

**Smooth minimization
of non-smooth functions**

Yu.Nesterov

CORE-INMA

Catholic University of Louvain
Louvain-la-Neuve, Belgium

Subgradient methods for non-smooth minimization

Advantages:

- Very simple iteration scheme.
- Low memory requirements.
- Optimal rate of convergence (uniformly in the dimension).
- Interpretation of the process.
- Extensions (saddle points, variational inequalities, stochastic optimization, etc.)

Objections:

- Low rate of convergence. (Confirmed by theory!)
- No acceleration.
- Very high sensitivity to the step-size strategy.
- Absence of dual information.
- No reliable stopping criterion.

Non-smooth unconstrained minimization

Problem:

$$\min \{ f(x) : x \in R^n \} \Rightarrow x^*, f^* = f(x^*),$$

where $f(x)$ is a non-smooth convex function.

Subgradients:

$$g \in \partial f(x) \Leftrightarrow f(y) \geq f(x) + \langle g, y - x \rangle \quad \forall y \in R^n.$$

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:

$$f(x) = \max_{1 \leq j \leq m} \{ \langle a_j, x \rangle + b^{(j)} \}$$

$$\partial f(x) = \text{Conv} \{ a_j : \langle a_j, x \rangle + b^{(j)} = f(x) \}.$$

Lower complexity bounds (Nemirovsky, Yudin 1976)

Model of the problem:

$f(x)$ is given by a *black-box* local oracle.

- We can see only $\mathcal{O}(x) = (f(x), g(x) \in \partial f(x))$.
- Change of f at y does not imply a change at x .

Note: “We” \equiv “method”.

Theorem:

In a black-box local setting it is impossible to converge faster than $O\left(\frac{1}{\sqrt{k}}\right)$ uniformly in the dimension of space of variables.

(Proof: look at $\max_{1 \leq i \leq n} x^{(i)}$.)

Here:

- k is the number of calls of oracle.
- Any auxiliary computation is allowed.
- No memory limitations.

Note: Convergence is very slow.

Question:

We want to find an approximate solution of the problem

$$\max_{1 \leq j \leq m} \left\{ \langle a_j, x \rangle + b^{(j)} \right\} \rightarrow \min_x : x \in R^n, \quad (1)$$

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

What rate of convergence we can guarantee?

Answer (Complexity Theory): In order to find an ϵ -solution we need at least

$$O\left(\frac{1}{\epsilon^2}\right)$$

calls of oracle (\equiv iterations).

Goal of the talk: a gradient scheme for solving (1) with efficiency estimate

$$O\left(\frac{1}{\epsilon}\right)$$

iterations.

Reason of speed up: (1) *is not* a black box.

Complexity of smooth minimization

Problem:

$$f(x) \rightarrow \min_x : x \in R^n, \quad (2)$$

where f is a convex function with Lipschitz-continuous gradient:

$$\|\nabla f(x) - \nabla f(y)\|^* \leq L(f)\|x - y\|$$

for all $x, y \in R^n$. (Notation: $f \in C^{1,1}(R^n)$.)

Rate of convergence:

1. Gradient method: $O\left(\frac{L(f)}{k}\right)$
2. Optimal method: $O\left(\frac{L(f)}{k^2}\right)$
[Nesterov 1983]

Complexity: $O\left(\sqrt{\frac{L(f)}{\epsilon}}\right)$.

Note: The difference with $O\left(\frac{1}{\epsilon^2}\right)$ is enormous.

Main questions:

1. Given by a non-smooth convex $f(x)$, can we find for it a smooth ϵ -approximation $f_\epsilon(x)$ with

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

If yes, we need only $O\left(\frac{1}{\epsilon}\right)$ iterations.

2. Can we do that in a systematic way?

Conclusion:

We need a convenient model of our problem.

(Compare with the theory of self-concordant functions.)

Adjoint problem

Primal problem:

$$\text{Find } f^* = \min_x \{f(x) : x \in Q_1\},$$

where $Q_1 \subset E_1$ is convex closed and bounded.

Model of objective function:

$$f(x) = \hat{f}(x) + \max_u \{ \langle Ax, u \rangle_2 - \hat{\phi}(u) : u \in Q_2 \},$$

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 $A : E_1 \rightarrow E_2^*$.

Adjoint problem:

$$\max_u \{ \phi(u) : u \in Q_2 \},$$

$$\phi(u) = -\hat{\phi}(u) + \min_x \{ \langle Ax, u \rangle_2 + \hat{f}(x) : x \in Q_1 \}.$$

Note: Adjoint problem is not unique!

Example: Consider

$$f(x) = \max_{1 \leq j \leq m} |\langle a_j, x \rangle_1 - b^{(j)}|.$$

1. $Q_2 \equiv E_2 = E_1^*$, $A = I$,

$$\begin{aligned} \hat{\phi}(u) &\equiv f_*(u) = \max_x \{ \langle u, x \rangle_1 - f(x) : x \in E_1 \} \\ &= \min_{s \in R^m} \left\{ \sum_{j=1}^m s^{(j)} b^{(j)} : u = \sum_{j=1}^m s^{(j)} a_j, \sum_{j=1}^m |s^{(j)}| \leq 1 \right\}. \end{aligned}$$

2. $E_2 = R^m$, $\hat{\phi}(u) = \langle b, u \rangle_2$,

$$\begin{aligned} f(x) &= \max_{1 \leq j \leq m} |\langle a_j, x \rangle_1 - b^{(j)}| \\ &= \max_{u \in R^m} \left\{ \sum_{j=1}^m u^{(j)} [\langle a_j, x \rangle_1 - b^{(j)}] : \sum_{j=1}^m |u^{(j)}| \leq 1 \right\}. \end{aligned}$$

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

$$\begin{aligned} f(x) &= \max_{u \in R^{2m}} \left\{ \sum_{j=1}^m (u_1^{(j)} - u_2^{(j)}) \cdot [\langle a_j, x \rangle_1 - b^{(j)}] : \right. \\ &\quad \left. \sum_{j=1}^m (u_1^{(j)} + u_2^{(j)}) = 1, u \geq 0 \right\}. \end{aligned}$$

Rule: Increase in $\dim E_2 \Rightarrow$ decrease in 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 *smoothness* 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 the constant

$$L_\mu = \frac{1}{\mu\sigma_2} \|A\|_{1,2}^2,$$

where $\|A\|_{1,2} = \max_{x,u} \{ \langle Ax, u \rangle_2 : \|x\|_1 = 1, \|u\|_2 = 1 \}$.

Note: 1. for any $\mu \geq 0$ and $x \in E_1$ 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.

Optimal method

Problem: $\min_x \{f(x) : x \in Q_1\}$ with $f \in C^{1,1}(Q_1)$.

Prox-function: strongly convex $d_1(x)$, $x \in Q_1$:

$$d_1(x_0) = 0, \quad d_1(x) \geq 0 \quad \forall x \in Q_1.$$

Gradient mapping:

$$T(x) = \arg \min_{y \in Q_1} \left\{ \langle \nabla f(x), y - x \rangle_1 + \frac{1}{2} L(f) \|y - x\|_1^2 \right\}.$$

Method: For $k \geq 0$ do

1. Compute $f(x_k), \nabla f(x_k)$.
2. Find $y_k = T(x_k)$.
3. Find $z_k = \arg \min_{x \in Q_1} \left\{ \frac{L(f)}{\sigma} d_1(x) + \sum_{i=0}^k \frac{i+1}{2} \langle \nabla f(x_i), x \rangle_1 \right\}$.
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:

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

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 $N \geq 1$. 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

$$0 \leq f(\hat{x}) - \phi(\hat{u}) \leq \frac{4\|A\|_{1,2}}{N+1} \cdot \sqrt{\frac{D_1 D_2}{\sigma_1 \sigma_2}} + \frac{4L(\hat{f})D_1}{\sigma_1 \cdot (N+1)^2}.$$

Corollary. Let $L(\hat{f}) = 0$. To get ϵ -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 1: Equilibrium in matrix games

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

$$\|x\|_1 = \left[\sum_{i=1}^n (x^{(i)})^2 \right]^{1/2}, \quad \|u\|_2 = \left[\sum_{j=1}^m (u^{(j)})^2 \right]^{1/2},$$

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

$$\|A\|_{1,2} = \lambda_{\max}^{1/2}(A^T A)$$

and

$$f(\hat{x}) - \phi(\hat{u}) \leq \frac{4\lambda_{\max}^{1/2}(A^T A)}{N+1}.$$

2. Entropy distance. Let us choose

$$\|x\|_1 = \sum_{i=1}^n |x^{(i)}|, \quad d_1(x) = \ln n + \sum_{i=1}^n x^{(i)} \ln x^{(i)},$$

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

Then

$$\sigma_1 = \sigma_2 = 1, \quad D_1 = \ln n, \quad D_2 = \ln m.$$

Moreover, since

$$\begin{aligned} \|A\|_{1,2} &= \max_x \left\{ \max_{1 \leq j \leq m} |\langle a_j, x \rangle| : \|x\|_1 = 1 \right\} \\ &= \max_{i,j} |A^{(i,j)}|, \end{aligned}$$

we have

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

Note: 1. Usually $\max_{i,j} |A^{(i,j)}| \ll \lambda_{\max}^{1/2}(A^T A)$.

2. $\bar{f}_\mu(x)$ is easily computable:

$$\bar{f}_\mu(x) = \langle c, x \rangle_1 + \mu \ln \left(\frac{1}{m} \sum_{j=1}^m e^{[\langle a_j, x \rangle + b^{(j)}]/\mu} \right).$$

Example 2: Continuous location problem

Problem: p cities with population m_j are located at

$$c_j \in R^n, \quad j = 1, \dots, p.$$

Construct a service center at point x^* , which minimizes the total distance to the center. That is

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

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: $E_2 = (E_1^*)^p$, $\|u\|_2^2 = \sum_{j=1}^p m_j (\|u_j\|_1^*)^2$,

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

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

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

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

Ex.3: Variational inequalities (linear operator)

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

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

Problem:

$$\text{Find } w^* \in Q : \quad \langle B(w^*), w - w^* \rangle \geq 0 \quad \forall w \in Q, \quad (3)$$

where Q is a bounded convex closed set.

Merit function:

$$\psi(w) = \max_v \{ \langle B(v), w - v \rangle : v \in Q \}. \quad (4)$$

- $\psi(w)$ is convex on E_1 .
- $\psi(w) \geq 0$ for all $w \in Q$.
- $\psi(w) = 0$ if and only if w solves (3).
- $\langle B(v), v \rangle$ is a *convex* function. Thus, (4) 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:

$$\phi_\mu(v) = \min_w \{ \langle B(v), w - v \rangle + \mu d_1(w) : w \in Q \}.$$

(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 $\|x\|_1^2 = \sum_{i=1}^n (x^{(i)})^2$, $d_1(x) = \frac{1}{2}\|x\|^2$.

It is convenient to choose $E_2 = R^{2m}$,

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

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

Complexity:

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

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

$$\bar{f}_\mu(x) = \mu \ln \left(\frac{1}{m} \sum_{j=1}^m \xi \left(\frac{1}{\mu} [\langle a_j, x \rangle + b^{(j)}] \right) \right)$$

2. Sum of absolute values. Consider

$$\min_x \left\{ f(x) = \sum_{j=1}^m |\langle a_j, x \rangle_1 - b^{(j)}| : x \in Q_1 \right\}. \quad (5)$$

Let us choose

$$E_2 = R^m, \quad Q_2 = \{u \in R^m : |u^{(j)}| \leq 1, j = 1, \dots, m\},$$

$$d_2(u) = \frac{1}{2} \|u\|_2^2 = \frac{1}{2} \sum_{j=1}^m \|a_j\|_1^* \cdot (u^{(j)})^2.$$

Then

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

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

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

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

Note: The bound and the scheme allow $m \rightarrow \infty$.

Computational experiments

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: Pentium 4, 2.6GHz.

Iteration: $2mn$ operations.

Results for $\epsilon = 0.01$. Table 1

$m \backslash n$	100	300	1000	3000	10000
100	808 0''	1011 0''	1112 3''	1314 12''	1415 44''
300	910 0''	1112 2''	1415 10''	1617 35''	1819 135''
1000	1112 2''	1213 8''	1415 32''	1718 115''	2020 451''

Number of iterations: 40 – 50% of predicted values.

Results for $\epsilon = 0.001$. Table 2

$m \backslash n$	100	300	1000	3000	10000
100	6970 2''	8586 8'	9394 29'',	10000 91''	10908 349''
300	7778 8''	10101 27''	12424 97''	14242 313''	15656 1162''
1000	8788 30''	11010 105''	13030 339''	15757 1083''	18282 4085''

Results for $\epsilon = 0.0001$. Table 3

$m \backslash n$	100	300	1000	3000
100	67068 25''	72073 80''	74075 287''	80081 945''
300	85086 89'', 42%	92093 243''	101102 914''	112113 3302''
1000	97098 331''	100101 760''	116117 2936''	139140 11028''

Comparing the bounds

Gradient: $2 \cdot 4 \cdot \frac{mn}{\epsilon} \sqrt{\ln n \ln m}.$

Short-step path-following method ($n \geq m$):

$$\left(7.2\sqrt{n} \ln \frac{1}{\epsilon}\right) \cdot \frac{m(m+1)}{2} n.$$

Right digits

m	n	2	3	4	5
100	100	g	g	b	b
100	300	g	g	b	b
100	1000	g	g	b	b
100	3000	g	g	b	b
100	10000	g	g	b	b
300	300	g	g	b	b
300	1000	g	g	b	b
300	3000	g	g	=	b
300	10000	g	g	g	b
1000	1000	g	g	g	b
1000	3000	g	g	g	b
1000	10000	g	g	g	=

g - gradient method, b - barrier method