Convex Optimization for Data Science

Gasnikov Alexander

gasnikov.av@mipt.ru

Lecture 6. Gradient-free methods. Coordinate descent

February, 2017

Main books:

Spall J.C. Introduction to stochastic search and optimization: estimation, simulation and control. Wiley, 2003.

Nesterov Yu. Random gradient-free minimization of convex functions // CORE Discussion Paper 2011/1. 2011.

Nesterov Y.E. Efficiency of coordinate descent methods on large scale optimization problem // SIAM Journal on Optimization. 2012. V. 22. № 2. P. 341–362.

Fercoq O., *Richtarik P.* Accelerated, Parallel and Proximal Coordinate Descent // e-print, 2013. arXiv:1312.5799

Duchi J.C., *Jordan M.I.*, *Wainwright M.J.*, *Wibisono A.* Optimal rates for zero-order convex optimization: the power of two function evaluations // IEEE Trans. of Inf. 2015. V. 61. № 5. P. 2788–2806.

Wright S.J. Coordinate descent algorithms // e-print, 2015. arXiv:1502.04759

Gasnikov A.V. Searching equilibriums in large transport networks. Doctoral Thesis. MIPT, 2016. arXiv:1607.03142

Structure of Lecture 6

- Two-points gradient free methods and directional derivative methods (Preliminary results)
- Stochastic Mirror Descent and gradient-free methods
- The principal difference between one-point and two-points feedbacks
- Non smooth case (double-smoothing technique)
- Randomized Similar Triangles Method
- Randomized coordinate version of Similar Triangles Method
- Explanations why coordinate descent methods can works better in practice then its full-gradient variants
- Nesterov’s examples
- Typical Data Science problem and its consideration from the (primal / dual) randomized coordinate descent point of view

Two points gradient-free methods and directional derivative methods

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

All the results can be generalized for composit case (Lecture 3). We assume that

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

N – number of required iterations (oracle calls): calculations of f (realizations) / directional derivative of f.

R – “distance” between starting point and the nearest solution.

N	$E\left[\left\|\partial_x f(x,\xi)\right\|_2^2\right] \le M_2^2$	$\left\|\nabla f(y) - \nabla f(x)\right\|_2 \le L_2 \left\|y - x\right\|_2$	$E\left[\left\|\nabla_x f(x,\xi) - \nabla f(x)\right\|_2^2\right] \le D$
$f(x)$ convex	$n \cdot \frac{M_2^2 R^2}{\varepsilon^2}$	$n \cdot \sqrt{\frac{L_2 R^2}{\varepsilon}}$	$n \cdot \max\left\{\sqrt{\frac{L_2 R^2}{\varepsilon}}, \frac{DR^2}{\varepsilon^2}\right\}$
$f(x)$ – μ_2-strongly convex in $\left\| \, \right\|_2$	$n \cdot \frac{M_2^2}{\mu_2 \varepsilon}$	$n \cdot \sqrt{\frac{L_2}{\mu_2}} \left\lceil \ln\left(\frac{\mu_2 R^2}{\varepsilon}\right) \right\rceil$	$n \cdot \max\left\{\sqrt{\frac{L_2}{\mu_2}} \left\lceil \ln\left(\frac{\mu_2 R^2}{\varepsilon}\right) \right\rceil, \frac{D}{\mu_2 \varepsilon}\right\}$

Stochastic Mirror Descent (SMD) (Lectures 3, 4)

Consider convex optimization problem

$$f\left(x\right)\to\min_{x\in Q}, \qquad (1)$$

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

$$E_{\xi}\left[\nabla_x f\left(x,\xi\right)\right]=\nabla f\left(x\right). \qquad (2)$$

We introduce norm p-norm ($p\in\left[1,2\right]$) with $1/p+1/q=1$ and assume that

$$E_{\xi}\left[\left\|\nabla_x f\left(x,\xi\right)\right\|_q^2\right]\le M^2,\ q\in\left[2,\infty\right]. \qquad (3)$$

We introduce prox-function $d\left(x\right)\ge 0$ ($d\left(x^0\right)=0$) which is 1-strongly convex due to the p-norm and Bregman’s divergence (Lecture 3)

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

Method is

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

We put $R^2 = V\left(x_*, x^0\right)$, where x_* – is the solution of (1) (if x_* isn’t unique then we assume that x_* is minimized $V\left(x_*, x^0\right)$). If $\left\{\xi^k\right\}$ – i.i.d. and

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

Then, after (all the result cited below in this Lecture can be expressed in terms of probability of high deviations bounds, see Lecture 4)

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

iterations (oracle calls)

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

Idea (randomization!)

$$\tilde{\nabla}_x f\left(x^k,\xi^k := \left(\xi^k,e^k\right)\right) := \frac{n}{\tau} f\left(x^k+\tau e^k,\xi^k\right)e^k, \text{ (one-point feedback)} \quad (4)$$

$$\tilde{\nabla}_x f\left(x^k,\xi^k\right) := \frac{n}{\tau}\left(f\left(x^k+\tau e^k,\xi^k\right) - f\left(x^k,\xi^k\right)\right)e^k, \text{ (two-points feedback)} \ (5)$$

$$\nabla_x f\left(x^k,\xi^k\right) := n\left\langle \nabla_x f\left(x^k,\xi^k\right),e^k\right\rangle e^k. \text{ (directional derivative feedback)} \quad (6)$$

Assume that $f\left(x^k,\xi^k\right)$ available with (non stochastic) small noise of level δ.

How to choose i.i.d. e^k**?** Two main approaches:

$e^k \in RS_2^n\left(1\right)$ – e^k is equiprobable distributed on a unit Euclidian sphere in $\mathbb{R}^n$;

$e^k = \left(\underbrace{0,...,0,1}_{i},0,...,0 \right)$ – with probability $1/n$ (coordinate descent) for (5), (6).

Note, that (we can't tend $\tau \to 0+$ in (5) because of δ/τ in (7))

$$E_{e^k}\left[n\left\langle \nabla_x f\left(x^k,\xi^k\right),e^k\right\rangle e^k\right]=\nabla_x f\left(x^k,\xi^k\right), \text{(see (2))}$$

$$E\left[\left\|\frac{n}{\tau}\left(f\left(x^k+\tau e^k,\xi^k\right)-f\left(x^k,\xi^k\right)\right)e^k\right\|_q^2\right]\le\frac{3}{4}n^2\tau^2L_2^2E_{e^k}\left[\left\|e^k\right\|_q^2\right]+$$

$$+3n^2E\left[\left\langle \nabla_x f\left(x^k,\xi^k\right),e^k\right\rangle^2\left\|e^k\right\|_q^2\right]+12\frac{\delta^2n^2}{\tau^2}E_{e^k}\left[\left\|e^k\right\|_q^2\right]. \text{(see (3))} \qquad (7)$$

If $E_{\xi^k}\left[\left|f\left(x,\xi^k\right)\right|^2\right]\le B^2$ then

$$E\left[\left\|\frac{n}{\tau}f\left(x^k+\tau e^k,\xi^k\right)e^k\right\|_q^2\right]\le\frac{n^2B^2}{\tau^2}E_{e^k}\left[\left\|e^k\right\|_q^2\right]. \text{(see (3))} \qquad (8)$$

For coordinate descent randomization it's optimal to choose $p=q=2$. The results will be the same as for $e^k\in RS_2^n\left(1\right)$. Since that we concentrate on $e^k\in RS_2^n\left(1\right)$.

If $e \in RS_2^n(1)$ then due to the measure concentration phenomena (I. Usmanova)

$$E\left[\|e\|_q^2\right] \le \min\{q-1, 4\ln n\}\cdot n^{\frac{2}{q}-1},\ E\left[\langle c,e\rangle^2\right] \le \|c\|_2^2 n^{-1},\ 2\le q\le\infty,$$

$$E\left[\langle c,e\rangle^2\|e\|_q^2\right] \le \frac{4}{3}\|c\|_2^2 \min\{q-1, 4\ln n\}\cdot n^{\frac{2}{q}-2},\ 2\le q\le\infty.$$

So the choice of $p\in[1,2]$ ($q\in[2,\infty]$) is already nontrivial! For example, for $Q=S_n(1)$ – unit simplex in $\mathbb{R}^n$, it’s natural to choose $p=1$ ($q=\infty$).

For the function’s values feedback ((4), (5)) we have biased estimation of gradient ((2) isn’t still the truth). So one’ve to generalize mentioned above approach

$$E_{e^k}\left[\frac{n}{\tau}f\left(x^k+\tau e^k,\xi^k\right)e^k\right]\underset{\text{if } \delta=0}{=}E_{e^k}\left[\frac{n}{\tau}\left(f\left(x^k+\tau e^k,\xi^k\right)-f\left(x^k,\xi^k\right)\right)e^k\right]\ne$$

$$\ne E_{e^k}\left[n\left\langle\nabla_x f\left(x^k,\xi^k\right),e^k\right\rangle e^k\right].\ \text{// because } \tau\not\to 0+ \text{ and } \delta>0$$

Assume, that instead of real (unbiased) stochastic gradients $\nabla_x f\left(x^k,\xi^k\right)$ (see (2)) it's only available biased ones $\tilde{\nabla}_x f\left(x^k,\xi^k\right)$, that satisfy (3) and additionally

$$\sup_{\left\{x^k=x^k\left(\xi^1,\dots,\xi^{k-1}\right)\right\}_{k=1}^N} E\left[\frac{1}{N}\sum_{k=1}^N\left\langle E_{\xi^k}\left[\tilde{\nabla}_x f\left(x^k,\xi^k\right)-\nabla_x f\left(x^k,\xi^k\right)\middle|\xi^1,\dots,\xi^{k-1}\right],x^k-x_*\right\rangle\right]\le\sigma,$$

then

$$\boxed{E\left[f\left(\bar{x}^N\right)\right]-f_*\le\varepsilon+\sigma}.$$

If δ is small enough, then one can show (by the optimal choice of τ) that for (4):

N $\left(R^2=\tilde{O}\left(\left\lVert x_*-x^0\right\rVert_p^2\right)\right)$	$E\left[\left\lVert\partial_x f\left(x,\xi\right)\right\rVert_2^2\right]\le M_2^2$	$\left\lVert\nabla f\left(y\right)-\nabla f\left(x\right)\right\rVert_2\le L_2\left\lVert y-x\right\rVert_2$ (stochastic)
$f\left(x\right)$ convex	$\tilde{O}\left(\frac{B^2M_2^2R^2n^{1+2/q}}{\varepsilon^4}\right)$	$\tilde{O}\left(\frac{B^2L_2R^2n^{1+2/q}}{\varepsilon^3}\right)$
$f\left(x\right)-\mu_2$-strongly convex in $\lVert\ \rVert_2$	$\tilde{O}\left(\frac{B^2M_2^2n^2}{\mu_2\varepsilon^3}\right)$	$\tilde{O}\left(\frac{B^2L_2n^2}{\mu_2\varepsilon^2}\right)$

For directional derivative feedback (6) one can obtain:

$N\ \left(R^2 = \tilde{O}\left(\left\|x_* - x^0\right\|_p^2\right)\right)$	$E\left[\left\|\partial_x f(x,\xi)\right\|_2^2\right] \le M_2^2$	$\left\|\nabla f(y) - \nabla f(x)\right\|_2 \le L_2 \left\|y - x\right\|_2$ (stochastic)
$f(x)$ convex	$\tilde{O}\left(\frac{M_2^2 R^2 n^{2/q}}{\varepsilon^2}\right)$	$\tilde{O}\left(\frac{M_2^2 R^2 n^{2/q}}{\varepsilon^2}\right)$
$f(x) - \mu_2$-strongly convex in $\left\| \ \right\|_2$	$\tilde{O}\left(\frac{M_2^2 n}{\mu_2 \varepsilon}\right)$	$\tilde{O}\left(\frac{M_2^2 n}{\mu_2 \varepsilon}\right)$

But for the two-points feedback (5) if δ is small enough

$$\delta \le \min\left\{\frac{\varepsilon\tau}{16R\sqrt{n}}, \frac{M_2\tau}{\sqrt{96n}}\right\},$$

then by the optimal choice of τ:

$$\tau = \min\left\{\max\left\{\frac{\varepsilon}{2M_2}, \sqrt{\frac{\varepsilon}{L_2}}\right\}, \frac{M_2}{L_2}\sqrt{\frac{1}{6n}}\right\}, \ // \ \delta \le \frac{\varepsilon^{3/2}}{16R\sqrt{L_2 n}}.$$

one can prove that only the last column of this table is truth. As for non-smooth case one should replace (5) by (see arXiv:1701.03821)

$$\tilde{\nabla}_x f\left(x^k,\xi^k\right) := \frac{n}{\tau_2}\left(f\left(x^k+\tau_1\tilde{e}_1^k+\tau_2 e_2^k,\xi^k\right) - f\left(x^k+\tau_2\tilde{e}_1^k,\xi^k\right)\right)e_2^k,$$

where $\tilde{e}_1^k \in RB_2^n(1)$ ($\tilde{e}_1^k$ is equiprobable distributed on a unit euclidian sphere in $\mathbb{R}^n$), $e_2^k \in RS_2^n(1)$ and $\left\{\tilde{e}_1^k, e_2^k, \xi^k\right\}_k$ are independent in total. If δ is small enough

$$\delta \le \frac{\varepsilon^2}{56M_2Rn^{3/2}}, \text{// compare with } \delta \le \frac{\varepsilon^{3/2}}{16R\sqrt{L_2 n}}$$

then by the optimal choice of τ_1, τ_2:

$$\tau_1 = \frac{\varepsilon}{4M_2},\ \tau_2 = \frac{\varepsilon}{4M_2 n},$$

one can prove that the middle column of the table above is also truth.

Conclusions and Remarks for SMD approach

- One-point feedback is much worse than two-points feedback. Two-points feedback (under rather small noise) is equivalent to the directional derivative feedback. The last one (in the worth case) is n-times slower (in terms of the oracle calls) then full (sub-)gradient approach. Moreover, k-points feedback is $2n/k$-times slower than full (sub-)gradient approach.
- In non-euclidian set-up ($p \in [1,2)$) this additional n-factor (multiplier) is reduced up to a $\ln n$-factor when $p=1$ ($q=\infty$), but $M = M_2$.
- All the estimations in the last table are unimprovable up to a $\ln n$-factor. Note that denotation $\tilde{\mathrm{O}}(\ )$ (we've used above) is equivalent to $\mathrm{O}(\ )$ up to a $\ln n$-factor. For one-points feedback one can improve the results in terms of ε by degradations in terms of n (arXiv:1502.06398 , arXiv:1607.03084).
- All the results will be also true in the online context (arXiv:1607.03142).
- Using the restart-technique (see Lecture 5) one can generalize the results for non-euclidian set-up (in non online context).

Similar Triangles Methods (STM) $Q=\mathbb{R}^n$, $p=2$ (Lecture 3)

STM	Randomized STM
$y^{k+1}=\frac{\alpha_{k+1}u^k+A_kx^k}{A_{k+1}}$, $u^{k+1}=\text{Mirr}_{u^k}\left(\alpha_{k+1}\nabla f\left(y^{k+1}\right)\right)$, $x^{k+1}=\frac{\alpha_{k+1}u^{k+1}+A_kx^k}{A_{k+1}}$.	$y^{k+1}=\frac{\alpha_{k+1}u^k+A_kx^k}{A_{k+1}}$, $u^{k+1}=\text{Mirr}_{u^k}\left(\alpha_{k+1}\tilde{\nabla}_y f\left(y^{k+1},\xi^{k+1}\right)\right)$, $x^{k+1}=\frac{\alpha_{k+1}u^{k+1}+A_kx^k}{A_{k+1}}$.
$\alpha_0=L^{-1},\ A_k=\alpha_k^2L$, $\alpha_{k+1}=\frac{1}{2L}+\sqrt{\frac{1}{4L^2}+\alpha_k^2}$.	$\alpha_0=\left(Ln^2\right)^{-1},\ A_k=\alpha_k^2Ln^2$, $\alpha_{k+1}=\frac{1}{2Ln^2}+\sqrt{\frac{1}{4\left(Ln^2\right)^2}+\alpha_k^2}$.

$\tilde{\nabla}_y f\left(y^{k+1},\xi^{k+1}\right)$ is determines by (4) – (6) (in practice interesting only (5), (6)).

This method works (with (5) and (6)) by the formula in yellow cell

N	$E\left[\left\|\partial_x f(x,\xi)\right\|_2^2\right]\le M_2^2$	$\left\|\nabla f(y)-\nabla f(x)\right\|_2\le L_2\left\|y-x\right\|_2$	$E\left[\left\|\nabla_x f(x,\xi)-\nabla f(x)\right\|_2^2\right]\le D$
$f(x)$ convex	$n\cdot\frac{M_2^2R^2}{\varepsilon^2}$	$n\cdot\sqrt{\frac{L_2R^2}{\varepsilon}}$	$n\cdot\max\left\{\sqrt{\frac{L_2R^2}{\varepsilon}},\frac{DR^2}{\varepsilon^2}\right\}$
$f(x)-\mu_2$-strongly convex in $\|\ \|_2$	$n\cdot\frac{M_2^2}{\mu_2\varepsilon}$	$n\cdot\sqrt{\frac{L_2}{\mu_2}}\left\lceil\ln\left(\frac{\mu_2R^2}{\varepsilon}\right)\right\rceil$	$n\cdot\max\left\{\sqrt{\frac{L_2}{\mu_2}}\left\lceil\ln\left(\frac{\mu_2R^2}{\varepsilon}\right)\right\rceil,\frac{D}{\mu_2\varepsilon}\right\}$

Using restart-technique one can obtain method that works by the formula in green cell. Based on these two methods by using mini-batch'ing technique (see Lecture 5) one can obtain methods that work by the formula in blue cells.

Here it doesn't matter what kind of two described above ways of choosing e^k we will use. If we use (5) one should say that δ is small enough.

Unfortunately, here in the prove it is significant that $u^{k+1}-u^k$ is collinear to $\tilde{\nabla}_y f$. Sufficient condition for that is $Q=\mathbb{R}^n$.

An **open question** is to generalize these results for arbitrary convex set Q and $p \in [1,2]$.

Hypothesis is that – in colored cells in the table above multiplier n for $p \in [1,2]$ ($q \in [2,\infty]$) should be replaced by $n^{1/q}$. Note that by using SMD we've already shown that in the grey cells multiplier n for $p \in [1,2]$ ($q \in [2,\infty]$) should be replaced by $n^{1/q+1/2}$.

Following by the Lectures 3, 5 one can generalize mentioned above results (obtained around STM) on USTM and its intermediate variant.

Now we lead a general randomized block-coordinate descent scheme, based on STM, that allows us to obtain more precise results.

In the following we concentrated only on coordinate descent randomization because this typically allows to fulfill one iteration for $O(n)$ and if $f(x) = \sum_{k=1}^{m} f_k\left(a_k^T x\right)$ with $\left\{a_k^T\right\}_{k=1}^{m}$ – s-sparse in average – for $O(s)$ (Lee–Sidford).

Block-Coordinate Randomized Similar Triangles Method (CSTM)

Suppose that $Q=\prod_{i=1}^{n} Q_i$, where $Q_i \subseteq \mathbb{R}^{n_i}$. Let's put $\|x\|^2=\sum_{i=1}^{n} L_i^{1-2\beta}\|x_i\|_i^2$,

$V(x,y)=\sum_{i=1}^{n} L_i^{1-2\beta} V_i(x_i,y_i)$, $\beta\in[0,1]$, where $\|\ \|_i$ – norm in the corresponding i-block $\mathbb{R}^{n_i}$, $V_i(x_i,y_i)$ – corresponding Bregman's divergence and

$$\left\|\nabla_i f(x+h\tilde{e}_i)-\nabla_i f(x)\right\|_{*,i}\le L_i h\left\|[\tilde{e}_i]_i\right\|_i.$$

Let's introduce vector $\nabla_i f(x)$ that has zero's components except the positions correspond to block i for these components: $\nabla_i f=\nabla f$. We put $\tilde{n}_L=\sum_{i=1}^{n} L_i^{\beta}$, $p_i=L_i^{\beta}/\tilde{n}_L$. For $\beta=0$ we have $\tilde{n}_L=n$, $p_i=1/n$. This case (with $n_i\equiv 1$ and simpler prox-structure) we've already considered above.

CSTM

Choose independently at random i^{k+1} $\left(P\left(i^{k+1}=i\right)=p_i\right)$

$$\boxed{\begin{aligned} y^{k+1} &= \frac{\alpha_{k+1}u^k + A_k x^k}{A_{k+1}},\\ u^{k+1} &= \mathrm{Mirr}_{u^k}\left(\alpha_{k+1}\nabla_{i^{k+1}} f\left(y^{k+1}\right)\right),\\ x^{k+1} &= y^{k+1} + \frac{1}{p_{i^{k+1}}}\frac{\alpha_{k+1}}{A_{k+1}}\left(u^{k+1}-u^k\right).\end{aligned}}$$

$$\alpha_0 = \left(\tilde{n}_L^2\right)^{-1},\; A_k = \alpha_k^2\tilde{n}_L^2,\; \alpha_{k+1} = \frac{1}{2\tilde{n}_L^2} + \sqrt{\frac{1}{4\left(\tilde{n}_L^2\right)^2} + \alpha_k^2}.$$

Note that for $\beta = 0$

$$x^{k+1} = y^{k+1} + \frac{1}{p_{i^{k+1}}}\frac{\alpha_{k+1}}{A_{k+1}}\left(u^{k+1}-u^k\right) \sim x^{k+1} = \frac{\alpha_{k+1}u^{k+1} + A_k x^k}{A_{k+1}}.$$

The rate of convergence

$$\boxed{N = \mathrm{O}\left(\tilde{n}_L \cdot \sqrt{\frac{\tilde{R}_L^2}{\varepsilon}} \right), \ \tilde{R}_L^2 = V\left(x_*, y^0\right)}.$$

One can generalize this result for strongly convex case. Nontrivial (but possible – A. Turin & P. Dvurechensky, 2016) to generalize CSTM on adaptive variant (the values $\left\{L_i\right\}_{i=1}^{n}$ are not available a priori). This can be combined with block-separable composite type optimization (Lecture 3). As far as we know in this case it would be the most general block-coordinate primal-dual descent method with optimal rate of convergence. Moreover one can postpone this method for stochastic optimization problems (with general inexact oracle – see Lecture 5; say, for (5) error in f could be $\delta \sim \varepsilon^3/n$, $\tau \sim \sqrt{\delta}$).

Typically one should use $\beta = 0$, $\beta = 1/2$ (Yu. Nesterov, 2010, 2015).

Why coordinate descent method works good in practice?

Answer: Because of the cheap iteration!

Let’s explain this fact. Due to the (fast) automatic differentiation (AD) arXiv:1502.05767 and http://www.ccas.ru/personal/evtush/p/198.pdf

it seems that the cost of one iteration (the main part of this cost is oracle call) is of order of calculation of the gradient of f, because typically gradient can be calculate at most 4-times expensive then the value of f. But first of all AD requires a big memory (and sometimes it could be a serious problem, see arXiv:1701.02595), secondly for CSTM we need partial derivative (not the value of the function). For example, for $f(x) = x^T Ax$, $x \in \mathbb{R}^n$, with dense matrix A, $\nabla f(x)$ can be calculated for $2n^2$ a.o. but $\partial f(x)/\partial x_1$ – for $2n$ a.o. But this is not general situation: see, for example, $f(x) = \ln\left(\sum_{k=1}^{n} \exp(x_k)\right)$.

But the main thing is that – we need recalculation of block coordinates, instead of calculation as it would be for the first time!

Example 1 (Yu. Nesterov, 2015). Assume that

$$f\left(x\right)=F\left(Ax,x\right),\ x\in\mathbb{R}^n,\ y=Ax\in\mathbb{R}^m.$$

The value $F\left(y,x\right)$ (and due to AD also $\nabla F\left(y,x\right)$) can be calculated for $\mathrm{O}\left(m+n\right)$. Let at least one of the following conditions is true:

1) $n=\mathrm{O}\left(m\right)$;

2) calculation of $\nabla_y F\left(y,x\right)$ costs $\mathrm{O}\left(m\right)$ and $\dfrac{\partial F\left(y,x\right)}{\partial x_j}$ – costs $\mathrm{O}\left(m\right)$.

Then the average cost of one iteration of CSTM ($n_i\equiv 1$) is $\mathrm{O}\left(m\right)$. □

Example 2 (Yu. Nesterov, 2015). Assume that

$$f\left(x\right)=\frac{1}{2}\langle x,Sx\rangle-\langle b,x\rangle,$$

where S – positive semi-definite matrix with elements lies between 1 and 2. We use CSTM with $\beta = 1/2$ ($n_i = 1$). One can show that

$$L=\lambda_{\max}\left(S\right)\geq\lambda_{\max}\left(1_n 1_n^T\right)=n,$$

but $L_i = S_{ii} \leq 2$. So CSTM is faster STM $\sim\sqrt{n}$-times ($\tilde{n}_L = \sum_{i=1}^{n}\sqrt{L_i} \leq \sqrt{2}n$):

$$T_{CSTM}\left(\varepsilon\right)=\mathrm{O}\left(n\cdot n\sqrt{\frac{\left\|x_*-y^0\right\|_2^2}{\varepsilon}}\right),\ T_{STM}\left(\varepsilon\right)=\mathrm{O}\left(n^2\cdot\sqrt{\frac{n\left\|x_*-y^0\right\|_2^2}{\varepsilon}}\right).$$

In general it’s useful to note, that

$$\frac{1}{n}\operatorname{tr}(S) \le \lambda_{\max}(S) \le \operatorname{tr}(S),\ \frac{1}{n}\sum_{i=1}^{n}\sqrt{L_i} \le \sqrt{\frac{1}{n}\sum_{i=1}^{n}L_i} = \sqrt{\frac{1}{n}\operatorname{tr}(S)}.$$

Hence

$$T_{CSTM} = \tilde{\mathrm{O}}\left(n^2\sqrt{\frac{\left(\operatorname{tr}(S)/n\right)\Theta}{\varepsilon}}\right) \le \mathrm{O}\left(n^2\sqrt{\frac{\lambda_{\max}(S)\Theta}{\varepsilon}}\right) = T_{STM}\ .$$

Note that profit $\sim\sqrt{n}$-times is maximal possible and reach when $\lambda_{\max}(S)$ and $\operatorname{tr}(S)$ are close to each other. Say, if eigenvalues of S are $\{1,\dots,n\}$, then $\lambda_{\max}(S) = n$ and $\operatorname{tr}(S) \sim n^2$, so one need more asymmetry.

This is also can be generalized for the sparse matrix. □

Example 3 (strongly convex case). Let’s consider the problem (Q – simple structure convex set)

$$\sum_{k=1}^{m} f_k\left(A_k^T x\right)+g\left(x\right)\to\min_{x\in Q},$$

where $g\left(x\right)=\sum_{i=1}^{n} g_i\left(x_i\right)$. Gradients of the convex function f_k can be calculated for $\mathrm{O}\left(1\right)$ a.o. and all of these functions have Lipchitz constant of gradient L in 2-norm. Function $g\left(x\right)$ is assumed to be strongly convex in p-norm with constant μ. Let’s introduce matrix $A=\left[A_1,...,A_m\right]^T$. For simplicity we restrict ourselves here by the following two examples (see Lecture 2)

1) $\frac{L}{2}\left\|Ax-b\right\|_2^2+\frac{\mu}{2}\left\|x-x_g\right\|_2^2\to\min_{x\in\mathbb{R}^n}$, 2) $\frac{L}{2}\left\|Ax-b\right\|_2^2+\mu\sum_{k=1}^{n} x_k\ln x_k\to\min_{x\in S_n(1)}$.

One can build dual problems

1) $$\frac{1}{2\mu}\left(\left\|x_g - A^T y\right\|_2^2 - \left\|x_g\right\|_2^2\right) + \frac{1}{2L}\left(\left\|y+b\right\|_2^2 - \left\|b\right\|_2^2\right) \to \min_{y\in\mathbb{R}^m},$$

2) $$\frac{1}{\mu}\ln\left(\sum_{i=1}^{n}\exp\left(\frac{\left[-A^T y\right]_i}{\mu}\right)\right) + \frac{1}{2L}\left(\left\|y+b\right\|_2^2 - \left\|b\right\|_2^2\right) \to \min_{y\in\mathbb{R}^m}.$$

$$L_{STM} = \frac{1}{\mu}\max_{\|y\|_2\le 1,\,\|x\|_p\le 1}\left\langle A^T y, x\right\rangle^2 = \frac{1}{\mu}\max_{\|x\|_p\le 1}\left\|Ax\right\|_2^2 = \frac{1}{\mu}\begin{cases} 1)\ \lambda_{\max}\left(A^T A\right) \\ 2)\ \max\limits_{k=1,\dots,n}\left\|A^{\langle k\rangle}\right\|_2^2 \end{cases},$$

$$L_{CSTM} = \frac{1}{\mu}\max_{\|y\|_1\le 1,\,\|x\|_p\le 1}\left\langle A^T y, x\right\rangle^2 = \frac{1}{\mu}\max_{\|x\|_p\le 1}\left\|Ax\right\|_\infty^2 = \frac{1}{\mu}\begin{cases} 1)\ \max\limits_{k=1,..,m}\left\|A_k\right\|_2^2 \\ 2)\ \max\limits_{\substack{i=1,\dots,m \\ j=1,\dots,n}}\left|A_{ij}\right|^2 \end{cases}.$$

We will use CSTM with $\beta = 0$ $(n_i = 1)$, for the dual problems:

$$1)\quad T_1 = \tilde{O}\left(nm\sqrt{\frac{L\max\limits_{k=1,..,m}\left\|A_k\right\|_2^2}{\mu}} \right),\quad 2)\ T_2 = \tilde{O}\left(nm\sqrt{\frac{L\max\limits_{i,j}\left|A_{ij}\right|^2}{\mu}} \right).$$

Note that for the problem 1 we can also apply primal CSTM. Moreover, if $\left\{A_k\right\}_{k=1}^m$ have in average s nonzero elements in whole n-vector then

$$T_1^{dual} = \tilde{O}\left(sm\sqrt{\frac{L\max\limits_{k=1,..,m}\left\|A_k\right\|_2^2}{\mu}} \right),\quad T_1^{primal} = \tilde{O}\left(sm\sqrt{\frac{L\max\limits_{k=1,...,n}\left\|A^{\langle k\rangle}\right\|_2^2}{\mu}} \right).$$

If A is a bit-matrix, then: $T_1^{dual} = \tilde{O}\left(sm\sqrt{\frac{Ls}{\mu}} \right)$, $T_1^{primal} = \tilde{O}\left(sm\sqrt{\frac{Ls\cdot(m/n)}{\mu}} \right)$.

In Data Science application very often it is necessary to solve (see Lecture 2)

$$\frac{1}{m}\sum_{k=1}^{m} f_k\left(A_k^T x\right)+g\left(x\right)\to \min_{x\in Q},$$

1) $$\tilde{T}_1=\tilde{\mathrm{O}}\left(n\cdot\left(m+\min\left\{\sqrt{m\frac{L\max\limits_{k=1,..,m}\left\|A_k\right\|_2^2}{\mu}},\frac{L\max\limits_{k=1,..,m}\left\|A_k\right\|_2^2}{\mu}\right\}\right)\right),$$

2) $$\tilde{T}_2=\tilde{\mathrm{O}}\left(n\cdot\left(m+\min\left\{\sqrt{m\frac{L\max\limits_{i,j}\left|A_{ij}\right|^2}{\mu}},\frac{L\max\limits_{i,j}\left|A_{ij}\right|^2}{\mu}\right\}\right)\right).\ \square$$

The End?