems*, The University of Michigan Press, Ann Arbor, 1975.
9. M. Nei, *Molecular evolutionary genetics*, Columbia University Press, New York, 1987.
10. A. Trouvé, *Cycle decompositions and simulated annealing*, Preprint (1993).
11. A. Trouvé, *Parallélisation massive du recuit simulé*, PhD Thesis, University Paris XI (1993).
12. A. Trouvé, *Rough large deviation estimates for the optimal convergence speed exponent of generalized simulated annealing algorithms*, Preprint (1993).
13. Voelker – Schaffer – Mukai, *Spontaneous allozyme mutations in drosophilia melanogaster: rate of occurrence and nature of the mutants*, Genetics **94** (1980), 961–968.

Chapitre Deuxième

CONVERGENCE ASYMPTOTIQUE DES ALGORITHMES GÉNÉTIQUES

ASYMPTOTIC CONVERGENCE OF GENETIC ALGORITHMS

Raphaël CERF

Ecole Normale Supérieure, Paris

February 1993

Abstract. We study a markovian evolutionary process which encompasses the classical simple genetic algorithm. This process is obtained by perturbing randomly a very simple selection scheme. Using the Freidlin–Wentzell theory, we carry out a precise study of the asymptotic dynamics of the process as the perturbations disappear. We show how a delicate interaction between the perturbations and the selection pressure may force the convergence toward the global maxima of the fitness function. We derive several conditions to ensure convergence in the homogeneous as well as the inhomogeneous case: these provide the first mathematically well-founded convergence results for genetic algorithms.

1991 *Mathematics Subject Classification.* 60F10, 60J10, 92D15.
Key words and phrases. Freidlin–Wentzell theory, genetic algorithms, stochastic optimization.
This work was done during my last year as a student at the Ecole Normale Supérieure.
Professor Jean–Arcady Meyer first spoke to me of genetic algorithms.
I am deeply grateful to Professor Gérard Ben Arous for introducing me to the Freidlin–Wentzell theory.
I thank Professor Alain Berlinet for his useful advice and for having carefully read this paper.

I. Introduction

Stochastic optimization has been a field of intense research during the last decade: computers become more and more powerful and there is an increasing need for robust and efficient search techniques. In the jungle of optimization procedures, there exists a particular class of algorithms called genetic algorithms. Introduced by J.H. Holland in the seventies [3], these algorithms are based on the genetic mechanisms which guide natural evolution: mutation, crossover and selection. Today, they are widely used to handle a large spectrum of optimization problems ranging from the classical traveling salesman problem to the design of network architecture [2].

However the behavior of genetic algorithms relies heavily on several control parameters (for instance the population size, the probabilities of mutation and crossover). Until now, no firm result concerning the choice of these parameters and the dynamics of the genetic algorithm has been available: practical know–how and experimental simulations were the only guides for handling concrete problems.

In this paper, we study a slightly more general evolutionary scheme. We consider first a very simple selection algorithm which is extremely rapidly trapped in bad points of the search space and we perturb randomly the whole mechanism. We focus on the asymptotic dynamics of the process as the random perturbations vanish. We show how a delicate interaction between mutations, crossovers and selection pressure ensures the convergence of the algorithm toward the set of the global maxima of the fitness function. Our results seem to be the first well–founded convergence theorems of this kind concerning genetic algorithms.

To achieve this program, we use the powerful tools developed by Freidlin and Wentzell in a much more general framework for the study of random perturbations of dynamical systems [1].

This paper has the following structure.

We first describe our model. Then, we study the convergence of the algorithm in the homogeneous case and derive sufficient conditions to ensure that the population settles in a global maximum of the fitness function. We show that these conditions are fulfilled when the population size is large enough. In the inhomogeneous case, we give several conditions on the rate of decrease of the perturbations to ensure the convergence of the population to the set of global maxima in finite time. We finally apply our results to the classical simple genetic algorithm.

II. General notations and conventions

The cardinality of a set X will be noted indifferently $|X|$ or $\operatorname{card} X$ and its characteristic function 1_X. We adopt usual conventions concerning empty sets:

$$\min \emptyset = +\infty, \quad \max \emptyset = -\infty, \quad \prod_{\emptyset} = 1, \quad \sum_{\emptyset} = 0.$$

If s is a real number, $\lfloor s \rfloor$ denotes the unique integer such that $\lfloor s \rfloor \leq s < \lfloor s \rfloor + 1$.
We say that a real number s is positive if $s > 0$ and that it is non–negative if $s \geq 0$.
We consider a finite space of states E and a real-valued positive non-constant function f (which will be called the fitness function) defined on E. The letters i, j will denote elements of E. The set of global maxima of f is

$$f^* = \{\, i \in E : f(i) = \max_{j \in E} f(j) \,\}.$$

By $f(f^*)$ we mean the maximum of f over E i.e. $\max_{j\in E} f(j)$.
The Kronecker symbol $\delta(i,j)$ will be used to denote the identity matrix indexed by E:

$$\forall i,j \in E \qquad \delta(i,j) = 0 \quad \text{if} \quad i \neq j, \quad \delta(i,j) = 1 \quad \text{if} \quad i = j.$$

Let m be a positive integer. The state space of most of Markov chains under consideration will be E^m, the set of m–uples of elements of E: these m–uples are called populations and their components individuals. They will be denoted by the letters x, y, z.
If $x = (x_1, \cdots, x_m)$ belongs to E^m, $[x]$ is the set of individuals contained in x i.e.

$$[x] = \{\, x_k : 1 \leq k \leq m \,\}$$

and for i in E, $x(i)$ is the number of occurrences of i in the population x:

$$x(i) = \operatorname{card}\{\, k : 1 \leq k \leq m,\; x_k = i \,\}.$$

The group of permutations $\mathfrak{S}_m$ of the set $\{1 \cdots m\}$ operates on E^m in the following way:

$$\forall \sigma \in \mathfrak{S}_m \qquad \sigma \cdot x = \sigma \cdot (x_1, \cdots, x_m) = (x_{\sigma(1)}, \cdots, x_{\sigma(m)}).$$

The set $E^m/\mathfrak{S}_m$ is the set of equivalence classes associated with this group operation.
We associate with f two functions defined on E^m.
The first one is still noted f and takes its values in $(\mathbb{R}_+^*)^m$:

$$f(x) = f(x_1, \cdots, x_m) = (f(x_1), \cdots, f(x_m)).$$

The second one is noted $\widehat{f}$ and is real-valued:

$$\widehat{f}(x) = \widehat{f}(x_1, \cdots, x_m) = \max_{1 \leq k \leq m} f(x_k).$$

For x in E^m, $\hat{x}$ denotes the set of those elements of $[x]$ which realize the value $\widehat{f}(x)$:

$$\hat{x} = \{\, x_k : 1 \leq k \leq m,\; f(x_k) = \widehat{f}(x) \,\}.$$

For a point i of E, (i) is the m–uple whose m components are equal to i and A is the set of all such m–uples (which are called the uniform populations). By S we denote the set of equi–fitness populations, that is, populations whose individuals have the same fitness:

$$S = \{\, x \in E^m : f(x_1) = \cdots = f(x_m) \,\}.$$

We sometimes identify f^* with $\{\, (i) : i \in f^* \}$ so that f^* may be seen as a subset of A.

III. Description of the model

1. The unperturbed Markov chain (X_n^∞)

In the absence of perturbations, the process under study is a Markov chain $(X_n^\infty)_{n\geq 0}$ with state space E^m. The superscript ∞ reflects the fact that this process describes the limit behavior of our model, when all perturbations have disappeared. The transition probabilities of this chain are

$$P\left(X_{n+1}^\infty = z / X_n^\infty = y\right) = \frac{1}{(\operatorname{card}\hat{y})^m}\prod_{k=1}^{m} 1_{\hat{y}}(z_k) y(z_k) = \frac{1}{(\operatorname{card}\hat{y})^m}\prod_{i\in[z]} 1_{\hat{y}}(i) y(i)^{z(i)}$$

that is, the individuals of the population X_{n+1}^∞ are chosen randomly (under the uniform distribution) and independently among the elements of $\hat{X}_n^\infty$ which are the best individuals of X_n^∞ according to the fitness function f.
Suppose the chain starts up with the initial population $X_0 = x_0$. Then

$$\forall n \geq 1 \qquad [X_n^\infty] \subset \hat{x}_0$$

and with probability one, after a finite number of steps N, the chain is absorbed in a state (i) where i belongs to $\hat{x}_0$. In particular, if $\hat{x}_0$ is reduced to one point i, the chain is instantaneously absorbed in (i).

2. The perturbed Markov chain (X_n^l)

The previous Markov chain (X_n^∞) is randomly perturbed by three distinct mechanisms. The first two act directly upon the population and mimic the phenomena of mutation and crossover. The third one consists in loosening the selection of the individuals.
The intensity of the perturbations is governed by an integer parameter l. As l grows toward infinity, the perturbations progressively disappear.
The perturbed Markov chain (X_n^l) is obtained through overlapping of several other chains $(U_n^l), (V_n^l)$ which represent the successive populations obtained by applying the perturbing operators. More precisely, we decompose the transition from X_n^l to X_{n+1}^l in three stages:

$$X_n^l \xrightarrow{\text{mutation}} U_n^l \xrightarrow{\text{crossover}} V_n^l \xrightarrow{\text{selection}} X_{n+1}^l.$$

We now proceed to a more detailed description of these three operations.

2.1. $X_n^l \longrightarrow U_n^l$**: mutation.** The mutation operator is modelled by random independent perturbations of the individuals of the population X_n^l. Such a perturbation is described by a markovian kernel p_l on the space E, that is a function defined on $E \times E$ with values in $[0,1]$ verifying

$$\forall i \in E \qquad \sum_{j \in E} p_l(i,j) = 1.$$

The transition probabilities from X_n^l to U_n^l are then given by

$$P\left(U_n^l = u / X_n^l = x\right) = p_l(x_1, u_1) \cdots p_l(x_m, u_m). \tag{1}$$

This perturbation is small whenever the matrix $(p_l(i,j))_{(i,j) \in E \times E}$ is close to the identity matrix. To ensure the vanishing of mutations when l grows toward infinity, we will impose

$$\forall i, j \in E \qquad \lim_{l \to \infty} p_l(i,j) = \delta(i,j). \tag{2}$$

2.2. $U_n^l \longrightarrow V_n^l$**: crossover.** The crossover operator is modelled by random independent perturbations of the couples formed by consecutive individuals of the population (X_n^l). As in the mutation case, such a perturbation is described by a markovian kernel q_l on the space $E \times E$, that is a function defined on $(E \times E) \times (E \times E)$ with values in $[0,1]$ verifying:

$$\forall (i_1, j_1) \in E \times E \qquad \sum_{(i_2, j_2) \in E \times E} q_l\left((i_1, j_1), (i_2, j_2)\right) = 1.$$

The transition probabilities from U_n^l to V_n^l are

$$P\left(V_n^l = v / U_n^l = u\right) = \delta_m(u_m, v_m) \prod_{1 \le k \le m/2} q_l\left((u_{2k-1}, u_{2k}), (v_{2k-1}, v_{2k})\right) \tag{3}$$

where $\delta_m(i,j) = \delta(i,j)$ if m is odd (the last individual of the population remains unchanged after crossover) and $\delta_m(i,j) = 1$ if m is even.
To ensure the vanishing of crossovers when l grows toward infinity, we will impose

$$\forall (i_1, j_1) \in E \times E \quad \forall (i_2, j_2) \in E \times E \qquad \lim_{l \to \infty} q_l\left((i_1, j_1), (i_2, j_2)\right) = \delta(i_1, i_2)\delta(j_1, j_2). \tag{4}$$

2.3. $V_n^l \longrightarrow X_{n+1}^l$: selection. In order to build our selection operator, we will use a selection function.

Definition. A selection function of order m is a function F defined on $\{1\cdots m\}\times(\mathbb{R}_+^*)^m$ with values in $[0,1]$ satisfying for each $(f_1,\cdots,f_m)$ in $(\mathbb{R}_+^*)^m$:

$$a)\quad \sum_{k=1}^{m} F(k,f_1,\cdots,f_m)=1,$$

$$b)\quad \forall\sigma\in\mathfrak{S}_m\quad \forall k\in\{1\cdots m\}\qquad F(\sigma(k),f_{\sigma(1)},\cdots,f_{\sigma(m)})=F(k,f_1,\cdots,f_m),$$

$$c)\quad f_1\geq f_2\geq\cdots\geq f_m\Rightarrow F(1,f_1,\cdots,f_m)\geq F(2,f_1,\cdots,f_m)\geq\cdots\geq F(m,f_1,\cdots,f_m).$$

The value $F(k,f_1,\cdots,f_m)$ is the probability of choosing f_k among the values $f_1,\cdots,f_m$: this probability does not depend on the indices and increases with the relative value of f_k among $f_1,\cdots,f_m$. Now let F_l be a selection function. The m individuals $X_{n+1}^{l,1},\cdots,X_{n+1}^{l,m}$ who make up the population X_{n+1}^l are chosen randomly and independently in the population V_n^l from the law defined by F_l:

$$\forall\, r\in\{1\cdots m\}\quad \forall i\in E\qquad P\left(X_{n+1}^{l,r}=i\right)=\sum_{h:V_n^{l,h}=i} F_l\left(h,f(V_n^l)\right).$$

(recall that $f(V_n^l)$ is the m–uple $\left(f(V_n^{l,1}),\cdots,f(V_n^{l,m})\right)$ so that the transition probabilities from V_n^l to X_{n+1}^l are

$$(5)\qquad P\left(X_{n+1}^l=x/V_n^l=v\right)=\prod_{i\in[x]}\Bigg(\sum_{k:v_k=i}F_l(k,f(v))\Bigg)^{x(i)}=\prod_{r=1}^{m}\sum_{k:v_k=x_r}F_l(k,f(v)).$$

The selection pressure is maximal if the individuals of X_{n+1}^l are chosen randomly and uniformly from the fittest individuals of V_n^l. The unique selection function F_∞ which implements such a selection scheme is defined by

$$F_\infty(k,f(x))=\frac{1_{\hat{x}}(x_k)}{\operatorname{card}\hat{x}}$$

i.e. we have then the uniform distribution over $\hat{x}$.
To ensure the disappearance of the selection of individuals below peak fitness, we will impose the convergence of F_l toward F_∞ on the set $f(E)^m$:

$$(6)\qquad \forall x\in E^m\quad \forall k\in\{1\cdots m\}\qquad \lim_{l\to\infty}F_l(k,f(x))=F_\infty(k,f(x)).$$

The basic example of such a sequence of selection functions is given by

$$(7)\qquad F_l(k,f_1,\cdots,f_m)=\frac{\exp(f_k\ln w_l)}{\sum_{r=1}^{m}\exp(f_r\ln w_l)}$$

where $(w_l)_{l\geq 0}$ is an increasing sequence of positive real numbers which tends to infinity. We will only be concerned with such sequences of selection functions in the sequel.
We can now evaluate the

2.4. Transition probabilities of the chain (X_n^l).

$$P\left(X_{n+1}^l = z/X_n^l = y\right) = \sum_{(u,v)\in(E^m)^2} P\left(X_{n+1}^l = z/V_n^l = v\right) P\left(V_n^l = v/U_n^l = u\right) \times$$
$$P\left(U_n^l = u/X_n^l = y\right)$$

.

Conditions (2), (4) and (6) imply

$$\forall (y,z) \in E^m \times E^m \qquad \lim_{l\to\infty} P\left(X_{n+1}^l = z/X_n^l = y\right) = P\left(X_{n+1}^\infty = z/X_n^\infty = y\right)$$

so that the transition probabilities of (X_n^l) converge toward those of (X_n^∞) as l tends to infinity. Thus the Markov chain (X_n^l) appears as a perturbation of the Markov chain (X_n^∞). In order to deal with the asymptotic dynamics of the chain (X_n^l) we need more information about the way its transition probabilities converge to those of (X_n^∞). We will make the following assumptions on the sequences of kernels p_l and q_l.

Hypothesis H_p. There exist a sequence $(u_l)_{l\geq 0}$ of positive real numbers and a function α defined on $E \times E$ with values in $\mathbb{R}^+$ satisfying

a) $\lim\limits_{l\to\infty} u_l = +\infty$

b) α is an irreducible kernel, that is

$$\forall i,j \in E \quad \exists i_0, i_1, \cdots, i_r \quad i_0 = i, \quad i_r = j, \qquad \prod_{0\leq k\leq r-1} \alpha(i_k, i_{k+1}) > 0$$

c) p_l admits the development

$$\forall i,j \in E \quad \forall s > 0 \qquad p_l(i,j) = \begin{cases} \alpha(i,j)(u_l)^{-1} + o((u_l)^{-s}) & \text{if } i \neq j \\ 1 - \alpha(i,j)(u_l)^{-1} + o((u_l)^{-s}) & \text{if } i = j \end{cases}$$

Hypothesis H_q. There exist a sequence $(v_l)_{l\geq 0}$ of positive real numbers and a function β defined on $(E\times E)\times(E\times E)$ with values in $\mathbb{R}^+$ satisfying

a) $\lim\limits_{l\to\infty} v_l = +\infty$

b) q_l admits the development

$$\forall (i_1,j_1) \in E\times E \quad \forall (i_2,j_2) \in E \times E \quad \forall s > 0$$

$$q_l\left((i_1,j_1),(i_2,j_2)\right) = \begin{cases} \beta\left((i_1,j_1),(i_2,j_2)\right)(v_l)^{-1} + o\left((v_l)^{-s}\right) & \text{if } (i_1,j_1) \neq (i_2,j_2) \\ 1 - \beta\left((i_1,j_1),(i_2,j_2)\right)(v_l)^{-1} + o\left((v_l)^{-s}\right) & \text{if } (i_1,j_1) = (i_2,j_2) \end{cases}$$

In fact, the asymptotic dynamics of the chain will be governed by the kernels α and β: the irreducibility condition on the kernel α (which implies the irreducibility of the kernel p_l for l large enough) is essential in order to allow the population to visit all the space E even when the perturbations are small.

IV. Convergence of the homogeneous algorithm

This section is devoted to the study of the behavior of the chain (X_n^l) when n first goes to infinity and then l does. We will be interested in the quantities

$$\lim_{l\to\infty} \quad \lim_{n\to\infty} \quad P\left(X_n^l = z / X_0^l = y\right).$$

The kernel α is irreducible so that for l large enough, the chain (X_n^l) is irreducible. Furthermore, for l sufficiently large, we have also for all x in E^m

$$P\left(X_{n+1}^l = x / X_n^l = x\right) > 0$$

so that the chain (X_n^l) is aperiodic. In addition, the state space of (X_n^l) is finite. The chain (X_n^l) admits thus a unique invariant probability measure μ^l[6, Theorems 4.1 and 4.2]. This measure charges all points of the space E^m and we have

$$\forall\, y, z \in E^m \qquad \lim_{n\to\infty} P\left(X_n^l = z / X_0^l = y\right) = \mu^l(z).$$

We are thus interested in the convergence of the stationary measures $(\mu^l)_{l\geq 0}$.
We will note P_x for the probability measure associated with the chain (X_n^l) starting at $X_0^l = x$ and E_x will be the expectation with respect to P_x.

1. The time of entrance in the set S of X_n^l

Let τ^l be the first entrance time of the chain (X_n^l) into S:

$$\tau^l = \min\{\, n > 0 : X_n^l \in S \,\}.$$

Let

$$s_l = \min_{x\in E^m} \sum_{k : x_k \in \hat{x}} F_l(k, f(x))$$

so that

$$\forall n \in \mathbb{N} \quad \forall x \in E^m \qquad P\left(X_{n+1}^l \in S / X_n^l = x\right) \geq P\left([X_{n+1}^l] \subset \hat{x} / X_n^l = x\right) \geq (s_l)^m.$$

From this we deduce that

$$\forall q \in \mathbb{N} \quad \forall x \in E^m \qquad P_x\left(\tau^l > q\right) \leq \left(1-(s_l)^m\right)^q. \tag{8}$$

Summing up this inequality from $q = 0$ to ∞ we get

$$\forall x \in E^m \qquad E_x\left(\tau^l\right) \leq (s_l)^{-m}.$$

We proceed now to a more careful study of the sequence $(s_l)_{l\geq 0}$. With our choice for the sequence F_l, it is clear that s_l is strictly positive for every l. In addition, for a fixed population x, $F_l(k, f(x))$ increases with l (recall that w_l is an increasing sequence) whenever x_k belongs to $\hat{x}$. The sequence s_l, being the minimum over E^m of sums of such increasing sequences, is also increasing. Define

$$\delta = \min\{\, |f(i) - f(j)| : i, j \in E,\ f(i) \neq f(j)\,\}. \tag{9}$$

For x in E^m, we have

$$1 - \sum_{k:x_k\in\hat{x}} F_l(k, f(x)) = \frac{\sum_{r:x_r\notin\hat{x}} \exp((f(x_r) - \widehat{f}(x))\ln w_l)}{\operatorname{card}\hat{x} + \sum_{r:x_r\notin\hat{x}} \exp((f(x_r) - \widehat{f}(x))\ln w_l} \leq (m-1)(w_l)^{-\delta}$$

whence finally

$$1 - s_l \leq (m-1)(w_l)^{-\delta}. \tag{10}$$

2. Representation of the measure μ^l

We are now in position to apply the following well–known [4, Proposition 5.3]

Theorem 2.1. *Let $(X_n)_{n\geq 0}$ be a Markov chain in a measurable space $(M, \mathcal{B})$ with transition probabilities $P(x, \Gamma)$ having an invariant probability measure μ. Let $V \subset M$ be a measurable set such that*

$$\sup_{x\in M} E_x(\tau_V) < \infty$$

where

$$\tau_V = \inf\{\, n > 0 : X_n \in V\,\}.$$

Then $\mu(V) > 0$ and we can define another Markov chain ${}^V\!X_n$ (called the induced Markov chain) on V by its transition probabilities ${}^V\!P(x, \Gamma), x \in V$ having the form

$${}^V\!P(x, \Gamma) = P_x(X_{\tau_V} \in \Gamma)$$

where Γ is a measurable subset of V and $P_x(\quad)$ denotes the probability for the Markov chain X_n starting at x. Then the restriction μ_V of $(\mu(V))^{-1}\mu$ to V is the invariant probability measure of the Markov chain ${}^V\!X_n$ and for any measurable set $G \subset M$,

$$\mu(G) = \mu(V)\int_V d\mu_V(x)\, E_x\left[\sum_{k=0}^{\tau_V - 1} 1_G(X_k)\right] = \int_V d\mu(x) E_x\left[\sum_{k=0}^{\tau_V-1} 1_G(X_k)\right]$$

which gives the representation of μ via μ_V, where 1_G denotes the indicator of a set G.

We apply this result to our Markov chain (X_n^l) and the set S. We denote by (Z_n^l) the induced Markov chain $({}^S\!X_n^l)$ and by ν^l its invariant probability measure μ_S^l. We then have

$$\forall x \in E^m \qquad \mu^l(x) = \mu^l(S)\sum_{y\in S}\nu^l(y)E_y\left[\sum_{k=0}^{\tau^l-1} 1_{\{X_k^l = x\}}\right]. \tag{11}$$

The study of the asymptotic behavior of μ^l will entirely rely on this representation formula.

3. Concentration of μ^l on S

If x does not belong to S then for y in S

$$E_y\left[\sum_{k=0}^{\tau^l-1} 1_{\{X_k^l=x\}}\right] = E_y\left[\sum_{k=1}^{\tau^l-1} 1_{\{X_k^l=x\}}\right] \leq E_y[\tau^l]-1 \leq (s_l)^{-m}-1.$$

Reporting this inequality in the representation formula (11) yields

$$\mu^l(x) \leq \mu^l(S)\left((s_l)^{-m}-1\right).$$

The convergence of the selection functions $(F_l)_{l\geq 0}$ toward F_∞ implies $\lim_{l\to\infty} s_l = 1$ whence

$$\forall x \in E^m \setminus S \qquad \lim_{l\to\infty} \mu^l(x) = 0$$

and by the additivity property

$$\lim_{l\to\infty} \mu^l(E^m \setminus S) = 0.$$

If x belongs to S then for y in S

$$E_y\left[\sum_{k=0}^{\tau^l-1} 1_{\{X_k^l=x\}}\right] = \delta(x,y).$$

Reporting this equality in the representation formula (11) yields

$$\mu^l(x) = \mu^l(S)\nu^l(x).$$

We have just shown that

$$\lim_{l\to\infty} \mu^l(S) = 1.$$

Therefore it remains to study the convergence of the sequence of probability measures ν^l.

4. Asymptotic behavior of the measure ν^l

To study the convergence of the measure ν^l we will use the powerful machinery developed by Freidlin and Wentzell. The first step consists in evaluating the asymptotics of the transition probabilities of the induced Markov chain (Z_n^l). Thus our next task will be to obtain estimates of the quantities

$$P\left(Z_{n+1}^l = y / Z_n^l = x\right)$$

for x and y in S.
We will successively study the asymptotics of the transition probabilities

$$P\left(U_n^l = u / X_n^l = x\right), \quad P\left(V_n^l = v / U_n^l = u\right), \quad P\left(X_{n+1}^l = x / V_n^l = v\right).$$

4.1. Asymptotics of $P\left(U_n^l = u / X_n^l = x\right)$. Define for x and u in E^m

$$d(x,u) = \text{card}\,\{\, k : 1 \le k \le m, x_k \ne u_k \,\}$$

and extend α to $E^m \times E^m$ by putting

$$\alpha(x,u) = \prod_{k : x_k \ne u_k} \alpha(x_k, u_k).$$

Define for x in E^m the set $\mathcal{U}(x)$ as the set of populations reachable from x in one transition step through the kernel α:

$$\mathcal{U}(x) = \{\, u \in E^m : \alpha(x,u) > 0 \,\}.$$

Using the development of p_l given by hypothesis H_p in (1), we see that, as $l \to \infty$,

$$\begin{array}{lll} \text{If} \quad u \notin \mathcal{U}(x) & P\left(U_n^l = u / X_n^l = x\right) = o\left((u_l)^{-s}\right) & \forall s > 0 \\ \text{If} \quad u \in \mathcal{U}(x) & P\left(U_n^l = u / X_n^l = x\right) \sim \alpha(x,u)(u_l)^{-d(x,u)}. & \end{array}$$

4.2. Asymptotics of $P\left(V_n^l = v / U_n^l = u\right)$. Define for u and v in E^m

$$\bar{d}(u,v) = \text{card}\,\{\, k : 1 \le k \le m/2, \quad (u_{2k-1}, u_{2k}) \ne (v_{2k-1}, v_{2k}) \,\}$$

and extend β to $E^m \times E^m$ by putting

$$\beta(u,v) = \delta_m(u_m, v_m) \prod_{k : (u_{2k-1}, u_{2k}) \ne (v_{2k-1}, v_{2k})} \beta((u_{2k-1}, u_{2k}), (v_{2k-1}, v_{2k})).$$

For u in E^m, let $\mathcal{V}(u)$ be the set of populations reachable from u in one transition step through the kernel β:

$$\mathcal{V}(u) = \{\, v \in E^m : \beta(u,v) > 0 \,\}.$$

Using the development of q_l given by hypothesis H_q in (3), we obtain, as $l \to \infty$,

$$\begin{array}{lll} \text{If} \quad v \notin \mathcal{V}(u) & P\left(V_n^l = v / U_n^l = u\right) = o\left((v_l)^{-s}\right) & \forall s > 0 \\ \text{If} \quad v \in \mathcal{V}(u) & P\left(V_n^l = v / U_n^l = u\right) \sim \beta(u,v)(v_l)^{-\bar{d}(u,v)}. & \end{array}$$

4.3. Asymptotics of $P\left(X_{n+1}^l = z/V_n^l = y\right)$. Formulas (5) and (7) yield

$$P\left(X_{n+1}^l = z/V_n^l = y\right) = \prod_{i\in[z]} \left(\frac{y(i)\exp(f(i)\ln w_l)}{\sum_{k=1}^m \exp(f(y_k)\ln w_l)}\right)^{z(i)}$$

whence

$$P\left(X_{n+1}^l = z/V_n^l = y\right) \underset{l\to\infty}{\sim} \frac{1}{(\operatorname{card}\hat{y})^m} \prod_{i\in[z]} y(i)^{z(i)} \exp\left(\left(\sum_{i\in[z]} z(i)f(i) - m\hat{f}(y)\right)\ln w_l\right)$$

The important quantity in this formula is the exponent of w_l which gives the rate of decreasing of the transition probabilities: it may be written

$$m\left(\frac{1}{m}\sum_{k=1}^m f(z_k) - \hat{f}(y)\right)$$

i.e. m multiplied by the difference between the mean of f over the population z and the maximum of f over the population y.

4.4. Choice of the sequences u_l, v_l and w_l. We are interested in the dynamics of the chain (X_n^l) when the perturbations become smaller and smaller that is when l grows to infinity. In this situation, with overwhelming probability, the chain (X_n^l) behaves as the chain (X_n^∞) would do. However the occurrence of rare events (which have very low probability) allows the chain (X_n^l) to escape from the absorbing states of the chain (X_n^∞) and to visit all the space. These rare events are caused by the three distinct mechanisms of perturbations described earlier. To ensure that these three mechanisms actually play their own role in the asymptotic dynamics of the chain, we must give the same order of intensity to the three kinds of perturbations. If one perturbation is negligible compared to another one, its asymptotic influence will be null. We will therefore focus on the situation where the three sequences u_l, v_l and w_l which express the intensity of each perturbation have asymptotics logarithmically of the same order.
In the sequel we make the choice

$$u_l = l^a, \quad v_l = l^b, \quad w_l = l^c$$

where a, b, c are three positive real numbers.

Let z_1 and z_2 be two arbitrary populations. As high as l might be, the probability for the chain (X_n^l) to travel from z_1 to z_2 is non–zero. Let r be a positive integer. We will now study the

4.5. Asymptotics of $P\left(X^l_{n+r} = z_2 / X^l_n = z_1\right)$. This transition probability may be written as

$$\sum P\big(X^l_{n+r} = z_2, V^l_{n+r-1} = v^{r-1}, U^l_{n+r-1} = u^{r-1}, X^l_{n+r-1} = x^{r-1}, \cdots,$$
$$X^l_{n+1} = x^1, V^l_n = v^0, U^l_n = u^0, X^l_n = z_1\big)$$

the sum being taken over all possible values of sequences of populations in E^m

$$u^0, \cdots, u^{r-1}, v^0, \cdots, v^{r-1} \quad \text{and} \quad x^1, \cdots, x^{r-1}.$$

Anyway the sum is finite and to estimate it we need only to take into account the terms which prevail at infinity. Since the kernel α is irreducible, for r large enough, there is always a path leading from z_1 to z_2 which involves only transition probabilities of order l^{-s} for some positive s. Therefore at least one term in the above sum is asymptotically of order l^{-t} for some positive t.
Moreover, if the path $\mathcal{P}$

$$z_1 = x^0 \to u^0 \to v^0 \to x^1 \to u^1 \to \cdots \to x^{r-1} \to u^{r-1} \to v^{r-1} \to x^r = z_2$$

contains a transition of very low probability, the corresponding term in the sum will be negligible compared to any power of $1/l$. More precisely:
If there exists an index k in $\{0, \cdots, r-1\}$ such that either $u^k \notin \mathcal{U}(x^k)$ or $v^k \notin \mathcal{V}(u^k)$, the product

$$\prod_{k=0}^{r-1} P\left(X^l_{n+1} = x^{k+1} / V^l_n = v^k\right) P\left(V^l_n = v^k / U^l_n = u^k\right) P\left(U^l_n = u^k / X^l_n = x^k\right)$$

is negligible with respect to l^{-s} for any positive s.
If there exists an index k in $\{0, \cdots, r-1\}$ such that $[x^{k+1}] \not\subset [v^k]$, the above term vanishes.
Let $D_r(z_1, z_2)$ be the set of paths

$$x^0 \to u^0 \to v^0 \to x^1 \to u^1 \to \cdots \to x^{r-1} \to u^{r-1} \to v^{r-1} \to x^r$$

satisfying

$$x^0 = z_1, \ x^r = z_2 \quad \text{and} \quad \forall k \in \{0, \cdots, r-1\} \qquad u^k \in \mathcal{U}(x^k), \ v^k \in \mathcal{V}(u^k), \ [x^{k+1}] \subset [v^k]$$

(This set may be empty for small values of r).
It follows from the preceding discussion that

$$P\left(X^l_{n+r} = z_2 / X^l_n = z_1\right) =$$
$$\sum_{D_r(z_1, z_2)} P\left(X^l_{n+r} = z_2 / V^l_{n+r-1} = v^{r-1}\right) \cdots P\left(U^l_n = u^0 / X^l_n = z_1\right) + o\left(l^{-s}\right)$$

for any positive s. In the above sum, the term associated with the path $\mathcal{P}$ of $D_r(z_1, z_2)$

$$z_1 = x^0 \to u^0 \to v^0 \to x^1 \to u^1 \to \cdots \to x^{r-1} \to u^{r-1} \to v^{r-1} \to x^r = z_2$$

is asymptotically of order l^{-t} where t is equal to

$$V(\mathcal{P}) = a \sum_{k=0}^{r-1} d(x^k, u^k) + b \sum_{k=0}^{r-1} \bar{d}(u^k, v^k) + c \sum_{k=0}^{r-1} \left(m\widehat{f}(v^k) - \sum_{h=1}^{m} f(x_h^{k+1}) \right).$$

This quantity, which will be called the cost of the path, reflects the difficulty for the process to move along the path under consideration. We need only to consider the terms with the lowest exponents. Define the function V_r on $E^m \times E^m$ by

$$V_r(z_1, z_2) = \min_{\mathcal{P} \in D_r(z_1, z_2)} V(\mathcal{P})$$

and let $D_r^*(z_1, z_2)$ be the set of those elements in $D_r(z_1, z_2)$ which realize the minimum $V_r(z_1, z_2)$. We define the length $l(\mathcal{P})$ of a path $\mathcal{P}$ as the number of transitions of the chain (X_n^l) it involves. With each path $\mathcal{P}$ we associate a constant $C(\mathcal{P})$ defined by

$$C(\mathcal{P}) = \prod_{0 \le k < l(\mathcal{P})} \alpha(x^k, u^k) \beta(u^k, v^k) (\mathrm{card}\, \hat{v}^k)^{-m} \prod_{i \in [x^{k+1}]} (v^k(i))^{x^{k+1}(i)} \tag{12}$$

and we put

$$C_r(z_1, z_2) = \sum_{\mathcal{P} \in D_r^*(z_1, z_2)} C(\mathcal{P}).$$

In particular, $D_r(z_1, z_2)$ is exactly the set of paths $\mathcal{P}$ of length r such that $C(\mathcal{P}) > 0$.
With these notations we see that:
If $D_r(z_1, z_2)$ is empty,

$$\forall s > 0 \qquad P\left(X_{n+r}^l = z_2 / X_n^l = z_1\right) = o(l^{-s}).$$

If $D_r(z_1, z_2)$ is not empty,

$$P\left(X_{n+r}^l = z_2 / X_n^l = z_1\right) = C_r(z_1, z_2)\, l^{-V_r(z_1, z_2)} + o\left(l^{-V_r(z_1, z_2)}\right).$$

Remark that this formula is valid even if z_1 and z_2 differ only by a permutation (in this case $V_r(z_1, z_2)$ may vanish).
Our next objective is to study the way the chain (X_n^l) behaves when it escapes from the set S. We first examine how the chain may travel from z_1 to z_2 staying outside S.

4.6. Asymptotics of $P\left(X^l_{n+r}=z_2,\ \forall k\in\{1,\cdots,r-1\},\ X^l_{n+k}\notin S/X^l_n=z_1\right)$.
Let $\widetilde{D}_r(z_1,z_2)$ be the set of elements belonging to $D_r(z_1,z_2)$ such that

$$\forall k\in\{1,\cdots,r-1\}\qquad x^k\notin S.$$

Analogously define the function $\widetilde{V}_r$ on $E^m\times E^m$ by

$$\widetilde{V}_r(z_1,z_2)=\min_{\widetilde{D}_r(z_1,z_2)}\ a\sum_{k=0}^{r-1}d(x^k,u^k)+b\sum_{k=0}^{r-1}\bar{d}(u^k,v^k)+c\sum_{k=0}^{r-1}\left(m\widehat{f}(v^k)-\sum_{h=1}^{m}f(x_h^{k+1})\right)$$

and let $\widetilde{D}^*_r(z_1,z_2)$ be the set of those elements in $\widetilde{D}_r(z_1,z_2)$ which realize the minimum. Put

$$\widetilde{C}_r(z_1,z_2)=\sum_{\mathcal{P}\in\widetilde{D}^*_r(z_1,z_2)}C(\mathcal{P}).$$

Using exactly the same technique as in the preceding case we get

$$P\left(X^l_{n+r}=z_2,\ \forall k\in\{1,\cdots,r-1\},\ X^l_{n+k}\notin S/X^l_n=z_1\right)=\tag{13}$$
$$\widetilde{C}_r(z_1,z_2)\,l^{-\widetilde{V}_r(z_1,z_2)}+o\left(l^{-\widetilde{V}_r(z_1,z_2)}\right)$$

whenever $\widetilde{D}_r(z_1,z_2)$ is not empty.

4.7. The minimal costs V and $\widetilde{V}$. The most probable path followed by the chain (X^l_n) between two populations z_1 and z_2 will be a path of minimal cost. We consider two minimal costs, depending on whether the path is constrained outside S or not:

$$V(z_1,z_2)\ =\ \inf_{r\in\mathbb{N}}V_r(z_1,z_2),\qquad \widetilde{V}(z_1,z_2)\ =\ \inf_{r\in\mathbb{N}}\widetilde{V}_r(z_1,z_2)$$

(We make the convention that $V_0(z_1,z_2)=\widetilde{V}_0(z_1,z_2)=\infty$ if $z_1\neq z_2$ and 0 if $z_1=z_2$). Since $\widetilde{D}_r(z_1,z_2)\subset D_r(z_1,z_2)$, then $V_r(z_1,z_2)\leq\widetilde{V}_r(z_1,z_2)$ and taking the minimum over r we obtain

$$V(z_1,z_2)\leq\widetilde{V}(z_1,z_2).$$

We first state some elementary properties of the functions V_r and $\widetilde{V}_r$.

Lemma 4.1. *Let r be a positive integer.*
For every integers k_1 and k_2 such that $1\leq k_1\leq k_2\leq r-1$,

$$V_r(z_1,z_2)=\min_{z_{k_1},z_{k_2}\in\ E^m}\left\{V_{k_1-1}(z_1,z_{k_1})+V_{k_2-k_1}(z_{k_1},z_{k_2})+V_{r-k_2+1}(z_{k_2},z_2)\right\}$$
$$\widetilde{V}_r(z_1,z_2)=\min_{z_{k_1},z_{k_2}\in E^m\setminus S}\left\{\widetilde{V}_{k_1-1}(z_1,z_{k_1})+\widetilde{V}_{k_2-k_1}(z_{k_1},z_{k_2})+\widetilde{V}_{r-k_2+1}(z_{k_2},z_2)\right\}$$

Corollary 4.1. $\forall r \in \mathbb{N}^* \quad \forall z_1, z_2 \in E^m$

$$V_r(z_1,z_2) = \min_{x^1\cdots x^{r-1}\in\ E^m} \Big\{ V_1(z_1,x^1) + V_1(x^1,x^2) + \cdots + V_1(x^{r-1},z_2) \Big\}$$
$$\widetilde{V}_r(z_1,z_2) = \min_{x^1\cdots x^{r-1}\in E^m\setminus S} \Big\{ \widetilde{V}_1(z_1,x^1) + \widetilde{V}_1(x^1,x^2) + \cdots + \widetilde{V}_1(x^{r-1},z_2) \Big\}$$

Lemma 4.2. *Let r be a positive integer. We have*

$$\begin{array}{lll}
\forall z_1 \in E^m \quad \forall z_2 \in E^m\setminus S & \widetilde{V}_r(z_1,z_2) \geq V_r(z_1,z_2) > 0 \\
\forall z_1 \in E^m \quad \forall z_2 \in E^m & V_r(z_1,z_2) = 0 \implies [z_2] \subset \hat{z}_1 \quad \text{and} \quad z_2 \in S \\
\forall z_1 \in A \quad \forall z_2 \in E^m & z_1 \neq z_2 \implies V_r(z_1,z_2) > 0
\end{array}$$

The above results are immediate consequences of the definition of the functions V_r and $\widetilde{V}_r$. The next lemma is of direct interest for V and $\widetilde{V}$:

Lemma 4.3. *Let $r^* = card\ (E^m/\mathfrak{S}_m)$. For every r strictly greater than r^* we have*

$$\inf_{k\leq r^*} V_k(z_1,z_2) \leq V_r(z_1,z_2), \qquad \inf_{k\leq r^*} \widetilde{V}_k(z_1,z_2) < \widetilde{V}_r(z_1,z_2)$$

Proof. Let r be an integer strictly greater than r^* and consider an element of $D_r(z_1,z_2)$:

$$u^0,\cdots,u^{r-1},v^0,\cdots,v^{r-1},x^0,\cdots,x^r.$$

Since $r >$ card $(E^m/\mathfrak{S}_m)$ then necessarily the sequence $x^1,\cdots,x^r$ contains two populations which are equivalent modulo $\mathfrak{S}_m$:

$$\exists\, k_1 \in \{1\cdots r\} \quad \exists\, k_2 \in \{1\cdots r\} \quad k_1 < k_2 \quad \exists\, \sigma \in \mathfrak{S}_m \qquad \sigma\cdot x^{k_1} = x^{k_2}.$$

In each of the three sequences u^k, v^k, x^k we remove the elements whose index lies in $\{k_1,\cdots,k_2-1\}$. We obtain an element of $D_{r-(k_2-k_1)}(z_1,z_2)$ whose cost is less than the original one, whence

$$V_{r-(k_2-k_1)}(z_1,z_2) \leq V_r(z_1,z_2).$$

Suppose now that the original sequence was in fact in the set $\widetilde{D}_r(z_1,z_2)$: necessarily x^{k_1} and therefore x^{k_2} were in $E^m\setminus S$; by lemma 4.2, the cost of the path between x^{k_1} and x^{k_2} is strictly positive and the cost of the new path between z_1 and z_2 (i.e. the element of $\widetilde{D}_{r-(k_2-k_1)}(z_1,z_2)$ obtained by removing the cycle $x^{k_1}\to\cdots\to x^{k_2}$) is strictly smaller than the original one, whence

$$\widetilde{V}_{r-(k_2-k_1)}(z_1,z_2) < \widetilde{V}_r(z_1,z_2).$$

An immediate descending induction on r gives the desired inequalities. $\square$

As a consequence, we have the following

Corollary 4.2. *Let $r^* = card\,(E^m/\mathfrak{S}_m)$. Then*

$$V(z_1,z_2) = \inf_{k\le r^*} V_k(z_1,z_2), \qquad \widetilde{V}(z_1,z_2) = \inf_{k\le r^*} \widetilde{V}_k(z_1,z_2).$$

Let $I(z_1,z_2)$ (respectively $\widetilde{I}(z_1,z_2)$) be the set of integers k such that $V(z_1,z_2) = V_k(z_1,z_2)$ (respectively $\widetilde{V}(z_1,z_2) = \widetilde{V}_k(z_1,z_2)$). Let $D(z_1,z_2)$ and $\widetilde{D}(z_1,z_2)$ be the set of all paths realizing the minimum in the functions V and $\widetilde{V}$:

$$D(z_1,z_2) = \bigcup_{k\in I(z_1,z_2)} D_k^*(z_1,z_2), \qquad \widetilde{D}(z_1,z_2) = \bigcup_{k\in \widetilde{I}(z_1,z_2)} \widetilde{D}_k^*(z_1,z_2).$$

Lemma 4.3 and corollary 4.2 show that these sets are non–empty; furthermore the set $D(z_1,z_2)$ contains at least a path whose length is less than r^* and the length of each path in $\widetilde{D}(z_1,z_2)$ is less than r^* (so that $\widetilde{D}(z_1,z_2)$ is finite). Finally, we put

$$C(z_1,z_2) = \sum_{\mathcal{P}\in D(z_1,z_2)} C(\mathcal{P}), \qquad \widetilde{C}(z_1,z_2) = \sum_{\mathcal{P}\in \widetilde{D}(z_1,z_2)} C(\mathcal{P}).$$

5. Entrance in S

We are now in position to estimate the transition probabilities of the chain (Z_n^l).
Let y be in E^m and z in S. A straightforward application of the Markov property to the chain (Z_n^l) shows that

$$P\left(Z_{n+1}^l = z / Z_n^l = y\right) = P\left(X_{\tau^l}^l = z / X_0^l = y\right) = P_y\left(X_{\tau^l}^l = z\right).$$

We decompose this quantity according to the possible values of τ^l:

$$P_y\left(X_{\tau^l}^l = z\right) = \sum_{k=1}^{q} P_y\left(X_{\tau^l}^l = z,\, \tau^l = k\right) + P_y\left(X_{\tau^l}^l = z,\, \tau^l > q\right). \tag{14}$$

However we have already shown in (8) and (10) that

$$P_y\left(\tau^l > q\right) \le (1-(s_l)^m)^q \le m^q(1-s_l)^q \le m^q(m-1)^q\, l^{-c\delta q}.$$

We choose the integer q large enough to ensure that $q > r^*$ and $c\delta q > \widetilde{V}(y,z)$. Since

$$\begin{aligned} P_y\left(X_{\tau^l}^l = z,\, \tau^l = k\right) &= P_y\left(X_k^l = z,\, \forall h\in\{1\cdots k-1\},\, X_h^l \notin S\right) \\ &= \widetilde{C}_k(y,z)\, l^{-\widetilde{V}_k(y,z)} + o\left(l^{-\widetilde{V}_k(y,z)}\right) \end{aligned}$$

then for each fixed integer q greater than r^*,

$$\sum_{k=1}^{q} P_y\left(X^l_{\tau^l} = z,\, \tau^l = k\right) \underset{l\to\infty}{\sim} \widetilde{C}(y,z)\, l^{-\widetilde{V}(y,z)}.$$

In addition, the second term in the right–hand side of (14) is dominated by $l^{-c\delta q}$ which is negligible compared to $l^{-\widetilde{V}(y,z)}$ whence finally

$$P\left(Z^l_{n+1} = z / Z^l_n = y\right) = P_y\left(X^l_{\tau^l} = z\right) \underset{l\to\infty}{\sim} \widetilde{C}(y,z)\, l^{-\widetilde{V}(y,z)}. \tag{15}$$

Now, to evaluate the stationary measure of the chain (Z^l_n) we need

6. Some results from the Freidlin–Wentzell theory

In this section we restate word for word some key results from the Freidlin–Wentzell theory. This material is extracted from [1], chapter 6, Lemmas on Markov Chains.

Let H be a finite set and let a subset W be selected in H. A graph consisting of arrows $i \to j\,(i \in H \setminus W, j \in H, i \neq j)$ is called a W–graph if it satisfies the following conditions:

(1) every point $i \in H \setminus W$ is the initial point of exactly one arrow;

(2) there are no closed cycles in the graph.

We note that condition (2) can be replaced by the following condition:

(2′) for any point $i \in H \setminus W$ there exists a sequence of arrows leading from i to some point $j \in W$.

We denote by $G(W)$ the set of W–graphs; we shall use the letter g to denote graphs. If $p_{ij}\,(i,j \in H, j \neq i)$ are numbers, then $\prod_{(i\to j)\in g} p_{ij}$ will be denoted by $\pi(g)$.

Lemma 6.1. *Let us consider a Markov chain with set of states H and transition probabilities p_{ij} and assume that every state can be reached from any other state in a finite number of steps. Then the stationary distribution of the chain is $\{(\sum_{i\in H} Q_i)^{-1} Q_i,\ i \in H\}$, where*

$$Q_i = \sum_{g \in G\{i\}} \pi(g). \tag{16}$$

Lemma 6.2. *Let us be given a Markov chain on a phase space X divided into disjoint sets X_i, where i runs over a finite set H. Suppose that there exist non–negative numbers $p_{ij}\,(j \neq i, i, j \in H)$ and a number $a > 1$ such that*

$$a^{-1} p_{ij} \le P(x, X_j) \le a p_{ij} \qquad (x \in X_i, i \neq j) \tag{17}$$

for the transition probabilities of our chain. Furthermore, suppose that every set X_j can be reached from any state x sooner or later (for this it is necessary and sufficient that for any j there exist a $\{j\}$–graph g such that $\pi(g) > 0$). Then

$$a^{2-2h}\left(\sum_{i\in H} Q_i\right)^{-1} Q_i \ \le\ \nu(X_i) \ \le\ a^{2h-2}\left(\sum_{i\in H} Q_i\right)^{-1} Q_i \tag{18}$$

for any normalized invariant measure ν of our chain, where h is the number of elements in H and the Q_i are defined by formula (16).

7. Convergence of the measure ν^l

Put for y and z in S

$$p_{yz}^l = \widetilde{C}(y,z)\, l^{-\widetilde{V}(y,z)}.$$

Since S is finite, from the estimation (15) we deduce that for any positive ϵ there exists an integer L such that

$$\forall\, l \geq L \quad \forall y, z \in S \qquad (1+\epsilon)^{-1}\, p_{yz}^l \leq P\left(Z_{n+1}^l = z / Z_n^l = y\right) \leq (1+\epsilon)\, p_{yz}^l.$$

Put for x in S

$$Q_x^l = \sum_{g \in G\{x\}} \pi^l(g)$$

where $G\{x\}$ denotes the set of x–graphs over S and for g in $G\{x\}$,

$$\pi^l(g) = \prod_{(y \to z) \in g} p_{yz}^l.$$

Application of lemma 6.2 to the chain (Z_n^l) yields for $l \geq L$

$$(19) \qquad (1+\epsilon)^{2-2|S|} \frac{Q_x^l}{\sum_{y \in S} Q_y^l} \leq \nu^l(x) \leq (1+\epsilon)^{2|S|-2} \frac{Q_x^l}{\sum_{y \in S} Q_y^l}.$$

We study now the asymptotic behavior of Q_x^l. For g in $G\{x\}$,

$$\pi^l(g) = \Bigg(\prod_{(y \to z) \in g} \widetilde{C}(y,z) \Bigg) \exp \Bigg(- \sum_{(y \to z) \in g} \widetilde{V}(y,z) \ln l \Bigg)$$

so that the crucial quantity for Q_x^l is

$$W(x) = \min_{g \in G\{x\}} \sum_{(y \to z) \in g} \widetilde{V}(y,z).$$

Let $G^*(x)$ be the set of x–graphs which realize the above minimum.
If g belongs to $G^*(x)$, $\pi^l(g)$ is asymptotically of order $l^{-W(x)}$.
If g does not belong to $G^*(x)$, $\pi^l(g)$ is negligible compared to $l^{-W(x)}$.
Finally

$$Q_x^l \underset{l \to \infty}{\sim} \Bigg(\sum_{g \in G^*(x)} \prod_{(y \to z) \in g} \widetilde{C}(y,z) \Bigg)\, l^{-W(x)}.$$

Put

$$\widetilde{D}(x) = \sum_{g\in G^*(x)} \prod_{(y\to z)\in g} \widetilde{C}(y,z)$$

and let

$$W^* = \left\{\, x\in S : W(x) = \min_{y\in S} W(y) \,\right\}.$$

By $W(W^*)$ we mean the value $\min_{y\in S} W(y)$. With these notations, we have

$$\sum_{y\in S} Q^l_y \underset{l\to\infty}{\sim} \Big(\sum_{x\in W^*} \widetilde{D}(x)\Big)\, l^{-W(W^*)}$$

and

$$\frac{Q^l_x}{\sum_{y\in S} Q^l_y} \underset{l\to\infty}{\sim} \frac{\widetilde{D}(x)}{\sum_{y\in W^*} \widetilde{D}(y)}\, l^{W(W^*)-W(x)}.$$

Thus there exists an integer L' such that for $l\geq L'$,

$$(1+\epsilon)^{-1}\frac{\widetilde{D}(x)}{\sum_{y\in W^*} \widetilde{D}(y)}\, l^{W(W^*)-W(x)} \leq \frac{Q^l_x}{\sum_{y\in S} Q^l_y} \leq (1+\epsilon)\frac{\widetilde{D}(x)}{\sum_{y\in W^*} \widetilde{D}(y)}\, l^{W(W^*)-W(x)}$$

whence for $l\geq \max(L,L')$, by (19),

$$(1+\epsilon)^{1-2|S|}\frac{\widetilde{D}(x)}{\sum_{y\in W^*} \widetilde{D}(y)}\, l^{W(W^*)-W(x)} \leq \nu^l(x) \leq (1+\epsilon)^{2|S|-1}\frac{\widetilde{D}(x)}{\sum_{y\in W^*} \widetilde{D}(y)}\, l^{W(W^*)-W(x)}$$

We conclude that:
If x does not belong to W^* then $\lim_{l\to\infty}\nu^l(x)=0$.
If x belongs to W^* then

$$\lim_{l\to\infty}\nu^l(x) = \frac{\widetilde{D}(x)}{\sum_{y\in W^*} \widetilde{D}(y)} = \frac{\sum_{g\in G^*(x)}\prod_{(y\to z)\in g}\widetilde{C}(y,z)}{\sum_{y\in W^*}\sum_{g\in G^*(y)}\prod_{(z_1\to z_2)\in g}\widetilde{C}(z_1,z_2)}$$

We denote by ν^∞ this limit measure. The complete formula for ν^∞ is

$$\nu^\infty(x) = \frac{\sum_{g\in G^*(x)}\prod_{(y\to z)\in g}\sum_{\mathcal{P}\in\widetilde{D}(y,z)} C(\mathcal{P})}{\sum_{y\in W^*}\sum_{g\in G^*(y)}\prod_{(z_1\to z_2)\in g}\sum_{\mathcal{P}\in\widetilde{D}(z_1,z_2)} C(\mathcal{P})}$$

where $C(\mathcal{P})$ is defined by (12). The measure ν^∞ is concentrated on the set W^* and charges all points of W^*. The crucial question now is: what is the set W^*?

8. The function W and the uniform populations

We derive some properties of the functional W. First, W may equivalently be defined through V instead of $\widetilde{V}$:

Lemma 8.1. *The following equality holds for all x in E^m:*

$$W(x) = \min_{g\in G(x)} \sum_{(y\to z)\in g} \widetilde{V}(y,z) = \min_{g\in G(x)} \sum_{(y\to z)\in g} V(y,z). \tag{20}$$

Proof. The proof of this lemma can be found in [1, chapter 6, lemma 4.1]. □

The perturbed chain (X_n^l) is attracted by the populations of the set S: these populations play a role analogous to the compacta K_i's (the ω–limit sets of the trajectories of the solutions of the deterministic system) in the work of Freidlin and Wentzell. They distinguish two kinds of compacta according to their stability properties. In our situation, the populations in A will play the role of the stable compacta whereas the populations in $S\setminus A$ will behave as the unstable compacta.

Lemma 8.2. *If x belongs to $S\setminus A$ then for each i in $[x]$,*

$$V(x,(i)) = \widetilde{V}(x,(i)) = 0.$$

Proof. This result is straightforward and similar to lemma 4.2 of [1].

We restate now textually lemma 4.3 of [1].

Lemma 8.3. *a) Among the x–graphs for which the minimum in $W(x)$ is attained there is one in which from the index y, $y\neq x$ of each population in $S\setminus A$ an arrow $y\to z$ is issued with $V(y,z)=0$ and z in A.*
b) For a population x in A, the value $W(x)$ can be calculated according to (20), considering graphs on the set of populations in A.
c) If x is a population in $S\setminus A$ then

$$W(x) = \min_{y\in A}\left(W(y)+V(y,x)\right).$$

Proof. The proof of this lemma is only a rewriting of the proof of lemma 4.3 of [1].

As a consequence, the set W^* is included in the set of uniform populations A (lemmas 4.2 and 4.3 imply that for y in A and $x\neq y$, $V(y,x)>0$).

9. Some more notations

For λ in $\mathbb{R}^+$ we define

$$f_\lambda = f^{-1}(\{\lambda\}), \quad f_\lambda^+ = f^{-1}(]\lambda,\infty[), \quad f_\lambda^- = f^{-1}([0,\lambda[).$$

Since the study of W^* requires only consideration of uniform populations, we will often write i instead of (i). For instance, $V(i,j)$ stands for $V((i),(j))$.
For any graph g over A we define

$$V(g) = \sum_{(i\to j)\in g} V(i,j).$$

If X and Y are two subsets of S, we denote by $G_X(Y)$ the set of Y–graphs over $X \cup Y$. For instance, we have $G(X) = G_S(X) = G_{S\setminus X}(X)$. We put

$$W_X(Y) = \min_{g\in G_X(Y)} \sum_{(i_1\to i_2)\in g} V(i_1,i_2) = \min_{g\in G_X(Y)} V(g).$$

We note $G_X^*(Y)$ the set of graphs in $G_X(Y)$ which realize this minimum.
We define $G_X(\emptyset)$ as the union of all x–graphs over X for all x in X:

$$G_X(\emptyset) = \bigcup_{x\in X} G_X\{x\}.$$

If g is a graph over E, its restriction $g_{|\lambda}$ to the level λ is the graph

$$g_{|\lambda} = \{\,(i_1,i_2) \in g : i_1 \in f_\lambda\}.$$

10. Sufficient conditions to ensure $W^* \subset f^*$

Theorem 10.1. *If the inequality*

$$\sum_{\lambda\in f(E)} W_{f_\lambda}(f_\lambda^+) < \sum_{\lambda\in f(E)} W_{f_\lambda}(f_\lambda^+ \cup f_\lambda^-) - \max_{i\in E\setminus f^*} \min_{j:f(j)\neq f(i)} V(i,j) \tag{21}$$

is satisfied, then the set W^* of all points which may be settled ultimately by the homogeneous algorithm with positive probability is included in the set f^* of the global maxima of the fitness function i.e.

$$\forall x \in E^m \qquad \lim_{l\to\infty} \lim_{n\to\infty} \quad P\left([X_n^l] \subset f^*/X_0^l = x\right) = 1.$$

Proof. Let g be a graph over E. We may decompose the sum $V(g)$ in the following way:

$$V(g) = \sum_{\lambda \in f(E)} \sum_{\substack{i_1 \in f_\lambda \\ (i_1 \to i_2) \in g}} V(i_1, i_2) = \sum_{\lambda \in f(E)} \sum_{(i_1 \to i_2) \in g_{|\lambda}} V(i_1, i_2)$$

i.e.

$$V(g) = \sum_{\lambda \in f(E)} V(g_{|\lambda}).$$

Suppose now that g is in $G\{i\}$ for some i in E. Put $\theta = f(i)$.
Then $g_{|\lambda}$ belongs to $G_{f_\lambda}(f_\lambda^+ \cup f_\lambda^-)$ whenever $\lambda \neq \theta$ whence

$$V(g_{|\lambda}) \geq W_{f_\lambda}(f_\lambda^+ \cup f_\lambda^-). \tag{22}$$

Let j be any point outside f_θ; then $g_{|\theta} \cup \{(i \to j)\}$ is a graph of $G_{f_\theta}(f_\theta^+ \cup f_\theta^-)$ whence

$$V(g_{|\theta}) \geq W_{f_\theta}(f_\theta^+ \cup f_\theta^-) - V(i, j)$$

and the inequality being valid for all j outside f_θ, we have

$$V(g_{|\theta}) \geq W_{f_\theta}(f_\theta^+ \cup f_\theta^-) - \min_{j : f(j) \neq f(i)} V(i, j). \tag{23}$$

Summing up inequalities (22) and (23) yields

$$V(g) \geq \sum_{\lambda \in f(E)} W_{f_\lambda}(f_\lambda^+ \cup f_\lambda^-) - \min_{j : f(j) \neq f(i)} V(i, j)$$

for any graph g in $G\{i\}$. Taking the minimum over all g in $G\{i\}$ we obtain

$$W(i) \geq \sum_{\lambda \in f(E)} W_{f_\lambda}(f_\lambda^+ \cup f_\lambda^-) - \min_{j : f(j) \neq f(i)} V(i, j)$$

so that finally

$$\min_{i \in E \setminus f^*} W(i) \geq \sum_{\lambda \in f(E)} W_{f_\lambda}(f_\lambda^+ \cup f_\lambda^-) - \max_{i \in E \setminus f^*} \min_{j : f(j) \neq f(i)} V(i, j). \tag{24}$$

We build now a near–optimal graph g.
For each λ in $f(E)$, select a graph g_λ in $G^*_{f_\lambda}(f^+_\lambda)$. If $\lambda = f(f^*)$, g_λ is an i^*–graph over f^* for some i^* in f^* (in this case, f^+_λ is empty).
We define the graph g as the union of the graphs g_λ:

$$(i_1 \to i_2) \in g \iff \exists\, \lambda \in f(E) \quad (i_1 \to i_2) \in g_\lambda.$$

It is easy to see that g is in $G\{i^*\}$. Furthermore we have by construction

$$\forall \lambda \in f(E) \qquad g_{|\lambda} = g_\lambda$$

whence

$$V(g) = \sum_{\lambda \in f(E)} V(g_{|\lambda}) = \sum_{\lambda \in f(E)} W_{f_\lambda}(f^+_\lambda)$$

from which we deduce, for the point i^* of f^*,

$$W(i^*) \leq \sum_{\lambda \in f(E)} W_{f_\lambda}(f^+_\lambda). \tag{25}$$

Putting together the inequalities (21), (24) and (25), we obtain

$$W(i^*) < \min_{i \in E \setminus f^*} W(i)$$

where i^* is a point of f^*, which clearly implies $W^* \subset f^*$. $\square$

The condition (21) depends strongly on the structure of the optimization problem: the essential ingredients which are hidden behind the functional W and V are the function f and the kernels α and β which determine the asymptotic dynamics of the chain (X^l_n).
We derive now some stronger but more explicit conditions for the inequality (21) to hold. Clearly it is sufficient to have both

$$\forall \lambda \in f(E) \setminus f(f^*) \qquad W_{f_\lambda}(f^+_\lambda) \leq W_{f_\lambda}(f^+_\lambda \cup f^-_\lambda) \tag{26}$$

and

$$W_{f^*}(\emptyset) < W_{f^*}(E \setminus f^*) - \max_{i \in E \setminus f^*} \; \min_{j : f(j) \neq f(i)} V(i,j). \tag{27}$$

We next prove two fundamental lemmas describing the behavior of the function V with respect to the size of the population m.

Lemma 11.1. *Let R be the smallest integer such that*

$$\forall i,j\in E\quad \exists\, r\le R\quad \exists i_0,\cdots,i_r\in E\quad \text{such that}$$
$$i_0=i,\ i_r=j,\quad \forall k\in\{0,\cdots,r-1\}\quad \alpha(i_k,i_{k+1})>0$$

i.e. R is the minimal number of transitions necessary to join two arbitrary points of E through the kernel α (since α is irreducible and E is finite, R is finite). Put

$$\Delta^\circ=\max_{i\in E\setminus f^*}\min\Big\{\max_{0\le k<r}(f(i)-f(i_k)):i_0=i,\ r\le R,\ f(i_r)>f(i),\ \prod_{0\le k<r}\alpha(i_k,i_{k+1})>0\Big\}$$
$$\Delta^*=\max_{i,j\in f^*}\min\Big\{\max_{0\le k<r}(f(i)-f(i_k)):i_0=i,\ r\le R,\ i_r=j,\ \prod_{0\le k<r}\alpha(i_k,i_{k+1})>0\Big\}$$

Let i be any point in $E\setminus f^*$. With these notations we have

$$\forall m\in\mathbb{N}^*\qquad \min_{j\in E,f(j)>f(i)}V(i,j)\le aR+c(R-1)\Delta^\circ. \tag{28}$$

Let i,j be two points in f^*. We have also

$$\forall m\in\mathbb{N}^*\qquad V(i,j)\le aR+c(R-1)\Delta^*. \tag{29}$$

Corollary 11.1. *Let $\Delta^\circledast=\max(\Delta^\circ,\Delta^*)$. Then,*

$$\sup_{m\in\mathbb{N}^*}\Big\{\max_{i\in E\setminus f^*}\ \min_{j\in E,f(j)>f(i)}V(i,j),\quad \max_{i,j\in f^*}\ V(i,j)\Big\}\ \le\ aR+c(R-1)\Delta^\circledast, \tag{30}$$

$$\sup_{m\in\mathbb{N}^*}\quad\max_{y\in S\setminus(f^*)^m}\quad\min_{z\in S,\widehat{f}(z)>\widehat{f}(y)}\quad V(y,z)\ \le\ aR+c(R-1)\Delta^\circ. \tag{31}$$

Remark. Putting

$$\Delta=\max\{\,|f(i)-f(j)|:i,j\in E\,\},$$

we have clearly $\Delta^\circledast=\max(\Delta^\circ,\Delta^*)\le\Delta$. All the right–hand side members of the preceding inequalities are thus smaller than $(a+c\Delta)|E|$.

Proof of lemma. Let i be a point of $E\setminus f^*$. By the very definition of Δ°, there exists a sequence $i_0,\cdots,i_r$ of points of E such that:

$$i_0=i,\ r\leq R,\ f(i_r)>f(i),\ \prod_{0\leq k<r}\alpha(i_k,i_{k+1})>0\qquad\text{and}\qquad\max_{0\leq k<r}(f(i)-f(i_k))\leq\Delta^\circ.$$

Among these sequences, we choose a sequence of minimal length, so that necessarily

$$\forall k\in\{0,\cdots,r-1\}\qquad f(i_k)\leq f(i).$$

We put $j=i_r$ and we define

$$\begin{gathered}x^0=(i_0),\ x^1=(i,\cdots,i,i_1),\ \cdots,\ x^{r-1}=(i,\cdots,i,i_{r-1}),\ x^r=(j),\\ u^0=v^0=x^1,\cdots,u^{r-2}=v^{r-2}=x^{r-1},\\ u^{r-1}=v^{r-1}=(i,\cdots,i,i_r).\end{gathered}$$

The path

$$x^0\to u^0\to v^0\to x^1\to u^1\to\cdots\to v^{r-2}\to x^{r-1}\to u^{r-1}\to v^{r-1}\to x^r$$

clearly belongs to $D_r((i),(j))$ and its cost is $ar+c\sum_{k=0}^{r-2}(f(i)-f(i_{k+1}))$. It follows that

$$V(i,j)\leq V_r(i,j)\leq ar+c\sum_{k=0}^{r-2}\left[f(i)-f(i_{k+1})\right].$$

from which we obtain immediately

$$V(i,j)\leq aR+c(R-1)\Delta^\circ$$

and since $f(j)>f(i)$, inequality (28) is proved.
Let i,j be two points in f^*. By the very definition of Δ^*, there exists a sequence $i_0,\cdots,i_r$ leading from i to j through the kernel α such that $r\leq R$ and

$$\max_{0\leq k<r}(f(i)-f(i_k))\leq\Delta^*.$$

We build the sequences u^k,v^k,x^k as above and thus obtain a similar inequality for $V(i,j)$ (with Δ^* instead of Δ°), which proves inequality (29). $\square$

Proof of corollary. Inequality (30) is a direct consequence of lemma 11.1. To prove inequality (31), consider an element y of $S\setminus(f^*)^m$. Pick up a point i in $\hat{y}$. Using inequality (28), we see that there exists a point j in E such that $f(j)>f(i)$ and $V(i,j)\leq aR+c(R-1)\Delta^\circ$. We build as above a path of length smaller than R which transforms by successive mutations the individual i in j and leaves unchanged all others individuals of y and obtain

$$V(y,(j))\leq aR+c(R-1)\Delta^\circ$$

whence

$$\min_{j\in E,f(j)>\widehat{f}(y)}V(y,(j))\leq aR+c(R-1)\Delta^\circ.$$

Taking the maximum over y in $S\setminus(f^*)^m$ and then over m in $\mathbb{N}^*$ yields inequality (31). $\square$

12. Decreasing of fitness

Lemma 12.1. *Let $\rho = \min(a, b/2, c\delta)$ where δ is defined by (9). Then*

$$(32) \qquad \forall\, z^1 \in E^m \quad \forall\, z^2 \in E^m \qquad V(z^1, z^2) \geq \rho\, \mathrm{card}\{\, k : 1 \leq k \leq m \quad f(z^1_k) > \widehat{f}(z^2)\,\}$$

Remark. The above inequality, being true for the function V, is a fortiori true for the function $\widetilde{V}$.

Proof. We show by induction on r that V_r satisfies this inequality.
Consider first the case $r = 0$. Either $z^1 = z^2$ and the right–hand member vanishes or $z^1 \neq z^2$ and $V_0(z^1, z^2)$ is infinite.
Suppose the result is true at rank r.
Let u^k, v^k, x^k be an element of $D_{r+1}(z^1, z^2)$. We consider two situations:

- If $\widehat{f}(v^0) \leq \widehat{f}(z^2)$ then all elements of the set

$$\{\, z^1_k : 1 \leq k \leq m,\ f(z^1_k) > \widehat{f}(z^2)\,\}$$

have been destroyed during the transition between z^1 and v^0 whence

$$d(z^1, u^0) + 2\bar{d}(u^0, v^0) \geq \mathrm{card}\{\, k : 1 \leq k \leq m,\ f(z^1_k) > \widehat{f}(z^2)\,\}.$$

- If $\widehat{f}(v^0) > \widehat{f}(z^2)$ then putting

$$h = \mathrm{card}\{\, k : 1 \leq k \leq m,\ \widehat{f}(v^0) > f(x^1_k)\,\}$$

we have

$$\mathrm{card}\{\, k : 1 \leq k \leq m,\ f(x^1_k) > \widehat{f}(z^2)\,\} \geq m - h$$

and

$$c\Big(m\widehat{f}(v^0) - \sum_{k=1}^{m} f(x^1_k)\Big) \geq ch\delta \geq \rho h.$$

The induction hypothesis at rank r implies

$$V_r(x^1, z^2) \geq \rho(m - h).$$

Now

$$a\sum_{k=0}^{r} d(x^k, u^k) + b\sum_{k=0}^{r} \bar{d}(u^k, v^k) + c\sum_{k=0}^{r}\Big(m\widehat{f}(v^k) - \sum_{h=1}^{m} f(x^{k+1}_h)\Big)$$
$$\geq c\Big(m\widehat{f}(v^0) - \sum_{k=1}^{m} f(x^1_k)\Big) + V_r(x^1, z^2) \geq \rho h + \rho(m - h) = \rho m$$

and the desired inequality is true in both cases for the path under consideration. Taking the minimum over all elements of $D_{r+1}(z^1, z^2)$ yields the result at rank $r + 1$ and the induction is completed. $\square$

13. Our last condition

Theorem 13.1. *Let m be an integer such that*

$$\rho m > \max_{i \in E \setminus f^*} \min_{j \in E, f(j) > f(i)} V(i,j) , \tag{33}$$

$$\rho m \geq \max_{i,j \in f^*} V(i,j). \tag{34}$$

For this integer m we have

$$\lim_{l \to \infty} \mu^l(f^*) = \lim_{l \to \infty} \nu^l(f^*) = 1.$$

Remark. Inequality (30) shows that such an integer m always exists: the quantities on the right–hand side are bounded by $(a + c\Delta)|E|$ which is a constant independent of m. In particular, the conditions are fulfilled as soon as $m > (aR + c(R-1)\Delta^{\circledast})/\min(a, b/2, c\delta)$.

Proof. We prove that for such an m the preceding sufficient conditions (26) and (27) are satisfied. Let λ be in $f(E) \setminus f(f^*)$ and let g be a graph in $G_{f_\lambda}(f_\lambda^+ \cup f_\lambda^-)$. This graph may contain a finite number of transitions from f_λ to f_λ^- : $i_1 \to i'_1, \cdots, i_r \to i'_r$. The first inequality (33) implies that for each i_k, $1 \leq k \leq r$, there exists a state j_k in f_λ^+ such that $V(i_k, j_k) < \rho m$ and inequality (32) of lemma 12.1 implies $V(i_k, j_k) < V(i_k, i'_k)$. Let g' be the graph obtained by replacing the r arrows $i_1 \to i'_1, \cdots, i_r \to i'_r$ by $i_1 \to j_1, \cdots, i_r \to j_r$. The graph g' is in $G_{f_\lambda}(f_\lambda^+)$ and $V(g') \leq V(g)$. This construction being valid for any graph in $G_{f_\lambda}(f_\lambda^+ \cup f_\lambda^-)$, we have

$$W_{f_\lambda}(f_\lambda^+) \leq W_{f_\lambda}(f_\lambda^+ \cup f_\lambda^-)$$

for all λ in $f(E) \setminus f(f^*)$ and the first inequality (26) is proved.
Now let g be a graph in $G_{f^*}(E \setminus f^*)$. This graph contains a finite number of transitions from f^* to $E \setminus f^*$: $i_1 \to j_1, \cdots, i_r \to j_r$. Let g' be the graph obtained from g by replacing the arrows $i_2 \to j_2, \cdots, i_r \to j_r$ by the $(r-1)$ arrows $i_2 \to i_1, \cdots, i_r \to i_1$. Inequality (34) implies that $V(g') \leq V(g)$. In addition the graph g' is still in $G_{f^*}(E \setminus f^*)$ and may be decomposed as the union of a graph of $G_{f^*}(\emptyset)$ and the arrow $i_1 \to j_1$ whence

$$V(g) \geq V(g') \geq W_{f^*}(\emptyset) + V(i_1, j_1) \geq W_{f^*}(\emptyset) + \rho m$$

Taking the minimum over all graphs in $G_{f^*}(E \setminus f^*)$ yields

$$W_{f^*}(E \setminus f^*) \geq W_{f^*}(\emptyset) + \rho m.$$

Since by (33)

$$\rho m > \max_{i \in E \setminus f^*} \min_{j \in E, f(j) > f(i)} V(i,j) \geq \max_{i \in E \setminus f^*} \min_{j \in E, f(j) \neq f(i)} V(i,j)$$

then

$$W_{f^*}(E \setminus f^*) > W_{f^*}(\emptyset) + \max_{i \in E \setminus f^*} \min_{j \in E, f(j) \neq f(i)} V(i,j)$$

and the second inequality (27) is proved. $\square$

14. Existence of a critical population size m^*

Our fundamental result may be stated as follows.

Theorem 14.1. *Fix the space E, the fitness function f, the mutation and crossover kernels α and β and the positive constants a, b, c.*
There exists a critical population size m^* depending upon all these objects such that, when m is greater than m^*, the set W^* of the minima of the virtual energy W is included in the set f^* of the global maxima of the fitness function f. Moreover we have

$$m^* \;\leq\; \frac{aR + c(R-1)\Delta^{\circledast}}{\min\big(a, b/2, c\delta\big)}.$$

Clearly, the algorithm can't possibly work with one single individual: when $m = 1$, the selection operator does nothing at all and the fitness function f does not intervene in the transition mechanism.
Perhaps two individuals is enough to solve all the optimization problems, i.e., $m^* = 2$? The answer is no. We give now examples where the critical size m^* is arbitrarily large. Consider the space $E = \{\,0, 1, \cdots, N\,\}$ and let f be a function defined on E, with values in $\mathbb{R}_+^*$ such that

$$f(1) = f(2) = \cdots = f(N-1), \quad f(0) = f(1) + \gamma, \quad f(N) = f(N-1) + \Gamma,$$

where $0 < \gamma < \Gamma$. Let α be a markovian mutation kernel on E such that

$$\forall i \in \{1 \cdots N-1\} \qquad \alpha(i,i-1) + \alpha(i,i+1) + \alpha(i,i) = 1, \quad \alpha(i,i-1) > 0, \quad \alpha(i,i+1) > 0,$$

(the possible mutations from i are $i \to i-1$ and $i \to i+1$), and for the points 0 and N,

$$\alpha(0,1) + \alpha(0,0) = 1, \quad \alpha(0,1) > 0, \quad \alpha(N,N-1) + \alpha(N,N) = 1, \quad \alpha(N,N-1) > 0.$$

For the crossover kernel β, we choose the identity matrix over $E \times E$. We have

$$\forall i \in \{1 \cdots N-1\} \qquad V(i,i-1) = V(i,i+1) = a,$$

(these transitions need only the mutation of one individual: the remainder of the population, driven by the process (X_n^∞), then comes on the new point)

$$V(0,1) = \min\big(am, a + cm\gamma\big), \quad V(N,N-1) = \min\big(am, a + cm\Gamma\big).$$

The transition from 0 to 1, for instance, takes place either trough a general mutation, of cost am, or by a general anti-selection after an individual has mutated to 1, of cost $a + cm\gamma$. Finally

$$V(0,N) = \min\big(am + a(N-1), a + cm\gamma + a(N-1), aN + c(N-1)\gamma\big).$$

There are two possible trajectories to go from 0 to N; either the whole population go from 0 to 1 (cost $V(0,1)$) and then from 1 to N (cost $a(N-1)$), or an explorer starts alone from 0 and mutates until the point N, while the remainder of the population waits in 0 (cost $aN + c(N-1)\gamma$): when the explorer reaches the point N, the whole population throw themselves on N, since $\gamma < \Gamma$.
Similarly

$$V(N,0) = \min\big(am + a(N-1), a + cm\Gamma + a(N-1), aN + c(N-1)\Gamma + cm(\Gamma-\gamma)\big).$$

Notice yet a crucial difference for the passage from N to 0: for the second kind of trajectories, whenever the explorer reaches 0, it is necessary to perform a general anti–selection of cost $cm(\Gamma-\gamma)$ to bring everyone in 0. In particular, we check that

$$\lim_{m\to\infty} V(0,N) \;=\; aN + c(N-1)\gamma \;<\; \infty$$

and

$$\lim_{m\to\infty} \frac{V(N,0)}{m} \;=\; \min\big(a, c(\Gamma-\gamma)\big) \;>\; 0.$$

We evaluate now the virtual energy W.
For i in $\{1\cdots N-1\}$, the i–graph of minimal cost is

$$0 \to 1 \cdots \to i-1 \to i \leftarrow i+1 \leftarrow \cdots \leftarrow N-1 \leftarrow N$$

so that

$$W(i) \;=\; a(N-2) \;+\; V(0,1) \;+\; V(N,N-1).$$

For the point 0, there are two possible 0–graphs:

$$N \to N-1 \to \cdots \to 1 \to 0 \quad \text{et} \quad N-1 \to \cdots \to 1 \to 0 \leftarrow N$$

whence

$$\begin{aligned} W(0) \;&=\; a(N-1) \;+\; \min\big(V(N,N-1), V(N,0)\big) \\ &=\; a(N-1) \;+\; \min\big(am, a+cm\Gamma, aN + c(N-1)\Gamma + cm(\Gamma-\gamma)\big). \end{aligned}$$

Similarly, we have

$$W(N) \;=\; a(N-1) + \min\big(am, a + cm\gamma, aN + c(N-1)\gamma\big).$$

Since $N > 1$ and $V(0,1) > a$, $V(N,N-1) > a$, for all i in $\{1\cdots N-1\}$, we have $W(0) < W(i)$, $W(N) < W(i)$, so that $W^* \subset \{0,N\}$. Suppose

$$am \le a + cm\gamma \quad \text{et} \quad am \le aN + c(N-1)\gamma.$$

For these inequalities to hold, it is enough that $a \leq c\gamma$ and $m \leq N$.
In this situation

$$W(0) = W(N) = a(N-1) + am$$

and the set $W^* = \{0, N\}$ is not included in $f^* = \{N\}$.
Thus, whenever $a \leq c\gamma$, we have $m^* > N$.
Let us try to explain this phenomenon: the exits from the attraction basins of 0 and N take place with a general mutation, of cost am, which brings the whole population in 1 and $N-1$. The heights γ and Γ do not intervene and the dynamics can't discriminate between the points 0 and N. The limiting law is then concentrated on both points $\{0, N\}$.

Whenever the mutation kernel α is not symmetric, we may in addition build examples where $f^* \cap W^* = \emptyset$ whenever $m < m^*$.
Consider the space $E = \{\, 0, 1, \cdots, N, N+1, \cdots, N+M-1 \,\}$ and let $f : E \mapsto \mathbb{R}_+^*$ be such that

$$f(1) = f(2) = \cdots = f(N-1) = f(N+1) = \cdots = f(N+M-1),$$
$$f(0) = f(1) + \gamma, \quad f(N) = f(N-1) + \Gamma,$$

where $0 < \gamma < \Gamma$. Let α be a markovian mutation kernel satisfying

$$\forall i \in \{0, \cdots, N+M-2\} \qquad \alpha(i, i+1) \,=\, 1$$

and $\alpha(N+M-1, 0) = 1$. We have

$$\begin{aligned} V(N,0) \,&=\, \min\big(am + a(M-1), a + cm\Gamma + a(M-1), aM + c(M-1)\Gamma + cm(\Gamma-\gamma)\big), \\ V(0,N) \,&=\, \min\big(am + a(N-1), a + cm\gamma + a(N-1), aN + c(N-1)\gamma\big), \end{aligned}$$

and

$$\begin{aligned} W(0) \,&=\, a(N+M-2) + \min\big(am, a + cm\Gamma, aM + c(M-1)\Gamma + cm(\Gamma-\gamma)\big), \\ W(N) \,&=\, a(N+M-2) + \min\big(am, a + cm\gamma, aN + c(N-1)\gamma\big). \end{aligned}$$

To have $W(0) < W(N)$ (and $W^* = \{0\}$), it is enough that

$$aM + c(M-1)\Gamma + cm(\Gamma-\gamma) \,<\, \min\big(am, a + cm\gamma, aN + c(N-1)\gamma\big).$$

Equivalently

$$\begin{aligned} aM + c(M-1)\Gamma \,&<\, m(a - c(\Gamma-\gamma)), \\ aM + c(M-1)\Gamma \,&<\, a + cm(2\gamma - \Gamma), \\ aM + c(M-1)\Gamma + cm(\Gamma-\gamma) \,&<\, aN + c(N-1)\gamma. \end{aligned}$$

We choose γ, Γ such that $0 < \Gamma/2 < \gamma < \Gamma$ and a, c such that $a > c(\Gamma - \gamma)$.
Fix an integer M. The first two inequalities are fulfilled for m sufficiently large. Fix such a value of m. The parameter N appears only in the third inequality. All other parameters being fixed, this inequality is satisfied for N sufficiently large.
We have thus built a space E, a function $f : E \mapsto \mathbb{R}_+^*$ and a set of parameters (a, c, m) (where the population size m may be arbitrarily large) such that $W^* \cap f^* = \emptyset$.

V. Convergence of the inhomogeneous algorithm

From now onwards we assume that l and n increase simultaneously. More precisely, we take l as an increasing function of n with $\lim_{n\to\infty} l(n) = \infty$.
The Markov chains of our model become thus time inhomogeneous and the transition probabilities depend on the time via the function $l(n)$. We suppress the index l in our notations; for instance X_n stands for X_n^l.
Let τ_n be the first time of entrance in S strictly greater than n:

$$\tau_n = \inf\{\, k : k > n,\, X_k \in S \,\}$$

and denote by $T_1, \cdots, T_n, \cdots$ the instants of the successive visits of the chain in S. Putting $T_0 = 0$, we have in particular $T_{n+1} = \tau_{T_n}$ for each integer n.
We are mostly interested in the behavior of the chain $(X_{T_n})_{n\ge 0}$.

1. Additional assumptions

To carry out the asymptotic estimations of the various transition probabilities, we need stronger assumptions than in the homogeneous case. We make the following

Hypothesis H_i. There exists a sequence $(L(n))_{n\ge 0}$ of positive real numbers such that

$$\forall r \in \mathbb{N} \quad \forall s \in \mathbb{R}^* \qquad l(n+r)^s = o(L(n)) \quad \text{as} \quad n \to \infty$$

and $p_{l(n)}$ and $q_{l(n)}$ admit the developments

$$\forall i, j \in E \qquad p_{l(n)}(i,j) = \begin{cases} \alpha(i,j) l(n)^{-a} + o\left(L(n)^{-1}\right) & \text{if } i \ne j \\ 1 - \alpha(i,j) l(n)^{-a} + o\left(L(n)^{-1}\right) & \text{if } i = j \end{cases}$$

$\forall (i_1, j_1) \in E \times E \quad \forall (i_2, j_2) \in E \times E$

$$q_{l(n)}\left((i_1,j_1),(i_2,j_2)\right) = \begin{cases} \beta\left((i_1,j_1),(i_2,j_2)\right) l(n)^{-b} + o\left(L(n)^{-1}\right) & \text{if } (i_1,j_1) \ne (i_2,j_2) \\ 1 - \beta\left((i_1,j_1),(i_2,j_2)\right) l(n)^{-b} + o\left(L(n)^{-1}\right) & \text{if } (i_1,j_1) = (i_2,j_2) \end{cases}$$

In order to avoid numerous occurrences of the usual sentence “for n sufficiently large” we impose that for all n in $\mathbb{N}$,

$$\begin{aligned} &\forall\, i \in E \qquad p_{l(n)}(i,i) > 0 \\ &\forall\, i, j \in E \qquad q_{l(n)}((i,j),(i,j)) > 0 \\ &\forall\, i, j \in E \quad i \ne j \qquad \alpha(i,j) > 0 \Rightarrow p_{l(n)}(i,j) > 0 \\ &\forall\, (i_1,j_1) \in E \times E \quad \forall (i_2,j_2) \in E \times E \quad (i_1,j_1) \ne (i_2,j_2) \\ &\qquad\qquad \beta\left((i_1,j_1),(i_2,j_2)\right) > 0 \Rightarrow q_{l(n)}\left((i_1,j_1),(i_2,j_2)\right) > 0. \end{aligned}$$

The asymptotic developments of $p_{l(n)}$ and $q_{l(n)}$ imply that these properties are satisfied “for n sufficiently large”. Since we deal only with the asymptotic dynamics of the process, these conditions are not restrictive.

2. Time of entrance in S

We proceed in the same way as in the homogeneous case. With

$$s_{l(n)} = \min_{x \in E^m} \sum_{k:x_k \in \hat{x}} F_{l(n)}(k, f(x))$$

we obtain

$$P\left(\tau_n > n + q / X_n = y\right) \leq \big(1 - (s_{l(n)})^m\big) \cdots \big(1 - (s_{l(n+q)})^m\big)$$

and

$$0 \leq 1 - s_{l(n)} \leq \frac{m-1}{l(n)^{\,c\delta}} \tag{35}$$

whence

$$P\left(\tau_n > n + q / X_n = y\right) \leq \frac{(m(m-1))^{q+1}}{(l(n) l(n+1) \cdots l(n+q))^{\,c\delta}}. \tag{36}$$

Multiplying by $P_x\left(X_n = y\right)$ and summing up over all y in E^m yields

$$P_x\left(\tau_n > n + q\right) \leq \frac{(m(m-1))^{q+1}}{(l(n) l(n+1) \cdots l(n+q))^{\,c\delta}}.$$

In particular, $E_x(\tau_n) < \infty$ for all n in $\mathbb{N}$ and the sequence $(T_n)_{n \geq 0}$ is well defined (i.e. all the times T_n are finite) with probability one.

3. Transition between two populations in r steps

Let z_1 and z_2 be two populations in E^m. We define the sets

$$D_r(z_1, z_2), \quad \widetilde{D}_r(z_1, z_2), \quad D_r^*(z_1, z_2), \quad \widetilde{D}_r^*(z_1, z_2)$$

and the functions $V_r(z_1, z_2), \widetilde{V}_r(z_1, z_2)$ exactly as in the homogeneous case.
Let r be a fixed integer. Using the asymptotic developments of hypothesis H_i, we have

$$P\left(X_{n+r} = z_2 / X_n = z_1\right) = \\ \sum_{D_r(z_1, z_2)} P\big(X_{n+r} = z_2, V_{n+r-1} = v^{r-1}, \cdots, V_n = v^0, U_n = u^0, X_n = z_1\big) + o\left(L(n)^{-1}\right)$$

Consider the term

$$P\big(X_{n+r} = z_2, V_{n+r-1} = v^{r-1}, U_{n+r-1} = u^{r-1}, X_{n+r-1} = x^{r-1}, \cdots, \\ X_{n+1} = x^1, V_n = v^0, U_n = u^0, X_n = z_1\big) \tag{37}$$

associated with a path $\mathcal{P} = (u^k, v^k, x^k)$ of $D_r(z_1, z_2)$. It is asymptotically equivalent to

$$C(\mathcal{P}) \exp\left(-\Big(ad(z_1, u^0) + b\bar{d}(u^0, v^0) + c\big(m\widehat{f}(v^0) - \sum_{h=1}^{m} f(x_h^1)\big)\Big)\ln l(n)\right) \times \cdots$$
$$\times \exp\left(-\Big(ad(x^{r-1}, u^{r-1}) + b\bar{d}(u^{r-1}, v^{r-1}) + c\big(m\widehat{f}(v^{r-1}) - \sum_{h=1}^{m} f(z_h^2)\big)\Big)\ln l(n+r-1)\right)$$

so that there exist two constants γ_1 and γ_2, $0 < \gamma_1 \leq \gamma_2$, such that for all n, the above term is greater than

$$\gamma_1 \exp\left(-\sum_{k=0}^{r-1}\Big(ad(x^k, u^k) + b\bar{d}(u^k, v^k) + c\big(m\widehat{f}(v^k) - \sum_{h=1}^{m} f(x_h^{k+1})\big)\Big)\ln l(n+r-1)\right)$$

and smaller than

$$\gamma_2 \exp\left(-\sum_{k=0}^{r-1}\Big(ad(x^k, u^k) + b\bar{d}(u^k, v^k) + c\big(m\widehat{f}(v^k) - \sum_{h=1}^{m} f(x_h^{k+1})\big)\Big)\ln l(n)\right).$$

Since the number of paths in the set $D_r(z_1, z_2)$ is finite we can even suppose γ_1 and γ_2 independent of the term under consideration. We sum up all these inequalities over the whole set $D_r(z_1, z_2)$. Considering once more the asymptotics of the bounding terms, we see that there exist two constants γ_1^r and γ_2^r such that $0 < \gamma_1^r \leq \gamma_2^r$ and for all n in $\mathbb{N}$,

$$\gamma_1^r\, l(n+r-1)^{-V_r(z_1,z_2)} \leq P_x\left(X_{n+r} = z_2 / X_n = z_1\right) \leq \gamma_2^r\, l(n)^{-V_r(z_1,z_2)}.$$

A similar argument yields the existence of two constants $\tilde{\gamma}_1^r$ and $\tilde{\gamma}_2^r$ with $0 < \tilde{\gamma}_1^r \leq \tilde{\gamma}_2^r$ and for all n in $\mathbb{N}$,

$$\begin{aligned}(38)\quad &\tilde{\gamma}_1^r\, l(n+r-1)^{-\widetilde{V}_r(z_1,z_2)} \\ &\leq P_x\left(X_{n+r} = z_2, \forall\, k \in \{1, \cdots, r-1\}\ X_{n+k} \notin S / X_n = z_1\right) \leq \\ &\qquad \tilde{\gamma}_2^r\, l(n)^{-\widetilde{V}_r(z_1,z_2)}.\end{aligned}$$

4. Law of the entrance point in S

Now we study how the chain most probably enters the set S after time n, that is we evaluate for y in E^m and z in S

$$P_x\left(X_{\tau_n} = z / X_n = y\right).$$

The quantities which are of fundamental importance to achieve this are the set $\widetilde{D}$ and the functional $\widetilde{V}$ which are defined as in the homogeneous case. We decompose the transition probability according to the value of τ_n:

$P_x\left(X_{\tau_n} = z/X_n = y\right) =$

$$\sum_{k=n+1}^{n+q} P_x\left(X_k = z,\, \tau_n = k/X_n = y\right) + P_x\left(X_{\tau_n} = z,\, \tau_n > n+q/X_n = y\right). \tag{39}$$

We choose the integer q large enough to ensure that

$$q > r^* = \text{card}\left(E^m/\mathfrak{S}_m\right) \quad \text{and} \quad c\delta(q - r^* + 2) > \widetilde{V}(y,z).$$

Yet, by (38),

$$\sum_{k=n+1}^{n+q} P_x\left(X_k = z,\, \tau_n = k/X_n = y\right) \geq \sum_{k=1}^{r^*} \tilde{\gamma}_1^k l(n+k-1)^{-\widetilde{V}_k(y,z)} \geq \tilde{\Gamma}_1 l(n+r^*-1)^{-\widetilde{V}(y,z)}$$

for some positive constant $\tilde{\Gamma}_1$. Similarly, there exists a constant $\tilde{\Gamma}_2$ such that

$$\sum_{k=n+1}^{n+q} P_x\left(X_k = z,\, \tau_n = k/X_n = y\right) \leq \tilde{\Gamma}_2 l(n)^{-\widetilde{V}(y,z)}$$

However, we deduce from (36) that

$$P\left(\tau_n > n+q/X_n = y\right) \leq \frac{(m(m-1))^{q+1}}{\left(l(n)\cdots l(n+r^*-2)\right)^{\delta c}\left(l(n+r^*-1)\right)^{c\delta(q-r^*+2)}}$$

and with our choice for the integer q, the second term in the sum (39) is thus negligible before the first one. Finally, there exist two positive constants Γ_1 and Γ_2 verifying for all n in $\mathbb{N}$

$$\Gamma_1 l(n+r^*-1)^{-\widetilde{V}(y,z)} \leq P_x\left(X_{\tau_n} = z/X_n = y\right) \leq \Gamma_2 l(n)^{-\widetilde{V}(y,z)} \tag{40}$$

and E being finite they can be chosen so as to be independent of y and z.
Our aim is to evaluate the transition probabilities of the chain (X_{T_n}). We thus have to evaluate conditional probabilities of the form

$$P_x\left(X_{T_{n+1}} = z/X_{T_n} = y\right)$$

instead of

$$P_x\left(X_{\tau_n} = z/X_n = y\right).$$

To compute conditional probabilities with respect to the event $\{X_{T_n} = y\}$ and to apply the preceding technique, we need some information about the value of T_n. We first evaluate for t real and strictly greater than one

$$P_x\left(T_n > tn/X_{T_n} = y\right).$$

5. A lower bound for $P\left(X_{T_n}=y/X_0=x\right),\ y\in S$

Let r be an integer such that $\widetilde{V}_r(x,y)<\infty$. We have

$$
\begin{aligned}
P_x\left(X_{T_n}=y\right) &\geq P_x\left(T_1=r,\ X_r=y,\ \forall k\in\{2,\cdots,n\}\ X_{r+k-1}=y\right)\\
&=P\left(X_{r+n-1}=y/X_{r+n-2}=y\right)\cdots P\left(X_{r+1}=y/X_r=y\right)P_x\left(T_1=r,X_{T_1}=y\right).
\end{aligned}
$$

If during the n–th transition of the chain there are neither mutations nor crossovers and the selection causes no loss of diversity then $X_{n+1}=X_n$. Thus for each y in S and n in $\mathbb{N}$,

$$
P\left(X_{n+1}=y/X_n=y\right)\geq m^{-m}\prod_{k=1}^{m}p_{l(n)}(y_k,y_k)\prod_{k=1}^{\lfloor m/2\rfloor}q_{l(n)}\left((y_{2k-1},y_{2k}),(y_{2k-1},y_{2k})\right).
$$

Hence

$$
\begin{aligned}
&P_x\left(X_{T_n}=y\right)\geq P_x\left(T_1=r,\ X_{T_1}=y\right)\times\\
&\qquad m^{-m(n-1)}\prod_{k=1}^{m}\prod_{h=r}^{r+n-2}p_{l(h)}(y_k,y_k)\prod_{k=1}^{\lfloor m/2\rfloor}\prod_{h=r}^{r+n-2}q_{l(h)}\left((y_{2k-1},y_{2k}),(y_{2k-1},y_{2k})\right).
\end{aligned}
$$

Let i and j be two fixed points of E. We study the products

$$
\prod_{h=r}^{r+n-2}p_{l(h)}(i,i)\quad\text{and}\quad\prod_{h=r}^{r+n-2}q_{l(h)}\left((i,j),(i,j)\right).
$$

Using the development given by hypothesis H_i, we have the following cases.

- If the series $\sum_{n\geq 0}l(n)^{-a}$ converges, the infinite product $\prod_{h=r}^{\infty}p_{l(h)}(i,i)$ is convergent whence

$$
\liminf_{n\to\infty}\prod_{h=r}^{n+r-2}p_{l(h)}(i,i)>0.
$$

- If the series $\sum_{n\geq 0}l(n)^{-a}$ diverges, the infinite product is divergent and

$$
\ln\prod_{h=r}^{r+n-2}p_{l(h)}(i,i)=\sum_{h=r}^{r+n-2}\ln p_{l(h)}(i,i)\underset{l\to\infty}{\sim}-\sum_{h=r}^{r+n-2}\frac{\alpha(i,i)}{l(h)^a}\geq-(n-1)\frac{\alpha(i,i)}{l(1)^a}
$$

whence

$$
\forall n\in\mathbb{N}\qquad\ln\prod_{h=r}^{r+n-2}p_{l(h)}(i,i)\geq-K_1n
$$

for some positive constant K_1 (the preceding inequality is certainly true for n large; using our assumptions on the sign of $p_{l(n)}(i,i)$ and taking K_1 large enough, it is also true for small values of n; such reasonings will be made implicitly in the sequel).
Similar arguments apply for the second product: either the series $\sum_{n\geq 0} l(n)^{-b}$ converges and

$$\liminf_{n\to\infty} \prod_{h=r}^{r+n-2} q_{l(h)}\left((i,j),(i,j)\right) > 0$$

or the series diverges and there exists a positive constant K_2 such that

$$\forall n \in \mathbb{N} \qquad \ln \prod_{h=r}^{r+n-2} q_{l(h)}\left((i,j),(i,j)\right) \geq -K_2 n.$$

Finally, there exists a positive constant K_3 such that

$$\forall n \in \mathbb{N} \qquad \ln \prod_{h=r}^{r+n-2} p_{l(h)}(i,i) + \ln \prod_{h=r}^{r+n-2} q_{l(h)}\left((i,j),(i,j)\right) \geq -K_3 n$$

whence in all cases, for some positive constant K_4,

$$(41) \qquad \forall y \in S \quad \forall n \in \mathbb{N}^* \qquad P\left(X_{T_n} = y / X_0 = x\right) \geq \exp(-K_4 n).$$

6. An upper bound for $P\left(T_n > tn/X_0 = x\right)$, $t > 1$

If $T_n > tn$ there are at least $\lfloor tn - n + 1 \rfloor$ transitions from V_n to X_{n+1} for which the selected individuals do not all belong to $\hat{V}_n$. The probability of such a transition at time n is smaller than $1 - \left(s_{l(n)}\right)^m$. In addition $s_{l(n)}$ is an increasing function of n. Necessarily

$$P_x\left(T_n > tn\right) \leq \left(1 - (s_{l(1)})^m\right) \cdots \left(1 - (s_{l(\lfloor tn-n+1 \rfloor)})^m\right).$$

Using inequality (35), we obtain

$$P_x\left(T_n > tn\right) \leq \frac{\left(m(m-1)\right)^{\lfloor tn-n+1 \rfloor}}{\left(l(1)l(2)\cdots l(\lfloor tn-n+1 \rfloor)\right)^{c\delta}}$$

Taking the logarithm, we have

$$\ln P_x\left(T_n > tn\right) \leq \lfloor tn - n + 1 \rfloor \ln(m(m-1)) - \delta c \sum_{k=1}^{\lfloor tn-n+1 \rfloor} \ln l(k).$$

However

$$\sum_{k=1}^{\lfloor tn-n+1\rfloor} \ln l(k) \geq \frac{\lfloor tn-n\rfloor}{2} \ln l(\lfloor (tn-n)/2\rfloor)$$

whence

$$\forall n\in\mathbb{N}\qquad \ln P_x\left(T_n > tn\right) \leq -K_5 n \ln l(\lfloor (tn-n)/2\rfloor) + K_6$$

for some positive constants K_5 and K_6. Finally, there exist positive constants K_7, K_8 and K_9 such that

$$\forall n\in\mathbb{N}\qquad P_x\left(T_n > tn\right) \leq K_7 \exp(-K_8 n \ln l(\lfloor K_9 n\rfloor)). \tag{42}$$

7. An upper bound for $P_x\left(T_n > tn/X_{T_n} = y\right)$

Using the two preceding inequalities (41) and (42), we obtain

$$P_x\left(T_n > tn/X_{T_n} = y\right) \leq \frac{P_x\left(T_n > tn\right)}{P_x\left(X_{T_n} = y\right)} \leq K_7 \exp\left(K_4 n - K_8 n \ln l(\lfloor K_9 n\rfloor)\right).$$

There exist positive constants C and K such that

$$\forall n\in\mathbb{N}\qquad K_4 n - K_8 n \ln l(\lfloor K_9 n\rfloor) \leq \ln C - Kn$$

whence finally

$$\forall n\in\mathbb{N}\qquad P_x\left(T_n > tn/X_{T_n} = y\right) \leq C\exp(-Kn).$$

In particular, for all $t>1$,

$$\lim_{n\to\infty} P_x\left(T_n > tn/X_{T_n} = y\right) = 0. \tag{43}$$

8. Transitions of the chain (X_{T_n})

All the tools needed to evaluate the order of the asymptotics of

$$P_x\left(X_{T_{n+1}} = z/X_{T_n} = y\right)$$

for y and z in S are now available. Again, we decompose this quantity according to the value of T_n.

$$\begin{aligned} P_x\left(X_{T_{n+1}} = z/X_{T_n} = y\right) &= \sum_{k=n}^{\infty} P_x\left(X_{T_{n+1}} = z,\, T_n = k/X_{T_n} = y\right) \\ &= \sum_{k=n}^{\infty} P_x\left(X_{T_{n+1}} = z/T_n = k,\, X_{T_n} = y\right) P_x\left(T_n = k/X_{T_n} = y\right). \end{aligned}$$

However
$$P_x\left(X_{T_{n+1}}=z/T_n=k,\,X_{T_n}=y\right)=P_x\left(X_{\tau_k}=z/X_k=y\right).$$
By (40), for any $t>1$ we have for all k in $\{n,\cdots,\lfloor tn\rfloor\}$

$$\Gamma_1 l(\lfloor tn\rfloor+r^*-1)^{-\widetilde{V}(y,z)}\leq P_x\left(X_{\tau_k}=z/X_k=y\right)\leq\Gamma_2 l(n)^{-\widetilde{V}(y,z)}.$$

We multiply this inequality by $P_x\left(T_n=k/X_{T_n}=y\right)$ and we sum up from n to $\lfloor tn\rfloor$:

$$\begin{aligned}
&\Gamma_1 l(\lfloor tn\rfloor+r^*-1)^{-\widetilde{V}(y,z)}P_x\left(T_n\leq tn/X_{T_n}=y\right)\\
&\qquad\leq\sum_{k=n}^{\lfloor tn\rfloor}P_x\left(X_{\tau_k}=z/X_k=y\right)P_x\left(T_n=k/X_{T_n}=y\right)\\
&\qquad\qquad\leq\Gamma_2 l(n)^{-\widetilde{V}(y,z)}P_x\left(T_n\leq tn/X_{T_n}=y\right).
\end{aligned}\tag{44}$$

In addition, we see from (40) that

$$\forall k\geq n\qquad P_x\left(X_{\tau_k}=z/X_k=y\right)\leq\Gamma_2 l(n)^{-\widetilde{V}(y,z)}.$$

Multiplying by $P_x\left(T_n=k/X_{T_n}=y\right)$ and summing up from $\lfloor tn\rfloor+1$ to ∞ gives

$$\begin{aligned}
&\sum_{k=\lfloor tn\rfloor+1}^{\infty}P_x\left(X_{T_{n+1}}=z/T_n=k,\,X_{T_n}=y\right)P_x\left(T_n=k/X_{T_n}=y\right)\\
&\qquad\qquad\leq\Gamma_2 l(n)^{-\widetilde{V}(y,z)}P_x\left(T_n>tn/X_{T_n}=y\right)
\end{aligned}$$

from which, together with the inequalities (44), we deduce

$$\begin{aligned}
&\Gamma_1 l(\lfloor tn\rfloor+r^*-1)^{-\widetilde{V}(y,z)}P_x\left(T_n\leq tn/X_{T_n}=y\right)\\
&\qquad\leq\sum_{k=n}^{\infty}P_x\left(X_{T_{n+1}}=z/T_n=k,\,X_{T_n}=y\right)\leq\Gamma_2 l(n)^{-\widetilde{V}(y,z)}.
\end{aligned}$$

Using (43), we see that there exists a positive constant Γ_1^t, depending on t, satisfying

$$\forall n\in\mathbb{N}\qquad\Gamma_1^t\leq\Gamma_1 P_x\left(T_n\leq tn/X_{T_n}=y\right)$$

whence finally, for all n in $\mathbb{N}$ and all y,z in S,

$$\Gamma_1^t l(\lfloor tn\rfloor+r^*-1)^{-\widetilde{V}(y,z)}\leq P_x\left(X_{T_{n+1}}=z/X_{T_n}=y\right)\leq\Gamma_2 l(n)^{-\widetilde{V}(y,z)}$$

or, putting $l_t(n) = l(\lfloor tn \rfloor + r^* - 1)$,

$$\Gamma_1^t l_t(n)^{-\widetilde{V}(y,z)} \leq P_x \left(X_{T_{n+1}} = z / X_{T_n} = y \right) \leq \Gamma_2 l(n)^{-\widetilde{V}(y,z)}. \tag{45}$$

From this estimation, we will derive numerous interesting results concerning the asymptotic dynamics of the chain (X_{T_n}). Most of our conditions will deal with the nature of the series $\sum_{n\geq 0} l(n)^{-\theta}$ for different values of θ. However, the sequence $l(n)$ being increasing, it is clear that for any value of θ and $t > 1$, the series

$$\sum_{n\geq 0} l(n)^{-\theta} \qquad \text{and} \qquad \sum_{n\geq 0} l_t(n)^{-\theta}$$

are of the same nature: thus the above inequality (45) with $l_t(n)$ on the left–hand side will be enough for our immediate purposes.
Recall that the convergence exponent [5, chapter 3] of the increasing sequence $l(n)$ is defined as the unique non–negative real number λ having the following property: the series

$$l(1)^{-\theta} + l(2)^{-\theta} + \cdots + l(n)^{-\theta} + \cdots$$

converges for $\theta > \lambda$ and diverges for $\theta < \lambda$. The preceding discussion shows that the sequences $l_t(n)$ and $l(n)$ have the same convergence exponent for any $t > 1$.

9. A criterion for the settling in a given subset of E

Proposition 9.1. *Let I be a subset of E. Define*

$$\begin{aligned} \phi_I &= \min_{y\in I^m\cap S} \quad \min_{z\in S\setminus I^m} \quad \widetilde{V}(y,z), \\ \psi_I &= \max_{y\in I^m\cap S} \quad \min_{z\in S\setminus I^m} \quad \widetilde{V}(y,z). \end{aligned}$$

If $\sum_{n\geq 0} l(n)^{-\psi_I} = \infty$, *the chain* (X_{T_n}) *is not trapped in* I *with probability one:*

$$P\left(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset I / X_0 = x\right) = 0.$$

If $\sum_{n\geq 0} l(n)^{-\phi_I} < \infty$, *the probability for the chain* (X_{T_n}) *to be trapped in* I *is non–zero:*

$$P\left(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset I / X_0 = x\right) > 0.$$

Remark. We can give a slightly stronger condition in term of the convergence exponent λ of the sequence $l(n)$: the hypothesis of the proposition are obviously fulfilled whenever $\lambda < \phi_I$ or $\lambda > \psi_I$.

Proof. Let I be a subset of E. We first consider the event

$$\bigcap_{N\le n\le M}\left\{[X_{T_n}]\subset I\right\}$$

for $N \le M$. We have

$$P_x\left(\forall n\in\{N\cdots M\}\quad [X_{T_n}]\subset I\right)=\sum_{x_N\cdots x_M\in I^m\cap S} P_x\left(X_{T_M}=x_M,\cdots,X_{T_N}=x_N\right)$$

$$(46)\qquad =\sum_{x_N\cdots x_M\in I^m\cap S} P_x\left(X_{T_N}=x_N\right)\prod_{k=N}^{M-1} P_x\left(X_{T_{k+1}}=x_{k+1}/X_{T_k}=x_k\right).$$

Yet

$$\sum_{x_M\in I^m\cap S} P_x\left(X_{T_M}=x_M/X_{T_{M-1}}=x_{M-1}\right)=P_x\left([X_{T_M}]\subset I/X_{T_{M-1}}=x_{M-1}\right).$$

Let y be a population in $I^m\cap S$. We study the quantity

$$P_x\left([X_{T_{n+1}}]\subset I/X_{T_n}=y\right).$$

Summing (45) over all z in $S\setminus I^m$, we obtain

$$\Gamma_1^t\sum_{z\in S\setminus I^m} l_t(n)^{-\widetilde{V}(y,z)}\le P_x\left([X_{T_{n+1}}]\not\subset I/X_{T_n}=y\right)\le \Gamma_2\sum_{z\in S\setminus I^m} l(n)^{-\widetilde{V}(y,z)}.$$

Once more, we need to take into account only the terms with the lowest exponents in the sums: thus there exist two constants Γ_1' and Γ_2', $0<\Gamma_1'\le\Gamma_2'$, such that

$$\forall n\in\mathbb{N}\qquad \Gamma_1' l_t(n)^{-\theta(y)}\le P_x\left([X_{T_{n+1}}]\not\subset I/X_{T_n}=y\right)\le \Gamma_2' l(n)^{-\theta(y)}$$

where $\theta(y)=\min_{z\in S\setminus I^m}\widetilde{V}(y,z)$.
By the very definition of ϕ_I and ψ_I we have $\phi_I\le\theta(y)\le\psi_I$ for all y in I^m whence

$$\Gamma_1' l_t(n)^{-\psi_I}\le P_x\left([X_{T_{n+1}}]\not\subset I/X_{T_n}=y\right)\le\Gamma_2' l(n)^{-\phi_I}$$

or

$$1-\Gamma_2' l(n)^{-\phi_I}\le P_x\left([X_{T_{n+1}}]\subset I/X_{T_n}=y\right)\le 1-\Gamma_1' l_t(n)^{-\psi_I}.$$

Thus, reporting these inequalities in (46),

$$\begin{aligned}\left(1-\Gamma'_2 l_t(M-1)^{-\phi_I}\right) P_x\left(\forall n\in\{N,\cdots,M-1\}\quad [X_{T_n}]\subset I\right)&\\ \le P_x\left(\forall n\in\{N,\cdots,M\}\quad [X_{T_n}]\subset I\right)\le&\\ \left(1-\Gamma'_1 l(M-1)^{-\psi_I}\right) P_x\left(\forall n\in\{N,\cdots,M-1\}\quad [X_{T_n}]\subset I\right).&\end{aligned}$$

A straightforward induction then yields

$$\begin{aligned}\left(1-\Gamma'_2 l_t(M-1)^{-\phi_I}\right)\cdots\left(1-\Gamma'_2 l_t(N)^{-\phi_I}\right) P_x\left([X_{T_N}]\subset I\right)&\\ \le P_x\left(\forall n\in\{N,\cdots,M\}\quad [X_{T_n}]\subset I\right)\le&\\ \left(1-\Gamma'_1 l(M-1)^{-\psi_I}\right)\cdots\left(1-\Gamma'_1 l(N)^{-\psi_I}\right) P_x\left([X_{T_N}]\subset I\right).&\end{aligned}$$

Passing to the limit when M goes to the infinity gives

$$\begin{aligned}P_x\left([X_{T_N}]\subset I\right)\prod_{n=N}^{\infty}\left(1-\Gamma'_2 l_t(n)^{-\phi_I}\right)&\\ \le P_x\left(\forall n\ge N\quad [X_{T_n}]\subset I\right)\le&\\ P_x\left([X_{T_N}]\subset I\right)\prod_{n=N}^{\infty}\left(1-\Gamma'_1 l(n)^{-\psi_I}\right).&\end{aligned}$$

If $\quad \sum_{n\ge 0} l(n)^{-\psi_I}=\infty$ we obtain for each N in $\mathbb{N}$

$$P\left(\forall n\ge N\quad [X_{T_n}]\subset I/X_0=x\right)=0$$

whence

$$P\left(\exists N\quad \forall n\ge N\quad [X_{T_n}]\subset I/X_0=x\right)=0.$$

If $\quad \sum_{n\ge 0} l(n)^{-\phi_I}<\infty$ then for each N in $\mathbb{N}^*$

$$P\left(\forall n\ge N\quad [X_{T_n}]\subset I/X_0=x\right)>0$$

since both

$$P_x\left([X_{T_N}]\subset I\right)\quad\text{and}\quad \prod_{n=N}^{\infty}\left(1-\Gamma'_2 l(n)^{-\phi_I}\right)$$

are positive. It follows that

$$P\left(\exists N\quad \forall n\ge N\quad [X_{T_n}]\subset I/X_0=x\right)>0$$

and the proposition is proved. $\square$

Corollary 9.1. *Let I be a subset of E. Define*

$$\tilde{\psi}_I = \max_{y \in I^m \cap S} \min_{z \in S \setminus I^m} V(y,z).$$

If $\sum_{n \geq 0} l(n)^{-\tilde{\psi}_I} = \infty$*, the chain (X_{T_n}) is not trapped in I with probability one:*

$$P\left(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset I / X_0 = x\right) = 0.$$

Remark. Since $\tilde{\psi}_I$ is defined through V instead of $\widetilde{V}$ we have in general $\tilde{\psi}_I < \psi_I$ and the above result is stronger than the first assertion of proposition 9.1. Furthermore $\tilde{\psi}_I$ is positive when $I \neq \emptyset$ (by lemma 4.2 and corollary 4.2) and is finite when I is a strict subset of E (since $V(y,z)$ is always finite), whereas ψ_I may be infinite (for instance if there exists a point y in I^m which can't reach $S \setminus I^m$ through a path constrained outside S).

Proof. Let r be an integer greater than r^*. We are going to apply proposition 9.1 to the chain $(X_{T_{rn}})_{n \in \mathbb{N}}$. To do this, we essentially need asymptotic estimates of the form (45). We decompose the transition from $X_{T_{rn}}$ to $X_{T_{r(n+1)}}$

$$P\left(X_{T_{r(n+1)}} = z / X_{T_{rn}} = y\right) = \\ \sum_{x^1, \cdots, x^{r-1} \in S} P\big(X_{T_{rn}} = y, X_{T_{rn+1}} = x^1, X_{T_{rn+2}} = x^2, \cdots, X_{T_{rn+r-1}} = x^{r-1}, X_{T_{r(n+1)}} = z\big)$$

and once more, we need only to take into account the terms which prevail at infinity. Putting for y and z in S

$$\widetilde{V}^{(r)}(y,z) = \min_{x^1 \cdots x^{r-1} \in S} \left\{ \widetilde{V}(y,x^1) + \widetilde{V}(x^1,x^2) + \cdots + \widetilde{V}(x^{r-1},z) \right\},$$

we deduce from (45) the existence of two positive constants Γ_3 and Γ_4 such that

$$\Gamma_3 l_t(r(n+1))^{-\widetilde{V}^{(r)}(y,z)} \leq P_x\left(X_{T_{r(n+1)}} = z / X_{T_{rn}} = y\right) \leq \Gamma_4 l(rn)^{-\widetilde{V}^{(r)}(y,z)}.$$

Since $r \geq r^*$, lemma 4.3 implies that $\widetilde{V}^{(r)}(y,z) = V(y,z)$ for y and z in S.
Proposition 9.1 applied to the chain $(X_{T_{rn}})$ and the set I yields

$$P\left(\exists N \quad \forall n \geq N \quad [X_{T_{rn}}] \subset I / X_0 = x\right) = 0.$$

Since

$$P\left(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset I / X_0 = x\right) \leq P\left(\exists N \quad \forall n \geq N \quad [X_{T_{rn}}] \subset I / X_0 = x\right),$$

the conclusion follows. $\square$

As a consequence, we derive

10. A necessary condition for the settling in f^*

Theorem 10.1. *For the chain (X_{T_n}) to be trapped in f^* after a finite number of transitions, i.e. to have*

$$\forall x \in E^m \qquad P(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset f^* / X_0 = x) = 1$$

it is necessary that, putting

$$\begin{aligned}\psi &= \max_{I \subset f^*} \phi_I = \max_{I \subset f^*} \min_{y \in I^m \cap S} \min_{z \in S \setminus I^m} \widetilde{V}(y,z),\\ \phi &= \max_{I \subset E \setminus f^*} \tilde{\psi}_I = \max_{I \subset E \setminus f^*} \max_{y \in I^m \cap S} \min_{z \in S \setminus I^m} V(y,z)\end{aligned}$$

we have

$$\sum_{n \geq 0} l(n)^{-\phi} = \infty \qquad \text{and} \qquad \sum_{n \geq 0} l(n)^{-\psi} < \infty. \tag{47}$$

Remark 1. The above condition implies in particular that $\phi < \psi$. Since $\psi \geq \rho m$ (take $I = f^*$ and use inequality (32)) whereas $\phi \leq (a + c\Delta)|E|$ (by inequality (31)), this condition is fulfilled for m large enough. However when m is small, we may have $\phi \geq \psi$. Thus, to ensure the settling of the population in f^* with an increasing sequence $l(n)$, the population size m has to be at least large enough to ensure that $\phi < \psi$.

Remark 2. We have always $\psi < \infty$ (since $f^* \neq E$, ϕ_I is finite when $I \subset f^*$) and $\phi > 0$ (since $f^* \neq E$, there is at least one subset I of $E \setminus f^*$ such that $\tilde{\psi}_I$ is positive). In term of the convergence exponent λ of the sequence $l(n)$, the above condition (47) implies $\phi \leq \lambda \leq \psi$. In particular, $0 < \lambda < \infty$.

Proof. The theorem is a direct consequence of proposition 9.1 and corollary 9.1. It only says that

- for every subset I of $E \setminus f^*$, we must have

$$\forall x \in E^m \qquad P(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset I / X_0 = x) = 0,$$

- there exists at least a subset I of f^* such that

$$\forall x \in E^m \qquad P(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset I / X_0 = x) > 0$$

which are clearly necessary conditions for the chain (X_{T_n}) to be trapped in f^* in finite time independently of the starting point and with probability one. $\square$

We give now

11. A sufficient condition for the settling in f^*

Theorem 11.1. *Define*

$$\theta_1 = \max_{y \in S \setminus (f^*)^m} \min_{z \in S, \widehat{f}(z) > \widehat{f}(y)} V(y,z) \,,$$

$$\theta_2 = \min_{y \in S, z \in E^m \,:\, \widehat{f}(z) < \widehat{f}(y)} \widetilde{V}(y,z).$$

If

$$\sum_{n \geq 0} l(n)^{-\theta_1} = \infty \qquad \text{and} \qquad \sum_{n \geq 0} l(n)^{-\theta_2} < \infty$$

then, with probability one, after a finite number of transitions, the chain (X_{T_n}) *is trapped in* f^**, that is,*

$$\forall x \in E^m \qquad P\left(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset f^* / X_0 = x\right) = 1.$$

Moreover, with probability one, after a finite number of transitions, the population (X_n) *permanently contains individuals belonging to* f^* *i.e.*

$$\forall x \in E^m \qquad P\Big(\exists N \quad \forall n \geq N \quad \hat{X}_n \subset f^* / X_0 = x\Big) = 1.$$

Remark. Inequality (31) of corollary 11.1 shows that θ_1 is bounded by a constant independent of m whereas inequality (32) implies $\theta_2 \geq \rho m$. For m large enough, the interval $]\theta_1, \theta_2[$ is non–empty and sequences $l(n)$ with the desired properties do exist.

Proof. Suppose the conditions of the theorem are fulfilled. We first show that, with probability one, the sequence $(f(X_{T_n}))_{n \geq 0}$ is increasing from some rank onwards. We study the event $\{f(X_{T_{n+1}}) < f(X_{T_n})\}$:

$$P_x\left(f(X_{T_{n+1}}) < f(X_{T_n})\right) = \sum_{y \in S} \sum_{\substack{z \in S \\ \widehat{f}(z) < \widehat{f}(y)}} P_x\left(X_{T_{n+1}} = z / X_{T_n} = y\right) P_x\left(X_{T_n} = y\right).$$

Using the right–hand inequality of (45) and the fact that $\widetilde{V}(y,z) \geq \theta_2$ when $\widehat{f}(z) < \widehat{f}(y)$ we obtain

$$P_x\left(f(X_{T_{n+1}}) < f(X_{T_n})\right) \leq \sum_{y \in S} \sum_{\substack{z \in S \\ \widehat{f}(z) < \widehat{f}(y)}} \Gamma_2 \, l(n)^{-\theta_2} \, P_x\left(X_{T_n} = y\right) \leq \Gamma_2 \, (\operatorname{card} S) \, l(n)^{-\theta_2}.$$

The series $\sum_{n \geq 0} l(n)^{-\theta_2}$ converges; a direct application of Borel–Cantelli lemma yields

$$P_x\left(f(X_{T_{n+1}}) < f(X_{T_n}) \quad \text{finitely often}\right) = 1.$$

Therefore, with probability one, there exists a rank N_1 (which is random) such that

$$\forall n \geq N_1 \qquad f(X_{T_{n+1}}) \geq f(X_{T_n}).$$

Since the set $f(E)$ is finite, the sequence $(f(X_{T_n}))_{n\geq 0}$ is in fact stationary:

$$\exists \lambda \in f(E) \quad \exists N_2 \quad \forall n \geq N_2 \quad f(X_{T_n}) = \lambda.$$

Denote by S_λ the set of populations whose individuals belong to f_λ and put

$$V(\lambda) \quad = \quad \max_{y \in S_\lambda} \quad \min_{z \in S \setminus S_\lambda} \quad V(y, z).$$

Suppose λ is not equal to $f(f^*)$. Then

$$V(\lambda) \quad \leq \quad \max_{y \in S \setminus (f^*)^m} \quad \min_{z \in S, \widehat{f}(z) > \widehat{f}(y)} \quad V(y, z) = \theta_1$$

Since the series $\sum_{n\geq 0} l(n)^{-\theta_1}$ diverges, the series $\sum_{n\geq 0} l(n)^{-V(\lambda)}$ is also divergent, and by corollary 9.1, the probability for the chain (X_{T_n}) to be trapped in S_λ is zero. Thus $\lambda = f(f^*)$ and the first part of the theorem is proved.

To prove the second assertion, we show that, with probability one, for n large enough,

$$\min_{T_n \leq N < T_{n+1}} \quad \widehat{f}(X_N) \quad \geq \quad \widehat{f}(X_{T_n}). \tag{48}$$

We will need the following

Lemma 11.1. *Suppose that for some $\theta > 0$ the series $\sum_{n\geq 0} l(n)^{-\theta}$ converges. Then, there exists a positive integer q such that*

$$\sum_{n=0}^{\infty} P_x\left(T_{n+1} - T_n > q\right) < \infty.$$

Proof of lemma. We have for $q \in \mathbb{N}^*$

$$P_x\left(T_{n+1} - T_n > q\right) = \sum_{k=n}^{\infty} \sum_{y \in S} P_x\left(T_{n+1} - T_n > q,\, T_n = k,\, X_k = y\right). \tag{49}$$

Yet

$$\begin{aligned} &P_x\left(T_{n+1} - T_n > q,\, T_n = k,\, X_k = y\right) \\ &\qquad = P_x\left(T_{n+1} - T_n > q / T_n = k,\, X_k = y\right) P_x\left(T_n = k,\, X_k = y\right) \\ &\qquad = P\left(\tau_k > k + q / X_k = y\right) P_x\left(T_n = k,\, X_k = y\right). \end{aligned}$$

However, by (36), for $k \geq n$,

$$P\left(\tau_k > k+q / X_k = y\right) \leq \frac{(m(m-1))^{q+1}}{(l(k)l(k+1)\cdots l(k+q))^{c\delta}} \leq \frac{(m(m-1))^{q+1}}{l(n)^{c\delta(q+1)}}.$$

Reporting this inequality in the sum (49), we obtain

$$P_x\left(T_{n+1}-T_n > q\right) \leq \frac{(m(m-1))^{q+1}}{l(n)^{c\delta(q+1)}}.$$

Any integer q such that $\delta c(q+1) > \theta$ solves the problem. $\square$

Now put

$$A_n = \{\exists N \quad T_n \leq N < T_{n+1} \quad \widehat{f}(X_N) < \widehat{f}(X_{T_n})\}.$$

Pick up an integer q large enough to ensure the convergence of the series

$$\sum_{n\geq 0} P_x\left(T_{n+1}-T_n > q\right)$$

(this is possible according to lemma 11.1).
We have

$$P_x\left(A_n\right) = P_x\left(A_n,\, T_{n+1}-T_n \leq q\right) + P_x\left(A_n,\, T_{n+1}-T_n > q\right)$$

so that we need only to examine the convergence of the series

$$\sum_{n\geq 0} P_x\left(A_n,\, T_{n+1}-T_n \leq q\right).$$

Yet

$$\begin{aligned}
&P_x\left(A_n,\, T_{n+1}-T_n \leq q\right)\\
&= \sum_{k=n}^{\infty}\sum_{y\in S} P_x\left(\exists N\ T_n \leq N < T_{n+1},\, \widehat{f}(X_N) < \widehat{f}(X_{T_n}),\, T_n = k,\, T_{n+1}-T_n \leq q,\, X_{T_n} = y\right)\\
&\leq \sum_{k=n}^{\infty}\sum_{y\in S} P_x\left(\exists N\ \ k \leq N < k+q,\, \widehat{f}(X_N) < \widehat{f}(y),\, T_n = k,\, T_{n+1} > N,\, X_{T_n} = y\right)\\
&= \sum_{k=n}^{\infty}\sum_{y\in S} P_x\left(\exists N\ \ k \leq N < k+q,\, \widehat{f}(X_N) < \widehat{f}(y),\, \tau_k > N / X_k = y\right) P_x\left(T_n = k,\, X_k = y\right)
\end{aligned}$$

Now

$$P_x\left(\exists N\ \ k \le N < k+q,\ \widehat{f}(X_N) < \widehat{f}(y),\ \tau_k > N / X_k = y\right)$$
$$\le \sum_{r=0}^{q-1} P_x\left(\widehat{f}(X_{k+r}) < \widehat{f}(y),\ \tau_k > k+r / X_k = y\right)$$
$$= \sum_{r=0}^{q-1} \sum_{\substack{z \in E^m \\ \widehat{f}(z) < \widehat{f}(y)}} P_x\left(X_{k+r} = z,\ \tau_k > k+r / X_k = y\right).$$

However, for two populations y in S and z in E^m such that $\widehat{f}(z) < \widehat{f}(y)$, the right–hand inequality of (38) together with the definition of θ_2 show that for $r < q$,

$$P_x\left(X_{k+r} = z,\ \tau_k > k+r / X_k = y\right) \le C\, l(k)^{-\theta_2}$$

where $C = \max\{\, \tilde{\gamma}_2^r : 1 \le r \le q-1\,\}$. Reporting this in the preceding inequalities, we obtain successively

$$P_x\left(\exists N\ \ k \le N < k+q,\ \widehat{f}(X_N) < \widehat{f}(y),\ \tau_k > N / X_k = y\right) \le C q\, (\operatorname{card} E^m)\, l(k)^{-\theta_2}$$

and

$$P_x\left(A_n,\ T_{n+1} - T_n \le q\right) \le C q\, (\operatorname{card} E^m)\, l(n)^{-\theta_2}.$$

Since the series $\sum_{n\ge 0} l(n)^{-\theta_2}$ converges, then the series $\sum_{n\ge 0} P_x\left(A_n,\ T_{n+1} - T_n \le q\right)$ is convergent. Hence the series $\sum_{n \ge 0} P_x\left(A_n\right)$ is also convergent, and a last application of Borel–Cantelli lemma yields (48). $\square$

12. The hierarchy of cycles

To obtain more precise conditions on the sequence $l(n)$ to ensure the settling in f^*, a deeper description of the asymptotic dynamics of the chain (X_n^l) is required. Here again, we follow the road opened by Freidlin and Wentzell. This section will be allusive; we just outline the main ideas and refer the reader to [1, chapter 6]. We first need logarithmic estimates of the transition probabilities of the chain (X_{T_n}). In order to use directly the inequalities (45) we suppose in addition that the sequence $l(n)$ satisfies the following

Property R. There exists a real number t strictly greater than one such that for all r in $\mathbb{N}$, the sequences $l(\lfloor tn \rfloor + r)$ and $l(n)$ are logarithmically equivalent, i.e.

$$\exists\, t \in \mathbb{R}\quad t > 1 \quad \forall r \in \mathbb{N} \quad \forall \epsilon > 0 \quad \exists N \in \mathbb{N} \quad \forall n \ge N$$
$$(1-\epsilon)\ln l(n) \ \le\ \ln l(\lfloor tn \rfloor + r) \ \le\ (1+\epsilon)\ln l(n).$$

This property gives us some information on the rate of increasing of the sequence $l(n)$ and is fulfilled as soon as $l(n)$ is logarithmically equivalent to a sequence of the form n^θ. It will allow us to obtain easily logarithmic estimates of the transition probabilities of the chain (X_{T_n}) and to apply various results from the Freidlin–Wentzell theory. It could be replaced by a much weaker condition : however this wouldn't change the spirit of the computations and would require much more technical tricks. We have then

$\forall \epsilon > 0 \quad \forall y, z \in S \quad \exists N \quad \forall n \geq N$

$$l(n)^{-\widetilde{V}(y,z)-\epsilon} \leq P_x\left(X_{T_{n+1}} = z / X_{T_n} = y\right) \leq l(n)^{-\widetilde{V}(y,z)+\epsilon}.$$

Thus

$$\lim_{n\to\infty} \frac{1}{\ln l(n)} \ln P_x\left(X_{T_{n+1}} = z / X_{T_n} = y\right) \;=\; -\widetilde{V}(y,z)$$

and we are in position to apply Freidlin–Wentzell estimates concerning exit times for the chain (X_{T_n}). The space S may be decomposed into a hierarchy of cycles (roughly speaking, a cycle is a set of populations such that the process will make a large number of visits to each element of the cycle before leaving it). With each cycle π is associated a constant $C(\pi)$, defined by

$$C(\pi) = \min_{g \in G(S\setminus\pi)} \sum_{(y\to z)\in g} V(y,z) - \min_{x\in\pi} \min_{g\in G_\pi(x)} \sum_{(y\to z)\in g} V(y,z).$$

In the homogeneous case, the mean time spent by the chain (X_n^l) in a cycle π before exit has asymptotics which are logarithmically of order $l^{C(\pi)}$. To ensure that

$$\forall x \in E^m \qquad \lim_{n\to\infty} P\left([X_n] \subset f^* / X_0 = x\right) = 1$$

it is necessary and sufficient that

- the set W^* is included in f^* (cf the sufficient conditions given in the homogeneous case)
- the chain (X_n) escapes from each cycle π disjoint from f^* with probability one, i.e. the rate of increasing of the sequence $l(n)$ is sufficiently slow to let the chain the time to leave all these cycles. More precisely, we should have

$$\liminf_{n\to\infty} \frac{\ln l(n)}{\ln n} \;>\; \max\{\, C(\pi) : \pi \text{ cycle}, \pi\cap f^* = \emptyset \,\}.$$

In addition, if there exists at least a cycle π included in f^* such that

$$\limsup_{n\to\infty} \frac{\ln l(n)}{\ln n} \;<\; C(\pi).$$

the chain will be trapped in f^* after a finite number of transitions. Anyway, the hierarchy of cycles depends strongly on the structure of the optimization problem, i.e. on the function f and the kernels α and β which determine the asymptotic dynamics of the chain (X_n). It seems therefore quite difficult to derive precise results (for instance results above the optimal population size) at this level of generality.

VI. Application to the classical simple genetic algorithm

We specialize our general model in order to apply our results to

1. The simple genetic algorithm

We take $E = \{0,1\}^N$ for some integer N. A point i of E is a word of length N over the alphabet $\{0,1\}$ and is noted $i = i_1 \cdots i_N$ where $i_k \in \{0,1\}$. The Hamming distance $H(i,j)$ between two points i,j of E is the number of letters where i and j differ:

$$H(i,j) = \operatorname{card}\{\, k : 1 \le k \le m,\ i_k \neq j_k \,\}.$$

The mutation kernel p_l is defined by

$$p_l(i,j) = \begin{cases} 0 & \text{if } H(i,j) > 1, \\ l^{-a} & \text{if } H(i,j) = 1, \\ 1 - Nl^{-a} & \text{if } H(i,j) = 0. \end{cases}$$

The associated kernel α is

$$\alpha(i,j) = \begin{cases} 0 & \text{if } H(i,j) > 1, \\ 1 & \text{if } H(i,j) = 1, \\ N & \text{if } H(i,j) = 0. \end{cases}$$

It is irreducible: the minimal number of transitions necessary to join two arbitrary points of E through the kernel α is $R = N$.
In order to build the crossover operator, we define now a cutting operator T_k for k in $\{1 \cdots N-1\}$; T_k maps $E \times E$ onto $E \times E$ and for i,j in E, we put $T_k(i,j) = (i',j')$ where

$$i' = i_1 \cdots i_k j_{k+1} \cdots j_N, \qquad j' = j_1 \cdots j_k i_{k+1} \cdots i_N.$$

The kernel β is then defined to be

$$\beta((i,j),(i',j')) = \operatorname{card}\{\, k : 1 \le k \le N-1,\ T_k(i,j) = (i',j') \,\}$$

and the kernel q_l is

$$q_l((i,j),(i',j')) = \beta((i,j),(i',j'))\, l^{-b} \qquad \text{if } (i,j) \neq (i',j')$$

and

$$q_l((i,j),(i,j)) = 1 - \sum_{(i',j') \neq (i,j)} \beta((i,j),(i',j'))\, l^{-b}.$$

Both hypothesis H_p and H_q are clearly fulfilled, as well as the additional assumption H_i concerning the asymptotic developments in the inhomogeneous case. We put as before

$$\Delta^\circ = \max_{i \in E\setminus f^*} \min \left\{ \max_{0\leq k<r} (f(i)-f(i^k)) : i^0 = i,\ r \leq R,\ f(i^r) > f(i),\ \prod_{0\leq k<r} H(i^k, i^{k+1}) > 0 \right\}$$

$$\Delta^* = \max_{i,j \in f^*} \min \left\{ \max_{0\leq k<r} (f(i)-f(i^k)) : i^0 = i,\ r \leq R,\ i^r = j,\ \prod_{0\leq k<r} H(i^k, i^{k+1}) > 0 \right\}$$

and let $\Delta^{\circledast} = \max(\Delta^\circ, \Delta^*)$.
Within this framework, all convergence results proved in part IV and V do apply. We restate and comment only the rough results, which take a more precise form in this context.

2. The homogeneous case

Theorem 2.1. *(Convergence of the homogeneous simple genetic algorithm)*
If

$$m \quad > \quad \frac{aN + c(N-1)\Delta^{\circledast}}{\min(a, b/2, c\delta)} \tag{50}$$

then

$$\forall x \in E^m \qquad \lim_{l\to\infty} \quad \lim_{n\to\infty} \quad P\left([X_n^l] \subset f^*/X_0^l = x\right) \;=\; 1.$$

Our condition shows that the optimization problem may be solved with a sufficiently large population size. In addition, we are completely free for the choice of the parameters a, b and c: if we take a very large c, so that $\min(a, b/2) \leq c\delta$, we see that δ does not intervene any more in the above condition. The parameter $\Delta^{\circledast}$ describes the difficulty for an individual to travel from one point to a better point. If $\Delta^{\circledast}$ is fixed and N becomes large, a population size of order CN for some constant C will always suffice to handle the problem. Finally, let us remark that our condition on the population size m is very rough: neither the fine structure of the optimization problem nor the possibility of using crossovers to travel between two populations have been taken into account to obtain it. The crucial point is that the perturbation mechanism allows the process to visit all the space, even when the random perturbations are very small. The role of the crossover is thus not fundamental (the algorithm without crossover corresponds to the case $b = \infty$): however this operator creates a lot of new possible transitions for the chain (X_n) and thus certainly decreases the values of the functions V and $\widetilde{V}$ as well as the optimal population size.

3. The inhomogeneous case

Theorem 3.1. *(Convergence of the inhomogeneous simple genetic algorithm)*
1) For the chain (X_{T_n}) to be trapped in f^* after a finite number of transitions, i.e. to have

$$\forall x \in E^m \qquad P\left(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset f^* / X_0 = x\right) = 1$$

the convergence exponent of the sequence $l(n)$ must be a positive real number; that is, there must exist two positive real numbers θ_1 and θ_2 such that

$$\sum_{n \geq 0} l(n)^{-\theta_1} = \infty \qquad \text{and} \qquad \sum_{n \geq 0} l(n)^{-\theta_2} < \infty.$$

2) If the convergence exponent λ of the sequence $l(n)$ and the population size m satisfy the inequalities

$$aN + c(N-1)\Delta^{\circledast} \; < \; \lambda \; < \; \min\left(a, b/2, c\delta\right) m$$

then, with probability one, the chain (X_{T_n}) is trapped in f^* after a finite number of transitions, i.e.,

$$\forall x \in E^m \qquad P\left(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset f^* / X_0 = x\right) = 1.$$

3) Suppose the sequence $l(n)$ verifies property R. If the convergence exponent λ of the sequence $l(n)$ and the population size m both satisfy the inequalities

$$aN + c(N-1)\Delta^{\circledast} \; < \; \min\left(a, b/2, c\delta\right) m \; \leq \; \lambda$$

then

$$\forall x \in E^m \qquad \lim_{n \to \infty} \quad P\left([X_n] \subset f^* / X_0 = x\right) \; = \; 1.$$

To ensure the convergence toward f^* in the inhomogeneous case, we must first ensure that the homogeneous algorithm converges (condition (50)) and that the rate of increasing of the sequence $l(n)$ is carefully adapted to avoid the process to be trapped in any sub–optimal population.

References

[1] M.I. Freidlin–A.D. Wentzell, *Random perturbations of dynamical systems*, Springer–Verlag, New York, 1984.

[2] D. Goldberg, *Genetic algorithms in search, optimization and machine learning*, Addison–Wesley, 1989.

[3] J.H. Holland, *Adaptation in natural and artificial systems*, The University of Michigan Press, Ann Arbor, 1975.

[4] Y. Kifer, *Random perturbations of dynamical systems*, Birkhäuser, Boston–Basel, 1988.

[5] G. Pólya–G. Szegö, *Problems and theorems in analysis I*, Springer–Verlag, Berlin Heidelberg, 1972.

[6] E. Seneta, *Non–negative matrices and Markov chains*, Springer–Verlag, New York, 1981.

Chapitre Troisième

LA DYNAMIQUE DES ALGORITHMES DE MUTATION–SÉLECTION AVEC DE GRANDES TAILLES DE POPULATION

THE DYNAMICS OF MUTATION-SELECTION ALGORITHMS WITH LARGE POPULATION SIZES

LA DYNAMIQUE DES ALGORITHMES DE MUTATION–SÉLECTION AVEC DE GRANDES TAILLES DE POPULATION

Raphaël CERF

Université de Montpellier

September 1993

Abstract. We build the mutation–selection algorithm by perturbing randomly a very simple selection scheme. Our process enters the class of the generalized simulated annealing algorithms studied by Trouvé. When the population size m is large, the various quantities associated with the algorithm are affine functions of m and the hierarchy of cycles on the set of uniform populations stabilizes. If the mutation kernel is symmetric, the limiting distribution is the uniform distribution over the set of the global maxima of the fitness function. The optimal convergence exponent defined by Azencott, Catoni and Trouvé is an affine strictly increasing function of m.

Résumé. Nous construisons l'algorithme de mutation–sélection en perturbant aléatoirement un mécanisme de sélection très simple. Le processus obtenu entre dans la classe des algorithmes de recuit simulé généralisés étudiés par Trouvé. Lorsque la taille de la population m est grande, les différentes quantités associées à l'algorithme sont des fonctions affines de m et la hiérarchie des cycles sur l'ensemble des populations uniformes se stabilise. Si le noyau de mutation est symétrique, la distribution limite est la distribution uniforme sur l'ensemble des maxima globaux de la fonction fitness. L'exposant optimal de convergence défini par Azencott, Catoni et Trouvé est une fonction affine strictement croissante de m.

1991 *Mathematics Subject Classification.* 60F10, 60J10, 92D15.

Key words and phrases. Freidlin–Wentzell theory, evolutionary algorithms, stochastic optimization.

Professor Jean–Arcady Meyer first spoke to me of evolutionary algorithms.

I am deeply grateful to Professor Gérard Ben Arous for introducing me to the Freidlin–Wentzell theory.

I thank Professor Alain Berlinet for his useful advice and for having carefully read this paper.

I thank Alain Trouvé for several discussions about the generalized simulated annealing and for comments which improved this work.

1. Introduction

Evolutionary algorithms are optimization techniques based on the mechanics of natural selection [6]. Experimental simulations have demonstrated their efficiency: they are robust, flexible and in addition extremely easy to implement on parallel computers [5].

Unfortunately, there is a critical lack of theoretical results for the convergence of this kind of algorithms. In a previous paper [3], the author proposed a Markovian model of Holland's Genetic Algorithm. This model was built by perturbing randomly a very simple selection scheme. The study of the asymptotic dynamics of the process was carried out with the powerful tools developed by Freidlin and Wentzell for the study of random perturbations of dynamical systems, yielding what seemed to be the first mathematically well–founded convergence results for genetic algorithms.

It was proved that the convergence toward the global maxima of the fitness function becomes possible when the population size is greater than a critical value (which depends on all the characteristics of the optimization problem). Surprisingly, the crossover did not play a fundamental role in this model.

In this paper, we deal with the two–operators mutation–selection algorithm. We focus on the dynamics and the critical constants of the algorithm when the population size m becomes very large. We describe the most probable trajectories of populations joining two uniform populations. We show that the decomposition in cycles of the set of uniform populations stabilizes; the structure of the limiting hierarchy of cycles depends on the relative costs of the mutations and the anti–selections and the fitness function is constant on the "bad" cycles (i.e. those which do not contain the minima of the virtual energy). Moreover, when the mutation kernel is symmetric, the limiting distribution is the uniform distribution over the set of the global maxima of the fitness function. Finally, the critical height is bounded (as a function of m) whereas the optimal convergence speed exponent defined by Azencott, Catoni and Trouvé [1,2,9,10,11] increases linearly with m.

Once more, we follow the road opened by Freidlin and Wentzell [4]. However we will use the concepts and tools introduced by Catoni in his precise study of the sequential simulated annealing [1,2]; these were put in a more general framework by Trouvé, who carried out a systematic study (initiated by Hwang and Sheu [7]) of a class of processes he baptized "generalized simulated annealing" [9,10,11]. Our algorithm belongs to this class of processes and our study will thus heavily rely on Catoni and Trouvé's work.

This paper has the following structure. First we give some notations and describe the model. Then we show how it enters the framework of the generalized simulated annealing. We examine the various quantities associated with our mutation–selection algorithm and their dependence upon the population size m: the virtual energy, the communication altitude, the communication cost, ... It turns out that they are affine function of m for m sufficiently large. We investigate the decomposition in cycles, as well as the critical constants describing the convergence of the algorithm.

2. General notations and conventions

The cardinality of a set X will be noted indifferently $|X|$ or $\operatorname{card} X$ and its characteristic function 1_X. We adopt usual conventions concerning empty sets:

$$\prod_{\emptyset} = 1, \quad \sum_{\emptyset} = 0, \quad \min \emptyset = +\infty, \quad \max \emptyset = -\infty.$$

If s is a real number, $\lfloor s \rfloor$ denotes the unique integer such that $\lfloor s \rfloor \leq s < \lfloor s \rfloor + 1$ and s^+ denotes the maximum $\max(s, 0)$. We say that a real number s is positive if $s > 0$ and that it is non–negative if $s \geq 0$.

We consider a finite space of states E and a real-valued positive non-constant function f (which will be called the fitness function) defined on E. The points of E will be called individuals and will be mostly denoted by the letters i, j, e. The set of global maxima of f is

$$f^* = \{\, i \in E : f(i) = \max_{j \in E} f(j) \,\}$$

and $f(f^*)$ is the maximum value of f over E, i.e. $\max_{j \in E} f(j)$. Symbols with a star $*$ in superscript will denote sets realizing the minimum or the maximum of a particular functional. We put

$$\delta = \min \{\, |f(i) - f(j)| : i, j \in E, f(i) \neq f(j) \,\}, \quad \Delta = \max \{\, |f(i) - f(j)| : i, j \in E \,\}.$$

The Kronecker symbol $\delta(i, j)$ will be used to denote the identity matrix indexed by E:

$$\forall i, j \in E \qquad \delta(i, j) = 0 \quad \text{if} \quad i \neq j, \quad \delta(i, j) = 1 \quad \text{if} \quad i = j.$$

The set $E^{(\mathbb{N})}$ is the set of all finite sequences of elements of E, that is $E^{(\mathbb{N})} = \bigcup_{m \in \mathbb{N}} E^m$. We will consider Markov chains with state space E^m, where m is the (variable) population size of our algorithm; the m–uples of elements of E, i.e. the points of the set E^m, are called populations and will be mostly denoted by the letters x, y, z, u.

We define a bracket operator $[\ ]$ from $E^{(\mathbb{N})}$ onto $\mathcal{P}(E)$, the set of all subsets of E, by

$$x = (x_1, \cdots, x_m) \in E^m \quad \longmapsto \quad [x] = \{\, x_k : 1 \leq k \leq m \,\}$$

i.e. $[x]$ is the set of all individuals present in the population x.

For x in E^m and i in E, $x(i)$ is the number of occurrences of i in the population x:

$$x(i) = \operatorname{card} \{\, k : 1 \leq k \leq m,\, x_k = i \,\}.$$

With f we associate a function $\widehat{f}$ defined on $E^{(\mathbb{N})}$ by

$$\widehat{f}(x) = \widehat{f}(x_1, \cdots, x_m) = \max \{\, f(x_k) : 1 \leq k \leq m \,\}.$$

For x in E^m, $\hat{x}$ denotes the set of those elements of $[x]$ which realize the value $\widehat{f}(x)$:

$$\hat{x} = \{\, x_k : 1 \leq k \leq m,\, f(x_k) = \widehat{f}(x) \,\}.$$

If i belongs to E, (i) is the m–uple whose m components are equal to i and U is the set of all such m-uples (which are called the uniform populations). We sometimes identify an element i of E with the uniform population (i). Thus f^* may be seen as a subset of U.

3. Description of the model

We model the mutation–selection algorithm as a Markov chain $(X_n^l)_{n\geq 0}$ with state space E^m (where l is a control parameter). The transition from X_n^l to X_{n+1}^l includes two stages corresponding to the mutation and the selection operators:

$$X_n^l \xrightarrow{\text{mutation}} Y_n^l \xrightarrow{\text{selection}} X_{n+1}^l.$$

We now describe more precisely these two operators.

3.1. $X_n^l \longrightarrow Y_n^l$: mutation. The mutation operator is modelled by random independent perturbations of the individuals of the population X_n^l. Let α be an irreducible Markov kernel on the space E, that is a function defined on $E \times E$ with values in $[0,1]$ satisfying

$$\forall i \in E \qquad \sum_{j\in E} \alpha(i,j) = 1,$$

$$\forall i,j \in E \quad \exists e_1, \cdots, e_r \in E \quad e_1 = i,\, e_r = j,\, \forall k \in \{1 \cdots r-1\} \quad \alpha(e_k, e_{k+1}) > 0.$$

Let a be a positive real number. Define for i,j in E

$$\alpha_l(i,j) = \begin{cases} \alpha(i,j)\, l^{-a} & \text{if } i \neq j \\ 1 - \sum_{e\neq i} \alpha(i,e)\, l^{-a} & \text{if } i = j \end{cases}$$

so that α_l is an irreducible Markov kernel on E.
The transition probabilities from X_n^l to Y_n^l are then given by

$$P\left(Y_n^l = y / X_n^l = x\right) = \alpha_l(x_1, y_1) \cdots \alpha_l(x_m, y_m).$$

Note that

$$\lim_{l\to\infty} P\left(Y_n^l = y / X_n^l = x\right) = \delta(x_1, y_1) \cdots \delta(x_m, y_m) \tag{1}$$

i.e. the mutations vanish when l goes to infinity.

3.2. $Y_n^l \longrightarrow X_{n+1}^l$: selection. The selection operator is built with a selection function [3]. Let c be a positive real number. The transition probabilities from Y_n^l to X_{n+1}^l are

$$P\left(X_{n+1}^l = x / Y_n^l = y\right) = \prod_{i\in[x]} \left(\frac{y(i)\exp(cf(i)\ln l)}{\sum_{k=1}^m \exp(cf(y_k)\ln l)} \right)^{x(i)} = \prod_{r=1}^m \frac{y(x_r)\exp(cf(x_r)\ln l)}{\sum_{k=1}^m \exp(cf(y_k)\ln l)}.$$

We have

$$\lim_{l\to\infty} P\left(X_{n+1}^l = x / Y_n^l = y\right) = \frac{1}{(\operatorname{card} \hat{y})^m} \prod_{k=1}^m 1_{\hat{y}}(x_k) y(x_k) \tag{2}$$

i.e. the selections of individuals below peak fitness tend to disappear when l goes to infinity.

3.3. The underlying scheme (X_n^∞). The previous model is built by perturbing randomly a Markov chain $(X_n^\infty)_{n\geq 0}$ on the space E^m and the parameter l governs the intensity of the perturbations; these vanish as l goes to infinity, unveiling the underlying scheme (X_n^∞):

$$\lim_{l\to\infty} P\left(X_{n+1}^l = z/X_n^l = y\right) = \prod_{k=1}^m \frac{1_{\hat{y}}(z_k)y(z_k)}{\operatorname{card}\hat{y}} = P\left(X_{n+1}^\infty = z/X_n^\infty = y\right).$$

The process (X_n^∞) is only a very simple selection scheme. The population X_{n+1}^∞ is formed by selecting independently and with the uniform distribution m individuals in the set $\hat{X}_n^\infty$. Notice that after a while, the chain (X_n^∞) starting from a population x is trapped in a uniform population selected in $\hat{x}$:

$$P\left(\exists i\in\hat{x}\quad \exists N\in\mathbb{N}\quad \forall n\geq N\quad X_n^\infty = (i)/X_0^\infty = x\right) = 1.$$

In fact, the absorbing states of the chain (X_n^∞) are exactly the uniform populations U.

To obtain the process (X_n^l), we perturb randomly (X_n^∞) in two ways: random perturbations of the individuals (mutations) and loosening of the selection scheme. A crucial point is to give the same intensity to these two kinds of perturbations in order that they interact properly when l goes to infinity. More precisely, the rate of convergence of the transition probabilities in formulas (1) and (2) should be logarithmically of the same order.

4. Asymptotic expansion of $P\left(X_{n+1}^l = z/X_n^l = y\right)$

By the very construction of the process (X_n^l), we have

$$P\left(X_{n+1}^l = z/X_n^l = y\right) = \sum_{x\in E^m} P\left(X_{n+1}^l = z/Y_n^l = x\right) P\left(Y_n^l = x/X_n^l = y\right).$$

For each x in E^m,

$$P\left(X_{n+1}^l = z, Y_n^l = x/X_n^l = y\right) \underset{l\to\infty}{\sim} \alpha(y,x)\gamma(x,z)\exp -\left(\left(ad(y,x)+c\sum_{k=1}^m(\widehat{f}(x)-f(z_k))\right)\ln l\right)$$

where we note for x, y, z in E^m

$$\alpha(y,x) = \prod_{k:y_k\neq x_k}\alpha(y_k,x_k),\quad \gamma(x,z) = \prod_{i\in[z]}\left(\frac{x(i)}{\operatorname{card}\hat{x}}\right)^{z(i)} = \prod_{r=1}^m \frac{x(z_r)}{\operatorname{card}\hat{x}}$$

and $d(y,x)$ is the Hamming distance between the vectors y and x, i.e.

$$d(y,x) = \operatorname{card}\{\,k : 1\leq k\leq m,\ y_k\neq x_k\,\}.$$

The above quantity vanishes whenever $\alpha(y,x)\gamma(x,z) = 0$. Let $\overline{D}_1(y,z)$ be the set of all three–uples (y,x,z) satisfying $\alpha(y,x)\gamma(x,z) > 0$.
We define next the communication cost V_1 on $E^m \times E^m$ by

$$V_1(y,z) = \min_{(y,x,z)\in\overline{D}_1(y,z)} ad(y,x) + c\sum_{k=1}^{m}(\widehat{f}(x) - f(z_k)).$$

Let $\overline{D}_1^*(y,z)$ be the elements of $\overline{D}_1(y,z)$ realizing the above minimum and put

$$q_1(y,z) = \sum_{(y,x,z)\in\overline{D}_1^*(y,z)} \alpha(y,x)\gamma(x,z).$$

With these notations, we see that

$$P\left(X_{n+1}^l = z/X_n^l = y\right) \underset{l\to\infty}{\sim} q_1(y,z)\exp\left(-V_1(y,z)\ln l\right).$$

Remark in addition that for each y, z in E^m,

$$P\left(X_{n+1}^l = z/X_n^l = y\right) = 0 \iff V_1(y,z) = \infty \iff q_1(y,z) = 0.$$

We are now in the framework of the generalized simulated annealing studied by Trouvé [8,9,10,11]. That is, the transition probabilities of the process (X_n^l) form a family of Markov kernels on the space E^m indexed by l which is admissible for the communication kernel q_1 and the cost function V_1 [9, Definition 3.1].

5. The virtual energy

Let g be a graph on E^m. The cost of g is

$$V_1(g) = \sum_{(x\to y)\in g} V_1(x,y).$$

We recall that an x–graph is a graph with no arrow starting from x and such that for any $y \neq x$ there exists a unique path in g leading from y to x. The set of all x–graphs is denoted by $G(x)$. For more details and notations concerning graphs, see [4, chapter 6].

Definition 5.1. The virtual energy W_1 associated with the cost function V_1 is defined by

$$\forall x \in E^m \qquad W_1(x) = \min\{\, V_1(g) : g \in G(x)\,\}.$$

We put also

$$W_1(E^m) = \min\{\, W_1(x) : x \in E^m\,\}, \qquad W_1^* = \{\, x \in E^m : W_1(x) = W_1(E^m)\,\}.$$

Proposition 5.1. *(Freidlin and Wentzell)*

$$\forall x \in E^m \qquad \lim_{l\to\infty} \quad \lim_{n\to\infty} \quad P\left(X_n^l \in W_1^*/X_0^l = x\right) \;=\; 1.$$

Theorem 5.1. *(convergence of the homogeneous algorithm)*
There exists a critical population size m^* such that

$$m \geq m^* \qquad \Longrightarrow \qquad W_1^* \subset f^*.$$

This theorem has been proved in [3] and various upper bounds for the value m^* were given there; for instance

$$m^* \;\leq\; \frac{aR + c(R-1)\Delta}{\min(a, c\delta)}$$

where R is the minimal number of transitions necessary to join two arbitrary points of E through the kernel α, i.e. R is the smallest integer such that

$$\forall i, j \in E \quad \exists\, r \leq R \quad \exists e_1, \cdots, e_{r+1} \in E \quad \text{such that}$$
$$e_1 = i,\; e_{r+1} = j, \quad \forall k \in \{1, \cdots, r\} \quad \alpha(e_k, e_{k+1}) > 0.$$

The preceding bound may be improved in several ways with more complicated constants: however, since we deal only with very large population sizes, we do not care about a precise estimation of m^*. Anyway, it is obvious that m^* depends strongly on the fitness function f and the kernel α.

We restate now Trouvé's result for the convergence in the inhomogeneous case, where the parameter l is an increasing function of n. We have then an inhomogeneous Markov chain $(X_n^{l(n)})_{n\geq 0}$ and we suppress the superscript l: X_n stands for $X_n^{l(n)}$.

Theorem 5.2. *(Trouvé,* [9, Theorem 3.23]*)*
There exists a constant H_1, called the critical height, such that for all increasing sequences $l(n)$, we have the equivalence

$$\forall x \in E^m \qquad P\left(X_n \in W_1^*/X_0 = x\right) \underset{n\to\infty}{\longrightarrow} 1 \qquad \Longleftrightarrow \qquad \sum_{n=0}^{\infty} l(n)^{-H_1} = \infty.$$

Note that H_1 may depend on m.

Corollary 5.1. *Suppose $m \geq m^*$ and $\sum_{n=0}^{\infty} l(n)^{-H_1} = \infty$. Then*

$$\forall x \in E^m \qquad P\left(X_n \in f^*/X_0 = x\right) \underset{n\to\infty}{\longrightarrow} 1.$$

All the tools introduced by Trouvé in [9,10,11] for the study of the generalized simulated annealing may be used for analyzing the mutation–selection algorithm. The particular interest of our process relies in the presence of several control parameters among which the population size m is of paramount importance. In the sequel, we will try to understand the behaviour of the mutation–selection algorithm for very large populations and we will examine how the crucial characteristics of the algorithm depend on m.

6. Some more notations

If $\mathcal{S}$ is an arbitrary set, $\mathcal{S}^{(\mathbb{N})}$ denotes the set of paths in $\mathcal{S}$, that is the finite sequences of elements of $\mathcal{S}$. A path s in $\mathcal{S}$ is noted indifferently

$$s = (s_1, \cdots, s_r), \quad s = (s^1, \cdots, s^r) \quad \text{or} \quad s = s_1 \to \cdots \to s_r$$

and its length is noted $|s|$ (r in the above example). A path s in $\mathcal{S}$ is said to join two elements t_1 and t_2 if $s_1 = t_1$ and $s_{|s|} = t_2$; the set of all paths in $\mathcal{S}$ joining the points t_1 and t_2 is noted $\mathcal{S}^{(\mathbb{N})}(t_1, t_2)$.
We will consider paths in the sets E, E^m and $\mathcal{P}(E)$. Paths of E^m will mostly be denoted by the letter p and paths of $\mathcal{P}(E)$ by the letter q.
By D^m we denote the paths in E^m which correspond to possible trajectories of the process (X_n^l), i.e. the paths p in E^m satisfying

$$\forall k \quad 1 \leq k < |p| \qquad V_1(p_k, p_{k+1}) < \infty.$$

The cost of such a path is

$$V_1(p) = \sum_{k=1}^{|p|-1} V_1(p_k, p_{k+1}).$$

This definition coincides with the cost of a graph if we consider the path as a graph over E^m. Notice that for the empty path (which has a null length), the cost is zero.
If p belongs to $E^{m(\mathbb{N})} \setminus D^m$, we put $V_1(p) = \infty$.
Similarly, by $\overline{D}^m$ we denote the paths in E^m which correspond to possible trajectories for the whole process

$$X_n^l \to Y_n^l \to X_{n+1}^l \to Y_{n+1}^l \to \cdots \to Y_{n+t-1}^l \to X_{n+t}^l$$

i.e. such a path p includes the intermediate populations Y_n^l, has an odd length and satisfies

$$\forall k \quad 1 \leq 2k < |p| \qquad \alpha(p^{2k-1}, p^{2k})\gamma(p^{2k}, p^{2k+1}) > 0.$$

The corresponding cost function $\overline{V}$ is defined by

$$\overline{V}(p) = \sum_{1 \leq 2k < |p|} a\, d(p^{2k-1}, p^{2k}) + c \sum_{h=1}^{m} \left(\widehat{f}(p^{2k}) - f(p_h^{2k+1}) \right)$$

if the path p belongs to $\overline{D}^m$ (here p_h^{2k+1} is the h–th component of the vector p^{2k+1}) and $\overline{V}(p) = \infty$ otherwise. We put also for y, z in E^m and i, j in E

$$D^m(y,z) = D^m \cap E^{m(\mathbb{N})}(y,z), \quad D(i,j) = \bigcup_{m \in \mathbb{N}^*} D^m((i),(j)), \quad D = \bigcup_{m \in \mathbb{N}^*} D^m$$

and we define similarly the sets $\overline{D}^m(y,z)$, $\overline{D}(i,j)$, $\overline{D}$ (just replace D by $\overline{D}$ in the above formulas).

The bracket operator [ ] provides a natural projection from the set $\cup_{m\in\mathbb{N}^*} E^{m(\mathbb{N})}$ onto $\mathcal{P}(E)^{(\mathbb{N})}$: with each path $p=(p_1,\cdots,p_r)$ in E^m we associate the path $[p]=([p_1],\cdots,[p_r])$ in $\mathcal{P}(E)$. We define a cost function $[V]$ on the set $\mathcal{P}(E)^{(\mathbb{N})}$ by

$$\forall q\in\mathcal{P}(E)^{(\mathbb{N})} \qquad [V](q)=\inf\{\,V_1(p):p\in\bigcup_{m\in\mathbb{N}^*}E^{m(\mathbb{N})},\,[p]=q\,\}.$$

The function $[\overline{V}]$ is defined similarly.
In the sequel, the population size m will vary and we will sometimes consider simultaneously several paths, possibly with different population sizes. Thus, if p is a path in E^m, we put $m(p)=m$.

7. General results about the virtual energy and the communication altitude

We define the minimal communication cost V for y and z in E^m by

$$V(y,z)=\inf\{\,V_1(p):p\in D^m(y,z)\,\}$$

and let $D^{m*}(y,z)$ be the paths of $D^m(y,z)$ realizing the above minimum. Similarly, we denote by $\overline{D}^{m*}(y,z)$ the elements p of $\overline{D}^m(y,z)$ such that $\overline{V}(p)=V(y,z)$.
Notice that for all x in E^m, we have $V(x,x)=0$.
We will now show that the minimal cost V contains all the relevant information to carry out the study of the dynamics of the generalized simulated annealing: the following results are not specific of the case of the mutation–selection algorithms and are valid in the general framework.
For a graph g on E^m, we note

$$V(g)=\sum_{(x\to y)\in g}V(x,y)$$

and we define the virtual energy for any element x of E^m by

$$W(x)=\min\,\{\,V(g):g\in G(x)\,\}.$$

Propositions 7.1 and 7.2 below are consequences of the more general lemma 5.6 of [10]. However we give here direct proofs which do not involve the hierarchy of cycles.

Proposition 7.1. $W=W_1$.

Proof. Since $V\le V_1$ we have clearly $W\le W_1$. The proof of the reverse inequality $W\ge W_1$ is similar to the proof of lemma 4.1 of [4, chapter 6]. □

Remark. We put $W(E^m)=\min\{\,W(x):x\in E^m\,\}$, $W^*=\{\,x\in E^m:W(x)=W_1(E^m)\,\}$. We have then $W(E^m)=W_1(E^m)$ and $W^*=W_1^*$.

The fundamental quantity to build the hierarchical decomposition of the space into cycles is the communication altitude.

Definition 7.1. (Trouvé, [9, Definition 3.15])
The communication altitude $A_1(x,y)$ between two distinct points x and y of E^m is

$$A_1(x,y) = \inf\{\max_{1\le k<|p|} W_1(p_k) + V_1(p_k,p_{k+1}) : p \in D^m(x,y)\}.$$

For any x in E^m, we put $A_1(x,x) = W_1(x)$.

The communication altitude may equivalently be defined through the cost function V.

Proposition 7.2. *Let for x and y two distinct points of E^m*

$$A(x,y) = \inf\{\max_{1\le k<|p|} W(p_k) + V(p_k,p_{k+1}) : p \in D^m(x,y)\}.$$

For x in E^m, put $A(x,x) = W(x)$.
Then $A = A_1$.

Proof. We have $W = W_1$ and $V \le V_1$: clearly $A \le A_1$.
Conversely, let p belong to $E^{m(\mathbb{N})}(x,y)$. For each k, $1 \le k < |p|$, let $p_1^k \to \cdots \to p_{n_k}^k$ be a path in E^m joining p_k and p_{k+1} such that

$$V(p_k,p_{k+1}) = V_1(p_1^k,p_2^k) + \cdots + V_1(p_{n_k-1}^k,p_{n_k}^k).$$

Consider the path $\bar p$ obtained by joining end to end all these paths:

$$p_1^1 \to \cdots \to p_{n_1-1}^1 \to p_1^2 \to \cdots \to p_{n_2-1}^2 \to \cdots \to p_1^{|p|-1} \to \cdots \to p_{n_{|p|-1}-1}^{|p|-1} \to p_{n_{|p|-1}}^{|p|-1}.$$

Let k and h be two integers such that $1 \le k < |p|$, $1 \le h < n_k$. We have

$$\begin{aligned} W_1(p_h^k) + V_1(p_h^k,p_{h+1}^k) &\le W(p_k) + V(p_k,p_h^k) + V_1(p_h^k,p_{h+1}^k) + V(p_{h+1}^k,p_{k+1}) \\ &= W(p_k) + V(p_k,p_{k+1}) \end{aligned}$$

whence

$$\max_{1\le k<|p|}\ \max_{1\le h<n_k} W_1(p_h^k) + V_1(p_h^k,p_{h+1}^k) \le \max_{1\le k<|p|} W(p_k) + V(p_k,p_{k+1}).$$

However, the left–hand side member of the above inequality coincides with

$$\max_{1\le k<|\bar p|} W_1(\bar p_k) + V_1(\bar p_k,\bar p_{k+1}).$$

Taking the minimum over all paths of $D^m(x,y)$, we obtain $A_1(x,y) \le A(x,y)$. $\square$

Another difficulty for the study of our algorithm is that the size of the state space E^m increases geometrically with m. We will therefore only examine the trace of the dynamics on the subset U of the uniform populations. Once more, the following definitions and results are not specific to mutation–selection algorithms.
Let H be a subset of E^m. We define a communication cost V_H by

$$V_H(x,y) = \inf\{\, V_1(p) : p \in D^m(x,y),\ \forall k \quad 1 < k < |p|,\ p_k \notin H\,\}.$$

We define a virtual energy W_H on H by

$$\forall x \in H \qquad W_H(x) = \min\{\, V_H(g) : g \in G_H(x)\,\}$$

where $G_H(x)$ is the set of x–graphs over H and for a graph g over H,

$$V_H(g) = \sum_{(x \to y) \in g} V_H(x,y).$$

Finally, we define a communication altitude A_H: if x and y are two distinct points of H,

$$A_H(x,y) = \inf\{\, \max_{1 \le k < |p|} W_H(p_k) + V_H(p_k, p_{k+1}) : p \in H^{(\mathbb{N})}(x,y)\,\}.$$

For any x in H, we put $A_H(x,x) = W_H(x)$.

Theorem 7.1. *Let H be a subset of E^m such that*

$$\forall x \in E^m \quad \exists y \in H \quad V(x,y) = 0.$$

Then $W_H = W$ and $A_H = A$ on the set H.

Proof. Let g belong to $G_H(x)$, where $x \in H$. By hypothesis, for each y in $E^m \setminus H$, there exists z in H such that $V(y,z) = 0$. Let $\bar{g}$ be an x–graph over E^m obtained by adding one such arrow for each point of $E^m \setminus H$. We have $V(\bar{g}) = V(g) \le V_H(g)$ whence, by taking the minimum over all x–graphs, $W(x) \le W_H(x)$.
Conversely, let g belong to $G(x)$, where x is in H, and consider the graph $\bar{g}$ of $G_H(x)$ defined by: the arrow $(y \to z)$ is in $\bar{g}$ if and only if y and z are in H and there exists a path $x^1 = y \to x^2 \to \cdots \to x^{r-1} \to x^r = z$ in E^m such that

$$\forall k \in \{1 \cdots r-1\} \quad (x^k \to x^{k+1}) \in g, \qquad \forall k \in \{2 \cdots r-1\} \quad x^k \in E^m \setminus H.$$

We have $V_H(\bar{g}) \le V_1(g)$. It follows that $W_H(x) \le W_1(x) = W(x)$ and the first equality $W_H = W$ is proved.

Let p belong to $H^{(\mathbb{N})}(x,y)$, where x,y are elements of H. For each k, $1\leq k<|p|$, let $p_1^k\to\cdots\to p_{n_k}^k$ be a path in E^m such that $p_1^k=p_k$, $p_{n_k}^k=p_{k+1}$ and

$$V_H(p_k,p_{k+1})=V_1(p_1^k,p_2^k)+\cdots+V_1(p_{n_k-1}^k,p_{n_k}^k).$$

Consider the path $\bar{p}$ obtained by joining end to end all these paths:

$$p_1^1\to\cdots\to p_{n_1-1}^1\to p_1^2\to\cdots\to p_{n_2-1}^2\to\cdots\to p_1^{|p|-1}\to\cdots\to p_{n_{|p|-1}-1}^{|p|-1}\to p_{n_{|p|-1}}^{|p|-1}.$$

Let k and h be two integers such that $1\leq k<|p|$, $1\leq h<n_k$. We have

$$\begin{aligned}W_1(p_h^k)+V_1(p_h^k,p_{h+1}^k)&\leq W_H(p_k)+V_1(p_k,p_2^k)+\cdots+V_1(p_{h-1}^k,p_h^k)+V_1(p_h^k,p_{h+1}^k)\\&\leq W_H(p_k)+V_H(p_k,p_{k+1})\end{aligned}$$

whence

$$\max_{1\leq k<|p|}\ \max_{1\leq h<n_k}\ W_1(p_h^k)+V_1(p_h^k,p_{h+1}^k)\leq\max_{1\leq k<|p|}\ W_H(p_k)+V_H(p_k,p_{k+1}).$$

However, the left–hand side member of the above inequality coincides with

$$\max_{1\leq k<|\bar{p}|}\ W_1(\bar{p}_k)+V_1(\bar{p}_k,\bar{p}_{k+1}).$$

Taking the minimum over all paths of $H^{(\mathbb{N})}(x,y)$, we obtain $A_1(x,y)=A(x,y)\leq A_H(x,y)$.
Conversely, let p belong to $E^{m(\mathbb{N})}(x,y)$. Let k belong to $\{1\cdots|p|\}$.
If p_k is in H, we put $p_1^k=p_k$ and $r_k=1$.
Suppose p_k is not in H and let p_1^k be an element of H such that $V(p_k,p_1^k)=0$ and there exist populations $p_1^\circ,\cdots,p_r^\circ$ in $E^m\setminus H$ such that $V_1(p_k,p_1^\circ)=V_1(p_1^\circ,p_2^\circ)=\cdots=V_1(p_r^\circ,p_1^k)=0$.
Let g belong to $G(p_k)$ such that $V_1(g)=W(p_k)$. We add the arrow $(p_k\to p_1^k)$ to g.
Let $\bar{g}$ be the graph on H defined by: the arrow $(y\to z)$ is in $\bar{g}$ if and only if y and z are in H and there exists a path $x^1=y\to x^2\to\cdots\to x^{r-1}\to x^r=z$ in E^m such that

$$\forall k\in\{1\cdots r-1\}\quad(x^k\to x^{k+1})\in g,\qquad\forall k\in\{2\cdots r-1\}\quad x^k\in E^m\setminus H.$$

We have $V_H(\bar{g})\leq V_1(g)=W(p_k)$.
Moreover, the graph $\bar{g}$ contains at most one cycle $p_1^k\to p_2^k\to\cdots\to p_{r_k}^k\to p_1^k$ (where the populations p_h^k, $2\leq h\leq r_k$ belong to H). If there is no cycle, we put $r_k=1$.
By the very construction of the graph $\bar{g}$, we have $W(p_k)=W(p_{r_k}^k)+V(p_{r_k}^k,p_k)$ and

$$\forall h\in\{1\cdots r_k-1\}\qquad W_H(p_h^k)+V_H(p_h^k,p_{h+1}^k)\leq V_H(\bar{g}).\tag{3}$$

Yet, there exists a sequence of populations $p^k_{r_k+1},\cdots,p^k_{n_k}$ in H such that

$$V(p_k,p_{k+1}) = V(p_k,p^k_{r_k+1}) + V(p^k_{r_k+1},p^k_{r_k+2}) + \cdots + V(p^k_{n_k-1},p^k_{n_k}) + V(p^k_{n_k},p_{k+1})$$

and each cost appearing in the right–hand side is realized by a path in $E^m \setminus H$. That is

$$\forall h \quad r_k+1 \leq h \leq n_k-1 \qquad V(p^k_h,p^k_{h+1}) = V_H(p^k_h,p^k_{h+1})$$

and there exist populations $p^-_1,\cdots,p^-_s,p^+_1,\cdots,p^+_t$ of $E^m \setminus H$ such that

$$\begin{aligned}
V(p_k,p^k_{r_k+1}) &= V_1(p_k,p^-_1) + V_1(p^-_1,p^-_2) + \cdots + V_1(p^-_s,p^k_{r_k+1}) \;, \\
V(p^k_{n_k},p_{k+1}) &= V_1(p^k_{n_k},p^+_1) + V_1(p^+_1,p^+_2) + \cdots + V_1(p^+_t,p_{k+1}) \;.
\end{aligned}$$

To obtain such a sequence $p^k_{r_k+1},\cdots,p^k_{n_k}$, we just take the successive populations of H which appear in a path of the set $D^{m*}(p_k,p_{k+1})$. Let $\bar p$ be the path obtained by joining end to end all these paths in the following way:

$$p^1_1,\cdots,p^1_{r_1},\cdots,p^1_{n_1},p^2_1,\cdots,p^2_{r_2},\cdots,p^2_{n_2},\cdots,p^{|p|-1}_1,\cdots,p^{|p|-1}_{r_{|p|-1}},\cdots,p^{|p|-1}_{n_{|p|-1}},p^{|p|}_1.$$

For each k in $\{1\cdots|p|-1\}$, we have

$$\begin{aligned}
&\forall h \qquad 1 \leq h < r_k \qquad W_H(p^k_h) + V_H(p^k_h,p^k_{h+1}) \leq W(p_k), \qquad \text{(by inequality (3))}\\
&\forall h \qquad r_k \leq h \leq n_k \qquad W_H(p^k_h) + V_H(p^k_h,p^k_{h+1}) \leq \\
&W(p^k_{r_k}) + V(p^k_{r_k},p_k) + V(p_k,p^k_{r_k+1}) + V(p^k_{r_k+1},p^k_{r_k+2}) + \cdots + V(p^k_{n_k-1},p^k_{n_k}) + V(p^k_{n_k},p_{k+1})\\
&\qquad\qquad = W(p_k) + V(p_k,p_{k+1}).
\end{aligned}$$

Thus

$$\max_{1\leq k<|p|}\ \max_{1\leq h\leq n_k}\ W_H(p^k_h) + V_H(p^k_h,p^k_{h+1}) \leq \max_{1\leq k<|p|}\ W(p_k) + V(p_k,p_{k+1})$$

(we put $p^k_{n_k+1} = p^{k+1}_1$ for k in $\{1\cdots|p|-1\}$).
However the left–hand side member of the above inequality is exactly

$$\max_{1\leq k<|\bar p|}\ W_H(\bar p_k) + V_H(\bar p_k,\bar p_{k+1}).$$

Taking the minimum over all paths of $E^{m(\mathbb{N})}(x,y)$ yields $A_H(x,y) \leq A(x,y)$. $\square$

8. Decomposition in cycles

The communication altitude yields the hierarchical decomposition of the space E^m into cycles. For the construction of cycles and the various related quantities, we refer the reader to [9,10]. Nevertheless, our notion of cycles differs slightly from Catoni and Trouvé's one. Whereas they consider as a cycle any set from which the process can't escape but without having forgotten its entrance point (when the perturbations are small), we merely consider the cycles as the most attractive and stable sets we can build for the perturbed process. We thus eliminate some cycles appearing in Trouvé's work: not all singletons are cycles (only those which are local minima of the virtual energy) and our cycles are all strict in the sense of Trouvé [9, Definition 3.20]. As a consequence, the unperturbed process will never leave a cycle. We believe that this point of view is closer to the initial notion introduced by Freidlin and Wentzell.

Definition 8.1. Let $\lambda \in \mathbb{R}$. We define an equivalence relation $\mathcal{R}_\lambda$ on the set

$$W_\lambda = \{\, x \in E^m : W(x) \leq \lambda \,\}$$

by

$$\forall y, z \in W_\lambda, \qquad y\, \mathcal{R}_\lambda\, z \quad \iff \quad A(y,z) \leq \lambda.$$

Proposition 8.1. *(Trouvé,* [9, Proposition 3.21]*)*
The set of cycles in E^m associated with the cost function V_1 and the kernel q_1 is

$$\mathcal{C}(E^m) \;=\; \bigcup_{\lambda \in \mathbb{R}^+} W_\lambda \,/\, \mathcal{R}_\lambda$$

where $W_\lambda \,/\, \mathcal{R}_\lambda$ is the quotient set of the equivalence classes of W_λ for the relation $\mathcal{R}_\lambda$.

Unfortunately, the set of cycles $\mathcal{C}(E^m)$ increases dramatically with m. Since we are mostly interested in the way the mutation–selection algorithm visits the set of uniform populations U (which are the attractors of the unperturbed process), we will study the projection of the cycles on U.
We define a projection T_U from the set $\mathcal{P}(E^m)$ onto E (or equivalently U, since we may identify U and E) by

$$\forall F \in \mathcal{P}(E^m) \qquad T_U(F) = \{\, i \in E : (i) \in F \,\}$$

and we put

$$\mathcal{C}_U(E^m) = \{\, T_U(\pi) : \pi \in \mathcal{C}(E^m) \,\}\,.$$

Since the set U of the uniform populations verify

$$\forall x \in E^m \quad \exists y \in U \qquad V(x,y) = 0$$

we are in position to apply theorem 7.1: the communication altitude and the virtual energy on the set U may be evaluated by considering only paths in U, with either the cost function V_U or V. As a consequence, we have the following

Theorem 8.1. *The set of cycles in U associated with the cost function V_U i.e.*

$$\mathcal{C}(U) = \bigcup_{\lambda\in\mathbb{R}^+} W_\lambda \cap U \,/\, \mathcal{R}_\lambda$$

coincides with the set $\mathcal{C}_U(E^m)$, which is the projection of the cycles in E^m on U.

Notice that the cardinality of the set U is equal to $|E|$ and does not depend on m. Our next task is to study the cost function V_U, or equivalently (by propositions 7.1 and 7.2 and theorem 7.1), the restriction of V to U, as a function of m. We will often omit the parenthesis when speaking of uniform populations: for instance $V(i,j)$ will stand for $V((i),(j))$.
The crucial result is that, for m sufficiently large,

9. $V(i,j)$ IS AN AFFINE FUNCTION OF m

Let us introduce some notations and definitions.
First, we put $[\overline{D}] = \{\, [p] : p \in \overline{D} \,\}$ and for i,j in E,

$$\mathcal{P}(i,j) = \big\{\, [p] : p \in \overline{D}(i,j) \,\big\}\,.$$

Let q be a path in $\mathcal{P}(E)$. We say that the path $e_1 \to \cdots \to e_r$ in E is admissible for q if $r = |q|$, $\forall k \in \{1\cdots|q|\}\quad e_k \in q_k$ and

$$\forall k \quad 1 \le 2k < |q| \qquad \text{either} \quad e_{2k-1} = e_{2k} \quad \text{or} \quad \alpha(e_{2k-1}, e_{2k}) > 0.$$

The set of all admissible paths for q is denoted by $\mathcal{A}(q)$. For instance, if p belongs to $\overline{D}$, for each h in $\{1\cdots m(p)\}$, the path $p_h^1 \to \cdots \to p_h^{|p|}$ is admissible for $[p]$ (here p_h^k denotes the h–th component of the vector p^k).
We define a function Ω on the set $\mathcal{P}(E)^{(\mathbb{N})}$ by: if $q \in \mathcal{P}(E)^{(\mathbb{N})} \setminus [\overline{D}]$, then $\Omega(q) = \infty$; if $q \in [\overline{D}]$, then we put

$$\Omega(q) = \min\Big\{ \sum_{1\le 2k<|q|} a(1-\delta(e_{2k-1},e_{2k})) + c\big(\widehat{f}(q_{2k}) - f(e_{2k+1})\big) : (e_k)_{1\le k\le|q|} \in \mathcal{A}(q) \Big\}.$$

We use the function Ω defined on $\mathcal{P}(E)^{(\mathbb{N})}$ to build a function Ω on $E\times E$:

$$\forall i,j \in E \qquad \Omega(i,j) = \inf\big\{\, \Omega(q) : q \in \mathcal{P}(i,j) \,\big\}.$$

Since the set $\mathcal{P}(i,j)$ is never empty (and is included in $\overline{D}$), the quantity $\Omega(i,j)$ is finite.
We denote by $\mathcal{P}^*(i,j)$ the elements of $\mathcal{P}(i,j)$ realizing the above minimum.
Let p be a path in E^m and $\tilde{p}$ a path in $E^{\tilde{m}}$.
We say that $\tilde{p}$ is included in p (noted $\tilde{p} \subset p$) if $|p| = |\tilde{p}|$ and

$$\forall k \in \{1\cdots|p|\} \quad \forall i \in E \qquad \tilde{p}^k(i) \le p^k(i).$$

Thus the path $\tilde{p}$ may be obtained from the path p by destroying some individuals in each population of the path.
Let q be a path in $\mathcal{P}(E)$. Among the paths p such that $[p] = q$, we single out the paths which are minimal with respect to the above inclusion relation:

Definition 9.1. The path p in $\overline{D}^m$ is a minimal path realizing the path q in $\mathcal{P}(E)$ if $[p]=q$ and for each $\tilde{m}$ in $\mathbb{N}^*$ and each path $\tilde{p}$ in $\overline{D}^{\tilde{m}}$, we have the implication

$$[\tilde{p}] = q \quad \text{and} \quad \tilde{p} \subset p \quad \Longrightarrow \quad m = \tilde{m}$$

(whence also $\forall k \in \{1 \cdots |p|\} \quad \forall i \in E \quad \tilde{p}^k(i) = p^k(i)$).
The path p in $\overline{D}^m$ is minimal if it is a minimal path realizing the path $[p]$.

This definition has the following interpretation: the path p is minimal if we can't destroy a fixed number of individuals in each population of p and still have a path belonging to $\overline{D}$ without altering $[p]$.
We have the following upper bound for the population sizes of minimal paths:

Lemma 9.1. *Let p be a minimal path realizing the path q in $\mathcal{P}(E)$. Then*

$$m(p) \quad \leq \quad \max_{1 \leq 2k < |q|} (\mathrm{card}\, q_{2k-1})(\mathrm{card}\, q_{2k}) \quad \leq \quad (\mathrm{card}\, E)^2.$$

Proof. Suppose

$$\forall k \quad 1 \leq 2k < |q| \qquad m(p) \; > \; (\mathrm{card}\, q_{2k-1})(\mathrm{card}\, q_{2k}).$$

For each k, $1 \leq 2k < |q|$, consider the pairs $(p_h^{2k-1}, p_h^{2k})_{1 \leq h \leq m(p)}$: they belong to the set $q_{2k-1} \times q_{2k}$ whose cardinality is strictly less than $m(p)$; necessarily at least two pairs are identical. We choose an index $h(k)$ such that the pair $(p_{h(k)}^{2k-1}, p_{h(k)}^{2k})$ is present twice. Let $\tilde{p}$ be a path obtained in the following way:
- for each k, $1 \leq 2k < |p|$, we remove the individuals $p_{h(k)}^{2k-1}$ and $p_{h(k)}^{2k}$ from p^{2k-1} and p^{2k} to obtain the populations $\tilde{p}^{2k-1}$ and $\tilde{p}^{2k}$;
- to build $\tilde{p}_{|q|}$, we remove from the population $p_{|q|}$ an individual which is present twice (such an individual necessarily exists since $[p_{|q|}] \subset [p_{|q|-1}]$ and the individuals of $p_{|q|-1}$ are not all distinct).

Finally we have for this path $\tilde{p}$

$$\tilde{p} \in \overline{D}, \quad m(\tilde{p}) = m(p) - 1, \quad [\tilde{p}] = q \quad \text{and} \quad \tilde{p} \subset p,$$

thus contradicting the minimality of p. $\square$

We define a function θ on $\mathcal{P}(E)^{(\mathbb{N})}$ which associates with each path q the value

$$\theta(q) = \min \left\{ \overline{V}(p) - m(p)\,\Omega([p]) : p \text{ minimal path realizing } q \right\}.$$

By the very definition of Ω, all paths p satisfy the inequality $\overline{V}(p) \geq m(p)\,\Omega([p])$ so that the quantity $\theta(q)$ is non–negative. More precisely, we have the following

Lemma 9.2. *For all paths p belonging to $\overline{D}$,*

$$\overline{V}(p) \quad \geq \quad \theta([p]) + m(p)\,\Omega([p]).$$

Proof. Let p be a path of $\overline{D}$. Necessarily, there exists a minimal path $\tilde{p}$ included in p which realizes the path $[p]$. We may rearrange the individuals of the populations $p^1, \cdots, p^{|p|}$ so that $\tilde{p}$ appears as the trajectories of the first $m(\tilde{p})$ components of the path p, that is $\tilde{p}^k = (p_1^k, \cdots, p_{m(\tilde{p})}^k)$ for each k in $\{1 \cdots |p|\}$. Although this reordering might be quite complex, it does not affect the cost of the path p. Yet

$$\begin{aligned}
\overline{V}(p) \quad &= \quad \sum_{h=1}^{m(p)} \quad \sum_{1\leq 2k<|p|} a\big(1-\delta(p_h^{2k-1},p_h^{2k})\big) + c\big(\widehat{f}(p^{2k}) - f(p_h^{2k+1})\big)\\
&= \quad \overline{V}(\tilde{p}) \quad + \quad \sum_{h=m(\tilde{p})+1}^{m(p)} \quad \sum_{1\leq 2k<|p|} a\big(1-\delta(p_h^{2k-1},p_h^{2k})\big) + c\big(\widehat{f}(p^{2k}) - f(p_h^{2k+1})\big).
\end{aligned}$$

By the very definition of Ω, we have for all h in $\{1 \cdots m(p)\}$

$$\Omega([p]) \quad \leq \quad \sum_{1\leq 2k<|p|} a\big(1-\delta(p_h^{2k-1},p_h^{2k})\big) + c\big(\widehat{f}(p^{2k}) - f(p_h^{2k+1})\big)$$

(remark that since p belongs to $\overline{D}$, the path $p_h^1 \to \cdots \to p_h^{|p|}$ is an admissible path for $[p]$). It follows that

$$\overline{V}(p) \;\geq\; \overline{V}(\tilde{p}) + (m(p) - m(\tilde{p}))\,\Omega([p]) \;=\; \overline{V}(\tilde{p}) - m(\tilde{p})\,\Omega([\tilde{p}]) + m(p)\Omega([p]).$$

Now $\tilde{p}$ is a minimal path realizing $[p]$, whence $\overline{V}(\tilde{p}) - m(\tilde{p})\,\Omega([\tilde{p}]) \geq \theta([p])$ and finally

$$\overline{V}(p) \geq \theta([p]) + m(p)\Omega([p]). \quad \square$$

Finally, we put for i, j in E

$$\theta(i,j) = \inf\,\big\{\,\theta(q) : q \in \mathcal{P}^*(i,j)\,\big\}$$

and

$$m(i,j) = \min\big\{\,m(p) : p \in \overline{D},\ [p] \in \mathcal{P}^*(i,j),\ p \text{ minimal},\ \overline{V}(p) - m(p)\Omega([p]) = \theta(i,j)\,\big\}.$$

Notice that $m(i,j)$, being the population size of a minimal path, is smaller than $|E|^2$. We are now in position to state our fundamental

Theorem 9.1. *For all pair (i,j) of points of E, there exists an integer M such that*

$$\forall m \geq M \qquad V(i,j) \;=\; \theta(i,j) \;+\; m\,\Omega(i,j).$$

Proof. Let i,j be two points of E. Let m be an integer greater than $m(i,j)$ and let p be an element of $\overline{D}^{m(i,j)}$ satisfying: $[p] \in \mathcal{P}^*(i,j)$, p is minimal, $\overline{V}(p) - m(p)\,\Omega([p]) = \theta(i,j)$. Let $e_1^* \to \cdots \to e_{|p|}^*$ be an admissible path for $[p]$ such that

$$\Omega([p]) \quad = \quad \sum_{1 \leq 2k < |q|} a(1 - \delta(e_{2k-1}^*, e_{2k}^*)) + c\big(\widehat{f}(p_{2k}) - f(e_{2k+1}^*)\big)$$

i.e. this path realizes the minimum defining $\Omega([p])$.
Consider the path $\bar{p}$ in E^m defined by

$$\forall k \in \{1 \cdots |p|\} \qquad \bar{p}^k = (\underbrace{\;p^k\;}_{E^{m(i,j)}},\; \underbrace{e_k^*,\; \cdots,\; e_k^*}_{E^{m-m(i,j)}}\;)$$

obtained by adding $m - m(i,j)$ copies of the path $(e_k^*)_{1 \leq k \leq |p|}$ to the path p. We have

$$\begin{aligned}
\overline{V}(\bar{p}) \;&=\; \overline{V}(p) \;+\; (m - m(i,j))\,\Omega(i,j) \\
&=\; \overline{V}(p) \;-\; m(p)\,\Omega([p]) \;+\; m\,\Omega(i,j) \\
&=\; \theta(i,j) \;+\; m\,\Omega(i,j)
\end{aligned}$$

whence clearly

$$\forall m \geq m(i,j) \qquad V(i,j) \;\leq\; \theta(i,j) + m\,\Omega(i,j).$$

Conversely, put

$$M(i,j) \quad = \quad \sup\; \left\{ \frac{\theta(i,j) - \theta(q)}{\Omega(q) - \Omega(i,j)} : q \in \mathcal{P}(i,j),\, \Omega(q) \neq \Omega(i,j) \right\}.$$

By the very definition of Ω, which is a sum of terms of the form a and $c\,|f(j_1) - f(j_2)|$, we have the implication

$$q \in \mathcal{P}(i,j),\, \Omega(q) \neq \Omega(i,j) \quad \Longrightarrow \quad \Omega(q) \geq \Omega(i,j) \;+\; \min(a, c\delta).$$

Since in addition $\theta(q)$ is non–negative, we see that

$$M(i,j) \;\leq\; \frac{\theta(i,j)}{\min(a, c\delta)} \;<\; \infty$$

so that $M(i,j)$ is finite (and independent of m: neither $\mathcal{P}(i,j)$, nor θ, nor Ω depend on m).

Let m be an integer such that $m \geq m(i,j)$ and $m > M(i,j)$. Let p belong to $\overline{D}^m(i,j)$. Suppose $\Omega([p]) \neq \Omega(i,j)$. Then, by definition of $M(i,j)$, we have

$$\theta(i,j) + m\,\Omega(i,j) \;<\; \theta([p]) + m\,\Omega([p]).$$

However we know already (by lemma 9.2) that $\overline{V}(p) \geq \theta([p]) + m\,\Omega([p])$ and (by the first part of the proof) $V(i,j) \leq \theta(i,j) + m\,\Omega(i,j)$, whence $\overline{V}(p) > V(i,j)$ and p does not belong to $\overline{D}^{m*}(i,j)$.
Henceforth each path p of $\overline{D}^{m*}(i,j)$ satisfies $\Omega([p]) = \Omega(i,j)$, or equivalently $[p] \in \mathcal{P}^*(i,j)$. For such a path p, we have also $\theta([p]) \geq \theta(i,j)$ (by the very definition of $\theta(i,j)$). Application of lemma 9.2 yields the desired inequality $\overline{V}(p) \geq \theta(i,j) + m\,\Omega(i,j)$ from which we deduce

$$V(i,j) \;\geq\; \theta(i,j) + m\,\Omega(i,j). \quad \square$$

Remark 1. This theorem may be interpreted in the following way: for $m \geq M$, each new individual added to the population follows the ideal path $e_1^* \to \cdots \to e_{|p|}^*$ which realizes the minimal cost $\Omega(i,j)$. However, it is necessary to have at least $m(i,j)$ individuals for such a path to become possible, i.e. the first $m(i,j)$ individuals are used to build the path p realizing the minimal cost $\Omega(i,j)$.

Remark 2. Another way to understand this result is the following: to travel between the uniform populations (i) and (j), some events of positive cost must necessarily take place, i.e. specific sequences of mutations and selections of individuals below peak fitness (these events will be called anti–selections). We distinguish informally two kinds of such events:
- the local events which affect only a limited number of individuals,
- the collective events which affect almost all the individuals.

As the reader would have guessed, $\theta(i,j)$ corresponds to the cost of the local events whereas $\Omega(i,j)$ corresponds to the cost of the collective events.

Remark 3. We have no useful practical information about the values $m(i,j)$ and $M(i,j)$: they depend strongly on the structure of the problem, that is the communication kernel α and the function f. Although they may be very large in theory, it is likely that in most cases they will be "reasonably" small: for instance if (i) and (j) are very close (i.e. the points i and j may be joined through the kernel α with a short path) or if $\Omega(i,j)$ is null.

Remark 4. We have also $V(i,j) \geq m\,\Omega(i,j)$ for each m in $\mathbb{N}$. This inequality is an improvement of lemma 12.1 of [3] which was the key to prove the existence of the critical population size m^*. In particular, it could be used to obtain better bounds on m^*. However, in this paper, we deal only with large values of m.

Of course, the coefficients $\Omega(i,j)$ and $\theta(i,j)$ are of crucial interest: the next two sections are devoted to their study.

10. The coefficient $\Omega(i,j)$

Proposition 10.1. *(upper bound with selection)*

Define

$$\omega(i,j)=\inf\big\{\max_{1\leq k\leq r} f(e_k)-f(j) : e_1=i,\, e_r=j,\, \forall k\in\{1\cdots r-1\}\quad \alpha(e_k,e_{k+1})>0\big\}.$$

We have

$$\Omega(i,j)\ \leq\ c\,\omega(i,j)$$

and

$$a\ >\ c\,\omega(i,j)\quad\Longrightarrow\quad \Omega(i,j)\ =\ c\,\omega(i,j).$$

Remark. Notice here something very queer: the quantity $\omega(i,j)$ plays also an essential role in the theory of the sequential simulated annealing, although this algorithm is always studied as a minimization procedure.

Proof. Let $e_1=i\to e_2\to\cdots\to e_r=j$ be a path in E such that $\alpha(e_k,e_{k+1})>0$ for all k in $\{1\cdots r-1\}$. Put $q_1=\{e_1\}$,

$$q_{2k}=q_{2k+1}=\{e_1,\cdots,e_k,e_{k+1}\}\quad\text{for } k \text{ in } \{1\cdots r-2\},$$

and $q_{2r-2}=\{e_1,\cdots,e_r\}$, $q_{2r-1}=\{j\}$.
Clearly, the path $q:q_1\to\cdots\to q_{2r-1}$ belongs to $\mathcal{P}(i,j)$. Let $(e_k^*)_{1\leq k\leq 2r-1}$ be a path in E such that $e_1^*=e_1$, $e_2^*=e_1^*$,

$$e_{2k-1}^*\in\hat{q}_{2k-1},\, e_{2k}^*=e_{2k-1}^*\qquad\text{for } k \text{ in } \{2\cdots r-1\},$$

and $e_{2r-1}^*=j$. For this path, which belongs to $\mathcal{A}(q)$ and which involves no mutations, we have

$$\sum_{1\leq 2k<|q|} a\left(1-\delta(e_{2k-1}^*,e_{2k}^*)\right)+c\left(\widehat{f}(q_{2k})-f(e_{2k+1}^*)\right)=$$
$$\sum_{1\leq 2k<|q|} c\left(\widehat{f}(q_{2k})-f(e_{2k+1}^*)\right)=c\left(\widehat{f}(q_{2r-2})-f(j)\right)=c\left(\max_{1\leq k\leq r} f(e_k)-f(j)\right)$$

from which we deduce $\Omega(i,j)\ \leq\ c\,\omega(i,j)$.
Suppose now $a>c\,\omega(i,j)$. Let the pair $(q,(e_k^*)_{1\leq k\leq |q|})$ realize the minimum in $\Omega(i,j)$ (where q belongs to $[\overline{D}]$ and $(e_k^*)_{1\leq k\leq |q|}$ is an admissible path for q). Since $\Omega(i,j)<a$, necessarily

$$\forall k\quad 1\leq 2k<|q|,\qquad e_{2k}^*=e_{2k-1}^*$$

and

$$\Omega(i,j) = \sum_{1\le 2k<|q|} \big(\widehat{f}(q_{2k}) - f(e^*_{2k+1})\big).$$

Since the path e^* realizes the minimum of the above quantity among all the paths admissible for q, we have in addition

$$\forall k \quad 1 \le 2k < |q|, \qquad e^*_{2k+1} \in \hat{q}_{2k+1}$$

so that

$$\Omega(i,j) = c \sum_{1\le 2k<|q|} \big(\widehat{f}(q_{2k}) - \widehat{f}(q_{2k+1})\big).$$

Yet $e^*_{2k-1} = e^*_{2k}$ whence $\widehat{f}(q_{2k}) \ge \widehat{f}(q_{2k-1})$ and $\widehat{f}(q_{2k}) - \widehat{f}(q_{2k+1}) \ge \big(\widehat{f}(q_{2k-1}) - \widehat{f}(q_{2k+1})\big)^+$.
Let r be an index such that $\widehat{f}(q_{2r-1}) = \max_{1\le 2k-1\le|q|} \widehat{f}(q_{2k-1})$.
We have (even if $2r-1 = |q|$)

$$\Omega(i,j) \;\ge\; c \sum_{2r-1\le 2k<|q|} \big(\widehat{f}(q_{2k-1}) - \widehat{f}(q_{2k+1})\big)^+ \;\ge\; c\,\big(\widehat{f}(q_{2r-1}) - f(j)\big)$$

whence

$$\Omega(i,j) \;\ge\; c\,\big(\max_{1\le 2k-1\le|q|} \widehat{f}(q_{2r-1}) - f(j)\big). \tag{4}$$

Necessarily, there exists a path $\tilde{e}$ admissible for q such that

$$\forall k \quad 1 \le 2k < |q|, \qquad \tilde{e}_{2k} = \tilde{e}_{2k+1}$$

(this path corresponds for instance to the creation of the individual j). Clearly

$$\max_{1\le k\le|q|} f(\tilde{e}_k) - f(j) \;\ge\; \omega(i,j) \tag{5}$$

(just delete all the elements $\tilde{e}_k$ of the path $(\tilde{e}_k)_{1\le k\le|q|}$ such that $\tilde{e}_k = \tilde{e}_{k-1}$ to obtain a path admissible for evaluating $\omega(i,j)$). However

$$\max_{1\le k\le|q|} f(\tilde{e}_k) \;=\; \max_{1\le 2k-1\le|q|} f(\tilde{e}_{2k-1}) \;\le\; \max_{1\le 2k-1\le|q|} \widehat{f}(q_{2k-1})$$

which, together with inequalities (4) and (5), imply $\Omega(i,j) \ge c\,\omega(i,j)$. $\square$

Proposition 10.2. *(upper bound with mutation)*

Define

$$\varpi(i,j) = \inf\ \big\{\ card\{\ k : 1 \leq k \leq r,\ e_{2k-1} \neq e_{2k}\ \}\ \big\}$$

where the infimum is taken over the set of paths $e_1 \to \cdots \to e_{2r+1}$ in E joining i and j (the length $2r+1$ is also variable) and satisfying

- $\forall k \in \{1 \cdots r\}$ *either* $e_{2k} = e_{2k-1}$
  or $\alpha(e_{2k-1}, e_{2k}) > 0, \quad f(e_{2k-1}) > \max\limits_{2k-1<h\leq 2r+1} f(e_h);$
- *for each k in $\{1 \cdots r\}$, $f(e_{2k+1}) \geq f(e_{2k})$ and there exists a path $j^k : j_1^k \to \cdots \to j_{n_k^k}^k$ of length n_k^k in E joining i and e_{2k+1} (i.e. $j_1^k = i$, $j_{n_k^k}^k = e_{2k+1}$) and $k-1$ integers $n_1^k, \cdots, n_{k-1}^k$ such that $(n_0^k =)1 < n_1^k < \cdots < n_{k-1}^k < n_k^k$ and*

$$\forall h \in \{0 \cdots k-1\} \quad \forall t \in \{n_h^k \cdots n_{h+1}^k - 1\} \qquad f(j_t^k) \leq f(e_{2h+1}).$$

Finally, for each h in $\{1 \cdots n_k^k - 1\}$, $\quad \alpha(j_h^k, j_{h+1}^k) > 0.$

We have

$$\Omega(i,j)\ \leq\ a\,\varpi(i,j)$$

and

$$c\delta\ >\ a\,\varpi(i,j) \quad \Longrightarrow \quad \Omega(i,j)\ =\ a\,\varpi(i,j).$$

Remark. To obtain an upper bound for $\Omega(i,j)$ we will build a path in E^m joining the populations (i) and (j), containing only $\varpi(i,j)$ collective events, which are mutations. The path $e_1 \to \cdots \to e_{2r+1}$ is the collective path followed by almost all individuals. The indexes k such that $e_{2k-1} \neq e_{2k}$ correspond to the collective mutations (that is almost all individuals mutate from e_{2k-1} to e_{2k}): such an event is justified if and only if $f(e_{2k-1}) > \max_{2k-1<h\leq 2r+1} f(e_h)$. Since our path should not contain collective anti–selections events we impose also $f(e_{2k+1}) \geq f(e_{2k})$. Finally, the path j^k corresponds to the sequence of mutations leading to the appearance of e_{2k+1}: this creation should not interfere with the collective behavior of the other individuals (i.e. it should involve only local events) and the condition required on j^k (i.e. the existence of the subdivision n^k and the linked inequalities $f(j_t^k) \leq f(e_{2h+1})$) exactly tells that e_{2k+1} may be created from i by a path of individuals whose fitness is always less than the maximal fitness of the current population. With the elements in hand, the only delicate problem to build the path is to stay the right amount of time in each state e_{2k+1} before leaving in order to let the time to the path j^k to be completed without disturbing the collective evolution of the population. This is the object of the (quite intricate) construction done in the first part of the proof.

Finally, this result, which describes the situation dual to proposition 10.1, will not be used in the sequel, and the following (painful) proof may well be skipped. Its main interest is to give some insight into the structure of the trajectories of the populations.

Proof. Let $e_1 = i \to \cdots \to e_{2r+1} = j$ be a path in E verifying the conditions imposed on the paths used for defining ϖ and let $(j^k)_{1\leq k\leq r}$ and $(n^k_h)_{1\leq k\leq h\leq r}$ be the associated paths and subdivisions. Put $n_0 = 1$, $n_1 = \max_{1\leq k\leq r} n^k_1$, define by induction for h in $\{2\cdots r-1\}$

$$n_h = n_{h-1} + \max_{h\leq k\leq r} n^k_h - n^k_{h-1}$$

and finally $n_r = n_{r-1} + n^r_r - n^r_{r-1}$. We define now a path p of length $2n_r - 1$ in E^{r+1}. First, we let $p^1 = (i)$. For h and s such that $0 \leq s < h \leq r$, we put

$$p_h^{2(n_s+t-1)} = \begin{cases} j^h_{n^h_s+t} & \text{if} \quad 1 \leq t < n^h_{s+1} - n^h_s \\ j^h_{n^h_{s+1}-1} & \text{if} \quad n^h_{s+1} - n^h_s \leq t < n_{s+1} - n_s \\ j^h_{n^h_{s+1}} & \text{if} \quad t = n_{s+1} - n_s \end{cases}$$

and

$$p_h^{2t+1} = p_h^{2t} \qquad 3 \leq 2t+1 \leq 2n_h - 1.$$

For h and s such that $1 \leq h \leq s < r$, we put $p_h^{2n_{s+1}-2} = e_{2s+2}$ and

$$p^t_h = e_{2s+1} \qquad 2n_s - 1 \leq t < 2n_{s+1} - 2.$$

For the $(r+1)$–th individual, for each s in $\{0\cdots r-1\}$,

$$p^t_{r+1} = \begin{cases} e_{2s+1} & \text{if} \quad 2n_s - 1 \leq t < 2n_{s+1} - 2 \\ e_{2s+2} & \text{if} \quad t = 2n_{s+1} - 2 \end{cases}$$

Finally, for each h in $\{1\cdots r+1\}$, $p_h^{2n_r-1} = e_{2r+1} = j$. We then consider the path $(e^*_k)_{1\leq k\leq 2n_r-1}$ followed by the last individual: $e^*_k = p^k_{r+1}$, $1 \leq k \leq 2n_r - 1$. We have

$$\begin{aligned} \sum_{1\leq 2k<|p|} a\left(1-\delta(e^*_{2k-1},e^*_{2k})\right) + c\left(\widehat{f}(p_{2k}) - f(e^*_{2k+1})\right) &= \sum_{s=0}^{r-1} \quad \sum_{2n_s-1\leq 2k<2n_{s+1}-1} \cdots \\ &= \sum_{s=0}^{r-1} a\left(1-\delta(e_{2s+1},e_{2s+2})\right) \end{aligned}$$

from which we deduce the inequality $\Omega(i,j) \;\leq\; a\,\varpi(i,j)$.

Suppose now $c\delta > a\,\varpi(i,j)$. Let $(q,(e^*_k)_{1\leq k\leq |q|})$ be a pair realizing the minimum in $\Omega(i,j)$ (where q belongs to $[\overline{D}(i,j)]$ and e^* is a path in E admissible for q). Thus

$$\Omega(i,j) = \sum_{1\leq 2k<|q|} a(1-\delta(e^*_{2k-1},e^*_{2k})) + c\big(\widehat{f}(q_{2k}) - f(e^*_{2k+1})\big)$$

and since $c\delta > a\,\varpi(i,j) \geq \Omega(i,j)$, necessarily for each k, $1\leq 2k < |q|$, we have $e^*_{2k+1}\in \hat{q}_{2k}$ (whence in particular $f(e^*_{2k+1}) \geq f(e^*_{2k})$) and $\widehat{f}(q_{2k+1}) = \widehat{f}(q_{2k})$.
Let us show that the path $(e^*_k)_{1\leq k\leq |q|}$ satisfies the conditions imposed on the paths defining the infimum in $\varpi(i,j)$. Let k be an integer such that $1\leq 2k<|q|$. Suppose $e^*_{2k}\neq e^*_{2k-1}$. Then $\alpha(e^*_{2k-1},e^*_{2k}) > 0$. Suppose by absurd there exists and index h, $2k-1 < 2h+1 \leq |q|$, such that $f(e^*_{2k-1}) \leq f(e^*_{2h+1})$. We define now a new pair $(\tilde{q},\tilde{e}^*)$; we put

$$\tilde{q}_t = \begin{cases} q_t & \text{if} \quad 1\leq t<2k, \quad 2h<t\leq |q| \\ q_t \cup \{e^*_{2k-1}\} & \text{if} \quad 2k\leq t\leq 2h \end{cases}$$

and we choose a sequence $(\tilde{e}^*)$ such that

$$\begin{array}{lll} \tilde{e}^*_t & = e^*_t & 1\leq t<2k, \quad 2h<t\leq |q| \\ \tilde{e}^*_{2t+1} & \in \hat{\tilde{q}}_{2t+1} & 2k\leq 2t+1\leq 2h \\ \tilde{e}^*_{2t} & = e^*_{2t-1} & 2k\leq 2t<2h \end{array}$$

The pair $(\tilde{q},\tilde{e}^*)$ satisfies the conditions related to the definition of the infimum in $\Omega(i,j)$ and

$$\sum_{1\leq 2k<|\tilde{q}|} a(1-\delta(\tilde{e}^*_{2k-1},\tilde{e}^*_{2k})) + c\big(\widehat{f}(\tilde{q}_{2k}) - f(\tilde{e}^*_{2k+1})\big)$$

$$\leq\ a\ +\sum_{1\leq 2k<|q|} a(1-\delta(e^*_{2k-1},e^*_{2k})) + c\big(\widehat{f}(q_{2k}) - f(e^*_{2k+1})\big)$$

which contradicts the fact that (q,e^*) realizes the value $\Omega(i,j)$.
Thus

$$f(e^*_{2k-1}) < \max_{2k-1<h\leq |q|} f(e^*_h).$$

The remaining condition, i.e. the existence of the paths (j^k) and the subdivision (n^h_k), is a consequence of the following facts:

- the path q belongs to $[\overline{D}]$
- for each k, $1\leq 2k<|q|$, the individual e^*_{2k+1} belongs to $\hat{q}_{2k}$.

To find the path j^k, we just look at the sequence of mutations which leads to the creation of the individual e^*_{2k+1} in the sequence of populations $q^1\to\cdots\to q^{2k}$.
Since e^* satisfies the conditions imposed on the paths defining the infimum in $\varpi(i,j)$, we obtain $\Omega(i,j)\geq a\,\varpi(i,j)$ whence finally $\Omega(i,j) = a\,\varpi(i,j)$. $\square$

Lemma 10.1. *Let i, j be two points of E. The coefficient $\Omega(i,j)$ vanishes if and only if there exists a path $e_1 \to \cdots \to e_r$ in E joining i and j such that*

$$\forall k \in \{1 \cdots r-1\} \quad \alpha(e_k, e_{k+1}) > 0 \quad \textit{and} \quad f(e_k) \leq f(j).$$

Proof. Let q be an element of $\mathcal{P}(i,j)$ such that $\Omega(q) = 0$ and let $(e_k^*)_{1\leq k\leq |q|}$ be a path of $\mathcal{A}(q)$ such that

$$\sum_{1\leq 2k<|q|} a\left(1-\delta(e_{2k-1}^*, e_{2k}^*)\right) + c\left(\widehat{f}(q_{2k}) - f(e_{2k+1}^*)\right) = 0.$$

Necessarily

$$\forall k \quad 1 \leq 2k < |q| \qquad e_{2k-1}^* = e_{2k}^*, \quad \widehat{f}(q_{2k}) = f(e_{2k+1}^*),$$

whence $\widehat{f}(q_{2k}) \leq \widehat{f}(q_{2k+1})$. Moreover, since $[q_{2k+1}] \subset [q_{2k}]$, we have also $\widehat{f}(q_{2k+1}) \leq \widehat{f}(q_{2k})$. It follows that $\widehat{f}(q_{2k}) = \widehat{f}(q_{2k+1})$, $1 \leq 2k < |q|$ and the individual e_{2k+1}^* belongs to $\hat{q}_{2k+1}$ (also $e_1^* \in \hat{q}_1$). Since for each k, $1 \leq 2k < |q|$, $e_{2k-1}^* = e_{2k}^*$, we have $\widehat{f}(q_{2k-1}) \leq \widehat{f}(q_{2k})$ and the sequence $\left(\widehat{f}(q_k)\right)_{1\leq k\leq |q|}$ is increasing. In particular

$$\forall k \in \{1 \cdots |q|\} \qquad \widehat{f}(q_k) \leq f(j).$$

Now, there exists a path $e_1 \to \cdots \to e_{|q|}$ in E which is admissible for q and such that

$$\forall k \quad 1 \leq 2k < |q| \qquad e_{2k} = e_{2k+1}$$

(this path is the path which leads to the creation of the individual j).
This path necessarily satisfies

$$\forall k \in \{1 \cdots |q|\} \qquad f(e_k) \leq f(j).$$

We suppress the elements e_k of this path such that $e_k = e_{k-1}$ to obtain a path with the desired properties.
The reverse implication is easy: we build a path p which contains an individual following the path $e_1 \to \cdots \to e_r$ given by the hypothesis of the lemma and we just let the remaining individuals evolve "naturally", i.e. there are no mutation and no selection of individuals below peak fitness apart from those of the path $e_1 \to \cdots \to e_r$ (this is explicitly done in the proof of lemma 11.1 below). $\square$

11. The coefficient $\Theta(i,j)$

Lemma 11.1. *Suppose $\Omega(i,j)=0$. Then*

$$\theta(i,j)=\inf\Big\{\,a(r-1)\,+\,c\,\sum_{k=1}^{r}\big(\max_{1\le l\le k} f(e_l)-f(e_k)\big)\Big\}$$

where the infimum is taken over the paths $e_1\to\cdots\to e_r$ in E (of variable length r) verifying $e_1=i$, $e_r=j$ and for each k in $\{1\cdots r-1\}$,

$$e_k\neq e_{k+1},\quad \alpha(e_k,e_{k+1})>0,\quad f(e_k)\le f(j).$$

Remark. The infimum does not change if we suppress the condition $e_k\neq e_{k+1}$: the infimum is attained with a path satisfying this additional condition.

Proof. Let $e_1\to\cdots\to e_r$ be a path in E joining i and j satisfying the above conditions. Let $(e_k^*)_{1\le k<r}$ be a sequence such that

$$\forall k\in\{1\cdots r-1\}\qquad e_k^*\in\{e_1\cdots e_k\}\quad\text{and}\quad f(e_k^*)=\max_{1\le l\le k} f(e_l)$$

and let p be the element of $\overline{D}^2(i,j)$ defined by

$$p^{2k-1}=\{e_k^*,e_k\},\quad p^{2k}=\{e_k^*,e_{k+1}\}\qquad\text{for } k \text{ in } \{1\cdots r-1\}$$

and $p^{2r-1}=\{e_r,e_r\}$. We have $\Omega([p])=0$ so that

$$\theta(i,j)\ \le\ \overline{V}(p)\ =\ a(r-1)\,+\,c\sum_{k=1}^{r}\big(\max_{1\le l\le k} f(e_l)-f(e_k)\big). \tag{6}$$

Conversely, let q belong to $\mathcal{P}^*(i,j)$: thus $\Omega(q)=0$ and the sequence $\big(\widehat{f}(q_k)\big),1\le k\le|q|$, is increasing (see the proof of lemma 10.1). Let p be an element of $\overline{D}(i,j)$ such that $[p]=q$. Necessarily, there exists a path $(e_k),1\le k\le|p|$ in $\mathcal{A}([p])$ such that $e_{2k}=e_{2k+1}$ and $(e_{2k-1},e_{2k})\in\{\,(p_h^{2k-1},p_h^{2k}):1\le h\le m(p)\,\}$ for each k, $1\le 2k<|p|$. Yet

$$\overline{V}(p)\ \ge\ \sum_{1\le 2k<|p|} a\big(1-\delta(e_{2k-1},e_{2k})\big)+c\big(\widehat{f}(p_{2k})-f(e_{2k+1})\big).$$

Since the sequence $\big(\widehat{f}(q_k)\big)_{1\le k\le|q|}$ is increasing, we have $\widehat{f}(p_{2k})\ge\max_{1\le l\le 2k+1} f(e_l)$ so that

$$\overline{V}(p)\ \ge\ \sum_{1\le 2k<|p|} a\big(1-\delta(e_{2k-1},e_{2k})\big)+c\big(\max_{1\le l\le 2k+1} f(e_l)-f(e_{2k+1})\big).$$

Let $\tilde{e}_1\to\cdots\to\tilde{e}_r$ be the path obtained by deleting from the path $e_1\to\cdots\to e_r$ all the individuals e_k such that $e_k=e_{k-1}$. We have

$$\overline{V}(p)\ \ge\ a(r-1)\,+\,c\,\sum_{k=1}^{r}\big(\max_{1\le l\le k} f(\tilde{e}_l)-f(\tilde{e}_k)\big)$$

which together with inequality (6) yields the desired result. $\square$

Corollary 11.1. *Suppose the kernel α is symmetric i.e.*

$$\forall i,j \in E \qquad \alpha(i,j)=\alpha(j,i)$$

and let i,j be two points of E such that $f(i)=f(j)$ and $\Omega(i,j)=0$.
Then $\Omega(j,i)=0$ and $\theta(i,j)=\theta(j,i)$.

Proof. In this situation, the symmetry of α together with lemma 10.1 yield $\Omega(j,i)=0$. In addition, the quantity

$$\theta(i,j) \;=\; \inf\Big\{\, a(r-1)\,+\,c\sum_{k=1}^{r}\big(f(i)-f(e_k)\big) : e_1=i,\; e_r=j,$$
$$\forall k\in\{1\cdots r-1\}\quad e_k\neq e_{k+1},\;\alpha(e_k,e_{k+1})>0,\;f(e_k)\leq f(j)\Big\}$$

becomes symmetric with respect to i and j. Thus $\theta(i,j)=\theta(j,i)$. $\square$

12. The virtual energy and the communication altitude as functions of m

We know that for m sufficiently large, the cost function $V(i,j)$ is affine. All the crucial quantities concerning our algorithm are defined through V as the minimum or the maximum of a set of finite sums involving V. The following elementary result shows immediately that this procedure yields also functions which become affine when m is large.

Lemma 12.1. *Let $(y_i)_{i\in I}$ be a finite family of affine functions. Put*

$$y(t)\;=\;\max\big\{\,y_i(t): i\in I\,\big\}\quad \text{for } t \text{ in } \mathbb{R}.$$

Then for t sufficiently large, y is affine and coincides with one of the functions y_i, i.e.

$$\exists T\in\mathbb{R}\quad \exists i\in I\quad \forall t\geq T\qquad y(t)=y_i(t).$$

Corollary 12.1. *The virtual energy restricted to U is an affine function of m for m sufficiently large:*

$$\forall i\in E\quad \exists M(i)\quad \forall m\geq M(i)\qquad W(i)=\theta(i)+m\Omega(i)$$

where

$$\Omega(i)=\inf\Big\{\sum_{(j_1\to j_2)\in g}\Omega(j_1,j_2): g\in G_U(i)\Big\}$$

and

$$\theta(i)=\inf\Big\{\sum_{(j_1\to j_2)\in g}\theta(j_1,j_2): g\in G_U(i)\,,\;\sum_{(j_1\to j_2)\in g}\Omega(j_1,j_2)=\Omega(i)\Big\}.$$

The next lemma will be used in the proofs of proposition 12.1 and corollary 12.2.

Lemma 12.2. $\forall i \in f^* \quad \forall j \in E \qquad \Omega(i) = 0 \quad$ *and* $\quad \Omega(j) = \Omega(i,j)$.

Proof. Let i belong to f^*. For each j_1 in $E\backslash f^*$, we choose a point j_2 such that $f(j_1) < f(j_2)$ and there exists a path $e_1 \to \cdots \to e_r$ in E joining j_1 and j_2 satisfying

$$\forall k \in \{1 \cdots r-1\} \qquad \alpha(e_k, e_{k+1}) > 0, \quad f(e_k) \leq f(j_2).$$

Let g be the i–graph built with all these arrows $j_1 \to j_2$ and the arrows $j \to i$, $j \in f^* \setminus \{i\}$. Lemma 10.1 shows that

$$\sum_{(j_1 \to j_2) \in g} \Omega(j_1, j_2) = 0$$

whence $\Omega(i) = 0$ and the first part of the lemma is proved.
Let j belong to E. Since $\Omega(i) + \Omega(i,j) \geq \Omega(j)$, we have $\Omega(j) \leq \Omega(i,j)$.
To prove the reverse inequality, first notice that Ω satisfies the triangular inequality

$$\forall j_1, j_2, e \in E \qquad \Omega(j_1, j_2) \leq \Omega(j_1, e) + \Omega(e, j_2)$$

(to prove this inequality, we just put end to end paths joining (j_1) to (e) and paths joining (e) to (j_2) to obtain paths joining (j_1) to (j_2) and then take the infimum defining Ω). Yet each j–graph contains a path $e_1 \to \cdots \to e_r$ joining i to j whence

$$\Omega(j) \geq \Omega(e_1, e_2) + \cdots + \Omega(e_{r-1}, e_r) \geq \Omega(i,j)$$

and finally $\Omega(j) = \Omega(i,j)$. $\square$

Proposition 12.1. *Suppose $a \geq c\big(f(f^*) - f(i)\big)$, where i is a point of E. Then*

$$\Omega(i) = c\big(f(f^*) - f(i)\big).$$

In particular, if $a \geq c\Delta$, then

$$\forall i \in E \qquad \Omega(i) = c\big(f(f^*) - f(i)\big).$$

Remark. When the cost of the mutations is much higher than the cost of the anti–selections, the coefficient Ω is obtained by applying an affine transformation on the fitness function f.

Proof. Let i belong to E and e to f^*. We have $\omega(e,i) = c\big(f(f^*) - f(i)\big)$ and the inequality $a \geq c\big(f(f^*) - f(i)\big)$ implies by proposition 10.1 that $\Omega(e,i) = c\big(f(f^*) - f(i)\big)$. Finally, it follows from lemma 12.2 that $\Omega(i) = c\left(f(f^*) - f(i)\right)$. $\square$

Corollary 12.2. *The communication altitude restricted to U is an affine function of m:*

$$\forall i,j\in E\quad \exists \rho(i,j)\quad \exists M\quad \forall m\geq M\qquad A(i,j)=\rho(i,j)+m\max\big(\Omega(i),\Omega(j)\big).$$

Proof. We already know that $A(i,j)$ is affine for m sufficiently large. It remains only to prove that

$$\lim_{m\to\infty}\frac{A(i,j)}{m}=\max\big(\Omega(i),\Omega(j)\big).$$

Since $A(i,j)\geq\max\big(W(i),W(j)\big)$, we have

$$\lim_{m\to\infty}\frac{A(i,j)}{m}\geq\max\Big(\lim_{m\to\infty}\frac{W(i)}{m},\lim_{m\to\infty}\frac{W(j)}{m}\Big)=\max\big(\Omega(i),\Omega(j)\big).$$

Now pick a point i^* in f^* and consider the path $i\to i^*\to j$. Lemma 12.2 yields

$$\lim_{m\to\infty}\frac{A(i,j)}{m}\leq\max\big(\Omega(i),\Omega(j)\big)$$

and the result of the corollary is proved. □

The relative order of a finite family of affine functions does not change any more when the variable is sufficiently large.

Corollary 12.3. *There exists an integer M such that for all i_1, j_1, i_2, j_2 in E, we have*

$$\begin{array}{lll} \textit{either} & \forall m\geq M & A(i_1,j_1)=A(i_2,j_2)\\ \textit{or} & \forall m\geq M & A(i_1,j_1)<A(i_2,j_2)\\ \textit{or} & \forall m\geq M & A(i_1,j_1)>A(i_2,j_2). \end{array}$$

Let M^* be the smallest integer such that the limit behavior described by the preceding results is achieved for $m\geq M^*$; that is V,W,A are affine functions of m and the relative positions of these affine functions does not change any more on $[M^*,\infty[$.

13. THE LIMITING DECOMPOSITION IN CYCLES

The preceding results yield the

Theorem 13.1. *(stabilization of the cycles of U)*
When m is greater than M^*, the set $\mathcal{C}(U)$ of cycles in U does not depend any more on m.

Proof. Let $\mathcal{A} = \{\, A(i,j) : i \in E,\, j \in E \,\}$. The cardinality of $\mathcal{A}$ is less than $|E|^2$.
Put $\mathcal{A} = \{\, a_1, \cdots a_s \,\}$ where s is a function of m and $0 \leq a_1 < a_2 < \cdots < a_{s-1} < a_s$.
Let $N : E \times E \longmapsto \mathbb{N}$ be the function defined by

$$\forall i,j \in E \qquad a_{N(i,j)} = A(i,j).$$

Consider the equivalence relation $\mathcal{T}_k$ defined on the set $\{\, i \in E : N(i,i) \leq k \,\}$ by $i\ \mathcal{T}_k\ j \iff A(i,j) \leq a_k \iff N(i,j) \leq k$. We have

$$\begin{aligned}\mathcal{C}(U) &= \bigcup_{\lambda \in \mathcal{A}} \{\, i : W(i) \leq \lambda \,\} \,/\, \mathcal{R}_\lambda \\ &= \bigcup_{k=1}^{s} \{\, i : W(i) \leq a_k \,\} \,/\, \mathcal{R}_{a_k} \\ &= \bigcup_{k=1}^{s} \{\, i : N(i,i) \leq k \,\} \,/\, \mathcal{T}_k.\end{aligned}$$

so that in fact $\mathcal{C}(U)$ depends only on the functions $N(i,j)$, the integer s and the equivalence relations $(\mathcal{T}_k)$: corollary 12.3 shows that these objects do not vary when $m \geq M^*$. $\square$

We focus on this limiting decomposition. From now onwards, the population size m is assumed to be greater than M^*.

14. PATHS BETWEEN POINTS OF f^* IN MINIMAL GRAPHS

We first state a general lemma about graphs.

Lemma 14.1. *Let i belong to E and let g be a graph on E such that for each j in $E \setminus \{i\}$ there exists a path $j = e_1 \to e_2 \to \cdots \to e_r = i$ in g leading from j to i. Then there exists an i–graph $\tilde{g}$ which is contained in g i.e.*

$$\forall j_1, j_2 \in E \qquad (j_1 \to j_2) \in \tilde{g} \implies (j_1 \to j_2) \in g.$$

Proof. We build the graph $\tilde{g}$ by removing arrows from g. First, we suppress all arrows starting at i. We then consider successively each point of $E \setminus \{i\}$. Let j be such a point. By hypothesis, there exists at least a path $j = e_1 \to e_2 \to \cdots \to e_r = i$ in g leading from j to i. We remove all arrows starting at j distinct from $(e_1 \to e_2)$. We continue this procedure until each point of $E \setminus \{i\}$ is the starting point of exactly one arrow. $\square$

We now prove two lemmas which describe very accurately the paths joining the points of f^* in minimal graphs.

Lemma 14.2. *Let i belong to f^* and g to $G^*_U(i)$. Suppose the arrow $(j_1 \to j_2)$ belong to g, where j_1 and j_2 are elements of f^*. Then any path $e_1 = j_1 \to \cdots \to e_r = j_2$ in E realizing the value*

$$\theta(j_1, j_2) \;=\; \inf\Big\{\, a(r-1) \,+\, c\sum_{k=1}^{r} \big(f(f^*) - f(e_k)\big) : e_1 = j_1,\, e_r = j_2,$$
$$\forall k \in \{1 \cdots r-1\} \quad e_k \neq e_{k+1},\ \alpha(e_k, e_{k+1}) > 0 \Big\}$$

is such that

$$\forall k \in \{2 \cdots r-1\} \qquad e_k \notin f^*.$$

Proof. Let i and g be as in the hypothesis and suppose the result is false: there exists an arrow $(j_1 \to j_2)$ in g, where j_1, j_2 belong to f^*, and a path $e_1 \to \cdots \to e_r$ in E joining j_1 and j_2 such that

$$\forall k \in \{1 \cdots r-1\} \quad e_k \neq e_{k+1},\ \alpha(e_k, e_{k+1}) > 0, \qquad \exists h \in \{2 \cdots r-1\} \qquad e_h \in f^*$$

$$\text{and} \qquad \theta(j_1, j_2) \;=\; a(r-1) \,+\, c\sum_{k=1}^{r} \big(f(f^*) - f(e_k)\big).$$

We have then

$$\theta(j_1, e_h) \;=\; a(h-1) \,+\, c\sum_{k=1}^{h} \big(f(f^*) - f(e_k)\big),$$

$$\theta(e_h, j_2) \;=\; a(r-h-1) \,+\, c\sum_{k=h+1}^{r} \big(f(f^*) - f(e_k)\big).$$

We build from g a graph $\tilde{g}$ in the following way:

- if $e_h = i$, we replace the arrow $(j_1 \to j_2)$ by $(j_1 \to i)$;
- if $e_h \neq i$, we replace the arrow $(j_1 \to j_2)$ by the arrows $j_1 \to e_h \to j_2$ and we remove the arrow starting at e_h.

In both cases, we obtain a graph $\tilde{g}$ such that $V(\tilde{g}) < V(g)$ and for each point j in $E \setminus \{i\}$, there exists a path $e_1 = j \to \cdots \to e_r = i$ in $\tilde{g}$ leading from j to i. Lemma 14.1 shows that $\tilde{g}$ contains an i–graph $\tilde{\tilde{g}}$, and necessarily $V(\tilde{\tilde{g}}) < V(g)$, which contradicts the fact that g belongs to $G^*_U(i)$. $\square$

Lemma 14.3. *Let i belong to f^* and g to $G_U^*(i)$. Suppose the arrow $(j_1 \to j_2)$ belong to g, where j_1 and j_2 are elements of f^*.*
A path p in E^m belongs to $D^{m*}(j_1, j_2)$ if and only if:
there exists a path $e_1 = j_1 \to \cdots \to e_r = j_2$ in E realizing the value $\theta(j_1, j_2)$ verifying

$$\forall k \in \{1 \cdots r-1\} \qquad e_k \neq e_{k+1},\ \alpha(e_k, e_{k+1}) > 0$$

and two integers $0 \leq t_1 < t_2 < |p|$ such that

$$\begin{array}{ll} p^s = (j_1) & 1 \leq s \leq t_1 + 1 \\ p^s(j_1) = m-1,\ p^s(e_{s-t_1}) = 1 & t_1 + 2 \leq s \leq t_1 + r - 1 \\ [p^s] = \{j_1, j_2\} & t_1 + r \leq s \leq t_2 \\ p^s = (j_2) & t_2 + 1 \leq s \leq |p| \end{array}$$

Proof. Any path p satisfying the above conditions belongs to $D^m(j_1, j_2)$, has a cost $\theta(j_1, j_2)$ and thus belongs to $D^{m*}(j_1, j_2)$.
Conversely, let p be an element of $D^{m*}(j_1, j_2)$. Put

$$t_1 = \min\{\, s : p^s \neq (j_1)\,\} - 2, \qquad t_2 = \max\{\, s : p^s \neq (j_2)\,\} + 1.$$

Necessarily, there exists a path $e_1 = j_1 \to \cdots \to e_r = j_2$ in E satisfying

$$\forall k \in \{1 \cdots r-1\} \qquad e_k = e_{k+1} \quad \text{or} \quad \alpha(e_k, e_{k+1}) > 0,$$

$$\forall k \in \{2 \cdots r-1\} \qquad e_k \in [p_{t_1+k}].$$

To obtain this path, we just look at the sequence of mutations which leads to the creation of j_2 in the path p. Thus the cost of the path p is greater than

$$a(r-1) + c\sum_{k=1}^{r} \big(f(f^*) - f(e_k)\big)$$

(necessarily, $\Omega([p]) = \Omega(j_1, j_2) = 0$ so that $\forall k \in \{2 \cdots r-1\}$ $\widehat{f}(p_{t_1+k}) = \widehat{f}(f^*)$, see for instance the proof of lemma 10.1). However, the cost of p is precisely $\theta(j_1, j_2)$: thus the path $e_1 \to \cdots \to e_r$ realizes the value $\theta(j_1, j_2)$ (whence in particular $e_k \neq e_{k+1}$ for each k and by the preceding lemma 14.2 $e_k \notin f^*$ for k in $\{2 \cdots r-1\}$) and the only events of positive cost in the path p are those concerning this individual path: that is there are no mutations nor selection of bad individuals apart from the individual which follows the path $e_1 \to \cdots \to e_r$. Thus

$$\begin{array}{lll} \forall s & t_1 + 2 \leq s \leq t_1 + r - 1 & p^s(j_1) = m-1,\ p^s(e_{s-t_1}) = 1 \\ \forall s & t_1 + r \leq s \leq t_2 & [p^s] = \{j_1, j_2\} \end{array}$$

and the path p satisfies the required conditions. $\square$

15. The limit distribution

We always suppose that m is greater than M^*. We are now in position to prove the

Theorem 15.1. *(limiting distribution)*
We have $W^* \subset f^*$. Suppose that the kernel α is symmetric. Then $W^* = f^*$ and the limit distribution ν is the uniform distribution on f^*:

$$\forall x \in E^m \quad \forall i \in f^* \qquad \nu(i) = \lim_{l \to \infty} \quad \lim_{n \to \infty} \quad P\left(X_n^l = (i)/X_0^l = x\right) = \frac{1}{|f^*|}.$$

Proof. The inclusion $W^* \subset f^*$ was proved in [3]. Suppose that the kernel α is symmetric. Let i and j be two points of f^*. Let g be a graph belonging to $G_U^*(j) : V(g) = W(j)$. There exists a unique path $e_1 = i \to e_2 \to \cdots \to e_r = j$ in g leading from i to j. Since $\Omega(j) = 0$, necessarily

$$\forall k \in \{1 \cdots r-1\} \qquad \Omega(e_k, e_{k+1}) = 0 \quad \Longrightarrow \quad f(e_k) \leq f(e_{k+1}).$$

Since i is in f^*, we have in fact

$$\forall k \in \{1 \cdots r\} \qquad e_k \in f^*.$$

Let $\Phi_{ij}(g)$ be the i–graph obtained from g by reversing the arrows of this path in the following way: $e_r = j \to e_{r-1} \to \cdots \to e_1 = i$. By corollary 11.1, V is symmetric on f^*, so that $V(g) = V(\Phi_{ij}(g))$. It follows that $W(i) \leq W(j)$ and by symmetry we conclude that $W(i) = W(j)$, i.e. the virtual energy is constant on f^* and $W^* = f^*$. Thus Φ_{ij} is a one-to-one map between $G_U^*(j)$ and $G_U^*(i)$. Moreover, $\Phi_{ij} \circ \Phi_{ji} = \mathrm{Id}_{G_U^*(i)}$. In addition, the operator Φ_{ij} reverses only arrows between points of f^*. Put

$$q(i,j) = \sum_{p \in D^{m*}(i,j)} \prod_{k=1}^{|p|-1} q_1(p_k, p_{k+1}).$$

The limit distribution ν may be expressed through Freidlin–Wentzell graphs [4, chapter 6]. For i in W^*, we have

$$\nu(i) = \frac{\sum_{g \in G_U^*(i)} \prod_{(j_1 \to j_2) \in g} q(j_1, j_2)}{\sum_{j \in f^*} \sum_{g \in G_U^*(j)} \prod_{(j_1 \to j_2) \in g} q(j_1, j_2)}.$$

The end of the proof rests on the

Lemma 15.1. *Let i belong to f^* and g to $G_U^*(i)$. Suppose the arrow $(j_1 \to j_2)$ belong to g, where j_1 and j_2 are elements of f^*.*
Then
$$q(j_1, j_2) = q(j_2, j_1).$$

Assume the lemma is true. It follows that

$$\forall i, j \in f^* \quad \forall g \in G_U^*(j) \qquad \prod_{(j_1 \to j_2) \in g} q(j_1, j_2) = \prod_{(j_1 \to j_2) \in \Phi_{ij}(g)} q(j_1, j_2)$$

whence

$$\sum_{g \in G_U^*(j)} \prod_{(j_1 \to j_2) \in g} q(j_1, j_2) = \sum_{g \in \Phi_{ij}\left(G_U^*(j)\right)} \prod_{(j_1 \to j_2) \in g} q(j_1, j_2) = \sum_{g \in G_U^*(i)} \prod_{(j_1 \to j_2) \in g} q(j_1, j_2)$$

and the above quantity is the same for all points of f^*: the limit distribution is thus uniform on f^*. □

Proof of lemma 15.1. Let i, j_1, j_2 be as in the hypothesis of the lemma. Let p be an element of $D^{m*}(j_1, j_2)$. By lemma 14.3, there exists a path $e_1 = j_1 \to \cdots \to e_r = j_2$ realizing the value $\theta(j_1, j_2)$ and satisfying

$$\forall k \in \{1 \cdots r-1\} \qquad e_k \neq e_{k+1}, \quad \alpha(e_k, e_{k+1}) > 0$$

and two integers $0 \le t_1 < t_2 < |p|$ such that

$$\begin{array}{ll} p^s = (j_1) & 1 \le s \le t_1 + 1 \\ p^s(j_1) = m-1,\ p^s(e_{s-t_1}) = 1 & t_1 + 2 \le s \le t_1 + r - 1 \\ [p^s] = \{j_1, j_2\} & t_1 + r \le s \le t_2 \\ p^s = (j_2) & t_2 + 1 \le s \le |p| \end{array}$$

With the path p we associate the path $\Phi(p) = \tilde{p}$ defined by
- $\forall s \quad 1 \le s \le |p| - t_2 \qquad \tilde{p}^s = (j_2)$

For each s such that $t_1 + 2 \le s \le t_1 + r - 1$, there exists a unique index $\sigma(s)$ such that $p^s_{\sigma(s)} = e_{s-t_1} \neq j_1$. We define
- $\forall s \quad |p| - t_2 + 1 \le s \le |p| - t_2 + r - 2$

$$\tilde{p}^s_h = \begin{cases} e_{r-|p|+t_2-s} & \text{if} \quad h = \sigma(s) \\ j_2 & \text{if} \quad h \neq \sigma(s) \end{cases}$$

Let $\tau_{j_1 j_2}$ be the transposition of E which exchanges the points j_1 and j_2.
For $x = (x_1, \cdots, x_m)$ in E^m, we put $\tau_{j_1 j_2} \cdot x = (\tau_{j_1 j_2}(x_1), \cdots, \tau_{j_1 j_2}(x_m))$.

We define
- $\forall s \quad |p|-t_2+r-1 \le s \le |p|-t_1-1 \qquad \tilde{p}^s = \tau_{j_1 j_2} \cdot p_{s-|p|+t_1+t_2+1}$

and finally
- $\forall s \quad |p|-t_1 \le s \le |p| \qquad \tilde{p}_s = (j_1)$.

The path $\tilde{p}$ is built by reversing the path $e_1 = j_1 \to \cdots \to e_r = j_2$ and by reproducing (with j_1 and j_2 exchanged) the portion of p which contains only the individuals j_1 and j_2. Since the kernel α is symmetric, the path $\Phi(p)$ belongs to $D^m(j_2, j_1)$ and its cost is $\theta(j_1, j_2)$; corollary 11.1 shows that θ is symmetric on f^*, so that $\Phi(p)$ is actually an element of $D^{m*}(j_2, j_1)$. Finally, the very definition of Φ yields the following facts:
- $\forall p \in D^{m*}(j_1, j_2) \qquad \Phi(\Phi(p)) = p$.
- Φ is one-to-one between $D^{m*}(j_1, j_2)$ and $D^{m*}(j_2, j_1)$.
- $\forall p \in D^{m*}(j_1, j_2)$

$$\prod_{k=1}^{|p|-1} q_1(p^k, p^{k+1}) = \prod_{k=1}^{|p|-1} q_1(\Phi(p)^k, \Phi(p)^{k+1})$$

(where $\Phi(p)^k$ denotes the k–th population of the path $\Phi(p)$).
As a consequence, we have

$$\begin{aligned} q(j_1, j_2) = & \sum_{p \in D^{m*}(j_1, j_2)} \prod_{k=1}^{|p|-1} q_1(p^k, p^{k+1}) = \sum_{p \in D^{m*}(j_1, j_2)} \prod_{k=1}^{|p|-1} q_1(\Phi(p)^k, \Phi(p)^{k+1}) \\ = & \sum_{p \in \Phi(D^{m*}(j_1, j_2))} \prod_{k=1}^{|p|-1} q_1(p^k, p^{k+1}) = \sum_{p \in D^{m*}(j_2, j_1)} \prod_{k=1}^{|p|-1} q_1(p^k, p^{k+1}) = q(j_2, j_1). \end{aligned}$$

and the lemma is proved. □

16. Structure of the limiting hierarchy of cycles

Theorem 16.1. *(structure of the cycles)*
Let π be a cycle of $\mathcal{C}(U)$. Suppose π is not included in W^*.
Then either Ω is constant on π or the set

$$\{\, i : \Omega(i) < \max_{j \in \pi} \Omega(j) \,\}$$

(which contains f^*) is included in the cycle π.

Proof. Let π be a cycle of $\mathcal{C}(U)$ not included in W^*. There exists a real number $\lambda(m)$ such that π is an equivalence class of $W_{\lambda(m)} \cap U$ for the relation $\mathcal{R}_{\lambda(m)}$ (for all $m \ge M^*$). Suppose Ω is not constant on π, so that there exist j and e in π such that

$$\Omega(j) = \max_{i \in \pi} \Omega(i), \quad \Omega(e) < \Omega(j).$$

Since $A(e,j) \le \lambda(m)$, necessarily

$$m \max\big(\Omega(e), \Omega(j)\big) = m\,\Omega(j) \le \lambda(m).$$

Let i be a point of E such that $\Omega(i) < \Omega(j)$. We have, as $m \to \infty$

$$\frac{A(i,e)}{m} \longrightarrow \max\big(\Omega(i), \Omega(e)\big) < \Omega(j)$$

so that for m sufficiently large $A(i,e) < m\,\Omega(j) \le \lambda(m)$ and i belongs to the cycle π. $\square$

Corollary 16.1. *Let π be a cycle of $\mathcal{C}(U)$ included in the set*

$$\{\, i : f(f^*) - f(i) \le \frac{a}{c} \,\}.$$

Either f is constant on π or π contains the set

$$\{\, i : \min_{j\in\pi} f(j) < f(i) \,\}$$

(whence f^* is included in π).
In particular, if $a \ge c\Delta$, then the function f is constant on all the cycles not containing f^*. The other cycles verify

$$\{\, i : \min_{j\in\pi} f(j) < f(i) \,\} \;\subset\; \pi.$$

Proof. This corollary is an easy consequence of the previous theorem and proposition 12.1.

Corollary 16.2. *(cycles of E^m)*
Suppose $a \ge c\Delta$. Let π belong to $\mathcal{C}(E^m)$. Either $\widehat{f}$ is constant on π or the set

$$\{\, (i) \in U : f(i) > \min_{j\in\pi} \widehat{f}(j) \,\}$$

is included in π (whence $f^* \subset \pi$).

Proof. It is enough to remark that

$$\forall \pi \in \mathcal{C}(E^m) \quad \forall x \in \pi \quad \forall i \in \hat{x} \qquad V(x,(i)) = 0 \quad \text{and} \quad (i) \in \pi$$

and to translate the preceding corollary on E^m with the projection T_U. $\square$

The remarkable fact is that for $a \ge c\Delta$, the structure of the cycles is essentially determined by the level sets of the function f (the kernel α just plays a role within the level sets).
We have the opposite phenomenon when $Ra \le c\delta$ (where R is the minimal number of transitions necessary to join two arbitrary points of E through the kernel α). The structure of the cycles is then essentially determined by the kernel α and the energy landscape (E, f, α). However, since there is no easy way of describing this dynamics ($\Omega(i,j)$ is then equal to $\varpi(i,j)$), we do not state the results dual to corollaries 16.1 and 16.2.
Finally, we prove a general fact concerning the "bad" cycles.

Theorem 16.2. *(cycles disjoint from W^*)*
The function f is constant over the cycles of $\mathcal{C}(U)$ not intersecting W^*.

Remark. The cycles not intersecting W^* are the "bad" cycles which slow down the convergence of the process toward W^* (the critical height H_1 is the maximal height of exit of these cycles [9,10,11]). The above result shows that these cycles are in a way "transverse" to the "good" cycles which intersect W^*.

Proof. Let π belong to $\mathcal{C}(U)$ be such that $\pi \cap W^* = \emptyset$. Let i, j be two points of π and let $e_1 \to \cdots \to e_r$ be a path in E joining i and j such that

$$A(i,j) = \max_{1 \le k < r} W(e_k) + V(e_k, e_{k+1}). \tag{7}$$

Once more, for $m \ge M^*$, the quantities involved in the above formula are affine functions of m so that the path $e_1 \to \cdots \to e_r$ actually realizes the value $A(i,j)$ for all values of m in $[M^*, \infty[$ (the minimizing path should a priori depends on m). Necessarily, all the points of this path are in the cycle π, so that by theorem 16.1,

$$\Omega(i) = \Omega(j) = \Omega(e_1) = \cdots = \Omega(e_r).$$

Identity (7) implies also

$$\max\big(\Omega(i), \Omega(j)\big) = \max_{1 \le k < r} \Omega(e_k) + \Omega(e_k, e_{k+1})$$

whence

$$\forall k \in \{1 \cdots r-1\} \qquad \Omega(e_k, e_{k+1}) = 0$$

and by lemma 10.1

$$\forall k \in \{1 \cdots r-1\} \qquad f(e_k) \le f(e_{k+1}).$$

Finally, we have $f(i) \le f(j)$ and by symmetry it follows that $f(i) = f(j)$. Thus f is constant over the cycle π. $\square$

Translating this result on E^m with the projection T_U, we obtain

Corollary 16.3. *The function $\widehat{f}$ is constant over the cycles of $\mathcal{C}(E^m)$ not intersecting W^*.*

Remark. When the process starts from a uniform population (i), it explores very intensively the neighbourhood of i until a point j of greater fitness is found. The cost to find such a point j corresponds to the mutations and anti–selections necessary to lead an individual from i to j while the remainder of the population waits in i. Then the process moves to the uniform population (j) and the mechanism starts again from j, until a global maxima of f is finally reached. With high probability, the process will thus visit f^* before coming back to (i). Therefore, as soon as a cycle contains two populations with distinct maximal fitness, it contains also points of f^*.

17. THE CRITICAL HEIGHT H_1

For the definitions and the properties of the quantities $H_e(\pi)$ and $H_m(\pi)$ (the height of exit and height of mixing of a cycle π), we refer the reader to Trouvé's work [9,10,11]. We use the following notations: if π is a cycle over the set U, its height of exit and of mixing are denoted by $H_e^U(\pi)$ and $H_m^U(\pi)$ (notice that they are evaluated with the cost function V_U); if π is a cycle of E^m, we note this quantities $H_e(\pi)$ and $H_m(\pi)$. The following lemma will be used in the proof of proposition 17.1.

Lemma 17.1. *Let π be a cycle and x an element of π. There exists a point i_x in E such that*

$$(i_x) \in \pi, \qquad W(x) = W((i_x)) + V((i_x), x).$$

Proof. Let g be an x–graph over E^m such that $V(g) = W(x)$ and each arrow $(j_1 \to j_2)$ of g starting from a point j_1 of $E^m \setminus U$ is such that: $j_2 \in U$, $V(j_1, j_2) = 0$ (for the existence of such a graph, see [4, chapter 6, Lemma 4.3]). Pick up a point i in $\hat{x}$. There exists a path $p_1 = (i) \to \cdots \to p_{r-1} \to p_r = x$ leading from (i) to x in the graph g. Necessarily, p_{r-1} belongs to U. We put $p_{r-1} = (i_x)$, where i_x is a point of E. We have

$$\forall k \in \{1 \cdots r-1\} \qquad W(p_k) + V(p_k, p_{k+1}) \leq W(x)$$

(to see this, just add the arrow $x \to (i)$ to g and remove the arrow $p_k \to p_{k+1}$)
whence $A((i), (i_x)) \leq W(x) = A(x, (i))$ and (i_x) belongs to π.
Moreover, the choice of g and i_x imply $W(x) = W((i_x)) + V((i_x), x)$. $\square$

The following proposition is not specific to our particular situation:

Proposition 17.1. *Let π be a cycle of E^m. We have the inequalities*

$$H_m^U(T_U(\pi)) \;\leq\; H_m(\pi) \;\leq\; H_e(\pi) \;\leq\; H_e^U(T_U(\pi)).$$

Proof. By definition ([9, Formula 23])

$$H_m(\pi) = \max_{x,y\in\pi} A(x,y) - \min_{x\in\pi} W(x)$$

and

$$H_m^U(T_U(\pi)) = \max_{(i),(j)\in\pi} A((i),(j)) - \min_{(i)\in\pi} W((i)).$$

For any cycle π of E^m, we have $\min_{x\in\pi} W(x) = \min_{(i)\in\pi} W(i)$ and

$$\max_{x,y\in\pi} A(x,y) \geq \max_{(i),(j)\in\pi} A((i),(j))$$

which yield the first inequality.

By definition ([10, Proposition 1.33])

$$H_e(\pi) = \max_{x\in\pi} \min_{y\notin\pi} \big(A(x,y) - W(x)\big)$$

and

$$H_e^U(T_U(\pi)) = \max_{(i)\in\pi} \min_{(j)\notin\pi} \big(A((i),(j)) - W((i))\big).$$

Let x belong to π and let (j) be an element of $U \setminus \pi$. By lemma 17.1, there exists i_x in E such that (i_x) is in π and $W(x) = W((i_x)) + V((i_x), x)$. We have

$$A\big(x,(j)\big) \leq \max\Big(A\big(x,(i_x)\big), A\big((i_x),(j)\big)\Big)$$

(to obtain this inequality, we just put end to end a path joining x to (i_x) and a path joining (i_x) to (j)). However $A\big(x,(i_x)\big) = W(x)$ and $A\big((i_x),(j)\big) \geq W((i_x))$. Thus

$$A\big(x,(j)\big) \leq A\big((i_x),(j)\big) + V((i_x), x)$$

and

$$A\big(x,(j)\big) - W(x) \leq A\big((i_x),(j)\big) - W((i_x)).$$

It follows that

$$\begin{aligned}
\min_{y\notin\pi} A(x,y) - W(x) &\leq \min_{(j)\notin\pi} A\big(x,(j)\big) - W(x)\\
&\leq \min_{(j)\notin\pi} A\big((i_x),(j)\big) - W((i_x))\\
&\leq \max_{(i)\in\pi} \min_{(j)\notin\pi} \big(A\big((i),(j)\big) - W((i))\big).
\end{aligned}$$

Taking the maximum of the left–hand side over x in π, we obtain the last inequality. □

Lemma 17.2. *Let π be a cycle of $\mathcal{C}(U)$. We have*

$$H_e^U(\pi) \;\leq\; \max_{i\in\pi} \min_{j\notin\pi} V(i,j)$$

and $H_e^U(\pi)$ is constant for m sufficiently large whenever $\pi \cap f^* = \emptyset$.

Proof. To obtain this inequality, we just notice that

$$\forall i,j \in E \qquad A(i,j) - W(i) \leq V(i,j).$$

Suppose $\pi \cap f^* = \emptyset$. We have then (see [3, Corollary 11.1])

$$\sup_{m\in\mathbb{N}^*} \quad \max_{i\in\pi} \min_{j\notin\pi} V(i,j) \;<\; \infty.$$

Moreover, the definition of $H_e^U(\pi)$ implies that (for m large enough), it is an affine function of m. Being bounded, it is necessarily constant. □

Corollary 17.1. *For each cycle* π *of* $\mathcal{C}(E^m)$, *we have*

$$H_e(\pi) \;\leq\; \max_{i\in\pi}\;\min_{j\notin\pi} V(i,j)$$

and the critical height H_1 *is bounded as a function of* m*:*

$$H_1 \;\leq\; \max_{i\notin f^*}\;\min_{j\in f^*} V(i,j).$$

Proof. This is an immediate consequence of proposition 17.1, the preceding lemma 17.2, the definition of H_1, $H_1 = \sup\big\{\, H_e(\pi) : \pi \in \mathcal{C}(E^m),\ \pi\cap W^* = \emptyset \,\big\}$ ([9, Definition 3.22]), and the fact that $\sup_{m\in\mathbb{N}^*} \quad \max_{i\notin f^*}\ \min_{j\in f^*}\ V(i,j) \;<\; \infty$ (see [3, Corollary 11.1]). $\square$

We now restate Trouvé's convergence result, which is an extension of a result by Hajek for the simulated annealing.

Theorem 17.1. *(Trouvé* [9, Theorem 3.23]*)*
Suppose α *is symmetric and* m *is large enough to have* $W^* = f^*$.
For all increasing sequences $l(n)$ *going to infinity, we have the equivalence*

$$\sup_{x\in E^m} \quad P\left(X_n \notin f^*/X_0 = x\right) \underset{n\to\infty}{\longrightarrow} 0 \qquad\Longleftrightarrow\qquad \sum_{n=0}^{\infty} l(n)^{-H_1} = \infty.$$

A remarkable fact is that we may adapt the sequence $l(n)$ in order to be trapped in f^*.

Theorem 17.2. *Define* $H_e^* = H_e(\{\, x\in E^m : [x]\cap f^* \neq \emptyset \,\})$.
Suppose m *is large enough to have* $W^* = f^*$ *and* $H_1 < H_e^*$.
For all increasing sequences $l(n)$*, we have the equivalence*

$$\forall x\in E^m \qquad P\left(\exists N \quad \forall n\geq N \quad [X_n]\cap f^* \neq \emptyset / X_0 = x\right) = 1$$

$$\Longleftrightarrow \qquad \sum_{n=0}^{\infty} l(n)^{-H_1} = \infty, \qquad \sum_{n=0}^{\infty} l(n)^{-H_e^*} < \infty.$$

Proof. The two above conditions on the sequence $l(n)$ exactly express that

- the chain (X_n) has a null probability of being trapped in a cycle disjoint from f^* i.e.

$$\forall x\in E^m \quad \forall I\subset E\setminus f^* \qquad P\left(\exists N \quad \forall n\geq N \qquad [X_n]\subset I/X_0 = x\right) = 0,$$

- the chain has a positive probability of being trapped in the set of populations containing an individual of f^* i.e.

$$\forall x\in E^m \qquad P\left(\exists N \quad \forall n\geq N \qquad [X_n]\cap f^* \neq \emptyset/X_0 = x\right) > 0. \quad \square$$

Remark. Since $H_e^* \geq m\,\min(a, c\delta^*)$, where $\delta^* = \min\big\{\, f(f^*) - f(i) : i\notin f^* \,\big\}$, the hypothesis of the theorem is fulfilled when m is large, and any increasing sequence $l(n)$ satisfying

$$\sum_{n=0}^{\infty} l(n)^{-H_1} = \infty, \qquad \sum_{n=0}^{\infty} l(n)^{-m\,\min(a,c\delta^*)} < \infty$$

will achieve the desired behavior.

18. The optimal rate of convergence

For the meaning and the properties of the optimal convergence exponent and the logarithmic convergence exponent we refer the reader to [1,2,9,10,11].

Proposition 18.1. *The optimal convergence exponent α_{opt} is an affine strictly increasing function of m for m large. The rate of increase of α_{opt} is*

$$\lim_{m\to\infty}\frac{\alpha_{\text{opt}}}{m} \;=\; \min\left\{\,\frac{\Omega(\pi)}{H_e(\pi)} : \pi\in\mathcal{C}(E^m),\ \pi\cap f^*=\emptyset\,\right\} \;\geq\; \frac{\min(a,c\delta^*)}{H_1}$$

where $\delta^*=\min\left\{\,f(f^*)-f(i) : i\notin f^*\,\right\}$.

Proof. The definition of α_{opt} is [9, Definition 3.22]

$$\alpha_{\text{opt}}=\min\left\{\,\frac{W(\pi)-W(W^*)}{H_e(\pi)} : \pi\in\mathcal{C}(E^m),\ \pi\cap f^*=\emptyset\,\right\}.$$

For π in $\mathcal{C}(E^m)$ with $\pi\cap f^*=\emptyset$, we have

$$\Omega(\pi)=\min\left\{\,\Omega(i) : i\in\pi\,\right\}\geq\min(a,c\delta^*)>0$$

so that $W(\pi)$ is an affine strictly increasing function of m. The result of the proposition follows easily. $\square$

We restate now Trouvé's result for the optimal convergence rate, which generalizes Catoni's work.

Theorem 18.1. *There exist two strictly positive constants R_1 and R_2 such that for all n*

$$\frac{R_1}{n^{\alpha_{\text{opt}}}} \;\leq\; \inf_{0\leq l(1)\leq\cdots\leq l(n)}\;\max_{x\in E^m}\; P\left(X_n\notin f^*/X_0=x\right) \;\leq\; \frac{R_2}{n^{\alpha_{\text{opt}}}}.$$

Proof. In order to apply Trouvé's result to the mutation–selection algorithm, we need only to check that condition C_1 (see [10,11]) concerning the transition probabilities is fulfilled; but these are precisely obtained as sums of fractions involving the powers of l. $\square$

Remark. The fact that the optimal convergence exponent α_{opt} increases linearly with m shows that the mutation–selection algorithm is intrinsically parallel: it involves only local independent computations. We have here a quantitative measurement of this parallelism.

19. The logarithmic convergence exponent

We consider here the best rate of convergence which can be achieved with sequences of the form $l(n) = n^{\kappa}$, $\kappa \in \mathbb{R}_+^*$. The exponent $\alpha_{\log}$, when computed on the space E^m, is bounded. We have

$$\alpha_{\log} = \min\left\{ \frac{W(x) - W(W^*)}{H_1} : x \notin F(E^m) \right\}.$$

Consider a state x of the form $(i, \cdots, i, j)$ where $i \in W^*$ and $j \notin W^*$, $\alpha(i,j) > 0$.
Clearly $W(x) \leq W(W^*) + a + c\big(f(i) - f(j)\big)$.
However, it is not fair to compare α_{opt} with this $\alpha_{\log}$ in this situation.
In fact, we are much more interested in the rate of convergence of

$$P\,(\widehat{f}(X_n) = f(f^*)/X_0 = x) = P\,([X_n] \cap f^* \neq \emptyset/X_0 = x)$$

than of $P\,([X_n] \subset f^*/X_0 = x)$. For $l(n) = n^{\frac{1}{H_1}}$, we obtain

$$P\,(\widehat{f}(X_n) < f(f^*)/X_0 = x) \quad \underset{n \to \infty}{\asymp} \quad \frac{1}{n^{\bar{\alpha}_{\log}}}$$

where $\bar{\alpha}_{\log}$ is an affine increasing function of m with

$$\lim_{m \to \infty} \frac{\bar{\alpha}_{\log}}{m} \;=\; \frac{\min(a, c\delta^*)}{H_1}.$$

The exponent $\bar{\alpha}_{\log}$ is the right object to compare with α_{opt}. Nevertheless, it may very well happen that

$$\min\left\{ \frac{\Omega(\pi)}{H_e(\pi)} : \pi \in \mathcal{C}(E^m),\ \pi \cap f^* = \emptyset \right\} \;>\; \frac{\min(a, c\delta^*)}{H_1}$$

and thus

$$\lim_{m \to \infty} \frac{\alpha_{\text{opt}}}{\bar{\alpha}_{\log}} \;>\; 1.$$

References

[1] O. Catoni, *Large deviations for annealing*, PhD Thesis, University Paris XI (1990).

[2] O. Catoni, *Rough large deviations estimates for simulated annealing. Application to exponential schedules*, The Annals of Probability **20** (1992), 1109–1146.

[3] R. Cerf, *Asymptotic convergence of genetic algorithms*, Preprint, submitted (1993).

[4] M.I. Freidlin–A.D. Wentzell, *Random perturbations of dynamical systems*, Springer–Verlag, New York, 1984.

[5] D. Goldberg, *Genetic algorithms in search, optimization and machine learning*, Addison–Wesley, 1989.

[6] J.H. Holland, *Adaptation in natural and artificial systems*, The University of Michigan Press, Ann Arbor, 1975.

[7] C.R. Hwang–S.J. Sheu, *Singular perturbed Markov chains and exact behaviours of simulated annealing process*, Journal of Theoretical Probability **Vol. 5 No. 2** (1992), 223–249.

[8] A. Trouvé, *Convergence optimale pour les algorithmes de recuits généralisés*, C. R. Acad. Sci. Paris **t.315 Série I** (1992), 1197–1202.

[9] A. Trouvé, *Cycle decompositions and simulated annealing*, Preprint (1993).

[10] A. Trouvé, *Parallélisation massive du recuit simulé*, PhD Thesis, University Paris XI (1993).

[11] A. Trouvé, *Rough large deviation estimates for the optimal convergence speed exponent of generalized simulated annealing algorithms*, Preprint (1993).

Chapitre Quatrième

UN NOUVEL ALGORITHME GÉNÉTIQUE

A NEW GENETIC ALGORITHM

Raphaël CERF

Université de Montpellier

November 1993

Abstract. Here is a new genetic algorithm. It is built by randomly perturbing a two operators crossover–selection scheme. Three conditions of biological relevance are imposed on the crossover. A new selection mechanism is used, which has the decisive advantage of preserving the diversity of the individuals in the population. The attractors of the unperturbed process are particular equi–fitness subsets of populations endowed with a rich structure. The random vanishing perturbations are twofold: local perturbations of the individuals (mutations) and loosening of the selection pressure. When the population size is greater than a critical value which depends strongly on the optimization problem, their delicate asymptotic interaction ensures the convergence (possibly in finite time) of the population toward the ideal attractor whose populations contain all the maxima of the fitness function. The process explores without respite the neighbourhoods of the best points found so far (instead of focusing on a particular point) and finds simultaneously all the global maxima of the fitness function; it seems to be the first cooperative search procedure of this kind.

1991 *Mathematics Subject Classification.* 60F10, 60J10, 92D15.

Key words and phrases. Freidlin–Wentzell theory, genetic algorithms, stochastic optimization.

I thank Alain Trouvé, Gérard Ben Arous and Alain Berlinet for useful discussions.

1. Introduction

In the realm of stochastic optimization, attention has focused essentially on two techniques during the last decade: simulated annealing [1] and evolutionary algorithms [8,9].

The fundamental problem may be stated as follows: given a finite but huge space (the size precludes any exhaustive search procedure) endowed with a transition mechanism and a real–valued function defined on this space, determine the set of its global maxima, or find at least sub–optimal points, as fast as possible.

The theory of the simulated annealing is now extensively developed and a great number of results describing the dynamics of this kind of algorithms in various settings are available. As far as we know, the most accurate work in this area has been achieved by Catoni, in the spirit of the Freidlin–Wentzell theory [2,3]. Nevertheless, the simulated annealing presents a fundamental drawback: it is sequential in nature.

In an attempt to investigate the theoretical aspects of the parallelization of this algorithm, Trouvé carried out a systematic study (initiated by Hwang and Sheu [10]) of a broader class of algorithms he baptized generalized simulated annealing [11,12,13,14]. As it turns out, this framework is also well adapted for evolutionary algorithms.

In a first paper [4], we proposed a Markovian model of Holland's simple genetic algorithm which is built by randomly perturbing a very simple selection scheme: mutations and crossovers are considered as vanishing random perturbations. We proved that the convergence toward the global maxima of the fitness function becomes possible when the population size is greater than a critical value (which depends strongly on the optimization problem). Surprisingly, the crossover is not fundamental to ensure this convergence: the crucial point is the delicate asymptotic interaction between the local perturbations of the individuals (i.e. the mutations) and the selection pressure.

In a second paper [5], we used the concepts introduced by Catoni [2,3] and further generalized by Trouvé [12,13,14] to fathom more deeply the dynamics of the two operators mutation–selection algorithm when the population size becomes very large. The key result lies in the structure of the trajectories of populations joining two uniform populations: a small group of individuals sacrifice themselves in order to create an ideal path which is then followed by all other individuals. As a consequence, the various quantities associated with the algorithm (such as the communication cost, the virtual energy, the communication altitude ...) are affine functions of the population size. We proved that the hierarchy of cycles on the set of the uniform populations stabilizes. Furthermore, if the mutation kernel is symmetric, the limiting distribution is the uniform distribution over the set of the global maxima of the fitness function.

In this third paper, we introduce two major modifications to the previous schemes. The crossover is now integrated in the unperturbed process and is not considered any more as a random vanishing perturbation: although this operator is not essential to ensure the desired convergence, it certainly increases the efficiency of the algorithm. Three conditions of biological relevance are imposed on the crossover:

• When two identical individuals mate, they produce offsprings identical to themselves.
• There is always a non–zero probability that nothing happens during a crossover.
• The two individuals of the mating pair play symmetric roles (the populations are asexual).
Amazingly enough, the first two conditions (especially the first one) are crucial for our algorithm to work.
Furthermore, we propose a new selection mechanism which has the decisive advantage of preserving the diversity of the individuals in the population.

The analysis of the algorithm follows the road opened by Freidlin and Wentzell [7]. Unlike the situations studied in [4,5], the structure of the set of the attractors of the unperturbed process is very rich: these are particular subsets of populations and they stand in one to one correspondence with the equi–fitness subsets of the state space whose cardinality is less than the population size. When the population size is large enough, there is a unique ideal attractor whose populations contain all the maxima of the fitness function. We study the communication cost between attractors: the costs of bad transitions (either those which decrease the maximal fitness of the population or those which lose some peak fitness individuals) increase linearly with the population size, whereas the costs of good transitions (those which create some new peak fitness individuals) remain bounded. As a consequence, when the population size is greater than a critical value, the minimum of the virtual energy corresponds to the ideal attractor previously described. Therefore the sequence of the stationary measures (associated with a fixed level of intensity of the perturbations) concentrates on this attractor as the perturbations vanish. The remaining problem is to adapt carefully the rate of decreasing of the perturbations in order to obtain an inhomogeneous Markov chain with the same limiting law. Besides, it is possible to ensure a stronger convergence: we may force the process to be forever trapped in the attraction basin of the ideal attractor after a finite number of transitions. Furthermore, when the population size is large, the cycles which do not contain the ideal attractor are reduced to one single attractor and the optimal convergence exponent increases faster than an affine function of the population size. We show also how our general model specializes to the case where the state space is $\{0,1\}^N$.

Finally, let us summarize the important aspects of this work.
Our algorithm finds simultaneously all the global maxima of the objective function in finite time and thus solves completely the optimization problem: it seems to be the first of this kind. The cornerstone of this cooperative search procedure is the delicate asymptotic interaction between the mutations and our enhanced selection mechanism; the process explores simultaneously and without respite the neighbourhoods of the best points found so far (instead of focusing on a particular point). Moreover, we hope to have found the right way of using the crossover operator.

2. General notations and conventions

The cardinality of a set X will be noted indifferently $|X|$ or $\operatorname{card} X$ and its characteristic function 1_X. We adopt usual conventions concerning empty sets:

$$\prod_{\emptyset} = 1, \quad \sum_{\emptyset} = 0, \quad \min\emptyset = +\infty, \quad \max\emptyset = -\infty.$$

If s is a real number, $\lfloor s \rfloor$ denotes the unique integer such that $\lfloor s \rfloor \le s < \lfloor s \rfloor + 1$.
We consider a finite space of states E and a real-valued positive non-constant function f (which will be called the fitness function) defined on E. The points of E will be called individuals and will be mostly denoted by the letters i, j, e. The set of global maxima of f is

$$f^* = \{\, i \in E : f(i) = \max_{j \in E} f(j) \,\}$$

and $f(f^*)$ is the maximum value of f over E, i.e. $\max_{j\in E} f(j)$. Symbols with a star $*$ in superscript will denote sets realizing the minimum or the maximum of a particular functional. We put

$$\delta = \min\{\, |f(i) - f(j)| : i, j \in E, f(i) \neq f(j) \,\}, \quad \Delta = \max\{\, |f(i) - f(j)| : i, j \in E \,\}.$$

The Kronecker symbol $\delta(i,j)$ will be used to denote the identity matrix indexed by E:

$$\forall i, j \in E \qquad \delta(i,j) = 0 \quad \text{if} \quad i \neq j, \quad \delta(i,j) = 1 \quad \text{if} \quad i = j.$$

We will consider Markov chains with state space E^m, where m is the (variable) population size of our algorithm; the m–uples of elements of E, i.e. the points of the set E^m, are called populations and will be mostly denoted by the letters x, y, z, u, v.
We define a bracket operator $[\ ]$ on E^m with values in $\mathcal{P}(E)$, the set of all the subsets of E, by

$$x = (x_1, \cdots, x_m) \in E^m \ \longmapsto\ [x] = \{\, x_k : 1 \le k \le m \,\}$$

i.e. $[x]$ is the set of all individuals present in the population x.
If F is a subset of E, $f(F)$ (or sometimes simply fF) denotes the set $\{\, f(i) : i \in F \,\}$.
With f we associate a function $\widehat{f}$ defined on E^m by

$$\widehat{f}(x) = \widehat{f}(x_1, \cdots, x_m) = \max\{\, f(x_k) : 1 \le k \le m \,\}.$$

For x in E^m, $\widehat{x}$ denotes the set of those elements of $[x]$ which realize the value $\widehat{f}(x)$:

$$\widehat{x} = \{\, x_k : 1 \le k \le m,\ f(x_k) = \widehat{f}(x) \,\}.$$

The previous definitions $[x], \widehat{f}(x), \widehat{x}$ are extended to $\mathcal{P}(E^m)$, the set of all the subsets of E^m, in a natural way: if K is a subset of E^m, we have

$$\begin{aligned}
[K] &= \{\, x_k : 1 \leq k \leq m,\ x = (x_1, \cdots, x_m) \in K \,\}, \\
\widehat{f}(K) &= \max \{\, f(x_k) : 1 \leq k \leq m,\ x = (x_1, \cdots, x_m) \in K \,\}, \\
\widehat{K} &= \{\, x_k : 1 \leq k \leq m,\ x = (x_1, \cdots, x_m) \in K,\ f(x_k) = \widehat{f}(K) \,\}.
\end{aligned}$$

Let λ be a positive real number. We put

$$f_\lambda = f^{-1}(\{\lambda\}), \quad f_\lambda^+ = f^{-1}(]\lambda, \infty[), \quad f_\lambda^- = f^{-1}([0, \lambda[)$$

and we define $\Lambda = \max \{\, |f_\lambda| : \lambda \in \mathbb{R}_+^* \,\}$.
For x in E^m, we let $x^\lambda = [x] \cap f_\lambda$ and we number the elements of the set x^λ

$$x^\lambda = \{\, x_1^\lambda, \cdots, x_{|x^\lambda|}^\lambda \,\}$$

in the order they appear in the sequence $(x_k, 1 \leq k \leq m)$. Equivalently, we have

$$\forall h, k \quad 1 \leq h < k \leq |x^\lambda| \qquad \min \{\, r : x_r = x_h^\lambda \,\} < \min \{\, r : x_r = x_k^\lambda \,\}.$$

Such a numbering is clearly unique. Whenever λ does not belong to $f[x]$, the set x^λ is empty. For the special case $\lambda = \widehat{f}(x)$, we use the notation $\widehat{x} = \{\, \widehat{x}_1, \cdots, \widehat{x}_{|\widehat{x}|} \,\}$ (i.e. we replace x^λ by $\widehat{x}$ in the preceding notation).
Finally, for x in E^m, we denote by $\lambda_1^x, \cdots, \lambda_{|f[x]|}^x$ the $|f[x]|$ elements of the set $f[x]$, where again the indexation respects the order of appearance of the elements of $f[x]$ in the sequence $(f(x_k), 1 \leq k \leq m)$, or equivalently

$$\forall h, k \quad 1 \leq h < k \leq |f[x]| \qquad \min \{\, r : f(x_r) = \lambda_h^x \,\} < \min \{\, r : f(x_r) = \lambda_k^x \,\}.$$

When the context is unambiguous, we will drop the superscript in the above notations.
For any integer r, the set of the permutations over $\{1 \cdots r\}$ is denoted by $\mathfrak{S}(r)$.

3. The unperturbed Process (X_n^∞)

When there is no random perturbation, the process under study is a Markov chain (X_n^∞) with state space E^m. The superscript ∞ reflects the fact that this process describes the limit behavior of our model, when all perturbations vanish. The transition mechanism from X_n^∞ to X_{n+1}^∞ is decomposed in two stages:

$$X_n^\infty \xrightarrow{\text{crossover}} Z_n^\infty \xrightarrow{\text{selection}} X_{n+1}^\infty.$$

We describe now in detail the crossover and selection operators.

3.1. $X_n^\infty \longrightarrow Z_n^\infty$**: crossover.** The phenomenon of crossover is modelled as a random operation on the couples formed by consecutive individuals of the population X_n^∞. Let β be a markovian kernel on the space $E\times E$, that is a function defined on $(E\times E)\times(E\times E)$ with values in $[0,1]$ verifying

$$\forall (i_1,j_1)\in E\times E \qquad \sum_{(i_2,j_2)\in E\times E} \beta((i_1,j_1),(i_2,j_2)) = 1.$$

The quantity $\beta((i_1,j_1),(i_2,j_2))$ is interpreted as the probability of producing the pair of individuals (i_2,j_2) by performing a crossover on the couple (i_1,j_1). We impose the following conditions on the kernel β:

$$\forall i\in E \qquad \beta((i,i),(i,i)) = 1, \tag{1}$$

$$\forall i,j\in E \qquad \beta((i,j),(i,j)) > 0, \tag{2}$$

$$\forall i_1,j_1,i_2,j_2\in E \qquad \beta((i_1,j_1),(i_2,j_2)) = \beta((j_1,i_1),(j_2,i_2)). \tag{3}$$

These three conditions have a natural biological interpretation.
Condition (1) states that when two identical individuals mate, they produce offsprings identical to themselves. Condition (2) states that there is a non–zero probability that nothing happens during the crossover. Condition (3) states that the two individuals of the mating pair play symmetric roles (our population is asexual).
Amazingly enough, we will see later that the first two conditions (especially the first one) are crucial for our algorithm to work.
The transition probabilities from X_n^∞ to Z_n^∞ are

$$P\left(Z_n^\infty = z / X_n^\infty = x\right) = \delta_m(x_m,z_m) \prod_{1\le k\le m/2} \beta\left((x_{2k-1},x_{2k}),(z_{2k-1},z_{2k})\right) \tag{4}$$

where $\delta_m(i,j)=\delta(i,j)$ if m is odd (the last individual of the population has no mating partner and remains unchanged after crossover) and $\delta_m(i,j)=1$ if m is even.
Notice here a fundamental difference with the model studied in [4]: the crossover is now incorporated into the unperturbed process.

3.2. $Z_n^\infty \longrightarrow X_{n+1}^\infty$: selection. We propose here an enhanced version of the selection mechanism used in the previous models [4,5] which has the decisive advantage of preserving the diversity of the individuals present in the population. This mechanism may be described roughly as follows. Suppose $Z_n^\infty = z$. To build the population $X_{n+1}^\infty = x$, we first select randomly with the uniform distribution a permutation σ of $\mathfrak{S}(|\widehat{z}|)$. We divide the set of indices $\{1 \cdots m\}$ in $|\widehat{z}|+1$ parts of approximately the same size (around $\lfloor m/(|\widehat{z}|+1) \rfloor$). The components of the r–th part (for $1 \le r \le |\widehat{z}|$) are set equal to $\widehat{z}_{\sigma(r)}$, that is, roughly, for the indices k such that

$$1+(r-1)\left\lfloor \frac{m}{|\widehat{z}|+1} \right\rfloor \le k \le r \left\lfloor \frac{m}{|\widehat{z}|+1} \right\rfloor$$

we put $x_k = \widehat{z}_{\sigma(r)}$. The components of the $(|\widehat{z}|+1)$–th part are chosen independently with the uniform distribution on the set $\widehat{z}$.
We now explicit precisely this transition mechanism.
We first define a triangular array of integers $\tau(k,h),\ 0 \le k \le h+1,\ 1 \le h \le m$, by

$$\begin{aligned} &\forall h \in \{1 \cdots m\} && \tau(0,h)=1, \quad \tau(h+1,h)=m, \\ &\forall k,h \quad 1 \le k \le h \le m && \tau(k,h) = 2k \left\lfloor \frac{m}{2(h+1)} \right\rfloor + 1. \end{aligned}$$

Let x and z be two elements of E^m. If there exists a permutation σ of $\mathfrak{S}(|\widehat{z}|)$ such that

$$\forall h \in \{1 \cdots |\widehat{z}|\} \quad \forall k \quad \tau(h-1,|\widehat{z}|) \le k < \tau(h,|\widehat{z}|) \qquad x_k = \widehat{z}_{\sigma(h)}$$

then

$$P\left(X_{n+1}^\infty = x / Z_n^\infty = z\right) = \frac{1}{|\widehat{z}|!} \prod_{\tau(|\widehat{z}|,|\widehat{z}|) \le k \le m} \frac{1_{\widehat{z}}(x_k)}{\operatorname{card} \widehat{z}}. \tag{5}$$

If no such permutation exists, then $P\left(X_{n+1}^\infty = x / Z_n^\infty = z\right) = 0$.

Remark 1. In formula (5), the $|\widehat{z}|!$ stands for the choice of a random permutation belonging to $\mathfrak{S}(|\widehat{z}|)$, and the product corresponds to the choice of the components of x whose indices belong to the $(|\widehat{z}|+1)$–th part.

Remark 2. The first $|\widehat{z}|$ parts of x have an even cardinal, so that the crossover can't act on a pair of individuals belonging to distinct parts. Since in addition each such part contains only one type of individual, condition (1) shows that the crossover operator will have no effect on the first $|\widehat{z}|$ parts of x. The main interest of the $(|\widehat{z}|+1)$–th part is to give the opportunity to distinct individuals of $|\widehat{z}|$ to mate without constraints.

Remark 3. If $m < 2(|\widehat{z}|+1)$, the first $|\widehat{z}|$ parts of x are empty, so that the components of x are chosen independently with the uniform distribution on the set $\widehat{z}$. However, the dynamics of the algorithm becomes particularly interesting when m is large (as will be shown later), and such situations won't then occur.

4. The attractors of the chain (X_n^∞)

According to the classical theory of Markov chains [6, XV.7, Theorem 3], the set of populations E^m is partitioned into the transient populations and a finite number of irreducible subsets: these subsets play the role of the attractors of the deterministic system in the Freidlin–Wentzell theory [7] and will thus be called the attractors in the sequel. The aim of this section is to investigate the zoology of the attractors of the chain (X_n^∞).

Definition–Proposition 4.1. (characterization of the attractors)
The attractors of the process (X_n^∞) are the sets of populations K such that

i) $[K] = \widehat{K}$

ii) a population $x = (x_1, \cdots, x_m)$ of E^m belongs to K if and only if

- $\forall r \in \{1 \cdots |\widehat{K}|\} \quad \forall k, h \qquad \tau(r-1, |\widehat{K}|) \le k, h < \tau(r, |\widehat{K}|) \quad \Longrightarrow \quad x_k = x_h,$
- $\{\, x_k : \ 1 \le k < \tau(|\widehat{K}|, |\widehat{K}|) \,\} = [K],$
- $\{\, x_k : \tau(|\widehat{K}|, |\widehat{K}|) \le k \le m \,\} \subset [K].$

The set of all the attractors is denoted by $\mathcal{K}$ (thus $\mathcal{K} \in \mathcal{P}(\mathcal{P}(E^m))$).

Proof. We do not give a precise formal proof of this statement: it is a quite straightforward consequence of the transition mechanism of the Markov chain (X_n^∞) and such a proof would be of little interest. $\square$

Notice that the populations of the attractors are equi-fitness populations (i.e. populations whose all individuals have the same fitness). More precisely, we have

$$\forall K \in \mathcal{K} \quad \exists \lambda \in \mathbb{R}_+^* \quad \forall x \in K \qquad [x] \subset f_\lambda$$

i.e. all the individuals belonging to the populations of a fixed attractor are in the same level set of f. We denote by $\mathcal{K}_\lambda$ (respectively $\mathcal{K}_\lambda^+, \mathcal{K}_\lambda^-$) the set of attractors K such that $[K] \subset f_\lambda$ (resp. $[K] \subset f_\lambda^+, [K] \subset f_\lambda^-$). For an attractor K, we denote by $f(K)$ the unique real number λ such that K belongs to $\mathcal{K}_\lambda$. We denote by $\mathcal{K}_*$ the attractors included in f^* (i.e. $\mathcal{K}_* = K_{f(f^*)}$) and by K^* the unique attractor (it exists for $m \ge |f^*|$) such that $[K^*] = f^*$. We let also $\mathcal{K}_*^- = \mathcal{K}_{f(f^*)}^-$.

The transition mechanism implies that the process (X_n^∞) is instantaneously absorbed in the set of the populations which belong to attractors, i.e.

$$\forall x \in E^m \qquad P\,(\forall n \ge 1 \quad \exists K \in \mathcal{K} \quad X_n^\infty \in K / X_0^\infty = x) = 1.$$

Moreover, as a consequence of condition (2) on the crossover kernel β, we see that, for each attractor K and each population x in K, the probability $P\left(X_{n+1}^\infty \in K / X_n^\infty = x\right)$ is positive so that the process has a non–zero probability of staying in an attractor.

We distinguish two kinds of attractors.
An attractor K is said to be unstable if

$$\exists x \in K \qquad P\left(X_{n+1}^{\infty} \in K / X_n^{\infty} = x\right) < 1.$$

An attractor K is said to be stable if

$$\forall x \in K \qquad P\left(X_{n+1}^{\infty} \in K / X_n^{\infty} = x\right) = 1.$$

Notice that the bracket operator $[\ ]$ provides a one to one correspondence between the set $\mathcal{K}$ of all the attractors and the subset $\mathcal{E}$ of $\mathcal{P}(E)$ defined by

$$\mathcal{E} = \{\, F \subset E : |F| \leq m,\ \widehat{F} = F \,\}.$$

In fact, if K is an attractor, for each x in K, we have $[x] = [K]$, i.e. the bracket operator is constant over K and thus characterizes K. As a consequence, two distinct attractors do not intersect.
When m is small, the set $\mathcal{E}$ clearly depends on m, but for $m \geq \Lambda$, it does not depend any more on m, nor does the structure of the set of attractors $\mathcal{K}$: only the size of the populations changes, and the composition of the attractors is stabilized.
We use the crossover kernel β to build an operator on the set $\mathcal{P}(E)$. If F is a subset of E, we define

$$\beta(F) = \{\, i \in E : \exists (i', j', j) \in F \times F \times E \quad \beta((i', j'), (i, j)) > 0 \,\}.$$

The operator $\widehat{\beta}$ is the composition of the operators β and $\widehat{\ }$, i.e.

$$\widehat{\beta}(F) = \{\, i \in \beta(F) : f(i) = \widehat{f}(\beta(F)) \,\}.$$

We have the following characterization of the stable attractors.

Lemma 4.1. *An attractor K of $\mathcal{K}$ is stable if and only if $\widehat{\beta}([K]) = [K]$ and $m \geq 2(|\widehat{K}|+1)$.*

Proof. This is an immediate consequence of the transition mechanism of the process (X_n^{∞}): if the above conditions are satisfied, we have

$$P\left([X_{n+1}^{\infty}] = [K] / X_n^{\infty} \in K\right) = 1.$$

If $\widehat{\beta}([K]) \neq [K]$, there is a non–zero probability that the crossover creates new individuals not belonging to $[K]$ with a fitness greater or equal than $\widehat{f}(K)$, so that X_{n+1}^{∞} has a positive probability of leaving K. If $m < 2(|\widehat{K}| + 1)$, the selection mechanism does not guarantee the survival of all the individuals of $[K]$ so that

$$P\left([X_{n+1}^{\infty}] \subsetneq [K] / X_n^{\infty} \in K\right) > 0. \quad \square$$

We define a partial relation $\underset{\infty}{\prec}$ on $\mathcal{K}$: for each pair K_1, K_2 of attractors, we have

$$K_1 \underset{\infty}{\prec} K_2 \qquad \Longleftrightarrow \qquad \exists x \in K_1 \quad \exists y \in K_2 \quad \exists r \in \mathbb{N} \qquad P\left(X^\infty_{n+r} = y / X^\infty_n = x\right) > 0.$$

This relation is reflexive and transitive. In addition, the process X^∞_n can leave an attractor only by creating with the crossover new individuals whose fitness is greater or equal to the fitness of the starting attractor. Thus

$$K_1 \underset{\infty}{\prec} K_2 \qquad \Longrightarrow \qquad [K_2] \subset \bigcup_{n=0}^{\infty} \underbrace{\widehat{\beta}(\cdots \widehat{\beta}(\widehat{\beta}}_{n \text{ times } \widehat{\beta}}([K_1])))$$

and the fitness necessarily increases during such a transition:

$$K_1 \underset{\infty}{\prec} K_2, \quad f(K_1) \neq f(K_2) \qquad \Longrightarrow \qquad f(K_1) < f(K_2).$$

Suppose $m \geq 2(\Lambda + 1)$. The selection mechanism then never causes a loss of diversity within a level set of f so that

$$K_1 \underset{\infty}{\prec} K_2, \quad f(K_1) = f(K_2) \qquad \Longrightarrow \qquad [K_1] \subset [K_2]$$

and the relation $\underset{\infty}{\precsim}$ is then a partial order on $\mathcal{K}$.
If x is a population of E^m which belongs to a (necessarily unique) attractor K, we put $K(x) = K$. It follows from the very definition of the relation $\underset{\infty}{\prec}$ that

$$\forall n \geq 1 \qquad K(X^\infty_n) \underset{\infty}{\prec} K(X^\infty_{n+1})$$

i.e. the sequence $(K(X^\infty_n))_{n \geq 1}$ is increasing in the ordered set $(\mathcal{K}, \underset{\infty}{\precsim})$. Since $\mathcal{K}$ is finite, this sequence is stationary; with probability one, the limit

$$\lim_{n \to \infty} K(X^\infty_n) = K_\infty$$

is a maximal element of $\mathcal{K}$ for the order $\underset{\infty}{\precsim}$. Yet, the maximal elements of $\mathcal{K}$ are precisely the stable attractors. Finally, let K be a stable attractor. For any populations x, y belonging to K, the probability transition $P\left(X^\infty_{n+1} = y / X^\infty_n = x\right)$ is independent of x and y (it is completely determined by the set $[K]$). As a consequence, the process (X^∞_n) admits a unique invariant probability measure on the attractor K, which is the uniform distribution.

Remark. The main role of the crossover is to make unstable some attractors. Suppose for instance we define the kernel β by

$$\forall i_1, j_1, i_2, j_2 \in E \qquad \beta((i_1, j_1), (i_2, j_2)) = \frac{1}{2}\delta(i_1, i_2)\delta(j_1, j_2) + \frac{1}{2}\delta(i_1, j_2)\delta(j_1, i_2)$$

so that the crossover can only exchange the individuals of the mating pair and never creates new individuals. Then, all the attractors are stable whenever $m \geq 2(|\widehat{K}| + 1)$.

5. The perturbed Markov chain (X_n^l)

The previous Markov chain (X_n^∞) is randomly perturbed by two distinct mechanisms. The first one acts directly upon the population and mimics the phenomenon of mutation. The second one consists in loosening the selection of the individuals. The intensity of the perturbations is governed by an integer parameter l: as l goes to infinity, the perturbations progressively disappear. The transition mechanism of the perturbed Markov chain (X_n^l) is decomposed in three stages:

$$X_n^l \quad \xrightarrow{\text{mutation}} \quad Y_n^l \quad \xrightarrow{\text{crossover}} \quad Z_n^l \quad \xrightarrow{\text{selection}} \quad X_{n+1}^l.$$

We now describe precisely these three phases.

5.1. $X_n^l \longrightarrow Y_n^l$**: mutation.** The mutations are modelled by random independent perturbations of the individuals of the population X_n^l. Let α be an irreducible Markov kernel on the space E, that is a function defined on $E \times E$ with values in $[0,1]$ satisfying

$$\forall i \in E \qquad \sum_{j \in E} \alpha(i,j) = 1,$$

$$\forall j_1, j_2 \in E \quad \exists e_1, \cdots, e_r \quad e_1 = j_1, \quad e_r = j_2, \quad \forall k \in \{1 \cdots r-1\} \quad \alpha(i_k, i_{k+1}) > 0.$$

Let a be a positive real number. Define

$$\alpha_l(i,j) = \begin{cases} \alpha(i,j) l^{-a} & \text{if } i \neq j \\ 1 - \sum_{e \neq i} \alpha(i,e) l^{-a} & \text{if } i = j \end{cases}$$

so that α_l is an irreducible Markov kernel on E.
The transition probabilities from X_n^l to Y_n^l are given by

$$P\left(Y_n^l = y / X_n^l = x\right) = \alpha_l(x_1, y_1) \cdots \alpha_l(x_m, y_m).$$

Note that

$$(6) \qquad \lim_{l \to \infty} P\left(Y_n^l = y / X_n^l = x\right) = \delta(x_1, y_1) \cdots \delta(x_m, y_m)$$

i.e. the mutations vanish when l goes to infinity.

5.2. $Y_n^l \longrightarrow Z_n^l$**: crossover.** The crossover is not perturbed in any way: this stage is exactly the same as the passage from X_n^∞ to Z_n^∞ (formula (4)). We define the crossover kernel β on $E^m \times E^m$ by

$$\forall y, z \in E^m \qquad \beta(y,z) = P\left(Z_n^l = z / Y_n^l = y\right) = P\left(Z_n^\infty = z / X_n^\infty = y\right).$$

5.3. $Z_n^l \longrightarrow X_{n+1}^l$**: selection.** We first describe informally the selection mechanism. Suppose $Z_n^l = z$ and we wish to build the vector $X_{n+1}^l = x$.
We first traverse $f[z]$; with each element λ of this set, we associate a sequence $\psi_1, \cdots, \psi_{n_\lambda}$ obtained by reordering randomly (all orders being equally probable) the set $f^{-1}(\{\lambda\}) \cap f[z]$. The population x is then built in the following way: for each component $x_k, 1 \le k \le m$, we draw a value λ under a distribution probability on the set $f[z]$ which is biased toward the high values. As l goes to infinity, this distribution concentrates on the value $\widehat{f}(z)$. With this value λ, we had previously associated a sequence $\psi_1, \cdots, \psi_{n_\lambda}$. We divide the set $\{1 \cdots m\}$ in $n_\lambda + 1$ parts. If the index k under consideration belongs to the r–th part, where $1 \le r \le n_\lambda$, we set $x_k = \psi_r$. If k belongs to the $(n_\lambda + 1)$–th part, we choose x_k randomly and uniformly over the set $\{ \psi_1, \cdots, \psi_{n_\lambda} \}$.
We now describe precisely this mechanism.
Suppose always $Z_n^l = z$. Let us recall some notations. The set $f[z]$ contains $|f[z]|$ distinct values $\lambda_1^z, \cdots, \lambda_{|f[z]|}^z$; throughout this section, when the context is unambiguous, we will drop the superscript z in this notation, so that λ_k will stand for λ_k^z. For each λ in $f[z]$, there are $|z^\lambda|$ distinct individuals in z whose fitness is equal to λ:

$$z^\lambda = \{ z_1^\lambda, \cdots, z_{|z^\lambda|}^\lambda \}.$$

The vector $X_{n+1}^l = x$ is built in the following way. For each h in $\{1 \cdots |f[z]|\}$, we select independently and randomly with the uniform distribution a permutation σ^h belonging to $\mathfrak{S}(|z^{\lambda_h}|) = \mathfrak{S}(|[z] \cap f_{\lambda_h}|)$. The law of each component x_k $(1 \le k \le m)$ of x is defined as follows: we select randomly a value λ in the set $f[z]$ with the distribution

$$\forall h \in \{1 \cdots |f[z]|\} \qquad P(\lambda = \lambda_h) = \frac{\exp(c\lambda_h \ln l)}{\sum_{r=1}^{|f[z]|} \exp(c\lambda_r \ln l)}$$

(where c is a positive real number).
The value of x_k is then chosen in the set z^λ according to the value of the index k:
- if $\tau(|z^\lambda|, |z^\lambda|) \le k \le \tau(|z^\lambda| + 1, |z^\lambda|) = m$, then x_k is chosen at random with the uniform distribution over the set $z^\lambda = [z] \cap f_\lambda$ (k lies in the last part of the set of indices).
- if there exists an index r in $\{1 \cdots |z^\lambda|\}$ such that $\tau(r-1, |z^\lambda|) \le k < \tau(r, |z^\lambda|)$, then we put $x_k = z_{\sigma^h(r)}^\lambda$, where h is the unique integer in $\{1 \cdots |z^\lambda|\}$ satisfying $\lambda = \lambda_h$ (the index k lies in the r–th part of the set of indices, where $1 \le r \le |z^\lambda|$).

Remark. The selection mechanism at work in the chain (X_n^∞) is obtained as a particular case: the distribution probability on $f[z]$ is degenerated and attributes mass one to $\widehat{f}(z)$.

We give now an explicit expression for the transition probabilities.
Let x and z belong to E^m. If $[x] \not\subset [z]$, we put $\sigma(z, x) = \emptyset$.
Suppose $[x] \subset [z]$. Let $\sigma(z, x)$ be the set of the $|f[z]|$–uples of permutations

$$(\sigma^1, \cdots, \sigma^{|f[z]|}) \;\in\; \mathfrak{S}(|z^{\lambda_1}|) \times \cdots \times \mathfrak{S}(|z^{\lambda_{|f[z]|}}|)$$

satisfying the following property:
for each k in $\{1\cdots m\}$, if h is the unique integer in $\{1\cdots|f[z]|\}$ such that $\lambda_h = f(x_k)$, and if for some r in $\{1\cdots|z^{\lambda_h}|\}$ we have $\tau(r-1,|z^{\lambda_h}|)\leq k<\tau(r,|z^{\lambda_h}|)$, then $x_k = z^{\lambda_h}_{\sigma^h(r)}$.
The transition probabilities from Z_n^l to X_{n+1}^l are then given by the intuitive formula:

$$P\left(X_{n+1}^l = x/Z_n^l = z\right) = \tag{7}$$

$$\frac{|\sigma(z,x)|}{\prod\limits_{\lambda\in f[z]}|z^\lambda|!}\left(\prod_{\lambda\in f[z]}\prod_{k=\tau(|z^\lambda|,|z^\lambda|)}^{m}\left(\frac{1_{z^\lambda}(x_k)}{|z^\lambda|}+1-1_{z^\lambda}(x_k)\right)\right)\prod_{k=1}^m\frac{\exp(cf(x_k)\ln l)}{\sum_{\lambda\in f[z]}\exp(c\lambda\ln l)}.$$

We define the kernel γ on $E^m\times E^m$ by the identity

$$P\left(X_{n+1}^l = x/Z_n^l = z\right) = \gamma(z,x)\prod_{k=1}^m\frac{\exp(cf(x_k)\ln l)}{\sum_{\lambda\in f[z]}\exp(c\lambda\ln l)}.$$

Only the last product in formula (7) depends upon l. It may be rewritten as

$$\frac{\exp(c\sum_{k=1}^m f(x_k)\ln l)}{\left(\sum_{\lambda\in f[z]}\exp(c\lambda\ln l)\right)^m}.$$

As l goes to infinity, this term is equivalent to

$$\exp\left(-c\big(m\widehat{f}(z)-\sum_{k=1}^m f(x_k)\big)\ln l\right)$$

which tends to zero whenever $[x]$ is not included in $\widehat{z}$. Thus

$$\lim_{l\to\infty}P\left(X_{n+1}^l = x/Z_n^l = z\right) = \frac{|\sigma(z,x)|}{\prod\limits_{\lambda\in f[z]}|z^\lambda|!}\left(\prod_{k=1}^m 1_{\widehat{z}}(x_k)\right)|\widehat{z}|^{\tau(|\widehat{z}|,|\widehat{z}|)-1-m}.$$

Yet, for x such that $[x]\subset\widehat{z}$, we have either $\sigma(z,x)=\emptyset$ and $P\left(X_{n+1}^l = x/Z_n^l = z\right)=0$, or

$$|\sigma(z,x)| = \prod_{\lambda\in f[z]\setminus\{\widehat{f}(z)\}}|z^\lambda|!$$

in which case (see formula (5))

$$\lim_{l\to\infty}P\left(X_{n+1}^l = x/Z_n^l = z\right) = \frac{|\widehat{z}|^{\tau(|\widehat{z}|,|\widehat{z}|)-1-m}}{|\widehat{z}|!} = P\left(X_{n+1}^\infty = x/Z_n^\infty = z\right) \tag{8}$$

i.e. the limiting selection mechanism is the one used for the Markov chain (X_n^∞).
Formulas (6) and (8) yield

$$\forall y,z\in E^m \qquad \lim_{l\to\infty}P\left(X_{n+1}^l = z/X_n^l = y\right) = P\left(X_{n+1}^\infty = z/X_n^\infty = y\right)$$

so that the process (X_n^l) is a perturbation of the process (X_n^∞). A crucial point is that the perturbations really interact as l goes to infinity. More precisely, the rates of convergence in (6) and (8) should be logarithmically of the same order.

6. Example: the case $E = \{0,1\}^N$

We specialize our general model to the case where the state space E is $\{0,1\}^N$ ($N \in \mathbb{N}$). A point i of E is a word of length N over the alphabet $\{0,1\}$ and is noted $i = i_1 \cdots i_N$ where $i_k \in \{0,1\}$. The Hamming distance $H(i,j)$ between two points i,j of E is the number of letters where i and j differ:

$$H(i,j) = \text{card}\,\{\, k : 1 \leq k \leq m,\ i_k \neq j_k \,\}.$$

The mutation kernel α is defined by

$$\alpha(i,j) = \begin{cases} 0 & \text{if } H(i,j) > 1, \\ 1/N & \text{if } H(i,j) = 1. \end{cases}$$

It is irreducible: the minimal number of transitions necessary to join two arbitrary points of E through the kernel α is N.
In order to build the crossover operator, we define now a cutting operator T_k for k in $\{0 \cdots N\}$; T_k maps $E \times E$ onto $E \times E$ and for i,j in E, we put $T_k(i,j) = (i',j')$ where

$$i' = i_1 \cdots i_k j_{k+1} \cdots j_N, \qquad j' = j_1 \cdots j_k i_{k+1} \cdots i_N.$$

Notice that T_0 only exchanges the two individuals of the mating pair, i.e. $T_0(i,j) = (j,i)$, whereas T_N is the identity map over $E \times E$.
For any pairs (i,j) and (i',j') of $E \times E$, we put

$$C((i,j),(i',j')) = \text{card}\,\{\, k : 1 \leq k \leq N,\ T_k(i,j) = (i',j') \,\}$$

and we define finally the crossover kernel β by

$$\beta((i,j),(i',j')) = \frac{C((i,j),(i',j')) + C((j,i),(i',j'))}{\sum_{(i'',j'') \in E \times E} C((i,j),(i'',j'')) + C((j,i),(i'',j''))}.$$

It is a straightforward matter to check that conditions (1), (2), (3) are satisfied.

7. A numerical example on the space $E = \{0,1\}^3$

We consider the space $E = \{0,1\}^3$ and we define the fitness function f by

$$f(\{001,011\}) = \{1\},\ f(\{010,101\}) = \{2\},\ f(\{100\}) = \{3\},\ f(\{000,110,111\}) = \{4\}.$$

We take $m = 10$. There are 15 attractors. If F is an equi–fitness subset of E, we denote by $K(F)$ the associated attractor. The attractors $K(010,101)$ and $K(000,111)$ are unstable. All others attractors are stable. Actually, we have

$$K(000,111) \underset{\infty}{\prec} K(000,110,111), \quad K(010,101) \underset{\infty}{\prec} K(100), \quad K(010,101) \underset{\infty}{\prec} K(110).$$

The ideal attractor $K^* = K(000,110,111)$ contains $3!\,3^4 = 486$ populations.
A typical example of such a population is

$$(110, 110, 000, 000, 111, 111, 000, 111, 110, 000).$$

8. Asymptotic expansion of $P\left(X_{n+1}^l = v / X_n^l = u\right)$

By the very construction of the process (X_n^l), we have

$$P\left(X_{n+1}^l = v/X_n^l = u\right) = \sum_{y,z\in E^m} P\left(X_{n+1}^l = v/Z_n^l = z\right) P\left(Z_n^l = z/Y_n^l = y\right) P\left(Y_n^l = y/X_n^l = u\right).$$

For each y, z in E^m,

$$P\left(X_{n+1}^l = v/Z_n^l = z\right) P\left(Z_n^l = z/Y_n^l = y\right) P\left(Y_n^l = y/X_n^l = u\right) \underset{l\to\infty}{\sim}$$

$$\alpha(u,y)\,\beta(y,z)\,\gamma(z,v)\exp\left(-\Big(ad(u,y) + c\sum_{k=1}^{m}(\widehat{f}(z) - f(v_k))\Big)\ln l\right) \tag{9}$$

where we note for u, y in E^m

$$\alpha(u,y) = \prod_{k:u_k\neq y_k} \alpha(u_k, y_k)$$

and $d(u,y)$ is the Hamming distance between the vectors u and y, i.e.

$$d(u,y) = \operatorname{card}\{\, k : 1 \leq k \leq m,\ u_k \neq y_k \,\}.$$

The above quantity (9) vanishes whenever $\alpha(u,y)\,\beta(y,z)\,\gamma(z,v) = 0$.
Let $\overline{D}_1(u,v)$ be the set of all four–uples (u,y,z,v) satisfying $\alpha(u,y)\,\beta(y,z)\,\gamma(z,v) > 0$.
We define next the communication cost V_1 on $E^m \times E^m$ by

$$V_1(u,v) = \min_{(u,y,z,v)\in\overline{D}_1(u,v)} ad(u,y) + c\sum_{k=1}^{m}(\widehat{f}(z) - f(v_k)).$$

and we note $\overline{D}_1^*(u,v)$ the elements of $\overline{D}_1(u,v)$ which realize the above minimum. Putting

$$q_1(u,v) = \sum_{(u,y,z,v)\in\overline{D}_1^*(u,v)} \alpha(u,y)\,\beta(y,z)\,\gamma(z,v),$$

we have

$$P\left(X_{n+1}^l = v/X_n^l = u\right) \underset{l\to\infty}{\sim} q_1(u,v)\exp\left(-V_1(u,v)\ln l\right).$$

Moreover, notice that for each u, v in E^m,

$$P\left(X_{n+1}^l = v/X_n^l = u\right) = 0 \iff V_1(u,v) = \infty \iff q_1(u,v) = 0.$$

We are now in the framework of the generalized simulated annealing studied by Trouvé [11,12,13,14]. That is, the transition probabilities of the process (X_n^l) form a family of Markov kernels on the space E^m indexed by l which is admissible for the communication kernel q_1 and the cost function V_1 [12, Definition 3.1].

9. More notations

If $\mathcal{S}$ is an arbitrary set, $\mathcal{S}^{(\mathbb{N})}$ denotes the set of paths in $\mathcal{S}$, that is the set of finite sequences of elements of $\mathcal{S}$. A path s in $\mathcal{S}$ is noted indifferently

$$s=(s_1,\cdots,s_r),\quad s=(s^1,\cdots,s^r)\quad\text{or}\quad s=s_1\to\cdots\to s_r$$

and its length is noted $|s|$ (r in the above example). A path s in $\mathcal{S}$ is said to join two elements t_1 and t_2 if $s_1=t_1$ and $s_{|s|}=t_2$; the set of all paths in $\mathcal{S}$ joining the points t_1 and t_2 is noted $\mathcal{S}^{(\mathbb{N})}(t_1,t_2)$.
We will consider paths in the sets E, E^m and $\mathcal{P}(E)$. Paths of E^m will mostly be denoted by the letter p and paths of $\mathcal{P}(E)$ by the letter q.
By D^m we denote the paths in E^m which correspond to possible trajectories of the process (X_n^l), i.e. the paths p in E^m satisfying

$$\forall k\quad 1\leq k<|p|\qquad V_1(p_k,p_{k+1})<\infty.$$

The V_1–cost of such a path is

$$V_1(p)=\sum_{k=1}^{|p|-1}V_1(p_k,p_{k+1}).$$

Notice that for the empty path (which has a null length), the cost is zero.
If p belongs to $E^{m(\mathbb{N})}\setminus D^m$, we put $V_1(p)=\infty$.
Similarly, by $\overline{D}^m$ we denote the paths in E^m which correspond to possible trajectories for the whole process (the number of transitions r being variable)

$$X_n^l\to Y_n^l\to Z_n^l\to X_{n+1}^l\to Y_{n+1}^l\to\cdots\to Z_{n+r-1}^l\to X_{n+r}^l$$

i.e. such a path p includes the intermediate populations Y_{n+k}^l and Z_{n+k}^l, $1\leq k<r$, has a length equal to 1 modulo 3 and satisfies

$$\forall k\quad 1\leq 3k<|p|\qquad \alpha(p^{3k-2},\,p^{3k-1})\,\beta(p^{3k-1},p^{3k})\,\gamma(p^{3k},\,p^{3k+1})>0.$$

The corresponding cost function $\overline{V}$ is defined by

$$\overline{V}(p)=\sum_{1\leq 3k<|p|}a\,d(p^{3k-2},\,p^{3k-1})+c\sum_{h=1}^{m}\big(\widehat{f}(p^{3k})-f(p_h^{3k+1})\big)$$

if the path p belongs to $\overline{D}^m$ (here p_h^{3k+1} is the h–th component of the vector p^{3k+1}) and $\overline{V}(p)=\infty$ otherwise. We put also for y,z in E^m and i,j in E

$$D^m(y,z)=D^m\cap E^{m(\mathbb{N})}(y,z),\quad D(i,j)=\bigcup_{m\in\mathbb{N}^*}D^m((i),(j)),\quad D=\bigcup_{m\in\mathbb{N}^*}D^m$$

and we define similarly $\overline{D}^m(y,z)$, $\overline{D}(i,j)$, $\overline{D}$ (just replace D by $\overline{D}$ in the above formulas). The bracket operator $[\ ]$ provides a natural projection from the set $\cup_{m\in\mathbb{N}^*}E^{m(\mathbb{N})}$ onto $\mathcal{P}(E)^{(\mathbb{N})}$: with each path $p=(p_1,\cdots,p_r)$ in E^m we associate the path $[p]=([p_1],\cdots,[p_r])$ in $\mathcal{P}(E)$. Finally, we put $[\overline{D}]=\{\,[p]:p\in\overline{D}\,\}$.

10. The minimal communication cost and the virtual energy

Definition 10.1. We define the minimal communication cost V for y and z in E^m by

$$V(y,z) = \inf \{ V_1(p) : p \in D^m(y,z) \}.$$

Notice that for all x in E^m, we have $V(x,x) = 0$.
If F, G are two subsets of E^m, we define

$$V(F,G) = \inf \{ V_1(p) : p \in D^m,\ p^1 \in F,\ p^{|p|} \in G \} = \inf \{ V(y,z) : y \in F,\ z \in G \}.$$

Let g be a graph on E^m. The cost of g is

$$V(g) = \sum_{(x \to y) \in g} V(x,y).$$

This definition works also for a path p in E^m if we consider the path as a graph over E^m. Notice that for the empty graph (which has no arrows), the cost is zero.
We recall that an x–graph is a graph with no arrow starting from x and such that for any $y \neq x$ there exists a unique path in g leading from y to x. The set of all x–graphs is denoted by $G(x)$. For more details and notations concerning graphs, see [7, chapter 6].

Definition 10.2. The virtual energy W associated with the cost function V is defined by

$$\forall x \in E^m \qquad W(x) = \min \{ V(g) : g \in G(x) \}.$$

We put also for any subset F of E^m

$$W(F) = \min \{ W(x) : x \in F \}, \qquad W^* = \{ x \in E^m : W(x) = W(E^m) \}.$$

Proposition 7.1 of [5] shows that this quantity could equivalently be defined through the cost function V_1 instead of V. The point is that the sequence of the stationary measures of the Markov chains $(X_n^l), l \in \mathbb{N}$, concentrates on the set W^* as l goes to infinity.

Proposition 10.1. *(Freidlin and Wentzell)*

$$\forall x \in E^m \qquad \lim_{l \to \infty} \ \lim_{n \to \infty} \quad P\left(X_n^l \in W^* / X_0^l = x \right) \ = \ 1.$$

We face here the same difficulty as in the case of the mutation–selection algorithm: the size of the state space increases geometrically with the population size m. Anyway, we are mostly interested in the way the chain (X_n^l) visits the populations of the attractors $\mathcal{K}$. We use the same technique as in [5]. Since

$$\forall x \in E^m \quad \exists K \in \mathcal{K} \qquad V(x, K) = 0,$$

then theorems 7.1 and 8.1 of [5] show that the dynamics of the process may be studied by considering only the restrictions of the communication cost V and the virtual energy W to the set of the populations belonging to the attractors. Furthermore, we have

$$\forall K \in \mathcal{K} \quad \forall x, y \in K \qquad V(x, y) = 0$$

so that

$$\forall K_1, K_2 \in \mathcal{K} \quad \forall x_1 \in K_1 \quad \forall x_2 \in K_2 \qquad V(x_1, x_2) = V(K_1, K_2)$$

and similarly

$$\forall K \in \mathcal{K} \quad \forall x \in K \quad W(x) = W(K).$$

The functionals W and V may thus be seen as defined on the set of attractors $\mathcal{K}$ rather than on the set of populations belonging to attractors.

For the sake of completeness, we recall also the definition of the communication altitude, as well as the construction of the hierarchy of cycles. These tools will only be used for Theorem 14.2. For the details, we refer the reader to [12,13].

Definition 10.3. (Trouvé, [12, Definition 3.15])
The communication altitude $A(K_1, K_2)$ between two distinct attractors K_1 and K_2 is

$$A(K_1, K_2) = \inf \{ \max_{1 \leq k < |p|} W(p_k) + V(p_k, p_{k+1}) : p \in D^m, p^1 \in K_1,\, p^{|p|} \in K_2 \}.$$

For any K in $\mathcal{K}$, we put $A(K, K) = W(K)$.

Definition 10.4. Let $\lambda \in \mathbb{R}$. We define an equivalence relation $\mathcal{R}_\lambda$ on the set

$$W_\lambda = \{ K \in \mathcal{K} : W(K) \leq \lambda \}$$

by

$$\forall K_1, K_2 \in W_\lambda, \qquad K_1 \, \mathcal{R}_\lambda \, K_2 \quad \Longleftrightarrow \quad A(K_1, K_2) \leq \lambda.$$

Proposition 10.2. *(Trouvé,* [12, Proposition 3.21]*)*
The set of cycles in $\mathcal{K}$ associated with the cost function V is

$$\mathcal{C}(\mathcal{K}) = \bigcup_{\lambda \in \mathbb{R}^+} W_\lambda \,/\, \mathcal{R}_\lambda$$

where $W_\lambda \,/\, \mathcal{R}_\lambda$ is the quotient set of the equivalence classes of W_λ for the relation $\mathcal{R}_\lambda$.

We will now study the costs of the transitions between the attractors. As it turns out, the costs of the bad transitions increase linearly with the population size m, whereas the costs of the good transitions remain bounded (as functions of m).

11. The costs of the bad transitions

We distinguish two kinds of bad transitions: those which lose some peak fitness individuals and those which decrease the fitness.

Lemma 11.1. *(loss of diversity in f^*)*
Let K_1, K_2 be two elements of $\mathcal{K}$ such that $[K_1] \subset f^*, [K_2] \subset f^*$ (i.e. $K_1,\ K_2 \in \mathcal{K}_*$).
We have

$$V(K_1, K_2) \ \geq\ \left\lfloor \frac{m}{2(|f^*|+1)} \right\rfloor \min\left(a, c\delta^*\right) card\left([K_1] \setminus [K_2]\right)$$

where $\delta^* = \min\{\, f(f^*) - f(i) : i \notin f^* \,\}$.

Proof. Let K_1, K_2 be two attractors of $\mathcal{K}_*$. Put $[K_1] \setminus [K_2] = \{\, e_1, \cdots, e_r \,\}$ and let p be a path joining the attractors K_1 and K_2, i.e. $p^1 \in K_1$, $p^{|p|} \in K_2$. We define for ι in $\{1 \cdots r\}$,

$$t(\iota) = \min\{\, k : 0 \leq k < |p|,\ k \equiv 0 \pmod 3,\ e_\iota \notin [p^{k+1}] \cap [p^{k+2}] \cap [p^{k+3}] \,\}$$

i.e. $t(\iota)$ is the last time equal to 0 modulo 3 before the disappearance of e_ι in the path p.
We make the conventions $p^0 = p^1$ and $p^{|p|+1} = p^{|p|+2} = p^{|p|}$.
The transition mechanism implies the following fact:
for each ι in $\{1 \cdots r\}$, there exists h_ι in $\{1 \cdots |\widehat{p}^{t(\iota)}|\}$ such that

$$\forall h \quad \tau(h_\iota - 1, |\widehat{p}^{t(\iota)}|) \leq h < \tau(h_\iota, |\widehat{p}^{t(\iota)}|) \qquad p_h^{t(\iota)+1} \in f^* \quad \Longrightarrow \quad p_h^{t(\iota)+1} = e_\iota.$$

Therefore, the r sets indexed by ι in $\{1 \cdots r\}$

$$\{\, t(\iota), t(\iota)+1, t(\iota)+2 \,\} \times \{\, h : \tau(h_\iota - 1, |\widehat{p}^{t(\iota)}|) \leq h < \tau(h_\iota, |\widehat{p}^{t(\iota)}|) \,\}$$

have pairwise empty intersections, and it follows that

$$\overline{V}(p) \ \geq\ \sum_{\iota=1}^{r} \sum_{h=\tau(h_\iota-1,|\widehat{p}^{t(\iota)}|)}^{\tau(h_\iota,|\widehat{p}^{t(\iota)}|)-1} c\left(\widehat{f}(p^{t(\iota)}) - f(p_h^{t(\iota)+1})\right) + a\left(1 - \delta(p_h^{t(\iota)+1},\ p_h^{t(\iota)+2})\right).$$

Let ι belong to $\{1 \cdots r\}$ and suppose

$$c\left(\widehat{f}(p^{t(\iota)}) - f(p_h^{t(\iota)+1})\right) + a\left(1 - \delta(p_h^{t(\iota)+1},\ p_h^{t(\iota)+2})\right) = 0$$

for some h satisfying $\tau(h_\iota - 1, |\widehat{p}^{t(\iota)}|) \leq h < \tau(h_\iota, |\widehat{p}^{t(\iota)}|)$. This implies $p_h^{t(\iota)+1} = p_h^{t(\iota)+2} = e_\iota$.
Suppose that

$$1 + \frac{1}{2}\left(\tau(h_\iota, |\widehat{p}^{t(\iota)}|) - \tau(h_\iota - 1, |\widehat{p}^{t(\iota)}|)\right) = \left\lfloor \frac{m}{2(|\widehat{p}^{t(\iota)}|+1)} \right\rfloor + 1$$

such terms vanish. Necessarily, there exists an odd index h such that

$$\tau(h_\iota - 1, |\widehat{p}^{t(\iota)}|) \le h < \tau(h_\iota, |\widehat{p}^{t(\iota)}|) \quad \text{and} \quad p_h^{t(\iota)+1} = p_h^{t(\iota)+2} = p_{h+1}^{t(\iota)+1} = p_{h+1}^{t(\iota)+2} = e_\iota.$$

The crossover operator does not affect the pair $(p_h^{t(\iota)+2}, p_{h+1}^{t(\iota)+2})$ (by condition (1)) and we have also $p_h^{t(\iota)+3} = p_{h+1}^{t(\iota)+3} = e_\iota$. Thus the individual e_ι is present in the populations $p^{t(\iota)+1}$, $p^{t(\iota)+2}$, $p^{t(\iota)+3}$, which contradicts the definition of $t(\iota)$.
We have proved that less than $\lfloor m/(2(|\widehat{p}^{t(\iota)}|+1))\rfloor$ terms vanish in each sum $(1 \le \iota \le r)$

$$\sum_{h=\tau(h_\iota - 1, |\widehat{p}^{t(\iota)}|)}^{\tau(h_\iota, |\widehat{p}^{t(\iota)}|)-1} c\left(\widehat{f}(p^{t(\iota)}) - f(p_h^{t(\iota)+1})\right) + a\left(1 - \delta(p_h^{t(\iota)+1}, p_h^{t(\iota)+2})\right).$$

Moreover, we know that $\widehat{f}(p^{t(\iota)}) = f(f^*)$ (since $e_\iota \in [p^{t(\iota)}]$), and each non–zero term is necessarily greater than $\min(a, c\delta^*)$, so that the above sum is greater than

$$\left(\tau(h_\iota - 1, |\widehat{p}^{t(\iota)}|) - \tau(h_\iota, |\widehat{p}^{t(\iota)}|) - \left\lfloor \frac{m}{2(|\widehat{p}^{t(\iota)}|+1)} \right\rfloor\right) \min(a, c\delta^*)$$
$$= \left\lfloor \frac{m}{2(|\widehat{p}^{t(\iota)}|+1)} \right\rfloor \min(a, c\delta^*).$$

Yet $\widehat{p}^{t(\iota)} \subset f^*$ so that $|\widehat{p}^{t(\iota)}| \le |f^*|$. Finally we obtain

$$\overline{V}(p) \;\ge\; \left\lfloor \frac{m}{2(|f^*|+1)} \right\rfloor \min(a, c\delta^*)\, r$$

(where $r = \operatorname{card}([K_1] \setminus [K_2])$) and taking the infimum over all paths p joining the attractors K_1 and K_2 yields the desired inequality. $\square$

To obtain a lower bound for the communication cost between two attractors, we will study the possible trajectories of a pair of contiguous individuals within a given path of populations joining the two attractors.

Definition 11.1. (admissible paths)
Let q be a path in $\mathcal{P}(E)$. We say that the pair $(i_k, j_k), 1 \le k \le |q|$, of paths in E is admissible for q if

i) $\forall k \in \{1 \cdots |q|\} \qquad i_k \in q^k, \quad j_k \in q^k$

ii) $\forall k \quad 1 \le 3k < |q|$

- either $\; i_{3k-2} = i_{3k-1} \;$ or $\; \alpha(i_{3k-2}, i_{3k-1}) > 0$
- either $\; j_{3k-2} = j_{3k-1} \;$ or $\; \alpha(j_{3k-2}, j_{3k-1}) > 0$
- $\beta((i_{3k-1}, j_{3k-1}), (i_{3k}, j_{3k})) > 0$
- $f(i_{3k+1}) = f(j_{3k+1}) \quad \Longrightarrow \quad i_{3k+1} = j_{3k+1}$.

The set of all pair of paths admissible for the path q is denoted by $\mathcal{A}(q)$.

We define the quantity $\rho(K_1,K_2)$ for K_1 and K_2 in $\mathcal{K}$ to be the infimum

$$\inf\Big\{\sum_{1\le 3k<|q|} a\,(2-\delta(i_{3k-2},\,i_{3k-1})-\delta(j_{3k-2},\,j_{3k-1}))\;+\;c\,(2\widehat{f}(q^{3k})-f(i_{3k+1})-f(j_{3k+1}))$$

$$\text{where}\qquad q\in[\overline{D}],\;q^1=[K_1],\;q^{|q|}=[K_2],\;(i_k,j_k)_{1\le k\le|q|}\in\mathcal{A}(q),\;i_1=j_1\Big\}.$$

We put

$$\rho=\min\big\{\,\rho(K_1,K_2):K_1,K_2\in\mathcal{K},\,\rho(K_1,K_2)>0\,\big\}.$$

Since the quantities $\rho(K_1,K_2)$ are finite sums involving terms of the form a or $c\big(f(i)-f(j)\big)$ (where $f(i)\ge f(j)$), we have $\rho\ge\min(a,\,c\delta)>0$.

Lemma 11.2. *Let K_1 and K_2 belong to $\mathcal{K}$. Suppose $\rho(K_1,K_2)=0$. Then there exists a path q in the set $[\overline{D}]$ such that*

$$q^1=[K_1],\,q^{|q|}=[K_2]\quad\textit{and}\quad\widehat{f}(q^1)\le\widehat{f}(q^4)\le\cdots\le\widehat{f}(q^{|q|-3})\le\widehat{f}(q^{|q|}),$$

i.e. the sequence $\big(\widehat{f}(q^{3k+1})\big),1\le 3k+1\le|q|$, is non–decreasing. In particular, we have $f(K_1)\le f(K_2)$.

Corollary 11.1. *Let K_1 and K_2 be two attractors such that $f(K_1)>f(K_2)$. Then $\rho(K_1,K_2)\ \ge\ \rho\ \ge\ \min(a,\,c\delta)\ >\ 0$.*

Proof. Let K_1 and K_2 be as in the hypothesis of lemma 11.2. By definition of $\rho(K_1,K_2)$, there exists a path q in $[\overline{D}]$ and a pair $(i_k,j_k),1\le k\le|q|$, of paths in E admissible for q such that $q^1=[K_1]$, $q^{|q|}=[K_2]$, $i_1=j_1$ and

$$\sum_{1\le 3k<|q|} a\,(2-\delta(i_{3k-2},\,i_{3k-1})-\delta(j_{3k-2},\,j_{3k-1}))\;+\;c\,(2\widehat{f}(q^{3k})-f(i_{3k+1})-f(j_{3k+1}))=0.$$

This relation implies that, for each k, $1\le 3k<|q|$, we have

$$(10)\qquad i_{3k-2}=i_{3k-1},\quad j_{3k-2}=j_{3k-1},\quad f(i_{3k+1})=f(j_{3k+1})=\widehat{f}(q^{3k}).$$

Since the pair of paths (i_k,j_k) is admissible for the path q, the last of the above equalities implies that $i_{3k+1}=j_{3k+1}$ for each k, $1\le 3k<|q|$. Notice also that $i_1=j_1$.
Moreover the crossover has no effect when two identical individual mates (condition (1)), so that

$$(11)\qquad\forall k\quad 1\le 3k<|q|\qquad i_{3k}=j_{3k}=i_{3k-1}=j_{3k-1}.$$

Therefore, we conclude from (10) and (11) that

$$\forall k\quad 1\le 3k<|q|\qquad\widehat{f}(q^{3k+1})=\widehat{f}(q^{3k})\ge\widehat{f}(q^{3k-2})$$

and the sequence $\big(\widehat{f}(q^{3k+1})\big),1\le 3k+1\le|q|$, is non–decreasing. $\square$

Lemma 11.3. *(lower bound for the communication cost)*
Let K_1 and K_2 belong to $\mathcal{K}$. We have the inequality

$$V(K_1,K_2)\ \geq\ \left(\frac{m}{4}-\frac{\Lambda+3}{2}\right)\,\rho(K_1,K_2).$$

(We recall that $\Lambda=\max\{\,|f_\lambda|:\lambda\in\mathbb{R}_+^*\,\}$.)

Proof. Let p be a path in $\overline{D}$ joining the attractors K_1 and K_2 (i.e. $p^1\in K_1$, $p^{|p|}\in K_2$). We make the convention $p^0=p^1$. Putting

$$\tau_k=\frac{1}{2}\big(\tau(|\widehat{p}^{3k-3}|,|\widehat{p}^{3k-3}|)-1\big)=|\widehat{p}^{3k-3}|\left\lfloor\frac{m}{2(|\widehat{p}^{3k-3}|+1)}\right\rfloor,\quad 1\leq 3k<|p|,$$

we have

$$\begin{aligned}
\overline{V}(p)\ &=\sum_{1\leq 3k<|p|}\ \sum_{h=1}^{m}\ a\left(1-\delta(p_h^{3k-2},\,p_h^{3k-1})\right)+c\left(\widehat{f}(p^{3k})-f(p_h^{3k+1})\right)\\
&\geq\sum_{1\leq 3k<|p|}\ \sum_{h=0}^{\tau_k-1} a\left(2-\delta(p_{2h+1}^{3k-2},\,p_{2h+1}^{3k-1})-\delta(p_{2h+2}^{3k-2},\,p_{2h+2}^{3k-2})\right)\\
&\qquad\qquad+\ c\left(2\widehat{f}(p^{3k})-f(p_{2h+1}^{3k+1})-f(p_{2h+2}^{3k+1})\right).
\end{aligned}\tag{12}$$

Let h belong to $\{1\cdots\Lambda\}$. Let $m=2(h+1)q+r$ be the euclidean division of m by $2(h+1)$; we have (since $h\geq 1$ and $0\leq r<2(h+1)$)

$$2h\left\lfloor\frac{m}{2(h+1)}\right\rfloor\ =\ 2hq\ =\ 2h\frac{m-r}{2(h+1)}\ =\ \frac{h}{h+1}(m-r)\ >\ \frac{1}{2}\big(m-2(h+1)\big)\ \geq\ \frac{m}{2}-\Lambda-1.$$

For each k, $1\leq 3k<|p|$, the very definition of Λ implies $1\leq|\widehat{p}^{3k-3}|\leq\Lambda$ whence

$$\tau(|\widehat{p}^{3k-3}|,|\widehat{p}^{3k-3}|)\ >\ \frac{m}{2}-\Lambda.$$

It follows from (12) that (the $\cdots$ stands for the summand in the last sum of (12))

$$\begin{aligned}
\overline{V}(p)\ &\geq\sum_{1\leq 3k<|p|}\ \sum_{0\leq h<\tau_k}\ \cdots\cdots\ =\sum_{1\leq 3k<|p|}\ \sum_{1\leq 2h+1<\tau(|\widehat{p}^{3k-3}|,|\widehat{p}^{3k-3}|)}\cdots\cdots\\
&\geq\sum_{1\leq 3k<|p|}\ \sum_{1\leq 2h+1<m/2-\Lambda}\cdots\cdots\ =\sum_{1\leq 2h+1<m/2-\Lambda}\ \sum_{1\leq 3k<|p|}\cdots\cdots
\end{aligned}$$

Yet, for each h such that $1\leq 2h+1<m/2-\Lambda$, the pair of paths (p_{2h+1}^k,p_{2h+2}^k), $1\leq k\leq|p|$, is admissible for $[p]$ and satisfy also $p_{2h+1}^1=p_{2h+2}^1$: as a consequence, the sum

$$\sum_{1\leq 3k<|p|} a\left(2-\delta(p_{2h+1}^{3k-2},\,p_{2h+1}^{3k-1})-\delta(p_{2h+2}^{3k-2},\,p_{2h+2}^{3k-2})\right)\ +\ c\left(2\widehat{f}(p^{3k})-f(p_{2h+1}^{3k+1})-f(p_{2h+2}^{3k+1})\right)$$

is greater than $\rho(K_1,K_2)$ and finally $\overline{V}(p)\ \geq\ (m/4-(\Lambda+3)/2)\ \rho(K_1,K_2)$. Taking the infimum over all the paths joining K_1 and K_2 gives the inequality of the lemma. $\square$

In the same flavor as the two preceding lemmas, we have the

Proposition 11.1. *Let K_1, K_2 be two elements of $\mathcal{K}$ such that $f(K_1) = f(K_2)$. Then*

$$V(K_1,K_2) \;\geq\; \min(a,\, c\delta) \min\left(\left\lfloor \frac{m}{2(\Lambda+1)} \right\rfloor \mathit{card}\left([K_1]\setminus[K_2]\right),\; \frac{m}{4} - \frac{\Lambda+3}{2} \right).$$

Proof. Let p be a path in E^m joining the attractors K_1 and K_2. We consider two cases.

- Suppose the sequence $(\widehat{f}(p^{3k+1})), 1 \leq 3k+1 \leq |p|$, is not non–decreasing: there exists an index $h, 1 < 3h+1 \leq |p|$, such that $f(K_1) > \widehat{f}(p^{3h+1})$ and an attractor K_3 satisfying

$$V(p^{3h+1},\, K_3) = 0, \qquad f(K_3) = \widehat{f}(p^{3h+1}).$$

Clearly $\overline{V}(p) \geq V(K_1,K_3)$ and corollary 11.1 and lemma 11.3 imply

$$V(K_1,K_3) \;\geq\; \min(a,\, c\delta) \left(\frac{m}{4} - \frac{\Lambda+3}{2} \right).$$

- Suppose the sequence $(\widehat{f}(p^{3k+1})), 1 \leq 3k+1 \leq |p|$, is non–decreasing. Then we have

$$\forall k \quad 1 \leq 3k+1 \leq |p| \qquad \widehat{f}(p^{3k+1}) = f(K_1) = f(K_2) = \theta.$$

We modify (just for the purpose of the proof!) the function f in the following way:
for each i in E, if $f(i)$ is strictly greater than θ, we set $f(i) = \theta - \delta$. The attractors K_1 and K_2 then become elements of $\mathcal{K}_*$ and we are in a situation analogous to lemma 11.1. Moreover the cost of the path p, i.e. the quantity

$$\overline{V}(p) = \sum_{1 \leq 3k < |p|} a\, d(p^{3k-2},\, p^{3k-1}) + c \sum_{h=1}^{m} \left(\widehat{f}(p^{3k}) - f(p_h^{3k+1}) \right)$$

decreases when evaluated with the new function f. Actually, for any indices h and k such that $1 \leq 3k < |p|$, $1 \leq h \leq m$, the value $f(p_h^{3k+1})$ remains unchanged (since it was lower or equal than θ) and the value $\widehat{f}(p^{3k})$, which was greater or equal than θ, becomes equal to θ (since $\widehat{f}(p^{3k+1}) = \theta$ and $[p^{3k+1}] \subset [p^{3k}]$, the set $[p^{3k}]$ necessarily intersects the set f_θ). Application of lemma 11.1 yields

$$\overline{V}(p) \;\geq\; \left\lfloor \frac{m}{2(|f_\theta|+1)} \right\rfloor \min(a, c\delta_\theta)\, \mathrm{card}\left([K_1]\setminus[K_2]\right)$$

where $\delta_\theta = \min\{\, \theta - f(i) : i \in E,\; f(i) < \theta \,\}$.
In both cases, the inequality of the lemma is satisfied. $\square$

12. The costs of the good transitions

We distinguish also two kinds of good transitions: those which create some new peak fitness individuals and those which increase the fitness.
We note by R the minimal number of transitions necessary to join two arbitrary points of E through the kernel α, i.e. R is the smallest integer satisfying

$$\forall i,j\in E\quad \exists\, r\leq R\quad \exists e_1,\cdots,e_{r+1}\in E\quad\text{such that}$$
$$e_1=i,\ e_{r+1}=j,\quad \forall k\in\{1,\cdots,r\}\quad \alpha(e_k,e_{k+1})>0.$$

Lemma 12.1. *(increasing of diversity in f^*)*
Let

$$V^* = \max_{i,j\in f^*}\ \min\Big\{ar+c\sum_{k=1}^{r+1} f(f^*)-f(e_k): e_1=i,\, e_{r+1}=j,\, \prod_{k=1}^{r}\alpha(e_k,e_{k+1})>0\Big\}.$$

We have

$$V^*\ \leq\ aR+c(R-1)\Delta\ \leq\ (a+c\Delta)|E|\ <\ \infty$$

and

$$\sup_{m\in\mathbb{N}^*}\ \max_{K_1\in\mathcal{K}_*,\,K_1\neq K^*}\ \min_{K_2\in\mathcal{K}_*,\,[K_1]\subsetneq[K_2]} V(K_1,K_2)\ \leq\ V^*.$$

(We recall that K^* is the unique attractor such that $[K^*]=f^*$.)

Proof. The inequalities $V^*\leq aR+c(R-1)\Delta\leq(a+c\Delta)|E|<\infty$ are straightforward. Let K_1 belong to $\mathcal{K}_*\setminus\{K^*\}$. There exists a point i in $[K_1]$, a point j in $f^*\setminus[K_1]$ and a path $e_1\to\cdots\to e_{r+1}$ in E joining i and j such that

$$\forall k\in\{1\cdots r\}\quad \alpha(e_k,e_{k+1})>0,\qquad \forall k\in\{2\cdots r\}\quad e_k\notin f^*,$$

$$ar+c\sum_{k=1}^{r+1} f(i)-f(e_k)\ \leq\ V^*.$$

Let p be a path in E^m defined by the extremal conditions

$$p^1\in K_1\quad\text{with}\quad p^1_m=i\quad\text{and}\quad p^{r+1}\in K_2\quad\text{where}\quad K_2\in\mathcal{K}_*,\quad [K_2]=[K_1]\cup\{j\}$$

(there exists a unique attractor K_2 satisfying $[K_2]=[K_1]\cup\{j\}$)
and having for intermediate populations

$$\forall k\in\{2\cdots r\}\quad \forall h\in\{1\cdots m-1\}\qquad p^k_h=p^1_h,\quad p^k_m=e_k.$$

In this path, the first $m-1$ components remain fixed and the last component follows the path $e_1\to\cdots\to e_{r+1}$. Clearly $[K_1]\subsetneq[K_2]$ and $V(K_1,K_2)\leq V(p)\leq V^*$. $\square$

Lemma 12.2. *(increasing of fitness)*
Let

$$V^+ \;=\; \max_{i\in E\setminus f^*}\; \min\left\{ ar + c\sum_{k=1}^{r} f(i)-f(e_k) : e_1=i,\; f(e_{r+1})>f(i),\; \prod_{k=1}^{r}\alpha(e_k,e_{k+1})>0 \right\}.$$

We have

$$V^+ \;\le\; aR + c(R-1)\Delta \;\le\; (a+c\Delta)|E| \;<\; \infty$$

and

$$\sup_{m\in\mathbb{N}^*} \quad \max_{K_1\in\mathcal{K}\setminus\mathcal{K}_*} \quad \min_{K_2\in\mathcal{K},\, f(K_2)>f(K_1)} V(K_1,K_2) \;\le\; V^+. \tag{13}$$

Proof. The proof is similar to the proof of the preceding lemma. Let K_1 belong to $\mathcal{K}\setminus\mathcal{K}_*$ and let i belong to $[K_1]$. By definition of V^+, there exists a path $e_1\to\cdots\to e_{r+1}$ in E such that $e_1=i$, $f(e_{r+1})>f(i)$ and

$$\forall k\in\{1\cdots r\}\qquad \alpha(e_k,e_{k+1})>0,\quad f(e_k)\le f(i),$$

$$ar + c\sum_{k=1}^{r} f(i)-f(e_k) \;\le\; V^+.$$

There exists a subdivision $k_1=1<k_2<\cdots<k_{s-1}<k_s=r+1$ of $\{1\cdots r+1\}$ such that $f(e_{k_1})=f(e_{k_2})=\cdots=f(e_{k_{s-1}})=f(i)$ and

$$\forall k\in\{1\cdots r+1\}\setminus\{\,k_1,\cdots,k_s\,\}\qquad f(e_k)<f(i).$$

For each ι, $1<\iota<s$, let K_ι be the unique attractor such that $[K_\iota]=[K_1]\cup\{\,e_{k_2},\cdots,e_{k_\iota}\,\}$. Let K_s be the attractor $\{\,(e_{r+1},\cdots,e_{r+1})\,\}$ (i.e. K_s contains only the population whose all components are equal to e_{r+1}). Clearly, the attractor K_s satisfies $f(K_s)>f(K_1)$. Let p be a path in E^m such that, for each ι in $\{1\cdots s\}$, p^{k_ι} belongs to K_ι and

$$\forall\iota\in\{1\cdots s-1\}\quad \forall k\in\{\,k_\iota+1,\cdots,k_{\iota+1}-1\,\}\quad p^k_m=e_k,\quad \forall h\in\{1\cdots m-1\}\quad p^k_h=p^{k_\iota}_h$$

(during the transition from K_ι to $K_{\iota+1}$, the first $m-1$ components remain fixed and the last component follows the mutation path $e_{k_\iota}\to\cdots\to e_{k_{\iota+1}}$). The path p belongs to D^m, joins the attractors K_1 and K_s, and its cost is less than V^+. It follows that $V(K_1,K_s)\le V^+$. $\square$

13. The asymptotic dynamics of the algorithm

The basic tools used to study the asymptotic dynamics of the process are the Freidlin-Wentzell graphs [7, chapter 6]. For completeness, we recall the basic definition.

Definition 13.1. (X–graphs)
Let H be a finite set and let X be a subset of H. An X–graph is a graph consisting of arrows $h_1 \to h_2$ ($h_1 \in H \setminus X$, $h_2 \in H$, $h_1 \neq h_2$) satisfying

(1) every point of $H \setminus X$ is the initial point of exactly one arrow;
(2) there are no closed cycles in the graph.

Condition (2) may be replaced by

(2′) for any point h_1 of $H \setminus X$ there exists a sequence of arrows leading from h_1 to some point h_2 of X.

The set of X–graphs is denoted by $G(X)$ and we will use the letter g to denote graphs.

We will consider graphs on the set of attractors $\mathcal{K}$.
For any graph g over $\mathcal{K}$, we define its cost by

$$V(g) = \sum_{(K_1 \to K_2) \in g} V(K_1, K_2).$$

If X and Y are two subsets of $\mathcal{K}$, we denote by $G_X(Y)$ the set of Y–graphs over $X \cup Y$. For instance, we have $G(X) = G_{\mathcal{K}}(X) = G_{\mathcal{K} \setminus X}(X)$.
We define

$$W_X(Y) = \min_{g \in G_X(Y)} \sum_{(K_1 \to K_2) \in g} V(K_1, K_2) = \min_{g \in G_X(Y)} V(g).$$

We note $G^*_X(Y)$ the set of graphs in $G_X(Y)$ which realize this minimum.
If g is a graph over $\mathcal{K}$, its restriction $g_{|\lambda}$ to the level λ is the graph

$$g_{|\lambda} = \{\, (K_1 \to K_2) \in g : K_1 \in \mathcal{K}_\lambda \}.$$

Theorem 13.1. *(sufficient condition to ensure $W^* = \{K^*\}$)*
If the inequality

$$\sum_{\lambda \in f(E \setminus f^*)} W_{\mathcal{K}_\lambda}(\mathcal{K}^+_\lambda) \; + \; W_{\mathcal{K}_*}(K^*) \; < \sum_{\lambda \in f(E \setminus f^*)} W_{\mathcal{K}_\lambda}(\mathcal{K}^-_\lambda \cup \mathcal{K}^+_\lambda) \; + \tag{14}$$
$$\min \Big(\min_{K \in \mathcal{K}_* \setminus \{K^*\}} W_{\mathcal{K}_*}(\{K\} \cup \mathcal{K}^-_*) \, , \;\; W_{\mathcal{K}_*}(\mathcal{K}^-_*) \; - \; \max_{K \in \mathcal{K}^-_*} \; \min_{K' ,\, f(K') \neq f(K)} V(K, K') \Big)$$

is satisfied, then the minimum of the virtual energy corresponds to the ideal attractor K^*, i.e. $W^* = \{K^*\}$, and therefore

$$\forall x \in E^m \qquad \lim_{l \to \infty} \; \lim_{n \to \infty} \; P\left([X^l_n] = f^* / X^l_0 = x\right) = 1.$$

Proof. Let g be a graph over $\mathcal{K}$. We decompose the sum $V(g)$ in the following way:

$$V(g) = \sum_{\lambda \in f(E)} \sum_{\substack{K_1 \in \mathcal{K}_\lambda \\ (K_1 \to K_2) \in g}} V(K_1, K_2) = \sum_{\lambda \in f(E)} \sum_{(K_1 \to K_2) \in g_{|\lambda}} V(K_1, K_2)$$

i.e.

$$V(g) = \sum_{\lambda \in f(E)} V(g_{|\lambda}).$$

Suppose now that g is in $G(K)$ for some K in $\mathcal{K}$. Put $\theta = f(K)$.
Then $g_{|\lambda}$ belongs to $G_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+)$ whenever $\lambda \neq \theta$ whence

$$V(g_{|\lambda}) \ \geq \ W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+). \tag{15}$$

We consider two cases, depending upon the value of θ.

- $\theta = f(f^*)$. In this case, $g_{|\theta}$ belongs to $G_{\mathcal{K}_*}(\{K\} \cup \mathcal{K}_*^-)$ so that

$$V(g_{|\theta}) \ \geq \ W_{\mathcal{K}_*}(\{K\} \cup \mathcal{K}_*^-). \tag{16}$$

Summing up inequalities (15) and (16) yields

$$V(g) \ \geq \sum_{\lambda \in f(E \setminus f^*)} W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+) \ + \ W_{\mathcal{K}_*}(\{K\} \cup \mathcal{K}_*^-).$$

Taking the minimum over all g in $G(K)$, we have

$$W(K) \ \geq \sum_{\lambda \in f(E \setminus f^*)} W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+) \ + \ W_{\mathcal{K}_*}(\{K\} \cup \mathcal{K}_*^-)$$

and taking again the minimum over all K in $\mathcal{K}_* \setminus \{K^*\}$, we obtain

$$\min_{K \in \mathcal{K}_* \setminus \{K^*\}} W(K) \ \geq \sum_{\lambda \in f(E \setminus f^*)} W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+) \ + \min_{K \in \mathcal{K}_* \setminus \{K^*\}} W_{\mathcal{K}_*}(\{K\} \cup \mathcal{K}_*^-). \tag{17}$$

- $\theta \neq f(f^*)$. Let K' be any element of $\mathcal{K}_\theta^- \cup \mathcal{K}_\theta^+$; then $g_{|\theta} \cup \{(K \to K')\}$ is a graph belonging to $G_{\mathcal{K}_\theta}(\mathcal{K}_\theta^- \cup \mathcal{K}_\theta^+)$ whence

$$V(g_{|\theta}) \ \geq \ W_{\mathcal{K}_\theta}(\mathcal{K}_\theta^- \cup \mathcal{K}_\theta^+) - V(K, K')$$

and this inequality being valid for all K' outside $\mathcal{K}_\theta$, we have

$$V(g_{|\theta}) \ \geq \ W_{\mathcal{K}_\theta}(\mathcal{K}_\theta^- \cup \mathcal{K}_\theta^+) - \min_{K',\, f(K') \neq f(K)} V(K, K'). \tag{18}$$

Summing up inequalities (15) and (18) yields

$$V(g) \;\geq\; \sum_{\lambda \in f(E)} W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+) \;-\; \min_{K',\, f(K') \neq f(K)} V(K,K').$$

Taking the minimum over all g in $G(K)$, we obtain

$$W(K) \;\geq\; \sum_{\lambda \in f(E)} W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+) \;-\; \min_{K',\, f(K') \neq f(K)} V(K,K')$$

and taking again the minimum over all K in $\mathcal{K} \setminus \mathcal{K}_* = \mathcal{K}_*^-$, we have

$$\min_{K \in \mathcal{K}_*^-} W(K) \;\geq\; \sum_{\lambda \in f(E)} W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+) \;-\; \max_{K \in \mathcal{K}_*^-} \quad \min_{K',\, f(K') \neq f(K)} V(K,K'). \tag{19}$$

Combining together inequalities (17) and (19), we see that

$$\min_{K \neq K^*} W(K) \;\geq\; \sum_{\lambda \in f(E \setminus f^*)} W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+) \;+ \tag{20}$$

$$\min \Big(\min_{K \in \mathcal{K}_* \setminus \{K^*\}} W_{\mathcal{K}_*}(\{K\} \cup \mathcal{K}_*^-)\;,\;\; W_{\mathcal{K}_*}(\mathcal{K}_*^-) \;-\; \max_{K \in \mathcal{K}_*^-} \quad \min_{K',\, f(K') \neq f(K)} V(K,K') \Big).$$

We build now a graph g of $G(K^*)$ which describes the most desirable dynamics of our algorithm. For each λ in $f(E \setminus f^*)$, we select a graph g_λ in the set $G^*_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^+)$. Let $g_{f(f^*)}$ be a graph of $G^*_{\mathcal{K}_*}(K^*)$. We define the graph g as the union of the graphs $(g_\lambda)_{\lambda \in f(E)}$:

$$(K_1 \to K_2) \in g \iff \exists\, \lambda \in f(E) \quad (K_1 \to K_2) \in g_\lambda.$$

Clearly g belongs to $G(K^*)$. Furthermore, we have by construction

$$V(g) = \sum_{\lambda \in f(E)} V(g_{|\lambda}) = \sum_{\lambda \in f(E \setminus f^*)} W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^+) + W_{\mathcal{K}_*}(K^*).$$

It follows that

$$W(K^*) \;\leq\; \sum_{\lambda \in f(E \setminus f^*)} W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^+) + W_{\mathcal{K}_*}(K^*). \tag{21}$$

Putting together inequalities (20) and (21) and the hypothesis (14) of the theorem, we see finally that $W^* = \{K^*\}$. $\square$

Of course, inequality (14) is strongly linked with the optimization problem: the quantities involved there are built with the fitness function f, the mutation kernel α, the crossover kernel β, the population size m and the parameters a, c. However, this inequality is of little practical interest: we will now derive stronger and simpler conditions to ensure $W^* = \{K^*\}$.

Corollary 13.1. *Suppose that*

$$W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^+) \;\leq\; W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+) \quad \text{for all} \quad \lambda \quad \text{in} \quad f(E \setminus f^*)\,, \tag{22}$$

$$W_{\mathcal{K}_*}(K^*) \;<\; W_{\mathcal{K}_*}(\mathcal{K}_*^-) \;-\; \max_{K \in \mathcal{K}_*^-} \quad \min_{K',\, f(K') \neq f(K)} V(K, K')\,, \tag{23}$$

$$W_{\mathcal{K}_*}(K^*) \;<\; \min_{K \in \mathcal{K}_* \setminus \{K^*\}} W_{\mathcal{K}_*}(\{K\} \cup \mathcal{K}_*^-))\,. \tag{24}$$

Then $W^* = \{K^*\}$.

Proof. Clearly, inequalities (22), (23), (24) imply inequality (14). □

We will show that these inequalities hold when the population size m is sufficiently large. First, the left–hand side member of (23) and (24) is bounded.

Proposition 13.1. *The quantity* $W_{\mathcal{K}_*}(K^*)$ *is bounded as a function of* m*:*

$$\sup_{m \in \mathbb{N}^*} W_{\mathcal{K}_*}(K^*) \;<\; |\mathcal{K}_*|\, V^* \;=\; 2^{|f^*|}\, V^* \;<\; \infty. \tag{25}$$

Proof. We build a graph g belonging to $G_{\mathcal{K}_*}(K^*)$ whose cost is less or equal than $|\mathcal{K}_*|\, V^*$. For each K in $\mathcal{K}_* \setminus \{K^*\}$, there exists by lemma 12.1 an attractor K' belonging to $\mathcal{K}_*$ such that $[K] \subsetneq [K']$ and $V(K, K') \leq V^*$. We consider successively each attractor K of $\mathcal{K}_* \setminus \{K^*\}$ and we add such an arrow $(K \to K')$ to the graph: the resulting graph g belongs to $G_{\mathcal{K}_*}(K^*)$ and its cost is less than $(|\mathcal{K}_*| - 1)\, V^*$. □

Theorem 13.2. *Let* m *be an integer such that*

$$\rho\left(\frac{m}{4} - \frac{\Lambda + 3}{2}\right) \;>\; \max_{K_1 \in \mathcal{K}_*^-} \quad \min_{K_2,\, f(K_2) > f(K_1)} V(K_1, K_2) \;+\; W_{\mathcal{K}_*}(K^*)\,, \tag{26}$$

$$\left\lfloor \frac{m}{2(|f^*| + 1)} \right\rfloor \min(a,\, c\delta^*) \;>\; W_{\mathcal{K}_*}(K^*)\,. \tag{27}$$

For this integer, we have $W^* = \{K^*\}$ *whence*

$$\forall x \in E^m \qquad \lim_{l \to \infty} \quad \lim_{n \to \infty} \quad P\left(X_n^l \in K^* / X_0^l = x\right) = 1.$$

Remark 1. Inequalities (13) and (25) show that such an integer m always exists: the quantities on the right–hand side are bounded by a constant independent of m. In particular, these conditions are fulfilled as soon as

$$\rho\left(\frac{m}{4} - \frac{\Lambda + 3}{2}\right) \;\geq\; 2^{|f^*|}\, V^* \;+\; V^+$$

$$\left\lfloor \frac{m}{2(|f^*| + 1)} \right\rfloor \min(a,\, c\delta^*) \;\geq\; 2^{|f^*|}\, V^*.$$

Remark 2. As a consequence, there exists a critical population size m^* depending on the function f, the kernels α, β and the parameters a, c such that

$$\forall m \geq m^* \qquad W^* = \{K^*\}.$$

Proof. We prove that if the integer m satisfies the inequalities (26) and (27), then the set of inequalities $(22), (23), (24)$ is fulfilled.
Let λ be in $f(E \setminus f^*)$ and let g be a graph belonging to $G_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+)$. This graph may contain a finite number of transitions from $\mathcal{K}_\lambda$ to $\mathcal{K}_\lambda^-$: $K_1 \to K_1', \cdots, K_r \to K_r'$. The first inequality (26) implies that for each $K_h, 1 \leq h \leq r$, there exists K_h'' in $\mathcal{K}_\lambda^+$ such that

$$\rho\left(\frac{m}{4} - \frac{\Lambda+3}{2}\right) \; > \; V(K_h, K_h'')$$

and lemma 11.3 yields $V(K_h, K_h'') < V(K_h, K_h')$.
Let $\tilde{g}$ be the graph obtained from g by replacing the r arrows $K_1 \to K_1', \cdots, K_r \to K_r'$ by $K_1 \to K_1'', \cdots, K_r \to K_r''$. The graph $\tilde{g}$ is in the set $G_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^+)$ and satisfies $V(\tilde{g}) \leq V(g)$. This construction being valid for any graph of $G_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+)$, we have

$$W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^+) \; \leq \; W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+)$$

for all λ in $f(E \setminus f^*)$ and the first condition (22) is proved.
Concerning the second condition (23), we notice that any graph g belonging to $G_{\mathcal{K}_*}(\mathcal{K}_*^-)$ contains at least an arrow starting from $\mathcal{K}_*$ and ending in $\mathcal{K}_*^-$. Thus

$$W_{\mathcal{K}_*}(\mathcal{K}_*^-) \; \geq \min_{K \in \mathcal{K}_*, K' \notin \mathcal{K}_*} V(K, K')$$

and lemma 11.3 implies

$$W_{\mathcal{K}_*}(\mathcal{K}_*^-) \; \geq \; \left(\frac{m}{4} - \frac{\Lambda+3}{2}\right)\rho$$

which, together with (26), yields (23).
Similarly, any graph belonging to $G_{\mathcal{K}_*}(\{K\} \cup \mathcal{K}_*^-)$ (where $K \neq K^*$) contains either a transition from K^* to an attractor K of f^* or to an attractor K of $\mathcal{K}_*^-$, so that

$$\min_{K \in \mathcal{K}_*, K \neq K^*} W_{\mathcal{K}_*}(\{K\} \cup \mathcal{K}_*^-) \; \geq \; \min \Big\{ V(K^*, K) : K \in \mathcal{K},\, K \neq K^* \Big\}$$

and lemmas 11.1 and 11.3 show that this quantity is greater than

$$\min\left(\left\lfloor \frac{m}{2(|f^*|+1)} \right\rfloor \min(a,\, c\delta^*),\, \left(\frac{m}{4} - \frac{\Lambda+3}{2}\right)\rho\right)$$

so that inequalities $(26), (27)$ imply the third and last condition (24). $\square$

Theorem 14.1. *(limit distribution on K^*)*
Suppose the population size m is large enough so that $W^* = \{K^*\}$.
Then the limit of the sequence of the stationary measures of the Markov chains (X_n^l), $l \in \mathbb{N}$, as l goes to infinity, is the uniform distribution over the ideal attractor K^*, that is

$$\forall x, y \in E^m \times K^* \qquad \lim_{l\to\infty} \lim_{n\to\infty} \quad P\left(X_n^l = y / X_0^l = x\right) = \frac{1}{|K^*|} = \frac{|f^*|^{2|f^*|\left\lfloor \frac{m}{2(|f^*|+1)}\right\rfloor - m}}{|f^*|!}$$

Proof. To prove this result, we could proceed as in [5] and use the representation formula of the stationary measure involving Freidlin–Wentzell graphs. However, this result is rather a consequence of the following fact: the virtual energy possesses a unique minimum at K^* and the unperturbed process (X_n^∞) admits a unique invariant probability measure on K^* (which is the uniform distribution over K^*). We are thus in a situation analogous to Theorem 4.2 of [7, chapter 6]. $\square$

Theorem 14.2. *(structure of the cycles)*
Let m be such that

$$\min\left(\left\lfloor \frac{m}{2(\Lambda+1)}\right\rfloor, \left(\frac{m}{4} - \frac{\Lambda+3}{2}\right)\right) \min(a,\, c\delta) \;>\; \max\left(V^+,\, V^*\right). \tag{28}$$

Then each cycle over the set of attractors $\mathcal{K}$ not containing the attractor K^* is reduced to one single attractor K.

Proof. We define an order $\prec$ on the set of attractors $\mathcal{K}$ by

$$\forall K_1, K_2 \in \mathcal{K} \quad K_1 \prec K_2 \quad \Longleftrightarrow \quad f(K_1) < f(K_2) \quad \text{or} \quad f(K_1) = f(K_2),\ [K_1] \subset [K_2].$$

Notice that this order is an extension of the previous order $\underset{\infty}{\prec}$ i.e.

$$\forall K_1, K_2 \in \mathcal{K} \qquad K_1 \underset{\infty}{\prec} K_2 \quad \Longrightarrow \quad K_1 \prec K_2.$$

Furthermore, the set $\mathcal{K}$ admits a greatest element which is precisely the attractor K^*:

$$\forall K \in \mathcal{K} \qquad K \prec K^*.$$

Now let π be a cycle over $\mathcal{K}$ not containing K^*. Suppose π is not reduced to one attractor. Let K be a maximal element of π for the order $\prec$. Then, for each K' in π distinct from K, we have either $f(K') < f(K)$ or $f(K') = f(K)$, $[K] \not\subset [K']$.

Proposition 11.1, corollary 11.1 and lemma 11.3 then imply

$$\forall K' \in \pi \setminus \{K\} \qquad V(K,K') \;\geq\; \min(a,\,c\delta)\,\min\left(\left\lfloor \frac{m}{2(\Lambda+1)} \right\rfloor,\,\frac{m}{4} - \frac{\Lambda+3}{2}\right).$$

Let K' be an element of $\pi \setminus \{K\}$ such that $V(K,K')$ is minimal. Then $A(K,K') = W(K) + V(K,K')$ (where A is the communication altitude) whence

$$A(K,K') \;\geq\; W(K) + \min(a,\,c\delta)\,\min\left(\left\lfloor \frac{m}{2(\Lambda+1)} \right\rfloor,\,\frac{m}{4} - \frac{\Lambda+3}{2}\right).$$

Yet lemmas 12.1 and 12.2 show that there exists an attractor K'' such that

$$K \prec K'', \quad K \neq K'', \quad V(K,K'') \;\leq\; \max(V^+,\,V^*)$$

whence

$$A(K,K'') \;\leq\; W(K) + \max(V^+,\,V^*).$$

Inequality (28) yields $A(K,K'') < A(K,K')$ so that necessarily the attractor K'' belongs to the cycle π: but this contradicts the maximality of K in the ordered set $(\pi, \prec)$. $\square$

Remark. Once more, this limiting structure is obtained as soon as the population size m is greater than a critical value, the other parameters being fixed. This structure of cycles is the most favorable one: the bad cycles, i.e. those which slow down the convergence, are reduced to one single attractor.

In the sequel, we suppose that the population size m is large so that the limit behavior described by theorems 14.1 and 14.2 is achieved. Once we know that the sequence of the stationary measures of the Markov chains (X_n^l), $l \in \mathbb{N}$, concentrates on the ideal attractor K^*, the remaining problem is to build a process with the same limiting law.
From now onwards, we consider the inhomogeneous algorithm. That is, the control parameter l is an increasing function of n and we deal with an inhomogeneous Markov chain $(X_n^{l(n)})_{n \in \mathbb{N}}$: we will suppress the superscript l in most notations.
The challenging problem is to adapt the sequence $(l(n))_{n \in \mathbb{N}}$ in order to have

$$\forall x \in E^m \qquad \lim_{n \to \infty} \quad P(X_n \in K^*/X_0 = x) \;=\; 1.$$

($l(n)$ should not increase too fast) and simultaneously, we wish to obtain the best rate of convergence ($l(n)$ should not increase too slowly). The answer to this now classical problem is given by Catoni and Trouvé's results [2,3,12,14].

15. The critical height H_1

For the definition and the properties of the height of exit H_e, we refer the reader to Trouvé's work [11,12,13,14]. The crucial constant for the convergence of the algorithm is the critical height H_1 defined by

$$H_1 \;=\; \sup\,\{\, H_e(\pi) : \pi \text{ cycle not containing } K^* \,\}\,.$$

Proposition 15.1. *The critical height H_1 is bounded as a function of m. More precisely, we have*

$$H_1 \;\leq\; \max\,(V^+,\, V^*).$$

Proof. To obtain an upper bound on H_1, it is enough to consider only the cycles over the set of attractors $\mathcal{K}$ [5, Proposition 17.1]. We also know that the attractors not containing K^* are reduced to one attractor. Therefore

$$H_1 \;\leq\; \max_{K\in\mathcal{K}\setminus\{K^*\}} \min\,\{\, V(K,K') : K'\in\mathcal{K},\ K'\neq K \,\}\,.$$

Lemmas 12.1 and 12.2 show that $H_1 \leq \max\,(V^+,\, V^*)$. $\square$

We now restate Trouvé's convergence result, which is an extension of a result by Hajek for the simulated annealing.

Theorem 15.1. *(Trouvé* [12, Theorem 3.23]*)*
Suppose m is large enough to have $W^*=\{K^*\}$.
For all increasing sequences $l(n)$ going to infinity, we have the equivalence

$$\sup_{x\in E^m} \quad P\,(X_n\notin K^*/X_0=x) \underset{n\to\infty}{\longrightarrow} 0 \qquad \Longleftrightarrow \qquad \sum_{n=0}^{\infty} l(n)^{-H_1}=\infty.$$

Furthermore, we may adapt the sequence $l(n)$ in order to be trapped in the basin of attraction of K^* after a finite number of transitions.

Theorem 15.2. *Define $H_e^* = H_e(\{\, x\in E^m : f^*\subset [x]\,\})$.*
Suppose m is large enough to have $W^*=\{K^*\}$ and $H_1 < H_e^*$.
For all increasing sequences $l(n)$, we have the equivalence

$$\begin{aligned} &\forall x\in E^m \qquad P\,(\exists N \quad \forall n\geq N \quad f^*\subset [X_n]\,/X_0=x)=1\\ &\Longleftrightarrow \qquad \sum_{n=0}^{\infty} l(n)^{-H_1}=\infty, \qquad \sum_{n=0}^{\infty} l(n)^{-H_e^*}<\infty. \end{aligned}$$

Remark. Proposition 1.33 of [13], proposition 11.1 and lemma 11.3 imply that H_e^* is greater than an affine increasing function of m, whereas H_1 is bounded. Henceforth, for m large enough, we have $H_1 < H_e^*$ and sequences $l(n)$ with the desired properties do exist.

16. The optimal rate of convergence

For the meaning and the properties of the optimal convergence exponent, we refer the reader to [2,3,12,13,14]. We restate now Trouvé's result for the optimal convergence rate, which generalizes Catoni's work.

Theorem 16.1. *There exist two strictly positive constants R_1 and R_2 such that for all n*

$$\frac{R_1}{n^{\alpha_{\mathrm{opt}}}} \quad\leq\quad \inf_{0\leq l(1)\leq\cdots\leq l(n)} \quad \max_{x\in E^m} \quad P\left(X_n\notin K^*/X_0=x\right) \quad\leq\quad \frac{R_2}{n^{\alpha_{\mathrm{opt}}}}.$$

Proposition 16.1. *The optimal convergence exponent α_{opt} is bounded between two affine strictly increasing functions of m, i.e. we have*

$$0 \;<\; \liminf_{m\to\infty}\frac{\alpha_{\mathrm{opt}}}{m} \;\leq\; \limsup_{m\to\infty}\frac{\alpha_{\mathrm{opt}}}{m} \;<\; \infty.$$

Proof. The definition of α_{opt} is [12, Definition 3.22]

$$\alpha_{\mathrm{opt}} = \min\left\{\,\frac{W(\pi)-W(K^*)}{H_e(\pi)} : \pi \text{ cycle not containing } K^*\,\right\}.$$

For a cycle π not containing K^*, the quantity $H_e(\pi)$ remains bounded whereas the virtual energy $W(\pi)$ is greater than an affine strictly increasing function of m (a graph belonging to $G_{\mathcal{K}}(\pi)$ contains necessarily a bad transition). Finally, the virtual energy of the ideal attractor $W(K^*)$ is bounded as a function of m. $\square$

Remark. The fact that the optimal convergence exponent α_{opt} increases linearly with m shows that our genetic algorithm is intrinsically parallel: it involves mostly local independent computations.

References

[1] E.H.L. Aarts–P.J.M. Van Laarhoven, *Simulated annealing: theory and applications*, D. Reidel, Dordrecht–Tokyo, 1987.

[2] O. Catoni, *Large deviations for annealing*, PhD Thesis, University Paris XI (1990).

[3] O. Catoni, *Rough large deviations estimates for simulated annealing. Application to exponential schedules*, The Annals of Probability **20** (1992), 1109–1146.

[4] R. Cerf, *Asymptotic convergence of genetic algorithms*, Preprint, submitted (1993).

[5] R. Cerf, *The dynamics of mutation-selection algorithms with large population sizes*, Preprint, submitted (1993).

[6] W. Feller, *An introduction to probability theory and its applications*, 3rd ed., vol. I, Wiley and Sons, New York–Singapore, 1950.

[7] M.I. Freidlin–A.D. Wentzell, *Random perturbations of dynamical systems*, Springer–Verlag, New York, 1984.

[8] D. Goldberg, *Genetic algorithms in search, optimization and machine learning*, Addison–Wesley, 1989.

[9] J.H. Holland, *Adaptation in natural and artificial systems*, The University of Michigan Press, Ann Arbor, 1975.

[10] C.R. Hwang–S.J. Sheu, *Singular perturbed Markov chains and exact behaviours of simulated annealing process*, Journal of Theoretical Probability **Vol. 5 No. 2** (1992), 223–249.

[11] A. Trouvé, *Convergence optimale pour les algorithmes de recuits généralisés*, C. R. Acad. Sci. Paris **t.315 Série I** (1992), 1197–1202.

[12] A. Trouvé, *Cycle decompositions and simulated annealing*, Preprint (1993).

[13] A. Trouvé, *Parallélisation massive du recuit simulé*, PhD Thesis, University Paris XI (1993).

[14] A. Trouvé, *Rough large deviation estimates for the optimal convergence speed exponent of generalized simulated annealing algorithms*, Preprint (1993).

Annexe A

BIBLIOGRAPHIE COMMENTÉE SUR LES ALGORITHMES GÉNÉTIQUES

BIBLIOGRAPHIE COMMENTÉE SUR LES ALGORITHMES GÉNÉTIQUES

Raphaël CERF

Octobre 1992

L'étude repose essentiellement sur une vingtaine d'articles datant des trois dernières années. Le but de cette bibliographie est de dégager les grandes directions de travail sur les algorithmes génétiques. Seuls des articles sur les problèmes théoriques posés par les algorithmes génétiques ont été retenus. Quatre grands thèmes sont ainsi apparus.

1. Les schemata
2. La transformée de Walsh
3. Le problème du codage
4. Les chaînes de Markov

1991 *Mathematics Subject Classification.* 60F10, 60J10, 92D15.
Key words and phrases. Freidlin–Wentzell theory, genetic algorithms, stochastic optimization.

1. Les schemata

La majeure partie des études sur les algorithmes génétiques repose sur la notion de schema (au pluriel, schemata).

Définition. *schema*
Soit $H = H_1 \cdots H_n$ un mot de longueur n sur l'alphabet $\{0, 1, \#\}$.
Une chaîne binaire $s = s_1 \cdots s_n$ est une instance du motif H si et seulement si

$$\forall p \in \{1 \cdots n\} \qquad s_p = H_p \quad \text{ou} \quad H_p = \#.$$

Un schema est l'ensemble des chaînes binaires de longueur n qui sont des instances d'un motif spécifié H. Le mot H est le nom du schema.

$$H = \left\{ s \in \{0,1\}^n : \forall p \in \{1 \cdots n\} \quad s_p = H_p \text{ ou } H_p = \# \right\}.$$

exemple: $n = 3$. $\quad \#01 = \{101, 001\} \qquad 1\#\# = \{100, 101, 110, 111\}$

Deux caractéristiques des schemata sont constamment utilisées: l'ordre et la taille.

Définition. *ordre d'un schema (order of a schema)*
L'ordre d'un schema H, noté $o(H)$, est le nombre de zéros et de un(s) dans H.

Définition. *positions spécifiées (defining positions)*
L'ensemble des positions spécifiées d'un schema H est l'ensemble des indices des zéros et des uns dans H:

$$\delta(H) = \{ p : H_p \neq \# \} \subset \{1 \cdots n\}.$$

Définition. *longueur d'un schema (length of a schema)*
La longueur d'un schema est la différence entre la dernière et la première position spécifiée de H:

$$l(H) = \max \delta(H) - \min \delta(H).$$

Considérons maintenant un algorithme génétique simple opérant sur une fonction f définie sur $\Omega = \{0,1\}^n$ à valeurs dans $\mathbb{R}_+^*$. Soit P une population de chaînes binaires de longueur n. Pour un schema H, l'utilité (utility) de H vis–à–vis de P est

$$f_P(H) = \frac{1}{|H \cap P|} \sum_{x \in P \cap H} f(x).$$

Remarque. La population P est un multiensemble, un élément de $P \cap H$ conserve la multiplicité qu'il avait dans P.

Notons $(P_t)_{t\in\mathbb{N}}$ les populations successives obtenues par l'algorithme génétique.
Le théorème des schemata est l'inégalité

$$E|H \cap P_{t+1}| \;\geq\; |H \cap P_t|\, \frac{f_{P_t}(H)}{f_{P_t}(\Omega)} \big(1 - \alpha(H,t) - \beta(H,t)\big)$$

où E est l'espérance et α, β approximent les probabilités qu'une instance de H soit détruite par crossing–over ou par mutation. Par exemple, si p_c est la probabilité d'avoir un crossing–over sur une chaîne donnée, et si p_m est la probabilité d'avoir une mutation, on a en linéarisant (i.e. en supposant p_c et p_m petits)

$$E|H \cap P_{t+1}| \;\geq\; |H \cap P_t|\, \frac{f_{P_t}(H)}{f_{P_t}(\Omega)} \big(1 - p_c \frac{l(H)}{n-1} - p_m o(H)\big).$$

La preuve n'utilise que des arguments combinatoires plutôt faciles et immédiats. C'est cette inégalité qui date de 1975, due à Holland, qui a ouvert la voie à toute une série d'études sur les schemata. L'inégalité indique que les schemata de valeur supérieure à la moyenne ($f_{P_t}(H) > f_{P_t}(\Omega)$) qui sont courts et d'ordre faible ($l(H)$ et $o(H)$ petits) vont proliférer dans la population: ces schemata sont appelés les "building blocks". Le succès ou non de l'algorithme génétique reposerait alors sur la possibilité de recombiner ces building blocks par crossing–over pour obtenir des individus de plus en plus performants. Si de plus on regarde l'algorithme génétique comme un opérateur sur l'espace des schemata, le nombre de schemata traités lorsque la population est de taille p est en moyenne de l'ordre de p^3. Ce phénomène est appelé dans la littérature le "parallélisme implicite". Voici un énoncé plus précis de ce résultat.

Théorème. *Considérons une population de $p = c_1 2^l = 2^{\beta l} (c_1 \geq 1, \beta \geq 1)$ individus choisis indépendamment et uniformément au hasard dans $\{0,1\}^n$; notons $\epsilon = 2l/n$.
Alors, pour $1 \leq \beta \leq 4/3$, le nombre moyen de schemata disjoints qui se propagent avec une probabilité inférieure à ϵ d'être détruit par application de l'opérateur de crossing–over est au moins*

$$\binom{2l}{[\beta l]} 2^{\beta l} (1 - e^{-l}).$$

Ceci donne une borne inférieure d'ordre p^3 pour $\beta = 1$ (i.e. $c_1 = 1$) et une borne inférieure d'ordre $p^{2.377}$ pour $\beta = 4/3$ ($c_1 = 2^{l/3}$). De plus, avec une probabilité supérieure à $1 - 2e^{-l}$, le nombre de schemata qui se propagent est au moins

$$\frac{1}{2} \binom{2l}{[\beta l]} 2^{\beta l} (1 - e^{-l}).$$

2. La transformée de Walsh

Les considérations sur les schemata suggèrent que les performances de l'algorithme génétique sont intimement liées à l'utilité relative des schemata. La transformée de Walsh s'avère être un outil commode pour calculer ces quantités.
Notons $\Omega = \{0,1\}^n$. La fonction de Walsh d'ordre j est

$$w_j(x) \;=\; (-1)^{j\cdot x} \qquad \text{où} \qquad j\cdot x \;=\; \sum_{i=1}^{n} j_i\, x_i.$$

Soit $f : \Omega \mapsto \mathbb{R}$. Le coefficient de Walsh d'ordre j de f est

$$\widehat{f}_j \;=\; 2^{-n/2} \sum_{x\in\Omega} f(x)\, w_j(x)$$

et on a la formule d'inversion (noter l'analogie avec l'analyse de Fourier)

$$f \;=\; 2^{-n/2} \sum_{j\in\Omega} \widehat{f}_j\, w_j$$

d'où pour un schema donné H

$$f_\Omega(H) \;=\; \frac{1}{|H|} \sum_{x\in H} f(x) \;=\; \frac{1}{|H|} 2^{-n/2} \sum_{j\in\Omega} \widehat{f}_j \sum_{x\in H} w_j(x).$$

Cette formule et quelques remarques simples permettent d'étudier les valeurs de f sur les schemata. Par exemple: posons $j = (j_1, \cdots, j_n)$.

- Si $j_k = 1$ et $H_k = \#$ alors $\widehat{f}_j$ n'intervient pas dans $f(H)$.
- Si $\sum_{k=1}^{n} j_k > o(H)$, alors $\widehat{f}_j$ n'intervient pas dans $f(H)$.

A partir de là, différentes notions de fonctions "trompeuses" ("GA-deceptive") ont été introduites puis construites à l'aide des fonctions de Walsh.

Définition. *schemata en concurrence (competing schemata)*
Deux schemata H et G sont concurrents s'ils ont même ensemble de positions spécifiées.

Définition. *fonction trompeuse (deceptive function)*
Soit f une fonction avec des maxima globaux aux points $\{x_1^*, \cdots, x_r^*\}$. Alors f est dite trompeuse d'ordre m s'il existe $x \notin \{x_1^*, \cdots, x_r^*\}$ tel que:
lorsque H et G sont deux schemata en concurrence d'ordre inférieur à m,

$$x \in H \quad \Longrightarrow \quad f(H) > f(G).$$

L'idée est que la recombinaison des building blocks dirige l'algorithme génétique vers des maxima locaux.

Définition. *fonction totalement trompeuse (fully deceptive function)*
Soit f une fonction avec un maximum global unique au point x^* et $\bar{x}$ le complément à 1 de x^* ("ou exclusif" avec $1 \cdots 1$). La fonction f est totalement trompeuse si, lorsque H et G sont deux schemata en concurrence d'ordre inférieur à m,

$$\bar{x} \in H \quad \Longrightarrow \quad f(H) > f(G).$$

Définition. *fonction extrêmement facile (fully easy function)*
Sous les mêmes hypothèses que ci–dessus, f est extrêmement facile si

$$x^* \in H \quad \Longrightarrow \quad f(H) > f(G).$$

Lemme. *Des fonctions complètement trompeuses existent dans tous les espaces $\{0,1\}^n$.*

Exemple. Si $o(x)$ désigne le nombre de 1 dans l'écriture binaire de x,

$$f(x) = \begin{cases} 1 - 1/2n & \text{si } o(x) = 0 \\ 1 - (1 + o(x))/n & \text{si } o(x) \neq 0, n \\ 1 & \text{si } o(x) = n \end{cases}$$

De nombreux exemples de fonctions trompeuses ont été ainsi construits.

3. Le codage

L'un des problèmes les plus délicats dans l'utilisation d'un algorithme génétique est le choix du codage. La théorie des schemata fournit quelques indications sur les caractéristiques souhaitables du codage: les schemata courts, de petits ordres, d'utilité supérieure à la moyenne, i.e. les building blocks, doivent pouvoir être recombinés afin de donner des schemata d'ordre supérieur co–adaptés. Ce problème se ramène à chercher des critères permettant de prédire si une fonction f sera ou non aisément traitée par l'algorithme génétique. La technique utilisée est essentiellement la transformée de Walsh (en relation avec la construction des fonctions trompeuses).
Notons également une autre tentative: "l'Epistasis Variance". Enfin des transformations élémentaires peuvent permettre de ramener un problème trompeur en un problème facile: par exemple des transformations affines du groupe dyadique. Un autre choix important réside dans la cardinalité de l'alphabet à employer. Quelques arguments qualitatifs proviennent là encore de la théorie des schemata.

Théorème. *Il y a $n_s = (k+1)^{\log_2 k}$ schemata par bit d'information pour des chaînes codées sur un alphabet à k lettres.*

Théorème. *Pour une quantité d'information fixée, les chaînes codées sur des alphabets de petite taille sont des instances d'un plus grand nombre de schemata que celles codées sur un alphabet de grande taille.*

4. Les chaînes de Markov

Une voie fort prometteuse réside dans la modélisation de l'algorithme génétique comme une chaîne de Markov. Soit Ω l'ensemble des mots à l lettres sur $\{0,1\}$ et $r = |\Omega| = 2^l$. Soit P une population d'éléments de Ω et $n = |P|$ la taille de la population. Soit Z une matrice de taille $r \times N$ dont les colonnes représentent les populations possibles de taille n. La ième colonne est le vecteur d'incidence de la ième population P_i.

Proposition. *Le nombre de populations possibles est* $N = \binom{n+r-1}{r-1}$.

Les populations successives obtenues par l'algorithme génétique forment une chaîne de Markov. La matrice de transition est gigantesque (elle est indexée par toutes les populations possibles) et les probabilités de transition sont fort compliquées. Peu de résultats ont pour l'instant été obtenus à l'aide de cette technique. Cependant la littérature concernant les chaînes de Markov est énorme et une étude fine de la matrice de transition peut permettre d'appliquer des résultats connus. De plus, cette approche fait ressortir certaines similitudes avec l'algorithme du recuit simulé. Lorsque les paramètres de l'algorithme génétique sont fixés, on obtient une chaîne de Markov homogène. Mais lorsqu'ils évoluent au cours du temps, on obtient une chaîne inhomogène. Il est plausible d'espérer que, comme dans le cas du recuit simulé, un choix adéquat d'une suite de paramètres (les probabilités de mutation et de crossing–over) fera apparaître une distribution limite stable pour les populations. Egalement, notons une tentative de construire un cadre général englobant simultanément les deux algorithmes.

5. Références

Cette étude est basée sur de nombreux preprints accumulés pendant deux années, de 1990 à 1992. Voici ceux qui ont été "sélectionnés".

5.1. Schemata analysis.

Schemata disruption – Vose, Liepins
Generalizing the notion of schema in GAs – Vose, 1991
Implicit parallelism in GAs – Bertoni, Dorigo

5.2. Walsh Transform.

GAs and Walsh functions, Part I and II – Goldberg, 1989
Construction of high–order deceptive functions using low–order Walsh coefficients – Goldberg, 1990
Sufficient conditions for deceptive and easy binary functions – Deb, Goldberg, 1992
Analyzing deception in trap functions – Deb, Goldberg, 1991
Deceptiveness and GA dynamic – Liepins, Vose, 1991
Polynomials, basis sets and deceptiveness in GAs – Liepins, Vose,1990
Walsh functions, schema variance and deception – Page, Richardson, 1992

5.3. Representational issue.

Epistasis variance: suitability of a representation to GAs – Davidor, 1990
Isomorphisms of GAs – Battle, Vose, 1991
Representational issues in genetic optimization – Liepins, Vose, 1990
Real–coded GAs, virtual alphabets and blocking – Goldberg, 1990

5.4. Markov chains.

Punctuated equilibria in genetic search – Vose, Liepins, 1991
Modeling GAs with Markov chains – Nix, Vose
Global convergence of GAs: a Markov chain analysis – Eiben, Aart, Van Hee, 1991
Finite Markov chain analysis of GAs – Goldberg, Segrest, 1987

5.5. Simulated annealing.

a note on Boltzmann tournament selection for GAs and population–oriented simulated annealing – Goldberg, 1990
a simulated annealing like convergence theory for the simple GA – Davis, Principe, 1991

5.6. Other topics.

Optimization and complexity in molecular biology and physics – Schuster
Expected time of random genetic drift of a population between stable phenotypic states – Lande, 1985
Global optimization of functions with the interval GA – Muselli, Ridella, 1992
GAs, noise and the sizing of populations – Goldberg, Deb, Clark, 1991
Tournament selection, niching and the preservation of diversity – Oei, Goldberg, Chang, 1991
Characterizing crossover in GAs – Liepins, Vose, 1991

6. Other works.

GAs as function optimizers – Bethke, 1980, PhD thesis
Adaptation in natural and artificial systems – Holland, 1975
Genetic algorithms and their applications – Proceedings of the second international conference on GAs, 1987

6.1. Mathematical population genetics.

Mathematics of genetic diversity – Kingman, 1980
Regenerative phenomena – Kingman, 1971
Mathematical population genetics – Ewens, 1978
Lecture notes in Biomathematics 15 – 17 – 19 – 31 – 36 – 53 – 70
Algèbres de Jordan et algèbres génétiques – Ouattara, 1988

Annexe B

CONVERGENCE ASYMPTOTIQUE D'UN ALGORITHME GÉNÉTIQUE

note soumise aux "Comptes Rendus de l'Académie des Sciences"

CONVERGENCE ASYMPTOTIQUE D'UN ALGORITHME GÉNÉTIQUE

(ASYMPTOTIC CONVERGENCE OF A GENETIC ALGORITHM)

Raphaël CERF

Ecole Normale Supérieure, Paris

Mars 1993

Résumé. Nous étudions un processus d'évolution markovien plus général que l'algorithme génétique simple classique. Ce processus est construit en perturbant aléatoirement un mécanisme de sélection. A l'aide de la théorie de Freidlin–Wentzell, nous étudions précisément la dynamique asymptotique du processus lorsque les perturbations tendent à disparaître. Nous montrons comment une délicate interaction entre les perturbations et la force sélective peut assurer la convergence vers les maxima globaux de la fonction score. Nous donnons différentes conditions de convergence dans les cas homogène et inhomogène.

Abstract. We study a markovian evolutionary process which encompasses the classical simple genetic algorithm. This process is built by perturbing randomly a selection scheme. Using the Freidlin–Wentzell theory, we carry out a precise study of the asymptotic dynamics of the process when the perturbations tend to disappear. We show how a delicate interaction between the perturbations and the selection pressure may force the convergence toward the global maxima of the fitness function. We derive several conditions to ensure convergence in homogeneous as well as inhomogeneous case.

1991 *Mathematics Subject Classification.* 60F10, 60J10, 92D15.

Key words and phrases. Freidlin–Wentzell theory, genetic algorithms, stochastic optimization.

This work was done during my last year as a student at the Ecole Normale Supérieure.
Professor Jean–Arcady Meyer first spoke to me of genetic algorithms.
I am deeply grateful to Professor Gérard Ben Arous for introducing me to the Freidlin–Wentzell theory.
I thank Professor Alain Berlinet for his precious advice. I thank the Referee for useful remarks.

Version française abrégée

Les algorithmes génétiques sont des procédures d'optimisation stochastique inspirées des mécanismes de l'évolution naturelle [3,4]. Nous donnons des résultats de convergence pour un modèle général d'algorithme génétique. Le texte de référence est [1].
Soit E un ensemble fini et f une application définie sur E à valeurs dans $\mathbb{R}^+$ dont on cherche à déterminer l'ensemble des maxima globaux f^*. Soit $m \in \mathbb{N}$ une taille de population. Nous considérons tout d'abord une chaîne de Markov (X_n^∞) d'espace d'états E^m construite ainsi: les éléments de la population (X_{n+1}^∞) sont choisis indépendamment suivant la loi uniforme parmi les meilleurs éléments de (X_n^∞) selon la fonction score f. Ce mécanisme est ensuite perturbé aléatoirement. L'intensité des perturbations est mesurée par un paramètre l et nous obtenons une suite de chaînes de Markov (X_n^l) en réalisant les étapes suivantes:

- un opérateur de mutation perturbe indépendamment les individus de la population,
- un opérateur de crossing–over perturbe les couples d'individus,
- la phase de sélection est relâchée et les éléments dont le score n'est pas maximal ont une probabilité non nulle de survivre.

Lorsque l croît vers l'infini, les perturbations aléatoires disparaissent progressivement et les probabilités de transition de la chaîne (X_n^l) convergent vers les probabilités de transition de la chaîne (X_n^∞). Le processus est alors attiré vers les populations d'équi–score (i.e. les populations dont tous les individus ont le même score). Des hypothèses adéquates sur la convergence des probabilités de transition (hypothèses H_p et H_q ci–après) permettent d'effectuer une analyse précise de la dynamique asymptotique du processus à l'aide de la théorie de Freidlin–Wentzell [2]. Nous étudions deux situations différentes. Dans le cas homogène, nous faisons d'abord tendre n vers l'infini puis l, i.e. nous étudions le comportement asymptotique de la mesure stationnaire de la chaîne $(X_n^l)_{n\geq 0}$. Nous donnons plusieurs conditions suffisantes pour que

$$\forall x \in E^m \qquad \lim_{l\to\infty} \quad \lim_{n\to\infty} \quad P\left([X_n^l] \subset f^*/X_0^l = x\right) = 1$$

où $[X_n^l]$ désigne l'ensemble des composantes de X_n^l. Nous montrons que ces conditions sont vérifiées lorsque la taille de la population est suffisamment élevée. Dans le cas inhomogène, nous faisons tendre simultanément n et l vers l'infini en prenant l comme une fonction croissante de n. Nous donnons des conditions sur la vitesse de croissance de la suite $l(n)$ pour assurer que la population (X_{T_n}) (où T_n est l'instant de la nième visite dans l'ensemble S des populations d'équi–score) soit piégée dans f^* après un nombre fini de transitions i.e.

$$\forall x \in E^m \qquad P\left(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset f^*/X_0 = x\right) = 1.$$

Lorsque la taille m de la population est suffisamment grande, il existe toujours des suites $l(n)$ de la forme n^θ qui conviennent; l'exposant θ doit se trouver dans un intervalle critique pour assurer la convergence. Nous spécialisons pour finir notre modèle au cas de l'algorithme génétique simple classique et obtenons ainsi les premiers résultats mathématiques significatifs sur la convergence des algorithmes génétiques.

We propose a general model of the simple genetic algorithm [3,4]. This model encompasses and enlightens the crucial elements necessary to ensure the desired convergence. It is analyzed by means of the Freidlin–Wentzell theory [2].
Let E be a finite set and f a positive function defined on E (which will be called the fitness function). We wish to find the set f^* of the global maxima of f. Let m be a population size. For an m–uple $x = (x_1, \cdots, x_m)$ of E^m we introduce the notations:

$$[x] = \{\, x_k : 1 \le k \le m \,\},$$
$$\hat{x} = \{\, x_k : 1 \le k \le m,\ f(x_k) = \max_{1 \le h \le m} f(x_h) \,\},$$
$$x(i) = \operatorname{card}\{\, k : 1 \le k \le m,\ x_k = i \,\} \quad \text{for} \quad i \in E.$$

The set A is the set of all m-uples whose m components are equal (such m–uples are called uniform populations). By S we denote the set of equi–fitness populations, that is, populations whose individuals have the same fitness:

$$S = \{\, x \in E^m : f(x_1) = \cdots = f(x_m) \,\}.$$

Let (X_n^∞) be the Markov chain with state space E^m having for transition probabilities

$$P\left(X_{n+1}^\infty = z / X_n^\infty = y\right) = \frac{1}{(\operatorname{card}\hat{y})^m} \prod_{k=1}^{m} 1_{\hat{y}}(z_k) y(z_k) = \frac{1}{(\operatorname{card}\hat{y})^m} \prod_{i \in [z]} 1_{\hat{y}}(i) y(i)^{z(i)}$$

that is, the individuals of the population X_{n+1}^∞ are chosen randomly (under the uniform distribution) and independently among the elements of $\hat{X}_n^\infty$ which are the best individuals of X_n^∞ according to the fitness function f. We build a sequence of Markov chains (X_n^l) by perturbing randomly (X_n^∞). The integer parameter l describes the intensity of the perturbations. The chain (X_n^l) is obtained through overlapping of several chain buildings:

$$X_n^l \xrightarrow{\text{mutation}} U_n^l \xrightarrow{\text{crossing–over}} V_n^l \xrightarrow{\text{selection}} X_{n+1}^l.$$

$X_n^l \longrightarrow U_n^l$**: mutation.** The mutation operator is modelled by random independent perturbations of the individuals of the population X_n^l. The transition probabilities from X_n^l to U_n^l are then given by

$$P\left(U_n^l = u / X_n^l = x\right) = p_l(x_1, u_1) \cdots p_l(x_m, u_m).$$

where p_l is a markovian kernel on the space E satisfying

Hypothesis H_p. There exist a function α defined on $E \times E$ with values in $\mathbb{R}^+$ which is an irreducible kernel i.e.

$$\forall i,j \in E \quad \exists i_0, i_1, \cdots, i_r \quad i_0 = i, \quad i_r = j, \quad \prod_{0 \le k \le r-1} \alpha(i_k, i_{k+1}) > 0$$

and a positive real number a such that p_l admits the development

$$\forall i,j \in E \quad \forall s \ge a \qquad p_l(i,j) = \begin{cases} \alpha(i,j) l^{-a} + o(l^{-s}) & \text{if } i \ne j \\ 1 - \alpha(i,j) l^{-a} + o(l^{-s}) & \text{if } i = j \end{cases}$$

$U_n^l \longrightarrow V_n^l$**: crossing–over.** The crossing–over operator is modelled by random independent perturbations of the couples formed by consecutive individuals of the population (V_n^l). The transition probabilities from U_n^l to V_n^l are

$$P\left(V_n^l = v / U_n^l = u\right) = \delta_m(u_m, v_m) \prod_{1 \le k \le m/2} q_l\left((u_{2k-1}, u_{2k}), (v_{2k-1}, v_{2k})\right)$$

where $\delta_m(i,j) = \delta(i,j)$ if m is odd (the last individual remains unchanged), $\delta_m(i,j) = 1$ if m is even, and q_l is a markovian kernel on the space $E \times E$ satisfying

Hypothesis H_q. There exist a function β defined on $(E \times E) \times (E \times E)$ with values in $\mathbb{R}^+$ and a positive real number b such that q_l admits the development

$$\forall (i_1, j_1) \in E \times E \quad \forall (i_2, j_2) \in E \times E \quad \forall s \ge b$$
$$q_l\left((i_1, j_1), (i_2, j_2)\right) = \begin{cases} \beta\left((i_1, j_1), (i_2, j_2)\right) l^{-b} + o(l^{-s}) & \text{if } (i_1, j_1) \ne (i_2, j_2) \\ 1 - \beta\left((i_1, j_1), (i_2, j_2)\right) l^{-b} + o(l^{-s}) & \text{if } (i_1, j_1) = (i_2, j_2) \end{cases}$$

$V_n^l \longrightarrow X_{n+1}^l$**: selection.** The transition probabilities from V_n^l to X_{n+1}^l are

$$P\left(X_{n+1}^l = x / V_n^l = v\right) = \prod_{i \in [x]} \left(\frac{v(i) \exp(cf(i) \ln l)}{\sum_{k=1}^m \exp(cf(v_k) \ln l)} \right)^{x(i)}$$

(where $c \in \mathbb{R}_*^+$ is a scaling parameter).
As l grows toward infinity the perturbations become smaller and smaller and with overwhelming probability, the chain (X_n^l) behaves as (X_n^∞) would do: it is attracted by the set S. Using the Freidlin–Wentzell theory [2], we carry out a precise study of the asymptotic dynamics of the chain when the perturbations disappear. Let T_n be the instant of the n–th visit to the set S and put $Z_n^l = X_{T_n}^l$. The chain $(Z_n^l)_{n \in \mathbb{N}}$ is the Markov chain induced by

the chain (X_n^l) on the set S. The study of this chain (Z_n^l) of successive visits to the set of attractors S is equivalent to the study of the whole process (X_n^l) [2, ch 6]. It follows from our assumptions that we have $P\left(Z_{n+1}^l = z / Z_n^l = y\right) \sim \widetilde{C}(y,z)\, l^{-\widetilde{V}(y,z)}$ as $l \to \infty$, where $\widetilde{C}$ and $\widetilde{V}$ are two non-negative kernels on $S \times S$. These estimates yield estimates of the invariant measure of the chain (Z_n^l) through the representation formula involving Freidlin–Wentzell graphs [2, ch 6]. The crucial quantity is the virtual energy W, defined for x in S by $W(x) = \min\{\sum_{(y\to z)\in g} \widetilde{V}(y,z) : g \in G\{x\}\}$ where $G\{x\}$ is the set of all x–graphs over S. We consider two situations. In the homogeneous case, we study the asymptotic behavior of the stationary measure of (X_n^l). As l goes to infinity, this stationary measure concentrates on the set $W^* = \left\{ x \in S : W(x) = \min_{y\in S} W(y) \right\}$. We show that $W^* \subset f^*$ for a sufficiently large population size so that

$$\forall x \in E^m \qquad \lim_{l\to\infty} \ \lim_{n\to\infty} \ P\left([X_n^l] \in f^* / X_0^l = x\right) = 1.$$

In the inhomogeneous case, where l is an increasing function of n, we derive several conditions on the population size and on the rate of increasing of the sequence $l(n)$, some necessary and some sufficient, to ensure the settling in f^* in finite time i.e.

$$\forall x \in E^m \qquad P\left(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset f^* / X_0 = x\right) = 1.$$

We finally specialize our general model in order to apply our results to the classical simple genetic algorithm [3]. We take $E = \{0,1\}^N$ for some integer N. A point i of E is a word of length N over the alphabet $\{0,1\}$ and is noted $i = i_1 \cdots i_N$ where $i_k \in \{0,1\}$. The Hamming distance $H(i,j)$ between two points i,j of E is the number of letters where i and j differ:

$$H(i,j) = \mathrm{card}\,\{\, k : 1 \le k \le m,\ i_k \neq j_k \,\}.$$

The mutation kernel p_l and its associated kernel α are

$$p_l(i,j) = \begin{cases} 0 & \text{if } H(i,j) > 1 \\ l^{-a} & \text{if } H(i,j) = 1 \\ 1 - Nl^{-a} & \text{if } H(i,j) = 0 \end{cases} \qquad \alpha(i,j) = \begin{cases} 0 & \text{if } H(i,j) > 1 \\ 1 & \text{if } H(i,j) = 1 \\ N & \text{if } H(i,j) = 0 \end{cases}$$

Two arbitrary points of E are connected through the kernel α in at most N transitions so that the irreducibility condition is satisfied.
In order to build the crossing–over operator, we define $N-1$ cutting operators $(T_k)_{1 \le k < N}$; T_k maps $E \times E$ onto $E \times E$ and for i,j in E, we put $T_k(i,j) = (i',j')$ where

$$i' = i_1 \cdots i_k j_{k+1} \cdots j_N, \qquad j' = j_1 \cdots j_k i_{k+1} \cdots i_N.$$

The kernel β is then defined to be

$$\beta((i,j),(i',j')) = \mathrm{card}\,\{\, k : 1 \le k \le N-1,\ T_k(i,j) = (i',j') \,\}$$

and the kernel q_l is defined by

$$q_l((i,j),(i',j')) = \beta((i,j),(i',j'))\, l^{-b} \qquad \text{if } (i,j) \neq (i',j'),$$
$$q_l((i,j),(i,j)) = 1 - \sum_{(i',j') \neq (i,j)} \beta((i,j),(i',j'))\, l^{-b}.$$

Both hypothesis H_p and H_q are clearly fulfilled, as well as a stronger assumption on the asymptotic developments needed to deal with the inhomogeneous case (see [1] for more details). We define

$$\Delta^\circ = \max_{i \in E \setminus f^*} \min \left\{ \max_{0 \le k < r} (f(i) - f(i^k)) : i^0 = i,\ r \le N,\ f(i^r) > f(i),\ \prod_{0 \le k < r} H(i^k, i^{k+1}) > 0 \right\}$$

$$\Delta^* = \max_{i,j \in f^*} \min \left\{ \max_{0 \le k < r} (f(i) - f(i^k)) : i^0 = i,\ r \le N,\ i^r = j,\ \prod_{0 \le k < r} H(i^k, i^{k+1}) > 0 \right\}$$

and let $\Delta^\circledast = \max(\Delta^\circ, \Delta^*)$, $\quad \delta = \min\{\, |f(i) - f(j)| : i,j \in E,\ f(i) \neq f(j) \,\}$.
Within this framework, all convergence results proved for the general model do apply. We restate only the rough results, which take a more precise form in this context.

Theorem. *(Convergence of the homogeneous simple genetic algorithm)*
If

$$m \quad > \quad \frac{aN + c(N-1)\Delta^{\circledast}}{\min(a, b/2, c\delta)}$$

then

$$\forall x \in E^m \qquad \lim_{l\to\infty} \quad \lim_{n\to\infty} \quad P\left([X_n^l] \subset f^*/X_0^l = x\right) \;=\; 1.$$

Recall that the convergence exponent of the increasing sequence $l(n)$ is the unique real number λ such that the series $\sum_{n\geq 0} l(n)^{-\theta}$ converges for $\theta > \lambda$ and diverges for $\theta < \lambda$.

Theorem. *(Convergence of the inhomogeneous simple genetic algorithm)*
1) For the chain (X_{T_n}) to be trapped in f^* after a finite number of transitions, i.e. to have

$$\forall x \in E^m \qquad P\,(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset f^*/X_0 = x) = 1$$

the convergence exponent of the sequence $l(n)$ must be a positive real number; that is, there must exist two positive real numbers θ_1 and θ_2 such that

$$\sum_{n\geq 0} l(n)^{-\theta_1} = \infty \qquad \text{and} \qquad \sum_{n\geq 0} l(n)^{-\theta_2} < \infty.$$

2) If the convergence exponent λ of the sequence $l(n)$ and the population size m satisfy the inequalities

$$aN + c(N-1)\Delta^{\circledast} \;<\; \lambda \;<\; \min(a, b/2, c\delta)\, m$$

then, with probability one, the chain (X_{T_n}) is trapped in f^* after a finite number of transitions, i.e.,

$$\forall x \in E^m \qquad P\,(\exists N \quad \forall n \geq N \quad [X_{T_n}] \subset f^*/X_0 = x) = 1.$$

3) Suppose there exists a real number t strictly greater than one such that for all r in $\mathbb{N}$, the sequences $l([tn]+r)$ and $l(n)$ are logarithmically equivalent. If the convergence exponent λ of the sequence $l(n)$ and the population size m satisfy the inequalities

$$aN + c(N-1)\Delta^{\circledast} \;<\; \min(a, b/2, c\delta)\, m \;\leq\; \lambda$$

then

$$\forall x \in E^m \qquad \lim_{n\to\infty} \quad P\,([X_n] \subset f^*/X_0 = x) \;=\; 1.$$

References

[1] R. Cerf, *Asymptotic convergence of genetic algorithms*, submitted (1993).
[2] M.I. Freidlin–A.D. Wentzell, *Random perturbations of dynamical systems*, Springer–Verlag, New York, 1984.
[3] D. Goldberg, *Genetic algorithms in search, optimization and machine learning*, Addison–Wesley, 1989.
[4] J.H. Holland, *Adaptation in natural and artificial systems*, The University of Michigan Press, Ann Arbor.

Annexe C

UN NOUVEL ALGORITHME GÉNÉTIQUE

note soumise aux "Comptes Rendus de l'Académie des Sciences"

UN NOUVEL ALGORITHME GÉNÉTIQUE

(A NEW GENETIC ALGORITHM)

RAPHAËL CERF

Université de Montpellier

November 1993

Résumé. Voici un nouvel algorithme génétique. Il est construit en perturbant aléatoirement un mécanisme à deux opérateurs (crossing–over et sélection). Un nouvelle procédure de sélection est utilisée. Les petites perturbations aléatoires sont de deux sortes: des perturbations locales des individus (mutations) et un relâchement de la force de sélection. Lorsque la taille de la population dépasse une valeur critique qui dépend fortement du problème d'optimisation, leur délicate interaction asymptotique force la convergence (éventuellement en temps fini) de la population vers l'attracteur idéal dont les populations contiennent tous les maxima globaux de la fonction d'adaptation.

Abstract. Here is a new genetic algorithm. It is built by perturbing randomly a two operators crossover–selection scheme. A new selection mechanism is used. The random vanishing perturbations are twofold: local perturbations of the individuals (mutations) and loosening of the selection pressure. When the population size is greater than a critical value which depends strongly on the optimization problem, their delicate asymptotic interaction ensures the convergence (possibly in finite time) of the population toward the ideal attractor whose populations contain all the maxima of the fitness function.

1991 *Mathematics Subject Classification.* 60F10, 60J10, 92D15.

Key words and phrases. Freidlin–Wentzell theory, genetic algorithms, stochastic optimization.

I thank Alain Trouvé, Gérard Ben Arous and Alain Berlinet for useful discussions.

Version française abrégée

Les algorithmes génétiques sont des procédures d'optimisation stochastique inspirées des mécanismes de l'évolution naturelle [3,4,7,8]. Nous présentons un nouvel algorithme génétique qui trouve simultanément tous les maxima globaux. La référence est [5].

Soit E un ensemble fini et f une application définie sur E à valeurs dans $\mathbb{R}_+^*$ dont on cherche à déterminer l'ensemble des maxima globaux f^*. Soit $m \in \mathbb{N}$ une taille de population. Nous considérons tout d'abord une chaîne de Markov (X_n^∞) d'espace d'états E^m construite à l'aide de deux opérateurs: le crossing–over, auquel nous imposons des conditions d'origine biologique, et un nouveau mécanisme de sélection qui a l'avantage décisif de préserver la diversité des individus dans la population. La chaîne (X_n^∞) est instantanément piégée dans des sous-ensembles particuliers de E^m que nous appelons les attracteurs. Lorsque m est assez grand, il existe un unique attracteur idéal K^* dont les populations contiennent tous les maxima globaux de la fonction f. Ce processus (X_n^∞) est ensuite perturbé aléatoirement de deux façons: des perturbations locales et indépendantes des individus (mutations) et un relâchement de la force de sélection.

Le mécanisme de sélection perturbé fonctionne de la manière suivante: pour construire la nouvelle population, nous commençons d'abord par regrouper les individus de l'ancienne population par niveau d'adaptation. Nous réordonnons aléatoirement et uniformément chacun de ces groupes. Le k–ième individu de la nouvelle population est choisi ainsi: un niveau d'adaptation est tiré parmi les valeurs observées dans l'ancienne population à l'aide d'une loi de probabilité biaisée vers les hautes valeurs. Supposons que le groupe précédemment associé à la valeur choisie contiennent r individus. Nous divisons l'ensemble d'indices $\{1 \cdots m\}$ en $r+1$ paquets. Si k est dans le h–ième paquet $(1 \leq h \leq r)$, nous choisissons le h–ième élément du groupe. Si k est dans le $(r+1)$–ième paquet, nous choisissons un individu uniformément au hasard dans le groupe.

L'intensité des perturbations est contrôlée par un paramètre l et nous obtenons une suite de chaînes de Markov $(X_n^l), l \in \mathbb{N}$. Le développement asymptotique des probabilités de transition (lorsque $l \to \infty$) donne $P\left(X_{n+1}^l = v / X_n^l = u\right) \sim q_1(u,v) l^{-V_1(u,v)}$ où $q_1 \in \mathbb{R}_+$ et V_1 est une fonction de coût de communication. Nous tombons dans la classe des algorithmes de recuit simulé généralisés étudiés par Trouvé [9,10,11]. La suite des mesures stationnaires des chaînes (X_n^l) se concentre sur l'ensemble des minima W^* de l'énergie virtuelle W. Nous étudions les coûts des transitions entre les attracteurs. Lorsque la taille de la population m dépasse une valeur critique m^* (qui dépend de toutes les caractéristiques du problème d'optimisation), l'ensemble W^* coïncide avec l'attracteur idéal K^*, de sorte que

$$\forall m \geq m^* \quad \forall x \in E^m \qquad \lim_{l \to \infty} \ \lim_{n \to \infty} \quad P\left([X_n^l] = f^* / X_0^l = x\right) = 1.$$

Dans le cas inhomogène où $l = l(n)$ est une suite croissante de n, nous utilisons les résultats de Catoni et Trouvé, qui permettent de décrire très précisément la trajectoire du processus en fonction de la vitesse de croissance de la suite $l(n)$. Il est possible d'ajuster les paramètres de sorte que la population soit piégée dans le bassin d'attraction de l'attracteur idéal K^* après un nombre fini de transitions.

We present a new genetic algorithm which finds simultaneously all the global maxima of the fitness function. We consider a finite space of states E and a positive function f (which will be called the fitness function) defined on E. We wish to find the set f^* of the global maxima of f. For a subset K of E^m we introduce the notations:

$$[K] = \{\, x_k : 1 \le k \le m,\ x = (x_1, \cdots, x_m) \in K \,\},$$
$$\widehat{f}(K) = \max \{\, f(x_k) : 1 \le k \le m,\ x = (x_1, \cdots, x_m) \in K \,\},$$
$$\widehat{K} = \{\, x_k : 1 \le k \le m,\ x = (x_1, \cdots, x_m) \in K,\ f(x_k) = \widehat{f}(K) \,\}.$$

If x is an element of E^m, we note $[x] = [\{x\}], \widehat{x} = \widehat{\{x\}}$. If F is a subset of E, $f(F)$ (or fF) denotes the set $\{\, f(i) : i \in F \,\}$. For $\lambda \in \mathbb{R}_+^*$, we put $f_\lambda = f^{-1}(\{\lambda\})$. For x in E^m, we let $x^\lambda = [x] \cap f_\lambda$ and we number the elements of the set x^λ: $x^\lambda = \{\, x_1^\lambda, \cdots, x_{|x^\lambda|}^\lambda \,\}$ in the order they appear in the sequence $(x_k^\lambda, 1 \le k \le |x^\lambda|)$. For the special case $\lambda = \widehat{f}(x)$, we use the notation $\widehat{x} = \{\, \widehat{x}_1, \cdots, \widehat{x}_{|\widehat{x}|} \,\}$. Finally, for x in E^m, we denote by $\lambda_1^x, \cdots, \lambda_{|f[x]|}^x$ the $|f[x]|$ elements of the set $f[x]$, where again the indexation respects the order of appearance of the elements of $f[x]$ in the sequence $(f(x_k), 1 \le k \le m)$.

The unperturbed Process (X_n^∞)

When there is no random perturbation, the process under study is a Markov chain (X_n^∞) with state space E^m. The transition from X_n^∞ to X_{n+1}^∞ is decomposed in two stages:

$$X_n^\infty \xrightarrow{\text{crossover}} Z_n^\infty \xrightarrow{\text{selection}} X_{n+1}^\infty.$$

$X_n^\infty \longrightarrow Z_n^\infty$**: crossover.** The phenomenon of crossover is modelled as a random operation on the couples formed by consecutive individuals of the population X_n^∞. Let β be a markovian kernel on the space $E \times E$. The quantity $\beta((i_1, j_1), (i_2, j_2))$ is interpreted as the probability of producing the pair of individuals (i_2, j_2) by performing a crossover on the couple (i_1, j_1). We impose three conditions on β, which are of biological relevance:

$$\begin{aligned} \forall i \in E \qquad & \beta((i,i),(i,i)) = 1, \\ \forall i, j \in E \qquad & \beta((i,j),(i,j)) > 0, \\ \forall i_1, j_1, i_2, j_2 \in E \qquad & \beta((i_1,j_1),(i_2,j_2)) = \beta((j_1,i_1),(j_2,i_2)). \end{aligned}$$

Amazingly enough, the first two conditions (especially the first one) are crucial for our algorithm to work. The transition probabilities from X_n^∞ to Z_n^∞ are

$$P\left(Z_n^\infty = z / X_n^\infty = x\right) = \delta_m(x_m, z_m) \prod_{1 \le k \le m/2} \beta\left((x_{2k-1}, x_{2k}), (z_{2k-1}, z_{2k})\right)$$

where $\delta_m(i,j) = \delta(i,j)$ if m is odd (the last individual of the population has no mating partner and remains unchanged after crossover) and $\delta_m(i,j) = 1$ if m is even.

$Z_n^\infty \longrightarrow X_{n+1}^\infty$**: selection.** We propose here an enhanced version of the selection mechanism used in the previous models [4,5] which has the decisive advantage of preserving the diversity of the individuals present in the population.
We define a triangular array of integers $\tau(k,h),\ 0 \leq k \leq h+1,\ 1 \leq h \leq m,$ by

$$\begin{aligned} &\forall h \in \{1 \cdots m\} && \tau(0,h)=1, \quad \tau(h+1,h)=m, \\ &\forall k,h \quad 1 \leq k \leq h \leq m && \tau(k,h) = 2k\left\lfloor \frac{m}{2(h+1)} \right\rfloor + 1. \end{aligned}$$

Suppose $Z_n^\infty = z$. To build the population $X_{n+1}^\infty = x$, we select randomly with the uniform distribution a permutation σ of $\mathfrak{S}(|\widehat{z}|)$. Using the subdivision $\tau(\cdot,|\widehat{z}|)$ of $\{1\cdots m\}$, we divide the set of indices $\{1\cdots m\}$ in $|\widehat{z}|+1$ parts. The components of the r–th part (for $1 \leq r \leq |\widehat{z}|$) are set equal to $\widehat{z}_{\sigma(r)}$, that is for the indices k such that $\tau(r-1,|\widehat{z}|) \leq k < \tau(r,|\widehat{z}|)$ we put $x_k = \widehat{z}_{\sigma(r)}$. The components of the $(|\widehat{z}|+1)$–th part (i.e. for the indices k such that $\tau(|\widehat{z}|,|\widehat{z}|) \leq k \leq m$) are chosen independently and uniformly on the set $\widehat{z}$.
The attractors of the process (X_n^∞) are the sets of populations K such that

$i)$ $[K] = \widehat{K}$

$ii)$ a population x of E^m belongs to K if and only if

- $\forall r \in \{1\cdots|\widehat{K}|\} \quad \forall k,h \qquad \tau(r-1,|\widehat{K}|) \leq k,h < \tau(r,|\widehat{K}|) \quad \Longrightarrow \quad x_k = x_h,$
- $\{\, x_k :\ 1 \leq k < \tau(|\widehat{K}|,|\widehat{K}|)\,\} = [K],$
- $\{\, x_k : \tau(|\widehat{K}|,|\widehat{K}|) \leq k \leq m \,\} \subset [K].$

Notice that the bracket operator [ ] provides a one to one correspondence between the set of all the attractors and the set $\{\, F \subset E : |F| \leq m,\ \widehat{F} = F \,\}$. We denote by K^* the unique attractor (it exists for $m \geq |f^*|$) such that $[K^*] = f^*$. The transition mechanism implies that the process (X_n^∞) is instantaneously absorbed in the set of the populations which belong to attractors.

The perturbed Process (X_n^l)

The previous Markov chain (X_n^∞) is randomly perturbed by two distinct mechanisms. The first one acts directly upon the population and mimics the phenomenon of mutation. The second one consists in loosening the selection of the individuals. The intensity of the perturbations is governed by an integer parameter l: as l goes to infinity, the perturbations progressively disappear. The transition mechanism of the perturbed Markov chain (X_n^l) is decomposed in three stages:

$$X_n^l \xrightarrow{\text{mutation}} Y_n^l \xrightarrow{\text{crossover}} Z_n^l \xrightarrow{\text{selection}} X_{n+1}^l.$$

$X_n^l \longrightarrow Y_n^l$: **mutation.** The mutations are modelled by random independent perturbations of the individuals of the population X_n^l. Let α be an irreducible Markov kernel on the space E. Let α_l be the irreducible Markov kernel on E defined by

$$\alpha_l(i,j) = \begin{cases} \alpha(i,j)l^{-a} & \text{if } i \neq j \\ 1 - \sum_{e\neq i}\alpha(i,e)l^{-a} & \text{if } i = j \end{cases} \qquad \text{(where } a \in \mathbb{R}_+^*\text{)}$$

The transition probabilities from X_n^l to Y_n^l are given by

$$P\left(Y_n^l = y / X_n^l = x\right) = \alpha_l(x_1, y_1)\cdots\alpha_l(x_m, y_m).$$

$Y_n^l \longrightarrow Z_n^l$: **crossover.** The crossover is not perturbed in any way: this stage is exactly the same as the passage from X_n^∞ to Z_n^∞.

$Z_n^l \longrightarrow X_{n+1}^l$: **selection.** The selection mechanism of the chain (X_n^∞) is perturbed by loosening randomly the selection pressure (individuals below peak fitness may survive).
Suppose $Z_n^l = z$ and we wish to build the vector $X_{n+1}^l = x$.
Let us summarize briefly and without formula this mechanism.
We first traverse $f[z]$; with each element λ of this set, we associate a sequence $\psi_1, \cdots, \psi_{n_\lambda}$ obtained by reordering randomly (all orders being equally probable) the set $f^{-1}(\{\lambda\}) \cap f[z]$.
The population x is then built in the following way: for each component $x_k, 1 \leq k \leq m$, we draw a value λ under a distribution probability on the set $f[z]$ which is biased toward the high values. With this value λ, we had previously associated a sequence $\psi_1, \cdots, \psi_{n_\lambda}$.
We divide the set $\{1\cdots m\}$ in $n_\lambda + 1$ parts. If the index k under consideration belongs to the r–th part, where $1 \leq r \leq n_\lambda$, we set $x_k = \psi_r$. If k belongs to the $(n_\lambda+1)$–th part, we choose x_k randomly and uniformly over the set $\{\psi_1, \cdots, \psi_{n_\lambda}\}$.
We now describe precisely the selection mechanism.
For each h in $\{1\cdots|f[z]|\}$, we select independently and randomly with the uniform distribution a permutation σ^h belonging to $\mathfrak{S}(|z^{\lambda_h}|) = \mathfrak{S}(|[z] \cap f_{\lambda_h}|)$. The law of each component $x_k\,(1 \leq k \leq m)$ of x is defined as follows: we select randomly a value λ in the set $f[z]$ with the distribution

$$\forall h \in \{1\cdots|f[z]|\} \qquad P(\lambda = \lambda_h) = \frac{\exp(c\lambda_h \ln l)}{\sum_{r=1}^{|f[z]|}\exp(c\lambda_r \ln l)} \qquad \text{(where } c \in \mathbb{R}_+^*\text{)}$$

The value of x_k is then chosen in the set z^λ according to the value of the index k:

- if $\tau(|z^\lambda|, |z^\lambda|) \leq k \leq \tau(|z^\lambda|+1, |z^\lambda|) = m$, then x_k is chosen at random with the uniform distribution over the set $z^\lambda = [z] \cap f_\lambda$ (k lies in the last part of the set of indices).
- if there exists an index r in $\{1\cdots|z^\lambda|\}$ such that $\tau(r-1, |z^\lambda|) \leq k < \tau(r, |z^\lambda|)$, then we put $x_k = z^\lambda_{\sigma^h(r)}$, where h is the unique integer in $\{1\cdots|z^\lambda|\}$ satisfying $\lambda = \lambda_h$ (the index k lies in the r–th part of the set of indices, where $1 \leq r \leq |z^\lambda|$).

We have

$$\forall y, z \in E^m \qquad \lim_{l\to\infty} P\left(X_{n+1}^l = z / X_n^l = y\right) = P\left(X_{n+1}^\infty = z / X_n^\infty = y\right)$$

so that the process (X_n^l) is a perturbation of the process (X_n^∞).

The asymptotic dynamics of the algorithm

We have the estimates $P\left(X_{n+1}^l = v / X_n^l = u\right) \sim q_1(u,v) l^{-V_1(u,v)}$ where q_1 is a non–negative kernel and V_1 is a communication cost corresponding to a discrete variational problem. We are thus in the framework of the generalized simulated annealing studied by Trouvé [9,10,11]. The crucial quantity is the virtual energy W, defined for x in E^m by $W(x) = \min\{\sum_{(y\to z)\in g} V_1(y,z) : g \in G(x)\}$ where $G(x)$ is the set of all x–graphs over E^m. As l goes to infinity, the sequence of the stationary measures of the chains $(X_n^l), l \in \mathbb{N}$, concentrates on the set W^* of the global minima of the virtual energy W.

Theorem. *There exists a critical population size m^* depending on the function f, the kernels α, β and the parameters a, c such that $W^* = K^*$ when m is greater than m^*, whence*

$$\forall m \geq m^* \quad \forall x \in E^m \qquad \lim_{l\to\infty} \quad \lim_{n\to\infty} \quad P\left([X_n^l] = f^* / X_0^l = x\right) = 1.$$

Furthermore, the limit distribution is the uniform distribution over the ideal attractor K^*.

We give various conditions (which depend strongly on the optimization problem) to ensure $W^* = K^*$ and upper bounds on m^*. We finally apply Catoni and Trouvé's tools for the inhomogeneous case [1,2,9,10,11] where $l = l(n)$ is an increasing function of n.

Theorem. *(Trouvé* [9, Theorem 3.23]*)*
Suppose m is large enough to have $W^* = K^*$. There exists a positive constant H_1 such that, for all increasing sequences $l(n)$ going to infinity, we have the equivalence

$$\sup_{x\in E^m} \quad P\left(X_n \notin K^* / X_0 = x\right) \underset{n\to\infty}{\longrightarrow} 0 \qquad \Longleftrightarrow \qquad \sum_{n=0}^{\infty} l(n)^{-H_1} = \infty.$$

Proposition. *The critical height H_1 is bounded as a function of m.*

We may adapt the sequence $l(n)$ in order to be trapped in the basin of attraction of K^*. For the definition and the properties of the height of exit H_e, see [9,10].

Theorem. *Define $H_e^* = H_e(\{\, x \in E^m : f^* \subset [x]\,\})$.*
Suppose m is large enough to have $W^* = \{K^*\}$ and $H_1 < H_e^*$.
For all increasing sequences $l(n)$, we have the equivalence

$$\forall x \in E^m \qquad P\left(\exists N \quad \forall n \geq N \quad f^* \subset [X_n] \, / X_0 = x\right) = 1$$

$$\Longleftrightarrow \qquad \sum_{n=0}^{\infty} l(n)^{-H_1} = \infty, \qquad \sum_{n=0}^{\infty} l(n)^{-H_e^*} < \infty.$$

Proposition. *The optimal convergence exponent α_{opt} is bounded between two affine strictly increasing functions of m.*

Theorem. *When the population size m is sufficiently large, each cycle over the set of attractors not containing the attractor K^* is reduced to one single attractor.*

References

[1] O. Catoni, *Large deviations for annealing*, PhD Thesis, University Paris XI (1990).

[2] O. Catoni, *Rough large deviations estimates for simulated annealing. Application to exponential schedules*, The Annals of Probability **20** (1992), 1109–1146.

[3] R. Cerf, *Asymptotic convergence of genetic algorithms*, Preprint, submitted (1993).

[4] R. Cerf, *The dynamics of mutation-selection algorithms with large population sizes*, Preprint, submitted (1993).

[5] R. Cerf, *A new genetic algorithm*, Preprint, submitted (1993).

[6] M.I. Freidlin–A.D. Wentzell, *Random perturbations of dynamical systems*, Springer–Verlag, New York, 1984.

[7] D. Goldberg, *Genetic algorithms in search, optimization and machine learning*, Addison–Wesley, 1989.

[8] J.H. Holland, *Adaptation in natural and artificial systems*, The University of Michigan Press, Ann Arbor, 1975.

[9] A. Trouvé, *Cycle decompositions and simulated annealing*, Preprint (1993).

[10] A. Trouvé, *Parallélisation massive du recuit simulé*, PhD Thesis, University Paris XI (1993).

[11] A. Trouvé, *Rough large deviation estimates for the optimal convergence speed exponent of generalized simulated annealing algorithms*, Preprint (1993).

Abstract

Genetic algorithms are search procedures based on the genetic mechanisms which guide natural evolution: mutation, crossover and selection. Today, they are widely used to handle a large spectrum of optimization problems ranging from the classical traveling salesman problem to the design of network architecture. Unfortunately, there is a critical lack of theoretical results for the convergence of this kind of algorithms.

The Freidlin–Wentzell theory deals with the study of random perturbations of dynamical systems. We build several models of genetic algorithms by randomly perturbing simple processes. The asymptotic dynamics of the resulting processes is analyzed with the powerful tools developed by Freidlin and Wentzell in a much more general framework.

In a first paper, a markovian model of Holland's simple genetic algorithm is built by randomly perturbing a very simple selection scheme: mutations and crossovers are considered as vanishing random perturbations. The convergence toward the global maxima of the fitness function becomes possible when the population size is greater than a critical value which depends strongly on the optimization problem. Surprisingly, the crossover is not fundamental to ensure this convergence: the crucial point is the delicate asymptotic interaction between the mutations and the selection pressure.

In a second paper, we use the concepts introduced by Catoni and further generalized by Trouvé to fathom more deeply the dynamics of the two operators mutation–selection algorithm when the population size becomes very large. The key result lies in the structure of the trajectories of populations joining two uniform populations: a small group of individuals sacrifice themselves in order to create an ideal path which is then followed by all other individuals. The various quantities associated with the algorithm are then affine functions of the population size. The hierarchy of cycles on the set of the uniform populations stabilizes. Furthermore, if the mutation kernel is symmetric, the limiting distribution is the uniform distribution over the set of the global maxima of the fitness function.

In a third paper, a new genetic algorithm is presented. It is built by randomly perturbing a two operators crossover–selection scheme. Three conditions of biological relevance are imposed on the crossover. A new selection mechanism is used, which has the decisive advantage of preserving the diversity of the individuals in the population. The attractors of the unperturbed process are particular equi–fitness subsets of populations endowed with a rich structure. The random vanishing perturbations are twofold: mutations and loosening of the selection pressure. When the population size is greater than a critical value, their delicate asymptotic interaction ensures the convergence of the population toward the ideal attractor whose populations contain all the global maxima of the fitness function.

Key words

Freidlin–Wentzell theory, evolutionary and genetic algorithms, stochastic optimization

AMS Subject Classification 60F10, 60J10, 92D15

AN ASYMPTOTIC THEORY FOR GENETIC ALGORITHMS

Raphaël Cerf

Genetic algorithms are search procedures based on the genetic mechanisms which guide natural evolution: mutation, crossover and selection. Today, they are widely used to handle a large spectrum of optimization problems ranging from the classical traveling salesman problem to the design of network architecture. Unfortunately, there is a critical lack of theoretical results for the convergence of this kind of algorithms.

The Freidlin–Wentzell theory deals with the study of random perturbations of dynamical systems. We build several models of genetic algorithms by randomly perturbing simple processes. The asymptotic dynamics of the resulting processes is analyzed with the powerful tools developed by Freidlin and Wentzell in a much more general framework.

In a first paper, a markovian model of Holland's simple genetic algorithm is built by randomly perturbing a very simple selection scheme: mutations and crossovers are considered as vanishing random perturbations. The convergence toward the global maxima of the fitness function becomes possible when the population size is greater than a critical value which depends strongly on the optimization problem. Surprisingly, the crossover is not fundamental to ensure this convergence: the crucial point is the delicate asymptotic interaction between the mutations and the selection pressure.

In a second paper, we use the concepts introduced by Catoni and further generalized by Trouvé to fathom more deeply the dynamics of the two operators mutation–selection algorithm when the population size becomes very large. The key result lies in the structure of the trajectories of populations joining two uniform populations: a small group of individuals sacrifice themselves in order to create an ideal path which is then followed by all other individuals. The various quantities associated with the algorithm are then affine functions of the population size. The hierarchy of cycles on the set of the uniform populations stabilizes. Furthermore, if the mutation kernel is symmetric, the limiting distribution is the uniform distribution over the set of the global maxima of the fitness function.

In a third paper, a new genetic algorithm is presented. It is built by randomly perturbing a two operators crossover–selection scheme. Three conditions of biological relevance are imposed on the crossover. A new selection mechanism is used, which has the decisive advantage of preserving the diversity of the individuals in the population. The attractors of the unperturbed process are particular equi–fitness subsets of populations endowed with a rich structure. The random vanishing perturbations are twofold: mutations and loosening of the selection pressure. When the population size is greater than a critical value, their delicate asymptotic interaction ensures the convergence of the population toward the ideal attractor whose populations contain all the global maxima of the fitness function.

Key words Freidlin–Wentzell theory, genetic algorithms, stochastic optimization

AMS Subject Classification 60F10, 60J10, 92D15