Convex Optimization for Data Science

Gasnikov Alexander

gasnikov.av@mipt.ru

Lecture 3. Complexity of optimization problems & Optimal methods for convex optimization problems

October, 2016

Complexity theory of convex optimization was built in 1976–1979 mainly in works of Arkadi Nemirovski

Main books:

Nemirovski A. Efficient methods in convex programming. Technion, 1995. http://www2.isye.gatech.edu/~nemirovs/Lec_EMCO.pdf

Nesterov Yu. Introduction Lectures on Convex Optimization. A Basic Course. Applied Optimization. – Springer, 2004.

Nemirovski A. Lectures on modern convex optimization analysis, algorithms, and engineering applications. – Philadelphia: SIAM, 2013.

Bubeck S. Convex optimization: algorithms and complexity // In Foundations and Trends in Machine Learning. – 2015. – V. 8. – no. 3-4. – P. 231–357.

Guzman C., Nemirovski A. On lower complexity bounds for large-scale smooth convex optimization // Journal of Complexity. 2015. V. 31. P. 1–14.

Gasnikov A., Nesterov Yu. Universal fast gradient method for stochastic composit optimization problems // Comp. Math. & Math. Phys. 2016. (in print) https://arxiv.org/ftp/arxiv/papers/1604/1604.05275.pdf

Structure of Lecture 3

- Pessimistic lower bound for non convex problems
- Resisting oracle
- Optimal estimation for convex optimization problems
- Lower complexity bounds
- Optimal and not optimal methods
- Mirror Descent
- Gradient Descent
- Similar Triangles Method (Fast Gradient Method)
- Open gap problem of A. Nemirovski
- Structural optimization (looking into the Black Box)
- Conditional problems
- Interior Point Method

Two practice examples (A. Nemirovski)

Stability number of graph

$$\sum_{i=1}^{n} x_i \to \max_{\substack{x_i^2 - x_i = 0 \\ x_i x_j = 0,\ (i,j)\in\Gamma}}, \quad n = 256.$$

La Tour Eiffel problem

$$x_0 \to \min_{\substack{\lambda_{\min}\begin{pmatrix} x_1, \dots\dots\dots, x_1^l \\ \dots\dots\dots\dots\dots\dots \\ \dots\dots\dots, x_m, x_m^l \\ x_1^l, \dots\dots, x_m^l, x_0 \end{pmatrix} \ge 0,\ l=1,\dots,k \\ \sum_{j=1}^{m} a_j x_j^l = b^l,\ l=1,\dots,k;\ \sum_{j=1}^{m} x_j = 1,}}, \quad k = 6,\ m = 160.$$

Which of these two problems harder to solve? Intuition says – the second. But the first problem is not convex and it's NP-hard. The best known method finds 0.5-solution required $2^n \simeq 10^{77}$ flop. The second problem is convex and one can find 10^{-6}-solution by CVX for few seconds (Lecture 1).

Pessimistic lower bound for non convex problems

Assume that we have to solve ($B_\infty^n(1)$ – unit cube in $\mathbb{R}^n$)

$$F(x) \to \min_{x \in B_\infty^n(1)},$$

in sense

$$F(x^N) - \min_{x \in B_\infty^n(1)} F(x) \le \varepsilon,$$

where $\left| d^k F(x+te) / dt^k \right| \le 1$ (k is fixed, $1 \le k \le n$) for all $e \in B_\infty^n(1)$.

For arbitrary method imposed with local oracle (this oracle in request for fixed point can return as high derivatives of $F(x)$ as we asked) we have that required number of (randomized) oracle calls is: $\boxed{N \succ \varepsilon^{-n/k}}$ and for one extremum problem for deterministic oracle is: $\boxed{N \succ \varepsilon^{-(n-1)/k}.}$

Resisting oracle: Uniform Grid method is worst-case optimal.

Resisting oracle (build online "bad" function for the method)

For simplicity consider 0-order oracle (return the value of the function). Divide $B_\infty^n(1)$ on m^n sub-cubes $B_\infty^n\left(1/(2m)\right)$. Assume that

$$\left|F(y)-F(x)\right| \le M\left\|y-x\right\|_\infty .$$

At each point reply $F\left(x^k\right)=0$. When $N<m^n$ there is ball $B_\infty^n\left(1/(2m)\right)$ with no question. Hence we can take

$$\min_{x\in B_\infty^n(1)} F(x) = -\frac{M}{2m}.$$

Thus $\varepsilon \ge M/(2m)$. Therefore, choosing $N=m^n-1$ one can obtain:

$$\boxed{N \ge \left(\frac{M}{2\varepsilon}\right)^n - 1}.$$

Optimal estimation for Convex Optimization problems ($N \geq n$)

$$F(x) \to \min_{x \in Q},$$

Q – compact (it's significant!) convex set, $n = \dim x$. We assume that $F(x^N) - F_* \leq \varepsilon$, where N – number of required iterations (calculations $\partial F(x)$ or separation hyperplane to Q or its cutting part).

$$\boxed{N \sim n \ln(\Delta F/\varepsilon),}$$

where $\Delta F = \sup_{x,y \in Q} \{F(y) - F(x)\}$. Additional iteration complexity is $\boxed{\tilde{O}(n^2)}$.

Lee Y.-T., Sidford A., Wong S.C-W. A faster cutting plane methods and its implications for combinatorial and convex optimization // e-print, 2015. https://arxiv.org/pdf/1508.04874v2.pdf

Ellipsoid method: $N \sim n^2 \ln(\varepsilon^{-1})$. Additional iteration complexity is $\tilde{O}(n^2)$.

LP in P by ellipsoid algorithm (L. Khachyan, 1978)

Assume we have to answer is $Ax \le b$ solvable ($n = \dim x$, $m = \dim b$)? We assume that all elements of A and b are integers. And we'd like to find one of **the exact** solutions x_*. This problem up to a logarithmic factor in complexity is equivalent to find the exact solution of LP problem $\langle c, x \rangle \to \min\limits_{Ax \le b}$ with integer A, b and c. To find the exact solution of $Ax = b$ one can use polynomial Gauss algorithm $\mathrm{O}\left(n^3\right)$. What is about $Ax \le b$? Let's introduce

$$L = \sum_{i,j=1,1}^{m,n} \log_2 \left|a_{ij}\right| + \sum_{i=1}^{m} \log_2 \left|b_i\right| + \log_2 \left(mn\right) + 1.$$

Useful properties: $\left\|x_*\right\|_\infty \le 2^L$; if $Ax - b \le 0$ is incompaitable then for all x $\left\|\left(Ax - b\right)_+\right\|_\infty \ge 2^{-(L-1)}$. Works in $\mathrm{O}\left(nL\right)$-bit arithmetic with $\tilde{\mathrm{O}}\left(mn + n^2\right)$ cost of PC memory one can find x_* (if it's exist) for $\tilde{\mathrm{O}}\left(n^3\left(n^2 + m\right)L\right)$ a.o.

LP in P? – is still an open question

Simplex Method (Kantorovich–Dantzig) solve (exactly since it's finite method) LP in polynomial time $\tilde{O}(m^3)$ only "in average" (Borgward, Smale, Vershik–Sporyshev; 1982–1986). Klee–Minty example (1972) shows that in worth case simplex methods required to get round all the vertexes of polyhedral (exponential number). At the very beginning of this century Spielman–Tseng (smooth analysis) show that if $A := A + \|A\| G$, where $G = \|g_{ij}\|_{i,j=1,1}^{m.n}$, i.i.d. $g_{ij} \in N(0, \tilde{\sigma}^2)$ and $T_{\tilde{\sigma}}(A)$ – time required by special version of Simplex Method to find exact solution, then

$$E_G\left[T_{\tilde{\sigma}}(A)\right] = \mathrm{Poly}\left(n, m, \tilde{\sigma}^{-1}\right). \text{ // } \log\left(\tilde{\sigma}^{-1}\right)? \text{ – \textit{an open question}}$$

In ideal arithmetic with real numbers it is still an open question (Blum–Shub–Smale): is it possible to find **the exact** solution of LP problem (with real numbers) in polynomial time in ideal arithmetic ($\pi \cdot e$ – costs $O(1)$).

Optimal estimations for Convex Optimization problems ($N \le n$)

$$F(x) \to \min_{x \in Q}.$$

We assume that

$$F(x^N) - F_* \le \varepsilon.$$

$N -$ number of required iterations (calculations of $F(x)$ and $\partial F(x)$).

$R -$ "distance" between starting point and nearest solution.

N	$\lvert F(y) - F(x) \rvert \le M \lVert y - x \rVert$	$\lVert \nabla F(y) - \nabla F(x) \rVert_* \le L \lVert y - x \rVert$
$F(x)$ convex	$\dfrac{M^2 R^2}{\varepsilon^2}$	$\sqrt{\dfrac{LR^2}{\varepsilon}}$
$F(x)$ μ-strongly convex	$\dfrac{M^2}{\mu\varepsilon}$	$\sqrt{\dfrac{L}{\mu}} \left\lceil \ln\left(\dfrac{\mu R^2}{\varepsilon}\right) \right\rceil$ $(\forall N)$

If norm is non euclidian then the last row is true up to $\mathrm{O}(\ln n)$-factor.

Lower complexity bound. Non smooth case ($N < n$)

Let's introduce

$$Q = B_2^n(2R),\ F_N(x) = M \max_{1 \le i \le N} x_i + \frac{\mu}{2}\|x\|_2^2,\ \mu = \frac{M}{R\sqrt{N}},$$

$$x^{k+1} = x^0 + \mathrm{Lin}\left\{\partial f\left(x^0\right),\ldots,\partial f\left(x^k\right)\right\}. \text{ // method}$$

Solving the problem

$$M\tau + \frac{\mu N}{2}\tau^2 \to \min_{\tau}$$

we get $\tau_* = -R\big/\sqrt{N}$, $\|x_*\|_2^2 = N\tau_*^2 = R^2$, $F_N^* = \min_{x \in Q} F_N(x) = -MR\big/\sqrt{N}$. If $x^0 = 0$ then after N iteration we can keep $x_i^N = 0$ for $i > N$. So we have

$$F_{N+1}\left(x^{N+1}\right) - F_{N+1}^* \ge -F_{N+1}^* = \left\{\frac{MR}{\sqrt{N+1}}, \frac{M^2}{2\mu\cdot(N+1)}\right\}.$$

Lower complexity bound. Smooth case

Let's introduce **(**$2N+1<n$**)**: $x^1=0,\ x^k\in \mathrm{Lin}\left\{\nabla f\left(x^1\right),...,\nabla f\left(x^k\right)\right\}$,

$$F_N\left(x\right)=\frac{L}{8}\left[x_1^2+\sum_{i=1}^{2N+1}\left(x_i-x_{i+1}\right)^2+x_{2N+1}^2\right]-\frac{L}{4}x_1,$$

Then $$\min_{1\le k\le N}F_N\left(x^k\right)-F_N^*\ge\frac{3L}{32}\frac{\left\|x^1-x_*\right\|_2^2}{\left(N+1\right)^2}.$$

Let's introduce $\chi=L/\mu$

$$F\left(x\right)=\frac{\mu\cdot\left(\chi-1\right)}{8}\left[x_1^2+\sum_{i=1}^{\infty}\left(x_i-x_{i+1}\right)^2-2x_1\right]+\frac{\mu}{2}\left\|x\right\|_2^2$$

Then $$F\left(x^N\right)-F_*\ge\frac{\mu}{2}\left(\frac{\sqrt{\chi}-1}{\sqrt{\chi}+1}\right)^{2\left(N-1\right)}\cdot\left\|x^1-x_*\right\|_2^2$$ (with arbitrary $N\ge1$).

Optimal method for non-smooth convex case (B. Polyak, N. Shor)

Let’s consider unconstrained convex case

$$f\left(x\right) \to \min_x . \qquad (1)$$

We search such $\bar{x}^N$ that

$$f\left(\bar{x}^N\right) - f_* \le \varepsilon ,$$

where $f_* = f\left(x_*\right)$ – optimal value of function in (1), x_* – solution of (1). Let’s introduce

$$\tilde{B}_2^n\left(x_*, R\right) = \left\{x \in \mathbb{R}^n : \left\|x - x_*\right\|_2 \le R\right\}.$$

The main iterative process is (for simplicity we’ll denote $\partial f\left(x\right) = \nabla f\left(x\right)$)

$$\boxed{x^{k+1} = x^k - h\nabla f\left(x^k\right)}. \qquad (2)$$

Assume that under $x \in \tilde{B}_2^n\left(x_*, \sqrt{2}R\right)$

$$\left\|\nabla f\left(x\right)\right\|_2 \le M\,,$$

where $R = \left\|x^0 - x_*\right\|_2$.

Hence from (2), (5) we have

$$\left\|x - x^{k+1}\right\|_2^2 = \left\|x - x^k + h\nabla f\left(x^k\right)\right\|_2^2 =$$

$$= \left\|x - x^k\right\|_2^2 + 2h\left\langle \nabla f\left(x^k\right), x - x^k\right\rangle + h^2\left\|\nabla f\left(x^k\right)\right\|_2^2 \le$$

$$\le \left\|x - x^k\right\|_2^2 + 2h\left\langle \nabla f\left(x^k\right), x - x^k\right\rangle + h^2 M^2. \qquad (3)$$

Here we choose $x = x_*$ (if x_* isn't unique, we choose the nearest x_* to x^0)

$$f\left(\frac{1}{N}\sum_{k=0}^{N-1}x^k\right)-f_*\le\frac{1}{N}\sum_{k=0}^{N-1}f\left(x^k\right)-f\left(x_*\right)\le\frac{1}{N}\sum_{k=0}^{N-1}\left\langle\nabla f\left(x^k\right),x^k-x_*\right\rangle\le$$

$$\le\frac{1}{2hN}\sum_{k=0}^{N-1}\left\{\left\|x_*-x^k\right\|_2^2-\left\|x_*-x^{k+1}\right\|_2^2\right\}+\frac{hM^2}{2}=$$

$$=\frac{1}{2hN}\left(\left\|x_*-x^0\right\|_2^2-\left\|x_*-x^N\right\|_2^2\right)+\frac{hM^2}{2}.$$

If
$$h=\frac{R}{M\sqrt{N}},\ \bar{x}^N=\frac{1}{N}\sum_{k=0}^{N-1}x^k,$$

then

$$\boxed{f\left(\bar{x}^N\right)-f_*\le\frac{MR}{\sqrt{N}}}. \qquad (4)$$

This means that

$$\boxed{N = \frac{M^2 R^2}{\varepsilon^2}}, \ \boxed{h = \frac{\varepsilon}{M^2}}.$$

Note that

$$0 \le \frac{1}{2hk}\left(\left\| x_* - x^0 \right\|_2^2 - \left\| x_* - x^k \right\|_2^2\right) + \frac{hM^2}{2},$$

Hence for all $k = 0, \ldots, N$

$$\left\| x_* - x^k \right\|_2^2 \le \left\| x_* - x^0 \right\|_2^2 + h^2 M^2 k \le 2\left\| x_* - x^0 \right\|_2^2,$$

therefore

$$\boxed{\left\| x^k - x_* \right\|_2 \le \sqrt{2}\left\| x^0 - x_* \right\|_2}, \ k = 0, \ldots, N. \qquad (5)$$

For general (constrained) case

$$f(x) \to \min_{x \in Q} \tag{6}$$

we introduce norm $\| \ \|$, prox-function $d(x) \geq 0$ ($d(x^0) = 0$) which is 1-strongly convex due to $\| \ \|$ and Bregman's divergence

$$V(x,z) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle.$$

We put $R^2 = V(x_*, x^0)$, where x_* – is solution of (6) (if x_* isn't unique then we assume that x_* is minimized $V(x_*, x^0)$). So instead of (3) we'll have

$$2V(x, x^{k+1}) \leq 2V(x, x^k) + 2h\langle \nabla f(x^k), x - x^k \rangle + h^2 M^2 \ (\|\nabla f(x)\|_* \leq M).$$

Mirror Descent (A. Nemirovski, 1977), for $k = 0, ..., N-1$

$$\boxed{x^{k+1} = \mathrm{Mirr}_{x^k}\left(h \partial f(x^k)\right), \ \mathrm{Mirr}_{x^k}(\mathrm{v}) = \arg\min_{x \in Q}\left\{\langle \mathrm{v}, x - x^k \rangle + V(x, x^k)\right\}.}$$

And analogues of formulas (4), (5) are also valid.

$$f\left(\bar{x}^N\right) - f_* \le \frac{\sqrt{2}MR}{\sqrt{N}},\ \left\|x^k - x_*\right\| \le 2\sqrt{V\left(x_*, x^0\right)},\ h = \frac{\varepsilon}{M^2}.$$

Typically, $\frac{1}{2}\left\|x_* - x^0\right\|^2 \le R^2 \le C\ln n \cdot \left\|x_* - x^0\right\|^2$.

Examples

Example 1. $Q = \mathbb{R}^n_+ = \left\{x \in \mathbb{R}^n :\ x \ge 0\right\},\ \left\|\nabla f\left(x\right)\right\|_2 \le M,\ x \in Q,$

$$\|\ \| = \|\ \|_2,\ d\left(x\right) = \frac{1}{2}\left\|x - \bar{x}\right\|_2^2,\ \bar{x} \in \operatorname{int} Q,\ h = \varepsilon/M^2,\ x^0 = \bar{x},$$

$$x^{k+1} = \left[x^k - h\nabla f\left(x^k\right)\right]_+ = \max\left\{x^k - h\nabla f\left(x^k\right), 0\right\},\ k = 1, \ldots, N-1,$$

where $\max\{\ \}$ is taken component-wise. □

Example 2. $Q = S_n(1) = \left\{ x \in \mathbb{R}^n_+ : \sum_{k=1}^{n} x_k = 1 \right\}$, $\|\nabla f(x)\|_\infty \le M$, $x \in Q$,

$\| \ \| = \| \ \|_1$, $d(x) = \ln n + \sum_{i=1}^{n} x_i \ln x_i$, $h = M^{-1}\sqrt{2\ln n / N}$, $x_i^0 = 1/n$, $i = 1,\dots,n$,

For $k = 0,\dots,N-1$, $i = 1,\dots,n$

$$x_i^{k+1} = \frac{\exp\left(-h\sum_{r=1}^{k}\nabla_i f\left(x^r\right)\right)}{\sum_{l=1}^{n}\exp\left(-h\sum_{r=1}^{k}\nabla_l f\left(x^r\right)\right)} = \frac{x_i^k \exp\left(-h\nabla_i f\left(x^k\right)\right)}{\sum_{l=1}^{n} x_l^k \exp\left(-h\nabla_l f\left(x^k\right)\right)},$$

$$f\left(\bar{x}^N\right) - f_* \le M\sqrt{\frac{2\ln n}{N}} \quad (\bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k).\square$$

Optimal method for non-smooth strongly convex case

Assume that $f(x)$ is additionally μ-strongly convex in $\| \; \|_2$ norm:

$$f(y)+\langle \nabla f(y), x-y\rangle + \frac{\mu}{2}\|x-y\|_2^2 \le f(x) \text{ (for all } x, y \in Q\text{)}.$$

Introduce $$x^{k+1} = \mathrm{Mirr}_{x^k}\left(h_k \nabla f\left(x^k\right)\right),$$

$$h_k = \frac{2}{\mu\cdot(k+1)},\; d(x) = \frac{1}{2}\left\|x-x^0\right\|_2^2,\; \left\|\nabla f(x)\right\|_2 \le M,\; x \in Q.$$

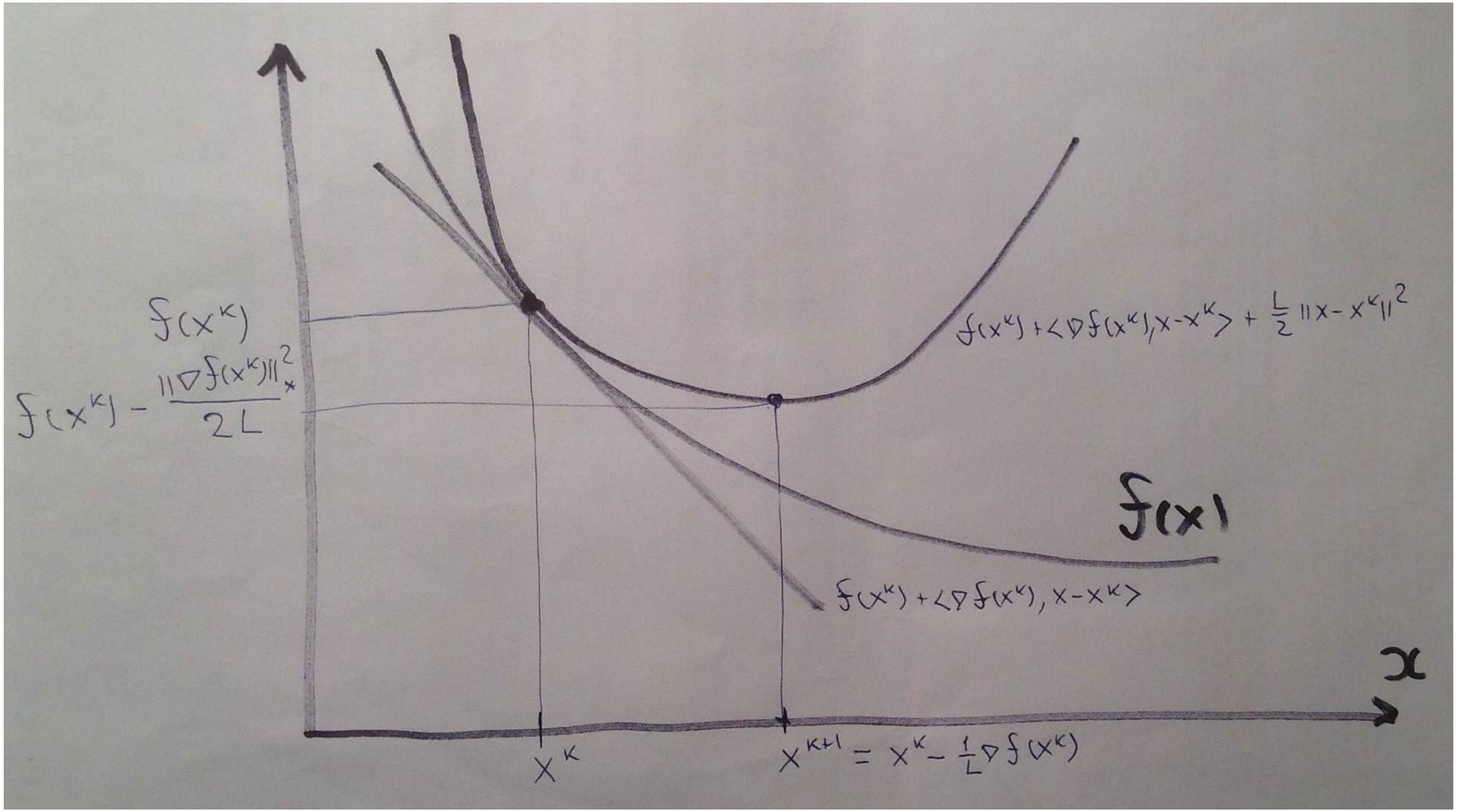
Then (Lacoste-Julien–Schmidt–Bach, 2012)

$$\boxed{f\left(\sum_{k=1}^{N}\frac{2k}{k(k+1)}x^k\right) - f_* \le \frac{2M^2}{\mu\cdot(k+1)}}.$$

Hence $$\boxed{N \simeq \frac{2M^2}{\mu\varepsilon}}.$$

Gradient descent is not optimal method for smooth convex case

$$\left\|\nabla f\left(y\right)-\nabla f\left(x\right)\right\|_{*}\leq L\left\|y-x\right\|$$



$$x^{k+1} = \arg\min_{x \in Q} \left\{ f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), x - x^k \right\rangle + \frac{L}{2}\left\| x - x^k \right\|^2 \right\},$$

$$f\left(x^N\right) - f_* \le \frac{2L\tilde{R}^2}{N},\ \tilde{R}^2 = \max_{x \in Q,\ f(x) \le f(x_0)} \left\| x - x_* \right\|.$$

In Euclidian case (2-norm) one can simplify

$$x^{k+1} = \pi_Q\left(x^k - \frac{1}{L}\nabla f\left(x^k\right) \right),$$

$$f\left(x^N\right) - f_* \le \frac{2LR^2}{N},\ R = \left\| x^0 - x_* \right\|_2.$$

If $Q = \mathbb{R}^n$ one has

$$x^{k+1} = x^k - \frac{1}{L}\nabla f\left(x^k\right).$$

Unfortunately, convergence of simple gradient descent isn’t optimal!

Polyak’s heavy ball method

Gradient descent (Cauchy, 1847):

$$\frac{dx}{dt} = -\nabla f(x); \; x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k).$$

Lyapunov’s functions: $V(x) = f(x) - f_*$, $V(x) = \|x - x_*\|_2^2$ (convex case).

Heavy ball method (Polyak, 1964):

$$\frac{dx}{dt} = y, \; \frac{dy}{dt} = -ay - b\nabla f(x); \; x^{k+1} = x^k - \alpha\nabla f(x^k) + \beta\cdot(x^k - x^{k-1}).$$

Lyapunov’s function: $V(x) = f(x) + \frac{1}{2b}\|y\|_2^2$ – full energy (convex case).

Wilson A., Recht B., Jordan M. arXiv:1611.02635; see also arXiv:1702.06751

Local convergence is optimal. Now we describe global optimal method.

Optimal method for smooth convex case

Estimation functions technique (Yu. Nesterov)

$$d\left(y^0\right)=0,\ d\left(x\right)\geq 0,\ V\left(x,z\right)=d\left(x\right)-d\left(z\right)-\left\langle \nabla d\left(z\right),x-z\right\rangle,$$

$$\varphi_0\left(x\right)=V\left(x,y^0\right)+\alpha_0\left[f\left(y^0\right)+\left\langle \nabla f\left(y^0\right),x-y^0\right\rangle\right],$$

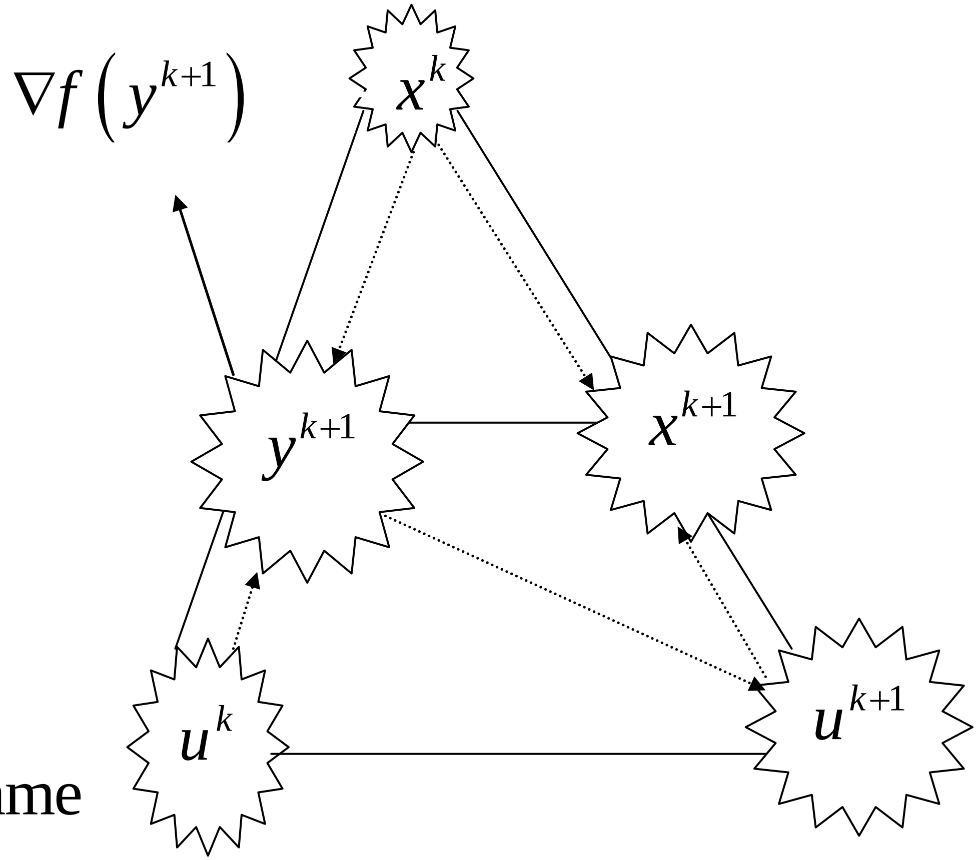
$$\varphi_{k+1}\left(x\right)=\varphi_k\left(x\right)+\alpha_{k+1}\left[f\left(y^{k+1}\right)+\left\langle \nabla f\left(y^{k+1}\right),x-y^{k+1}\right\rangle\right]$$

$$x^0=u^0=\arg\min_{x\in Q}\varphi_0\left(x\right),\ A_k=\sum_{i=0}^{k}\alpha_i,\ \ \alpha_0=L^{-1},\ A_k=\alpha_k^2L,$$

$$\alpha_{k+1}=\frac{1}{2L}+\sqrt{\frac{1}{4L^2}+\alpha_k^2},\ A_k\geq\frac{\left(k+1\right)^2}{4L},\ k=0,1,2,\ldots$$

Similar Triangles Method (Yu. Nesterov; 1983, 2016)

$$\boxed{\begin{aligned} y^{k+1} &= \frac{\alpha_{k+1}u^k + A_k x^k}{A_{k+1}}, \\ u^{k+1} &= \arg\min_{x\in Q} \varphi_{k+1}(x), \\ x^{k+1} &= \frac{\alpha_{k+1}u^{k+1} + A_k x^k}{A_{k+1}}. \end{aligned}}$$



$$u^{k+1} = \mathrm{Mirr}_{y^0}\left(\sum_{i=0}^{k+1}\alpha_i \nabla f\left(y^i\right)\right)$$ the same

$$u^{k+1} = \mathrm{Mirr}_{u^k}\left(\alpha_{k+1}\nabla f\left(y^{k+1}\right)\right)$$ mirror version (Alexander Turin, HSE)

Assume that

$$\left\|\nabla f(y) - \nabla f(x)\right\|_* \le L\left\|y - x\right\| \text{ (for all } x, y \in Q\text{)}.$$

Then

$$\boxed{f\left(x^N\right) - \min_{x\in Q} f(x) \le \frac{4LR^2}{\left(N+1\right)^2}}.$$

That is $\boxed{N \simeq 2\sqrt{LR^2/\varepsilon}}$. And for all $k = 0,1,2,\ldots$

$$\left\|u^k - x_*\right\|^2 \le 2V\left(x_*, y^0\right),$$

$$\max\left\{\left\|x^k - x_*\right\|^2, \left\|y^k - x_*\right\|^2\right\} \le 4V\left(x_*, y^0\right) + 2\left\|x^0 - y^0\right\|^2.$$

Primal-duality: $A_N f\left(x^N\right) \le \min_{x\in Q}\left\{V\left(x, y^0\right) + \sum_{k=0}^{N}\alpha_k\left[f\left(y^k\right) + \left\langle\nabla f\left(y^k\right), x - y^k\right\rangle\right]\right\}.$

Optimal method for smooth strongly convex case

$$\varphi_0(x) = V(x, y^0) + \alpha_0\left[f(y^0) + \langle \nabla f(y^0), x - y^0 \rangle + \frac{\mu}{2}\|x - y^0\|_2^2 \right],$$

$$\varphi_{k+1}(x) = \varphi_k(x) + \alpha_{k+1}\left[f(y^{k+1}) + \langle \nabla f(y^{k+1}), x - y^{k+1} \rangle + \frac{\mu}{2}\|x - y^k\|_2^2 \right],$$

$$A_k = \sum_{i=0}^{k}\alpha_i, \;\; \alpha_0 = L^{-1}, \; A_{k+1}\cdot(1 + A_k\mu) = \alpha_{k+1}^2 L, \; x^0 = u^0 = \arg\min_{x\in Q}\varphi_0(x),$$

$$\alpha_{k+1} = \frac{1 + A_k\mu}{2L} + \sqrt{\frac{1 + A_k\mu}{4L^2} + \frac{A_k\cdot(1 + A_k\mu)}{L}}, \; A_{k+1} = A_k + \alpha_{k+1},$$

$$A_k \geq \frac{1}{L}\left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{2k} \geq \exp\left(\frac{k}{2}\sqrt{\frac{\mu}{L}}\right), \; k = 0,1,2,\ldots$$

Then using Similar Triangles Method with new estimating functions sequence and new step size policy one can obtain (continuous on $\mu \geq 0$)

$$f\left(x^N\right) - \min_{x\in Q} f\left(x\right) \leq \min\left\{\frac{4LR^2}{\left(N+1\right)^2}, LR^2 \exp\left(-\frac{N}{2}\sqrt{\frac{\mu}{L}}\right)\right\}.$$

In other “words”

$$\boxed{N \simeq 2\sqrt{\frac{L}{\mu}}\ln\left(\frac{LR^2}{\varepsilon}\right)}.$$

Unfortunately here and before, in strongly convex case we were significantly restricted by Euclidian norm/prox-structure. Generalization requires another approach: restarts technique (Lecture 5).

For $Q = \mathbb{R}^n$ one can simplify method (Yu. Nesterov; 1983, 2001)

$$x^0 = y^0,$$

$$x^{k+1} = y^k - \frac{1}{L}\nabla f\left(y^k\right),$$

$$y^{k+1} = x^{k+1} + \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\left(x^{k+1} - x^k\right).$$

Unfortunately, this method isn't continuous on $\mu > 0$.

Note: In smooth case from $f\left(x^N\right) - f\left(x_*\right) \le \varepsilon$ one has that

$$\left\|\nabla f\left(x^N\right)\right\|^2 \le 2L\varepsilon \ (\left\|\nabla f\left(x_*\right)\right\|^2 = 0).$$

and in strongly convex case (geometric convergence in argument)

$$\left\|x^N - x_*\right\|_2^2 \le 2\varepsilon/\mu.$$

Open gap problem (A. Nemirovski, 2015)

Assume that $Q = B_1^n(R)$ (ball in $\mathbb{R}^n$ of radius R in 1-norm),

$$\|\nabla f(y) - \nabla f(x)\|_2 \le L_2 \|y - x\|_2 .$$

Then for $N \le n$ and arbitrary method with local first order oracle

$$f(x^N) - f_* \ge \frac{C_1 L_2 R^2}{N^3}.$$

When $\|\nabla f(y) - \nabla f(x)\|_\infty \le L_1 \|y - x\|_1$ Similar Triangles Methods takes us

$$f(x^N) - f_* \le \frac{C_2 L_1 R^2}{N^2},$$

where $L_2/n \le L_1 \le L_2$. Unfortunately, we can't say that there is no gap between lower and upper bounds.

Optimality

Meanwhile, for the most interesting convex sets Q there exists such a norm $\|\ \|$ and appropriate prox-structure $d(x)$ that Mirror Descent and Similar Triangles Method (and theirs restart-strongly convex variants, Lecture 5) lead (up to a logarithmic factor) to unimprovable estimations, collected in the table below (we assume that all parameters M, L, R, μ we choose correspond to the norm $\|\ \|$ – this isn't true for A. Nemirovski example):

N	$\lvert F(y)-F(x)\rvert \le M\|y-x\|$	$\|\nabla F(y)-\nabla F(x)\|_* \le L\|y-x\|$
$F(x)$ convex	$\frac{M^2R^2}{\varepsilon^2}$	$\sqrt{\frac{LR^2}{\varepsilon}}$
$F(x)$ μ-strongly convex	$\frac{M^2}{\mu\varepsilon}$	$\sqrt{\frac{L}{\mu}}\left\lceil \ln\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil$

If norm is non euclidian then the last row is true up to $O(\ln n)$-factor.

How to choose norm and prox function?

Arkadi Nemirovski, 1979

$$a = \frac{2\log n}{2\log n - 1} \simeq 1 + \frac{1}{2\log n}$$

$Q = B_p^n(1)$	$1 \le p \le a$	$a \le p \le 2$	$2 \le p \le \infty$
$\\| \ \\|$	$\\| \ \\|_a$ or $\\| \ \\|_1$	$\\| \ \\|_p$	$\\| \ \\|_2$
$d(x)$	$d(x) = \frac{1}{2(a-1)} \\|x\\|_a^2$	$d(x) = \frac{1}{2(p-1)} \\|x\\|_p^2$	$\frac{1}{2} \\|x\\|_2^2$
R^2	$O(\log n)$	$O\left((p-1)^{-1}\right)$	$O(1)$

Structural optimization (looking into the Black Box)

Composite optimization (Yu. Nesterov, 2007) $\boxed{f(x)+h(x)\to\min_{x\in Q},}$

$$\mathrm{Mirr}_{x^k}\left(\alpha\partial f\left(x^k\right)\right):=\arg\min_{x\in Q}\left\{\left\langle\alpha\partial f\left(x^k\right),x-x^k\right\rangle+\alpha h(x)+V\left(x,x^k\right)\right\},$$

$$\varphi_{k+1}(x):=\varphi_k(x)+\alpha_{k+1}\left[f\left(y^{k+1}\right)+\left\langle\nabla f\left(y^{k+1}\right),x-y^{k+1}\right\rangle+h(x)\right].$$

Rates of convergences of MD and STM don't change and determine only by properties of function $f(x)$ as it was previously (without $h(x)$).

Example (L1 optimization). $h(x)=\lambda\|x\|_1$, $d(x)=\|x\|_2^2/2$, $Q=\mathbb{R}^n$,

$$\left\{\left(\left|x_i^k-\frac{1}{L}\nabla_i f\left(x^k\right)\right|-\frac{\lambda}{L}\right)_+\operatorname{sign}\left(x_i^k-\frac{1}{L}\nabla_i f\left(x^k\right)\right)\right\}_{i=1}^n=$$

$$=\arg\min_{x\in\mathbb{R}^n}\left\{\left\langle\nabla f\left(x^k\right),x-x^k\right\rangle+\lambda\|x\|_1+(L/2)\left\|x-x^k\right\|_2^2\right\}.$$

Structural optimization (looking into the Black Box)

MinMax problem (idea A. Nemirovski, 1979; Yu. Nesterov, 2004)

$$\boxed{F(x)=\max_{l=1,\dots,m} f_l(x)+h(x)\to\min_{x\in Q},}$$

$$u^{k+1}=\arg\min_{u\in Q}\left\{\alpha_{k+1}\left\{\max_{l=1,\dots,m}\left[f_l\left(y^{k+1}\right)+\left\langle\nabla f_l\left(y^{k+1}\right),u-y^{k+1}\right\rangle\right]+h(u)\right\}+V\left(u,u^k\right)\right\}.$$

Unfortunately in general this sub-problem isn't simple enough. But the number of such iteration of Turins' variant of STM will be the same (up to a constant) as in the case of previous slide

$$F\left(x^N\right)-F_*\le\frac{8LR^2}{(N+1)^2},\ \left\|\nabla f_l(y)-\nabla f_l(x)\right\|_*\le L\|y-x\|,\ x,y\in Q,\ l=1,..,m.$$

One can also generalize this result further:

Lan G. Bundle-level methods uniformly optimal for smooth and non-smooth convex optimization // Math. program. Ser. A. 2015. V. 149. no. 1. P. 1–45.

Note that $F(x)$ isn't necessarily smooth even with $h(x) \equiv 0$. So if we can calculate at each iteration only $\partial F\left(x^k\right)$ then one can think that such a methods (that used only this information) can't converges faster then $\mathrm{O}\left(MR/\sqrt{N}\right)$ according to lower bound from the table above. But there is no contradiction with the previous slide since there we have more information $\left\{\nabla f_l\left(x^k\right)\right\}_{l=1}^m$ and we allow ourselves to solve at each iteration non trivial problem (in general). Nevertheless, estimation $\mathrm{O}\left(MR/\sqrt{N}\right)$ is not the right lower bound, for example, when $f_l(x) = \langle c_l, x\rangle$, because the problem has a special structure (functional has a simple Fenchel's type representation). This structure allows to replace the problem by (Nesterov's smoothing, 2005)

$$F_\gamma(x) = \gamma \ln\left(\sum_{l=1}^m \exp\left(\langle c_l, x\rangle/\gamma\right)\right) \to \min_{x\in Q}, \ \gamma = \varepsilon/(2\ln m).$$

If one can find such x^N that

$$F_\gamma\left(x^N\right) - F_\gamma^* \le \varepsilon/2$$

then for the same x^N one will have (Lecture 5)

$$F\left(x^N\right) - F_* \le \varepsilon .$$

The above is obvious from the dual regularized representation

$$F(x) = \max_{y \in S_m(1)} \sum_{l=1}^{m} y_l \langle c_l, x \rangle;\ F_\gamma(x) = \max_{y \in S_m(1)} \left\{ \sum_{l=1}^{m} y_l \langle c_l, x \rangle - \gamma \sum_{l=1}^{n} y_l \ln y_l \right\}.$$

Since $\sum_{l=1}^{n} y_l \ln y_l$ is 1-strongly convex in 1-norm then

$$\left\| \nabla F_\gamma(y) - \nabla F_\gamma(x) \right\|_2 \le L_{F_\gamma} \left\| y - x \right\|_2,\ \ L_{F_\gamma} = \frac{1}{\gamma} \max_{l=1,\dots,m} \left\| c_l \right\|_2^2 .$$

So we have $N = \mathrm{O}\left(\max_{l=1,\dots,m} \left\| c_l \right\|_2 R_2 \sqrt{\ln m} \Big/ \varepsilon \right)$ instead of $N = \mathrm{O}\left(\max_{l=1,\dots,m} \left\| c_l \right\|_2^2 R_2^2 \Big/ \varepsilon^2 \right)$.

Conditional problems

In **smooth case** the main trick is to reduce

$$f_0(x) \to \min_{f_l(x)\le 0,\ l=1,\dots,m;\ x\in Q}$$

to the searching of

$$F(t) = \min_{x\in Q} \max\left\{ f_0(x) - t,\ f_1(x), \dots,\ f_m(x) \right\}.$$

The last problem (with fixed t is considered above). Our task is to find the minimal t_* such that $F(t_*) = 0$. Since $F(t)$ convex and decrease one can do it with precision ε using $\sim \log\left(\varepsilon^{-1}\right)$ recalculations of $F(t)$.

Nesterov Yu. Introduction Lectures on Convex Optimization. A Basic Course. Applied Optimization. – Springer, 2004.

In **non smooth** case

$$f(x) \to \min_{g(x) \le 0;\; x \in Q},$$

where $\left\|\partial f(x)\right\|_* \le M_f$, $\left\|\partial g(x)\right\|_* \le M_g$. Let's $h_g = \varepsilon_g / M_g^2$, $h_f = \varepsilon_g / \left(M_f M_g\right)$,

$$\boxed{\begin{aligned} x^{k+1} &= \mathrm{Mirr}_{x^k}\left(h_f \partial f\left(x^k\right)\right), \;\; \text{if } g\left(x^k\right) \le \varepsilon_g, \\ x^{k+1} &= \mathrm{Mirr}_{x^k}\left(h_g \partial g\left(x^k\right)\right), \;\; \text{if } g\left(x^k\right) > \varepsilon_g, \end{aligned}} \quad k = 1, \dots, N,$$

and the set I of such indexes k, that $g\left(x^k\right) \le \varepsilon_g$. Then for $\boxed{N \simeq 2M_g^2 R^2 / \varepsilon_g^2}$

$$f\left(\bar{x}^N\right) - f_* \le \varepsilon_f = M_f \varepsilon_g / M_g, \; g\left(\bar{x}^N\right) \le \varepsilon_g,$$

where $\bar{x}^N = \frac{1}{N_I} \sum_{k \in I} x^k$, $N_I = |I|$, $N_I \ge 1$.

High-order methods

In 1989 Nemirovski–Nesterov propose a general (Newton's type) method to solve large class of convex optimization problems of these type

$$\langle c, x\rangle \to \min_{x\in Q}, \text{ // Note: } F(x) \to \min_{x} \ \sim \ y \to \min_{F(x)\le y},$$

Idea (inner penalty): $t\langle c, x\rangle + F(x) \to \min_{x}$, $t \to \infty$. // central path

Convex (but rather complex for projections) set Q imposed by ν-self-concordant barrier function $F_Q(x)$. As we have already known (see Lecture 1) many interesting convex problems have such representation with $Q = \mathbb{R}^n_+$, $Q = S^n_+$ (up to affine transformation). For this sets

$$F_Q(x) = -\sum_{i=1}^{n} x_i \ln x_i \text{ and } F_Q(X) = -\ln\det(X)$$

are corresponding n-self-concordant barrier (and in general $\nu = O(n)$).

Interior Point Method (inserted in CVX)

The proposed method looks as follows

$$t^{k+1} = \left(1 + \frac{1}{13\sqrt{\nu}}\right) t^k,\ x^{k+1} = x^k - \left[\nabla^2 F_Q\left(x^k\right)\right]^{-1}\left(t^{k+1} \cdot c + \nabla F_Q\left(x^k\right)\right).$$

With proper choice of starting point (these procedure costs $\mathrm{O}\left(\sqrt{\nu}\log\nu\right)$) described IPM has the following rate of convergence $\boxed{N = \mathrm{O}\left(\sqrt{\nu}\log\left(\nu/\varepsilon\right)\right)}$. This estimation is better (since $\nu = \mathrm{O}(n)$) than lower bound $\sim n\log\left(\varepsilon^{-1}\right)$ (we consider here the case $N \geq n$). There is no contradiction here, because of additional assumption about the structure of the problem. This estimation is accurate, but in real live IPM is much faster (30 iteration is typically enough). IPM works much better for $n \geq 10^2$ then ellipsoid's type methods.

Can one obtain something better?

The question is natural since local convergence of Newton method is $\sim \log\log\left(\varepsilon^{-1}\right)$. As it was shown by A. Nemirovski (1979) this rate of convergence could be in principal be realized globally. But the price should be to high – rather complex iterations. Even in IPM realization we have in principle the following complexity of one iteration $\mathrm{O}\left(n^3\right)$ (this can be reduced for the special cases). Moreover, it was also shown that even in $\mathbb{R}^1$ for the function $f\left(x\right)$, with $1 \le f''\left(x\right) \le 2$, $\left|f^{(k)}\left(x\right)\right| \le 1$, $k = 1, 3, \ldots, m$, $m \gg 1$ the lower bound will be $\sim c_m \log\log\left(\varepsilon^{-1}\right)$ (here we can asked oracle as much derivatives $k \le m$ as we want). Local convergence can be faster (Chebyshev's type methods http://www.ccas.ru/personal/evtush/p/198.pdf)!

IPM is a powerful tool that finds applications to real large scale convex problems ($n \le 10^4$). Especially for Semi Definite Programming (see CVX).

Semi Definite Relaxation (MAX CUT)

$$f\left(x\right)=\frac{1}{2}\sum_{i,j=1,1}^{n,n} A_{ij}\left(x_i-x_j\right)^2 \to \max_{x\in\{-1,1\}^n},$$

where $A=\left\|A_{ij}\right\|_{i,j=1,1}^{n,n}$ $(A=A^T)$.

Let’s introduce

$$L=\operatorname{diag}\left\{\sum_{j=1}^{n} A_{ij}\right\}_{i=1}^{n}-A,$$

ς – random vector, uniformly distributed on a Hamming cube $\{-1,1\}^n$.

Note, that

$$f\left(x\right)=\left\langle x, Lx\right\rangle.$$

Simple observation:

$$E\left\langle \varsigma, L\varsigma \right\rangle \geq 0.5 \max_{x\in\{-1,1\}^n} \left\langle x, Lx \right\rangle.$$

Could we do better?

$$\max_{x\in\{-1,1\}^n} \left\langle x, Lx \right\rangle = \max_{x\in\{-1,1\}^n} \left\langle L, xx^T \right\rangle \leq \max_{\substack{X\in S_+^n \\ X_{ii}=1,\ i=1,\dots,n}} \left\langle L, X \right\rangle \text{ // SDP problem!}$$

The book of Goemans–Williamson, 1995

Let Σ be the solution of SDP problem. Let

$$\xi \in N\left(0,\Sigma\right),\ \varsigma = \operatorname{sign}\left(\xi\right).$$

Then (the constant is unimprovable if $P \neq NP$ – Unique Games Conjecture)

$$E\left\langle \varsigma, L\varsigma \right\rangle \geq 0.878 \max_{x\in\{-1,1\}^n} \left\langle x, Lx \right\rangle.$$

“Optimal” methods aren’t always optimal indeed

Due to Lecture 2 we can reduce Google problem to

$$f\left(x\right)=\frac{1}{2}\left\|Ax\right\|_2^2\to\min_{x\in S_n\left(1\right)},$$

We will use not optimal (in terms the number of oracle calls) conditional gradient method (Frank–Wolfe, 1956). But we assume that the number of nonzero elements at each row and each column smaller then $s\ll\sqrt{n}$.

We choose starting point at one of the simplex vertex x^1. Induction step

$$\left\langle\nabla f\left(x^k\right),y\right\rangle\to\min_{y\in S_n\left(1\right)}\quad.$$

Let’s denote the solution of this problem by

$$y^k = \left(0,\ldots,0,1,0,\ldots,0\right),$$

where 1 is posed at the position

$$i_k = \arg\min_{i=1,\ldots,n} \partial f\left(x_k\right) / \partial x^i \; .$$

The main algorithm looks as follows

$$x^{k+1} = \left(1-\gamma_k\right)x^k + \gamma_k y^k,\; \gamma_k = \frac{2}{k+1},\; k = 1,2,\ldots,$$

One can obtain that (here we also used that $f_* = 0$)

$$f\left(x^N\right) = f\left(x^N\right) - f_* \le \frac{2L_p R_p^2}{N+1}, \text{ // for optimal method } \mathrm{O}\left(\frac{LR^2}{N^2}\right)$$

$$R_p^2 = \max_{x,y\in S_n(1)} \|y-x\|_p^2,\ L_p = \max_{\|h\|_p\le 1} \langle h, A^T A h\rangle = \max_{\|h\|_p\le 1} \|Ah\|_2^2,\ 1\le p\le\infty.$$

Since we work on a simplex we choose $p=1$. As a result

$$R_1^2 = 4,\ L_1 = \max_{i=1,\dots,n} \left\|A^{\langle i\rangle}\right\|_2^2 \le 2.$$

Hence for $f\left(x^N\right)\le \varepsilon^2/2$ ($\left\|Ax^N\right\|_2 \le \varepsilon$) we have to do $N = 32\varepsilon^{-2}$ iterations ($N\le n \Rightarrow \varepsilon \ge n^{-1/2}$, but since $\left\|\left(n^{-1},\dots,n^{-1}\right)\right\|_2 = n^{-1/2}$, here we are interested in $n^{-1} \ll \varepsilon \ll n^{-1/2}$). One can show that after $\mathrm{O}(n)$ preprocessing each iteration will costs $\mathrm{O}\left(s^2\ln\left(n/s^2\right)\right)$. So the total complexity will be

$$\mathrm{O}\left(n + s^2\ln\left(n/s^2\right)/\varepsilon^2\right),$$

instead of total complexity of “optimal” method STM $\mathrm{O}\left(sn/\varepsilon\right)$.

To be continued...