Convex Optimization for Data Science

Alexander Gasnikov (MIPT), Soomin Lee (Yahoo),

Angelia Nedic (Univ. of Illinois), Cesar Uribe (Univ. of Illinois)

Lecture 7. Decentralized distributed optimization

June, 2017

Main books:

Bertsekas D.P., Tsitsiklis J.N. Parallel and distributed optimization: numerical methods. Prentice-Hall International, 1989.

Nedic A., Ozdaglar A. Cooperative distributed multi-agent optimizations, 2009.

https://asu.mit.edu/sites/default/files/documents/publications/Dist-chapter.pdf

Boyd S. et al. Distributed optimization and statistical learning via the alternating direction method of multipliers // Foundations and Trends in Machine Learning. 2011. V. 3(1). P. 1–122. http://web.stanford.edu/~boyd/papers/admm_distr_stats.html

Richtarik P. http://www.maths.ed.ac.uk/~prichtar/

Lan G., Lee S., Zhou Y. Communication-efficient algorithms for decentralized and stochastic optimization // e-print, 2017. arXiv:1701.03961

Scaman K. et al. Optimal algorithms for smooth and strongly convex distributed optimization in networks // e-print, 2017. arXiv:1702.08704

Structure of Lecture 7

- How to solve convex optimization problems with affine equality constraints in case when we have possibility to build explicitly dual problem
- How to solve convex optimization problems with affine restrictions in case when we don’t have possibility to build explicitly dual problem
- How to get acceleration in solving convex optimization problems with affine restrictions in case when we have proximal-friendly functional
- Decentralized distributed optimization. Consensus
- How to improve conditional number
- About the cost of gossip step
- Time-varying gossip graphs

How to solve convex optimization problems with affine restrictions in the case when we have a possibility to build explicitly a dual problem

1.a Strongly convex case (most important case because of regularization)

We have to solve the following convex optimization problem

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

where $f\left(x\right)$ is μ-strongly convex function in p-norm $\left(1\le p\le 2\right)$. We build dual problem (by Demyanov–Danskin's formula $\nabla\varphi\left(y\right)=Ax\left(y\right)-b$)

$$\varphi\left(y\right)=\max_{x\in Q}\left\{\left\langle y,Ax-b\right\rangle-f\left(x\right)\right\}=\left\langle y,Ax\left(y\right)-b\right\rangle-f\left(x\left(y\right)\right)\to\min_{y}. \quad (2)$$

In many applications the main contribution in computational complexity of one iteration comes from calculations of Ax, $A^T y$. To find $x\left(y\right)$ one can use optimal (including randomized) numerical methods (A. Nemirovski, Yu. Nesterov). But we assume first that we can calculate $x\left(y\right)$ explicitly.

To solve problem (2) let us use Adaptive Similar Triangles Methods (ASTM) of Yu. Nesterov, 2016 (in STM L is known and fixed)

Initialization $(k=0, y^0=0)$ arXiv:1604.05275; arXiv:1706.07622

Put

$$A_0 = \alpha_0 = 1/L_0^0 = 1,\ k=0,\ j_0 = 0;\ \tilde{y}^0 := u^0 := y^0 - \alpha_0 \nabla\varphi\left(y^0\right).$$

While

$$\{\varphi\left(\tilde{y}^0\right) > \varphi\left(y^0\right) + \left\langle \nabla\varphi\left(y^0\right), \tilde{y}^0 - y^0 \right\rangle + \frac{L_0^{j_0}}{2}\left\|\tilde{y}^0 - y^0\right\|^2\}$$

Do

$$\{\ j_0 := j_0 + 1;\ L_0^{j_0} := 2^{j_0} L_0^0;\ \left(A_0 :=\right)\alpha_0 := \frac{1}{L_0^{j_0}},\ \tilde{y}^0 := u^0 := y^0 - \alpha_0 \nabla\varphi\left(y^0\right)\ \}.$$

```
1. Put $L_{k+1}^{0} = L_{k}^{j_k} \big/ 2,\ j_{k+1} = 0.$
    $\alpha_{k+1} := \frac{1}{2L_{k+1}^{0}} + \sqrt{\frac{1}{4\left(L_{k+1}^{0}\right)^{2}} + \frac{A_k}{L_{k+1}^{0}}},\ A_{k+1} := A_k + \alpha_{k+1},$
    $y^{k+1} = \frac{\alpha_{k+1}u^{k} + A_k\tilde{y}^{k}}{A_{k+1}},\ u^{k+1} = u^{k} - \alpha_{k+1}\nabla\varphi\left(y^{k+1}\right),\ \tilde{y}^{k+1} = \frac{\alpha_{k+1}u^{k+1} + A_k\tilde{y}^{k}}{A_{k+1}}.$
2. While
    $\{\varphi\left(y^{k+1}\right) + \left\langle\nabla\varphi\left(y^{k+1}\right), \tilde{y}^{k+1} - y^{k+1}\right\rangle + \frac{L_{k+1}^{j_{k+1}}}{2}\left\|\tilde{y}^{k+1} - y^{k+1}\right\|_{2}^{2} < \varphi\left(\tilde{y}^{k+1}\right)\}$
Do
    $\{\ j_{k+1} := j_{k+1} + 1;\ L_{k+1}^{j_{k+1}} = 2^{j_{k+1}} L_{k+1}^{0};$
```

$$\alpha_{k+1} := \frac{1}{2L_{k+1}^{j_{k+1}}} + \sqrt{\frac{1}{4\left(L_{k+1}^{j_{k+1}}\right)^2} + \frac{A_k}{L_{k+1}^{j_{k+1}}}},\ A_{k+1} := A_k + \alpha_{k+1};$$

$$y^{k+1} := \frac{\alpha_{k+1}u^k + A_k\tilde{y}^k}{A_{k+1}},\ u^{k+1} := u^k - \alpha_{k+1}\nabla\varphi\left(y^{k+1}\right),\ \tilde{y}^{k+1} := \frac{\alpha_{k+1}u^{k+1} + A_k\tilde{y}^k}{A_{k+1}}\}.$$

3. If stopping rule is not satisfied, put $k := k+1$ and **go to** 1.

Put $$x^N = \sum_{k=0}^{N} \lambda_k x\left(y^k\right),\ \lambda_k = \alpha_k / A_N\ .$$

Since (x_* – solution of (1))

$$f\left(x^N\right) - f\left(x_*\right) \le \varphi\left(\tilde{y}^N\right) + f\left(x^N\right).$$

Note: ASTM (because of adaptability) can be applied below only for centralized context.

Theorem 1. *Let we solve problem (1) by passing to the dual problem (2), according to the formulas mentioned above. Let us choose the following stopping rule for ASTM*

$$\varphi\left(\tilde{y}^N\right) + f\left(x^N\right) \le \varepsilon, \left\|Ax^N - b\right\|_2 \le \tilde{\varepsilon}.$$

Then ASTM is terminated by making no more than ($L_\varphi = \mu^{-1} \cdot \max_{\|x\|_p \le 1} \|Ax\|_2^2$)

$$6 \cdot \max\left\{\sqrt{\frac{L_\varphi R^2}{\varepsilon}}, \sqrt{\frac{L_\varphi R}{\tilde{\varepsilon}}}\right\} \qquad (3)$$

iterations ("$6\cdot$"$\rightarrow$"$15\cdot$" *calculations of* Ax, $A^T y$), *where* $R^2 = \|y_*\|_2^2$, y_* – *solution of the problem (2) (if the solution is not unique, we can choose such a solution* y_* *that minimizes* R^2). arXiv:1602.01686 & arXiv:1606.08988

1.b Strongly convex case. Primal-duality via regularization

Let us introduce μ-strongly convex in 2-norm problem ($\mu > 0$)

$$F^{\mu}\left(x\right) = F\left(x\right) + \frac{\mu}{2}\left\|x - x^{0}\right\|_{2}^{2} \to \min_{x \in Q}. \qquad (4)$$

Let F_{*}^{μ} – is optimal value in (4), $F_{*} = F\left(x_{*}\right)$ – optimal value in (4) with $\mu = 0$.

Proposition 1 (regularization). *Let*

$$\mu \le \frac{\varepsilon}{\left\|x_{*} - x^{0}\right\|_{2}^{2}} = \frac{\varepsilon}{R^{2}},$$

and there exists such $x^{N} \in Q$ *that*

$$F^{\mu}\left(x^{N}\right) - F_{*}^{\mu} \le \varepsilon/2.$$

Then

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

Idea: regularize dual problem (2)

$$\varphi^{\mu}(y) = \varphi(y) + \frac{\mu}{2}\|y\|_2^2 \to \min_y, \tag{5}$$

where $\mu \simeq \varepsilon / R^2$. Since we typically don't know $R^2 = \|y_*\|_2^2$, we can restart on μ. The cost of that is just 8-multiplicative factor in the final estimate.

To solve (5), one can use Nesterov's Fast Gradient Descent (FGM)

$$\tilde{y}^0 = y^0 = 0,$$

$$y^{k+1} = \tilde{y}^k - \frac{1}{L_\varphi}\nabla\varphi^{\mu}\left(\tilde{y}^k\right),$$ (we use $x^N = x\left(y^N\right)$ for solution of (1))

$$\tilde{y}^{k+1} = y^{k+1} + \frac{\sqrt{L_\varphi} - \sqrt{\mu}}{\sqrt{L_\varphi} + \sqrt{\mu}}\left(y^{k+1} - y^k\right).$$

One can generalize this approach to adapt one in L_φ (arXiv:1604.05275).

Note, that

$$\frac{1}{2L_\varphi}\left\|\nabla\varphi^\mu(y)\right\|_2^2 \le \varphi^\mu(y)-\varphi_*^\mu \le \frac{1}{2\mu}\left\|\nabla\varphi^\mu(y)\right\|_2^2,$$

$$f\left(x(y)\right)-f\left(x_*\right)\le \|y\|_2\left\|Ax(y)-b\right\|_2.$$

We use the **stopping rule**:

$$\left\|y^N\right\|_2\left\|Ax\left(y^N\right)-b\right\|_2 \le \varepsilon,\ \left\|Ax\left(y^N\right)-b\right\|_2 \le \tilde{\varepsilon}.$$

Number of oracle calls (calculations of Ax, $A^T y$) does not exceed

$$N \simeq \sqrt{\frac{2L_\varphi\cdot\left(\varepsilon+2R\tilde{\varepsilon}\right)}{\tilde{\varepsilon}^2}}\ln\left(\frac{4L_\varphi\cdot\left(\min\limits_{x\in Q,\ Ax=b} f(x)-\min\limits_{x\in Q} f(x)\right)\cdot\left(\varepsilon+2R\tilde{\varepsilon}\right)}{\varepsilon\cdot\tilde{\varepsilon}^2}\right). \quad (6)$$

https://arxiv.org/ftp/arxiv/papers/1410/1410.7719.pdf

2. Smooth & Strongly convex case (Euclidian set up)

We have to solve the following convex optimization problem

$$f(x) \to \min_{Ax=0}, \tag{7}$$

where $f(x)$ is μ-strongly convex function in 2-norm, which has L-Lipchitz continuous gradient in 2-norm:

$$\left\|\nabla f(y) - \nabla f(x)\right\|_2 \le L\left\|y - x\right\|_2.$$

System $Ax = 0$ is assumed to be compatible. Matrix A might be not a full row rank matrix. This causes the multiple dual solution in the form $y_* + \mathrm{Ker}\left(A^T\right)$. Denote

$$\sigma_{\max}(A) = \lambda_{\max}\left(A^T A\right) = \max\left\{\lambda > 0 :\ \exists\ \ x \ne 0 : A^T Ax = \lambda x\right\} \text{ and}$$

$$\sigma_{\min}(A) = \min\left\{\lambda > 0 :\ \exists\ \ x \ne 0 : A^T Ax = \lambda x\right\}.$$

Our goal is to build a dual problem of the problem (one can easily generalize all the results in case $b \neq 0$)

$$\varphi(y) = f^*\left(A^T y\right) = \max_x \left\{\left\langle A^T y, x\right\rangle - f(x)\right\} =$$

$$= \left\langle A^T y, x\left(A^T y\right)\right\rangle - f\left(x\left(A^T y\right)\right) \to \min_y .$$

We assume that function $f^*(z)$ is $1/L$-strongly convex function in 2-norm and has $1/\mu$-Lipchitz continuous gradient in 2-norm. Then function $\varphi(y)$ is $\mu_\varphi = \sigma_{\min}(A)/L$-strongly convex function in 2-norm in $\operatorname{Ker}\left(A^T\right)^\perp$ and has $L_\varphi = \sigma_{\max}(A)/\mu$-Lipchitz continuous gradient in 2-norm. Denote

$$\chi\left(A^T A\right) = \sigma_{\max}(A)/\sigma_{\min}(A).$$

We use **stopping rule**: $\left\|y^N\right\|_2\left\|Ax\left(y^N\right)\right\|_2 \leq \varepsilon$, $\left\|Ax\left(y^N\right)\right\|_2 \leq \tilde{\varepsilon}$.

Nesterov’s FGM give us for this case the following number of calls of oracle:

$$N \simeq 2\sqrt{L_\varphi / \mu_\varphi} \cdot \ln\left(\max\left\{4L_\varphi R^2/\varepsilon, 2L_\varphi R/\tilde{\varepsilon}\right\}\right) \sim \sqrt{(L/\mu)\cdot \chi\left(A^T A\right)}. \quad (8)$$

3. Smooth case (Euclidian set up)

Now we assume that $f(x)$ has L-Lipchitz continuous gradient in 2-norm. In this case, we can regularize (see (4)) primal problem (7) with $\mu = \varepsilon / R_x^2$, $R_x^2 = \left\|x_* - x^0\right\|_2^2$. Then from (8) we obtain (up to a logarithmic factor) that

$$N \sim \sqrt{\left(LR_x^2/\varepsilon\right)\cdot \chi\left(A^T A\right)}. \quad (9)$$

4. Convex case without assumptions (Euclidian set up)

Now we assume that $f(x)$ is just a convex function. In this case, we can also regularize (see (4)) primal problem (7) with $\mu = \varepsilon / R_x^2$. And then we can use (3) or (6). We obtain correspondently (up to constant factor and log-factor)

$$N \sim \max\left\{ \sqrt{\frac{\sigma_{\max}(A) R^2 R_x^2}{\varepsilon^2}}, \sqrt{\frac{\sigma_{\max}(A) R R_x^2}{\tilde{\varepsilon}\varepsilon}} \right\}, \tag{10}$$

$$N \sim \sqrt{\frac{\sigma_{\max}(A) R_x^2 \cdot (\varepsilon + 2R\tilde{\varepsilon})}{\varepsilon\tilde{\varepsilon}^2}}. \tag{11}$$

Now we explain how we can compare formulas (3), (6) with (8) and (10), (11) with (9). We use Slater's arguments (for general problem (1))

$$R^2 = \|y_*\|_2^2 \le \|\nabla f(x_*)\|_2^2 \Big/ \sigma_{\min}(A) = M^2 \Big/ \sigma_{\min}(A). \qquad (12)$$

http://pwp.gatech.edu/guanghui-lan/wp-content/uploads/sites/330/2017/02/DCS20170131.pdf Theorem 3

Formulas (3), (8), (10) are unimpovable up to a constant factor, formulas (6), (9), (11) are unimpovable up to a logarithmic factor (in precision).

So let us **summarize** all the results in a compact form. We consider

$$f(x) \to \min_{Ax=0}.$$

If $f(x)$ is μ-strongly convex function in 2-norm and has L-Lipchitz continuous gradient in 2-norm then $\boxed{N \sim \sqrt{L_\varphi / \mu_\varphi} \sim \sqrt{(L/\mu)\cdot \chi\left(A^T A\right)}}$. If $\mu = 0$ one can take $\mu = \varepsilon / R_x^2$, $\boxed{N \sim \sqrt{\left(LR_x^2/\varepsilon\right)\cdot \chi\left(A^T A\right)}}$. If $L = \infty$ then $\mu_\varphi = 0$ and one can take $\mu_\varphi \simeq \varepsilon / R^2 \sim \varepsilon \sigma_{\min}(A)/M^2$, $\boxed{N \sim \sqrt{\left(M^2/(\mu\varepsilon)\right)\cdot \chi\left(A^T A\right)}}$ (here and in the next formula we assume $\tilde{\varepsilon} \sim \varepsilon / R$). If $\mu = 0$ and $L = \infty$ one can take $\mu = \varepsilon / R_x^2$, $\mu_\varphi \simeq \varepsilon / R^2 \sim \varepsilon \sigma_{\min}(A)/M^2$, $\boxed{N \sim \sqrt{\left(M^2 R_x^2/\varepsilon^2\right)\cdot \chi\left(A^T A\right)}}$.

One can generalize case 1), 4) to non Euclidian setup (we skip the details).

How to solve convex optimization problems with affine restrictions in the case when we don’t have a possibility to build explicitly dual problem

In **cases 2, 3** (see above), one can solve auxiliary problem

$$\max_{x\in Q}\left\{\langle y, Ax-b\rangle - f(x)\right\} = \langle y, Ax(y)-b\rangle - f\left(x(y)\right) \qquad (13)$$

to find $x(y)$ by fast gradient methods (in case 3 one should make additional regularization $\mu = \varepsilon / R_x^2$) applied to the strongly convex (concave) problem. So we can find $x(y)$ in a logarithmic number of iteration in the desired relative precision δ. This fact allows us to consider $x(y)$ to be almost the precise and don’t think about the inaccuracy in of $x(y)$ calculation. In case 2, we can solve (13) in $N \sim \sqrt{L/\mu}\ln\left(\delta^{-1}\right)$ oracle calls (here oracle call is a calculation of $\nabla f(x)$), in case 3 this can be done in $N \sim \sqrt{LR_x^2/\varepsilon}\ln\left(\delta^{-1}\right)$

calls. Both estimates are optimal up to a logarithmic factor. So, in cases 2, 3 we propose totally optimal methods.

Unfortunately this is not the true in **cases 1, 4**. In these cases one might use another approach, that gives optimal estimates in both sense: total number of $\nabla f(x)$ calculations and total number of Ax, $A^T y$ multiplications.

Let us start with the **case 4** and applied Nesterov's smoothing technique

$$\min_{Ax=0} f(x) = \min_x \left\{ \max_y \left\{ \langle y, Ax \rangle - f(x) \right\} \right\} =$$

$$= \max_y \left\{ \min_x \left\{ f(x) + \langle y, Ax \rangle \right\} \right\} \text{ – one of the dual problems.} \qquad (14)$$

Since we have the bound (12), we can replace $F(Ax) = \max_y \langle y, Ax \rangle$ by $F_\varepsilon(Ax) = \max_y \left\{ \langle y, Ax \rangle - \left(\varepsilon / 2R^2 \right) \|y\|_2^2 \right\} = \left(R^2 / 2\varepsilon \right) \|Ax\|_2^2$. One can show that the function $F_\varepsilon(Ax)$ has $\left(\sigma_{\max}\left(A^T\right)R^2/\varepsilon\right)$–Lipschitz continuous gradient in 2-norm. Note, that $\sigma_{\max}\left(A^T\right)=\sigma_{\max}(A)$.

So we have to solve composite type mixed smooth/non-smooth type problem

$$\underbrace{F_\varepsilon(Ax)}_{\sim 1/\varepsilon\text{-Lipchitz gradient}} + \underbrace{f(x)}_{M\text{-Lipchitz}} \to \min_{\|x\|_2\le R_x}, \tag{15}$$

where the gradient oracle for $F_\varepsilon(Ax)$ requires $\mathrm{O}(1)$ Ax multiplications (since we can write explicit formula for $\nabla F_\varepsilon(z)$) and gradient oracle for $f(x)$ require one $\nabla f(x)$ calculation. Using Lan's accelerated gradient sliding

http://pwp.gatech.edu/guanghui-lan/wp-content/uploads/sites/330/2016/02/GS-nonsmooth-stochastic6-11-submit.pdf

one can find ε-solution (in functional value) of (15) (without any auxiliary dual problems) after (see summarized slide above):

$$N_{[Ax]} \sim \sqrt{\frac{\left(\sigma_{\max}\left(A\right)R^2/\varepsilon\right)R_x^2}{\varepsilon}} = \sqrt{\frac{M^2R_x^2}{\varepsilon^2}\chi\left(A^TA\right)} \text{ and } N_{[\nabla f(x)]} \sim M^2R_x^2/\varepsilon^2 \quad (16)$$

Ax-multiplications and gradient $\nabla f\left(x\right)$-calculations. Unfortunately in this approach we can guarantee $\left\|Ax^N\right\|_2 \le \varepsilon/R$ only in the best case.

Using restart technique

https://hal.archives-ouvertes.fr/hal-00508933v1/document

http://www2.isye.gatech.edu/~nemirovs/MLOptChapterI.pdf

one can posptpone Lan’s accelerated gradient sliding for μ-strongly convex in 2-norm $f\left(x\right)$ (**case 1**). At k-th restart, $N_{[Ax]} \sim \sqrt{\sigma_{\max}\left(A\right)R^2/\left(\mu\varepsilon\right)}$ and $N_{[\nabla f(x)]} \sim 2^k M^2/\left(\mu^2R_x^2\right)$. This trick allows to improve estimates (16) for the problem (15) in the following manner (see summarized slide above):

$$N_{[Ax]} \sim \sqrt{\frac{M^2}{\mu\varepsilon}\chi\left(A^T A\right)} \ln\left(\frac{\mu R_x^2}{\varepsilon}\right) \text{ and } N_{[\nabla f(x)]} \sim M^2 / (\mu\varepsilon) \qquad (17)$$

Ax-multiplications and gradient $\nabla f\left(x\right)$-calculations.

Estimations (16), (17) are also unimprovable up to a logarithmic factor.

One can spread formulas (16), (17) for stochastic convex optimization problems (these estimations don’t changes)

http://pwp.gatech.edu/guanghui-lan/wp-content/uploads/sites/330/2017/02/DCS20170131.pdf

How to get acceleration in solving convex optimization problems with affine restrictions in case when we have proximal-friendly functional

Let us consider case 4 and return to (14)

$$\min_{Ax=0} f(x) = \max_{y} \min_{x} \left\{ f(x) + \langle y, Ax \rangle \right\} =$$

$$= \max_{y} \min_{z} \underbrace{\min_{x} \left\{ f(x) + \langle y, Ax \rangle + \frac{1}{2} \|x - z\|_2^2 \right\}}_{G(y,z)}.$$

Due to the assumptions, we can solve auxiliary strongly convex problem

$$\operatorname{prox}_{f, A^T y}(z) = \arg\min_{x} \left\{ f(x) + \left\langle A^T y, x \right\rangle + \frac{1}{2} \|x - z\|_2^2 \right\}$$

explicitly or in a cheap way.

Since

$$\left\|\nabla G(y',z')-\nabla G(y,z)\right\|_2 \le L_y \left\|y'-y\right\|_2 + L_z \left\|z'-z\right\|_2,$$

one can find ε-solution (in terms of duality gap) of a saddle-point problem

$$\max_{\|y\|_2\le R}\min_{z} G(y,z)$$

after $N_{[Ax]} \sim 1/\varepsilon$ Ax-multiplications

https://arxiv.org/pdf/1405.4980.pdf, item 5.2.

So we've obtained the well known result (see case 4 above) but in another manner. But unfortunately, we haven't gained any acceleration. Moreover, to find $\operatorname{prox}_{f,A^T y}(z)$ typically one have to find ε-solution of strongly convex optimization problem. It can be done in $1/\varepsilon$ $\nabla f(x)$-calculations. So, the total number of iterations will be of the order $N_{[\nabla f(x)]} \sim 1/\varepsilon^2$.

Now let us propose another approach

$$\min_{Ax=0} f(x) = \min_{Ax=0}\left\{ f(x) + \frac{1}{2}\|Ax\|_2^2 \right\} = \min_{x,z:\,Az=Ax}\left\{ f(x) + \max_y \langle -y, Az\rangle + \frac{1}{2}\|Az\|_2^2 \right\} =$$

$$= \min_{x,z}\left\{ f(x) + \max_y \langle -y, Az\rangle + \max_{y'}\langle y', Az - Ax\rangle + \frac{1}{2}\|Az\|_2^2 \right\} \overset{Az=u}{=}$$

$$\overset{Az=u}{=} -\min_{y,y'}\left\{ \max_x\left[\left\langle A^T y', x\right\rangle - f(x)\right] + \max_{u\in \operatorname{Im} A}\left\{\left\langle y - y', u\right\rangle - \frac{1}{2}\|u\|_2^2\right\}\right\} =$$

$$= -\min_{y,y'}\left\{ f^*\left(A^T y'\right) + \frac{1}{2}\left\|\operatorname{proj}_{\operatorname{Ker}\left(A^T\right)^{\perp}}\left(y - y'\right)\right\|_2^2 \right\}.$$

If $f^*\left(A^T y'\right)$ is proximal-friendly, we can solve auxiliary problem (this problem is just proximal version of standard dual problem) explicitly or in a cheap manner, then using accelerated proximal gradient descent (in space y)

http://web.stanford.edu/~boyd/papers/prox_algs.html

https://arxiv.org/pdf/1506.02186.pdf (allows to postpone on randomized methods)

one can find ε-solution of dual problem (it would be useful to spread this result for the discussion of the duality gap) in $N \sim \varepsilon^{-1/2}$ proximal steps. We obtain acceleration! But the natural question here as is foolows: since we have such a “magic” proximal oracle, why we can’t solve the standard dual problem? The answer as is follows: typically we don’t have a possibility to calculate explicitly $\operatorname{prox}_{f^*\left(A^T y'\right)}\left(y\right)$. But if we additionally assume that $f\left(x\right)$ is 1-strongly convex function in 2-norm, then we can find $\operatorname{prox}_{f^*\left(A^T y'\right)}\left(y\right)$ with the relative precision δ after $N_{[Ax]} \sim \ln \delta^{-1}$. However, the total number of Ax-multiplications and $\nabla f\left(x\right)$-calculations (up to a logarithmic factor) will be the same as above in case 1 (but for another approach).

https://arxiv.org/pdf/1507.06243.pdf

Decentralized distributed optimization. Preliminaries

https://arxiv.org/pdf/1702.08704.pdf

Assume we have a connected undirected graph $G=\langle V,E\rangle$ with n vertexes (nodes). Let A be adjacency matrix of this graph: $\mathrm{A}_{ij}=1$, $(i,j)\in E$; $\mathrm{A}_{ij}=0$, $(i,j)\neq E$. Let us introduce gossip matrix W (one can generalize the results mentioned below for general weighted symmetric communication's matrix)

$$W_{ij}=\begin{cases}-\mathrm{A}_{ij},\, i\neq j\\ \sum_{j=1}^{n}\mathrm{A}_{ij},\, i=j\end{cases}.$$

One can verify that W is nonnegative semidefinite matrix, with the following properties: $Wv=0\Leftrightarrow v_1=...=v_m$, $\sigma_{\max}\left(\sqrt{W}\right)=\lambda_{\max}\left(W\right)$.

Consensus

Assume that initially at k-th node ($k = 1,...,m$) of the graph $G = \langle V, E\rangle$ stored unique number v_k. Each node can obtain at each iteration average of its neighbors. How many iterations N required (and what is the proper algorithm) to reach the consensus $\left\{v_i^N\right\}_{i=1}^m$:

$$\sqrt{\sum_{i=1}^m \left(v_i^N - \frac{1}{n}\sum_{j=1}^m v_j \right)^2} \le \varepsilon \sqrt{\sum_{i=1}^m \left(v_i - \frac{1}{n}\sum_{j=1}^m v_j \right)^2} \text{ ?}$$

To answer for this questions let us consider convex optimization problem

$$\frac{1}{2}\langle \nu, W\nu \rangle \to \min_{\nu}. \tag{18}$$

One can solve this problem by Nesterov’s FGM for strongly convex problem. So one can obtain $N \sim \sqrt{\chi(W)}\ln\varepsilon^{-1}$ ($\sqrt{\chi(W)}$ often reflects the diameter of the graph). This estimate is unimprovable up to a constant factor. The most important thing here is that FGM ($v_i^0 = \tilde{v}_i^0 = v_i$),

$$v^{k+1} = \tilde{v}^k - \frac{1}{\sigma_{\max}\left(\sqrt{W}\right)} W\tilde{v}^k,$$

$$\tilde{v}^{k+1} = v^{k+1} + \frac{\sqrt{\sigma_{\max}\left(\sqrt{W}\right)} - \sqrt{\sigma_{\min}\left(\sqrt{W}\right)}}{\sqrt{\sigma_{\max}\left(\sqrt{W}\right)} + \sqrt{\sigma_{\min}\left(\sqrt{W}\right)}}\left(v^{k+1} - v^k\right).$$

satisfy the condition “each node obtains at each iteration average of its neighbors” because of matrix-vector multiplication $W\tilde{v}^k$ (all other calculations are fully separable).

Note: one can consider $W = D - \tilde{P} = \operatorname{diag}\left\{\sum_{j=1}^{n} \mathrm{A}_{ij}\right\}_{i=1}^{m} - \tilde{P}$ to be Laplacian matrix. So one can applied simple power method $\nu^{k+1} = P\nu^{k} = D^{-1}\tilde{P}\nu^{k}$ (see ergodic theorem for Markov chain). Here "each node also obtains at each iteration average of its neighbors" because of matrix-vector multiplication $P\nu^{k}$. Unfortunately, the number of required iterations will be $N \sim \chi(W)\ln\varepsilon^{-1}$. But this result is still be the truth when $G = \langle V, E\rangle$ is directed graph ($\tilde{P}$ is not symmetric matrix). One should also mention that procedure $\nu^{k+1} = P\nu^{k}$ can be considered as a (non accelerated) weighted gradient descent for the problem (18).

The next important step is to consider v_k to be vectors from $\mathbb{R}^n$. In this case $Wv = 0 \Leftrightarrow v_1 = ... = v_m$, where $W := W \otimes I^{n\times n}$ – Kronecker product, $I_{ij}^{n\times n} = 1$.

Let us consider convex optimization problem

$$f\left(x\right)=\sum_{k=1}^{m} f_k\left(x\right)\to\min_{x\in\mathbb{R}^n}.$$

Let us introduce $x=\left(x_1,\dots,x_m\right)\in\mathbb{R}^{mn}$ $(R_x^2\to mR_x^2,\ R^2\to mR^2)$

$$f\left(x\right)=\sum_{k=1}^{m} f_k\left(x_k\right)\to\min_{x_1=\dots=x_m}.$$

One can rewrite this problem in two different ways

$$f\left(x\right)=\sum_{k=1}^{m} f_k\left(x_k\right)\to\min_{Wx=0}, \tag{19}$$

$$f\left(x\right)=\sum_{k=1}^{m} f_k\left(x_k\right)\to\min_{\sqrt{W}x=0}. \tag{20}$$

If $f_k\left(x_k\right)$ is μ_k-strongly convex function in 2-norm and has L_k-Lipchitz continuous gradient in 2-norm then $f\left(x\right)$ is $\mu = \min\limits_{k=1,...,m} \mu_k$-strongly convex function in 2-norm and has $L = \max\limits_{k=1,...,m} L_k$-Lipchitz continuous gradient in 2-norm.

The main observations ($W = W^T$**):**

1) If one put $A = W$ then all the results on summarized slide can be applied to (19);

2) If one put $A = \sqrt{W}$ then all the results mentioned above (not only collected on summarized slide) can be applied to (20) if we change (when it's required) dual variables $\sqrt{W}y \to z$:

$$\sqrt{W}\cdot\left| y^{k+1}=\tilde{y}^{k}-\frac{1}{L_{\varphi}}\nabla\varphi\left(\tilde{y}^{k}\right)=\tilde{y}^{k}-\frac{1}{L_{\varphi}}\sqrt{W}x\left(\sqrt{W}\tilde{y}^{k}\right),\right.$$

$$\sqrt{W}\cdot\left|\tilde{y}^{k+1}=y^{k+1}+\frac{\sqrt{L_{\varphi}}-\sqrt{\mu_{\varphi}}}{\sqrt{L_{\varphi}}+\sqrt{\mu_{\varphi}}}\left(y^{k+1}-y^{k}\right).\right.$$

This leads to the

$$z^{k+1}=\tilde{z}^{k}-\frac{1}{L_{\varphi}}Wx\left(\tilde{z}^{k}\right),\ \tilde{z}^{k+1}=z^{k+1}+\frac{\sqrt{L_{\varphi}}-\sqrt{\mu_{\varphi}}}{\sqrt{L_{\varphi}}+\sqrt{\mu_{\varphi}}}\left(z^{k+1}-z^{k}\right).$$

The last formulas can be fulfilled in a distributed manner since we have only *Wx* interaction term. It’s obvious that one should use (20) because in this case we have $\chi\left(A^{T}A\right)=\chi\left(W\right)$ instead of $\chi\left(W^{T}W\right)=\chi\left(W^{2}\right)\gg\chi\left(W\right)$ (in case (19)). Note, that for star topology $\sqrt{\chi\left(W\right)}\sim\sqrt{m}$, for totally connected graphs

$\sqrt{\chi(W)} \sim \operatorname{diam}(G) \sim 1$, for linear graphs $\sqrt{\chi(W)} \sim \operatorname{diam}(G) \sim m$, for regular networks $\sqrt{\chi(W)} \ge \operatorname{diam}(G)/\ln m$. So we can observe that $\sqrt{\chi(W)}$ typically corresponds to the diameter of the graph G and square root of spectral gap of stochastic matrix P. Note that for star topology $\operatorname{diam}(G) = 2$ (one master and $m-1$ slaves, master connect to each slave; slaves don't connect to each other). Is it possible to change $\sqrt{\chi(W)}$ to $\operatorname{diam}(G)$ in cases 1, 4? Yes, but in general without gossip consensus communication!;

3) The main auxiliary problem (in both of the cases 1) and 2))

$$\arg\max_{x}\left\{\left\langle A^T y, x\right\rangle - f(x)\right\} = \arg\max_{x_k, k=1,\dots,m}\left\{\left\langle \left[A^T y\right]_k, x_k\right\rangle - f_k(x_k)\right\}$$

can be also spitted in a distributed manner.

Unfortunately, in the case 2 in described above approach L/μ can be too big!

How to improve conditional number L/μ (case 2)

As we've already mentioned above $\mu = \min\limits_{k=1,\dots,m} \mu_k$, $L = \max\limits_{k=1,\dots,m} L_k$. This is not good in general, because L/μ can be too large if one of the μ_k is small. To eliminate this drawback one can reformulate (20) as

$$f_\alpha(x) = \sum_{k=1}^{m} f_k(x_k) + \frac{\alpha}{2}\langle x, Wx\rangle \to \min_{\sqrt{W}x=0}. \qquad (21)$$

Note that $f_\alpha(x)$ is $\mu \geq \min\left\{\sum\limits_{k=1}^{m}\mu_k, \alpha\lambda_{\min}(W)\right\}$-strongly convex in 2-norm and has $L \leq \left(\max\limits_{k=1,\dots,m} L_k + \alpha\lambda_{\max}(W)\right)$-Lipchitz continuous gradient in 2-norm. If we put $\alpha \simeq \sum\limits_{k=1}^{m}\mu_k \Big/ \lambda_{\min}(W)$, then one can solve (21) with relative precision ε

(see first half of this presentation to understand what does it mean, but due to the additional term in functional this is not the same) after (see also https://arxiv.org/pdf/1607.03218.pdf)

$$N_{[Wx]} \sim \left(\frac{\max\limits_{k=1,\dots,m} L_k}{\sum\limits_{k=1}^{m} \mu_k} + \chi(W) \right) \sqrt{\chi(W)} \cdot \ln^2\left(\varepsilon^{-1}\right) \qquad (22)$$

consensus (gossip/communication) steps and $N_{[\nabla f(x)]} \sim N_{[Wx]} \cdot \ln^{-1}\left(\varepsilon^{-1}\right)$ gradient $\nabla f(x)$ calculations.

Using regularization technique (with $\mu_k = \varepsilon \big/ \left(mR_{x_k}^2\right) = \varepsilon / R_x^2$) one can postpone this result to the case 3.

How one can understand mentioned above results? Naive explanation

Let us look once again at the estimations on summarized slide

$$\boxed{N \sim \sqrt{\left(L/\mu\right)\cdot \chi\left(W\right)}},\ \boxed{N \sim \sqrt{\left(LR_x^2/\varepsilon\right)\cdot \chi\left(W\right)}}, \tag{23}$$

$$\boxed{N \sim \sqrt{\left(M^2/\left(\mu\varepsilon\right)\right)\cdot \chi\left(W\right)}},\ \boxed{N \sim \sqrt{\left(M^2R_x^2/\varepsilon^2\right)\cdot \chi\left(W\right)}}. \tag{24}$$

In the smooth cases $L < \infty$ (cases 2, 3) these estimations follows (up to a logarithmic factor) from the classical (non-distributive) estimations

$$\boxed{N \sim \sqrt{L/\mu}\ln\left(\mu R_x^2/\varepsilon\right)},\ \boxed{N \sim \sqrt{LR_x^2/\varepsilon}},$$

and the fact that one has to additionally do $\sqrt{\chi\left(W\right)}\ln \varepsilon^{-1}$ consensus (communications) steps on each iteration of classical (non-distributive) FGM (analogously (22)). We’ve talked about it above (see the consensus slides).

Almost in all the calculations above we significantly use the fact that $W = W^T$. Unfortunately, at the moment we know how to use reduction philosophy described above only in this ($W = W^T$) case. But one can split primal and dual steps in different networks (dual to each other). In this case networks can be directed. But this is a very special case (we have two communication networks dual to each other). But if one returns to the consensus slides and simple power method (with matrix P, not necessarily symmetric) one can work on directed graph. The payment for that is

$$\boxed{N \sim \sqrt{L/\mu} \cdot \chi(W)}, \ \boxed{N \sim \sqrt{LR_x^2/\varepsilon} \cdot \chi(W)}.$$

Unfortunately, in non-smooth cases (cases 1, 4) at the moment we don’t have such a simple explanation in general case (in special cases, i.e. $f(x)$ admit explicitly calculated Nesterov’s smoothing representation, it is possible).

Now let us return to the question: how to reduce in (23) (cases 2, 3) $\sqrt{\chi(W)}$ to $\operatorname{diam}(G)$? Do to this we have to change the philosophy from decentralize to centralize. That is we fixed one node to be a master node. This node from $\operatorname{diam}(G)$ communications steps can collect full gradient from the slave's nodes and then make a standard iteration of FGM. Then master spread new values to the slaves (it also takes $\operatorname{diam}(G)$ communication steps). But this philosophy typically required more assumptions about synchronization delays (the slowest slave determine the performance), sensitivity to errors in master node and computational power of master node. Moreover in time-varying graph (it is typical for some applications) it is hardly possible to organize efficiently such a procedure. At the end let us mentioned that in cases 1, 4 (see also (24)) it seems impossible to reduce $\sqrt{\chi(W)}$ to $\operatorname{diam}(G)$ because here we have that the number of gossip steps is smaller the number of gradient-oracle calls (as far as we know it's an open question to explain it).

About the cost of gossip step

CPUs in these days can read and write the memory at over 10 GB per second whereas communication over TCP/IP is about 10 MB per second. Therefore, the gap between intra-node computation and inter-node communication is about 3 orders of magnitude. Communication start-up cost itself is also not negligible as it usually takes a few milliseconds.

So let us to consider that one node can calculate $\nabla f_k\left(x_k\right)$ (or even calculate $x_k\left(y\right)$) to the 1 unite of time and the communication (gossip) step takes τ unites of time. In case $\tau \gg 1$ all the results above seem reasonable, because we first think of communication steps. But if $\tau \ll 1$ one should use Chebyshev acceleration (see https://arxiv.org/pdf/1702.08704.pdf): $W \to \mathrm{Poly}_{Cheb}^{\sqrt{\chi(W)}}\left(W\right)$, $\tilde{W} = \mathrm{Poly}_{Cheb}^{\sqrt{\chi(W)}}\left(W\right)$ is also gossip matrix with $\chi\left(\tilde{W}\right) \sim 1$.

Time-varying gossip graphs

Let us return to consensus problem

$$\frac{1}{2}\langle \nu, W\nu \rangle \to \min_{\nu}.$$

But now we assume that from time to time matrix W changes, remaining every time the gossip matrix, hence every time $Wv = 0 \Leftrightarrow v_1 = ... = v_m$. So we have a family of nonnegative semi-definite quadratic functions with the same $\mathrm{Ker}(W)$ (in our case this kernel described by $v_1 = ... = v_m$). How to find a projection of $v_1^0,...,v_m^0$ on this set working at each iteration with different matrixes W? One can solve (18) (with relative precision ε) by the simple gradient descent method for $\chi(W)\ln\varepsilon^{-1}$ gossip steps because of this dynamic has Lyapunov function: square distance between current point and $\mathrm{Ker}(W)$.

But we know that if W is fixed one can accelerate to $\sqrt{\chi(W)}\ln\varepsilon^{-1}$ gossip steps. Is it possible in general case? We don’t know the exact answer for the moment. All known to us different versions of Nesterov’s accelerated gradient descent (even for the variants with line and 2-plane search):

https://ie.technion.ac.il/~mcib/sesop_report_version301005.pdf ; https://arxiv.org/pdf/1405.4980.pdf, items 3.6, 3.7 https://arxiv.org/pdf/1407.1537.pdf ; https://arxiv.org/pdf/1506.02186.pdf ; https://arxiv.org/pdf/1512.07516.pdf ; https://www.lccc.lth.se/media/LCCC2017/WorkshopOptimization/slides/Lessard%20LCCC%20slides.pdf

don’t allows such a generalization directly. But if one can detect the moment of changes of W and such changes don’t happen very often one can use restart technique that allows to obtain the following estimation of gossip steps

$$\sqrt{\max_{W}\left(\sigma_{\max}\left(\sqrt{W}\right)\Big/\sigma_{\min}\left(\sqrt{W}\right)\right)}\ln\varepsilon^{-1}.$$

https://arxiv.org/pdf/1609.07358.pdf ; https://arxiv.org/ftp/arxiv/papers/1703/1703.00267.pdf Remark 1

One can show that this result postpones to the general dual problem (14).

To be continued…