Structure of the course

Main goals:
Theoretical justification of efficiency of optimization methods.
No gap between theory and practice.

Part 1: Black-Box Optimization
Lecture 1.
Complexity of Black-Box Optimization
Difficult problems
Lower complexity bounds for Convex Optimization
Optimal methods
Lecture 2.
Second order methods. Systems of nonlinear equations
Globally convergent second-order schemes
Cubic regularization for Newton Method
Modified Gauss-Newton method
Main goals:

- Theoretical justification of efficiency of optimization methods.
- No gap between theory and practice.

Part 1: Black-Box Optimization

Lecture 1.
- Complexity of Black-Box Optimization
  - Difficult problems
  - Lower complexity bounds for Convex Optimization

Optimal methods

Lecture 2.
- Second order methods. Systems of nonlinear equations
  - Globally convergent second-order schemes
  - Cubic regularization for Newton Method
  - Modified Gauss-Newton method
Main goals:

- Theoretical justification of efficiency of optimization methods.
Main goals:

- Theoretical justification of *efficiency* of optimization methods.
- No gap between theory and practice.
Main goals:

- Theoretical justification of *efficiency* of optimization methods.
- No gap between theory and practice.

Part 1: Black-Box Optimization
Structure of the course

Main goals:

- Theoretical justification of *efficiency* of optimization methods.
- No gap between theory and practice.

Part 1: Black-Box Optimization

Lecture 1. *Complexity of Black-Box Optimization*
Structure of the course

Main goals:

- Theoretical justification of *efficiency* of optimization methods.
- No gap between theory and practice.

Part 1: Black-Box Optimization

Lecture 1. *Complexity of Black-Box Optimization*

- Difficult problems
Main goals:

- Theoretical justification of *efficiency* of optimization methods.
- No gap between theory and practice.

Part 1: Black-Box Optimization

Lecture 1. *Complexity of Black-Box Optimization*

- Difficult problems
- Lower complexity bounds for Convex Optimization
Main goals:

- Theoretical justification of *efficiency* of optimization methods.
- No gap between theory and practice.

Part 1: Black-Box Optimization

Lecture 1. *Complexity of Black-Box Optimization*

- Difficult problems
- Lower complexity bounds for Convex Optimization
- Optimal methods
Main goals:

- Theoretical justification of *efficiency* of optimization methods.
- No gap between theory and practice.

Part 1: Black-Box Optimization

**Lecture 1.** *Complexity of Black-Box Optimization*

- Difficult problems
- Lower complexity bounds for Convex Optimization
- Optimal methods

**Lecture 2.** *Second order methods. Systems of nonlinear equations*
Main goals:

- Theoretical justification of *efficiency* of optimization methods.
- No gap between theory and practice.

Part 1: Black-Box Optimization

**Lecture 1.** *Complexity of Black-Box Optimization*

- Difficult problems
- Lower complexity bounds for Convex Optimization
- Optimal methods

**Lecture 2.** *Second order methods. Systems of nonlinear equations*

- Globally convergent second-order schemes
Main goals:
- Theoretical justification of efficiency of optimization methods.
- No gap between theory and practice.

Part 1: Black-Box Optimization

Lecture 1. *Complexity of Black-Box Optimization*
- Difficult problems
- Lower complexity bounds for Convex Optimization
- Optimal methods

Lecture 2. *Second order methods. Systems of nonlinear equations*
- Globally convergent second-order schemes
- Cubic regularization for Newton Method
Structure of the course

Main goals:

- Theoretical justification of efficiency of optimization methods.
- No gap between theory and practice.

Part 1: Black-Box Optimization

Lecture 1. *Complexity of Black-Box Optimization*

- Difficult problems
- Lower complexity bounds for Convex Optimization
- Optimal methods

Lecture 2. *Second order methods. Systems of nonlinear equations*

- Globally convergent second-order schemes
- Cubic regularization for Newton Method
- Modified Gauss-Newton method
Part 2: Structural Optimization
Part 2: Structural Optimization

Lecture 3. *Interior-point methods*
Lecture 3. *Interior-point methods*

- Self-concordant functions
Part 2: Structural Optimization

Lecture 3. Interior-point methods

- Self-concordant functions
- Self-concordant barriers
Part 2: Structural Optimization

Lecture 3. *Interior-point methods*

- Self-concordant functions
- Self-concordant barriers
- Application examples
Part 2: Structural Optimization

Lecture 3. *Interior-point methods*
- Self-concordant functions
- Self-concordant barriers
- Application examples

Lecture 4. *Smoothing Technique*
Part 2: Structural Optimization

Lecture 3. *Interior-point methods*
- Self-concordant functions
- Self-concordant barriers
- Application examples

Lecture 4. *Smoothing Technique*
- Explicit model of objective function
Part 2: Structural Optimization

**Lecture 3.** *Interior-point methods*
- Self-concordant functions
- Self-concordant barriers
- Application examples

**Lecture 4.** *Smoothing Technique*
- Explicit model of objective function
- Smoothing
Part 2: Structural Optimization

Lecture 3. *Interior-point methods*
- Self-concordant functions
- Self-concordant barriers
- Application examples

Lecture 4. *Smoothing Technique*
- Explicit model of objective function
- Smoothing
- Application examples
Part 2: Structural Optimization

**Lecture 3.** *Interior-point methods*
- Self-concordant functions
- Self-concordant barriers
- Application examples

**Lecture 4.** *Smoothing Technique*
- Explicit model of objective function
- Smoothing
- Application examples

**Lecture 5.** *Huge-scale optimization*
Part 2: Structural Optimization

**Lecture 3. Interior-point methods**
- Self-concordant functions
- Self-concordant barriers
- Application examples

**Lecture 4. Smoothing Technique**
- Explicit model of objective function
- Smoothing
- Application examples

**Lecture 5. Huge-scale optimization**
- Sparsity in optimization problems
Part 2: Structural Optimization

Lecture 3. *Interior-point methods*
- Self-concordant functions
- Self-concordant barriers
- Application examples

Lecture 4. *Smoothing Technique*
- Explicit model of objective function
- Smoothing
- Application examples

Lecture 5. *Huge-scale optimization*
- Sparsity in optimization problems
- Coordinate-descent schemes
Part 2: Structural Optimization

**Lecture 3.** *Interior-point methods*
- Self-concordant functions
- Self-concordant barriers
- Application examples

**Lecture 4.** *Smoothing Technique*
- Explicit model of objective function
- Smoothing
- Application examples

**Lecture 5.** *Huge-scale optimization*
- Sparsity in optimization problems
- Coordinate-descent schemes
- Gradient methods with sublinear cost of iteration
Books:


Lecture 1. Intrinsic complexity of Black-Box Optimization

Yurii Nesterov, CORE/INMA (UCL)

January 20-22, 2016 (Ecole Polytechnique, Paris)
Outline

1 Basic NP-hard problem

2 NP-hardness of some popular problems

3 Lower complexity bounds for Global Minimization

4 Nonsmooth Convex Minimization. Subgradient scheme.

5 Smooth Convex Minimization. Lower complexity bounds

6 Methods for Smooth Minimization with Simple Constraints
Standard Complexity Classes

Let data be coded in matrix $A$, and $n$ be dimension of the problem.

### Combinatorial Optimization

- **NP-hard problems**: $2^n$ operations. Solvable in $O(p(n)\|A\|)$.  
- **Fully polynomial approximation schemes**: $O\left(\frac{p(n)}{\epsilon^k} \ln^\alpha \|A\|\right)$.  
- **Polynomial-time problems**: $O(p(n)\ln^\alpha \|A\|)$.  

### Continuous Optimization

- **Sublinear complexity**: $O\left(\frac{p(n)}{\epsilon^\alpha} \|A\|^\beta\right)$, $\alpha, \beta > 0$.  
- **Polynomial-time complexity**: $O\left(p(n)\ln\left(\frac{1}{\epsilon} \|A\|\right)\right)$. 
Basic NP-hard problem: Problem of stones

Given $n$ stones of integer weights $a_1, \ldots, a_n$, decide if it is possible to divide them on two parts of equal weight.

Mathematical formulation

Find a Boolean solution $x_i = \pm 1$, $i = 1, \ldots, n$, to a single linear equation

$$\sum_{i=1}^{n} a_i x_i = 0.$$ 

Another variant: $\sum_{i=2}^{n} a_i x_i = a_1$.

NB: Solvable in $O\left(\ln n \cdot \sum_{i=1}^{n} |a_i|\right)$ by FFT transform.
Theorem: Minimization of quartic polynomial of \( n \) variables is NP-hard.

Proof: Consider the following function:

\[
    f(x) = \sum_{i=1}^{n} x_i^4 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i^2 \right)^2 + \left( \sum_{i=1}^{n} a_i x_i \right)^4 + (1 - x_1)^4.
\]

The first part is \( \langle A[x]^2, [x]^2 \rangle \), where \( A = I - \frac{1}{n} e_n e_n^T \succeq 0 \) with \( Ae_n = 0 \), and \( [x]^2_i = x_i^2, \ i = 1, \ldots, n \).

Thus, \( f(x) = 0 \) iff all \( x_i = \tau \), \( \sum_{i=1}^{n} a_i x_i = 0 \), and \( x_1 = 1 \).

Corollary: Minimization of convex quartic polynomial over the unit sphere is NP-hard.
Problem: \( \min_u \{ f(x(1)) : x' = g(x, u), \ 0 \leq t \leq 1, \ x(0) = x_0 \} \).

Consider \( g(x, u) = \frac{1}{n} x \cdot \langle x, u \rangle - u \).

Lemma. Let \( \|x_0\|^2 = n \). Then \( \|x(t)\|^2 = n, \ 0 \leq t \leq 1 \).

Proof. Consider \( \tilde{g}(x, u) = \left( \frac{xx^T}{\|x\|^2} - I \right) u \) and let \( x' = \tilde{g}(x, u) \). Then

\[
\langle x', x \rangle = \langle \left( \frac{xx^T}{\|x\|^2} - I \right) u, x \rangle = 0.
\]

Thus, \( \|x(t)\|^2 = \|x_0\|^2 \). Same is true for \( x(t) \) defined by \( g \). \( \square \)

Note: We have enough degrees of freedom to put \( x(1) \) at any position of the sphere.

Hence, our problem is: \( \min \{ f(y) : \|y\|^2 = n \} \).
Consider $\phi(x) = \left(1 - \frac{1}{\gamma}\right) \max_{1 \leq i \leq n} |x_i| - \min_{1 \leq i \leq n} |x_i| + |\langle a, x \rangle|$, where $a \in \mathbb{Z}^n_+$ and $\gamma \overset{\text{def}}{=} \sum_{i=1}^{n} a_i \geq 1$. Clearly, $\phi(0) = 0$.

**Lemma.** It is NP-hard to decide if $\phi(x) < 0$ for some $x \in \mathbb{R}^n$.

**Proof:**
1. Assume that $\sigma \in \mathbb{R}^n$ with $\sigma_i = \pm 1$ satisfies $\langle a, \sigma \rangle = 0$. Then $\phi(\sigma) = -\frac{1}{\gamma} < 0$.
2. Assume $\phi(x) < 0$ and $\max_{1 \leq i \leq n} |x_i| = 1$. Denote $\delta = |\langle a, x \rangle|$.

Then $|x_i| > 1 - \frac{1}{\gamma} + \delta$, $i = 1, \ldots, n$.

Denoting $\sigma_i = \text{sign}x_i$, we have $\sigma_i x_i > 1 - \frac{1}{\gamma} + \delta$. Therefore, $|\sigma_i - x_i| = 1 - \sigma_i x_i < \frac{1}{\gamma} - \delta$, and we conclude that

$$
|\langle a, \sigma \rangle| \leq |\langle a, x \rangle| + |\langle a, \sigma - x \rangle| \leq \delta + \gamma \max_{1 \leq i \leq n} |\sigma_i - x_i| < (1 - \gamma)\delta + 1 \leq 1.
$$

Since $a \in \mathbb{Z}^n$, this is possible iff $\langle a, \sigma \rangle = 0$. 
Black-box optimization

**Oracle:** Special unit for computing function value and derivatives at test points. (0-1-2 order.)

**Analytic complexity:** Number of calls of oracle, which is necessary (sufficient) for solving any problem from the class. (Lower/Upper complexity bounds.)

**Solution:** $\epsilon$-approximation of the minimum.

**Resisting oracle:** creates the worst problem instance for a particular method.

- Starts from “empty” problem.
- Answers must be compatible with the description of the problem class.
- The bad problem is created after the method stops.
Bounds for Global Minimization

**Problem:** \( f^* = \min_x \{ f(x) : x \in B_n \}, \ B_n = \{ x \in \mathbb{R}^n : 0 \leq x \leq e_n \} \).

**Problem Class:** \( |f(x) - f(y)| \leq L \|x - y\|_{\infty} \ \forall x, y \in B_n. \)

**Oracle:** \( f(x) \) (zero order).

**Goal:** Find \( \bar{x} \in B_n : f(\bar{x}) - f^* \leq \epsilon. \)

**Theorem:** \( N(\epsilon) \geq (\frac{L}{2\epsilon})^n. \)

**Proof.** Divide \( B_n \) on \( p^n \) \( l_\infty \)-balls of radius \( \frac{1}{2p} \).

*Resisting oracle:* at each test point reply \( f(x) = 0. \)

Assume, \( N < p^n. \) Then, \( \exists \) ball with no questions. Hence, we can take \( f^* = -\frac{L}{2p}. \) Hence, \( \epsilon \geq \frac{L}{2p}. \)

**Corollary:** Uniform Grid method is worst-case optimal.
Nonsmooth Convex Minimization (NCM)

**Problem:** $f^* = \min_{x} \{ f(x) : x \in Q \}$, where

- $Q \subseteq \mathbb{R}^n$ is a convex set: $x, y \in Q \Rightarrow [x, y] \in Q$. It is simple.
- $f(x)$ is a sub-differentiable convex function:

$$f(y) \geq f(x) + \langle f'(x), y - x \rangle, \quad x, y \in Q,$$

for certain subgradient $f'(x) \in \mathbb{R}^n$.

**Oracle:** $f(x), f'(x)$ (first order).

**Solution:** $\epsilon$-approximation in function value.

**Main inequality:** $\langle f'(x), x - x^* \rangle \geq f(x) - f^* \geq 0, \forall x \in Q$.

**NB:** Anti-subgradient decreases the distance to the optimum.
Denote by $\partial f(x)$ the subdifferential of $f$ at $x$. This is the set of all subgradients at $x$.

1. For $f = \alpha_1 f_1 + \alpha_2 f_2$ with $\alpha_1, \alpha_2 > 0$, we have
   \[ \partial f(x) = \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x). \]

2. For $f = \max\{f_1, f_2\}$, we have
   \[ \partial f(x) = \text{Conv} \left\{ \partial f_1(x), \partial f_2(x) \right\}. \]
Let $Q \equiv \{ \|x\| \leq 2R \}$ and $x^{k+1} \in x^0 + \text{Lin}\{f'(x^0), \ldots, f'(x^k)\}$.

Consider the function $f_m(x) = L \max_{1 \leq i \leq m} x_i + \frac{\mu}{2} \|x\|^2$ with $\mu = \frac{L}{Rm^{1/2}}$.

From the problem: $\min_{\tau} (L\tau + \frac{\mu m}{2} \tau^2)$, we get

$$\tau^* = -\frac{L}{\mu m} = -\frac{R}{m^{1/2}}, \quad f_m^* = -\frac{L^2}{2\mu m} = -\frac{LR}{m^{1/2}}, \quad \|x^*\|^2 = m\tau^*_2 = R^2.$$

**NB:** If $x^0 = 0$, then after $k$ iterations we can keep $x_i = 0$ for $i > k$.

**Lipschitz continuity:** $f_{k+1}(x^k) - f_{k+1}^* \geq -f_{k+1}^* = \frac{LR}{(k+1)^{1/2}}$.

**Strong convexity:** $f_{k+1}(x^k) - f_{k+1}^* \geq -f_{k+1}^* = \frac{L^2}{2(k+1)\mu}$.

Both lower bounds are **exact**!
**Subgradient Method (SG)**

**Problem:** \( \min_{x \in Q} \{ f(x) : g(x) \leq 0 \} \),

where \( Q \) is a closed convex set, and convex \( f, g \in C^0_0(Q) \).

**SG:** If \( \frac{g(x^k)}{\|g'(x^k)\|} > h \)

\( a) \) \( x^{k+1} = \pi_Q \left( x^k - \frac{g(x^k)}{\|g'(x^k)\|^2} g'(x^k) \right) \),

\( b) \) \( x^{k+1} = \pi_Q \left( x^k - \frac{h}{\|f'(x^k)\|} f'(x^k) \right) \).

Denote \( f_N^* = \min_{0 \leq k \leq N} \{ f(x^k) : k \in b) \} \). Let \( N = N_a + N_b \).

**Theorem:** If \( N > \frac{1}{h^2} \|x^0 - x^*\|^2 \), then \( f_N^* - f^* \leq hL \). \((h = \frac{c}{L}).\)

**Proof:** Denote \( r_k = \|x^k - x^*\| \).

\( a) \): \( r_{k+1}^2 - r_k^2 \leq -\frac{2g(x^k)}{\|g'(x^k)\|^2} \langle g'(x^k), x^k - x^* \rangle + \frac{g^2(x^k)}{\|g'(x^k)\|^2} \leq -h^2 \).

\( b) \): \( r_{k+1}^2 - r_k^2 \leq -\frac{2h \langle f'(x^k), x^k - x^* \rangle}{\|f'(x^k)\|} + h^2 \leq -\frac{2h}{L} (f(x^k) - f^*) + h^2 \).

Thus, \( Nb \frac{2h}{L} (f_N^* - f^*) \leq r_0^2 + h^2 (N_b - N_a) = r_0^2 + h^2 (2N_b - N) \).
Smooth Convex Minimization (SCM)

Lipschitz-continuous gradient: \[ \| f'(x) - f'(y) \| \leq L \| x - y \|. \]

Geometric interpretation: for all \( x, y \in \text{dom } F \) we have

\[
0 \leq f(y) - f(x) - \langle f'(x), y - x \rangle \\
= \int_0^1 \langle f'(x + \tau(y - x)) - f'(x), y - x \rangle d\tau \leq \frac{L}{2} \| x - y \|^2.
\]

Sufficient condition: \( 0 \leq f''(x) \leq L \cdot I_n, \ x \in \text{dom } f. \)

Equivalent definition:

\[
f(y) \geq f(x) + \langle f'(x), y - x \rangle + \frac{1}{2L} \| f'(x) - f'(y) \|^2.
\]

Hint: Prove first that \( f(x) - f^* \geq \frac{1}{2L} \| f'(x) \|^2. \)
Consider the family of functions $(k \leq n)$:

$$f_k(x) = \frac{1}{2} \left[ x_1^2 + \sum_{i=1}^{k-1} (x_i - x_{i+1})^2 + x_k^2 \right] - x_1 \equiv \frac{1}{2} \langle A_k x, x \rangle - x_1.$$ 

Let $R^n_k = \{ x \in R^n : x_i = 0, \ i > k \}$. Then $f_{k+p}(x) = f_k(x)$, $x \in R^n_k$.

Clearly, $0 \leq \langle A_k h, h \rangle \leq h_1^2 + \sum_{i=1}^{k-1} 2(h_i^2 + h_{i+1}^2) + h_k^2 \leq 4\|h\|^2$,

$$A_k = \begin{pmatrix}
2 & -1 & 0 & \cdots & 0 \\
-1 & 2 & -1 & 0 & \cdots \\
0 & -1 & 2 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & -1 & 2 & -1 \\
0 & \cdots & 0 & -1 & 2 \\
& & & & 0_{n-k,k} \\
& & & & 0_{n-k,n-k}
\end{pmatrix},$$

$k$ lines
Hence, \( A_kx = e_1 \) has the solution 
\[
\bar{x}_i^k = \begin{cases} 
\frac{k+1-i}{k+1}, & 1 \leq i \leq k, \\
0, & i > k.
\end{cases}
\]
Thus \( f_k^* = \frac{1}{2} \langle A_k \bar{x}^k, \bar{x}^k \rangle - \langle e_1, \bar{x}^k \rangle = -\frac{1}{2} \langle e_1, \bar{x}^k \rangle = -\frac{k}{2(k+1)} \), and
\[
\| \bar{x}^k \|^2 = \sum_{i=1}^{k} \left( \frac{k+1-i}{k+1} \right)^2 = \frac{1}{(k+1)^2} \sum_{i=1}^{k} i^2 = \frac{k(2k+1)}{6(k+1)}.
\]

Let \( x^0 = 0 \) and \( p \leq n \) is fixed.

**Lemma.** If \( x^k \in \mathcal{L}_k \overset{\text{def}}{=} \text{Lin}\{f'_p(x^0), \ldots, f'_p(x^{k-1})\} \), then \( \mathcal{L}_k \subseteq R^n_k \).

**Proof:** \( x^0 = 0 \in R^n_0, f'_p(0) = -e_1 \in R^n_1 \Rightarrow x^1 \in R^n_1, f'_p(x^1) \in R^n_2, \Box \)

**Corollary 1:** \( f_p(x^k) = f_k(x^k) \geq f_k^* \).

**Corollary 2:** Take \( p = 2k + 1 \). Then
\[
\frac{f_p(x^k) - f_p^*}{L\|x^0 - \bar{x}^p\|^2} \geq \left[ -\frac{k}{2(k+1)} + \frac{2k+1}{2(2k+2)} \right] / \left[ \frac{(2k+1)(4k+3)}{3(k+1)} \right] = \frac{3}{4(2k+1)(4k+3)}.
\]
\[
\| x^k - \bar{x}^p \|^2 \geq \sum_{i=k+1}^{2k+1} (\bar{x}_i^{2k+1})^2 = \frac{(2k+3)(k+2)}{24(k+1)} \geq \frac{1}{8} \| \bar{x}^p \|^2.
\]
Some remarks

1. The rate of convergence of any Black-Box gradient methods as applied to $f \in C^{1,1}$ cannot be higher than $O\left(\frac{1}{k^2}\right)$.

2. We cannot guarantee any rate of convergence in the argument.

3. Let $A = LL^T$ and $f(x) = \frac{1}{2}\langle Ax, x \rangle - \langle b, x \rangle$. Then
   
   $$f(x) - f^* = \frac{1}{2}\|L^T x - d\|^2,$$

   where $d = L^T x^*$.

   Thus, the residual of the linear system $L^T x = b$ cannot be decreased faster than with the rate $O\left(\frac{1}{k}\right)$ (provided that we are allowed to multiply by $L$ and $L^T$).

4. Optimization problems with nontrivial linear equality constraints cannot be solved faster than with the rate $O\left(\frac{1}{k}\right)$.
Consider the problem: \( \min_x \{ f(x) : x \in Q \} \),
where convex \( f \in C^{1,1}_L(Q) \), and \( Q \) is a simple closed convex set (allows projections).

**Gradient mapping:** for \( M > 0 \) define
\[
T_M(x) = \arg\min_{y \in Q} \left[ f(x) + \langle f'(x), y - x \rangle + \frac{M}{2} \|x - y\|^2 \right].
\]
If \( M \geq L \), then
\[
f(T_M(x)) \leq f(x) + \langle f'(x), T_M(x) - x \rangle + \frac{M}{2} \|x - T_M(x)\|^2.
\]

**Reduced gradient:** \( g_M(x) = M \cdot (x - T_M(x)) \).

Since \( \langle f'(x) + M(T_M(x) - x), y - T_M(x) \rangle \geq 0 \) for all \( y \in Q \),
\[
f(x) - f(T_M(x)) \geq \frac{M}{2} \|x - T_M(x)\|^2 = \frac{1}{2M} \|g_M(x)\|^2, \quad (\rightarrow 0)
\]
\[
f(y) \geq f(x) + \langle f'(x), T_M(x) - x \rangle + \langle f'(x), y - T_M(x) \rangle
\]
\[
\geq f(T_M(x)) - \frac{1}{2M} \|g_M(x)\|^2 + \langle g_M(x), y - T_M(x) \rangle.
\]
Primal Gradient Method (PGM)

Main scheme: \( x^0 \in Q, \quad x^{k+1} = T_L(x^k), \ k \geq 0. \)

Primal interpretation: \( x^{k+1} = \pi_Q (x^k - \frac{1}{L} f'(x^k)). \)

Rate of convergence. \[ f(x^k) - f(x^{k+1}) \geq \frac{1}{2L} \| g_L(x^k) \|^2. \]
\[
 f(T_L(x)) - f^* \leq \frac{1}{2L} \| g_L(x) \|^2 + \langle g_L(x), T_L(x) - x^* \rangle \\
 \leq \frac{1}{2L} (\| g_L(x) \| + LR)^2 - \frac{L}{2} R^2.
\]

Hence, \( \| g_L(x) \| \geq \frac{2L(f(T_L(x)) - f^*) + L^2 R^2}{2L(f(T_L(x)) - f^*)} \) \( 1/2 - LR \)
\[
 = \frac{2L(f(T_L(x)) - f^*)}{[2L(f(T_L(x)) - f^*) + L^2 R^2]^{1/2} + LR} \geq \frac{c}{R} \cdot (f(T_L(x)) - f^*).
\]

Thus, \( f(x^k) - f(x^{k+1}) \geq \frac{c^2}{LR^2} (f(x^{k+1}) - f^*)^2. \)

Similar situation: \( a'(t) = -a^2(t) \Rightarrow a(t) \approx \frac{1}{t}. \)

Conclusion: PGM converges as \( O\left(\frac{1}{k}\right)\). This is far from the lower complexity bounds.
Dual Gradient Method (DGM)

**Model:** Let $\lambda_i^k \geq 0$, $i = 0, \ldots, k$, and $S_k \overset{\text{def}}{=} \sum_{i=0}^{k} \lambda_i^k$. Then

$$S_k f(y) \geq \mathcal{L}_{\lambda^k}(y) \overset{\text{def}}{=} \sum_{i=0}^{k} \lambda_i^k [f(x^i) + \langle f'(x^i), y - x^i \rangle], \quad y \in Q.$$

**DGM:**

$$x^{k+1} = \arg \min_{y \in Q} \left\{ \psi_k(y) \overset{\text{def}}{=} \mathcal{L}_{\lambda^k}(y) + \frac{M}{2} \| y - x^0 \|^2 \right\}.$$

Let us choose $\lambda_i^k \equiv 1$ and $M = L$. We prove by induction

$$\star: \quad F^*_k = \sum_{i=0}^{k} f(y^i) \leq \psi^*_k \overset{\text{def}}{=} \min_{y \in Q} \psi_k(y). \quad (\leq (k+1)f^* + \frac{L}{2}R^2)$$

1. $k = 0$. Then $y^0 = T_L(x^0)$.
2. Assume ($\star$) is true for some $k \geq 0$. Then

$$\psi^*_{k+1} = \min_{y \in Q} \left[ \psi_k(y) + f(x^k) + \langle f'(x^k), y - x^k \rangle \right] \geq \min_{y \in Q} \left[ \psi^*_k + \frac{L}{2} \| y - x^k \|^2 + f(x^k) + \langle f'(x^k), y - x^k \rangle \right].$$

We can take $y^{k+1} = T_L(x_k)$. Thus, $\frac{1}{k+1} \sum_{i=0}^{k} f(y^i) \leq f^* + \frac{LR^2}{2(k+1)}$. 
Some remarks

1. Dual gradient method works with the model of the objective function.

2. The minimizing sequence \( \{y^k\} \) is not necessary for the algorithmic scheme. We can generate it if necessary.

3. Both primal and dual method have the same rate of convergence \( O\left(\frac{1}{k}\right) \). It is not optimal.

May be we can combine them in order to get a better rate?
Comparing PGM and DGM

<table>
<thead>
<tr>
<th>Primal Gradient method</th>
</tr>
</thead>
<tbody>
<tr>
<td>■ Monotonically improves the current state using the local model of the objective.</td>
</tr>
<tr>
<td>■ <strong>Interpretation:</strong> Practitioners, industry.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dual Gradient Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>■ The main goal is to construct a model of the objective.</td>
</tr>
<tr>
<td>■ It is updated by a new experience collected around the predicted test points ( x_k ).</td>
</tr>
<tr>
<td>■ Practical verification of the advices ( y_k ) is not essential for the procedure.</td>
</tr>
<tr>
<td>■ <strong>Interpretation:</strong> Science.</td>
</tr>
</tbody>
</table>

**Hint:** Combination of theory and practice should give better results
Def. A sequences \( \{ \phi_k(x) \}_{k=0}^{\infty} \) and \( \{ \lambda_k \}_{k=0}^{\infty} \), \( \lambda_k \geq 0 \) are called the *estimating sequences* if \( \lambda_k \to 0 \) and \( \forall x \in Q, k \geq 0, \)
\[
(\ast): \quad \phi_k(x) \leq (1 - \lambda_k)f(x) + \lambda_k \phi_0(x).
\]

**Lemma:** If (\(\ast\ast\)) : \( f(x^k) \leq \phi_k^* = \min_{x \in Q} \phi_k(x) \), then
\[
f(x^k) - f^* \leq \lambda_k [\phi_0(x^*) - f^*] \to 0.
\]

**Proof.** \( f(x^k) \leq \phi_k^* = \min_{x \in Q} \phi_k(x) \leq \min_{x \in Q} [(1 - \lambda_k)f(x) + \lambda_k \phi_0(x)] \)
\[
\leq (1 - \lambda_k)f(x^*) + \lambda_k \phi_0(x^*). \quad \square
\]

Rate of \( \lambda_k \to 0 \) defines the rate of \( f(x^k) \to f^* \).

**Questions**

- How to construct the estimating sequences?
- How we can ensure (\(\ast\ast\))?
Let $\phi_0(x) = \frac{L}{2} \|x - x^0\|^2$, $\lambda_0 = 1$, $\{y^k\}_{k=0}^{\infty}$ is a sequence in $Q$, and $\{\alpha_k\}_{k=0}^{\infty} : \alpha_k \in (0, 1)$, $\sum_{k=0}^{\infty} \alpha_k = \infty$. Then $\{\phi_k(x)\}_{k=0}^{\infty}$, $\{\lambda_k\}_{k=0}^{\infty}$:

$$
\lambda_{k+1} = (1 - \alpha_k)\lambda_k,
$$

$$
\phi_{k+1}(x) = (1 - \alpha_k)\phi_k(x) + \alpha_k[f(y^k) + \langle f'(y^k), x - y^k \rangle]
$$

are estimating sequences.

**Proof:** $\phi_0(x) \leq (1 - \lambda_0)f(x) + \lambda_0\phi_0(x) \equiv \phi_0(x)$.

If (*) holds for some $k \geq 0$, then

$$
\phi_{k+1}(x) \leq (1 - \alpha_k)\phi_k(x) + \alpha_k f(x)
$$

$$
= (1 - (1 - \alpha_k)\lambda_k)f(x) + (1 - \alpha_k)(\phi_k(x) - (1 - \lambda_k)f(x))
$$

$$
\leq (1 - (1 - \alpha_k)\lambda_k)f(x) + (1 - \alpha_k)\lambda_k\phi_0(x)
$$

$$
= (1 - \lambda_{k+1})f(x) + \lambda_{k+1}\phi_0(x). \quad \Box
$$
Denote $\phi^*_k = \min_{x \in Q} \phi_k(x)$, $v^k = \arg\min_{x \in Q} \phi_k(x)$. Suppose $\phi^*_k \geq f(x^k)$.

$$
\phi^*_{k+1} = \min_{x \in Q} \left\{ (1 - \alpha_k) \phi_k(x) + \alpha_k [f(y^k) + \langle f'(y^k), x - y^k \rangle] \right\} \geq \\
\min_{x \in Q} \left\{ (1 - \alpha_k) \left[ \phi^*_k + \frac{\lambda_k L}{2} \|x - v_k\|^2 \right] + \alpha_k [f(y^k) + \langle f'(y^k), y - y^k \rangle] \right\}
$$

$$
\geq \min_{x \in Q} \left\{ f(y^k) + \frac{(1 - \alpha_k) \lambda_k L}{2} \|x - v_k\|^2 \\
+ \langle f'(y^k), \alpha_k (x - y^k) + (1 - \alpha_k) (x^k - y^k) \rangle \right\}
$$

$$(y_k) \overset{\text{def}}{=} (1 - \alpha_k) x^k + \alpha_k v^k = x^k + \alpha_k (v^k - x^k)$$

$$
= \min_{x \in Q} \left\{ f(y^k) + \frac{(1 - \alpha_k) \lambda_k L}{2} \|x - v_k\|^2 + \alpha_k \langle f'(y^k), x - v_k \rangle \right\}
$$

$$
= \min_{y = x^k + \alpha_k (x - x^k)} \left\{ f(y^k) + \frac{(1 - \alpha_k) \lambda_k L}{2 \alpha_k^2} \|y - y_k\|^2 + \langle f'(y^k), y - y^k \rangle \right\} \overset{(?)}{=} f(x^{k+1})
$$

**Answer:** $\alpha_k^2 = (1 - \alpha_k) \lambda_k$. $x_{k+1} = T_L(y_k)$. 
Choose \( v^0 = x^0 \in Q \), \( \lambda_0 = 1 \), \( \phi_0(x) = \frac{L}{2} \| x - x^0 \|^2 \).

For \( k \geq 0 \) iterate:

- Compute \( \alpha_k : \alpha_k^2 = (1 - \alpha_k) \lambda_k \equiv \lambda_{k+1} \).
- Define \( y_k = (1 - \alpha_k)x^k + \alpha_k v^k \).
- Compute \( x^{k+1} = T_L(y^k) \).
- \( \phi_{k+1}(x) = (1 - \alpha_k)\phi_k(x) + \alpha_k [f(y^k) + \langle f'(y^k), x - y^k \rangle] \).

**Convergence:** Denote \( a_k = \lambda_k^{-1/2} \). Then

\[
 a_{k+1} - a_k = \frac{\lambda_k^{1/2} - \lambda_{k+1}^{1/2}}{\lambda_k^{1/2} \lambda_{k+1}^{1/2}} = \frac{\lambda_k - \lambda_{k+1}}{2 \lambda_k^{1/2} \lambda_{k+1}^{1/2} (\lambda_k^{1/2} + \lambda_{k+1}^{1/2})} \geq \frac{\lambda_k - \lambda_{k+1}}{2 \lambda_k \lambda_{k+1}} = \frac{\alpha_k}{2 \lambda_{k+1}^{1/2}} = \frac{1}{2}.
\]

Thus, \( a_k \geq 1 + \frac{k}{2} \). Hence, \( \lambda_k \leq \frac{4}{(k+2)^2} \).
Interpretation

1. $\phi_k(x)$ accumulates all previously computed information about the objective. This is a current *model* of our problem.
2. $v^k = \arg\min_{x \in Q} \phi_k(x)$ is a prediction of the optimal strategy.
3. $\phi^*_k = \phi_k(v^k)$ is an estimate of the optimal value.
4. **Acceleration condition:** $f(x^k) \leq \phi^*_k$. We need a firm, which is at least as good as the best theoretical prediction.
5. Then we create a startup $y^k = (1 - \alpha_k)x^k + \alpha_k v^k$, and allow it to work one year.

6. **Theorem:** Next year, its performance will be at least as good as the new theoretical prediction. And we can continue!

**Acceleration result:** 10 years instead 100.

Who is in a right position to arrange 5? Government, political institutions.
Lecture 2. Second-order methods. Solving systems of nonlinear equations

Yurii Nesterov, CORE/INMA (UCL)

January 20-22, 2016 (Ecole Polytechnique, Paris)
Outline

1. Historical remarks
2. Trust region methods
3. Cubic regularization of second-order model
4. Local and global convergence
5. Accelerated Cubic Newton
6. Solving the system of nonlinear equations
7. Numerical experiments
Historical remarks

Problem: \( f(x) \rightarrow \min \quad x \in \mathbb{R}^n \)
is treated as a non-linear system \( f'(x) = 0 \).

Newton method: \[ x_{k+1} = x_k - \left[ f''(x_k) \right]^{-1} f'(x_k). \]

Standard objections:

- The method is not always well defined (det \( f''(x_k) = 0 \)).
- Possible divergence.
- Possible convergence to saddle points or even to local maximums.
- Chaotic global behavior.
Pre-History (see Ortega, Rheinboldt [1970].)

- **Bennet [1916]:** Newton’s method in general analysis.
- **Levenberg [1944]:** Regularization. If $f''(x) \not\succ 0$, then use $d = G^{-1}f'(x)$ with $G = f''(x) + \gamma I \succ 0$. (See also Marquardt [1963].)
- **Kantorovich [1948]:** Proof of local quadratic convergence. Assumptions:
  a) $f \in C^3(R^n)$.
  b) $\|f''(x) - f''(y)\| \leq L_2\|x - y\|$.
  c) $f''(x^*) \succ 0$.
  d) $x_0 \approx x^*$.

**Global convergence:** Use line search (good advice).

**Global performance:** Not addressed.
Main idea: Trust Region Approach.

1. By some norm $\| \cdot \|_k$ define the trust region $\mathcal{B}_k = \{ x \in R^n : \| x - x_k \|_k \leq \Delta_k \}$.

2. Denote $m_k(x) = f(x_k) + \langle f'(x_k), x - x_k \rangle + \frac{1}{2} \langle G_k(x - x_k), x - x_k \rangle$. Variants: $G_k = f''(x_k), G_k = f''(x_k) + \gamma_k I \succ 0$, etc.

3. Compute the trial point $\hat{x}_k = \arg \min_{x \in \mathcal{B}_k} m_k(x)$.

4. Compute the ratio $\rho_k = \frac{f(x_k) - f(\hat{x}_k)}{f(x_k) - m_k(\hat{x}_k)}$.

5. In accordance to $\rho_k$ either accept $x_{k+1} = \hat{x}_k$ or update the value $\Delta_k$ and repeat the steps above.
Advantages:

- More parameters ⇒ Flexibility
- Convergence to a point, which satisfies second-order necessary optimality condition:
  \[ f'(x^*) = 0, \quad f''(x^*) \succeq 0. \]

Disadvantages:

- Complicated strategies for parameters’ coordination.
- For certain \( \| \cdot \|_k \) the auxiliary problem is difficult.
- Line search abilities are quite limited.
- Unselective theory.
- Global complexity issues are not addressed.
Development of numerical schemes

**Classical style:** Problem formulation $\Rightarrow$ Method

Examples:
- Gradient and Newton methods in optimization.
- Runge-Kutta method for ODE, etc.

**2. Modern style:**

```
\{ Problem formulation, Problem class \} $\Rightarrow$ Method
```

Examples:
- Non-smooth convex minimization.
- Smooth minimization: $\min_{x \in Q} f(x)$, with $f \in C^{1,1}$.

*Gradient mapping (Nemirovsky & Yudin 77):*

\[
\begin{align*}
x_+ &= T(x) \equiv \arg \min_{y \in Q} m_1(y), \\
m_1(y) &\equiv f(x) + \langle f'(x), y - x \rangle + \frac{L_1}{2} \|y - x\|^2.
\end{align*}
\]

Justification: $f(y) \leq m_1(y)$ for all $y \in Q$. 
Using the second-order model

**Problem:** $f(x) \min x \in \mathbb{R}^n$.

**Assumption:** Let $\mathcal{F}$ be an open convex set. Then

$$\|f''(x) - f''(y)\| \leq L_2 \|x - y\| \quad \forall x, y \in \mathcal{F},$$

$$\mathcal{L}(x_0) = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\} \subset \mathcal{F}.$$ 

Define

$$m_2(x, y) = f(x) + \langle f'(x), y - x \rangle + \frac{1}{2} \langle f''(x)(y - x), y - x \rangle,$$

$$m'_2(x, y) = f'(x) + f''(x)(y - x).$$

**Lemma 1.** For any $x, y \in \mathcal{F}$

$$\|f'(y) - m'_2(x, y)\| \leq \frac{1}{2} L_2 \|y - x\|^2,$$

$$|f(y) - m_2(x, y)| \leq \frac{1}{6} L_2 \|y - x\|^3.$$

**Corollary:** For any $x$ and $y$ from $\mathcal{F}$,

$$f(y) \leq m_2(x, y) + \frac{1}{6} L_2 \|y - x\|^3.$$
For $M > 0$ define $\hat{f}_M(x, y) = m_2(x, y) + \frac{1}{6} M \|y - x\|^3$,
$$T_M(x) \in \text{Arg min}_y \hat{f}_M(x, y),$$
where “Arg” indicates that $T_M(x)$ is a global minimum.

**Computability:** If $\| \cdot \|$ is a Euclidean norm, then $T_M(x)$ can be computed from a convex problem.

For $r \in \mathcal{D} \equiv \{ r \in \mathbb{R} : f''(x) + \frac{M}{2} r I \succ 0, r \geq 0 \}$, denote $v(r) = -\frac{1}{2} \langle (f''(x) + \frac{Mr}{2} I)^{-1} f'(x), f'(x) \rangle - \frac{M}{12} r^3$.

**Lemma.** For $M > 0$, \[ \min_{h \in \mathbb{R}^n} \hat{f}_M(x, x + h) = \sup_{r \in \mathcal{D}} v(r). \]
If the sup is attained at $r^* : f''(x) + \frac{Mr^*}{2} I \succ 0$, then
$$h^* = -(f''(x) + \frac{Mr^*}{2} I)^{-1} f'(x)$$
where $r^* > 0$ is a unique solution to $r = \|(f''(x) + \frac{Mr}{2} I)^{-1} f'(x)\|$. 
Simple properties

1. Denote \( r_M(x) = \|x - T_M(x)\| \). Then
\[
f'(x) + f''(x)(T_M(x) - x) + \frac{Mr_M(x)}{2}(T_M(x) - x) = 0,
\]
\[
f''(x) + \frac{1}{2} M r_M(x) I \succeq 0.
\]

2. We have \( \langle f'(x), x - T_M(x) \rangle \geq 0 \), and
\[
f(x) - \bar{f}_M(x) \geq \frac{M}{12} r_M^3(x),
\]
\[
r_M^2(x) \geq \frac{2}{L+M} \|f'(x)\|.
\]

3. If \( M \geq L \) then \( \bar{f}_M(x) \geq f(T_M(x)) \).

4. \( \bar{f}_M(x) \leq \min_y \left[ f(y) + \frac{L+M}{6} \|y - x\|^3 \right] \).

Compare with \textit{prox-method}: \( x_+ = \min_y \left[ f(y) + \frac{1}{2} M \|y - x\|^2 \right] \).
Cubic regularization of Newton method

Consider the process: \( x_{k+1} = T_L(x_k), \quad k = 0, 1, \ldots \)

Note that \( f(x_{k+1}) \leq f(x_k) \).

**Saddle points.** Let \( f'(x^*) = 0 \) and \( f''(x^*) \not\leq 0 \). Then \( \exists \epsilon, \delta > 0 \) such that

\[
\|x - x^*\| \leq \epsilon, \quad f(x) \geq f(x^*) \Rightarrow f(T_L(x)) \leq f(x^*) - \delta
\]

**Local convergence.** If \( \mathcal{L}(x_0) \) is bounded, then

\[
X^* \equiv \lim_{k \to \infty} \{x_k\} \neq \emptyset.
\]

For any \( x^* \in X^* \) we have \( f(x^*) = f^*, \quad f'(x^*) = 0, \quad f''(x^*) \geq 0 \).

**Global convergence:** \( g_k \equiv \min_{1 \leq i \leq k} \|f'(x_i)\| \leq O \left( \frac{1}{k^{2/3}} \right) \).

For gradient method we can guarantee only \( g_k \leq O \left( \frac{1}{k^{1/2}} \right) \).

**Local rate of convergence:** Quadratic.
Global performance: Star-convex functions

Def. For any $x^* \in X^*$ and any $x \in \mathcal{F}$, $\alpha \in [0,1]$ we have $f(\alpha x^* + (1 - \alpha)x) \leq \alpha f(x^*) + (1 - \alpha)f(x)$.

Th 1. Let $\text{diam } \mathcal{F} \leq D$. Then

1. If $f(x_0) - f^* \geq \frac{3}{2}LD^3$, then $f(x_1) - f^* \leq \frac{1}{2}LD^3$.
2. If $f(x_0) - f^* \leq \frac{3}{2}LD^3$, then $f(x_k) - f^* \leq \frac{3LD^3}{2(1 + \frac{1}{3}k)^2}$.

Let $X^*$ be non-degenerate: $f(x) - f^* \geq \frac{\gamma}{2}\rho^2(x, X^*)$. Denote $\bar{\omega} = \frac{1}{L^2}(\frac{\gamma}{2})^3$.

Th 2. Denote $k_0$ the first number for which $f(x_{k_0}) - f^* \leq \frac{4}{9}\bar{\omega}$.

If $k \leq k_0$, then $f(x_k) - f^* \leq \left[\frac{(f(x_0) - f^*)^{1/4} - \frac{k}{6}\sqrt{\frac{2}{3}\bar{\omega}^{1/4}}}{\bar{\omega}}\right]^4$.

For $k \geq k_0$ we have $f(x_{k+1}) - f^* \leq \frac{1}{2}(f(x_k) - f^*)\sqrt{\frac{f(x_k) - f^*}{\bar{\omega}}}$.

NB The Hessian $f''(x^*)$ can be degenerate!
Global performance: Gradient-dominated functions

**Definition.** For any \( x \in \mathcal{F} \) and \( x^* \in X^* \) we have
\[
f(x) - f(x^*) \leq \tau_f \| f'(x) \|^p
\]
with \( \tau_f > 0 \) and \( p \in [1, 2] \) (degree of domination).

**Example 1.** *Convex functions:*
\[
f(x) - f^* \leq \langle f'(x), x - x^* \rangle \leq R \| f'(x) \|
\]
for \( \| x - x^* \| \leq R \). Thus, \( p = 1, \tau_f = \frac{1}{2} D \).

**Example 2.** *Strongly convex functions:*
\[
\forall x, y \in \mathbb{R}^n \quad f(x) \leq f(y) + \langle f'(y), x - y \rangle + \frac{1}{2\gamma} \| f'(x) - f'(y) \|^2.
\]
Thus, \( f(x) - f^* \leq \frac{1}{2\gamma} \| f'(x) \|^2 \quad \Rightarrow \quad p = 2, \tau_f = \frac{1}{2\gamma} \).
Example 3. *Sum of squares.* Consider the system
\[ g(x) = 0 \in \mathbb{R}^m, \quad x \in \mathbb{R}^n. \]
Assume that \( m \leq n \) and the Jacobian \( J(x) = (g'_1(x), \ldots, g'_m(x)) \) is 
uniformly non-degenerate:
\[ \sigma \equiv \inf_{x \in \mathcal{F}} \lambda_{\min}(J^T(x)J(x)) > 0. \]
Consider the function \( f(x) = \sum_{i=1}^{m} g_i^2(x) \). Then
\[ f(x) - f^* \leq \frac{1}{2\sigma} \| f'(x) \|^2. \]
Thus, \( p = 2 \) and \( \tau_f = \frac{1}{2\sigma} \).
Gradient dominated functions: rate of convergence

**Theorem 3.** Let $p = 1$. Denote $\hat{\omega} = \frac{2}{3} L (6\tau_f)^3$. Let $k_0$ be defined as $f(x_{k_0}) - f^* \leq \xi^2 \hat{\omega}$ for some $\xi > 1$. Then for $k \leq k_0$ we have

$$\ln \left( \frac{1}{\hat{\omega}} (f(x_k) - f^*) \right) \leq \left( \frac{2}{3} \right)^k \ln \left( \frac{1}{\hat{\omega}} (f(x_0) - f^*) \right).$$

Otherwise, $f(x_k) - f^* \leq \hat{\omega} \cdot \frac{\xi^2 (2 + \frac{3}{2} \xi)^2}{(2 + (k + \frac{3}{2}) \cdot \xi)^2}.$

**Theorem 4.** Let $p = 2$. Denote $\tilde{\omega} = \frac{1}{(144L)^2 \tau_f^3}$. Let $k_0$ be defined as $f(x_{k_0}) - f^* \leq \tilde{\omega}$. Then for $k \leq k_0$ we have

$$f(x_k) - f^* \leq (f(x_0) - f^*) \cdot e^{-k\sigma}$$

with $\sigma = \frac{\tilde{\omega}^{1/4}}{\tilde{\omega}^{1/4} + (f(x_0) - f^*)^{1/4}}$. Otherwise,

$$f(x_{k+1}) - f^* \leq \tilde{\omega} \cdot \left( \frac{f(x_k) - f^*}{\tilde{\omega}} \right)^{4/3}.$$

**NB:** Superlinear convergence without direct nondegeneracy assumption for the Hessian.
Transformations of convex functions

Let $u(x) : R^n \rightarrow R^n$ be non-degenerate. Denote by $v(u)$ its inverse: $v(u(x)) \equiv x$.

Consider the function $f(x) = \phi(u(x))$, where $\phi(u)$ is a convex function. Denote

$$
\sigma = \max_u \{ \| v'(u) \| : \phi(u) \leq f(x_0) \},
$$

$$
D = \max_u \{ \| u - u^* \| : \phi(u) \leq f(x_0) \}.
$$

**Theorem 5.**

1. If $f(x_0) - f^* \geq \frac{3}{2} L(\sigma D)^3$, then $f(x_1) - f^* \leq \frac{1}{2} L(\sigma D)^3$.

2. If $f(x_0) - f^* \leq \frac{3}{2} L(\sigma D)^3$, then $f(x_k) - f^* \leq \frac{3L(\sigma D)^3}{2(1 + \frac{1}{3} k)^2}$.

**Example.**

$$
\begin{align*}
u_1(x) &= x_1, \\ u_2(x) &= x_2 + \phi_1(x_1), \\ &\vdots \\ u_n(x) &= x_n + \phi_{n-1}(x_1, \ldots , x_{n-1}),
\end{align*}
$$

where $\phi_i(\cdot)$ are arbitrary functions.
Accelerated Newton: Cubic prox-function

Denote \( d(x) = \frac{1}{3} \| x - x_0 \| ^3 \).

**Lemma.** Cubic prox-function is *uniformly convex*: for all \( x, y \in \mathbb{R}^n \),

\[
\langle d'(x) - d'(y), x - y \rangle \geq \frac{1}{2} \| x - y \| ^3,
\]

\[
d(x) - d(y) - \langle d'(y), x - y \rangle \geq \frac{1}{6} \| x - y \| ^3,
\]

Moreover, its Hessian is Lipschitz continuous:

\[
\| d''(x) - d''(y) \| \leq 2 \| x - y \|, \; x, y \in \mathbb{R}^n.
\]

**Remark.** In our constructions, we are going to use \( d(x) \) instead of the standard *strongly convex* prox-functions.
We recursively update the following sequences.

- Sequence of estimate functions $\psi_k(x) = l_k(x) + \frac{N}{2} d(x)$, $k \geq 1$, where $l_k(x)$ are linear, and $N > 0$.
- A minimizing sequence $\{x_k\}_{k=1}^{\infty}$.
- A sequence of scaling parameters $\{A_k\}_{k=1}^{\infty}$: $A_{k+1} \overset{\text{def}}{=} A_k + a_k$, $k \geq 1$.

These objects have to satisfy the following relations:

\[ A_k f(x_k) \leq \psi_k^* \equiv \min_x \psi_k(x), \]
\[ (*) : \quad \psi_k(x) \leq A_k f(x) + (L_2 + \frac{1}{2}N) d(x), \quad \forall x \in \mathbb{R}^n, \]

for all $k \geq 1$. (⇒ $A_k(f(x_k) - f(x^*)) \leq (L + \frac{N}{2})d(x^*)$.)

For $k = 1$, we can choose $x_1 = T_{L_2}(x_0)$, $l_1(x) \equiv f(x_1)$, $A_1 = 1$. 
Denote $v_k = \arg\min_x \psi_k(x)$.

For some $a_k > 0$ and $M \geq 2L_2$, define

$$\alpha_k = \frac{a_k}{A_k + a_k} \in (0, 1),$$

$$y_k = (1 - \alpha_k)x_k + \alpha_k v_k,$$

$$x_{k+1} = T_M(y_k),$$

$$\psi_{k+1}(x) = \psi_k(x) + a_k[f(x_{k+1}) + \langle f'(x_{k+1}), x - x_{k+1} \rangle].$$

**Theorem.** For $M = 2L_2$, $N = 12L_2$, and $a_k = \frac{(k+1)(k+2)}{2}$, $k \geq 1$, relations (*) hold recursively.

**Corollary.** For any $k \geq 1$ we have $f(x_k) - f(x^*) \leq \frac{14L_2\|x_0 - x^*\|^3}{k(k+1)(k+2)}$. 
**Initialization:** Set $x_1 = T_{L_2}(x_0)$. Define $\psi_1(x) = f(x_1) + 6L_2 \cdot d(x)$.

**Iteration $k$, ($k \geq 1$):**

$v_k = \arg \min_{x \in \mathbb{R}^n} \psi_k(x),

y_k = \frac{k}{k+3}x_k + \frac{3}{k+3}v_k, \quad x_{k+1} = T_{2L_2}(y_k),

\psi_{k+1}(x) = \psi_k(x) + \frac{(k+1)(k+2)}{2} \left[ f(x_{k+1}) + \langle f'(x_{k+1}), x - x_{k+1} \rangle \right]

**Remark:**

Instead of recursive computation of $\psi_k(x)$, we can update only one vector:

$s_1 = 0, \quad s_{k+1} = s_k + \frac{(k+1)(k+2)}{2} f'(x_{k+1}), \quad k \geq 1.$

Then $v_k$ can be computed by an explicit expression.
Global non-degeneracy

**Standard setting:** for convex \( f \in C^2(R^n) \) define positive constants \( \sigma_1 \) and \( L_1 \) such that

\[
\sigma_1 \| h \|^2 \leq \langle f''(x)h, h \rangle \leq L_1 \| h \|^2
\]

for all \( x, y, h \in R^n \). The value \( \gamma_1(f) = \frac{\sigma_1}{L_1} \) is called the condition number of \( f \).

(Compatible with definition in Linear Algebra.)

**Geometric interpretation:**

\[
\frac{\langle f'(x), x - x^* \rangle}{\|f'(x)\| \cdot \|x - x^*\|} \geq \frac{2\sqrt{\gamma_1(f)}}{1+\gamma_1(f)}, \quad x \in R^n.
\]

**Complexity:** (1st-order methods)

\[
\text{PGM: } O \left( \frac{1}{\gamma_1(f)} \cdot \ln \frac{1}{\epsilon} \right), \quad \text{FGM: } O \left( \frac{1}{\sqrt{\gamma_1(f)}} \cdot \ln \frac{1}{\epsilon} \right).
\]

It does not work for 2nd-order schemes:

\[
f(x_k) - f^* \leq \frac{14 L_2 R^3}{k(k+1)(k+2)}.
\]
Global 2nd-order non-degeneracy

**Assumption:** for any \( x, y \in \mathbb{R}^n \), function \( f \in C^2(\mathbb{R}^n) \) satisfies inequalities

\[
\|f''(x) - f''(y)\| \leq L_2\|x - y\|,
\]
\[
\langle f'(x) - f'(y), x - y \rangle \geq \sigma_2\|x - y\|^3,
\]

where \( \sigma_2 > 0 \). We call the value \( \gamma_2(f) = \frac{\sigma_2}{L_2} \in (0, 1) \) the *2nd-order condition number* of function \( f \).

(Invariant w.r.t. addition of convex quadratic functions.)

**Example:** \( \gamma_2(d) = \frac{1}{4} \).

**Justification:** \( \frac{\sigma_2}{3} \|x_k - x^*\|^3 \leq f(x_k) - f^* \leq \frac{14L_2\|x_0 - x^*\|^3}{k(k+1)(k+2)} \).

Hence, in \( O\left(\frac{1}{[\gamma_2(f)]^{1/3}}\right) \) iterations we halve the distance to \( x^* \).

**Complexity bound:** (Accelerated CNM with restart)

\( O\left(\frac{1}{[\gamma_2(f)]^{1/3} \cdot \ln \frac{1}{\epsilon}}\right) \) iterations.
Open questions

1. Problem classes.
2. Lower complexity bounds and optimal methods.
3. Non-degenerate problems: geometric interpretation?
4. Complexity of strongly convex functions. (1st-order schemes?)
5. Consequences for polynomial-time methods.
1. Standard Gauss-Newton method

**Problem:** Find \( x \in \mathbb{R}^n \) satisfying the system \( F(x) = 0 \in \mathbb{R}^m \).

**Assumption:** \( \forall x, y \in \mathbb{R}^n \) \( \| F'(x) - F'(y) \| \leq L \| x - y \| \).

**Gauss-Newton method:** Choose a merit function \( \phi(u) \geq 0 \), \( \phi(0) = 0 \), \( u \in \mathbb{R}^m \).

Compute \( x_+ \in \text{Arg min}_y [\phi(F(x) + F'(x)(y - x))] \).

**Usual choice:** \( \phi(u) = \sum_{i=1}^{m} u_i^2 \). (Justification: Why not?)

**Remarks**

- Local quadratic convergence (\( m \geq n \), non-degeneracy and \( F(x^*) = 0 \)).
- If \( m < n \), then the method is not well-defined.
- No global complexity results.
Modified Gauss-Newton method

Lemma. For all $x, y \in \mathbb{R}^n$ we have

$$\|F(y) - F(x) - F'(x)(y - x)\| \leq \frac{1}{2}L\|y - x\|^2.$$ 

Corollary. Denote $f(y) = \|F(y)\|$. Then

$$f(y) \leq \|F(x) + F'(x)(y - x)\| + \frac{1}{2}L\|y - x\|^2.$$ 

Modified method:

$$x_{k+1} = \arg\min_y [\|F(x_k) + F'(x_k)(y - x_k) + \frac{1}{2}L\|y - x_k\|^2].$$ 

Remarks

- The merit function is non-smooth.
- Nevertheless, $f(x_{k+1}) < f(x_k)$ unless $x_k$ is a stationary point.
- Quadratic convergence for non-degenerate solutions.
- Global efficiency bounds.
- Problem of finding $x_{k+1}$ is convex.
- Different norms in $\mathbb{R}^n$ and $\mathbb{R}^m$ can be used.
Testing CNM: Chebyshev oscillator

Consider \( f(x) = \frac{1}{4}(1 - x^{(1)})^2 + \sum_{i=1}^{n-1} (x^{(i+1)} - p_2(x^{(i)}))^2 \), with \( p_2(\tau) = 2\tau^2 - 1 \).

Note that \( p_2 \) is a Chebyshev polynomial: \( p_k(\tau) = \cos(k \arccos(\tau)) \).

Hence, the equations for the “central path” is

\[
x^{(i+1)} = p_2(x^{(i)}) = p_4(x^{(i-1)}) = \cdots = p_{2^i}(x^{(1)}).
\]

This is an exponential oscillation! However, all coefficients and derivatives are small.

**NB:** \( f(x) \) is unimodular and \( x^* = (1, \ldots, 1) \).

In our experiments we usually take \( x_0 = (-1, 1, \ldots, 1) \).

Drawback: \( x_0 - 2\nabla f(x_0) = x^* \). Hence, sometimes we use \( x_0 = (-1, 0.9, \ldots, 0.9) \).
Solving Chebyshev oscillator by CN: $\|\nabla f(x)\|_2(2) \leq 10^{-8}$

<table>
<thead>
<tr>
<th>n</th>
<th>Iter</th>
<th>DF</th>
<th>GNorm</th>
<th>NumF</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>14</td>
<td>$7.0 \cdot 10^{-19}$</td>
<td>$4.2 \cdot 10^{-09}$</td>
<td>18</td>
<td>0.032</td>
</tr>
<tr>
<td>3</td>
<td>33</td>
<td>$1.1 \cdot 10^{-24}$</td>
<td>$7.5 \cdot 10^{-12}$</td>
<td>51</td>
<td>0.031</td>
</tr>
<tr>
<td>4</td>
<td>82</td>
<td>$1.7 \cdot 10^{-20}$</td>
<td>$9.3 \cdot 10^{-10}$</td>
<td>148</td>
<td>0.047</td>
</tr>
<tr>
<td>5</td>
<td>207</td>
<td>$4.5 \cdot 10^{-19}$</td>
<td>$1.2 \cdot 10^{-09}$</td>
<td>395</td>
<td>0.078</td>
</tr>
<tr>
<td>6</td>
<td>541</td>
<td>$1.0 \cdot 10^{-17}$</td>
<td>$5.6 \cdot 10^{-09}$</td>
<td>1062</td>
<td>0.266</td>
</tr>
<tr>
<td>7</td>
<td>1490</td>
<td>$1.4 \cdot 10^{-18}$</td>
<td>$2.9 \cdot 10^{-09}$</td>
<td>2959</td>
<td>0.609</td>
</tr>
<tr>
<td>8</td>
<td>4087</td>
<td>$2.7 \cdot 10^{-17}$</td>
<td>$9.1 \cdot 10^{-09}$</td>
<td>8153</td>
<td>1.782</td>
</tr>
<tr>
<td>9</td>
<td>11205</td>
<td>$1.6 \cdot 10^{-16}$</td>
<td>$9.6 \cdot 10^{-09}$</td>
<td>22389</td>
<td>5.922</td>
</tr>
<tr>
<td>10</td>
<td>30678</td>
<td>$2.7 \cdot 10^{-15}$</td>
<td>$9.6 \cdot 10^{-09}$</td>
<td>61335</td>
<td>18.89</td>
</tr>
<tr>
<td>11</td>
<td>79292</td>
<td>$7.7 \cdot 10^{-14}$</td>
<td>$1.0 \cdot 10^{-08}$</td>
<td>158563</td>
<td>57.813</td>
</tr>
<tr>
<td>12</td>
<td>171522</td>
<td>$9.7 \cdot 10^{-13}$</td>
<td>$9.9 \cdot 10^{-09}$</td>
<td>343026</td>
<td>144.266</td>
</tr>
<tr>
<td>13</td>
<td>385353</td>
<td>$1.3 \cdot 10^{-11}$</td>
<td>$9.9 \cdot 10^{-09}$</td>
<td>770691</td>
<td>347.094</td>
</tr>
<tr>
<td>14</td>
<td>938758</td>
<td>$2.1 \cdot 10^{-11}$</td>
<td>$1.0 \cdot 10^{-08}$</td>
<td>1877500</td>
<td>1232.953</td>
</tr>
<tr>
<td>15</td>
<td>2203700</td>
<td>$7.8 \cdot 10^{-11}$</td>
<td>$1.0 \cdot 10^{-08}$</td>
<td>4407385</td>
<td>3204.359</td>
</tr>
</tbody>
</table>
## Other methods

<table>
<thead>
<tr>
<th>$n$</th>
<th>Trust region</th>
<th>Knitro</th>
<th>Minos 5.5</th>
<th>Snopt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Inner Iter</td>
<td>Iter</td>
<td>Iter NFG</td>
<td>Iter# NFG</td>
</tr>
<tr>
<td>3</td>
<td>129 50</td>
<td>30</td>
<td>44 120</td>
<td>106 78</td>
</tr>
<tr>
<td>4</td>
<td>431 123</td>
<td>80</td>
<td>136 309</td>
<td>268 204</td>
</tr>
<tr>
<td>5</td>
<td>1310 299</td>
<td>203</td>
<td>339 793</td>
<td>647 509</td>
</tr>
<tr>
<td>6</td>
<td>3963 722</td>
<td>531</td>
<td>871 2022</td>
<td>1417 1149*</td>
</tr>
<tr>
<td>7</td>
<td>12672 1921</td>
<td>1467</td>
<td>2291 5404</td>
<td>* ***</td>
</tr>
<tr>
<td>8</td>
<td>40036 5234</td>
<td>4040</td>
<td>6109 14680</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>120873 13907</td>
<td>11062</td>
<td>11939 28535</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>358317 36837</td>
<td>29729*</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>842368 78854</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>2121780 182261</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notation:** * early termination, (*** ) numerical difficulties/inaccurate solution, # needs an alternative starting point.

**Trust region:** very reliable, but $T(12) = 2577$ sec (Matlab), $T(n) = Const \times (4.5)^n$. 
Lecture 3. Structural Optimization: Interior Point Methods

Yuri Nesterov, CORE/INMA (UCL)

January 20-22, 2016 (Ecole Polytechnique, Paris)
Outline

1. Interior-point methods: standard problem
2. Newton method
3. Self-concordant functions and barriers
4. Minimizing self-concordant functions
5. Conic optimization problems
6. Primal-dual barriers
7. Primal-dual central path and path-following methods
Black-Box Methods: Main assumptions represent the bounds for the size of certain derivatives.

Example

Consider the function \( f(x_1, x_2) = \begin{cases} \frac{x_2^2}{x_1}, & x_1 > 0, \\ 0, & x_1 = x_2 = 0. \end{cases} \)

It is closed, convex, but discontinuous at the origin.

However, its epigraph \( \{ x \in \mathbb{R}^3 : x_1 x_3 \geq x_2^2 \} \) is a simple convex set:

\[
x_1 = u_1 + u_3, \quad x_2 = u_2, \quad x_3 = u_1 - u_3 \quad \Rightarrow \quad u_1 \geq \sqrt{u_2^2 + u_3^2}.
\]

(Lorentz cone)

Question: Can we always replace the functional components by convex sets?
Problem: \( f^* = \min_{x \in Q} \langle c, x \rangle \), where \( Q \subset E \) is a closed convex set with nonempty interior.

How we can measure the quality of \( x \in Q \)?

1. The residual \( \langle c, x \rangle - f^* \) is not very informative since it does not depend on position of \( x \) inside \( Q \).

2. The boundary of a convex set can be very complicated.

3. It is easy to travel inside provided that we keep a sufficient distance to the boundary.

Conclusion: we need a barrier function \( f(x) \):

- \( \text{dom } f = \text{int } Q \),
- \( f(x) \to \infty \) as \( t \to \partial Q \).
Central path: for $t > 0$ define $x^*(t), \quad tc + f'(x^*(t)) = 0$

(\text{hence } x^*(t) = \arg \min_x \left[ \Psi_t(x) \overset{\text{def}}{=} t\langle c, x \rangle + f(x) \right].)\]

Lemma. Suppose $\langle f'(x), y - x \rangle \leq A$ for all $x, y \in \text{dom } Q$. Then

$\langle c, x^*(t) - x^* \rangle = \frac{1}{t} \langle f'(x^*(t)), x^* - x^*(t) \rangle \leq \frac{1}{t} A.$

Method: $t_k > 0, x^k \approx x^*(t_k) \Rightarrow t_{k+1} > t_k, \quad x^{k+1} \approx x^*(t^{k+1}).$

For approximating $x^*(t^{k+1})$, we need a powerful minimization scheme.

**Main candidate:** Newton Method.

(\text{Very good local convergence.})
Classical results on the Newton Method

Method: \[ x^{k+1} = x^k - [f''(x^k)]^{-1} f'(x^k). \]

Assume that:

- \( f''(x^*) \geq \ell \cdot I_n \)
- \[ \|f''(x) - f''(y)\| \leq M\|x - y\|, \quad \forall x, y \in \mathbb{R}^n. \]
- The starting point \( x^0 \) is close to \( x^* \): \( \|x^0 - x^*\| < \bar{r} = \frac{2\ell}{3M}. \)

Then \( \|x^k - x^*\| < \bar{r} \) for all \( k \), and the Newton method converges quadratically: \[ \|x^{k+1} - x^*\| \leq \frac{M\|x^k - x^*\|^2}{2(\ell - M\|x^k - x^*\|)}. \]

Note:

- The description of the region of quadratic convergence is given in terms of the metric \( \langle \cdot, \cdot \rangle \).
- The resulting neighborhood is changing when we choose another metric.
Let $f(x)$ satisfy our assumptions. Consider $\phi(y) = f(Ay)$, where $A$ is a non-degenerate $(n \times n)$-matrix.

**Lemma:** Let $\{x^k\}$ be a sequence, generated by Newton Method for function $f$.
Consider the sequence $\{y^k\}$, generated by the Newton Method for function $\phi$ with $y^0 = A^{-1}x^0$.
Then $y^k = A^{-1}x^k$ for all $k \geq 0$.

**Proof:** Assume $y^k = A^{-1}x^k$ for some $k \geq 0$. Then

$$y^{k+1} = y^k - \left[\phi''(y^k)\right]^{-1}\phi'(y^k)$$

$$= y^k - \left[A^T f''(Ay^k)A\right]^{-1}A^T f'(Ay^k)$$

$$= A^{-1}x^k - A^{-1}[f''(x^k)]^{-1}f'(x^k) = A^{-1}x^{k+1}. \quad \square$$

**Conclusion:** The method is *affine invariant*. Its region of quadratic convergence *does not depend on the metric!*
What was wrong?

**Old assumption:** \[ \| f''(x) - f''(y) \| \leq M \| x - y \|. \]

Let \( f \in C^3(R^n) \). Denote \( f'''(x)[u] = \lim_{\alpha \to 0} \frac{1}{\alpha} [f''(x + \alpha u) - f''(x)]. \)

This is a matrix!

Then the old assumption is equivalent to: \[ \| f'''(x)[u] \| \leq M \| u \|. \]

Hence, at any point \( x \in R^n \) we have

\[ (\star) : \quad |\langle f'''(x)[u]v, v \rangle| \leq M \| u \| \cdot \| v \|^2 \text{ for all } u, v \in R^n. \]

**Note:**

- The LHS of (\( \star \)) is an *affine invariant* directional derivative.
- The norm \( \| \cdot \| \) has nothing common with our particular \( f \).
- However, there exists a local norm, which is closely related to \( f \). This is \( \| u \|_{f''(x)} = \langle f''(x)u, u \rangle^{1/2} \).
- Let us make a similar assumption in terms of \( \| \cdot \|_{f''(x)}. \)
Let $f(x) \in C^3(\text{dom } f)$ be a closed and convex, with open domain. Let us fix a point $x \in \text{dom } f$ and a direction $u \in R^n$.

Consider the function $\phi(x; t) = f(x + tu)$. Denote

$$Df(x)[u] = \phi'_t(x; 0) = \langle f'(x), u \rangle,$$

$$D^2 f(x)[u, u] = \phi''_{tt}(x; 0) = \langle f''(x)u, u \rangle = \| u \|_{f''(x)}^2,$$

$$D^3 f(x)[u, u, u] = \phi'''_{ttt}(x; 0) = \langle f'''(x)[u]u, u \rangle.$$

**Def.** We call function $f$ self-concordant if the inequality $| D^3 f(x)[u, u, u] | \leq 2 \| u \|_{f''(x)}^3$ holds for any $x \in \text{dom } f, u \in R^n$.

**Note:**

- We cannot expect that these functions are very common.
- We hope that they are good for the Newton Method.
Examples

1. Linear function is s.c. since $f''(x) \equiv 0$, $f'''(x) \equiv 0$

2. Convex quadratic function is s.c. ($f'''(x) \equiv 0$).

3. Logarithmic barrier for a ray $\{x > 0\}$:
   
   $f(x) = -\ln x$, $f'(x) = -\frac{1}{x}$, $f''(x) = \frac{1}{x^2}$, $f'''(x) = -\frac{2}{x^3}$.

4. Logarithmic barrier for a quadratic region. Consider a concave function $\phi(x) = \alpha + \langle a, x \rangle - \frac{1}{2} \langle Ax, x \rangle$. Define $f(x) = -\ln \phi(x)$.

   
   $Df(x)[u] = -\frac{1}{\phi(x)} [\langle a, u \rangle - \langle Ax, u \rangle] \overset{\text{def}}{=} \omega_1$,

   $D^2f(x)[u]^2 = \frac{1}{\phi^2(x)} [\langle a, u \rangle - \langle Ax, u \rangle]^2 + \frac{1}{\phi(x)} \langle Au, u \rangle$,

   $D^3f(x)[u]^3 = -\frac{2}{\phi^3(x)} [\langle a, u \rangle - \langle Ax, u \rangle]^3 - \frac{3 \langle Au, u \rangle}{\phi^2(x)} [\langle a, u \rangle - \langle Ax, u \rangle]$.

   $D_2 = \omega_1^2 + \omega_2$, $D_3 = 2\omega_1^3 - 3\omega_1\omega_2$. Hence, $|D_3| \leq 2|D_2|^{3/2}$. 
Simple properties

1. If $f_1, f_2$ are s.c.f., then $f_1 + f_2$ is s.c. function.

2. If $f(y)$ is s.c.f., then $\phi(x) = f(Ax + b)$ is also a s.c. function.

**Proof:** Denote $y = y(x) = Ax + b$, $v = Au$. Then

$$D\phi(x)[u] = \langle f'(y(x)), Au \rangle = \langle f'(y), v \rangle,$$

$$D^2\phi(x)[u]^2 = \langle f''(y(x))Au, Au \rangle = \langle f''(y)v, v \rangle,$$

$$D^3\phi(x)[u]^3 = D^3f(y(x))[Au]^3 = D^3f(y)[v]^3.\Box$$

**Example:** $f(x) = \langle c, x \rangle - \sum_{i=1}^{m} \ln(a_i - \|A_ix - b_i\|^2)$ is a s.c.-function.
Let $x \in \text{dom } f$ and $u \in \mathbb{R}^n$, $u \neq 0$. For $x + tu \in \text{dom } f$, consider
\[ \phi(t) = \frac{1}{\langle f''(x+tu)u, u \rangle^{1/2}}. \]

**Lemma 1.** For all feasible $t$ we have: $|\phi'(t)| \leq 1$.

**Proof:** Indeed, $\phi'(t) = -\frac{f'''(x+tu)[u]^3}{2\langle f''(x+tu)u, u \rangle^{3/2}}$. \(\square\)

**Corollary 1:** $\text{dom } \phi$ contains the interval $(-\phi(0), \phi(0))$.

**Proof:** Since $f(x + tu) \to \infty$ as $x + tu \to \partial \text{dom } f$, the same is true for $\langle f''(x + tu)u, u \rangle$. Hence $\text{dom } \phi(t) \equiv \{ t \mid \phi(t) > 0 \}$. \(\square\)

Denote $\|h\|^2_x = \langle f''(x)h, h \rangle$, $W^0(x; r) = \{ y \in \mathbb{R}^n \mid \| y - x \|_x < r \}$. Then
\[ W^0(x; r) \subseteq \text{dom } f \text{ for } r < 1. \]
Denote $W(x; r) = \{y \in \mathbb{R}^n \mid \|y - x\|_x < r\}$.

**Theorem.** For all $x, y \in \text{dom } f$ the following inequality holds:

$$\| y - x \|_y \geq \frac{\|y - x\|_x}{1 + \|y - x\|_x}.$$ 

If $\| y - x \|_x < 1$ then $\| y - x \|_y \leq \frac{\|y - x\|_x}{1 - \|y - x\|_x}$.

**Proof.** 1. Let us choose $u = y - x$. Then

$$\phi(1) = \frac{1}{\|y - x\|_y}, \quad \phi(0) = \frac{1}{\|y - x\|_x},$$

and $\phi(1) \leq \phi(0) + 1$ in view of Lemma 1.

2. If $\| y - x \|_x < 1$, then $\phi(0) > 1$, and in view of Lemma 1, $\phi(1) \geq \phi(0) - 1$. 

□
Theorem. For any $x, y \in \text{dom } f$ we have:
\[
\langle f'(y) - f'(x), y - x \rangle \geq \frac{\|y - x\|^2_x}{1 + \|y - x\|_x},
\]
\[
f(y) \geq f(x) + \langle f'(x), y - x \rangle + \omega(\|y - x\|_x),
\]
where $\omega(t) = t - \ln(1 + t)$.

Proof. Denote $y_\tau = x + \tau(y - x)$, $\tau \in [0, 1]$, and $r = \|y - x\|_x$.
\[
\langle f'(y) - f'(x), y - x \rangle = \int_0^1 \langle f''(y_\tau)(y - x), y - x \rangle d\tau
\]
\[
= \int_0^1 \frac{1}{\tau^2} \| y_\tau - x \|_y^2 d\tau \geq \int_0^1 \frac{r^2}{(1 + \tau r)^2} d\tau = r \int_0^r \frac{1}{(1 + t)^2} dt = \frac{r^2}{1 + r}
\]
\[
f(y) - f(x) - \langle f'(x), y - x \rangle = \int_0^1 \langle f'(y_\tau) - f'(x), y - x \rangle d\tau
\]
\[
\geq \int_0^1 \frac{\|y_\tau - x\|^2_x}{\tau(1 + \|y_\tau - x\|_x)} d\tau = \int_0^1 \frac{\tau r^2}{1 + \tau r} d\tau = r \int_0^r \frac{tdt}{1 + t} = \omega(r). \quad \Box
**Theorem.** Let $x \in \text{dom } f$ and $\| y - x \|_x < 1$. Then
\[
\langle f'(y) - f'(x), y - x \rangle \leq \frac{\|y-x\|_x^2}{1-\|y-x\|_x},
\]
\[
f(y) \leq f(x) + \langle f'(x), y - x \rangle + \omega_*(\| y - x \|_x),
\]
where $\omega_*(t) = -t - \ln(1-t)$.

**Main Theorem:** for any $y \in W(x; r), \ r \in [0, 1)$, we have
\[
(1 - r)^2F''(x) \preceq F''(y) \preceq \frac{1}{(1-r)^2}F''(x).
\]

**Corollary.** For $G = \int_0^1 f''(x + \tau(y - x))d\tau$, we have
\[
(1 - r + \frac{r^2}{3})f''(x) \preceq G \preceq \frac{1}{1-r}f''(x).
\]

**Observation:** If $\text{dom } f$ contains no straight line, then $f''(x) \succ 0$ for any $x \in \text{dom } f$. (If not, then $W(x, 1)$ is unbounded.)
Minimizing the self-concordant function

Consider the problem: \( \min \{ f(x) \mid x \in \text{dom } f \} \). Assume \( \text{dom } f \) contains no straight line.

Theorem. Let \( \lambda_f(x) < 1 \) for some \( x \in \text{dom } f \). Then the solution of this problem \( x_f^* \), exists and unique.

Proof. Indeed, for any \( y \in \text{dom } f \) we have:

\[
\begin{align*}
f(y) & \geq f(x) + \langle f'(x), y - x \rangle + \omega(\| y - x \|_x) \\
& \geq f(x) - \| f'(x) \|_x^* \cdot \| y - x \|_x + \omega(\| y - x \|_x) \\
& = f(x) - \lambda_f(x) \cdot \| y - x \|_x + \omega(\| y - x \|_x).
\end{align*}
\]

Since \( \omega(t) = t - \ln(1 + t) \), the level sets are bounded. \( \Rightarrow \exists x_f^* \).

It is unique since in since \( f(y) \geq f(x_f^*) + \omega(\| y - x_f^* \|_{x_f^*}) \), and \( f''(x_f^*) \) is nondegenerate.

Example: \( f(x) = (1 - \epsilon)x - \ln x \) with \( \epsilon \in (0, 1) \) and \( x = 1 \).
Consider the following scheme: \( x_0 \in \text{dom } f \),

\[ x_{k+1} = x_k - \frac{1}{1+\lambda_f(x_k)} [f''(x_k)]^{-1} f'(x_k). \]

**Theorem.** For any \( k \geq 0 \) we have \( f(x_{k+1}) \leq f(x_k) - \omega(\lambda_f(x_k)) \).

**Proof.** Denote \( \lambda = \lambda_f(x_k) \). Then \( \| x_{k+1} - x_k \|_x = \frac{\lambda}{1+\lambda} \). Therefore,

\[
\begin{align*}
f(x_{k+1}) &\leq f(x_k) + \langle f'(x_k), x_{k+1} - x_k \rangle + \omega_*(\| x_{k+1} - x_k \|_x) \\
&= f(x_k) - \omega(\lambda). \quad \square
\end{align*}
\]

**Consequence:** we come to the region \( \lambda_f(x_k) \leq \text{const in } O(f(x_0) - f^*) \) iterations.
For \(x\) close to \(x^*\), \(f'(x^*) = 0\), function \(f(x)\) is almost quadratic:

\[
f(x) \approx f^* + \frac{1}{2} \langle f''(x^*)(x - x^*), x - x^* \rangle.
\]

Therefore, \(f(x) - f^* \approx \frac{1}{2} \|x - x^*\|_{x^*}^2 \approx \frac{1}{2} \|x - x^*\|_x^2\)

\[\approx \frac{1}{2} \langle f'(x), [f''(x)]^{-1}f'(x) \rangle \overset{\text{def}}{=} \frac{1}{2}(\|f'(x)\|_{x^*})^2 \overset{\text{def}}{=} \lambda_f^2(x).
\]

The last value is the local norm of the gradient. It is computable.

**Theorem:** Let \(x \in \text{dom } f\) and \(\lambda_f(x) < 1\).

Then the point \(x_+ = x - [f''(x)]^{-1}f'(x)\) belongs to \(\text{dom } f\) and

\[\lambda_f(x_+) \leq \left(\frac{\lambda_f(x)}{1 - \lambda_f(x)}\right)^2.\]
Proof

Denote \( p = x_+ - x \), \( \lambda = \lambda_f(x) \). Then \( \| p \|_x = \lambda < 1 \), \( x_+ \in \text{dom } f \).

\[
\lambda_f(x_+) \leq \frac{1}{1-\|p\|_x} \| f'(x_+) \|_x = \frac{1}{1-\lambda} \| f'(x_+) \|_x.
\]

Note that \( f'(x_+) = f'(x_+) - f'(x) - f''(x)(x_+ - x) = Gp \), where

\[
G = \int_0^1 [f''(x + \tau p) - f''(x)] d\tau.
\]

Therefore

\[
\| f'(x_+) \|_x^2 = \langle [f''(x)]^{-1} Gp, Gp \rangle \leq \| H \|^2 \cdot \| p \|^2_x,
\]

where \( H = [f''(x)]^{-1/2} G[f''(x)]^{-1/2} \). In view of Corollary,

\[
(-\lambda + \frac{1}{3} \lambda^2) f''(x) \leq G \leq \frac{\lambda}{1-\lambda} f''(x).
\]

Therefore \( \| H \| \leq \max \left\{ \frac{\lambda}{1-\lambda}, \lambda - \frac{1}{3} \lambda^2 \right\} = \frac{\lambda}{1-\lambda} \), and

\[
\lambda_f^2(x_+) \leq \frac{1}{(1-\lambda)^2} \| f'(x_+) \|_x^2 \leq \frac{\lambda^4}{(1-\lambda)^4}.
\]

\[\square\]

\textbf{NB:} Region of quadratic convergence is \( \lambda_f(x) < \bar{\lambda} \), \( \frac{\bar{\lambda}}{(1-\lambda)^2} = 1 \).

It is affine-invariant!
Following the central path

Consider $\Psi_t(x) = t\langle c, x \rangle + f(x)$ with s.c. function $f$.

- For $\Psi_t$, Newton Method has local quadratic convergence.
- The region of quadratic convergence (RQC) is given by
  $$\lambda_{\Psi_t}(x) \leq \beta < \bar{\lambda}.$$  

Assume we know $x = x^*(t)$. We want to update $t$, $t_+ = t + \Delta$, keeping $x$ in RQC of function $\Psi_{t+\Delta}$: $\lambda_{\Psi_{t+\Delta}}(x) \leq \beta$.

**Question:** How large can be $\Delta$? Since $tc + f'(x) = 0$, we have:

$$\lambda_{\Psi_{t+\Delta}}(x) = \| t + c + f'(x) \|^*_x = |\Delta| \cdot \| c \|^*_x = \frac{|\Delta|}{t} \| f'(x) \|^*_x \leq \beta.$$  

**Conclusion:** for the linear rate, we need to assume that

$$\langle [f''(x)]^{-1} f'(x), f'(x) \rangle$$  

is uniformly bounded on $\text{dom} \ f$.

Thus, we come to the definition of self-concordant barrier.
Let $F(x)$ be a s.c.-function. It is a $\nu$-self-concordant barrier, if
\[
\max_{u \in \mathbb{R}^n} [2\langle F'(x), u \rangle - \langle F''(x)u, u \rangle] \leq \nu \text{ for all } x \in \text{dom } F.
\]
The value $\nu$ is called the parameter of the barrier.

If $F''(x)$ is non-degenerate, then $\langle F'(x), [F''(x)]^{-1}F'(x) \rangle \leq \nu$.

Another form: $\langle F'(x), u \rangle^2 \leq \nu \langle F''(x)u, u \rangle$.

**Main property:** $\langle F'(x), y - x \rangle \leq \nu$, $x, y \in \text{int } Q$.

**NB:** $\nu$ is responsible for the rate of p.-f. method: $t_+ = t \pm \frac{\alpha \cdot t}{\nu^{1/2}}$.

**Complexity:** $O\left(\sqrt{\nu} \ln \frac{\nu}{\varepsilon}\right)$ iterations of the Newton method.

**Calculus:**
1. Affine transformations do not change $\nu$.
2. Restriction on a subspace can only decrease $\nu$.
3. $F = F_1 + F_2 \Rightarrow \nu = \nu_1 + \nu_2$. 
Examples

1. Barrier for a ray: \( F(t) = -\ln t \), \( F'(t) = -\frac{1}{t} \), \( F''(t) = \frac{1}{t^2} \), \( \nu = 1 \).

2. Polytop \( \{ x : \langle a_i, x \rangle \leq b_i \} \), \( F(x) = -\sum_{i=1}^{m} \ln(b_i - \langle a_i, x \rangle) \), \( \nu = m \).

3. \( l_2 \)-ball: \( F(x) = -\ln(1 - \| x \|^2) \), \( D_1 = \omega_1 \), \( D_2 = \omega_1^2 + \omega_2 \), \( \nu = 1 \).

4. Intersection of ellipsoids: \( F(x) = -\sum_{i=1}^{m} \ln(r_i^2 - \| A_i x - b_i \|^2) \), \( \nu = m \).

5. Epigraph \( \{ t \geq e^x \} \), \( F(x, t) = -\ln(t - e^x) - \ln(\ln t - x) \), \( \nu = 4 \).

6. Universal barrier. Define the polar set
\[
P(x) = \{ s : \langle s, y - x \rangle \leq 1, \ y \in Q \}.
\]
Then \( F(x) = -\ln \text{vol}_n P(x) \) is an \( O(n) \)-s.c. barrier for \( Q \).

7. Lorentz cone \( \{ t \geq \| x \| \} \), \( F(x, t) = -\ln(t^2 - \| x \|^2) \), \( \nu = 2 \).

8. LMI-cone \( \{ X = X^T \succeq 0 \} \), \( F(X) = -\ln \det X \), \( \nu = n \).
Problem: \( f^* = \min \{ \langle c, x \rangle : Ax = b, \ x \in K \} \), where

\[ A \in R^{m \times n} : R^n \to R^m, \ m < n, \ c \in R^n, \ b \in R^m. \]

\( K, \ \text{int} K \neq \emptyset, \) is a closed convex pointed cone:

1. \( \forall x_1, x_2 \in K \implies x_1 + x_2 \in K. \)
2. \( \forall x \in K, \ \tau \geq 0 \implies \tau x \in K. \)
3. \( K \) contains no straight line.

Assumptions:

\[ A \text{ is nondegenerate, } b \neq 0. \]

\[ \text{There is no } y \in R^m \text{ such that } c = A^T y. \]

Explanations:

\[ \text{If } b = 0 \text{ then either } f^* = 0 \text{ or } f^* = -\infty. \]

\[ \text{If } \exists y \in R^m : c = A^T y \text{ then } \forall x, \ Ax = b, \ \text{we have:} \]
\[ \langle c, x \rangle = \langle A^T y, x \rangle = \langle y, Ax \rangle = \langle b, y \rangle. \]
Main Assumption:

There exist a computable $\nu$-normal barrier $F(x)$ for $K$ such that

- $F(x)$ is a $\nu$-self-concordant barrier for $K$.
- $F(x)$ is logarithmically homogeneous: $\forall x \in \text{int} K, \tau > 0$
  \[
  (*) \quad F(\tau x) = F(x) - \nu \ln \tau.
  \]

Examples:

1. Positive orthant: $K = \mathbb{R}^n_+ \overset{\text{def}}{=} \{ x \in \mathbb{R}^n : x(i) \geq 0, \ i = 1 \ldots n \}$,
   
   $F(x) = -\sum_{i=1}^{n} \ln x(i), \quad \nu = n.$

2. Cone of positive semidefinite matrices:
   
   $K = S^n_+ \overset{\text{def}}{=} \{ X \in \mathbb{R}^{n \times n} : X = X^T, \langle Xu, u \rangle \geq 0 \ \forall u \in \mathbb{R}^n \}$,
   
   $F(X) = -\ln \det X, \quad \nu = n.$

3. 2nd order cone: $K = \mathcal{L}_n \overset{\text{def}}{=} \{ z = (x, \tau) \in \mathbb{R}^{n+1} : \tau \geq \| x \| \}$,
   
   $F(z) = -\ln(\tau^2 - \| x \|^2), \quad \nu = 2.$

4. Direct sums of these cones.
Then for any $x \in \text{int} \ K$ and $\tau > 0$ we have:

(1) : $F'(\tau x) = \frac{1}{\tau} F'(x)$,  \hspace{1cm} (2) : $F''(\tau x) = \frac{1}{\tau^2} F''(x)$.
(3) : $\langle F'(x), x \rangle = -\nu$, \hspace{1cm} (4) : $F''(x)x = -F'(x)$.
(5) : $\langle F''(x)x, x \rangle = \nu$, \hspace{1cm} (6) : $\langle [F''(x)]^{-1} F'(x), F'(x) \rangle = \nu$.

Proof:
1. Differentiate (*) in $x$: $\tau F'(\tau x) = F'(x)$.
2. Differentiate 1) in $x$: $\tau F''(\tau x) = \frac{1}{\tau} F''(x)$.
3. Diff. (*) in $\tau$: $\langle F'(\tau x), x \rangle = -\frac{\nu}{\tau}$. Take $\tau = 1$ and we get 3).
4. Differentiate 3) in $x$: $F''(x)x + F'(x) = 0$.
5. Substitute 4) in 3).
**Definition.** Let $K$ be a closed convex cone. The set
$$K^* = \{ s : \langle s, x \rangle \geq 0 \ \forall x \in K \}$$
is called the *dual cone* to $K^*$.

**Theorem.** If $K$ is a proper cone, then $K^*$ is also proper and $(K^*)^* = K$.

**Proof:** $K^*$ is closed and convex as an intersection of half-spaces. If $K^*$ contains a straight line $\{ s = \tau \bar{s}, \tau \in \mathbb{R} \}$, then $\langle \bar{s}, x \rangle = 0 \ \forall x \in K$ (contradiction).

For all $s \in K^*$ and $x \in K$ we have: $\langle s, x \rangle \geq 0$. Therefore $K \subseteq (K^*)^*$. If $\exists u \in (K^*)^* \setminus K$ then
$$\exists \bar{s} : \langle \bar{s}, u \rangle < \langle \bar{s}, x \rangle \ \forall x \in K.$$Hence, $\bar{s} \in K^*$ and $u \notin (K^*)^*$. Contradiction.

If $\text{int } K^* = \emptyset$ then $\exists \bar{x} : \langle s, \bar{x} \rangle = 0, \ \forall s \in K^*$. Therefore $\pm \bar{x} \in (K^*)^* \equiv K^*$. Contradiction.
**Conjugate barriers**

**Definition.** Let $K$ be a proper cone and $F(x)$ be a $\nu$-s.c.b. for $K$. The function

$$F_*(s) = \max\{-\langle s, x \rangle - F(x) : x \in K\}$$

is called conjugate (or dual) barrier.

**Main properties**

- $\text{dom } F_*(s) \equiv \text{int } K^*$.
- $F_*(s)$ is a $\nu$-normal barrier for $K^*$.
- For any $x \in \text{int } K$ and $s \in \text{int } K^*$ we have:

  $$F(x) + F_*(s) \geq -\nu \ln \langle s, x \rangle - \nu + \nu \ln \nu.$$  

  Equality is attained iff $s = -\tau F'(x)$ for some $\tau > 0$.

**Examples:** The barriers for $R^n_+$, $\mathcal{L}_n$ and $S^n_+$ are *self-dual*. 
Primal–Dual Problems

**Primal problem:** \( f^* = \min_x \{ \langle c, x \rangle : Ax = b, \ x \in K \} \).

**Dual problem:** \( f_* = \max_{y, s} \{ \langle b, y \rangle : s + A^T y = c, \ s \in K^* \} \).

Denote \( F_D \) the feasible set of the dual problem.

**Note:**

- For any \( x \in F_P \) and \((s, y) \in F_D\) we have:
  \[
  0 \leq \langle s, x \rangle = \langle c - A^T y, x \rangle = \langle c, x \rangle - \langle b, y \rangle.
  \]
- Therefore we always have: \( f^* \geq f_* \).
There exists a strictly feasible primal-dual solution \((\bar{x}, \bar{s}, \bar{y})\):
\[
A\bar{x} = b, \quad \bar{x} \in \text{int} \ K, \quad \bar{s} + A^T \bar{y} = c, \quad \bar{s} \in \text{int} \ K^*.
\]

**Primal central path:** \(x(t) = \arg \min \{t\langle c, x \rangle + F(x) : Ax = b\} \).

**Dual central path:**
\[
(s(t), y(t)) = \arg \min \{-t\langle b, y \rangle + F_*(s) : s + A^T y = c\}.
\]

**Primal–dual central path:** \((x(t), s(t), y(t)), t > 0\).

**Lemma.** The primal-dual central path is well defined.

**Proof:** Note that \(\forall x \in \mathcal{F}_P\)
\[
F(x) \geq -t\langle \bar{s}, x \rangle - F_*(t\bar{s}) = -t(\langle c, x \rangle - \langle b, \bar{y} \rangle) - F_*(t\bar{s}).
\]
Therefore \(t\langle c, x \rangle + F(x) \geq t\langle b, \bar{y} \rangle - F_*(t\bar{s})\).

Thus, \(x(t)\) exists. The proof for the dual path is symmetric. \(\Box\)
Properties of primal-dual central path

**Theorem.** For any $t > 0$ we have the following:

\[
\langle c, x(t) \rangle - \langle b, y(t) \rangle = \frac{\nu}{t},
\]

\[
F'(x(t)) = -\frac{1}{t} F'(s(t)), \quad F'(s(t)) = -\frac{1}{t} F'(x(t)),
\]

\[
F(x(t)) + F_*(s(t)) = -\nu + \nu \ln t.
\]

**Proof:** Let us write down the optimality conditions for $x(t)$:

\[
tc + F'(x(t)) = A^T \tilde{y}(t), \quad Ax(t) = b.
\]

Denote $\hat{s}(t) = -\frac{1}{t} F'(x(t)), \hat{y}(t) = \frac{1}{t} \tilde{y}(t)$. Then

\[
x(t) = -F'(t \hat{s}(t)) = -\frac{1}{t} F'_*(\hat{s}(t)).
\]

Thus, $c = \hat{s}(t) + A^T \hat{y}(t), \quad tb + AF'_*(\hat{s}(t)) = 0$.

This is the optimality conditions for the dual path. In view of uniqueness, we have $s(t) = -\frac{1}{t} F'(x(t))$.

The rest: $\langle c, x(t) \rangle - \langle b, y(t) \rangle = \langle s(t), x(t) \rangle = \frac{\nu}{t},$

\[
F(x(t)) + F_*(s(t)) = F(x(t)) + F_*(-F'(x(t))) + \nu \ln t = -\nu + \nu \ln t. \qed
\]
Remarks

- Under our Main Assumption $f^* = f_*$.
- The set $\mathcal{F}_P \times \mathcal{F}_D$ is never bounded.
- We have complete characterization for the duality gap $\langle c, x \rangle - \langle b, y \rangle$ and the barrier $F(x) + F_*(s)$ along the central path.
- This information forms the basis for all primal-dual schemes.
- That is not for free: *We assume that $F_*(s)$ is computable.*
Primal–dual potential

\[ \Phi(x, s) = 2\nu \ln \langle s, x \rangle + F(x) + F^*(s) \]
\[ = 2\nu \ln [\langle c, x \rangle - \langle b, y \rangle] + F(x) + F^*(s). \]

**Lemma.** For any \((x, s, y) \in \mathcal{F}_{PD}^0 \equiv \mathcal{F}_P^0 \times \mathcal{F}_D^0\) we have:
\[ \langle c, x \rangle - \langle b, y \rangle \leq \frac{1}{\nu} \exp \left\{ 1 + \frac{1}{\nu} \Phi(x, s) \right\}. \]

**Proof:**
\[ \Phi(x, s) = 2\nu \ln \langle s, x \rangle + F(x) + F^*(s) \]
\[ \geq 2\nu \ln \langle s, x \rangle - \nu + \nu \ln \nu - \nu \ln \langle s, x \rangle \]
\[ = \nu \ln [\langle c, x \rangle - \langle b, y \rangle] - \nu + \nu \ln \nu. \quad \Box \]

**Main question:** What can be the rate of decrease of \(\Phi(x, s)\)?
Decrease along the central path \( z(t) = (x(t), s(t), y(t)) \)

We want to have \( \| z(t + \Delta t) - z(t) \|_{z(t)} \leq 1 \).

This is approximately \( | \Delta t | \leq \frac{1}{\| z'(t) \|_{z(t)}} \).

Note that \( \langle s'(t), x'(t) \rangle = 0 \) and
\[
\begin{align*}
    s'(t) &= -\frac{1}{t} s(t) - \frac{1}{t} F''(x(t)) x'(t), \\
    x'(t) &= -\frac{1}{t} x(t) - \frac{1}{t} F''(s(t)) s'(t).
\end{align*}
\]

Therefore
\[
\begin{align*}
    \langle F''(x(t)) x'(t), x'(t) \rangle &= -\langle s(t), x'(t) \rangle, \\
    \langle F''(s(t)) s'(t), s'(t) \rangle &= -\langle s'(t), x(t) \rangle.
\end{align*}
\]

Hence
\[
\| z'(t) \|_{z(t)}^2 = \langle F''(x(t)) x'(t), x'(t) \rangle + \langle F''(s(t)) s'(t), s'(t) \rangle
\]
\[
= -\left( \langle s(t), x(t) \rangle \right)'_t = - \left( \frac{\nu}{t} \right)'_t = \frac{\nu}{t^2}.
\]

Thus, we can take \( \Delta t = \frac{t}{\sqrt{\nu}} \).

For potential:
\[
\Delta t \cdot (\Phi(x(t), s(t)))'_t = \frac{t}{\sqrt{\nu}} \cdot \left( 2 \nu \ln \frac{\nu}{t} + \nu \ln t \right)'_t
\]
\[
= \frac{t}{\sqrt{\nu}} \cdot (-\nu \ln t)'_t = \frac{t}{\sqrt{\nu}} \cdot \left( -\frac{\nu}{t} \right) = -\sqrt{\nu}.
\]
Proximity measure

\[ \Omega(x, s) = \nu \ln \langle s, x \rangle + F(x) + F_*(s) + \nu - \nu \ln \nu \]
\[ = \nu \ln [\langle c, x \rangle - \langle b, y \rangle] + F(x) + F_*(s) + \nu - \nu \ln \nu. \]

Properties:
- \( \Omega(x, s) \geq 0 \) for all \( (x, s, y) \in \mathcal{F}_P^0 \).
- \( \Omega(x, s) = 0 \) only along the central path.
- The restriction of \( \Omega(x, s) \) onto the hyperplane \( \langle c, x \rangle - \langle b, y \rangle = \text{const} \) is a convex self-concordant function.

Note: \( (x(t), s(t), y(t)) = \arg \min_{x, s, y} \{ F(x) + F_*(s) : Ax = b, s + A^T y = c, \langle c, x \rangle - \langle b, y \rangle = \frac{\nu}{t} \} \).

Proof: \( z(t) \) is feasible and \( F(x(t)) + F_*(s(t)) = -\nu + \nu \ln t \).

On the other hand, for any feasible \( (x, s, y) \) we have:
\[
F(x) + F_*(s) \geq -\nu + \nu \ln \nu - \nu \ln \langle s, x \rangle \\
= -\nu + \nu \ln \nu - \nu \ln \frac{\nu}{t} = -\nu + \nu \ln t.
\]

The minimum is unique since \( \mathcal{F}_P \) contains no straight line.
Primal-dual path-following scheme

- If we close to the central path, we can move along the *tangent* direction up to the moment \( \Omega(x, s) \leq \beta \).
- Then we fix \( \langle c, x \rangle - \langle b, y \rangle \) and go back to the central path by minimizing the barrier.

Efficiency estimate: \( O\left(\sqrt{\nu} \ln \frac{1}{\epsilon}\right) \).

**Advantages:**

- The tangent step typically is large.
- The level \( \beta \) bounds the number of Newton steps for the corrector process by an absolute constant.

**NB.** These schemes are the most efficient now for solving Linear and Quadratic Optimization problems and Linear Matrix Inequalities of moderate size.
Lecture 4. Structural Optimization: Smoothing Technique

Yuri Nesterov, CORE/INMA (UCL)

January 20-22, 2016 (Ecole Polytechnique, Paris)
Outline

1 Nonsmooth Optimization

2 Smoothing technique

3 Application examples
**Problem:** \( \min \left\{ f(x) : x \in \mathbb{R}^n \right\} \Rightarrow x^*, \ f^* = f(x^*) \), where \( f(x) \) is a nonsmooth convex function.

**Subgradients:** \( g \in \partial f(x) \iff f(y) \geq f(x) + \langle g, y - x \rangle \ \forall y \in \mathbb{R}^n. \)

**Main difficulties:**

- \( g \in \partial f(x) \) is *not* a descent direction at \( x \).
- \( g \in \partial f(x^*) \) does not imply \( g = 0. \)

**Example**

\[
\begin{align*}
   f(x) &= \max_{1 \leq j \leq m} \left\{ \langle a_j, x \rangle + b_j \right\}, \\
   \partial f(x) &= \text{Conv} \left\{ a_j : \langle a_j, x \rangle + b_j = f(x) \right\}.
\end{align*}
\]
## Advantages

- Very simple iteration scheme.
- Low memory requirements.
- Optimal rate of convergence (uniformly in the dimension).
- Interpretation of the process.

## Objections:

- Low rate of convergence. (Confirmed by theory!)
- No acceleration.
- High sensitivity to the step-size strategy.
Lower complexity bounds

Nemirovsky, Yudin 1976

If \( f(x) \) is given by a local *black-box*, it is impossible to converge faster than \( O \left( \frac{1}{\sqrt{k}} \right) \) uniformly in \( n \). (\( k \) is the \# of calls of oracle.)

**NB:** Convergence is very slow.

**Question:** We want to find an \( \epsilon \)-solution of the problem

\[
\max_{1 \leq j \leq m} \left\{ \langle a_j, x \rangle + b_j \right\} \rightarrow \min_x : x \in \mathbb{R}^n,
\]

by a gradient scheme (\( n \) and \( m \) are big).

*What is the worst-case complexity bound?*

“Right answer” (Complexity Theory): \( O \left( \frac{1}{\epsilon^2} \right) \) calls of oracle.

**Our target:** A gradient scheme with \( O \left( \frac{1}{\epsilon} \right) \) complexity bound.

**Reason of speed up:** our problem *is not* in a black box.
Complexity of Smooth Minimization

**Problem:** \( f(x) \rightarrow \min_{x} x \in \mathbb{R}^n \), where \( f \) is a convex function and \( \| \nabla f(x) - \nabla f(y) \|_* \leq L(f) \| x - y \| \) for all \( x, y \in \mathbb{R}^n \).

(For measuring gradients we use dual norms: \( \| s \|_* = \max_{\| x \| = 1} \langle s, x \rangle \).)

**Rate of convergence:** Optimal method gives \( O \left( \frac{L(f)}{k^2} \right) \).

**Complexity:** \( O \left( \sqrt{\frac{L(f)}{\epsilon}} \right) \). The difference with \( O \left( \frac{1}{\epsilon^2} \right) \) is very big.
For function $f$ define its Fenchel conjugate:

$$f_*(s) = \max_{x \in \mathbb{R}^n} [\langle s, x \rangle - f(x)].$$

It is a closed convex function with $\text{dom } f_* = \text{Conv}\{f'(x) : x \in \mathbb{R}^n\}$.

Moreover, under very mild conditions $(f_*(s))_* \equiv f(x)$.

Define $f_\mu(x) = \max_{s \in \text{dom } f_*} [\langle s, x \rangle - f_*(s) - \frac{\mu}{2}\|s\|_*^2]$, where $\| \cdot \|_*$ is a Euclidean norm.

Note: $f'_\mu(x) = s_\mu(x)$, and $x = f'_*(s_\mu(x)) + \mu s_\mu(x)$. Therefore,

$$\|x^1 - x^2\|^2 = \|f'_*(s^1) - f'_*(s^2)\|^2 + 2\mu \langle f'_*(s^1) - f'_*(s^2), s^1 - s^2 \rangle + \mu^2\|s^1 - s^2\|^2 \geq \mu^2\|s^1 - s^2\|^2.$$ 

Thus, $f_\mu \in C^{1,1}_{1/\mu}$ and $f(x) \geq f_\mu(x) \geq f(x) - \mu D^2$, where $D = \text{Diam}(\text{dom } f_*)$. 
Main questions

1. Given by a non-smooth convex $f(x)$, can we form its computable smooth $\varepsilon$-approximation $f_\varepsilon(x)$ with

$$L(f_\varepsilon) = O\left(\frac{1}{\varepsilon}\right)?$$

If yes, we need only $O\left(\sqrt{\frac{L(f_\varepsilon)}{\varepsilon}}\right) = O\left(\frac{1}{\varepsilon}\right)$ iterations.

2. Can we do this in a systematic way?

Conclusion: We need a convenient model of our problem.
Primal problem: Find \( f^* = \min_x f(x) : x \in Q_1 \), where \( Q_1 \subset E_1 \) is convex closed and bounded.

Objective: \( f(x) = \hat{f}(x) + \max_u \{ \langle Ax, u \rangle \nabla - \hat{\phi}(u) : u \in Q_2 \} \), where

- \( \hat{f}(x) \) is differentiable and convex on \( Q_1 \).
- \( Q_2 \subset E_2 \) is a closed convex and bounded.
- \( \hat{\phi}(u) \) is continuous convex function on \( Q_2 \).
- linear operator \( A : E_1 \to E_2^* \).

Adjoint problem: \( \max_u \{ \phi(u) : u \in Q_2 \} \), where

\[ \phi(u) = -\hat{\phi}(u) + \min_x \{ \langle Ax, u \rangle \nabla + \hat{f}(x) : x \in Q_1 \} \]

NB: Adjoint problem is not unique!
Consider \( f(x) = \max_{1 \leq j \leq m} |\langle a_j, x \rangle_1 - b_j| \).

1. \( Q_2 = E_1^*, A = I, \hat{\phi}(u) \equiv f_*(u) = \max_x \{ \langle u, x \rangle_1 - f(x) : x \in E_1 \} \)

\[
= \min_{s \in \mathbb{R}^m} \left\{ \sum_{j=1}^m s_j b_j : u = \sum_{j=1}^m s_j a_j, \sum_{j=1}^m |s_j| \leq 1 \right\}.
\]

2. \( E_2 = \mathbb{R}^m, \hat{\phi}(u) = \langle b, u \rangle_2, f(x) = \max_{1 \leq j \leq m} |\langle a_j, x \rangle_1 - b_j| \)

\[
= \max_{u \in \mathbb{R}^m} \left\{ \sum_{j=1}^m u_j [\langle a_j, x \rangle_1 - b_j] : \sum_{j=1}^m |u_j| \leq 1 \right\}.
\]

3. \( E_2 = \mathbb{R}^{2m}, \hat{\phi}(u) \) is a linear, \( Q_2 \) is a simplex:

\[
f(x) = \max_{u \in \mathbb{R}^{2m}} \left\{ \sum_{j=1}^m (u_j^1 - u_j^2) [\langle a_j, x \rangle_1 - b_j] : \sum_{j=1}^m (u_j^1 + u_j^2) = 1, u \geq 0 \right\}.
\]

**NB:** Increase in dim \( E_2 \) decreases the complexity of representation.
Smooth approximations

**Prox-function:** $d_2(u)$ is continuous and strongly convex on $Q_2$: 
$$d_2(v) \geq d_2(u) + \langle \nabla d_2(u), v - u \rangle_2 + \frac{1}{2} \sigma_2 \|v - u\|_2^2.$$  
Assume: $d_2(u_0) = 0$ and $d_2(u) \geq 0 \ \forall u \in Q_2$.

Fix $\mu > 0$, the smoothing parameter, and define 
$$f_\mu(x) = \max_u \{ \langle Ax, u \rangle_2 - \hat{\phi}(u) - \mu d_2(u) : u \in Q_2 \}.$$  
Denote by $u(x)$ the solution of this problem.

**Theorem:** $f_\mu(x)$ is convex and differentiable for $x \in E_1$. Its gradient $\nabla f_\mu(x) = A^* u(x)$ is Lipschitz continuous with 
$$L(f_\mu) = \frac{1}{\mu \sigma_2} \|A\|_{1,2}^2,$$
where $\|A\|_{1,2} = \max_{x,u} \{ \langle Ax, u \rangle_2 : \|x\|_1 = 1, \|u\|_2 = 1 \}$.

**NB:** 1. For any $x \in E_1$ we have $f_0(x) \geq f_\mu(x) \geq f_0(x) - \mu D_2$, where $D_2 = \max_u \{ d_2(u) : u \in Q_2 \}$.

2. All norms are very important.
Problem: \( \min_{x} \{ f(x) : x \in Q_1 \} \) with \( f \in C^{1,1}(Q_1) \).

Prox-function: strongly convex \( d_1(x) \), \( d_1(x^0) = 0 \), \( d_1(x) \geq 0 \), \( x \in Q_1 \).

Gradient mapping:
\[
T_L(x) = \arg \min_{y \in Q_1} \{ \langle \nabla f(x), y - x \rangle_1 + \frac{1}{2} L \| y - x \|_1^2 \}. 
\]

Method. For \( k \geq 0 \) do:
1. Compute \( f(x^k), \nabla f(x^k) \).
2. Find \( y^k = T_L(f)(x^k) \).
3. Find \( z^k = \arg \min_{x \in Q_1} \{ \frac{L(f)}{\sigma} d_1(x) + \sum_{i=0}^{k} \frac{i+1}{2} \langle \nabla f(x^i), x \rangle_1 \} \).
4. Set \( x^{k+1} = \frac{2}{k+3} z^k + \frac{k+1}{k+3} y^k \).

Convergence: \( f(y^k) - f(x^*) \leq \frac{4L(f)d_1(x^*)}{\sigma_1(k+1)^2} \), where \( x^* \) is the optimal solution.
Applications

Smooth problem: \( \bar{f}_\mu(x) = \hat{f}(x) + f_\mu(x) \rightarrow \min : x \in Q_1. \)

Lipschitz constant: \( L_\mu = L(\hat{f}) + \frac{1}{\mu \sigma_2} \|A\|_{1,2}^2. \) Denote \( D_1 = \max_x \{d_1(x) : x \in Q_1\}. \)

**Theorem:** Let us choose \( \mu \geq 1. \) Define
\[
\mu = \mu(N) = \frac{2\|A\|_{1,2}}{N+1} \cdot \sqrt{\frac{D_1}{\sigma_1 \sigma_2 D_2}}.
\]
After \( N \) iterations set \( \hat{x} = y^N \in Q_1 \) and
\[
\hat{u} = \sum_{i=0}^{N} \frac{2(i+1)}{(N+1)(N+2)} u(x^i) \in Q_2.
\]
Then \( 0 \leq f(\hat{x}) - \phi(\hat{u}) \leq \frac{4\|A\|_{1,2}}{N+1} \cdot \sqrt{\frac{D_1 D_2}{\sigma_1 \sigma_2}} + \frac{4L(\hat{f})D_1}{\sigma_1 (N+1)^2}. \)

**Corollary.** Let \( L(\hat{f}) = 0. \) For getting an \( \epsilon \)-solution, we choose
\[
\mu = \frac{\epsilon}{2D_2}, \quad L = \frac{D_2}{2\sigma_2} \cdot \frac{\|A\|_{1,2}^2}{\epsilon}, \quad N \geq 4\|A\|_{1,2} \sqrt{\frac{D_1 D_2}{\sigma_1 \sigma_2}} \cdot \frac{1}{\epsilon}.
\]
Example: Equilibrium in matrix games (1)

Denote $\Delta_n = \{ x \in \mathbb{R}^n : x \geq 0, \sum_{i=1}^{n} x^{(i)} = 1 \}$. Consider the problem

$$\min_{x \in \Delta_n} \max_{u \in \Delta_m} \{ \langle Ax, u \rangle_2 + \langle c, x \rangle_1 + \langle b, u \rangle_2 \}.$$

**Minimization form:**

$$\min_{x \in \Delta_n} f(x), \quad f(x) = \langle c, x \rangle_1 + \max_{1 \leq j \leq m} [\langle a_j, x \rangle_1 + b_j],$$

$$\max_{u \in \Delta_m} \phi(u), \quad \phi(u) = \langle b, u \rangle_2 + \min_{1 \leq i \leq n} [\langle \hat{a}_i, u \rangle_2 + c_i],$$

where $a_j$ are the rows and $\hat{a}_i$ are the columns of $A$.

1. **Euclidean distance:** Let us take

$$\|x\|_1^2 = \sum_{i=1}^{n} x_i^2, \quad \|u\|_2^2 = \sum_{j=1}^{m} u_j^2,$$

$$d_1(x) = \frac{1}{2} \|x - \frac{1}{n} e_n\|_1^2, \quad d_2(u) = \frac{1}{2} \|u - \frac{1}{m} e_m\|_2^2.$$

Then $\|A\|_{1,2} = \lambda_{\max}^{1/2}(A^T A)$ and $f(\hat{x}) - \phi(\hat{u}) \leq \frac{4\lambda_{\max}^{1/2}(A^T A)}{N+1}$. 
Example: Equilibrium in matrix games (2)

2. Entropy distance. Let us choose

\[ \|x\|_1 = \sum_{i=1}^{n} |x_i|, \quad d_1(x) = \ln n + \sum_{i=1}^{n} x_i \ln x_i, \]
\[ \|u\|_2 = \sum_{j=1}^{m} |u_j|, \quad d_2(u) = \ln m + \sum_{j=1}^{m} u_j \ln u_j. \]

**LM:** \( \sigma_1 = \sigma_2 = 1. \) (Hint: \( \langle d''_1(x)h, h \rangle = \sum_{i=1}^{n} \frac{h_i^2}{x_i} \rightarrow \min_{x \in \Delta_n} = \|h\|_1^2. \))

Moreover, since \( D_1 = \ln n, \ D_2 = \ln m, \) and

\[ \|A\|_{1,2} = \max_{x} \max_{1 \leq j \leq m} |\langle a_j, x \rangle| : \|x\|_1 = 1 \} = \max_{i,j} |A_{i,j}|, \]

we have \( f(\hat{x}) - \phi(\hat{u}) \leq \frac{4\sqrt{\ln n \ln m}}{N+1} \cdot \max_{i,j} |A_{i,j}|. \)

**NB:** 1. Usually \( \max_{i,j} |A_{i,j}| \ll \lambda_{\max}^{1/2}(A^TA). \)

2. We have \( \tilde{f}_\mu(x) = \langle c, x \rangle_1 + \mu \ln \left( \frac{1}{m} \sum_{j=1}^{m} e^{[\langle a_j, x \rangle + b_j]/\mu} \right). \)
Example 2: Continuous location problem

Problem: \( p \) cities with populations \( m_j, j = 1, \ldots, m \), are located at
\[
c_j \in \mathbb{R}^n, \quad j = 1, \ldots, p.
\]

Goal: Construct a service center at point \( x^* \), which minimizes the total distance to the center.

That is Find \( f^* = \min_x \left\{ f(x) = \sum_{j=1}^{p} m_j \|x - c_j\|_1 : \|x\|_1 \leq \bar{r} \right\} \).

Primal space:
\[
\|x\|_1^2 = \sum_{i=1}^{n} (x^{(i)})^2, \quad d_1(x) = \frac{1}{2}\|x\|_1^2, \quad \sigma_1 = 1, \quad D_1 = \frac{1}{2}\bar{r}^2.
\]
Adjoint space: $E_2 = (E_1^*)^p$, $\|u\|_2^2 = \sum_{j=1}^p m_j (\|u_j\|_1^*)^2$,

$$Q_2 = \{ u = (u_1, \ldots, u_p) \in E_2 : \|u_j\|_1^* \leq 1, j = 1, \ldots, p \},$$

$$d_2(u) = \frac{1}{2} \|u\|_2^2, \quad \sigma_2 = 1, \quad D_2 = \frac{1}{2} P,$$

with $P \equiv \sum_{j=1}^p m_j$, the total size of population.

Operator norm: $\|A\|_{1,2} = P^{1/2}$.

Rate of convergence: $f(\hat{x}) - f^* \leq \frac{2P\bar{r}}{N+1}$.

$$f_\mu(x) = \sum_{j=1}^p m_j \psi_\mu(\|x - c_j\|_1), \quad \psi_\mu(\tau) = \begin{cases} \frac{\tau^2}{2\mu}, & \tau \leq \mu, \\ \tau - \frac{\mu}{2}, & \mu \leq \tau. \end{cases}$$
Example 3: Variational inequalities (linear operator)

Consider $B(w) = Bw + c: E \to E^*$, which is monotone:

$$\langle Bh, h \rangle \geq 0 \quad \forall h \in E.$$  

**Problem:** Find $w^* \in Q : \langle B(w^*), w - w^* \rangle \geq 0 \quad \forall w \in Q,$ where $Q$ is a bounded convex closed set.

**Merit function:** $\psi(w) = \max_v \{ \langle B(v), w - v \rangle : v \in Q \}.$

- $\psi(w)$ is convex on $E_1$.
- $\psi(w) \geq 0$ for all $w \in Q$.
- $\psi(w) = 0$ if and only if $w$ solves VI-problem.
- $\langle B(v), v \rangle$ is a convex function. Thus, $\psi$ is exactly in our form.

**Primal smoothing:**

$$\psi_\mu(w) = \max_v \{ \langle B(v), w - v \rangle - \mu d_2(v) : v \in Q \}.$$  

**Dual smoothing:**

$$\phi_\mu(v) = \min_w \{ \langle B(v), w - v \rangle + \mu d_1(w) : w \in Q \}. \quad \text{(Looks better.)}$$
Example 4: Piece-wise linear functions

1. Maximum of absolute values. Consider

\[
\min_x \left\{ f(x) = \max_{1 \leq j \leq m} |\langle a_j, x \rangle_1 - b^{(j)}| : x \in Q_1 \right\}.
\]

For simplicity choose \[\|x\|^2_1 = \sum_{i=1}^{n} (x^{(i)})^2, \quad d_1(x) = \frac{1}{2} \|x\|^2.\]

It is convenient to choose \[E_2 = \mathbb{R}^{2m},\]

\[\|u\|^2_2 = \sum_{j=1}^{2m} |u^{(j)}|, \quad d_2(u) = \ln(2m) + \sum_{j=1}^{2m} u^{(j)} \ln u^{(j)}.\]

Denote by \(A\) the matrix with the rows \(a_j\). Then

\[f(x) = \max_u \{ \langle \hat{A}x, u \rangle_2 - \langle \hat{b}, u \rangle_2 : u \in \Delta_{2m} \},\]

where \(\hat{A} = \begin{pmatrix} A \\ -A \end{pmatrix}\) and \(\hat{b} = \begin{pmatrix} b \\ -b \end{pmatrix}\).
Thus, $\sigma_1 = \sigma_2 = 1$,

$$D_2 = \ln(2m), \quad D_1 = \frac{1}{2} \bar{r}^2, \quad \bar{r} = \max_x \{\|x\|_1 : x \in Q_1\}.$$ 

**Operator norm:** $\|\hat{A}\|_{1,2} = \max_{1 \leq j \leq m} \|a_j\|_1^\ast$. 

**Complexity:** $2\sqrt{2} \bar{r} \max_{1 \leq j \leq m} \|a_j\|_1^\ast \sqrt{\ln(2m)} \cdot \frac{1}{\epsilon}$. 

**Approximation:** for $\xi(\tau) = \frac{1}{2}[e^{\tau} + e^{-\tau}]$ define

$$\bar{f}_\mu(x) = \mu \ln \left( \frac{1}{m} \sum_{j=1}^{m} \xi \left( \frac{1}{\mu} [\langle a_j, x \rangle + b(j)] \right) \right).$$
Piece-wise linear functions: Sum of absolute values.

$$\min_x \left\{ f(x) = \sum_{j=1}^{m} |\langle a_j, x \rangle_1 - b(j)| : x \in Q_1 \right\}. $$

Let us choose $E_2 = \mathbb{R}^m$, $Q_2 = \{ u \in \mathbb{R}^m : |u(j)| \leq 1, j = 1, \ldots, m \}$, and $d_2(u) = \frac{1}{2}\|u\|_2^2 = \frac{1}{2} \sum_{j=1}^{m} \|a_j\|_1^* \cdot (u(j))^2$.

Then $f_\mu(x) = \sum_{j=1}^{m} \|a_j\|_1^* \cdot \psi_\mu \left( \frac{|\langle a_j, x \rangle_1 - b(j)|}{\|a_j\|_1^*} \right)$,

$$\|A\|_{1,2}^2 = P \equiv \sum_{j=1}^{m} \|a_j\|_1^*. $$

On the other hand, $D_2 = \frac{1}{2} P$ and $\sigma_2 = 1$. Thus, we get the following complexity bound: 

$$\frac{1}{\epsilon} \cdot \sqrt{\frac{8D_1}{\sigma_1}} \cdot \sum_{j=1}^{m} \|a_j\|_1^*. $$

**NB:** The bound and the scheme allow $m \to \infty$. 
Test problem: \[ \min_{x \in \Delta_n} \max_{u \in \Delta_m} \langle Ax, u \rangle_2. \]

Entries of \( A \) are uniformly distributed in \([-1, 1] \).

**Goal:** Test of computational stability. **Computer:** 2.6GHz.

**Complexity of iteration:** \( 2mn \) operations.

<table>
<thead>
<tr>
<th>( m \setminus n )</th>
<th>100</th>
<th>300</th>
<th>1000</th>
<th>3000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>808</td>
<td>1011</td>
<td>1112</td>
<td>1314</td>
<td>1415</td>
</tr>
<tr>
<td></td>
<td>0”</td>
<td>0”</td>
<td>3”</td>
<td>12”</td>
<td>44”</td>
</tr>
<tr>
<td>300</td>
<td>910</td>
<td>1112</td>
<td>1415</td>
<td>1617</td>
<td>1819</td>
</tr>
<tr>
<td></td>
<td>0”</td>
<td>2”</td>
<td>10”</td>
<td>35”</td>
<td>135”</td>
</tr>
<tr>
<td>1000</td>
<td>1112</td>
<td>1213</td>
<td>1415</td>
<td>1718</td>
<td>2020</td>
</tr>
<tr>
<td></td>
<td>2”</td>
<td>8”</td>
<td>32”</td>
<td>115”</td>
<td>451”</td>
</tr>
</tbody>
</table>

**Results for** \( \epsilon = 0.01 \). **Table 1**

**Number of iterations:** 40 – 50\% of predicted values.
**Results for** $\epsilon = 0.001$.  

### Table 2

<table>
<thead>
<tr>
<th>$m$</th>
<th>$n$</th>
<th>100</th>
<th>300</th>
<th>1000</th>
<th>3000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>6970</td>
<td>8586</td>
<td>9394</td>
<td>10000</td>
<td>10908</td>
<td></td>
</tr>
<tr>
<td>2’’</td>
<td>8’’</td>
<td>29’’</td>
<td>91’’</td>
<td>349’’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>7778</td>
<td>10101</td>
<td>12424</td>
<td>14242</td>
<td>15656</td>
<td></td>
</tr>
<tr>
<td>8’’</td>
<td>27’’</td>
<td>97’’</td>
<td>313’’</td>
<td>1162’’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>8788</td>
<td>11010</td>
<td>13030</td>
<td>15757</td>
<td>18282</td>
<td></td>
</tr>
<tr>
<td>30’’</td>
<td>105’’</td>
<td>339’’</td>
<td>1083’’</td>
<td>4085’’</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Results for** $\epsilon = 0.0001$.  

### Table 3

<table>
<thead>
<tr>
<th>$m$</th>
<th>$n$</th>
<th>100</th>
<th>300</th>
<th>1000</th>
<th>3000</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>67068</td>
<td>72073</td>
<td>74075</td>
<td>80081</td>
<td></td>
</tr>
<tr>
<td>25’’</td>
<td>80’’</td>
<td>287’’</td>
<td>945’’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>85086</td>
<td>92093</td>
<td>101102</td>
<td>112113</td>
<td></td>
</tr>
<tr>
<td>89’’, 42'</td>
<td>243’’</td>
<td>914’’</td>
<td>3302’’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>97098</td>
<td>100101</td>
<td>116117</td>
<td>139140</td>
<td></td>
</tr>
<tr>
<td>331’’</td>
<td>760’’</td>
<td>2936’’</td>
<td>11028’’</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Comparing the bounds

**Smoothing + FGM:** \[2 \cdot 4 \cdot \frac{mn}{\epsilon} \sqrt{\ln n \ln m}.

**Short-step p.-f. method** (\(n \geq m\)): \((7.2 \sqrt{n \ln \frac{1}{\epsilon}}) \cdot \frac{m(m+1)}{2} n\).

Right digits

<table>
<thead>
<tr>
<th>(m)</th>
<th>(n)</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>100</td>
<td>(g)</td>
<td>(g)</td>
<td>(b)</td>
<td>(b)</td>
</tr>
<tr>
<td>300</td>
<td>300</td>
<td>(g)</td>
<td>(g)</td>
<td>(b)</td>
<td>(b)</td>
</tr>
<tr>
<td>300</td>
<td>1000</td>
<td>(g)</td>
<td>(g)</td>
<td>(b)</td>
<td>(b)</td>
</tr>
<tr>
<td>300</td>
<td>3000</td>
<td>(g)</td>
<td>(g)</td>
<td>(=)</td>
<td>(b)</td>
</tr>
<tr>
<td>300</td>
<td>10000</td>
<td>(g)</td>
<td>(g)</td>
<td>(g)</td>
<td>(b)</td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td>(g)</td>
<td>(g)</td>
<td>(g)</td>
<td>(b)</td>
</tr>
<tr>
<td>1000</td>
<td>3000</td>
<td>(g)</td>
<td>(g)</td>
<td>(g)</td>
<td>(b)</td>
</tr>
<tr>
<td>1000</td>
<td>10000</td>
<td>(g)</td>
<td>(g)</td>
<td>(=)</td>
<td>(=)</td>
</tr>
</tbody>
</table>

\(g\) - S+FGM, \(b\) - barrier method
Outline

1. Problems sizes
2. Random coordinate search
3. Confidence level of solutions
4. Sparse Optimization problems
5. Sparse updates for linear operators
6. Fast updates in computational trees
7. Simple subgradient methods
8. Application examples
Nonlinear Optimization: problems sizes

<table>
<thead>
<tr>
<th>Class</th>
<th>Operations</th>
<th>Dimension</th>
<th>Iter. Cost</th>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small-size</td>
<td>All</td>
<td>$10^0 - 10^2$</td>
<td>$n^4 \rightarrow n^3$</td>
<td>Kilobyte: $10^3$</td>
</tr>
<tr>
<td>Medium-size</td>
<td>$A^{-1}$</td>
<td>$10^3 - 10^4$</td>
<td>$n^3 \rightarrow n^2$</td>
<td>Megabyte: $10^6$</td>
</tr>
<tr>
<td>Large-scale</td>
<td>$Ax$</td>
<td>$10^5 - 10^7$</td>
<td>$n^2 \rightarrow n$</td>
<td>Gigabyte: $10^9$</td>
</tr>
<tr>
<td>Huge-scale</td>
<td>$x + y$</td>
<td>$10^8 - 10^{12}$</td>
<td>$n \rightarrow \log n$</td>
<td>Terabyte: $10^{12}$</td>
</tr>
</tbody>
</table>

Sources of Huge-Scale problems

- Internet (New)
- Telecommunications (New)
- Finite-element schemes (Old)
- Partial differential equations (Old)
Very old optimization idea: Coordinate Search

**Problem:** \[ \min_{x \in \mathbb{R}^n} f(x) \quad (f \text{ is convex and differentiable}). \]

**Coordinate relaxation algorithm**

For \( k \geq 0 \) iterate

1. Choose active coordinate \( i_k \).
2. Update \( x_{k+1} = x_k - h_k \nabla_{i_k} f(x_k) e_i \) ensuring \( f(x_{k+1}) \leq f(x_k) \).
   
   \((e_i \text{ is } i\text{th coordinate vector in } \mathbb{R}^n.)\)

**Main advantage:** Very simple implementation.
Possible strategies

1. Cyclic moves. (Difficult to analyze.)
2. Random choice of coordinate (Why?)
3. Choose coordinate with the maximal directional derivative.

**Complexity estimate:** assume
\[ \| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \|, \quad x, y \in \mathbb{R}^n. \]
Let us choose \( h_k = \frac{1}{L} \). Then

\[
f(x_k) - f(x_{k+1}) \geq \frac{1}{2L} |\nabla_{i_k} f(x_k)|^2 \geq \frac{1}{2nL} \| \nabla f(x_k) \|^2 \\
\geq \frac{1}{2nLR^2} (f(x_k) - f^*)^2.
\]

Hence, \( f(x_k) - f^* \leq \frac{2nLR^2}{k}, \quad k \geq 1. \) (For Grad.Method, drop \( n \).

This is the only known theoretical result known for CDM!
Theoretical justification:

- Complexity bounds are not known for the most of the schemes.
- The only justified scheme needs computation of the whole gradient. (Why don’t use GM?)

Computational complexity:

- Fast differentiation: if function is defined by a sequence of operations, then \( C(\nabla f) \leq 4C(f) \).
- Can we do anything without computing the function’s values?

Result: CDM are almost out of the computational practice.
Let $E \in \mathbb{R}^{n \times n}$ be an incidence matrix of a graph. Denote $e = (1, \ldots, 1)^T$ and

$$
\bar{E} = E \cdot \text{diag} \left( (E^T e)^{-1} \right).
$$

Thus, $\bar{E}^T e = e$. Our problem is as follows:

Find $x^* \geq 0 : \bar{E} x^* = x^*$.

**Optimization formulation:**

$$
f(x) \overset{\text{def}}{=} \frac{1}{2} \|\bar{E} x - x\|^2 + \frac{\gamma}{2} [\langle e, x \rangle - 1]^2 \rightarrow \min_{x \in \mathbb{R}^n}
$$
### Huge-scale problems

#### Main features
- The size is very big \((n \geq 10^7)\).
- The data is distributed in space.
- The requested parts of data are not always available.
- The data may be changing in time.

#### Consequences
Simplest operations are expensive or infeasible:
- Update of the full vector of variables.
- Matrix-vector multiplication.
- Computation of the objective function’s value, etc.
Let us look at the gradient of the objective:

$$\nabla_i f(x) = \langle a_i, g(x) \rangle + \gamma [\langle e, x \rangle - 1], \ i = 1, \ldots, n,$$

$$g(x) = \bar{E}x - x \in R^n, \ (\bar{E} = (a_1, \ldots, a_n)).$$

**Main observations:**

- The coordinate move $$x_+ = x - h_i \nabla_i f(x)e_i$$ needs $$O(p_i)$$ a.o. ($$p_i$$ is the number of nonzero elements in $$a_i$$.)
- $$d_i \overset{\text{def}}{=} \text{diag} \left( \nabla^2 f \overset{\text{def}}{=} \bar{E}^T \bar{E} + \gamma ee^T \right)_i = \gamma + \frac{1}{p_i}$$ are available.

We can use them for choosing the step sizes ($$h_i = \frac{1}{d_i}$$).

**Reasonable coordinate choice strategy?** Random!
Random coordinate descent methods (RCDM)

\[
\min_{x \in \mathbb{R}^N} f(x), \quad (f \text{ is convex and differentiable})
\]

**Main Assumption:**

\[|f'_i(x + h_i e_i) - f'_i(x)| \leq L_i |h_i|, \quad h_i \in \mathbb{R}, \quad i = 1, \ldots, N,\]

where \(e_i\) is a coordinate vector. Then

\[f(x + h_i e_i) \leq f(x) + f'_i(x) h_i + \frac{L_i}{2} h_i^2. \quad x \in \mathbb{R}^N, \quad h_i \in \mathbb{R}.
\]

Define the coordinate steps: \(T_i(x) \overset{\text{def}}{=} x - \frac{1}{L_i} f'_i(x) e_i\). Then,

\[f(x) - f(T_i(x)) \geq \frac{1}{2L_i} [f'_i(x)]^2, \quad i = 1, \ldots, N.
\]
Random choice for coordinates

We need a special random counter $\mathcal{R}_\alpha$, $\alpha \in R$:

$$\text{Prob}[i] = p^{(i)}_{\alpha} = L_i^\alpha \cdot \left[ \sum_{j=1}^{N} L_j^\alpha \right]^{-1}, \quad i = 1, \ldots, N.$$ 

**Note:** $\mathcal{R}_0$ generates uniform distribution.

Method $RCDM(\alpha, x_0)$

For $k \geq 0$ iterate:

1) Choose $i_k = \mathcal{R}_\alpha$.

2) Update $x_{k+1} = T_{i_k}(x_k)$. 
Complexity bounds for RCDM

We need to introduce the following norms for $x, g \in R^N$:

$$\|x\|_\alpha = \left[ \sum_{i=1}^{N} L_i^\alpha [x^{(i)}]^2 \right]^{1/2}, \quad \|g\|_\alpha^* = \left[ \sum_{i=1}^{N} \frac{1}{L_i^\alpha} [g^{(i)}]^2 \right]^{1/2}.$$ 

After $k$ iterations, $RCDM(\alpha, x_0)$ generates random output $x_k$, which depends on $\xi_k = \{i_0, \ldots, i_k\}$. Denote $\phi_k = E_{\xi_{k-1}} f(x_k)$.

**Theorem.** For any $k \geq 1$ we have

$$\phi_k - f^* \leq \frac{2}{k} \cdot \left[ \sum_{j=1}^{N} L_j^\alpha \right] \cdot R^2_{1-\alpha}(x_0),$$

where $R_\beta(x_0) = \max \left\{ \max_{x \in X^*} \|x - x^*\|_\beta : f(x) \leq f(x_0) \right\}$. 

Interpretation

Denote \( S_\alpha = \sum_{i=1}^{N} L_i^\alpha \).

1. \( \alpha = 0 \). Then \( S_0 = N \), and we get

\[
\phi_k - f^* \leq \frac{2N}{k} \cdot R_1^2(x_0).
\]

Note

- We use the metric \( \|x\|_1^2 = \sum_{i=1}^{N} L_i[x^{(i)}]^2 \).
- Matrix with diagonal \( \{L_i\}_{i=1}^{N} \) can have its norm equal to \( n \).
- Hence, for GM we can guarantee the same bound.
  But its cost of iteration is much higher!
2. $\alpha = \frac{1}{2}$. Denote

$$D_\infty(x_0) = \max_x \left\{ \max_{y \in X^*} \max_{1 \leq i \leq N} |x(i) - y(i)| : f(x) \leq f(x_0) \right\}.$$ 

Then, $R^2_{1/2}(x_0) \leq S_{1/2}D_\infty^2(x_0)$, and we obtain

$$\phi_k - f^* \leq \frac{2}{k} \cdot \left[ \sum_{i=1}^{N} L_i^{1/2} \right]^2 \cdot D_\infty^2(x_0).$$

Note:

- For the first order methods, the worst-case complexity of minimizing over a box depends on $N$.
- Since $S_{1/2}$ can be bounded, RCDM can be applied in situations when the usual GM fail.
3. \( \alpha = 1 \). Then \( R_0(x_0) \) is the size of the initial level set in the standard Euclidean norm. Hence,

\[
\phi_k - f^* \leq \frac{2}{k} \cdot \left[ \sum_{i=1}^{N} L_i \right] \cdot R_0^2(x_0) \equiv \frac{2N}{k} \cdot \left[ \frac{1}{N} \sum_{i=1}^{N} L_i \right] \cdot R_0^2(x_0).
\]

Rate of convergence of GM can be estimated as

\[
f(x_k) - f^* \leq \frac{\gamma}{k} R_0^2(x_0),
\]

where \( \gamma \) satisfies condition \( f''(x) \leq \gamma \cdot I, \ x \in \mathbb{R}^N \).

**Note:** maximal eigenvalue of symmetric matrix can reach its trace.

In the worst case, the rate of convergence of GM is the same as that of \( RCDM \).
Minimizing the strongly convex functions

**Theorem.** Let \( f(x) \) be strongly convex with respect to \( \| \cdot \|_{1-\alpha} \) with convexity parameter \( \sigma_{1-\alpha} > 0 \). Then, for \( \{x_k\} \) generated by \( \text{RCDM}(\alpha, x_0) \) we have

\[
\phi_k - \phi^* \leq \left( 1 - \frac{\sigma_{1-\alpha}}{S_{\alpha}} \right)^k (f(x_0) - f^*).
\]

**Proof:** Let \( x_k \) be generated by \( \text{RCDM} \) after \( k \) iterations. Let us estimate the expected result of the next iteration.

\[
f(x_k) - E_{i_k}(f(x_{k+1})) = \sum_{i=1}^{N} p^{(i)}_{\alpha} \cdot [f(x_k) - f(T_i(x_k))]
\]

\[
\geq \sum_{i=1}^{N} \frac{p^{(i)}_{\alpha}}{2L_i} \| f'(x_k) \|^2 = \frac{1}{2 S_{\alpha}} \left( \| f'(x_k) \|_{1-\alpha}^* \right)^2
\]

\[
\geq \frac{\sigma_{1-\alpha}}{S_{\alpha}} (f(x_k) - f^*).
\]

It remains to compute expectation in \( \xi_{k-1} \).
Confidence level of the answers

**Note:** We have proved that the expected values of random $f(x_k)$ are good.

*Can we guarantee anything after a single run?*

**Confidence level:** Probability $\beta \in (0, 1)$, that some statement about random output is correct.

**Main tool:** Chebyshev inequality ($\xi \geq 0$):

$$\text{Prob} \left[ \xi \geq T \right] \leq \frac{E(\xi)}{T}.$$

**Our situation:**

$$\text{Prob} \left[ f(x_k) - f^* \geq \epsilon \right] \leq \frac{1}{\epsilon} [\phi_k - f^*] \leq 1 - \beta.$$

We need $\phi_k - f^* \leq \epsilon \cdot (1 - \beta)$. Too expensive for $\beta \to 1$?
Regularization technique

Consider \( f_\mu(x) = f(x) + \frac{\mu}{2} \| x - x_0 \|_{1-\alpha}^2 \). It is strongly convex. Therefore, we can obtain \( \phi_k - f_\mu^* \leq \epsilon \cdot (1 - \beta) \) in

\[
O \left( \frac{1}{\mu} S_\alpha \ln \frac{1}{\epsilon(1-\beta)} \right) \text{ iterations.}
\]

**Theorem.** Define \( \alpha = 1, \mu = \frac{\epsilon}{4R_0^2(x_0)} \), and choose

\[
k \geq 1 + \frac{8S_1R_0^2(x_0)}{\epsilon} \left[ \ln \frac{2S_1R_0^2(x_0)}{\epsilon} + \ln \frac{1}{1-\beta} \right].
\]

Let \( x_k \) be generated by \( RCDM(1, x_0) \) as applied to \( f_\mu \). Then

\[
\text{Prob} \left( f(x_k) - f^* \leq \epsilon \right) \geq \beta.
\]

**Note:** \( \beta = 1 - 10^{-p} \Rightarrow \ln 10^p = 2.3p \).
Implementation details: Random Counter

Given the values \( L_i, i = 1, \ldots, N \), generate efficiently random \( i \in \{1, \ldots, N\} \) with probabilities \( \text{Prob}[i = k] = \frac{L_k}{\sum_{j=1}^{N} L_j} \).

Solution: a) Trivial \( \Rightarrow \) \( O(N) \) operations.
b). Assume \( N = 2^p \). Define \( p + 1 \) vectors \( S_k \in \mathbb{R}^{2^{p-k}} \), \( k = 0, \ldots, p \):

\[
S_0^{(i)} = L_i, \ i = 1, \ldots, N.
\]
\[
S_k^{(i)} = S_{k-1}^{(2i)} + S_{k-1}^{(2i-1)}, \ i = 1, \ldots, 2^{p-k}, \ k = 1, \ldots, p.
\]

Algorithm: Make the choice in \( p \) steps, from top to bottom.

- If the element \( i \) of \( S_k \) is chosen, then choose in \( S_{k-1} \) either \( 2i \) or \( 2i - 1 \) in accordance to probabilities \( \frac{S_{k-1}^{(2i)}}{S_k^{(i)}} \) or \( \frac{S_{k-1}^{(2i-1)}}{S_k^{(i)}} \).

Difference: for \( n = 2^{20} > 10^6 \) we have \( p = \log_2 N = 20 \).
Problem: \( \min_{x \in Q} f(x) \), where \( Q \) is closed and convex in \( \mathbb{R}^N \), and \( f(x) = \Psi(Ax) \), where \( \Psi \) is a simple convex function:

\[
\Psi(y_1) \geq \Psi(y_2) + \langle \Psi'(y_2), y_1 - y_2 \rangle, \quad y_1, y_2 \in \mathbb{R}^M,
\]

- \( A : \mathbb{R}^N \to \mathbb{R}^M \) is a sparse matrix.

Let \( p(x) \overset{\text{def}}{=} \# \text{ of nonzeros in } x \). **Sparsity coefficient:**

\[
\gamma(A) \overset{\text{def}}{=} \frac{p(A)}{MN}.
\]

Example 1: Matrix-vector multiplication

- Computation of vector \( Ax \) needs \( p(A) \) operations.
- Initial complexity \( MN \) is reduced in \( \gamma(A) \) times.
Gradient Method

\[ x_0 \in Q, \quad x_{k+1} = \pi_Q(x_k - hf'(x_k)), \quad k \geq 0. \]

Main computational expenses

- Projection onto a simple set \( Q \) needs \( O(N) \) operations.
- Displacement \( x_k \rightarrow x_k - hf'(x_k) \) needs \( O(N) \) operations.
- \( f'(x) = A^T \Psi'(Ax) \). If \( \Psi \) is simple, then the main efforts are spent for two matrix-vector multiplications: \( 2p(A) \).

Conclusion: As compared with full matrices, we accelerate in \( \gamma(A) \) times.

Note: For Large- and Huge-scale problems, we often have \( \gamma(A) \approx 10^{-4} \ldots 10^{-6} \). Can we get more?
Sparse updating strategy

Main idea

- After update $x_+ = x + d$ we have $y_+ \overset{\text{def}}{=} Ax_+ = Ax + Ad$.
- What happens if $d$ is sparse?

Denote $\sigma(d) = \{j : d^{(j)} \neq 0\}$. Then $y_+ = y + \sum_{j \in \sigma(d)} d^{(j)} \cdot Ae_j$.

Its complexity, $\kappa_A(d) \overset{\text{def}}{=} \sum_{j \in \sigma(d)} p(Ae_j)$, can be VERY small!

$$\kappa_A(d) = \sum_{j \in \sigma(d)} \gamma(Ae_j) = \gamma(d) \cdot \frac{1}{p(d)} \sum_{j \in \sigma(d)} \gamma(Ae_j) \cdot MN \leq \gamma(d) \max_{1 \leq j \leq m} \gamma(Ae_j) \cdot MN.$$ 

If $\gamma(d) \leq c \gamma(A)$, $\gamma(Ae_j) \leq c \gamma(A)$, then

$$\kappa_A(d) \leq c^2 \cdot \gamma^2(A) \cdot MN.$$

Expected acceleration: $(10^{-6})^2 = 10^{-12} \Rightarrow 1 \text{ sec} \approx 32 000 \text{ years}$
When it can work?

- Simple methods: No full-vector operations! (Is it possible?)
- Simple problems: Functions with sparse gradients.

**Examples**

1. Quadratic function $f(x) = \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle$. The gradient $f'(x) = Ax - b$, $x \in \mathbb{R}^N$, is not sparse even if $A$ is sparse.

2. Piece-wise linear function $g(x) = \max_{1 \leq i \leq m} [\langle a_i, x \rangle - b(i)]$. Its subgradient $f'(x) = a_{i(x)}$, $i(x) : f(x) = \langle a_{i(x)}, x \rangle - b(i(x))$, can be sparse if $a_i$ is sparse!

**But:** We need a fast procedure for updating max-operations.
**Def:** Function $f(x)$, $x \in \mathbb{R}^n$, is *short-tree representable*, if it can be computed by a short binary tree with the height $\approx \ln n$.

Let $n = 2^k$ and the tree has $k + 1$ levels: $v_{0,i} = x^{(i)}$, $i = 1, \ldots, n$.

Size of the next level halves the size of the previous one:

$$v_{i+1,j} = \psi_{i+1,j}(v_{i,2j-1}, v_{i,2j}), \quad j = 1, \ldots, 2^{k-i-1}, \quad i = 0, \ldots, k - 1,$$

where $\psi_{i,j}$ are some bivariate functions.
Main advantages

■ Important examples (symmetric functions)

\[ f(x) = \|x\|_p, \quad p \geq 1, \quad \psi_{i,j}(t_1, t_2) \equiv \left[ |t_1|^p + |t_2|^p \right]^{1/p}, \]

\[ f(x) = \ln \left( \sum_{i=1}^{n} e^{x(i)} \right), \quad \psi_{i,j}(t_1, t_2) \equiv \ln (e^{t_1} + e^{t_2}), \]

\[ f(x) = \max_{1 \leq i \leq n} x^{(i)}, \quad \psi_{i,j}(t_1, t_2) \equiv \max \{t_1, t_2\}. \]

■ The binary tree requires only \(n - 1\) auxiliary cells.

■ Its value needs \(n - 1\) applications of \(\psi_{i,j}(\cdot, \cdot)\) (\(\equiv\) operations).

■ If \(x_+\) differs from \(x\) in one entry only, then for re-computing \(f(x_+\) we need only \(k \equiv \log_2 n\) operations.

Thus, we can have pure subgradient minimization schemes with \textit{Sublinear Iteration Cost}. 
Simple subgradient methods

1. Problem: \[ f^* \overset{\text{def}}{=} \min_{x \in Q} f(x), \] where

- \( Q \) is a closed and convex and \( \|f'(x)\| \leq L(f), x \in Q \),
- the optimal value \( f^* \) is known.

Consider the following optimization scheme (B.Polyak, 1967):

\[ x_0 \in Q, \quad x_{k+1} = \pi_Q \left( x_k - \frac{f(x_k) - f^*}{\|f'(x_k)\|^2} f'(x_k) \right), \quad k \geq 0. \]

Denote \( f_k^* = \min_{0 \leq i \leq k} f(x_i) \). Then for any \( k \geq 0 \) we have:

\[ f_k^* - f^* \leq \frac{L(f)\|x_0 - \pi x^*_0(x_0)\|}{(k+1)^{1/2}}, \]

\[ \|x_k - x^*\| \leq \|x_0 - x^*\|, \quad \forall x^* \in X^*. \]
Proof:

Let us fix $x^* \in X_*$. Denote $r_k(x^*) = \|x_k - x^*\|$. Then

$$r_{k+1}^2(x^*) \leq \left\| x_k - \frac{f(x_k) - f^*}{\|f'(x_k)\|^2} f'(x_k) - x^* \right\|^2,$$

$$= r_k^2(x^*) - 2 \frac{f(x_k) - f^*}{\|f'(x_k)\|^2} \langle f'(x_k), x_k - x^* \rangle + \frac{(f(x_k) - f^*)^2}{\|f'(x_k)\|^2},$$

$$\leq r_k^2(x^*) - \frac{(f(x_k) - f^*)^2}{\|f'(x_k)\|^2} \leq r_k^2(x^*) - \frac{(f_k - f^*)^2}{L^2(f)}.$$

From this reasoning, $\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2$, $\forall x^* \in X^*$.

**Corollary:** Assume $X_*$ has recession direction $d_*$. Then

$$\|x_k - \pi_{X_*}(x_0)\| \leq \|x_0 - \pi_{X_*}(x_0)\|, \quad \langle d_*, x_k \rangle \geq \langle d_*, x_0 \rangle.$$

(Proof: consider $x^* = \pi_{X_*}(x_0) + \alpha d_*$, $\alpha \geq 0$.)
Consider the following method. It has step-size parameter $h > 0$.

If $g(x_k) > h \|g'(x_k)\|$, then (A):
$$x_{k+1} = \pi_Q \left(x_k - \frac{g(x_k)}{\|g'(x_k)\|^2} g'(x_k)\right),$$
else (B):
$$x_{k+1} = \pi_Q \left(x_k - \frac{h}{\|f'(x_k)\|} f'(x_k)\right).$$

Let $\mathcal{F}_k \subseteq \{0, \ldots, k\}$ be the set (B)-iterations, and $f^*_k = \min_{i \in \mathcal{F}_k} f(x_i)$.

**Theorem:** If $k > \frac{\|x_0 - x^*\|^2}{h^2}$, then $\mathcal{F}_k \neq \emptyset$ and
$$f^*_k - f(x) \leq hL(f), \quad \max_{i \in \mathcal{F}_k} g(x_i) \leq hL(g).$$
1. Constants \( L(f), L(g) \) are known \((\text{e.g. Linear Programming})\)

We can take \( h = \frac{\epsilon}{\max\{L(f), L(g)\}} \). Then we need to decide on the number of steps \( N \) (easy!).

**Note:** The standard advice is \( h = \frac{R}{\sqrt{N+1}} \) (much more difficult!)

2. Constants \( L(f), L(g) \) are not known

- Start from a guess.
- Restart from scratch each time we see the guess is wrong.
- The guess is doubled after restart.

3. Tracking the record value \( f_k^* \)

Double run. Other ideas are welcome!
Application examples

Observations:

1. Very often, Large- and Huge- scale problems have repetitive sparsity patterns and/or limited connectivity.
   - Social networks.
   - Mobile phone networks.
   - Truss topology design (local bars).
   - Finite elements models (2D: four neighbors, 3D: six neighbors).

2. For $p$-diagonal matrices $\kappa(A) \leq p^2$. 
Nonsmooth formulation of Google Problem

Main property of spectral radius \((A \geq 0)\)

If \(A \in \mathbb{R}_{+}^{n \times n}\), then \(\rho(A) = \min_{x \geq 0} \max_{1 \leq i \leq n} \frac{1}{x(i)} \langle e_i, Ax \rangle\).

The minimum is attained at the corresponding eigenvector.

Since \(\rho(\bar{E}) = 1\), our problem is as follows:

\[
 f(x) \overset{\text{def}}{=} \max_{1 \leq i \leq N} \left[ \langle e_i, \bar{E}x \rangle - x(i) \right] \rightarrow \min_{x \geq 0}.
\]

**Interpretation:** Maximizing the self-esteem!

Since \(f^* = 0\), we can apply Polyak’s method with sparse updates.

**Additional features:** the optimal set \(X^*\) is a *convex cone*.

If \(x_0 = e\), then the whole sequence is separated from zero:

\[
 \langle x^*, e \rangle \leq \langle x^*, x_k \rangle \leq \|x^*\|_1 \cdot \|x_k\|_{\infty} = \langle x^*, e \rangle \cdot \|x_k\|_{\infty}.
\]

**Goal:** Find \(\bar{x} \geq 0\) such that \(\|\bar{x}\|_{\infty} \geq 1\) and \(f(\bar{x}) \leq \epsilon\).

(First condition is satisfied automatically.)
Computational experiments: Iteration Cost

We compare Polyak’s GM with sparse update ($GM_s$) with the standard one ($GM$).

**Setup:** Each agent has exactly $p$ random friends. Thus, $\kappa(A) \approx p^2$.

**Iteration Cost:** $GM_s \approx p^2 \log_2 N$, $GM \approx pN$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\kappa(A)$</th>
<th>$GM_s$</th>
<th>$GM$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024</td>
<td>1632</td>
<td>3.00</td>
<td>2.98</td>
</tr>
<tr>
<td>2048</td>
<td>1792</td>
<td>3.36</td>
<td>6.41</td>
</tr>
<tr>
<td>4096</td>
<td>1888</td>
<td>3.75</td>
<td>15.11</td>
</tr>
<tr>
<td>8192</td>
<td>1920</td>
<td>4.20</td>
<td>139.92</td>
</tr>
<tr>
<td>16384</td>
<td>1824</td>
<td>4.69</td>
<td>408.38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\kappa(A)$</th>
<th>$GM_s$</th>
<th>$GM$</th>
</tr>
</thead>
<tbody>
<tr>
<td>131072</td>
<td>576</td>
<td>0.19</td>
<td>213.9</td>
</tr>
<tr>
<td>262144</td>
<td>592</td>
<td>0.25</td>
<td>477.8</td>
</tr>
<tr>
<td>524288</td>
<td>592</td>
<td>0.32</td>
<td>1095.5</td>
</tr>
<tr>
<td>1048576</td>
<td>608</td>
<td>0.40</td>
<td>2590.8</td>
</tr>
</tbody>
</table>

1 sec $\approx$ 100 min!
Convergence of \( GM_s \): Medium Size

Let \( N = 131072, \ p = 16, \ \kappa(A) = 576, \) and \( L(f) = 0.21. \)

<table>
<thead>
<tr>
<th>Iterations</th>
<th>( f - f^* )</th>
<th>Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 1.0 \cdot 10^5 )</td>
<td>0.1100</td>
<td>16.44</td>
</tr>
<tr>
<td>( 3.0 \cdot 10^5 )</td>
<td>0.0429</td>
<td>49.32</td>
</tr>
<tr>
<td>( 6.0 \cdot 10^5 )</td>
<td>0.0221</td>
<td>98.65</td>
</tr>
<tr>
<td>( 1.1 \cdot 10^6 )</td>
<td>0.0119</td>
<td>180.85</td>
</tr>
<tr>
<td>( 2.2 \cdot 10^6 )</td>
<td>0.0057</td>
<td>361.71</td>
</tr>
<tr>
<td>( 4.1 \cdot 10^6 )</td>
<td>0.0028</td>
<td>674.09</td>
</tr>
<tr>
<td>( 7.6 \cdot 10^6 )</td>
<td>0.0014</td>
<td>1249.54</td>
</tr>
<tr>
<td>( 1.0 \cdot 10^7 )</td>
<td>0.0010</td>
<td>1644.13</td>
</tr>
</tbody>
</table>

Dimension and accuracy are sufficiently high, but the time is still reasonable.
Convergence of $GM_s$: Large Scale

Let $N = 1048576$, $p = 8$, $\kappa(A) = 192$, and $L(f) = 0.21$. 

<table>
<thead>
<tr>
<th>Iterations</th>
<th>$f - f^*$</th>
<th>Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.000000</td>
<td>0.00</td>
</tr>
<tr>
<td>$1.0 \cdot 10^5$</td>
<td>0.546662</td>
<td>7.69</td>
</tr>
<tr>
<td>$4.0 \cdot 10^5$</td>
<td>0.276866</td>
<td>30.74</td>
</tr>
<tr>
<td>$1.0 \cdot 10^6$</td>
<td>0.137822</td>
<td>76.86</td>
</tr>
<tr>
<td>$2.5 \cdot 10^6$</td>
<td>0.063099</td>
<td>192.14</td>
</tr>
<tr>
<td>$5.1 \cdot 10^6$</td>
<td>0.032092</td>
<td>391.97</td>
</tr>
<tr>
<td>$9.9 \cdot 10^6$</td>
<td>0.016162</td>
<td>760.88</td>
</tr>
<tr>
<td>$1.5 \cdot 10^7$</td>
<td>0.010009</td>
<td>1183.59</td>
</tr>
</tbody>
</table>

**Final point** $\bar{x}_*$: $\|\bar{x}_*\|_\infty = 2.941497$, $R_0^2 \overset{\text{def}}{=} \|\bar{x}_* - e\|_2^2 = 1.2 \cdot 10^5$. 

**Theoretical bound:** $\frac{L^2(f)R_0^2}{\epsilon^2} = 5.3 \cdot 10^7$. **Time for GM:** $\approx 1$ year!
1. Sparse GM is an efficient and reliable method for solving Large- and Huge- Scale problems with uniform sparsity.

2. We can treat also dense rows. Assume that inequality $\langle a, x \rangle \leq b$ is dense. It is equivalent to the following system:

   \[
   y^{(1)} = a^{(1)} x^{(1)}, \quad y^{(j)} = y^{(j-1)} + a^{(j)} x^{(j)}, \quad j = 2, \ldots, n,
   \]

   \[
   y^{(n)} \leq b.
   \]

   We need new variables $y^{(j)}$ for all nonzero coefficients of $a$.

   - Introduce $p(a)$ additional variables and $p(A)$ additional equality constraints. (No problem!)
   - Hidden drawback: the above equalities are satisfied with errors.
   - May be it is not too bad?

3. Similar technique can be applied to dense columns.
Assume that $\kappa(A) \approx \gamma^2(A)n^2$. Compare three methods:

- **Sparse updates (SU).** Complexity $\gamma^2(A)n^2\frac{L^2R^2}{\epsilon^2} \log n$ operations.
- **Smoothing technique (ST).** Complexity $\gamma(A)n^2\frac{LR}{\epsilon}$ operations.
- **Polynomial-time methods (PT).** Complexity $(\gamma(A)n + n^3)n \log \frac{LR}{\epsilon}$ operations.

There are three possibilities.

- **Low accuracy:** $\gamma(A)\frac{LR}{\epsilon} < 1$. Then we choose SU.
- **Moderate accuracy:** $1 < \gamma(A)\frac{LR}{\epsilon} < n^2$. We choose ST.
- **High accuracy:** $\gamma(A)\frac{LR}{\epsilon} > n^2$. We choose PT.

**NB:** For Huge-Scale problems usually $\gamma(A) \approx \frac{1}{n} \Rightarrow \frac{LR}{\epsilon} \lor n$