Convex Optimization for Data Science

Gasnikov Alexander

gasnikov.av@mipt.ru

Lecture 4. Stochastic optimization. Randomized methods

November, 2016

Main books:

Polyak B.T., Juditsky A.B. Acceleration of stochastic approximation by averaging // SIAM J. Control Optim. – 1992. – V. 30. – P. 838–855.
Sridharan K. Learning from an optimization viewpoint. PhD Thesis, 2011.
Juditsky A., Nemirovski A. First order methods for nonsmooth convex large-scale optimization, I, II. // Optimization for Machine Learning. // Eds. S. Sra, S. Nowozin, S. Wright. – MIT Press, 2012.
Shapiro A., Dentcheva D., Ruszczynski A. Lecture on stochastic programming. Modeling and theory. – MPS-SIAM series on Optimization, 2014.
Guiges V., Juditsky A., Nemirovski A. Non-asymptotic confidence bounds for the optimal value of a stochastic program // e-print, 2016 arXiv:1601.07592
Duchi J.C. http://stanford.edu/~jduchi/PCMIConvex/Duchi16.pdf
Gasnikov A.V. Searching equilibriums in large transport networks. Doctoral Thesis. MIPT, 2016. https://arxiv.org/ftp/arxiv/papers/1607/1607.03142.pdf
https://www.youtube.com/user/PreMoLab (see course of A.V. Gasnikov)

Structure of Lecture 4

- Auxiliary facts (Azuma–Hoeffding's inequality; Heavy-tails, large deviations; Le Cam lower bound)
- Stochastic Mirror Descent
- Rate of convergence
- Lower bounds
- Nesterov's problem about Mage and Experts (Parallelization)
- Conditional Stochastic optimization
- SAA *vs* SA
- Acceleration of Stochastic Approximation by proper Averaging
- Randomized MD for huge QP
- Randomized MD for Antagonistic matrix game

Auxiliary facts

Azuma–Hoeffding's inequality: Let $\{\chi_t\}_t$ – a scalar random sequence is martingale-difference

$$\chi_t = Y_t - Y_{t-1},\ E\left[Y_t \middle| F_{\sigma-\text{algebra}}\left(Y_1,\dots,Y_{t-1}\right)\right] = Y_{t-1},$$

such that

$$E\left[\exp\left(\chi_t^2/M^2\right)\middle|\chi_1,\dots,\chi_{t-1}\right] \le \exp(1) \text{ for all } t = 1,2,\dots,N.$$

Then ($s > 0$)

$$P\left(\sum_{t=1}^{N}\gamma_t\chi_t \ge sM\sqrt{\sum_{t=1}^{N}\gamma_t^2}\right) \le \exp\left(-s^2/3\right),$$

$$P\left(\sum_{t=1}^{N}\gamma_t\chi_t^2 \ge M^2\sum_{t=1}^{N}\gamma_t + M^2\max\left\{\sqrt{6.6s\sum_{t=1}^{N}\gamma_t^2}, 6.6s\frac{1}{N}\sum_{t=1}^{N}\gamma_t\right\}\right) \le \exp(-s).$$

Heavy-tails, large deviations: Let scalar random sequence $\{\chi_t\}_t$ – i.i.d.,

$E[\chi_t]=0$, $\mathrm{Var}[\chi_t]=D$, $P(\chi_t > s)=V(s)=\mathrm{O}(s^{-\alpha})$, $\alpha>2$.

Then $\boxed{P\left(\sum_{t=1}^{N}\chi_t \ge s\right) \underset{N\gg 1}{\simeq} 1-\Phi\left(\frac{s}{\sqrt{DN}}\right)+N\cdot V(s)}$, $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}dy$,

$$P\left(\sum_{t=1}^{N}\chi_t \ge s\right) \underset{N\gg 1}{\simeq} 1-\Phi\left(\frac{s}{\sqrt{DN}}\right),\ s\le\sqrt{(\alpha-2)DN\ln N}, \text{ (CLT regime)}$$

$$P\left(\sum_{t=1}^{N}\chi_t \ge s\right) \underset{N\gg 1}{\simeq} N\cdot V(s),\ s>\sqrt{(\alpha-2)DN\ln N}. \text{ (heavy-tails regime)}$$

Note: $0.2e^{-2x^2/\pi} \le 1-\Phi(x) \le e^{-x^2/2},\ x\gg 1.$

These estimations can be generalized for the weighted sums of scalar martingale-differences and weighted sums of squares of martingale-differences.

Two coins comparison: Consider two coins: $p = 0.5$ and $p = 0.5 + \varepsilon$. How many observations $y = \left(y^1, y^2, ..., y^N\right)$ we have to do to decide with probability $\geq 1 - \sigma$ what is a best coin? Let's introduce some decision rule $\varphi\left(y\right)$ that takes values $\left[0,1\right]$ (we interpret $\varphi\left(y\right)$ as a probability to decide in favor of the second coin if we observe y). Then the probability of right decision is

$$\left| E\left[\varphi\left(y\right)\middle| p = 0.5 + \varepsilon\right] - E\left[\varphi\left(y\right)\middle| p = 0.5\right]\right| \geq 2 - 2\sigma.$$

Since for all measurable $0 \leq \varphi\left(y\right) \leq 1$ (Pinsker's inequality + chain rule)

$$\left| E_{P^N}\left[\varphi\left(y\right)\right] - E_{Q^N}\left[\varphi\left(y\right)\right]\right| \leq \left\| P^N - Q^N \right\|_1^2 \leq 2KL\left(P^N, Q^N\right) = 2N \cdot KL\left(P, Q\right),$$

$$KL\left(P, Q\right) = \left(0.5 + \varepsilon\right)\ln\left(\left(0.5 + \varepsilon\right)/0.5\right) + \left(0.5 - \varepsilon\right)\ln\left(\left(0.5 - \varepsilon\right)/0.5\right) \simeq 4\varepsilon^2,$$

we have that $N \geq C\varepsilon^{-2}$. One can show that indeed: $\boxed{N \geq C\ln\left(\sigma^{-1}\right)\varepsilon^{-2}}$.

Another way to use Rao–Cramer's inequality for Bernoulli scheme (Lect. 2).

Stochastic Mirror Descent

Consider convex optimization problem (see Lecture 3)

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

with stochastic oracle, returns such stochastic subgradient $\partial_x f(x,\xi)$ that:

$$E_\xi\left[\partial_x f(x,\xi)\right] = \partial f(x),\ E_\xi\left[\left\|\partial_x f(x,\xi)\right\|_*^2\right] \le M^2.$$

Method (the main tools for numerical stochastic programming!)

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

The main property of MD-step ($\left\{\xi^k\right\}$ – i.i.d.)

$$2V\left(x,x^{k+1}\right) \le 2V\left(x,x^k\right) + 2h\left\langle \partial_x f\left(x^k,\xi^k\right), x - x^k\right\rangle + h^2\left\|\partial_x f\left(x^k,\xi^k\right)\right\|_*^2.$$

$$f\left(x^k\right)-f\left(x\right)\le\left\langle \partial f\left(x^k\right),x^k-x\right\rangle\le\left\langle \partial f\left(x^k\right)-\partial_x f\left(x^k,\xi^k\right),x^k-x\right\rangle+$$

$$+\frac{1}{h}\left(V\left(x,x^k\right)-V\left(x,x^{k+1}\right)\right)+\frac{h}{2}\left\|\partial_x f\left(x^k,\xi^k\right)\right\|_*^2\Big| \quad E\left[\,\cdot\,|\xi^1,...,\xi^{k-1}\right],$$

$$f\left(x^k\right)-f\left(x\right)\le\left\langle \partial f\left(x^k\right),x^k-x\right\rangle\le$$

$$\le\frac{1}{h}\left(V\left(x,x^k\right)-E\left[V\left(x,x^{k+1}\right)\Big|\xi^1,...,\xi^{k-1}\right]\right)+\frac{h}{2}\underbrace{E\left[\left\|\partial_x f\left(x^k,\xi^k\right)\right\|_*^2\Big|\xi^1,...,\xi^{k-1}\right]}_{\le M^2}.$$

If we sum all these inequalities from $k=0,...,N-1$ and take the total mathematical expectation from the both sides of the result with $x=x_*$, then due to the convexity of $f\left(x\right)$ we obtain (as in deterministic case)

$$E\left[f\left(\bar{x}^N\right)\right]-f_*\le\left(hN\right)^{-1}V\left(x_*,x^0\right)+M^2h/2\le\sqrt{2M^2R^2/N},$$

where

$$R^2 = V\left(x_*, x^0\right),\ \bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k,\ h = \frac{R}{M}\sqrt{\frac{2}{N}} = \frac{\varepsilon}{M^2}.$$

In other words, after $\boxed{N = 2M^2R^2/\varepsilon^2}$ oracle calls $\boxed{E\left[f\left(\bar{x}^N\right)\right] - f_* \le \varepsilon}$.

Absolutely the same result (even constants) as it was in deterministic case!

If one will use adaptive stepsize policy

$$x^{k+1} = \mathrm{Mirr}_{x^k}\left(h_k \partial_x f\left(x^k, \xi^k\right)\right),\ h_k = \frac{R}{\sqrt{\sum_{i=0}^{k}\left\|\partial_x f\left(x^i, \xi^i\right)\right\|_*^2}},\ R = \max_{x \in Q} V\left(x_*, x\right),$$

Then after $N = 9M^2R^2/\varepsilon^2$ oracle calls $E\left[f\left(\bar{x}^N\right)\right] - f_* \le \varepsilon$.

In deterministic case one can take $h_k = \varepsilon \Big/ \left\|\partial_x f\left(x^k\right)\right\|_*^2$.

From the convergence in average due to the Markov's inequality

$$P\left(f\left(\bar{x}^N\right) - f_* \geq 2\varepsilon \right) \leq \frac{E\left[f\left(\bar{x}^N\right)\right] - f_*}{2\varepsilon} \leq \frac{1}{2}.$$

So we can run in parallel $\sim \log_2\left(\sigma^{-1}\right)$ MD-trajectories. Let's denote by $\bar{x}_{\min}^N$ such $\bar{x}^N$ from these trajectories that minimize $f\left(\bar{x}^N\right)$. Here we assume that we have an oracle for the value of function $f\left(x\right)$.

So after (see formulas in frame on the previous slide)

$$N = \frac{8M^2R^2}{\varepsilon^2}\log_2\left(\sigma^{-1}\right)$$

oracle calls one can obtain

$$P\left(f\left(\bar{x}_{\min}^N\right) - f_* \geq 2\varepsilon \right) \leq \sigma .$$

But what we should do if there is no oracle for the value of the function?

Assume that $\left\|\partial_x f\left(x,\xi\right)\right\|_* \le M$ a.s. for ξ, then

$$P\left(f\left(\overline{x}^N\right) - f_* \le M\sqrt{\frac{2}{N}}\left(R + 2\tilde{R}\sqrt{\ln\left(2/\sigma\right)}\right)\right) \ge 1-\sigma,$$

where $\tilde{R} = \sup_{x\in\tilde{Q}}\left\|x - x_*\right\|$, $\tilde{Q} = \left\{x \in Q:\ \left\|x - x_*\right\|^2 \le 65R^2\ln\left(4N/\sigma\right)\right\}$.

More generally, one can show (using Azuma–Hoeffding's inequality) that

- if $\left\|\partial_x f\left(x,\xi\right)\right\|_* \le M$, then

$$\boxed{N \sim \frac{M^2R^2\ln\left(\sigma^{-1}\right)}{\varepsilon^2}};$$

• if $E\left(\exp\left(\left\|\partial_x f(x,\xi)\right\|_*^2 \big/ M^2\right)\right) \le \exp(1)$ and $\varepsilon \le MR$ then

$$\boxed{N \sim \frac{M^2 R^2 \ln\left(\sigma^{-1}\right)}{\varepsilon^2}}.$$

Using heavy-tails large deviations estimations one can obtain

• if $P\left(\left\|\partial_x f(x,\xi)\right\|_*^2 \big/ M^2 \ge s\right) = O\left(s^{-\alpha}\right)$, $\alpha > 2$ then

$$\boxed{N \sim M^2 R^2 \max\left\{\frac{\ln\left(\sigma^{-1}\right)}{\varepsilon^2}, \left(\frac{1}{\sigma\varepsilon^{\alpha}}\right)^{\frac{2}{3\alpha-2}}\right\}}.$$

All these bounds are optimal up to a multiplicative constants.

Using the restarts technique (see Lecture 5) one can generalize all the results mentioned above to μ**-strongly convex functions** in norm $\| \;\|$. In all the estimations we leave non-euclidian prox-factor $\omega_n = \mathrm{O}\left(\ln^{\beta} n\right)$ ($Q \subseteq \mathbb{R}^n$).

- if $\left\|\partial_x f\left(x,\xi\right)\right\|_* \le M$, then

$$N \sim \frac{M^2 \ln\left(\left(\ln N\right)/\sigma\right)}{\mu\varepsilon};$$

- if $E\left(\frac{\left\|\partial_x f\left(x,\xi\right)\right\|_*^2}{M^2}\right) \le \exp\left(1\right)$ and $\varepsilon \le MR$ then

$$N \sim \frac{M^2 \ln\left(\left(\ln N\right)/\sigma\right)}{\mu\varepsilon};$$

- if $P\left(\left\|\partial_x f\left(x,\xi\right)\right\|_*^2 \big/ M^2 \ge s\right) = \mathrm{O}\left(s^{-\alpha}\right)$, $\alpha > 2$ then

$$N \sim \max\left\{\frac{M^2 \ln\left(\left(\ln N\right)/\sigma\right)}{\mu\varepsilon}, \left(\frac{M^2}{\mu\varepsilon}\right)^{\frac{\alpha}{3\alpha-2}}\left(\frac{\ln N}{\sigma}\right)^{\frac{2}{3\alpha-2}}\right\}.$$

All these bounds are optimal up to a $\ln N$-factor of σ. We don't know at the moment is it possible to eliminate this factor and the ω_n-factor.

Juditsky A., Nesterov Yu. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization // Stoch. System. – 2014. – V. 4. – no. 1. – P. 44–80.

Is Markov's inequality always rough?

Consider sum-type convex optimization problem

$$f(x) = \frac{1}{m}\sum_{k=1}^{m} f_k(x) + h(x) \to \min_{x\in Q},$$

where $\left\|\nabla f_k(y) - \nabla f_k(x)\right\|_2 \le L\left\|y - x\right\|_2$ and $h(x)$ is μ-strongly convex in $\|\ \|_2$. As we've seen later (Lecture 6) one can obtain $E\left[f\left(x^{N(\varepsilon)}\right)\right] - f_* \le \varepsilon$ after $N(\varepsilon) \sim \left(m + \min\left\{L/\mu, \sqrt{mL/\mu}\right\}\right)\ln\left(\Delta f/\varepsilon\right)$ iterations (calculations of $\nabla f_k(x)$ solely). Using rough Markov's inequality

$$P\left(f\left(x^{N(\varepsilon\sigma)}\right) - f_* \ge \varepsilon\sigma/\sigma\right) \le \frac{E\left[f\left(x^{N(\varepsilon\sigma)}\right)\right] - f_*}{\varepsilon\sigma/\sigma} \le \sigma,$$

one can obtain unimprovable large deviations bound $\sim \ln\left(\sigma^{-1}\right)$.

Simple lower bounds

Consider **non strongly convex case**

$$\varepsilon x \to \min_{x \in [-1,1]}.$$

Assume that the oracle return $\nabla f(x,\xi) = \varepsilon + \xi,\ \xi \in N(0,1)$. At each call ξ chooses independently. Assume we know in advance all the details except of ε sign – but we can observe $y^k = \varepsilon + \xi^k$. So we know in advanced that we should choose $x = \pm 1$. How many oracle's calls we need to determine with probability $\geq 1-\sigma$ the right sign? Due to Neyman–Pirson's lemma the best strategy is $\hat{x}_N = -\text{sign} \sum_{k=1}^{N} y^k$. $P\left(\hat{x}_N = 1 \middle| \varepsilon > 0\right) = P\left(\sum_{k=1}^{N} y^k < 0\right) \simeq Ce^{-\varepsilon^2 N}$, when $\varepsilon > 0$, we have the following lower bound $\boxed{N \geq C \ln\left(\sigma^{-1}\right) / \varepsilon^2}$.

Consider **strongly convex case**. Probabilistic model:

$$y^k = x + \xi^k,\ \xi^k \in N(0,1) \text{ // loglikelihood: } -(y-x)^2 \big/ 2;$$

$$x_* = \arg\min_x (x - x_*)^2 \big/ 2 = \arg\min_x E\left[(y-x)^2 \big/ 2 \right], y \in N(x_*, 1). \qquad (*)$$

One can consider (*) to be the stochastic programming problem with the oracle returns stochastic gradients $y^k - x$, $y^k \in N(x_*, 1)$. Due to Rao–Cramer's inequality (Lecture 2) we have $E\left[\left(\hat{x}_N \left(y^1, ..., y^N \right) - x_* \right)^2 \right] \geq N^{-1}$. Since normal distribution (with mathematical expectation as parameter) belongs to Exponential family, for MLE $\hat{x}_N = \arg\min_x \frac{1}{2} \sum_{k=1}^{N} \left(y^k - x \right)^2 = \frac{1}{N} \sum_{k=1}^{N} y^k$ we have equality in Rao–Cramer's inequality. Since that we have a precise lower bound for that case $\boxed{N \simeq C \ln\left(\sigma^{-1}\right) \big/ \varepsilon}$. The other example – Bernoulli scheme (here one can also use lower bound for two coins comparison).

General lower bounds (A. Nemirovski)

Consider convex optimization problem

$$f(x) \to \min_{x \in B_p^n(R)}$$

with stochastic oracle, return such $\partial f(x,\xi)$ that:

$$E_\xi\left[\partial f(x,\xi)\right] = \partial f(x),\ E_\xi\left[\left\|\partial f(x,\xi)\right\|_q^2\right] \le M_p^2\ (1/p + 1/q = 1).$$

We'd like to obtain lower bound for the oracle calls N, that guarantee x^N

$$E\left[f\left(x^N\right)\right] - f_* \le \varepsilon.$$

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

Lower bounds for the **Stochastic Oracle** are (MD achieves these bounds)

- $N \geq c_p M_p^2 R^2 \big/ \varepsilon^{\max(2,p)}$, under $N \ll n$, where $c_p = \mathrm{O}(\ln n)$ (this estimation of c_p become precise when $p \to 1+0$);
- $N \geq c_p M_p^2 R^2 n^{1-2/\max(2,p)} \big/ \varepsilon^2$, under $N \gg n$.

For the **Deterministic Oracle** (when oracle returns subgradient $\partial f(x)$ with the property $\|\partial f(x)\|_p \leq M_p$) we have lower bound

- $N \geq cn \ln(M_p R/\varepsilon)$, under $N \gg n$. // differs only in this regime

Agarwal A., Bartlett P.L., Ravikumar P., Wainwright M.J. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization // IEEE Trans. of Inform. – 2012. – V. 58. – № 5. – P. 3235–3249.

Nesterov's problem about Mage and Experts (Parallelization)

Assume that the optimal configuration determines by convex problem

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

But each day one can only observe independent stochastic subgradients

$$\partial_x f(x,\xi):\ E_\xi\left[\partial_x f(x,\xi)\right] = \partial f(x),\ \left\|\partial_x f(x,\xi)\right\|_* \le M.$$

Mage can live $N \sim M^2R^2\ln\left(\sigma^{-1}\right)/\varepsilon^2$ iterations and Expert $N \sim M^2R^2/\varepsilon^2$.

What is better to ask a solution from Mage or from $K \sim \ln\left(\sigma^{-1}\right)$ Experts?

Answer (arXiv:1701.01830): In both of the cases we obtain (up to constant factors) the same (ε,σ)-quality.

Indeed, as we've already known clever Mage (this Mage know MD algorithm) can give us (ε,σ)-solutions. That is return such a point that

$$P\left(f\left(\overline{x}^{N}\right) - f_* \le \varepsilon \right) \ge 1 - \sigma .$$

On the other hand clever Expert returns such $\overline{x}^{N,i}$ that $E\left[f\left(\overline{x}^{N,i}\right)\right] - f_* \le \varepsilon$.

Therefore without loss of generality one can assume that (see above)

$$f\left(\overline{x}^{N,i}\right) - f_* \in N\left(\varepsilon,\varepsilon^2\right).$$

Since we assume Experts to be independent and $f(x)$ is convex

$$f\left(\overline{x}^{K}\right) - f_* \le \frac{1}{K}\sum_{i=1}^{K}\left(f\left(\overline{x}^{N,i}\right) - f_*\right) \in N\left(\varepsilon,\frac{\varepsilon^2}{K}\right), \quad \overline{x}^{K} = \frac{1}{K}\sum_{i=1}^{K}\overline{x}^{N,i}$$

Hence, $P\left(f\left(\overline{x}^{K}\right) - f_* \le \varepsilon \right) \ge 1 - \exp(-K) \simeq 1 - \sigma$.

It'd be interesting to generalize this result for the other cases (see above).

Conditional Stochastic optimization

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

where

$$E_\xi\left[\partial_x f(x,\xi)\right] = \partial f(x),\; E_\xi\left[\partial_x g(x,\xi)\right] = \partial g(x),$$

$$E_\xi\left[\left\|\partial_x f(x,\xi)\right\|_*^2\right] \le M_f^2,\; E_\xi\left[\left\|\partial_x g(x,\xi)\right\|_*^2\right] \le M_g^2.$$

Let’s

$$h_g = \varepsilon_g / M_g^2,\; h_f = \varepsilon_g / \left(M_f M_g\right),$$

$$\boxed{\begin{array}{l} x^{k+1} = \mathrm{Mirr}_{x^k}\left(h_f \partial_x f\left(x^k,\xi^k\right)\right), \text{ if } g\left(x^k\right) \le \varepsilon_g, \\ x^{k+1} = \mathrm{Mirr}_{x^k}\left(h_g \partial_x g\left(x^k,\xi^k\right)\right), \text{ if } g\left(x^k\right) > \varepsilon_g, \end{array}} \quad k = 1,\ldots,N,$$

and the set I ($N_I = |I|$) of such indexes k, that $g\left(x^k\right) \le \varepsilon_g$.

Then if $\boxed{N \ge 2M_g^2 R^2 / \varepsilon_g^2}$ then $N_I \ge 1$ with probability $\ge 1/2$ and

$$E\left[f\left(\bar{x}^N\right)\right] - f_* \le \varepsilon_f = \frac{M_f}{M_g}\varepsilon_g,\ g\left(\bar{x}^N\right) \le \varepsilon_g\ ,\ \bar{x}^N = \frac{1}{N_I}\sum_{k\in I} x^k .$$

If additionally $\left\|\partial_x f\left(x,\xi\right)\right\|_* \le M_f$, $\left\|\partial_x g\left(x,\xi\right)\right\|_* \le M_g$, then for all

$$\boxed{N \ge \frac{81 M_g^2 \tilde{R}^2}{\varepsilon_g^2}\ln\left(\sigma^{-1}\right)}$$

up to a constant factor and $R \to \tilde{R}$ the same as it was in unconditional case (see above)

with probability $\ge 1-\sigma$ it's true $N_I \ge 1$ and

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

where $\tilde{R}^2 = \sup_{x,y\in Q} V\left(x,y\right)$.

A. Bayandina generalizes it to strongly convex case, using restarts technique.

Here we have still an open problem: to generalize on composite optimization.

SAA *vs* SA (Nemirovski–Juditsky–Lan–Shapiro, 2007)

Stochastic Average Approximation (Empirical Risk Minimization, Monte Carlo) approach proposes to change Stochastic convex optimization problem

$$E_{\xi}\left[f\left(x,\xi\right)\right]\to\min_{x\in Q}$$

by **non stochastic** sum-type **SAA-problem** ($\left\{\xi^k\right\}_{k=1}^m$ – i.i.d. realizations from ξ)

$$\frac{1}{m}\sum_{k=1}^{m}f\left(x,\xi^k\right)\to\min_{x\in Q}.$$

Unfortunately, for the absolutely accurate solution of SAA-problem to be $\left(\varepsilon,\sigma\right)$-solution of initial one, one should take at least ($\left\|\partial_x f\left(x,\xi\right)\right\|_* \le M$)

$$m \ge C\cdot M^2R^2\left(n\ln\left(MR/\varepsilon\right)+\ln\left(\sigma^{-1}\right)\right)\Big/\varepsilon^2 \text{ terms.}$$

Stochastic Approximation approach (Robbins–Monro, 1951) in our sense is nothing more than Mirror Descent. So we can find (ε,σ)-solution of initial stochastic programming problem for

$$N \sim M^2 R^2 \ln\left(\sigma^{-1}\right)/\varepsilon^2 \ll m \text{ // SA is better SAA}$$

oracle calls (i.e. calculations of stochastic subgradients $\partial_x f\left(x,\xi\right)$). It seems too strange (n-factor in m can be eliminated via regularization, N. Srebro)! But it should be mentioned that one can find (ε,σ)-solution of SAA-problem for

$$N \sim M^2 R^2 \ln\left(\sigma^{-1}\right)/\varepsilon^2$$

calculations of stochastic subgradients of the terms of the sum chose at random. Indeed, let's introduce

$$f(x,\eta)=\begin{cases} f\left(x,\xi^1\right), \text{ with probability } 1/m \\ \text{...} \\ f\left(x,\xi^m\right), \text{ with probability } 1/m \end{cases}$$

Non stochastic sum-type SAA-problem can be considered as simple stochastic problem (bootstrap trick)

$$E_\eta\left[f(x,\eta)\right]\to\min_{x\in Q},$$

with stochastic subgradient: $\partial_x f(x,\eta)=\partial_x f\left(x,\xi^\eta\right)$, $\eta\in R[1,...,m]$. One can generate η for $\mathrm{O}(\log_2 m)$ arithmetic operations. Since $\left\|\partial_x f(x,\eta)\right\|_* \le M$ one can easily obtain that $N\sim M^2R^2\ln\left(\sigma^{-1}\right)/\varepsilon^2$ QED. But sometimes SAA-approach isn't substantial at all instead of SA (K. Sridharan's example).

Acceleration of Stochastic Approximation by proper Averaging

Let $\mathrm{x}_k, k = 1,\dots,N$ – i.i.d. with density function $p_{\mathrm{x}}\left(\mathrm{x}\middle|\theta\right)$ (supp. doesn't depend on θ), depends on unknown vector of parameters θ. Then for all statistics $\tilde{\theta}(\mathrm{x})$ ($E_{\mathrm{x}}\left[\tilde{\theta}(\mathrm{x})^2\right] < \infty$): $E_{\mathrm{x}}\left[\left(\tilde{\theta}(\mathrm{x}) - \theta\right)\left(\tilde{\theta}(\mathrm{x}) - \theta\right)^T\right] \succ \left[I_{p,N}\right]^{-1}$,

$$I_{p,N} = E_{\mathrm{x}}\left[\nabla_\theta \ln p_{\mathrm{x}}\left(\mathrm{x}\middle|\theta\right)\left(\nabla_\theta \ln p_{\mathrm{x}}\left(\mathrm{x}\middle|\theta\right)\right)^T\right] = N I_{p,1} \text{ (see Lecture 2).}$$

In 1990 B. Polyak (see also Polyak–Juditsky, 1992) showed that for

$$\theta^{k+1} = \theta^k + \gamma_k \nabla_\theta \ln p_{\mathrm{x}}\left(\mathrm{x}_k\middle|\theta^k\right),\ \bar{\theta}^N = \frac{1}{N}\sum_{k=1}^{N}\theta^k,\ \gamma_k = \gamma\cdot k^{-\beta},\ \beta\in(0,1),$$

$$\sqrt{N}\cdot\left(\bar{\theta}^N - \theta_*\right) \xrightarrow{d} N\left(0, \left[I_{p,1}\right]^{-1}\right),\ E_{\mathrm{x}}\left[N\cdot\left(\bar{\theta}^N - \theta_*\right)\left(\bar{\theta}^N - \theta_*\right)^T\right] \to \left[I_{p,1}\right]^{-1}.$$

SAA approach leads to analogues result (Fisher's theorem, Lecture 2).

Randomized MD for huge QP (Juditsky–Nemirovski randomization)

Let's consider QP problem ($n\times n$ matrix $A\succ 0$ is fully completed, $\left|A_{ij}\right|\le M$)

$$\frac{1}{2}\langle x, Ax\rangle \to \min_{x\in S_n(1)}.$$

Using STM (see Lecture 3), one can find ε-solution for

$\mathrm{O}\left(n^2\sqrt{M\ln n/\varepsilon}\right)$ arithmetic operations. // not good since $n\gg 1$ is huge

But if one use randomized MD with stochastic gradient $A^{\langle i[x]\rangle}$ – $i[x]$-column of matrix A and $P\left(i[x]=j\right)=x_j$, $j=1,\ldots,n$ (one can generate $i[x]$ for $\mathrm{O}(n)$ arithmetic operations), than one can find (ε,σ)-solutions for

$$\mathrm{O}\left(nM^2\ln n\cdot\ln\left(\sigma^{-1}\right)/\varepsilon^2\right) \text{ arithmetic operations.}$$

Randomized MD for Antagonistic matrix game (Grigoriadis–Khachiyan)

As we've already known (see Lecture 2) Google problem can be reduced to the saddle-point problem ($\tilde{A}$ is s-row and s-column sparse, Lecture 3)

$$\min_{x \in S_n(1)} \max_{\omega \in S_{2n}(1)} \left\langle \omega, \tilde{A}x \right\rangle.$$

Assume that there are two players A and B. All the players know matrix $\tilde{A} = \left\| \tilde{a}_{ij} \right\|$, where $\left| \tilde{a}_{ij} \right| \le 1$, $\tilde{a}_{ij}$ – prize of A (loss of B) in case when A plays i and B plays j. We play for the player B. Assume that the game is repeated $N \gg 1$ times. Let's introduce loss-function at the step k

$$f_k(x) = \left\langle \omega^k, \tilde{A}x \right\rangle, \; x \in S_n(1),$$

where $\omega^k \in S_{2n}(1)$ – such a vector with all zero components except one component, that component corresponds to the A's choice at the step k –

this components equals 1. This vector in principle could depends on all the history for that moment (but it can't depends on the realization of the randomized strategy of player B at the step k). Analogously, vector x^k has only one non zero component, corresponds to the choice of player B at the step k. One can introduce the price of the game ($C = 0$)

$$C = \max_{\omega \in S_{2n}(1)} \min_{x \in S_n(1)} \left\langle \omega, \tilde{A}x \right\rangle = \min_{x \in S_n(1)} \max_{\omega \in S_{2n}(1)} \left\langle \omega, \tilde{A}x \right\rangle. \quad \text{(von Neumann theorem)}$$

The solution of the saddle-point problem (ω, x) is Nash equilibrium. Since that (Hannan)

$$\min_{x \in S_n(1)} \frac{1}{N} \sum_{k=1}^{N} f_k(x) \leq C.$$

So if we (player B) will choose $\left\{x^k\right\}$ at random according to the following randomized MD-strategy (randomization under KL-projection!):

1. $p^1 = \left(n^{-1},\ldots,n^{-1}\right)$;

2. Choose at random $j\left(k\right)$ such, that $P\left(j\left(k\right)=j\right)=p_j^k$;

3. Put $x_{j(k)}^k = 1,\ x_j^k = 0,\ j \ne j\left(k\right)$;

4. Recalculate

$$p_j^{k+1} \sim p_j^k \exp\left(-\sqrt{\frac{2\ln n}{N}}\tilde{a}_{i(k)j}\right),\ j = 1,\ldots,n,$$

where $i\left(k\right)$ – the choice of A at the step k;

then with probability $\geq 1-\sigma$ (see Lecture 3 for MD in a simplex)

$$\frac{1}{N}\sum_{k=1}^{N} f_k\left(x^k\right) - \min_{x\in S_n(1)} \frac{1}{N}\sum_{k=1}^{N} f_k\left(x\right) \leq \sqrt{\frac{2}{N}}\left(\sqrt{\ln n} + 2\sqrt{2\ln\left(\sigma^{-1}\right)}\right),$$

i.e. with probability $\geq 1-\sigma$ our (B's player) loss can be bounded

$$\frac{1}{N}\sum_{k=1}^{N} f_k\left(x^k\right) \leq C + \sqrt{\frac{2}{N}}\left(\sqrt{\ln n} + 2\sqrt{2\ln\left(\sigma^{-1}\right)}\right).$$

The worst case – when A is also know this strategy and use it when choosing $\left\{\omega^k\right\}$ (it should be mentioned that A solve max-type problem). If A and B will use this strategy then they converges to Nash's equilibrium according to the following estimation.

With probability $\geq 1-\sigma$

$$0 \leq \left\| A\bar{x}^N \right\|_\infty = \max_{\omega\in S_{2n}(1)} \left\langle \omega, \tilde{A}\bar{x}^N \right\rangle - \max_{\omega\in S_{2n}(1)} \min_{x\in S_n(1)} \left\langle \omega, \tilde{A}x \right\rangle \leq$$

$$\leq \max_{\omega\in S_{2n}(1)} \left\langle \omega, \tilde{A}\bar{x}^N \right\rangle - \min_{x\in S_n(1)} \left\langle \bar{\omega}^N, \tilde{A}x \right\rangle \leq$$

$$\leq \max_{\omega\in S_{2n}(1)} \left\langle \omega, \tilde{A}\bar{x}^N \right\rangle - \frac{1}{N}\sum_{k=1}^{N} \left\langle \omega^k, \tilde{A}x^k \right\rangle + \frac{1}{N}\sum_{k=1}^{N} \left\langle \omega^k, \tilde{A}x^k \right\rangle - \min_{x\in S_n(1)} \left\langle \bar{\omega}^N, \tilde{A}x \right\rangle \leq$$

$$\leq \sqrt{\frac{2}{N}}\left(\sqrt{\ln(2n)} + 2\sqrt{2\ln(2/\sigma)}\right) + \sqrt{\frac{2}{N}}\left(\sqrt{\ln n} + 2\sqrt{2\ln(2/\sigma)}\right) \leq$$

$$\leq 2\sqrt{\frac{2}{N}}\left(\sqrt{\ln(2n)} + 2\sqrt{2\ln(2/\sigma)}\right),$$

where

$$\bar{x}^N = \frac{1}{N}\sum_{k=1}^{N} x^k, \ \bar{\omega}^N = \frac{1}{N}\sum_{k=1}^{N} \omega^k.$$

So when

$$N = 16\frac{\ln(2n) + 8\ln(2/\sigma)}{\varepsilon^2},$$

then with probability $\geq 1-\sigma$ one can guarantee $\left\|A\bar{x}^N\right\|_\infty \leq \varepsilon$. The total number of arithmetic operations can be estimated as follows

$$\mathrm{O}\left(n + \frac{s\ln n \cdot \ln(n/\sigma)}{\varepsilon^2}\right).$$

To be continued…