LIVING LA VIDA LOCA: LEARNING IN INTERPOLATION REGIMES

CLAIRE BOYER

Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université, INRIA Paris

2022, APRIL 6TH

Contents

Chapter 1

Interpolation in parametric learning

Contents

This chapter heavily relies on the articles Hastie et al. (2019), Bartlett et al. (2021) and the lecture notes of our spiritual father Francis Bach (from which I have shamelessly extracted parts). The reader may also find interesting developments in the lecture of Matus Telgarsky https://mjt.cs.illinois.edu/dlt/.

The performance of deep learning is remarkable and surprising, especially since it seems to contradict the statistical theory that has guarded against overfitting for decades: while being complex models, NN seem to still provide excellent predictive accuracy.

The training of NN is usually performed via stochastic gradient strategies (SGD). The conjecture is therefore that overparametrization allows gradient methods to find "good" interpolating solutions: the overfitting would be then "benign" as not harmful for the optimization of NN nor for the generalization abilities of the found solution.

1.1 Implicit bias of (S)GD in interpolation regimes

Statistical wisdom suggests that a method that takes advantage of too many degrees of freedom by perfectly interpolating noisy training data will be poor at predicting new outcomes. In deep learning, training algorithms appear to induce a bias that breaks the equivalence among all the models that interpolate the observed data.

1.1.1 Preliminary on optimization

The goal of an optimization problem is generally to minimize a function F over some parameter space Θ. If the global minimizer $\theta^\star$ is unique, even if the initial goal is to minimize F, one should expect that that the t-th iterate θ_t given by some optimization algorithm converges to that $\theta^\star$. When there are multiple minimizers (preventing the function to minimize to be strongly convex), one can only expect that $F(\theta_t) - \inf_{\theta\in\Theta} F(\theta)$ is converging to zero (and only if a minimizer exists). Note that when F is a convex function, the set of minimizers is a convex set.

With some extra assumptions, one can show that the algorithm is converging to one of the multiple minimizers of F. But which one? This is what is referred to as the implicit regularization properties of optimization algorithms, and here gradient descent and its variants.

Imagine now that, for a learning purpose, F stands for an empirical loss associated to n observations with $\Theta \subset \mathbb{R}^p$ and p much larger than n. No regularization being used, there are multiple minimizers achieving a zero training error (usually referred as the overfitting regime). Therefore, an arbitrary empirical risk minimizer is not expected to work well on unseen data. To reduce the complexity of the model (embodied by p here), a classical way to prevent overfitting is to use explicit regularization (e.g. ℓ^2-norms - Ridge/Tikhonov penalties- or ℓ^1-norms -Lasso penalties).

In this section, we show that optimization algorithms have a similar regularizing effect, without appealing to explicit penalties. In a nutshell, gradient descent usually leads to particular solutions of minimum ℓ^2-norm, meaning that the chosen empirical risk minimizer is not arbitrary.

1.1.2 Quadratic loss and linear models

Setting. To better understand this phenomenon, we restrict ourselves in a first time to the case of linear models. Choosing a quadratic loss for the empirical risk minimization boils down in building a least-square estimator:

$$F(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - X_i^\top\theta)^2 = \frac{1}{2n}\left\|y - \mathbb{X}\theta\right\|_2^2 \tag{1.1}$$

where $\mathbb{X}^\top\mathbb{X} = (X_1|X_2|\ldots|X_n)^\top \in \mathbb{R}^{n\times p}$ with $n \ll p$. The (kernel) matrix $\mathbb{X}\mathbb{X}^\top$ is assumed to be invertible. Therefore there is an infinity of minimizers of F corresponding to the solutions of the system $y = \mathbb{X}\theta$: the set of minimizers is actually an affine space (given a solution θ_0 to $y = \mathbb{X}\theta$, $\theta_0 + \text{null}(\mathbb{X})$ is also solution).

Running GD. Let's do the thought experiment that you are not able to write a particular solution of $y = \mathbb{X}\theta$, one can use a gradient descent strategy instead to minimize F. F being convex, the GD is going to converge to a global minimizer (and we know there exists since there is an infinity of solutions).

The gradient algorithm (GD) used to minimize F can be written as follows: for some initial parameter θ_0, the iterates of GD are

$$\theta_{t+1} \longleftarrow \theta_t - \eta\nabla F(\theta_t).$$

When the function F is assumed to be L-smooth (meaning that its gradient is assumed to be L-Lipschitz), the gradient algorithm is a descent method provided that the step η is chosen such that $\eta \le 1/L$.

Therefore applying GD with $\theta_0 = 0$ (zero initialization) and $\eta \le 1/\lambda_{\max}(\mathbb{X}^\top\mathbb{X}/n)$ and considering a solution θ of $y = \mathbb{X}\theta$ leads to

$$\theta_t - \theta = \left(I - \frac{\eta}{n}\mathbb{X}^\top\mathbb{X}\right)^t(\theta_0 - \theta) = -\left(I - \frac{\eta}{n}\mathbb{X}^\top\mathbb{X}\right)^t\theta$$

and

$$\theta_t = \left[I - \left(I - \frac{\eta}{n} \mathbb{X}^\top \mathbb{X} \right)^t \right] \theta \tag{1.2}$$

Note that it is not entirely obvious that the formula above is independent of the choice of θ (but it is).

Proposition 1.1

The solution of $y = \mathbb{X}\theta$ with the minimal ℓ^2-norm is

$$\mathbb{X}^\dagger y = V \mathrm{diag}(s^{-1}) U^T y$$

where

- $\mathbb{X}^\dagger$ is the pseudo-inverse of X, and
- $\mathbb{X} = U \mathrm{diag}(s) V^T$ is the SVD decomposition of X, so that
  - $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{p \times p}$ are orthonormal ($U^T U = I_n$ and $V^T V = I_d$),
  - $\mathrm{diag}(s) \in \mathbb{R}^{n \times p}$ ⚠ with the singular values $(s_i)_{1 \le i \le n}$ of $\mathbb{X}$.
    NB: $\mathrm{diag}(s^{-1}) \in \mathbb{R}^{p \times n}$

Proof. Left in exercise. □

Proposition 1.2

Considering the gradient descent initialized at 0 (1.2), one gets

$$\left\| \theta_t - V \mathrm{diag}(s^{-1}) U^T y \right\|_2 \le \left(1 - \frac{\min_i s_i^2}{\max_i s_i^2} \right)^t \left\| V \mathrm{diag}(s^{-1}) U^T y \right\|_2 . \tag{1.3}$$

Proof. Choosing $\theta = V \mathrm{diag}(s^{-1}) U^T y$ in (1.2) leads to

$$\theta_t = V \mathrm{diag}\left(\left(1 - \left(1 - \frac{\eta s_i^2}{n} \right)^t \right) s_i^{-1} \right) U^\top y. \tag{1.4}$$

Since each $s_i > 0$ for $i = 1, \dots, n$ (X is assumed of full rank) and $\eta \le 1/\lambda_{\max}(\mathbb{X}^\top \mathbb{X}/n) = n/\max_i s_i^2$, one gets

$$0 \le \left(1 - \left(1 - \frac{\eta s_i^2}{n} \right)^t \right) s_i^{-1} \le s_i^{-1} \left(1 - \left(1 - \frac{\eta \min_i s_i^2}{n} \right)^t \right).$$

This shows that

$$\left\| \theta_t - V \mathrm{diag}(s^{-1}) U^T y \right\|_2 \le \left(1 - \frac{\eta \min_i s_i^2}{n} \right)^t \left\| V \mathrm{diag}(s^{-1}) U^T y \right\|_2 . \tag{1.5}$$

Fixing η to be the largest step size allowed, i.e. $\eta = n/\max_i s_i^2$ gives the result. □

Note that $\frac{\min_i s_i^2}{\max_i s_i^2}$ can be seen as the inverse of the conditioning number of $\mathbb{X}$.

In the case of overparameterized linear regression, the gradient descent (with constant step size, initialized at 0) linearly converges towards the solution of $y = \mathbb{X}\theta$ of minimal ℓ^2-norm.

Question: how important is the initialization at zero?

Exercise: http://fa.bianp.net/blog/2022/implicit-bias-regression/

Consider the optimization problem where the objective function is a generalized linear model with a data matrix $\mathbb{X} = (X_1|\dots|X_n)^\top \in \mathbb{R}^{n\times p}$ and a target vector $y \in \mathbb{R}^n$:

$$\min_{\theta\in\mathbb{R}^p} f(\theta) = \sum_{i=1}^n \ell(X_i^\top\theta, y_i) \tag{1.6}$$

where $\ell(z, y)$ is a differentiable real-valued function verifying the "unique finite root condition", which is that it has a unique minimizer at $z = y$. These losses are usually used for regression and includes the quadratic loss or Huber functions. Assume that $p > n$ and that $\mathbb{X}$ is of full-rank.

Problems of this form verify two key properties that make it easy to characterize the bias of gradient-based methods. By gradient-based methods I mean any method in which the updates are given by a linear combination of current and past gradients. This includes gradient descent, gradient descent with momentum, stochastic gradient descent (SGD), SGD with momentum, Nesterov's accelerated gradient method. It does not include however quasi-Newton methods or diagonally preconditioned methods such as Adagrad or Adam.

1. Show that iterates remain in the span of $\mathbb{X}$. The gradient of the i-th sample $\ell(X_i^\top\theta, y_i)$ has the same direction as its data sample X_i:

$$\nabla_\theta\left[\ell(X_i^\top\theta, y_i)\right] = \underbrace{X_i}_{\text{vector}}\underbrace{\ell'(X_i^\top\theta, y_i)}_{\text{scalar}}$$

   This implies that any gradient-based method generates updates that stay in the span of the vectors $\{X_1, \dots, X_n\}$.

   It's no surprise then that the vector space generated by the samples $X_1, \dots, X_n$ plays a crucial role here. For convenience we'll denote this subspace by

$$\mathcal{X} := \text{span}\,(X_1, \dots, X_n) = \text{Im}(\mathbb{X}^\top)$$

   and its orthogonal complement $\mathcal{X}^\perp$.

2. What can you say about minimizers of f? Minimizers solve the linear system $\mathbb{X}\theta = y$. By the unique root condition of ℓ, the global minimizer is achieved when $X_i^\top\theta = y_i$ for all i. In other words, the global minimizers are the solutions to the linear system $\mathbb{X}\theta = y$, a set that is non-empty by the under-specification assumption.

3. Starting from θ_0, characterize the limit iterate of a gradient-based method. The main argument here is to show that the limit iterate belongs to the intersection of two affine spaces and then compute their intersection.

   By Property 2, the limit iterate must solve the linear system $\mathbb{X}\theta = y$. A classical linear algebra result states that all solutions of this problem are of the form $\theta + c$, with θ any solution of $\mathbb{X}\theta = y$ and $c \in \mathcal{X}^\top = \text{Ker}(\mathbb{X})$. Let's take $\theta = \mathbb{X}^\dagger y$ so that

$$\theta_\infty = \mathbb{X}^\dagger y + c, \text{ for some } c \in \mathcal{X}^\top$$

Let P denote the orthogonal projection onto $\mathscr{X}$. Then we can decompose the initialization as $\theta_0 = P\theta_0 + (I-P)\theta_0$. By the first property all updates are in $\mathscr{X}$, so the limit iterate can be written as

$$\theta_\infty = (I-P)\theta_0 + x \qquad \text{for some } x \in \mathscr{X}.$$

Combining the previous two equations, we have that $c = (I-P)\theta_0$ and $x = \mathbb{X}^\dagger y$. Hence we have arrived at the characterization

$$\theta_\infty = \mathbb{X}^\dagger y + (I-P)\theta_0. \tag{1.7}$$

4. Show that the limit iterate is actually the projection of θ_0 on the set of solutions of $\mathbb{X}\theta = y$. Let $\theta^\star$ denote the solution to

$$\theta^\star = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \|\theta-\theta_0\|_2 \qquad \text{such that} \qquad \mathbb{X}\theta = y.$$

$\theta^\star$ is unique by strong convexity. We want to show that $\theta_\star = \theta_\infty$. For any solution θ of $\mathbb{X}\theta = y$, one has $\theta - \theta_\infty \in \mathscr{X}^\perp$ and

$$\begin{aligned}\|\theta-\theta_0\|_2 &= \|\theta-\theta_\infty+\theta_\infty-\theta_0\|_2 \\ &= \|\theta-\theta_\infty+\mathbb{X}^\dagger y - P\theta_0\|_2 \\ &= \sqrt{\|\theta-\theta_\infty\|_2^2 + \|\mathbb{X}^\dagger y - P\theta_0\|_2^2}\end{aligned}$$

where the last identity follows by orthogonality. Since $\theta^\star$ minimizes the distance $\|\theta-\theta_0\|_2$ on the set of solutions of $\mathbb{X}\theta = y$, we must have $\theta^\star - \theta_\infty$=0, and so $\theta^\star = \theta_\infty$. We have actually shown the following result.

Theorem 1.1. *Gradient-based methods started from θ_0 converge to the solution with smallest distance to θ_0. More precisely, assume that the iterates of a gradient-based method converge to a solution of (1.6), and let $\theta_\infty := \lim_{t\to+\infty}\theta_t$ denote this limit. Then θ_∞ solves*

$$\theta_\infty = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \|\theta-\theta_0\|_2 \qquad \textit{such that} \qquad \mathbb{X}\theta = y.$$

5. Conclude on the role of the initialization to converge towards the solution of minimal ℓ^2-norm. An immediate consequence of this Theorem is that when the initialization θ_0 is in $\mathscr{X}$, then its projection onto $\mathscr{X}^\perp$ is zero, and so from Eq. (1.7) we have $\theta_\infty = \mathbb{X}^\dagger y$ which corresponds to the minimal norm solution.

Corollary 1.2. *If $\theta_0 \in \mathscr{X}$, then the limit iterate θ_∞ solves the minimal norm problem*

$$\theta_\infty = \operatorname*{argmin}_{\theta} \|\theta\|_2 \qquad \textit{such that} \qquad \mathbb{X}\theta = y.$$

Alternative proof for convergence (in short). If started at $\theta_0 = 0$, gradient descent techniques (stochastic or not) will always have iterates θ_t which are linear combinations of rows of $\mathbb{X}$, that is, of the form $\theta_t = \mathbb{X}^\top \alpha_t$ for some $\alpha_t \in \mathbb{R}^n$. This is an alternative algorithmic version of the *representer theorem*.

If the method is converging, then we must have $\mathbb{X}\theta_t$ converging to y (because the standard squared Euclidean norm is strongly-convex, and $\mathbb{X}\theta$ is unique while θ may not be), and thus $\mathbb{X}\mathbb{X}^\top\alpha_t$ is converging to y. If $K = \mathbb{X}\mathbb{X}^\top$ is invertible, this means that α_t is converging to $K^{-1}y$, and thus $\theta_t = X^\top\alpha_t$ is converging to $\mathbb{X}^T K^{-1} y$.

For the story to be complete, one should check that $\mathbb{X}^T K^{-1} y$ is indeed the solution to $y = \mathbb{X}\theta$ of minimal ℓ^2-norm. By standard Lagrangian duality one gets

$$\begin{aligned}
\inf_{\theta\in\mathbb{R}^d} \frac{1}{2}\|\theta\|_2^2 \quad \text{such that} \quad y = \mathbb{X}\theta \qquad &= \qquad \inf_{\theta\in\mathbb{R}^d} \sup_{\alpha\in\mathbb{R}^n} \underbrace{\frac{1}{2}\|\theta\|_2^2 + \alpha^\top (y - \mathbb{X}\theta)}_{\text{Lagrangian function } L(\theta,\alpha)} \\
&= \qquad \sup_{\alpha\in\mathbb{R}^n} \alpha^\top y - \frac{1}{2}\left\|\mathbb{X}^\top \alpha\right\|_2^2 \quad \text{(with } \theta = \mathbb{X}^\top\alpha \text{ at the optimum)} \\
&= \qquad \sup_{\alpha\in\mathbb{R}^n} \alpha^\top y - \frac{1}{2}\alpha^\top K\alpha.
\end{aligned}$$

The last problem is exactly solved for $\alpha = K^{-1} y$.

What about SGD? Note that in the overparameterized regime, SGD will also converge to the minimum-norm interpolator, even with a fixed learning rate. In constrast, under-parameterized SGD with a fixed learning rate does not converge at all (indeed the stochastic noise at the optimum is 0 only in the over-parameterized setting).

1.1.3 Interpolation in logistic regression

Context. Consider now the setting of binary classification (for the output Y living in $\{-1, 1\}$), based on the model of logistic regression, i.e. the prior on the distribution of $Y|X$ is

$$\mathbb{P}(Y = +1|X = x) = \sigma(\varphi(x)^\top \beta)$$

where σ is the sigmoid function, φ is an encoding of the input variables X with $\varphi(x) \in \mathbb{R}^p$, and the model parameters are $\beta \in \mathbb{R}^p$. Given a dataset $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ i.i.d. copies of (X, Y), the estimation of β is usually performed via MLE, resulting in solving the following problem:

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp\left(-Y_i \varphi(X_i)^\top \beta\right)\right) =: F(\beta).$$

Define the design matrix as $\Phi := \left(\varphi(X_1)|\ldots|\varphi(X_n)\right)^\top \in \mathbb{R}^{n\times p}$ and consider the case where $d > n$, assuming in addition that $\Phi\Phi^\top$ is invertible.

Rewriting an SVM Since $\Phi\Phi^\top$ is invertible, there exists $\eta \in \mathbb{R}^p$ of unit-norm such that for all $i \in \{1, \ldots, n\}$, $Y_i \varphi(X_i)^\top \eta > 0$, meaning that the data is linearly separable[1]. The distance of any point $\varphi(x) \in \mathbb{R}^p$ to a hyperplane defined by $\{x' : \eta^\top \varphi(x') + b = 0\}$ is given by

$$\frac{|\langle \eta, x\rangle + b|}{\|\eta\|}$$

Therefore, the distance of a separating hyperplane to the closest points in the dataset, which is called the *margin* is given by

$$\min_{x\in\{X_1,\ldots,X_n\}} \frac{|\eta^\top x + b|}{\|\eta\|} = \min_{1\le i\le n} Y_i \varphi(X_i)^\top \eta,$$

when no intercept is considered. One can thus search for a direction η of unit ℓ^2-norm that maximizes the margin:

$$\eta^\star \in \operatorname*{argmax}_{\|\eta\|_2\le 1} \min_{1\le i\le n} Y_i \varphi(X_i)^\top \eta.$$

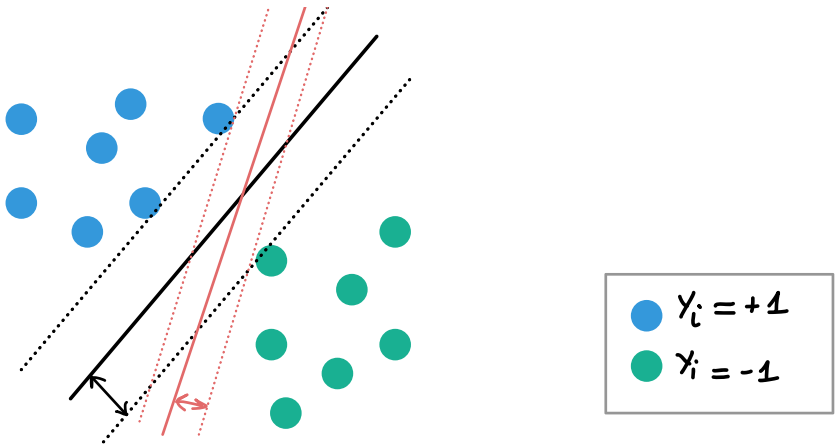


Figure 1.1: The maximum-margin classifier (in black) vs. a classifier based on an arbitrary separating hyperplane (in orange)

$\eta^\star$ corresponds to the max-margin classifier (SVM). By Lagrange duality,

$$\begin{aligned}\sup_{\|\eta\|_2\le 1} \inf_{i\in\{1,\dots,n\}} Y_i\varphi(X_i)^\top\eta &= \sup_{\|\eta\|_2\le 1} t \quad \text{such that} \quad \forall i\in\{1,\dots,n\}, Y_i\varphi(X_i)^\top\eta \ge t,\\ &= \inf_{\alpha\in\mathbb{R}^n_+} \sup_{\|\eta\|_2\le 1} t + \sum_{i=1}^n \alpha_i (Y_i\varphi(X_i)^\top\eta - t)\\ &= \inf_{\alpha\in\mathbb{R}^n_+} \left\| \sum_{i=1}^n \alpha_i Y_i\varphi(X_i)\right\|_2 \quad \text{such that} \quad \sum_{i=1}^n \alpha_i = 1.\end{aligned}$$

where in the last step we used:

1. (Lagrangian for the constrained sup) $L(t,\eta,\mu,\alpha) = t + \sum_i \alpha_i (Y_i\varphi(X_i)^\top\eta - t) + \mu(\|\eta\|_2^2 - 1)$
2. (KKT 1) $\nabla_t L = 0$, i.e. $\sum_i \alpha_i = 1$
3. (KKT 2) $\nabla_\eta L = 0$, i.e. $\sum_i \alpha_i Y_i\varphi(X_i) + 2\mu\eta = 0$
4. (KKT 3: complementary slackness 1) $\mu = 0$ or $\|\eta\|_2^2 = 1$
5. (KKT 4: complementary slackness 2) $\forall i$, $\alpha_i = 0$ or $t = Y_i\varphi(X_i)^\top\eta$

so that $\eta \propto \sum_{i=1}^n \alpha_i Y_i\varphi(X_i)$ at the optimum. Besides, by complementary slackness, non-negative α_i is non-zero only for i such that at the optimum $t = Y_i\varphi(X_i)^\top\eta$, i.e. for i attaining the minimum $\min_{1\le i\le n} Y_i\varphi(X_i)^\top\eta$, corresponding to the so-called support vectors, see Figure 1.2.

Link with the traditional SVM Because of homogeneity, we want $\min_{1\le i\le n} Y_i\varphi(X_i)^\top\eta$ to be large and $\|\eta\|_2$ to be small. We can therefore constrain the former, and minimize the latter. In other words,

[1] the invertibility of Φ prevents F to admit a critical point and then a minimum (only an infimum).

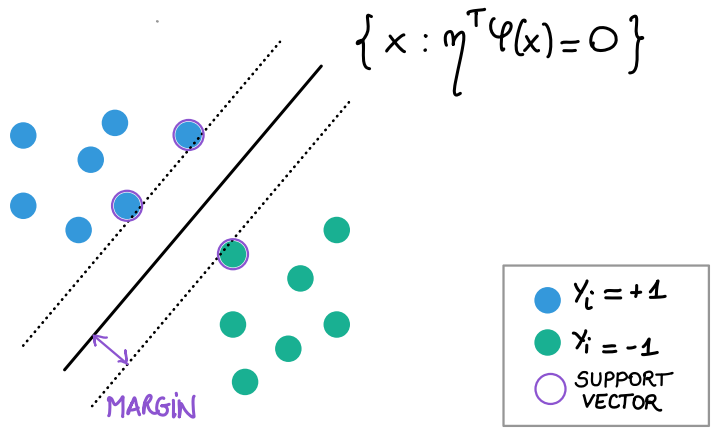


Figure 1.2: Support vectors.

we can see $\eta^\star$ as the direction of $\beta^\star$, solution of

$$\inf_{\beta\in\mathbb{R}^p} \frac{1}{2}\|\beta\|_2^2 \quad \text{such that} \quad \underbrace{\operatorname{diag}((Y_i)_i)\Phi\beta \geq \mathbb{1}_n}_{\substack{\text{perfect classifier} \\ \text{interpolating training data}}} = \inf_{\beta\in\mathbb{R}^p} \sup_{\alpha\in\mathbb{R}_+^n} \frac{1}{2}\|\beta\|_2^2 + \alpha^\top \left(\mathbb{1}_n - \operatorname{diag}((Y_i)_i)\Phi\beta\right)$$

$$= \sup_{\alpha\in\mathbb{R}_+^n} \alpha^\top \mathbb{1}_n - \frac{1}{2}\left\|\Phi^\top \operatorname{diag}((Y_i)_i)\alpha\right\|_2^2$$

with $\beta = \Phi^\top \operatorname{diag}((Y_i)_i)\alpha$ at the optimum.

Note that above, $\operatorname{diag}((Y_i)_i)\Phi\beta \geq \mathbb{1}_n$ is the compact formulation of: for all $i \in \{1,\dots,n\}$,

$$Y_i \varphi(X_i)^\top \beta \geq 1.$$

This amounts to take the following convention: the margin hyperplanes are shifted by 1 and -1, see Figure 1.3.

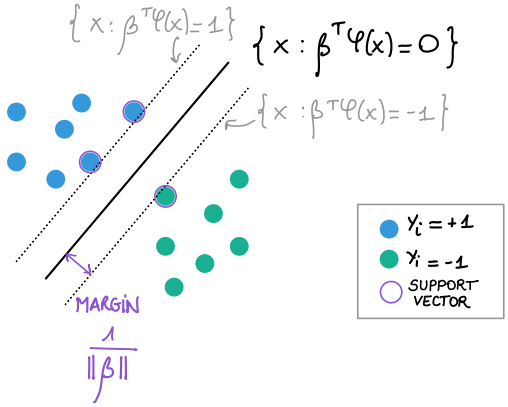


Figure 1.3: A traditional SVM with the margin hyperplanes shifted by 1 and -1.

Overall, the optimal $\beta^\star$ above is the solution of the separable SVM with vanishing regularization parameter, that is, of $\frac{1}{2}\|\beta\|_2 + C\sum_{i=1}^n (1 - Y_i\varphi(X_i)^\top\beta)_+$ for C large enough.

Divergence for the logistic regression with hands. The function F has an infimum equal to zero, which is not attained. However, for any sequence β_t such that all $Y_i\varphi(X_i)^\top\beta_t$ tend to infinity, we have $F(\beta_t) \to \inf_{\beta\in\mathbb{R}^d} F(\beta) = 0$.

In such a situation, gradient descent cannot converge to a point, and, to achieve small values of F, it has to diverge. It turns out that it diverges along a direction, that is, $\|\beta_t\|_2 \to +\infty$, with $\beta_t/\|\beta_t\|_2 \to \eta$ for some $\eta \in \mathbb{R}^d$ of unit ℓ_2-norm. See Soudry et al. (2018) for a proof. Here, we just show what the vector η is.

The gradient $\nabla F(\beta)$ is given by

$$\nabla F(\beta) = -\frac{1}{n}\sum_{i=1}^n \frac{\exp(-Y_i\varphi(X_i)^\top\beta)}{1+\exp(-Y_i\varphi(X_i)^\top\beta)} Y_i\varphi(X_i).$$

Asymptotically, β_t behaves as $\|\beta_t\|_2\eta$ with $\|\beta_t\|_2$ tending to infinity. By the structure of the sum of exponentials, the dominant term in $\nabla F(\beta_t)$ corresponds to the indices i for which $-Y_i\varphi(X_i)^\top\eta$ is the largest. Moreover, all of these values have to be negative (indeed we can only attain zero loss for well-classified training data). We denote by I this set. Thus,

$$\nabla F(\beta) \sim -\frac{1}{n}\sum_{i\in I} Y_i \exp\left(-\|\beta_t\|_2 Y_I\varphi(X_i)^\top\eta\right)\varphi(X_i).$$

Moreover, we must have $F(\beta_t)$ along $-u$ to diverge in the direction u, thus u has to be proportional to a vector $\sum_{i\in I}\alpha_i Y_i\varphi(X_i)$, where $\alpha \geq 0$ and $\alpha_i = 0$ as soon as i is not among the minimizers $Y_i\varphi(X_i)^\top\eta$. This is exactly the optimality condition for $\eta^\star$ above. Thus $\eta = \eta^\star$. Overall, we obtain a classifier corresponding to a minimum ℓ^2-norm.

See Lyu and Li (2019) for an extension beyond the linear classification case.

1.1.4 Implicit bias in neural network training

In this section we consider one-hidden-layer neural networks of the form

$$f(x) = \frac{1}{K}\sum_{k=1}^K a_k\sigma\left(x^\top b_k\right) = \frac{1}{K}\sum_{k=1}^K a_k\left(x^\top b_k\right)_+$$

with σ the ReLU activation function, K the number of neurons in the hidden layer, $(a_k)_k$ the weights between the hidden layer and the output layer, $(b_k)_k$ the weights between the input layer and the hidden layer.

Lifting the problem to the space of measures Setting $\omega_k = (a_k, b_k)$ and $\Phi(\omega) = a\sigma(\cdot^\top b)$, the NN can be rewritten,

$$f = \frac{1}{K}\sum_{k=1}^K \Phi(\omega_k).$$

Therefore the neural network parameterized by $\left(\mathbb{R}^{p+1}\right)^K$ can be represented by the discrete measure $\mu = \frac{1}{K}\sum_{k=1}^K \delta_{\omega_k} = \frac{1}{K}\sum_{k=1}^K \delta_{(a_k,b_k)}$, so that we embed $\left(\mathbb{R}^{p+1}\right)^K$ into the space of Radon measures, so that

$$f = f_\mu = \int \Phi(\omega)\,d\mu(\omega).$$

What do we gain by doing so? Note that the mapping $(a,b) \mapsto f_{(a,b)}$ is not linear, whereas $\mu \mapsto f_\mu$ is linear.

Therefore training a one-hidden-layer NN entails to find the best parameters $(a_k, b_k)_k$, and therefore to find the best measure μ. The key benefit is that the set of measures is convex and $\mu \mapsto f_\mu = \int \Phi(\omega) d\mu(\omega)$ is linear in the measure μ, so that the risk minimization problem has become convex:

$$\min_\mu \mathscr{R}\left(\int \Phi(\omega) d\mu(\omega)\right).$$

Gradient descent on such a space? Now we weed to define what a gradient descent is on the space of measure. To do so, we need a metrics: the Wasserstein metrics given by

$$W_2^2(\mu,\nu) = \inf_{\substack{(X,Y) \\ X\sim\mu\ Y\sim\nu}} \mathbb{E}\left[|X-Y|^2\right].$$

Usually a gradient descent with a discretization step h can be written as follows:

$$x_t^h = x_{t-1}^h - h\nabla\mathscr{R}(x_{t-1}^h) \qquad \text{(explicit)}$$
$$x_t^h = x_{t-1}^h - h\nabla\mathscr{R}(x_t^h) \qquad \text{(implicit)}$$

The last version is equivalent in the smooth case to find

$$x_t^h \in \operatorname*{argmin}_x \mathscr{R}(x) + \frac{1}{2h}|x - x_{t-1}^h|^2.$$

In our case,

$$\mu_t^h \in \operatorname*{argmin}_\mu \mathscr{R}(\mu) + \frac{1}{2h} W_2^2(\mu, \mu_{t-1}^h),$$

and let h go to 0 to obtain the Wasserstein gradient flow.

The squared 2-Wasserstein distance between two discrete measures with the same number of atoms is obtained by minimizing the pairwise distance between Dirac masses over the set of all permutations.

This can be extended to any pair of probability measures, and used within gradient flows, it has a very natural decoupling property: if μ is fixed, and ν is within a small distance of μ in Wasserstein distance, then the optimal permutation above will always be the same, that is, locally, the Wasserstein distance is a sum of squared Euclidean distances. Then, the Wasserstein gradient flow will lead to K independent local regular Euclidean gradient flows, which interact through the gradient term as:

$$\dot{\omega_k} = -\nabla\Phi(\omega_k)\nabla\mathscr{R}\left(\int \Phi d\mu\right)$$

where

- $\nabla\Phi$ is a linear operator from $\mathscr{F}$ to $\mathbb{R}^{p+1}$
- $\nabla\mathscr{R}$ is a the gradient operator of $\mathscr{R}$ from $\mathbb{R}^{p+1}$ to $\mathscr{F}$.

Main result Here we state the result of Chizat and Bach (2020) in a very informal way, and one can refer to a related blog post of Francis Bach.

Theorem 1.3. *Assume that for some $r > 0$, the hidden weights $(b_k)_k$ are initialized uniformly on the sphere of radius r and the output weights $(a_k)_k$ uniformly in $\{r, -r\}$. Let μ_t be the Wasserstein gradient flow μ_t for the unregularized exponential loss and $f_t = \int \Phi(\omega) d\mu_t(\omega)$ be the corresponding dynamics in predictor space. Under some technical assumptions, the normalized predictor $\frac{f_t}{\|f_t\|_{\mathscr{F}_1}}$ converges to*

$$\max_{\|f\|_{\mathscr{F}_1} \leq 1} \min_i Y_i f(X_i),$$

where

$$\|f\|_{\mathscr{F}_1} := \min_{\mu} \frac{1}{2} \int \|\omega\|_2^2 d\mu(\omega) \qquad \textit{such that} \qquad f = \int \Phi(\omega) d\mu(\omega).$$

1.2 Interpolation is no longer synonym of bad generalization

The aim of this section is to present recent developments on the generalisation capabilities of neural networks, which in practice seem fantastic and which classical generalisation error bounds struggle to explain.

1.2.1 Preliminaries

A first lecture on machine/statistical learning traditionally warns the reader to the evils of overfitting, see Figure 1.4.

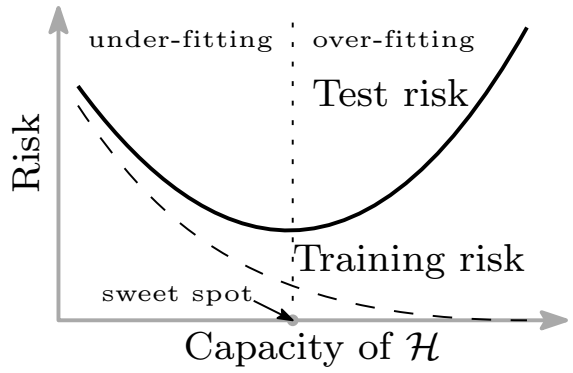


Figure 1.4: A typical learning curve about the bias-variance trade-off in the prediction when increasing the predictor complexity.

Typically the "capacity" of the space of learners $\mathscr{H}$ is controlled either by the number of parameters, or by some norms of its parameters. In particular, at the extreme right of the curve, when there is zero training error, the testing error may be arbitrarily large (bad), and the classical theoretical bound, such as Rademacher averages for $\mathscr{H}$ controlled by the ℓ^2-norm of some parameters (with a bound D), grows as $D/\sqrt{n}$, which can be typically quite large.

Proposition 1.4 (Estimation error)**.** *Assume a G-Lipschitz-continuous loss function ℓ, linear prediction functions with $\mathscr{F} = \{f_\theta(x) = \theta^\top \phi(x), \|\theta\|_2 \leq D\}$, where $\mathbb{E}\left[\|\phi(x)\|_2^2\right] \leq R^2$. Let $\hat{F} = f_{\hat{\theta}} \in \mathscr{F}$ be the minimizer of the empirical risk, then*

$$\mathbb{E}\left[\mathscr{R}(f_{\hat{\theta}})\right] \leq \inf_{\|\theta\|_2 \leq D} \mathscr{R}(f_\theta) + \frac{2GRD}{\sqrt{n}}.$$

Model	Year	Nb of layers	Nb of param	Error
Shallow	<2012	-	-	> 25%
AlexNet	2012	8	61M	16.4%
VGG19	2014	19	144M	7.3%
GoogleNet	2014	22	7M	6.7%
ResNet-152	2015	152	60M	3.6%

Table 1.1: Performances of different architectures on the ImageNet dataset ($n = 500000$) in regard of the learning complexity captured here through the number of parameters or layers.

Here is a table summarizing the performances of different learners on the ImageNet dataset ($n = 500000$).

When the model is over-parameterized (in other words, the capacity gets very large), that is, when the number of parameters is large or the norm constraint allows for exact fitting, a new phenomenon occurs, where after the test error explodes as the capacity grows, it goes down again: this is the so-called *double descent* curve.

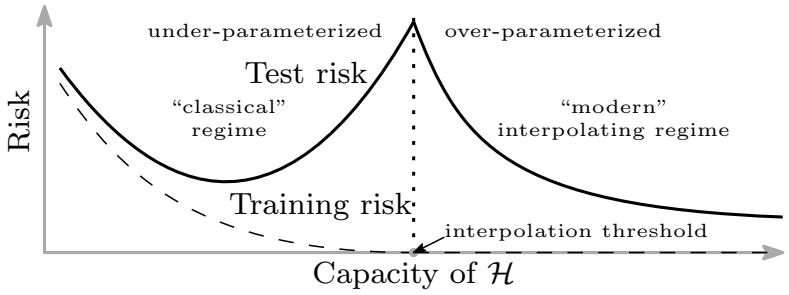


Figure 1.5: From Belkin et al. (2019) The learning story: to be continued.

1.2.2 Linear model

This paradox has been resolved in the case of linear models by Hastie et al. (2019), relying on non-asymptotic results for random matrices.

Consider a Gaussian random variable with mean 0 and covariance matrix identity, with n observations $X_1, \dots, X_n$, and responses $Y_i = X_i^\top \theta^\star + \varepsilon_i$, with ε_i normal with zero mean and variance $\sigma^2 I$.

We know the exact expression of the empirical risk minimizer (for which we know that gradient descent will converge to under proper initialization). Denote the design matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$, the non-centered covariance matrix $\hat{\Sigma} = \mathbb{X}^\top \mathbb{X}/n$, and the kernel matrix $K = \mathbb{X}\mathbb{X}^\top$.

The excess risk is

$$R(\hat{\theta}) = \mathbb{E}_X\left[(X^\top \hat{\theta} - X^\top \theta^\star)^2\right] = \mathbb{E}_X(\hat{\theta} - \theta^\star) X X^\top (\hat{\theta} - \theta^\star) = (\hat{\theta} - \theta^\star)\Sigma(\hat{\theta} - \theta^\star)$$
$$= \|\hat{\theta} - \theta^\star\|_2^2.$$

Underparameterized regime. In the underparameterized regime, then the minimum norm empirical risk minimizer is simply the ordinary least-squares estimator, which is unbiased, that is $\mathbb{E}\left[\hat{\theta}\right] = \theta^\star$, and we have an expected excess risk equal to

$$\mathbb{E}_{\mathcal{D}_n}\left[R(\hat{\theta})\right] = \frac{\sigma^2}{n}\mathbb{E}\left[tr\left(\Sigma\hat{\Sigma}^{-1}\right)\right]$$

Indeed,

$$\begin{aligned}
\mathbb{E}_{\mathscr{D}_n,\varepsilon}\big[R(\hat{\theta})\big] &= \mathbb{E}_{\mathscr{D}_n,\varepsilon}\big[\|\hat{\theta}-\theta^\star\|_{2,\Sigma}^2\big] \\
&= \mathbb{E}_{\mathscr{D}_n,\varepsilon}\big[\|(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top Y-\theta^\star\|_{2,\Sigma}^2\big] = \mathbb{E}_{\mathscr{D}_n,\varepsilon}\big[\|(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\big(\mathbb{X}\theta^\star+\varepsilon\big)-\theta^\star\|_{2,\Sigma}^2\big] \\
&= \mathbb{E}_{\mathscr{D}_n,\varepsilon}\big[\|(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\varepsilon\|_{2,\Sigma}^2\big] = \mathbb{E}_{\mathscr{D}_n,\varepsilon}\big[\varepsilon^\top\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\Sigma(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\varepsilon\big] \\
&= \mathbb{E}_{\mathscr{D}_n,\varepsilon}\big[\,tr\big(\varepsilon^\top\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\Sigma(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\varepsilon\big)\,\big] \\
&= \sigma^2\mathbb{E}_{\mathscr{D}_n}\big[\,tr\big(\Sigma(\mathbb{X}^\top\mathbb{X})^{-1}\big)\,\big] = \frac{\sigma^2}{n}\mathbb{E}\big[tr\big(\Sigma\hat{\Sigma}^{-1}\big)\big].
\end{aligned}$$

In our case, $\Sigma = I$, so that the expected risk boils down to

$$\mathbb{E}_{\mathscr{D}_n,\varepsilon}\big[R(\hat{\theta})\big] = \sigma^2\mathbb{E}_{\mathscr{D}_n}\big[\,tr(\mathbb{X}^\top\mathbb{X})^{-1}\big]$$

The matrix $\mathbb{X}\in\mathbb{R}^{n\times p}$ is Gaussian, so that the matrix $\mathbb{X}^\top\mathbb{X}\in\mathbb{R}^{p\times p}$ has a Wishart distribution, with n degrees of freedom:

- it is almost surely invertible if $n > p$,
- $\mathbb{E}_{\mathscr{D}_n}\big[\,tr\big((\mathbb{X}^\top\mathbb{X})^{-1}\big)\,\big] = \frac{p}{n-p-1}$ if $n \geq p+2$. The expectation is infinite for $n = p$ or $n = p+1$.

Proof. Here is a simple way to derive the expectation of an inverse Wishart matrix W^{-1} where $W = \sum_{i=1}^n \Sigma^{1/2} g_i g_i^T \Sigma^{1/2}$ for a covariance $\Sigma \in R^{p\times p}$ and i.i.d. standard vectors $g_i \sim N(0, I_p)$. The covariance Σ is assumed invertible. The point follows jlewk (https://math.stackexchange.com/users/484640/jlewk).

The first observation is that

$$E[W^{-1}] = \Sigma^{-1/2} E\left[\left(\sum_{i=1}^n g_i g_i^T\right)^{-1}\right]\Sigma^{-1/2}$$

so that it is enough to treat the case $\Sigma = I_p$. Assume $\Sigma = I_p$ here after.

Concerning the non-diagonal terms of $E[W^{-1}]$, note that with identity covariance, $\sum_i g_i g_i^T$ and $\sum_i \tilde{g}_i \tilde{g}_i^T$ with $\tilde{g}_i = Dg_i$ have the same distribution where $D = \mathrm{diag}(1,...,1,-1,1,..1)$ (only one sign changes). This implies that

$$E[W^{-1}] = D^{-1}E[W^{-1}]D^{-1}$$

so that $E[W^{-1}]_{ij} = 0$ for $i \neq j$ (outside of the diagonal).

Concerning the diagonal terms, by symmetry,

$$E[W^{-1}]_{ii} = \frac{1}{d}E[\mathrm{trace}[W^{-1}].$$

The trace is also the sum of the eigenvalues $\lambda_i(W^{-1})$ of W^{-1}, or the following Frobenius norm:

$$E\big[\,\mathrm{trace}[W^{-1}]\big] = E\sum_{i=1}^d \lambda_i(W)^{-1} = E[\|G^\dagger\|_F^2]$$

where $G \in R^{n\times p}$ is the matrix with n rows $g_1,...,g_n$, and $\dagger$ denotes the pseudo-inverse. At this point, if $c_1,...,c_p$ are the columns of $G^\dagger$, the above display is $\sum_{j=1}^p \|c_j\|_2^2$. Furthermore by definition of the pseudo-inverse, with $z_1,...,z_d$ the rows of G, we have $c_j^T z_j = 1$ and $c_j^T z_k 0$ for $j \neq k$. This implies that c_j belongs to the orthogonal complement of $\{z_k, k \in \{1,...,p\}\setminus j\}$. Since c_j belongs to the span of $z_1,...,z_p$, it must be that $c_j = \theta_j Q_j z_j$ with $Q_j \in R^{n\times n}$ the orthogonal projection onto $\{z_k, k \in \{1,...,p\}\setminus j\}^\perp$ and θ_j

a scalar. The condition $c_j^T z_j = 1$ then reveals $\theta_j = \|Q_j z_j\|_2^{-2}$. Finally, $\|Q_j z_j\|_2^2$ has χ^2_{n-p+1} distribution as Q_j and z_j are independent thanks to G having i.i.d. $N(0,1)$ entries, hence

$$E[\text{trace}[W^{-1}]] = E\sum_{j=1}^{d}\|c_j\|_2^2 = E\sum_{j=1}^{d}\|Q_j z_j\|_2^{-2} = \frac{p}{(n-p+1)-2} = \frac{p}{n-p-1}$$

provided that we already know that the expectation of an inverse χ^2_ν distribution has expectation $1/(\nu-2)$ for $\nu > 2$.

□

In conclusion, in the underparameterized regime when $n \geq p+2$, the excess risk is equal to

$$\mathbb{E}\big[R(\hat{\theta})\big] = \sigma^2\frac{p}{n-p-1}.$$

Overparameterized regime. In the overparameterized regime, when $n \leq p$, then the kernel matrix is almost surely invertible, and the minimum ℓ^2-norm interpolator $\hat{\theta}$ is equal to

$$\hat{\theta} = \mathbb{X}^\dagger Y = \mathbb{X}^\top\left(\mathbb{X}\mathbb{X}^\top\right)^{-1}Y = \mathbb{X}^\top\left(\mathbb{X}\mathbb{X}^\top\right)^{-1}\left(\mathbb{X}\theta^\star + \varepsilon\right)$$

The expected excess risk can be decomposed into a variance and a bias term:

(i) The variance term is equal to

$$\begin{aligned}\mathbb{E}\left[\varepsilon^\top\left(\mathbb{X}\mathbb{X}^\top\right)^{-1}\mathbb{X}\Sigma\mathbb{X}^\top\left(\mathbb{X}\mathbb{X}^\top\right)^{-1}\varepsilon\right] &= \sigma^2\mathbb{E}\left[tr\left(\mathbb{X}^\top\right)^{-1}\mathbb{X}\mathbb{X}^\top\left(\mathbb{X}\mathbb{X}^\top\right)^{-1}\right]\\ &= \sigma^2\mathbb{E}\left[tr\left(\mathbb{X}\mathbb{X}^\top\right)^{-1}\right]\\ &= \sigma^2\frac{n}{p-n+1}\end{aligned}$$

when $p \geq n+2$ since we recognize a similar Wishart matrix as before (with the role of n and p reversed).

(ii) The bias term is equal to

$$\begin{aligned}\mathbb{E}\left[(\theta^\star)^\top\Big(I - \mathbb{X}^\top\left(\mathbb{X}\mathbb{X}^\top\right)^{-1}\mathbb{X}\Big)\theta^\star\right] &= \mathbb{E}\left[\left\|\text{Proj}_{\text{span}(X_1,\dots,X_n)^\perp}\left(\theta^\star\right)\right\|_2^2\right] = \mathbb{E}\left[\left\|\text{Proj}_{\text{Im}(\mathbb{X}^\top)^\perp}\left(\theta^\star\right)\right\|_2^2\right]\\ &= \mathbb{E}\left[\left\|\text{Proj}_{\text{Ker}(\mathbb{X})}\left(\theta^\star\right)\right\|_2^2\right]\end{aligned}$$

Indeed, the matrix $\mathbb{X}^\top\left(\mathbb{X}\mathbb{X}^\top\right)^{-1}\mathbb{X} \in \mathbb{R}^{p\times p}$ is the projection matrix on the rowspace of $\mathbb{X}$, which is a random subspace of dimension n corresponding to the linear span of the p-dimensional vectors $\{X_1,\dots,X_n\}$. By rotational invariance of the Gaussian distribution, this random subspace is uniformly distributed among all subspaces, and therefore, by rotational invariance, we can replace $\theta^\star$ by $\|\theta^\star\|_2 e_j$, for any of the canonical basis vector e_j in dimension p, that is

$$\mathbb{E}\left[(\theta^\star)^\top\mathbb{X}^\top\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}\theta^\star\right] = \|\theta^\star\|_2^2\mathbb{E}\left[e_j^\top\mathbb{X}^\top\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}e_j\right]$$

and thus

$$\begin{aligned}\mathbb{E}\left[(\theta^\star)^\top\mathbb{X}^\top\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}\theta^\star\right] &= \frac{\|\theta^\star\|_2^2}{p}\sum_{j=1}^{p}\mathbb{E}\left[e_j^\top\mathbb{X}^\top\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}e_j\right]\\ &= \frac{\|\theta^\star\|_2^2}{p}\mathbb{E}\left[tr\left(\mathbb{X}^\top\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}\right)\right]\\ &= \|\theta^\star\|_2^2\frac{n}{p}.\end{aligned}$$

Thus the bias term leads to

$$\mathbb{E}\left[(\theta^\star)^\top\left(I-\mathbb{X}^\top\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}\right)\theta^\star\right]=\|\theta^\star\|_2^2\frac{p-n}{p}.$$

Therefore, the overall expected risk is

$$\mathbb{E}\left[R(\hat{\theta})\right]=\sigma^2\frac{n}{p-n+1}+\|\theta^\star\|_2^2\frac{p-n}{p}.$$

Wrapping up One gets

$$\begin{cases}\text{if } p\le n-2, & \mathbb{E}\left[R(\hat{\theta})\right]=\sigma^2\frac{p}{n-p-1},\\ \text{if } p\ge n+2 & \mathbb{E}\left[R(\hat{\theta})\right]=\sigma^2\frac{n}{p-n+1}+\|\theta^\star\|_2\frac{p-n}{p},\end{cases}$$

as illustrated on Figure 1.7. The interpretation of these bounds are taken from Hastie et al. (2019):

- The **bias** increases with p/n in the overparameterized regime, which is intuitive. When $p>n$, the min-norm least squares estimate is constrained to lie the row space of $\mathbb{X}$, the training feature matrix. This is a subspace of dimension n lying in a feature space of dimension p. Thus as p increases, so does the bias, since this row space accounts for less and less of the ambient p-dimensional feature space. Another way to see it is to note that the bias is nothing else than

$$\mathbb{E}\left[\left\|\text{Proj}_{\text{Ker}(\mathbb{X})}\left(\theta^\star\right)\right\|_2^2\right]$$

with $\dim\left(\text{Ker}(\mathbb{X})\right)=p-n$. Therefore

$$\min_{\theta\in\text{Ker}(\mathbb{X})}\|\theta-\theta^\star\|_2=\left\|\text{Proj}_{\text{Ker}(\mathbb{X})}\left(\theta^\star\right)-\theta^\star\right\|_2^2$$

which $\searrow$ when $p\nearrow$ (the minimum is least on a larger subspace), so that $\text{Proj}_{\text{Ker}(\mathbb{X})}$ increases with p.

- In the overparameterized regime, the **variance** decreases with p/n. This may seem counterintuitive at first, because it says, in a sense, that the min-norm least squares estimator becomes more regularized as p grows. However, this is explained as follows by the authors of Hastie et al. (2019): as p grows, the minimum ℓ^2-norm least squares solution—i.e., the minimum ℓ^2-norm solution to the linear system $\mathbb{X}\theta=Y$, for a training feature matrix $\mathbb{X}$ and response vector Y—will generally have decreasing ℓ^2-norm. Why? Compare two such linear systems: in each, we are asking for the min-norm solution to a linear system with the same Y, but in one instance we are given more columns in $\mathbb{X}$, so we can generally decrease the components of θ (by distributing them over more columns), and achieve a smaller ℓ^2-norm.

- Set the SNR= $\|\theta^\star\|_2^2/\sigma^2$. Note that the risk of the null estimator (i.e. $\hat{\theta}=0$) is $\|\theta^\star\|_2^2$, which can be called the null risk. In the overparameterized regime, with an infinite sample size, and with $p/n\to\gamma$,

  - when SNR ≤ 1, the min-norm least squares risk is always worse than the null risk. Moreover, it is monotonically decreasing, and approaches the null risk (from above).
  - When SNR >1, the min-norm least squares risk beats the null risk if and only if $\gamma>\text{SNR}/(\text{SNR}-1)$. It has a local minimum at $\gamma=\sqrt{\text{SNR}}/(\sqrt{SNR}-1)$, and approaches the null risk from below when $\gamma\to+\infty$.

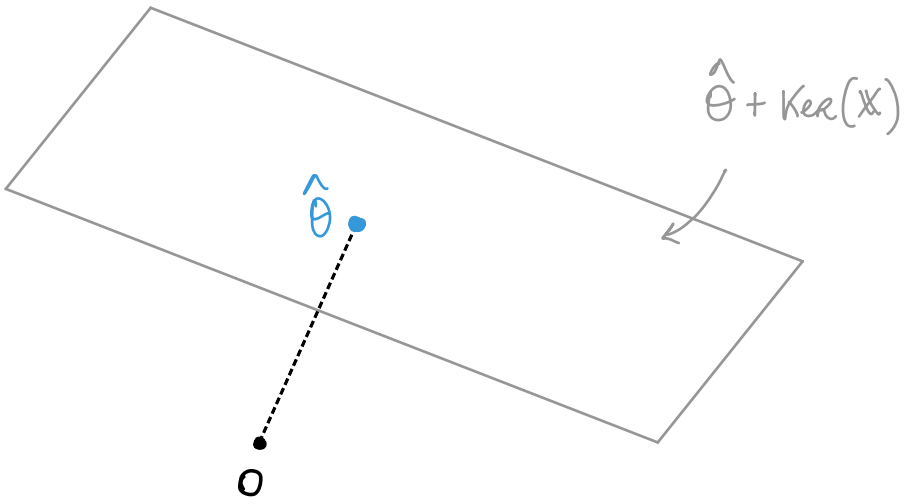


Figure 1.6: Intuition for the double descent phenomenon in linear models. When p increases, $\mathrm{Ker}(\mathbb{X})$ becomes larger, so there are more solutions to the system $Y = \mathbb{X}\theta$, so that $\|\hat{\theta}\|_2$ decreases, the bias increases with p/n and the variance decreases with p/n.

Strikingly, interpolating predictors such as those studied here have been historically overlooked, at least for noisy data. Indeed, a classical prescription is to regularize the predictor by e.g., adding a ridge penalty "$\lambda\|\cdot\|_2^2$" (which adds λI to $\mathbb{X}\mathbb{X}^\top$), and leads to non-interpolating predictors.

In conclusion, this simple setting misses the approximation/variance trade-off. But the results are the best for $\gamma = 0$. Therefore, one can wonder if there exists linear settings for which there is a true benefit to go in the over-parameterization regime? A partial answer is given in the misspecified case.

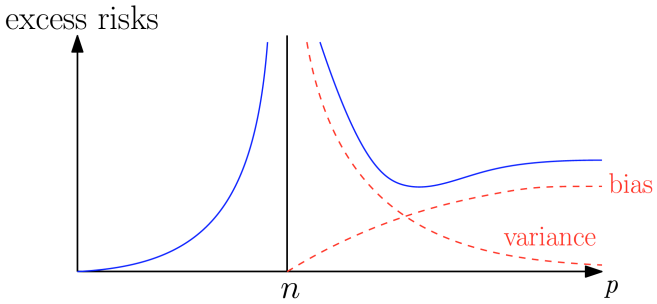


Figure 1.7: From Hastie et al. (2019), the double descent phenomenon in the linear case.

Conclusion on linear/kernel models in overparameterization regimes

Based on the previous sections, one can notice there is alignment of optimization and generalization in overparameterized linear (and kernel) models, since (S)GD converge to minimum norm interpolators having good generalization properties.

Remark 1.5 (No phenomenom when using regularization)**.** *When an extra (ridge) regularizer is used, then the double descent phenomenom is reduced (see Mei and Montanari (2019)). In particular, if the regularization parameter λ is adapted for each number of observations, then the phenomenom totally disappears (see Mei and Montanari (2019), for more details).*

1.2.3 Misspecified linear model

Suppose that the model is now

$$Y_i = X_i^\top \theta^\star + \underbrace{W_i^\top}_{\substack{\text{unobserved} \\ \text{i.i.d.}}} \zeta^\star + \varepsilon_i,$$

in which $(W_i)_i$'s are i.i.d. unobserved features that help to explain the outcome Y. In such a case, the risk is going to compare $X^\top\hat{\theta}$ to $\mathbb{E}[Y|X,W] = X^\top\theta^\star + W^\top\zeta^\star$.

$$\begin{aligned} R(\hat{\theta}) &:= \mathbb{E}\left[\left(X^\top\hat{\theta} - \mathbb{E}[Y|X,W]\right)^2 \middle| \mathbb{X}\right] \\ &= \mathbb{E}\left[\left(X^\top\hat{\theta} - \mathbb{E}[Y|X]\right)^2 \middle| \mathbb{X}\right] + \underbrace{\mathbb{E}\left[\left(\mathbb{E}[Y|X] - \mathbb{E}[Y|X,W]\right)^2 \middle| \mathbb{X}\right]}_{=:M_{\zeta^\star} \text{ approximation bias}} \text{ (by Pythagore)} \end{aligned}$$

where $M_{\zeta^\star}$ is complex in general.

When all entries of X and W are i.i.d. and isotropic,

$$\begin{aligned} M_{\zeta^\star} &= \mathbb{E}\left[\left(X^\top\theta^\star - \left(X^\top\theta^\star + W^\top\zeta^\star\right)\right)^2 \middle| \mathbb{X}\right] = \mathbb{E}\left[\left(W^\top\zeta^\star\right)^2 \middle| \mathbb{X}\right] = \|\zeta^\star\|^2 \\ &= r^2(1-\kappa) \end{aligned}$$

with

- $r^2 = \|\theta^\star\|_2^2 + \|\zeta^\star\|_2^2$ corresponds to the signal strength
- $\kappa = \|\theta^\star\|_2^2 / r^2$ is the fraction of the signal explained by the covariates X only.

Theorem 1.6. *Assume the misspecified linear model, and assume that (X,W) has i.i.d. entries with zero mean, unit variance, and a finite moment of order $8+\eta$, for some $\eta > 0$. Also assume that for all n, p, $r^2 = \|\theta^\star\|_2^2 + \|\zeta^\star\|_2^2$ and $\kappa = \|\theta^\star\|_2^2 / r^2$. Then for the min-norm least squares estimator $\hat{\theta}$, as $n, p \to \infty$, with $p/n \to \gamma$, it holds almost surely that*

$$\mathbb{E}\left[R(\hat{\theta})\right] \to \begin{cases} r^2(1-\kappa) + \left(r^2(1-\kappa) + \sigma^2\right)\frac{\gamma}{1-\gamma} & \text{for} \quad \gamma < 1 \\ r^2(1-\kappa) + r^2\kappa(1-\frac{1}{\gamma}) + \left(r^2(1-\kappa) + \sigma^2\right)\frac{1}{\gamma-1} & \text{for} \quad \gamma > 1 \end{cases}$$

In the independence setting, the dimension of the unobserved feature space does not play any role: we may equally well take it equal to ∞ for all n, p (i.e., infinitely many unobserved features). Note that

1. The first term $r^2(1-\kappa)$ is the misspecification bias (irreducible).

2. The second term equal to 0 when $\gamma < 1$ or to $r^2\kappa(1-\frac{1}{\gamma})$ is the bias.
3. The third term is the misspecification variance.
4. The last term is the variance.

By considering a polynomial decay for the approximation bias, i.e.

$$1-\kappa(\gamma) = (1+\gamma)^{-a}$$

for some $a > 0$, the global minimum of the risk is achieved in the overparameterized regime.

> In such a case, linear over-parameterized predictors are sometimes preferable to any "classical" under-parameterized model.

1.2.4 A first non-linear model with random features

Consider now a one-hidden-layer neural network, in which we optimize only the output weights, see Figure 1.8. The model is the following:

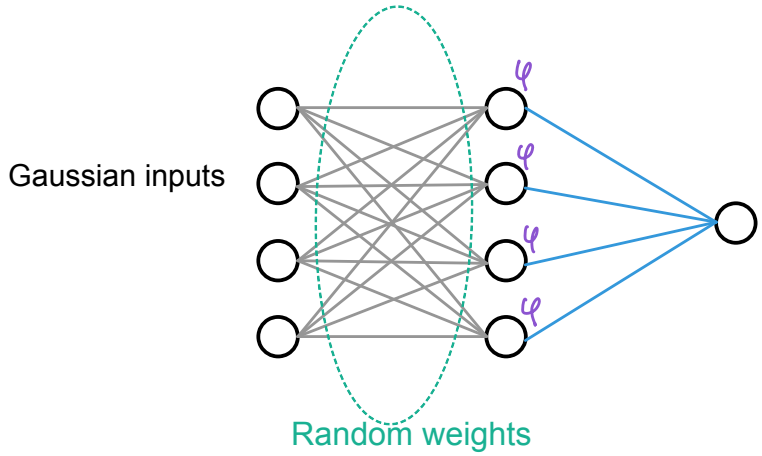


Figure 1.8: A first non-linear model, in which we optimize only the output weights (in blue). The input and the weights between the input layer and the hidden one are assumed to be Gaussian.

- assume the input vectors $X_i \in \mathbb{R}^p$ with i.i.d. centered Gaussian entries, $X_i \sim \mathcal{N}(0, I_p)$;
- assume that the weights $W \in \mathbb{R}^{q\times p}$ between the input layer and the hidden one are such that each entry of W is a random $\mathcal{N}(0, 1/d)$ variable;
- call φ the activation function used in the hidden layer, and assume it is purely non-linear, i.e.

$$\mathbb{E}\Big[\varphi(G)\Big] = \mathbb{E}\Big[G\varphi(G)\Big] = 0, \qquad \text{for} \qquad G \sim \mathcal{N}(0,1).$$

The hypothesis if purely non-linear is not common, it is satisfied for instance for $\varphi(t) = a(|t| - b)$, for $a = \sqrt{\pi}/\sqrt{\pi-2}$ and $b = \sqrt{2}/\sqrt{\pi}$.

We optimize the weights θ of the output layer in terms of quadratic risk minimization penalized by a ridge term:

$$\hat{\theta}_\lambda \in \operatorname*{argmin}_{\theta_\lambda} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \varphi(WX_i)^\top\theta\right)^2 + \lambda\|\theta\|_2^2.$$

This amounts to penalized linear regression with transformed features (the $\varphi(WX_i)$'s instead of the X_i's).

Theorem 1.7. *Assume that $|\varphi(x)| \le c_0(1+|x|)^{c_0}$ for a constant $c_0 > 0$. Also, for $G \sim N(0,1)$, assume that the standardization conditions hold: $\mathbb{E}[\varphi(G)] = 0$ and $\mathbb{E}[\varphi(G)^2] = 1$, $\mathbb{E}[G\varphi(G)] = 0$.*

Then for $\gamma := p/n > 1$, the variance satisfies, almost surely,

$$\lim_{\lambda\to 0^+}\lim_{n,p,q\to\infty} V_X(\hat{\theta}_\lambda;\theta) = \frac{\sigma^2}{\gamma-1},$$

which is precisely as in the case of linear isotropic features. Also, under a isotropic prior, namely $\mathbb{E}(\theta) = 0$, $Cov(\theta) = r^2 I_q/q$, the Bayes bias $B_X(\hat{\theta}_\lambda) := \mathbb{E}_\theta B_X(\hat{\theta}_\lambda;\theta)$ satisfies, almost surely

$$\lim_{\lambda\to 0^+}\lim_{n,p,d\to\infty} B_X(\hat{\theta}_\lambda) = \begin{cases} 0 & \text{for } \gamma < 1, \\ r^2(1-1/\gamma) & \text{for } \gamma > 1, \end{cases}$$

which is again as in the case of linear isotropic features

Note that this result is asymptotic, and heavily relies on the purely non-linear feature of the activation function that allows to retrieve standard asymptotics distribution got in the standard linear case.

When considering standard linear features, the out-of-sample risk can be decomposed in a bias and a variance terms:

$$\begin{aligned} R_{\mathbb{Z}}(\hat{\theta}) &= \mathbb{E}\left[(Z^\top\theta - Z^\top\theta^\star)^2|\mathbb{Z}\right] = \mathbb{E}\left[\|\hat{\theta}-\theta^\star\|_\Sigma^2|\mathbb{Z}\right] \\ &= \underbrace{\left\|\mathbb{E}\left[\hat{\theta}|\mathbb{Z}\right]-\theta^\star\right\|_\Sigma^2}_{\text{Bias}} + \underbrace{\operatorname{tr}\left[\operatorname{Cov}\left[\hat{\theta}|\mathbb{Z}\right]\right]}_{\text{Variance}} \end{aligned}$$

Ideas of proof for the variance. Focus on the variance term, for the regularized parameter:

$$\begin{aligned} V_{\mathbb{Z}}(\hat{\theta}_\lambda) &= \frac{\sigma^2}{n}\operatorname{tr}\left[\Sigma\frac{\mathbb{Z}^\top\mathbb{Z}}{n}\left(\lambda I + \frac{\mathbb{Z}^\top\mathbb{Z}}{n}\right)^{-2}\right] \qquad \text{for} \qquad \mathbb{Z} = \begin{pmatrix}\varphi(WX_1)^\top \\ \vdots \\ \varphi(WX_n)^\top\end{pmatrix} \\ &= \frac{\sigma^2}{n}p\sum_{i=1}^{p}\frac{1}{p}\frac{\mu_i}{(\mu_i+\lambda)^2} \qquad \text{for } (\mu_i)_i \text{ the singular values of } \mathbb{Z} \\ &\longrightarrow_{n,p\to\infty} \qquad \sigma^2\gamma\int\frac{t}{\lambda+t^2}dMP_\gamma(t) \end{aligned}$$

where MP_γ denotes the Marchenko-Pastur law of parameter γ. This is true by (Péché, 2019, Theorem 1.1), for purely non-linear activation functions (entailing $\theta_2(f) = 0$ in Péché (2019)), and by using the convergence of the spectral measure of $\frac{\mathbb{X}^\top\mathbb{X}}{n}$ towards the Marchenko-Pastur law.

We actually know an explicit form for the Stieltjes transform of this distribution, i.e.

$$m(-\lambda) = \int\frac{1}{t-\lambda}dMP_\gamma(t)$$

so we can deduce

$$\begin{aligned}
M(\lambda) &= \int \frac{t}{\lambda + t^2} dMP_\gamma(t) \\
&= \frac{\partial}{\partial \lambda} \int \frac{-t}{\lambda + t} dMP_\gamma(t) \\
&= \frac{\partial}{\partial \lambda} \left(\underbrace{\int \frac{-t-\lambda}{\lambda+t} dMP_\gamma(t)}_{=-1} + \int \frac{\lambda}{\lambda+t} dMP_\gamma(t) \right) \\
&= \frac{\partial}{\partial \lambda} \left(\lambda m(-\lambda) \right) \\
&= \frac{\partial}{\partial \lambda} \left(\frac{-(1-\gamma+\lambda) + \sqrt{(1-\gamma+\lambda)^2 + 4\gamma\lambda}}{2\gamma} \right) \\
&= -\frac{1}{2\gamma} + \frac{1}{2\gamma} \left((1-\gamma+\lambda)^2 + 4\gamma\lambda \right)^{-1/2} (1+\gamma+\lambda) \\
&\to_{\lambda \to 0^+} \frac{1}{\gamma(\gamma-1)}.
\end{aligned}$$

Finally, when $\gamma > 1$,

$$V_{\mathbb{X}}(\hat{\theta}_\lambda) \to_{\lambda \to 0^+} \gamma\sigma^2 \frac{1}{\gamma(\gamma-1)} = \frac{\sigma^2}{\gamma - 1}.$$

1.2.5 Analysis via Neuberger's theorem

This section presents the work of Caron and Chrétien (2020), in the case of Ridge functions. Let $Z_i = (X_i, Y_i) \in \mathbb{R}^{p+1}$ be observations drawn from the model

$$Y_i = f^\star(X_i) + \varepsilon_i, \qquad i = 1, \dots, n.$$

The noise vector $\varepsilon \in \mathbb{R}^n$ is assumed to be sub-Gaussian with parameter γ_ε[2].

The goal is to estimate $f^\star$ based on the observation $Z_1, \dots, Z_n$, restricting the search for a candidate in a subset $\mathscr{F}$ of functions in a Banach space $\mathscr{B}$.

To do so, we construct an empirical risk minimizer:

$$\hat{f}^{ERM} \in \operatorname*{argmin}_{f \in \mathscr{F}} \hat{R}_n(f), \qquad \text{with} \qquad \hat{R}_n(f) := \frac{1}{n} \sum_{i=1}^n \ell(Y_i - f(X_i)), \tag{1.8}$$

for some cost function ℓ such that $\ell'(0) = 0$ and ℓ'' is upper bounded by $L_{\ell'}$ (i.e. its derivative is $L_{\ell'}$-Lipschitz).

Ridge type functions Consider a statistical model of the form

$$\mathbb{E}[Y_i | X_i] = \varphi(X_i^\top \theta^\star).$$

where

- $\varphi : \mathbb{R} \to \mathbb{R}$ is assumed to be an increasing function

[2] meaning that its φ_2-norm is bounded by γ_ε, $\|\epsilon\|_{\varphi_2} := \inf\left\{ s > 0 : \mathbb{E}\left[e^{-\varepsilon/s^2} - 1 \le 1 \right] \right\} \le \gamma_\varepsilon$

- the X_i's are assumed to be isotropic (i.e. $\mathbb{E}(X_i X_i^\top) = I_d$), sub-Gaussian with parameter γ_X, and of ℓ^2-norm equal to $\sqrt{p}$ (think about Rademacher vectors, or random vectors on the sphere).

Note that such assumptions ensure that

$$\mathbb{X} = \begin{pmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{pmatrix}$$

is of full rank.

Theorem 1.8. *For some $\alpha > 0$, assume that p and n are such that*

$$C_{\gamma_X}^2 \; n < (1-\alpha)^2 p.$$

Let

$$r = 6\sqrt{C} C_{\ell''} C_{\varphi'} \gamma_\epsilon \delta^{-1} \frac{\sqrt{n}}{(1-\alpha)\sqrt{p} - C_{K_X}\sqrt{n}},$$

where C is a positive absolute constant. Assume, that ℓ and φ are such that,

$$|\ell''(Y_i - \varphi(X_i^\top z))\, \varphi'(X_i^\top z)^2 - \ell'(Y_i - \varphi(X_i^\top z))\, \varphi''(X_i^\top z)| \geq \delta > 0 \tag{1.9}$$

for all z in $\mathscr{B}_2(\theta^\star, r)$. Then,

1. *there exists a first order stationary point $\hat{\theta}$ to the optimisation problem (1.8) such that, with probability larger than or equal to*

$$1 - 2\exp(-c_{K_X}\alpha^2 p) - \exp\Big(-\frac{n}{2}\Big),$$

we have

$$\|\hat{\theta} - \theta^\star\|_2 \leq r.$$

2. *in terms of generalization error, one has for $\hat{\theta}^\circ$ the solution of minimal ℓ^2-norm of the system $\mathbb{X}\hat{\theta}^\circ = \mathbb{X}\hat{\theta}$, for any $\gamma > 0$,*

$$\mathbb{E}\left[|\varphi(X_{n+1}^\top \hat{\theta}^\circ) - \varphi(X_{n+1}^t \theta^\star)|^2 \mid X\right] \leq C_{\varphi'}^2 \left(\frac{1}{\sqrt{p}} \frac{\gamma p + (\sqrt{\gamma p} + \sqrt{n})^2}{((1-\alpha)\sqrt{p} - C_{K_X}\sqrt{n})^4} + \frac{\gamma p + (\sqrt{\gamma p} + \sqrt{n})^2}{\sqrt{p}} \|\theta^\star\|_2 \right.$$
$$\left. + \frac{2\sqrt{\gamma p}}{\sqrt{p}} + \sqrt{\gamma p}\, \frac{6\sqrt{C} C_{\ell''} C_{\varphi'} K_\epsilon n}{\delta((1-\alpha)\sqrt{p} - C_{K_X}\sqrt{n})} \right)^2 \tag{1.10}$$

with probability, at least

$$1 - 2\exp(-c_{K_X}\alpha^2 p) - \exp\Big(-\frac{p}{2}\Big) - (2+n)\exp(-c_{K_X} p) - 2\exp(-\gamma p) \tag{1.11}$$

Remark that when the number p of features tends to infinity, the radius tends to zero.

Ideas for the proof. We focus only on the first point of Theorem 1.8 (since the second point is a consequence of the first one). To establish such results, one should exhibit a stationary point for Problem (1.8), therefore one should exhibit a zero for the Jacobian associated to the ERM. This can be given by Neuberger's theorem.

Theorem 1.9 (Neuberger's theorem for ERM)**.** *Suppose that $r > 0$, that $\theta^\star \in \mathbb{R}^p$ and that the Jacobian $D\hat{R}_n(\cdot)$ is a continuous map on $B(\theta^\star, r)$ such that for each $\theta \in B(\theta^\star, r)$, there exists a vector $d \in B(0,r)$ such that*

$$\lim_{t \searrow 0} \frac{D\hat{R}_n(\theta + td) - \hat{R}_n(\theta)}{t} = -D\hat{R}_n(\theta^\star).$$

Then, there exists $u \in B(\theta^\star, r)$ such that $D\hat{R}_n(u) = 0$.

Neuberger's theorem assumes that there exists a direction d of norm less than r, such that for any point θ in a ball $B(\theta^\star, r)$, the directional derivative in θ w.r.t. d matches the gradient of $-D\hat{R}_n(\theta^\star)$ evaluated in $\theta^\star$. The question is then to find a radius r which is compatible with all these conditions.

Therefore, in order to use Neuberger's theorem, one has to show that for a particular d

$$\nabla^2 \hat{R}_n(\theta) d = -\nabla \hat{R}_n(\theta^\star). \tag{1.12}$$

1. (Gradient and Hessian) Since the loss is twice differentiable the empirical risk $\hat{R}_n$ is itself twide differentiable. The gradient is given by

$$\begin{aligned}\nabla \hat{R}_n(\theta) &= -\frac{1}{n}\sum_{i=1}^n \ell'(Y_i - \varphi(X_i^\top \theta))\varphi'(X_i^\top \theta) X_i \\ &= -\frac{1}{n}\mathbb{X}^\top \mathrm{diag}(\nu)\ell'(\varepsilon),\end{aligned}$$

   where $\ell'(\varepsilon)$ has to be understood componentwise, and $\nu_i := \varphi'(X_i^\top \theta)$ for all $i = 1, \dots, n$.

   The Hessian matrix is given by

$$\nabla^2 \hat{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \Big(\ell''(Y_i - \varphi(X_i^\top \theta))\varphi'(X_i^\top \theta)^2 - \ell'(Y_i - \varphi(X_i^\top \theta))\varphi''(X_i^\top \theta)\Big) X_i X_i^\top$$

   and can be actually rewritten as

$$\nabla^2 \hat{R}_n(\theta) = \frac{1}{n}\mathbb{X}^\top \mathrm{diag}(\mu)\mathbb{X} \tag{1.13}$$

   where $\mathrm{diag}(\mu) \in \mathbb{R}^{n \times n}$ is a diagonal matrix, with diagonal entries

$$\mu_i = \big(\ell''(Y_i - \varphi(X_i^\top \theta))\varphi'(X_i^\top \theta)^2 - \ell'(Y_i - \varphi(X_i^\top \theta))\varphi''(X_i^\top \theta)\big)$$

2. (Solving Neuberger's equation) One has to solve the Neuberger's equation

$$\frac{1}{n}\mathbb{X}^\top \mathrm{diag}(\mu)\mathbb{X} d = \frac{1}{n}\mathbb{X}^\top \mathrm{diag}(\nu)\ell'(\varepsilon),$$

   which can be solved by finding the least norm solution of the interpolation sub-problem

$$\mathrm{diag}(\mu)\mathbb{X} d = \mathrm{diag}(\nu)\ell'(\varepsilon),$$

   i.e.

$$d = \mathbb{X}^\dagger \mathrm{diag}(\mu)^{-1}\mathrm{diag}(\nu)\ell'(\varepsilon).$$

3. (Bounding $\|d\|_2$) Then one should ensure that $\|d\| \le r$. Given the SVD of $\mathbb{X} = U\Sigma V^\top$, one gets

$$d = V\Sigma^{-1}U^\top \mathrm{diag}(\mu^{-1})\mathrm{diag}(\nu)\ell'(\varepsilon),$$

so

$$\|d\|_2 = \left\|V\Sigma^{-1}U^\top \mathrm{diag}(\mu^{-1})\mathrm{diag}(\nu)\ell'(\varepsilon)\right\|_2 \le \frac{\left\|U^\top \mathrm{diag}(\mu^{-1})\mathrm{diag}(\nu)\ell'(\varepsilon)\right\|_2}{s_{\min}(\mathbb{X})}.$$

   (a) (Bounding $\left\|U^\top \mathrm{diag}(\mu^{-1})\mathrm{diag}(\nu)\ell'(\varepsilon)\right\|_2$). One can show that $U^\top \mathrm{diag}(\mu^{-1})\mathrm{diag}(\nu)\ell'(\varepsilon)$ is a subGaussian vector with variance proxy

$$C \max_{1\le i\le n} \left\|\frac{\nu_i}{\mu_i}\ell'(\varepsilon_i)\right\|_{\psi_2}^2,$$

   so that with probability higher than $1-e^{-u^2/2}$,

$$\left\|U^\top \mathrm{diag}(\mu^{-1})\mathrm{diag}(\nu)\ell'(\varepsilon)\right\|_2 \lesssim \max_{1\le i\le n} \left\|\frac{\nu_i}{\mu_i}\ell'(\varepsilon_i)\right\|_{\psi_2}^2 (\sqrt{p}+u).$$

   We then deduce that $\frac{\nu_i}{\mu_i}\ell'(\varepsilon_i)$ is a sugGaussian random variable with variance proxy

$$\left\|\frac{\nu_i}{\mu_i}\ell'(\varepsilon_i)\right\|_{\psi_2}^2 \le \frac{\nu_i}{\mu_i}\|\varepsilon_i\|_{\psi_2}^2 \tag{1.14}$$

$$\le C_{\ell''}\frac{\max_i \nu_i}{\max_i \mu_i}K_\varepsilon \tag{1.15}$$

$$\le C_{\ell''}\frac{C'_\varphi}{\delta}\gamma_\varepsilon \tag{1.16}$$

   where we used the boundedness of φ' in the last inequality.
   Overall,

$$\left\|U^\top \mathrm{diag}(\mu^{-1})\mathrm{diag}(\nu)\ell'(\varepsilon)\right\|_2 \lesssim \frac{C_{\ell''}C_{\varphi'}\gamma_\varepsilon}{\delta}(\sqrt{p}+u),$$

   with probability $1-\left(\exp(-u^2/2)+2n\exp\left(-\frac{t^2}{C_{\ell''}^2\gamma_\varepsilon^2}\right)\right)$.

   (b) (Bounding $s_{\min}(\mathbb{X})$). By sub-Gaussianity of the covariates, one has with probability larger than $1-2\exp(-c_{\gamma_X}\alpha^2 n)$,

$$s_{\min}(\mathbb{X}) \ge \sqrt{p} - (\alpha + C_{\gamma_X})\sqrt{n})\sqrt{n}.$$

4. (Finishing) Putting all this together, one gets,

$$\|d\|_2 \gtrsim \frac{\sqrt{n}}{(1-\alpha)\sqrt{p} - C_{K_X}\sqrt{n}} \tag{1.17}$$

   with probability larger than $1-2\exp(-c_{\gamma_X}\alpha^2 n)-\exp(-u^2/2)-2n\exp\left(-\frac{t^2}{C_{\ell''}^2\gamma_\varepsilon^2}\right)$. Choosing t and u of the order of $\sqrt{\log(n)}$ completes the proof.

1.2.6 Overparametrization/interpolation in neural network

Why overparameterizing in neural networks? It is often observed that for neural networks, depth efficiently helps to extract features in the dataset. Recent studies found that increasing both depth and width of a shallow model leads to very nice continuous limits, where PDE tools can be put in work. Besides, on the numerical side, one could argue that increasing the number of parameters could make harder the optimization/training of such complex architectures. However, networks with wide layers (larger than the sample size) can be shown to have no spurious minimizers Nguyen et al. (2018); Nguyen (2019) (i.e. no local optima with bad generalization properties).

Bad consequences of overparametrizartion in neural networks? Beware, when training a NN with layers not wide enough, overparametrization usually entails existence of many local minimizers with potentially different statisticial performances. Common practice advises to run stochastic gradient algorithm with random initialization and converges to parameters with very good practical prediction accuracy.Why is this simple approach actually often working? The goal of current research is to resolve these paradoxes.

Some empirical observations are reported in Huang et al. (2020) on the importance of "being flat". Flat minimizers (with a bad conditioned Hessian, and therefore with a flat attraction basin) are easier to reach and have better generalisation properties. Flatness seems to be nice for both generalization properties, and convergence of the used algorithms.

Bibliography

Bartlett, P. L., Montanari, A., and Rakhlin, A. (2021). Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177*.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Caron, E. and Chrétien, S. (2020). A finite sample analysis of the benign overfitting phenomenon for ridge function estimation. *arXiv preprint arXiv:2007.12882*.

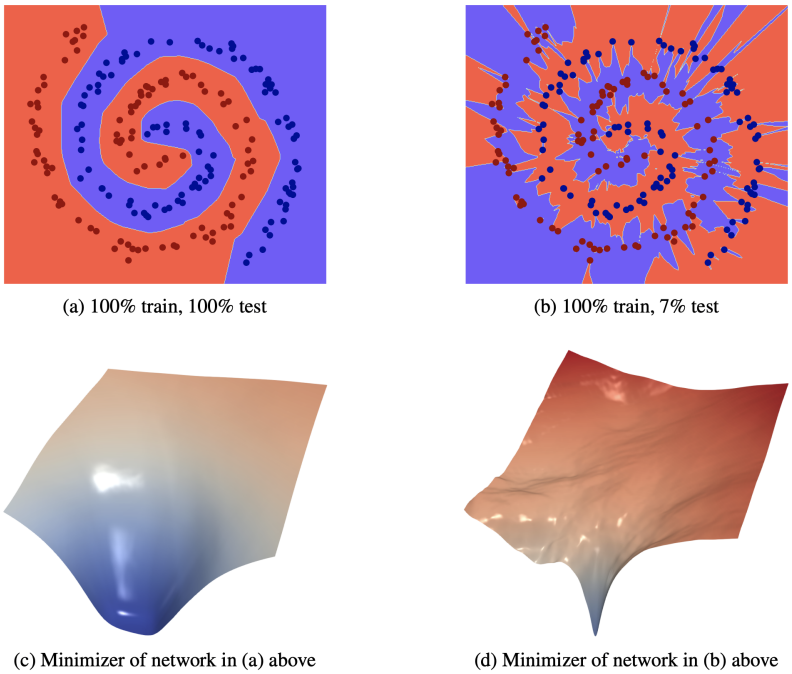
Chizat, L. and Bach, F. (2020). Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In Abernethy, J. and Agarwal, S., editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1305–1338. PMLR.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*.

Huang, W. R., Emam, Z., Goldblum, M., Fowl, L., Terry, J. K., Huang, F., and Goldstein, T. (2020). Understanding generalization through visualizations.

jlewk (https://math.stackexchange.com/users/484640/jlewk). Simple(r) way to derive the expectation of an inverse wishart? Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/3994137 (version: 2021-01-21).

Lyu, K. and Li, J. (2019). Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*.



(a) 100% train, 100% test

(b) 100% train, 7% test

(c) Minimizer of network in (a) above

(d) Minimizer of network in (b) above

Figure 1.9: From Huang et al. (2020). Top: Decision boundaries of two networks with different parameters. Network (a) generalizes well. Network (b) generalizes poorly (perfect train accuracy, bad test accuracy). The flatness and large volume of (a) make it likely to be found by SGD, while the sharpness and tiny volume of (b) make this minimizer unlikely. Red and blue dots correspond to the training data. See Bottom: A slice through the loss landscapes around these minima reveals sharpness/flatness.

Mei, S. and Montanari, A. (2019). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*.

Nguyen, Q. (2019). On connected sublevel sets in deep learning. In *International Conference on Machine Learning*, pages 4790–4799. PMLR.

Nguyen, Q., Mukkamala, M. C., and Hein, M. (2018). On the loss landscape of a class of deep neural networks with no bad local valleys. *arXiv preprint arXiv:1809.10749*.

Péché, S. (2019). A note on the pennington-worah distribution. *Electronic Communications in Probability*, 24:1–7.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. (2018). The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878.

Chapter 2

Interpolation in non-parametric learning

Contents

2.1 The nearest neighbour

The nearest neighbour (1-NN) rule Cover and Hart (1967) is the most simple classifier/regressor that always interpolates the training data by definition. Its training error is indeed always 0.

Warming: a toy classification setting Let $\mathscr{D}_n = \{(X_i, Y_i) \in [0,1]^d \times \{0,1\}, i = 1,\dots,n\}$ be a training set and assume that the X_i are uniformly distributed on $[0,1]^d$ and that for all $x \in [0,1]^d$,

$$\eta(x) = \mathbb{P}(Y = 1 | X = x) = \alpha > 1/2.$$

1. In all generality, the Bayes classifier $f^\star$ is defined as

$$f^\star(x) = \begin{cases} 0 & \text{if } \eta(x) \le 1/2 \\ 1 & \text{if } \eta(x) > 1/2 \end{cases}$$

In the particular case described above, give the expression of Bayes classifier.

Answer: In this case, the Bayes classifier is given by $f^\star(x) = 1$ for all $x \in [0,1]^d$.

2. Consider any classifier $f_n : [0,1]^d \to \{0,1\}$. Prove that its error in x satisfies

$$\mathbb{P}[f_n(X) \neq Y | X = x] = \alpha - (2\alpha - 1)\mathbb{E}[f_n(X)|X = x].$$

What is the value of the previous quantity for the Bayes classifier $f^\star$?

Answer: We have

$$\begin{aligned}
\mathbb{P}[f_n(X) \neq Y | X = x] &= \mathbb{P}[f_n(X) = 1, Y = 0 | X = x] + \mathbb{P}[f_n(X) = 0, Y = 1 | X = x] \\
&= \mathbb{P}[f_n(X) = 1 | X = x]\mathbb{P}[Y = 0 | X = x] \\
&\quad + \mathbb{P}[f_n(X) = 0 | X = x]\mathbb{P}[Y = 1 | X = x] \\
&= \mathbb{E}[f_n(X)|X = x](1 - \alpha) + (1 - \mathbb{E}[f_n(X)|X = x])\alpha \\
&= \alpha - (2\alpha - 1)\mathbb{E}[f_n(X)|X = x].
\end{aligned}$$

For the Bayes classifier $f^\star$, according to the previous question

$$\mathbb{P}[f^\star(X) \neq Y | X = x] = \alpha - (2\alpha - 1) = 1 - \alpha.$$

3. Now consider the 1 nearest neighbor estimate $f_{1,n}$. Let $B_i(x)$ be a Bernoulli variable equal to 1 if the i-th observation is the nearest neighbor of x and 0 otherwise. Write $f_{1,n}(x)$ as a sum of random variables. What is the value of $\sum_{i=1}^n B_i(x)$?

Answer: By definition, $f_{1,n}(x)$ takes value 1 if the label of the nearest neighbor of x is 1 and zero otherwise. Thus,

$$f_{1,n}(x) = \sum_{i=1}^n B_i(x) Y_i,$$

where, in the sum, only one term is nonzero. Note that exactly one $B_i(x)$ is nonzero and equal to one. Therefore $\sum_{i=1}^n B_i(x) = 1$.

4. Compute $\mathbb{E}[f_{1,n}(x)]$ and $\mathbb{P}[f_{1,n}(x) \neq Y]$.

Answer: According to the previous question, we have

$$\begin{aligned}
\mathbb{E}[f_{1,n}(x)] &= \sum_{i=1}^n \mathbb{E}[B_i(x) Y_i] \\
&= \sum_{i=1}^n \mathbb{E}[\mathbb{E}[B_i(x) Y_i | X_i]] \\
&= \sum_{i=1}^n \mathbb{E}[\mathbb{E}[B_i(x)|X_i] \underbrace{\mathbb{E}[Y_i|X_i]}_{=\alpha}] \\
&= \alpha \sum_{i=1}^n \mathbb{E}[B_i(x)] \\
&= \alpha \underbrace{\mathbb{E}[\sum_{i=1}^n B_i(x)]}_{=1} \\
&= \alpha.
\end{aligned}$$

Thus, using question 2, we get

$$\mathbb{P}[f_{1,n}(X) \neq Y | X = x] = \alpha - (2\alpha - 1)\alpha = 2\alpha(1 - \alpha).$$

5. We say that a classifier f_n is consistent if its risk $R(f_n) = \mathbb{P}[f_n(X) \neq Y]$ tends to the risk of the Bayes classifier $R(f^\star) = \mathbb{P}[f^\star(X) \neq Y]$. Prove that the nearest neighbor estimate is not consistent.

   Answer: Since $2\alpha > 1$,

   $$\mathbb{P}[f_{1,n}(X) \neq Y | X = x] - \mathbb{P}[f^\star(X) \neq Y | X = x] = \alpha(1-\alpha) > 0.$$

   Taking the expectation of the previous inequality with respect to x leads to

   $$R(f_{1,n}) - R(f^\star) \geq \alpha(1-\alpha) > 0.$$

   If $1/2 < \alpha < 1$, the 1 nearest neighbor estimate is not consistent since there is a gap of $\alpha(1-\alpha)$ between its risk and the risk of the Bayes classifier. This gap collapses only when $\alpha = 1$, i.e. in the noiseless case, showing the consistency of the 1-NN neighbour in such a case.

Regression Assume that the model is given by

$$Y_i = f^\star(X_i) + \varepsilon_i,$$

with $\mathbb{E}[Y_i^2] < \infty$ so that the regression function $f^\star(x) = \mathbb{E}[Y|X = x]$ achieves the minimal quadratic risk over all Borel measurable functions $g : \mathbb{R}^p \to \mathbb{R}$, that is

$$\mathbb{E}\left[\left|Y - f^\star(X)\right|^2\right] = \inf_f \mathbb{E}\left[\left|Y - f(X)\right|^2\right]$$

The nearest neighbour estimate is defined by

$$\hat{f}_n^{1\mathrm{NN}}(x) = Y_{(1)}(x)$$

where $(X_{(1)}(x), Y_{(1)}(x)), \ldots, (X_{(n)}(x), Y_{(n)}(x))$ is a reordering of the data according to the increasing values of $\|X_i - x\|_2$.

Theorem 2.1 (Biau and Devroye (2015), Theorem 9.1)**.** *Assume that* $\mathbb{E}[Y^2] < \infty$*. Then, the 1-nearest neighbour* $\hat{f}_n^{1\mathrm{NN}}$ *predictor satisfies*

$$\mathbb{E}\left[\left|\hat{f}_n^{1\mathrm{NN}}(X) - f^\star(X)\right|^2\right] \longrightarrow_{n\to+\infty} \underbrace{\mathbb{E}\left[\left|Y - f^\star(X)\right|^2\right]}_{\textit{residual variance}}.$$

The proof can be found in Chapter 9 of Biau and Devroye (2015).

This convergence is *universal*, in the sense that it happens for any distribution of (X, Y) with $\mathbb{E}\left[Y^2\right] < \infty$.

When considering the nearest neigbour estimate, the mean integrated squared error $\mathbb{E}\left[\left|\hat{f}_n^{1\mathrm{NN}}(X) - f^\star(X)\right|^2\right]$ converges to the residual variance

$$\mathbb{E}\left[\left|Y - f^\star(X)\right|^2\right] = \mathbb{E}\left[Y^2\right] - \mathbb{E}\left[\left(f^\star(X)\right)^2\right].$$

This residual variance is zero if and only if $Y = f^\star(X)$ with probability one, i.e. in the noiseless case.

Therefore, the 1-NN predictor is inconsistent, apart from the noiseless setting. The 1-NN is indeed very sensitive to noise.

2.2 Interpolating kernel estimates

The 1-NN neighbour estimate is a particular local-means estimate, that can be rewritten as

$$\hat{f}_n(x) = \sum_{i=1}^{n} W_{in} Y_i$$

given the training data points $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$.

More general local-means (Nadaraya-Watson) estimators can be constructed with the form

$$\hat{f}_n(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}.$$

K is the kernel function (not to be confused with kernel machines),and $h > 0$ can be called the bandwidth. In Devroye et al. (1998), the authors choose the so-called "Hilbert" kernel:

$$K(x) = \frac{1}{\|x\|^p}$$

with p the input dimension. Remark that $K(u) \xrightarrow[\|u\|\to 0^+]{} +\infty$ so that $\hat{f}_n(x) \xrightarrow[x\to X_i]{} Y_i$, so that there is interpolation of the training data. Note that there is no usual window parameter needed it cancels out in the numerator and denominator. They show in particular that for almost all point x such that $\varphi_X(x) > 0$ with φ_X the density of the input X and for bounded output Y,

$$\hat{f}_n(x) \xrightarrow[n\to+\infty]{} f^\star(x)$$

so that this predictor is universally consistent, concluding that the interpolation "introduces unnecessary noise, but at the same time, except for immediate regions around the data points, the estimate feels and behaves like a true kernel smoother" (1998).

In Belkin et al. (2019), the authors choose a singular kernel such as

$$K(x) = \frac{1}{\|x\|^\alpha} \mathbb{1}_{\|x\|\le 1}, \qquad \text{with } \alpha \quad \text{to be chosen,}$$

and manage to obtain convergence rates in this flavour: if $0 < \alpha < p/2$ and $h \sim \left(\frac{1}{n}\right)^{\frac{1}{2\beta+p}}$ when $f^\star$ is a β-Hölder regular function

$$\forall x, \qquad \mathbb{E}\left[\left(\hat{f}_n(x) - f^\star(x)\right)^2\right] \lesssim \left(\frac{1}{n}\right)^{\frac{2\beta}{2\beta+d}}.$$

By tuning the kernel bandwidth, the influence of the interpolation can be very limited and very localized around the training points. Anywhere else, the estimated function remains “smooth", see Figure 2.1. Indeed, the NW estimator with a singular kernel can be seen as a general smooth estimate that is given by averaging the data in a neighbourhood of size h, to which we add small “spikes” at the data points, allowing interpolation. As pointed in Bartlett et al. (2021), any estimator $\hat{f}$ could be turned into an interpolating one $\hat{f}^{\text{int}} = \hat{f} + \Delta$, where $\Delta(X_i) = Y_i - \hat{f}(X_i)$ but $\|\Delta\|_{L^2(\mathbb{P})} = o(1)$. This has been observed empirically by Wyner et al. (2017), in what they called “spiked-smooth" estimates.

2.3 What about random forests?

In this section, we discuss about recent results in a joint work with Ludovic Arnould (Sorbonne Université) and Erwan Scornet (Ecole Polytechnique).

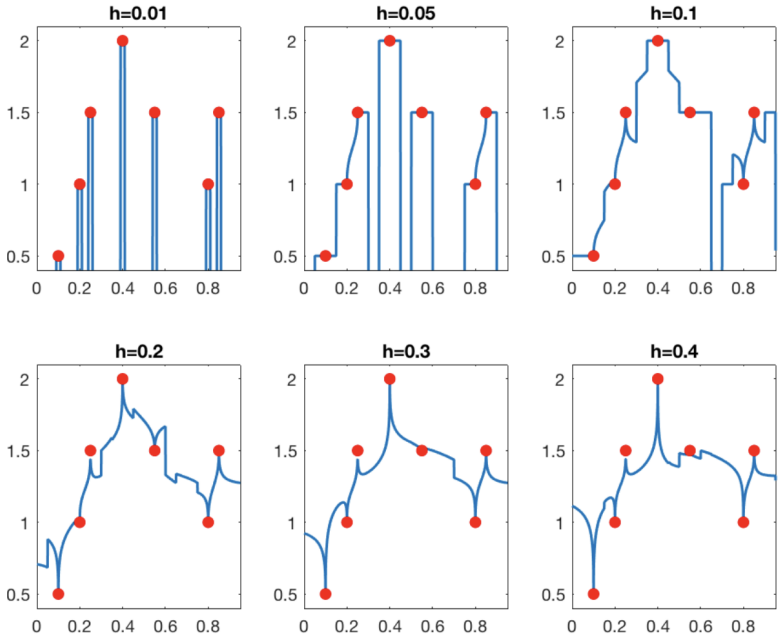


Figure 2.1: From Belkin et al. (2019), choosing $\alpha = 0.49$. The influence of interpolation remains very localized around the training points.

2.3.1 Setting

We assume to be given a *training set* $\mathcal{D}_n := ((X_1, Y_1), ..., (X_n, Y_n))$, composed of i.i.d. copies of the generic random variable (X, Y), where $X \in [0,1]^d$ is the input and $Y \in \mathbb{R}$ is the output. The underlying model is assumed to satisfy $Y = f^\star(X) + \varepsilon$, where $f^\star(x) = \mathbb{E}\big[Y|X = x\big]$ is the regression function and ε is a random centered noise of variance $\sigma^2 < \infty$. Given an input vector x, the goal is then to predict the associated square integrable random response by estimating $f^\star(x)$.

2.3.2 Preliminary on random forests

RF predictor A Random Forest (RF) is a predictor consisting of a collection of M randomized trees (see Breiman et al., 1984, for details about decision trees) that can be seen weak learners.

Trees are predictors

(a) that are built by partitioning the feature space into hyperrectangles (along the features axes);

(b) which, for a data point x, predict by "averaging" the labels $(Y_i)_i$ of the training points $(X_i)_s$ falling in the same partition cell as x.

Note that a fully-grown tree is "spiky", in the sense that there is no smoothing mechanism: everywhere in space, the prediction relies only on one data point.

RF aggregates the prediction of several trees. To improve prediction with aggregation, the constructed trees need to be as uncorrelated as possible. To this end, in RF,

1. trees are usually trained on different *bootstrap* samples from the original training sample.

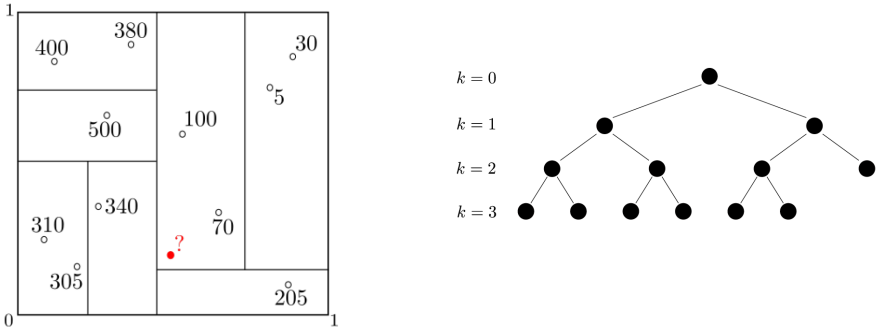


Figure 2.2: A partition made of hyperrectangles can always be coded by a binary tree. The prediction of a tree for a new data point is made by a majority vote in a classification setting, or by averaging in a regression setting, so that here, the prediction will be 85.

2. each time that a tree creates new partition cells, it will only cut over a random subset of features for splits.

Non-adaptive RF To build a forest, we generate $M \in \mathbb{N}^\star$ independent random variables $(\theta_1, \ldots, \theta_M)$, distributed as a generic random variable Θ, independent of $\mathscr{D}_n$. In our setting, θ_m actually represents the successive random splitting directions and the resampling data mechanism in the m-th tree. The predicted value at the query point x given by the m-th tree is defined as

$$f_n(x, \theta_j) = \sum_{i=1}^{n} \frac{\mathbb{1}_{X_i \in A_n(x,\theta_j)} Y_i}{N_n(x, \theta_j)} \mathbb{1}_{N_n(x,\theta_j)>0}$$

where $A_n(x, \theta_j)$ is the cell containing x and $N_n(x, \theta_j)$ is the number of points falling into $A_n(x, \theta_j)$. The (finite) forest estimate then results from the aggregation of M trees:

$$f_{M,n}(x, \Theta_M) = \frac{1}{M} \sum_{m=1}^{M} f_n(x, \theta_m),$$

where $\Theta_M := (\theta_1, ..., \theta_M)$. By making the number M of trees grows towards infinity, we can consider instead the *infinite* forest estimate, which has also played an important role in the theoretical understanding of forests:

$$f_{\infty,n}(x) = \mathbb{E}_\theta [f_n(x, \theta)],$$

where $\mathbb{E}_\theta$ denotes the expectation w.r.t. θ, conditional on $\mathscr{D}_n$.

Centered RF Centered Random Forests (Biau (2012)) are ensemble methods that are said to be non-adaptive since trees are built independently of the data: at each step of a centered tree construction, a feature is uniformly chosen among all possible d features and the split along the chosen feature is made at the center of the current cell. Then trees are aggregated to produce a CRF.

Choosing the depth of the order of $\log_2(n)$ characterizes another type of interpolation regime. To see this, consider a centered tree of depth k, whose leaves are denoted $L_1, \ldots, L_{2^k}$. The number of points falling into the leaf L_i is denoted $N_n(L_i)$. If X is uniformly distributed over $[0,1]^d$, then by construction, for a given leaf L_i,

$$\mathbb{P}(X \in L_i) = \frac{1}{2^k} \quad \text{and} \quad \mathbb{E}(N_n(L_i)) = \frac{n}{2^k}. \tag{2.1}$$

Definition 2.2 (Mean interpolation regime)**.** *A CRF $f_{M,n}^{\text{CRF}}$ satisfies the* mean interpolation regime *when each tree of $f_{M,n}^{\text{CRF}}$ has at least n leaves ($k \geq \log_2(n)$).*

2.3.3 Centered forests: watch the empty cells out

Proposition 2.3. *Suppose that $\mathbb{E}\left[f^\star(X)^2\right] > 0$. Then the infinite Centered Random Forest of depth $k_n \geq \lfloor \log_2 n \rfloor$ is inconsistent.*

The non-consistency of the CRF stems from the fact that the probability for a random point X to fall in an empty cell does not converge to zero, i.e. for a random tree θ,

$$\mathbb{P}\left(N_n(X,\theta) = 0 | X\right) \not\xrightarrow[n\to\infty]{} 0.$$

Indeed, denoting $\mathscr{E}$ the event "$N_n(X,\theta) = 0$" (or equivalently, "X falls into a non-empty leaf"), in such a case the infinite CRF outputs 0. Setting $\bar{m}_{n,\infty}(X) = \mathbb{E}\left[m_{n,\infty}(X)|X, X_1, ..., X_n\right]$,

$$\mathscr{R}(f_{n,\infty}^{\text{CRF}}(X)) = \mathbb{E}\left[\left(f_{n,\infty}^{\text{CRF}}(X) - f^\star(X)\right)^2)\right] \tag{2.2}$$

$$= \mathbb{E}_{X,\mathscr{D}_n}\left[\left(\sum_{i=1}^n W_{n,i}^\infty(X) f^\star(X_i) + W_{n,i}^\infty(X)\varepsilon_i - f^\star(X)\right)^2\right] \tag{2.3}$$

$$\geq \mathbb{E}_{X,\mathscr{D}_n}\left[\left(\sum_{i=1}^n W_{n,i}^\infty(X) f^\star(X_i) - f^\star(X)\right)^2\right] \tag{2.4}$$

$$\geq \mathbb{E}_{X,\mathscr{D}_n}\left[\left(\sum_{i=1}^n \mathbb{E}_\theta\left[\frac{\mathbb{1}_{X_i\in A_n(x,\theta)}}{N_n(x,\theta)}\mathbb{1}_{N_n(x,\theta)>0}\right] f^\star(X_i) - f^\star(X)\right)^2\right] \tag{2.5}$$

$$= \mathbb{E}_{X,\mathscr{D}_n}\left[\left(\sum_{i=1}^n \mathbb{E}_\theta\left[\frac{\mathbb{1}_{X_i\in A_n(X,\theta)}}{N_n(X,\theta)}\left(f^\star(X_i) - f^\star(X)\right)\mathbb{1}_{N_n(X,\theta)>0}\right]\right)^2\right] + \mathbb{E}\left[(f^\star)^2(X)\mathbb{1}_{N_n(X,\theta)>0}\right] \tag{2.6}$$

$$\geq \mathbb{E}\left[(f^\star)^2(X)\mathbb{1}_{N_n(X,\theta)>0}\right] = \mathbb{E}_X\left[(f^\star)^2(X)\mathbb{P}_{\mathscr{D}_n,\theta}\left(N_n(X,\theta) = 0|X\right)\right] \tag{2.7}$$

When the tree contains $\alpha_n n$ leaves with $\alpha_n \geq 1$

$$\mathbb{P}_{\mathscr{D}_n,\theta}\left(N_n(X,\theta) = 0|X\right) = \left(1 - \frac{1}{\alpha_n n}\right)^n \not\xrightarrow[n\to\infty]{} 0.$$

This emphasizes the poor generalisation capacities of the interpolating CRF (under any interpolating regime). Since controlling empty cells seems crucial for the consistency, this motivates the introduction of a modified version of a CRF that does not take into account the empty cells to make the final prediction:

Wiser CRF: aggregating only non-empty cells A finite CRF aggregating only non-empty cells is given by

$$f_{M,n}^{\text{wCRF}}(x,\Theta_M) = \frac{1}{\Lambda_n(x,\Theta_M)}\sum_{m=1}^M f_n(x,\theta_m)\mathbb{1}_{N_n(x,\theta_m)>0}.$$

with $\Lambda_n(x,\Theta_M) := |\{m : N_n(x,\theta_m) > 0\}|$.

An infinite version of a centered CRF that would only aggregate non-empty cells is given by

$$f_{\infty,n}^{\text{wCRF}}(x) = \mathbb{E}_\theta\left[f_n(x,\theta)|N_n(x,\theta) > 0\right] \tag{2.8}$$

$$= \mathbb{E}_\theta\left[\frac{f_n(x,\theta)\mathbb{1}_{N_n(x,\theta)>0}}{\mathbb{P}_\theta\left(N_n(x,\theta) > 0\right)}\right]. \tag{2.9}$$

Theorem 2.4. *Suppose that $f^\star$ is bounded and has bounded partial derivatives. Then, the infinite wiser-CRF of depth $k = \log_2 n$ is consistent in a noiseless setting. More precisely, if $\sigma = 0$,*

$$\mathbb{E}\left[\left(f_{\infty,n}^{\text{wCRF}}(X) - m(X)\right)^2\right] \leq 2d \sum_{j=1}^{d} ||\partial m_j||_\infty^2 n^{\log_2\left(1-\frac{1}{2d}\right)} + 2n^{-\frac{1}{\log(2)}}$$

The risk of the wiser-CRF estimator can be decomposed as the sum of bias and variance. In a noiseless setting, only the bias need to be controlled. To do so, one should control what happens on the event where the wCRF arbitrarily outputs 0, this corresponds to the case where "$\mathbb{P}_\theta\left[N_n(X,\theta) > 0\right] = 0$", i.e. when "$\mathbb{P}_\theta\left[N_n(X,\theta) = 0\right] = 1$". In such a case, we manage to prove that

$$\mathbb{P}_{X,\mathscr{D}_n}\Big(\mathbb{P}_\theta\big(N_n(X,\theta) = 0\big) = 1\Big) \xrightarrow[n\to\infty]{} 0$$

i.e.

$$\mathbb{P}_{X,\mathscr{D}_n}\Big(\forall\theta(\omega), \quad N_n(X,\theta(\omega)) = 0\Big) \xrightarrow[n\to\infty]{} 0.$$

Even though we proved consistency in a noiseless framework, we firmly believe that it holds in a noisy setting and that we could not demonstrate the result because of difficulties in the proof.

2.3.4 Kernel RF

In order to reach consistency in a noisy scenario, we now focus on the Kernel RF. Instead of averaging the predictions of all centered trees, the construction of a kernel RF (KeRF) is performed by growing all centered trees and then averaging along all points contained in the leaves in which x falls, i.e.

$$f_{M,n}^{\text{KeRF}}(x,\Theta_M) := \frac{\sum_{i=1}^n Y_i \sum_{m=1}^M \mathbb{1}_{X_i \in A_n(x,\theta_m)}}{\sum_{i=1}^n \sum_{m=1}^M \mathbb{1}_{X_i \in A_n(x,\theta_m)}}.$$

Letting $K_{M,n}$ be the connection function of the finite forest with M trees defined by

$$K_{M,n}(x,z) := \frac{1}{M}\sum_{m=1}^M \mathbb{1}_{z \in A_n(x,\theta_m)},$$

Scornet (2016) shows that the KeRF can be rewritten as

$$f_{M,n}^{\text{KeRF}}(x,\Theta_M) = \frac{\sum_{i=1}^n Y_i K_{M,n}(x,X_i)}{\sum_{i=1}^n K_{M,n}(x,X_i)},$$

hence the name of kernel RF. In addition, it is shown that

$$\lim_{M\to\infty} K_{M,n}(x,z) := K_n(x,z),$$

where $K_n(x,z) = \mathbb{P}_\theta\left[z \in A_n(x,\theta)\right]$ which can be seen as the empirical probability for x and z to be in the same cell w.r.t. a tree built according to θ. Consequently, for all $x \in [0,1]^d$, the infinite KeRF reads as

$$f_{\infty,n}^{\text{KeRF}}(x) = \frac{\sum_{i=1}^n Y_i K_n(x,X_i)}{\sum_{i=1}^n K_n(x,X_i)}.$$

Note that the mean interpolation regime is met for centered trees, and therefore for KeRF, as soon as $k_n \geq \log_2 n$.

Theorem 2.5. *Assume that $f^\star$ is Lipschitz continuous and that the additive noise ε is a centered Gaussian variable with a finite variance σ^2. Then, the risk of the infinite centered KeRF of depth $k_n = \lfloor \log_2(n) \rfloor$ verifies, for all $n \geq 2$,*

$$\mathscr{R}(f_{\infty,n}^{\mathrm{KeRF}}) \leq C_d \log(n)^{-(d-11)/6},$$

with $C_d > 0$ a constant depending on $\sigma, d, \|f^\star\|_\infty$.

If a “mean" overfitting regime is benign for the consistency of KeRF, it seems to be nonetheless malignant for the convergence rate.

2.3.5 RF & exact interpolation

Definition of AdaCRF Since consistency has been analyzed so far in the mean interpolation regime, we introduce a new adaptive tree which reaches the strict interpolation regime. This so-called *adaptive centered tree* is a modified version of a centered tree, built by taking into account the positions of the X_i's, and thereby reduces the number of empty leaves. It is recursively grown:

1. (splitting direction) at each node, a feature is uniformly chosen among the set of all *separable d* features (a feature is separable if cutting this feature produces two non-empty cells).

   Note that if there are more than one point in the current node and none of the feature separates them, the splitting direction is uniformly chosen among all the separable features of the previous cut.

2. (split) the split is made in the middle of the current node along the chosen feature.

3. (stop) The construction stops when all leaves contain 0 or 1 observation.

The *Adaptive Centered RF* (AdaCRF) results from a specific aggregation of such trees: for a given point x, the final prediction of the RF is given by averaging along all the trees for which x falls into a non-empty leaf.

Interpolation

Lemma 2.6 (Depth of an adaptive centered tree)**.** *For all $\alpha \in [0,1)$,*

$$k_n(X) \in [\log(n) \pm \log^{1-\alpha}(n)] \xrightarrow[n\to\infty]{} 1.$$

Lemma 2.6 states that the asymptotic behavior of $k_n(X)$ is equivalent to $\log n$ up to a negligible factor. The $\log(n)$ equivalent matches the condition for the mean interpolation regime in the case of CRF. Therefore, while AdaCRF has a depth of the same order as that of a classical CRF, its adaptivity nature ensures its interpolation.

Interpolation volume We start by studying the interpolation area of the RF.

Definition 2.7. *The **interpolation area** is the subspace of $[0,1]^d$ where the prediction of the forest depends on one training point only. For a given forest $m_{M,n}(.,\Theta_M)$, the interpolation area is denoted by*

$$\mathscr{A}(m_{M,n}(.,\Theta_M)) = \left\{ x \in [0,1]^d, \exists! X_i \in \mathscr{D}_n, X_i \in \bigcap_{m=1}^{M} A_n(x,\theta_m) \right\}.$$

Note that the partition of the RF consists in the intersection of the tree partitions.

The interpolation zone heavily depends on both the geometry of the training points X_i's and the construction of the trees. Analyzing the interpolation area for a finite AdaCRF turns out to be quite a challenging task. Therefore, we focus our study on the *core interpolation area* $\mathscr{A}_{min}$ written as

$$\mathscr{A}_{min} = \bigcap_{M\in\mathbb{N},\boldsymbol{\Theta}_M} \mathscr{A}(f_{M,n}(.,\boldsymbol{\Theta}_M)).$$

The area $\mathscr{A}_{min}$ is nothing but the intersection of the interpolation zones of all possible forests, or equivalently of a forest containing all the possible trees (and therefore all possible cuts). As an example note that in the case of centered trees, every cut may occur with a positive probability. Therefore, $\mathscr{A}_{min}$ matches the volume of the interpolation area of an infinite centered AdaCRF.

Proposition 2.8. *For all $n \geq 2$, for all $d \geq 2$, consider an infinite AdaCRF $f_{\infty,n}^{\text{AdaCRF}}$. Then,*

$$\mathbb{E}_{\mathscr{D}_n}\left[\mu(\mathscr{A}_{min})\right] \leq \left(\frac{2}{\log 2}\right)^d \frac{(1-2^{-n})^d}{n^{d-1}}.$$

This highlights the predominant *self-averaging* property of such forest architectures, and hence underpins the idea of good capabilities of AdaCRF in interpolation regimes apart from the empty cells. As we prove in the next section, the exact interpolation regime still produces too many empty cells that hinder the consistency property of AdaCRF.

Non-consistency

Proposition 2.9. *When $\mathbb{E}\left[f^\star(X)^2\right] > 0$, the infinite AdaCRF $f_{\infty,n}^{\text{AdaCRF}}$ is inconsistent in an exact interpolation regime (grown til pure leaves).*

The adaptiveness of AdaCRF is not sufficient to ensure consistency while reaching exact interpolation: it produces too many empty cells. In order to maintain both interpolation and consistency properties, it seems necessary to chose the threshold to split over between two points (for instance in the case of Breiman RF or Median RF) which produces a forest without empty leaves.

2.3.6 Breiman's forest

The widely-used Breiman RF is composed of several trees, built with CART methodology, each one trained on bootstrap samples, and for which the successive splitting directions and thresholds are chosen at each step (among a random subset of directions) in order to minimize the CART criterion (empirical variance for instance). Breiman forests are among the state-of-the-art ensemble methods in terms of predictive performance even if their adaptivity to the data remains a real hurdle to their theoretical analysis.

Proposition 2.10. *Consider an infinite Breiman forest constructed without bootstrap. Suppose that for a given configuration of the training data, all cuts have a probability strictly greater than 0 to appear. Then, the volume of the minimal interpolation zone verifies*

$$\mathbb{E}\left[\mu(\mathscr{A}_{min})\right] \leq \frac{1}{n^{d-1}}\left(1-2^{-n}\right)^d.$$

2.3.7 Conclusion

In particular, we show that *simple* models such as the vanilla CRF can still reach consistency within the mean interpolation regime if the aggregation rule does not take the empty cells into account. However, preserving both exact interpolation and consistency comes at the cost of a greater adaptivity of the RF.

Bibliography

Bartlett, P. L., Montanari, A., and Rakhlin, A. (2021). Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177*.

Belkin, M., Rakhlin, A., and Tsybakov, A. B. (2019). Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, 13:1063–1095.

Biau, G. and Devroye, L. (2015). *Lectures on the nearest neighbor method*, volume 246. Springer.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Devroye, L., Györfi, L., and Krzyżak, A. (1998). The hilbert kernel regression estimate. *Journal of Multivariate Analysis*, 65(2):209–227.

Scornet, E. (2016). Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500.

Wyner, A. J., Olson, M., Bleich, J., and Mease, D. (2017). Explaining the success of adaboost and random forests as interpolating classifiers. *The Journal of Machine Learning Research*, 18(1):1558–1590.