Introductory Course on Non-smooth Optimisation

Lecture 10 - Stochastic optimisation

Jingwei Liang

Department of Applied Mathematics and Theoretical Physics

Table of contents

Data n observations

$$(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, \ i = 1, ..., m.$$

Goal Find a prediction as a linear function of $\theta^T \Phi(x)$ of features $\Phi(x) \in \mathbb{R}^d$.

Regularised empirical risk minimisation find $\hat{\theta}$ via

$$\min_{\theta \in \mathbb{R}^n} \ \mu R(\theta) + \frac{1}{m} \sum_{i=1}^{m} \ell\big(y_i, \theta^T \Phi(x_i)\big).$$

Challenges

- Large scale, m is very large.
- Complexity.

Stochastic optimisation

Categories

- Solving stochastic problem with deterministic/stochastic methods.
- Solving deterministic problem with stochastic methods.

Developments

- Starts from 1950s.
- Tremendous growth since around 2010: machine learning...

Stochastic optimisation

- Stochastic optimisation is prevailing in data science, machine learning.
- The nature of machine learning tasks: no need to optimise below statistical error.
- Cost functions are averages: expectations.
- Test errors are more important than training error: over-fitting.

Deterministic optimisation

- Deterministic optimisation is still dominating fields including inverse problems, signal/image processing, variational inequalities...
- For many problems, stochastic methods won't provide any benefits: not expectations, hence big difference in step-size.
- However, the situation is changing: medical imaging.

Contents

Contents will be covered

- Incremental methods.
- Stochastic gradient descent.
- Variance reduction technique.
- Non-convex problem: escape bad local minimal or saddle points.

NB: Via the perturbation of deterministic methods perspective.

Outline

Problem

$$\min_{x \in \mathbb{R}^n} R(x) + F(x).$$

- R is regularisation function.
- $F(x) = \sum_{i=1}^{m} f_i(x)$ is a finite separable sum.

Deterministic gradient methods and their complexity

- Proximal gradient descent

$$x_{k+1} = \text{prox}_{\gamma R}(x_k - \gamma \nabla F(x_k)).$$

- Subgradient descent

$$x_{k+1} = x_k - \gamma_k g_k, \;\; g_k \in \nabla F(x_k) + \partial R(x_k).$$

Consider

$$\min_{x\in\mathbb{R}^n} F(x) \stackrel{\text{def}}{=} \sum_{i=1}^m f_i(x)$$

and

$$x_{k+1} = x_k - \gamma\nabla F(x_k) = x_k - \gamma\sum_i \nabla f_i(x_k).$$

Question at each step k, what if we **randomly** pick an integer

$$i_k \in \{1, 2, \cdots, m\}$$

and apply the following incremental updates

$$x_{k+1} = x_k - \gamma\nabla f_{i_k}(x_k).$$

- Instead of random sampling, i_k can go cyclically through $\{1, 2, \cdots, m\}$.
- Only the gradient of the sampled function is needed, m times faster than $\nabla F(x_k)$.

Incremental schemes

- Neural networks: back-propagation.
- In m steps, one path through all data, especially when cycling i_k.
- Effectiveness of the scheme depends on the variance of data.
- Usually effective when far from the solution, become very slow when close to the solution.

Sum of squares

$$\min_{x \in \mathbb{R}} F(x) \stackrel{\text{def}}{=} \sum_{i=1}^{m} \frac{1}{2}(a_i x - b_i)^2$$

- Solving $\nabla F(x^\star) = 0$, we get
$$x^\star = \frac{\sum_i a_i b_i}{\sum_i a_i^2}.$$
- For each $i = 1, \ldots, m$, $f_i(x) = \frac{1}{2}(a_i x - b_i)^2$ minimises at
$$\bar{x}_i = \frac{b_i}{a_i}.$$
- Note that
$$x^\star \in \Omega \stackrel{\text{def}}{=} \left[\min_i \bar{x}_i, \max_i \bar{x}_i \right].$$

Sum of squares

$$\min_{x \in \mathbb{R}} F(x) \stackrel{\text{def}}{=} \sum_{i=1}^{m} \frac{1}{2}(a_i x - b_i)^2$$

- Given a point x, then
$$\nabla f_i(x) = a_i(a_i x - b_i),$$
$$\nabla F(x) = \sum_i a_i(a_i x - b_i).$$
- If x is outside Ω, then $\nabla f_i(x)$ and $\nabla F(x)$ have the same sign, *i.e.* pointing at same direction.
- Once $x \in \Omega$, **no guarantee** that $\nabla f_i(x)$ will point toward $x^\star$.

Incremental proximal point algorithm

Consider applying PPA to solve

$$\min_{x\in\mathbb{R}^n} F(x) \stackrel{\text{def}}{=} \sum_{i=1}^m f_i(x).$$

That is

$$x_{k+1} = \text{prox}_{\gamma F}(x_k).$$

Incremental PPA randomly sample i_k from $\{1, 2, \cdots, m\}$, and

$$x_{k+1} = \text{prox}_{\gamma_k f_{i_k}}(x_k).$$

Regularised finite sum

$$\min_{x \in \mathbb{R}^n} R(x) + \big\{ F(x) \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^{m} f_i(x) \big\}.$$

Proximal gradient

$$x_{k+1} = \text{prox}_{\gamma R}\big(x_k - \nabla F(x_k)\big).$$

Incremental proximal gradient randomly sample i_k from $\{1, 2, \cdots, m\}$, and

$$x_{k+1} = \text{prox}_{\gamma R}\big(x_k - \nabla f_{i_k}(x_k)\big).$$

NB: i_k can also be chosen as $i_k = \text{mod}(k, m)$.

If moreover R is also a finite sum

$$R(x) \stackrel{\text{def}}{=} \sum_{i=1}^{m} r_i(x).$$

Increments both proximal mapping and gradient randomly sample i_k from $\{1, 2, \cdots, m\}$, and

$$x_{k+1} = \text{prox}_{\gamma r_{i_k}}\left(x_k - \nabla f_{i_k}(x_k)\right).$$

For sampled gradient

$$\nabla f_i(x) = \nabla F(x) + \big(\nabla f_i(x) - \nabla F(x)\big).$$

Define the (random) error

$$\epsilon_k \overset{\text{def}}{=} \nabla f_i(x) - \nabla F(x).$$

Perturbed proximal gradient randomly sample i_k from $\{1, 2, \cdots, m\}$, and

$$x_{k+1} = \text{prox}_{\gamma_k R}\big(x_k - \gamma_k(\nabla F(x_k) + \epsilon_k)\big).$$

NB: in terms of inexact Krasnosel'skiĭ-Mann iteration, for convergence

$$\sum_k \gamma_k \|\epsilon_k\| < +\infty.$$

Vanishing step-size $\|\epsilon_k\| \leq C$, hence $\gamma_k \epsilon_k \to 0$ which implies $\gamma_k \to 0$.

Incremental methods can be seen as perturbed gradient methods, same for the to be introduced stochastic methods.

- Need to control the coupled term $\gamma_k \epsilon_k$ for convergence.
- Error makes even smooth problems more like non-smooth ones.
- Convergence depends crucially on step-size γ_k.

Step-size choices

- Constant step-size $\gamma_k \equiv \gamma$, should be small enough. Not always work, except least square...
- $\gamma_k \to$ and $\sum_k \gamma_k = +\infty$.
- "constant $\to$ vanishing $\to$ constant $\to$ vanishing..."
- Let $a, b, c > 0$ and

$$\gamma_k = \min \left\{ c, \frac{a}{k+b} \right\}.$$

Suppose the m points $x_k, x_{k-1}, \ldots, x_{k-m+1}$ are very close to each other

$$\nabla F(x_k) = \sum_{i=1}^{m} \nabla f_i(x_k) \approx \sum_{i=1}^{m} \nabla f_i(x_{k-i+1}).$$

(Cyclic) incremental averaged gradient (IAG)

$$x_{k+1} = \operatorname{prox}_{\gamma_k R}\Big(x_k - \gamma_k \sum_i \nabla f_{\operatorname{mod}(k-i+1,m)}(x_{k-i-1})\Big).$$

- Perturbation error

  $$\epsilon_k = \sum_i \nabla f_{\operatorname{mod}(k-i+1,m)}(x_{k-i-1}) - \sum_i \nabla f_i(x_k)$$

  is vanishing.
- Guaranteed convergence under constant step-size, or $\inf_k \gamma_k > 0$.
- Increased memory cost: store the previous gradients.
- Almost no extra computational cost.

Assume F is quadratic, denote $j_{k,i} = \mathrm{mod}(k - i + 1, m)$

$$\begin{aligned}\sum_i \nabla f_i(x_k) - \sum_i \nabla f_{j_{k,i}}(x_{k-i-1}) &= \sum_i \left(\nabla f_{j_{k,i}}(x_k) - \nabla f_{j_{k,i}}(x_{k-i-1})\right) \\ &= \sum_i \nabla^2 f_{j_{k,i}}(x_k - x_{k-i-1}).\end{aligned}$$

Back to the incremental scheme

$$\begin{aligned}x_{k+1} &= \mathrm{prox}_{\gamma_k R}\left(x_k - \gamma_k \sum_i \nabla f_{j_{k,i}}(x_{k-i-1})\right) \\ &= \mathrm{prox}_{\gamma_k R}\left(x_k + \gamma_k \sum_i \nabla^2 f_{j_{k,i}}(x_k - x_{k-i-1}) - \gamma_k \nabla F(x_k)\right)\end{aligned}$$

- For each $j_{k,i}$, $\gamma_k \nabla^2 f_{j_{k,i}}(x_k - x_{k-i-1})$ can be treated as a generalisation of momentum.
- IAG is a multi-step inertial scheme.
- In practice, IAG has slight edge over proximal gradient.

Outline

Averaged finite sum

$$\min_{x \in \mathbb{R}^n} F(x) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} f_i(x).$$

Incremental gradient descent sample i_k randomly from $\{1, \ldots, m\}$,

$$x_{k+1} = x_k - \gamma_k \nabla f_{i_k}(x_k).$$

- $g_k = \nabla f_{i_k}(x_k)$ can be viewed as a **stochastic gradient**.
- $g_k = \nabla F(x_k) + \epsilon_k$, where ϵ_k is zero mean, *i.e.* $\mathbb{E}[\epsilon_k] = 0$.

- Index i_k is chosen uniformly from $\{1, \ldots, m\}$.
- Hence in expectation, $g = \nabla f_{i_k}(x)$

$$\mathbb{E}[g] = \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(x) = \nabla F(x).$$

- In terms of perturbation error

$$\mathbb{E}[g - \nabla F(x)] = \mathbb{E}[\epsilon_k] = 0.$$

- g is called an unbiased estimation of $\nabla F(x)$, otherwise biased estimation. Difference between the two...
- Otaining g in two steps
  - Pick i_k uniformly from $\{1, \ldots, m\}$, or based on certain distribution.
  - Compute the stochastic gradient based on i_k.

Consider minimising

$$\min_{x \in \mathbb{R}^n} F(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x).$$

Stochastic gradient desent sample i_k randomly from $\{1, \ldots, m\}$,

$$x_{k+1} = x_k - \gamma_k \nabla f_{i_k}(x_k).$$

Behaviour is similar to subgradient descent even for smooth F. Convergence properties?

- Define $\Delta_k = \|x_k - x^\star\|^2$ and $e_k = \mathbb{E}[\Delta_k]$.
- Note that x_k depends on $i_j, j = 1, ..., k-1$, and does not depend on i_k.
- Bounding Δ_k, denote $g_k = \nabla f_{i_k}(x_k)$

$$\Delta_{k+1} = \|x_k - x^\star - \gamma_k g_k\|^2 = \Delta_k + \gamma_k^2 \|g_k\|^2 - 2\gamma_k \langle g_k,\, x_k - x^\star \rangle.$$

- Assume $\Delta_k \leq C$. Eg $\min_{x \in \mathcal{X}} F(x)$.
- Taking expection

$$e_{k+1} \leq e_k + \gamma_k^2 C^2 - 2\gamma_k \mathbb{E}\big[\langle g_k,\, x_k - x^\star \rangle\big].$$

- Since x_k is independent of i_k

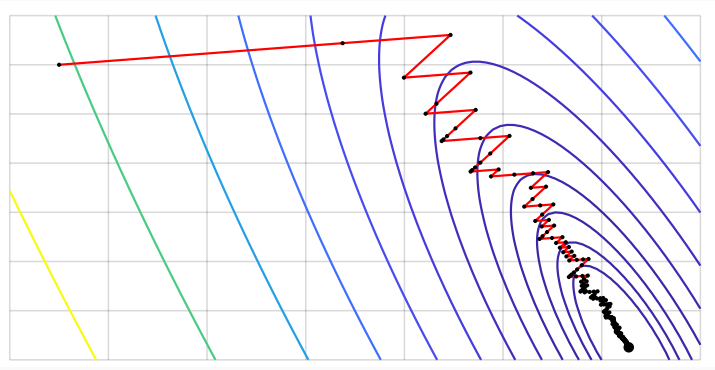
$$\begin{aligned}\mathbb{E}\big[\langle g_k,\, x_k - x^\star \rangle\big] &= \mathbb{E}\Big[\mathbb{E}\big[\langle g_k,\, x_k - x^\star \rangle\big] | i_1, ..., i_{k-1}\Big] \\ &= \mathbb{E}\big[\langle \nabla F(x_k),\, x_k - x^\star \rangle\big].\end{aligned}$$

- Need to bound $\mathbb{E}\big[\langle \nabla F(x_k),\, x_k - x^\star \rangle\big]$.

- Since F is convex,
$$F(x) \geq F(x_k) + \langle \nabla F(x_k),\, x - x_k \rangle.$$
- Let $x = x^\star$, then
$$2\gamma_k \mathbb{E}\big[F(x^\star) - F(x_k)\big] \geq 2\gamma_k \mathbb{E}\big(\langle \nabla F(x_k),\, x^\star - x_k \rangle\big).$$
- Put back into e_{k+1}
$$\begin{aligned} & e_{k+1} \leq e_k + \gamma_k^2 C^2 - 2\gamma_k \mathbb{E}\big[\langle g_k,\, x_k - x^\star \rangle\big] \\ \implies \quad & 2\gamma_k \mathbb{E}\big[\langle g_k,\, x_k - x^\star \rangle\big] \leq e_k - e_{k-1} + \gamma_k^2 C^2 \\ \implies \quad & 2\gamma_k \mathbb{E}\big[F(x_k) - F(x^\star)\big] \leq e_k - e_{k-1} + \gamma_k^2 C^2 \end{aligned}$$
- Sum over $k = 1, \ldots, T$
$$\textstyle\sum_{k=1}^T 2\gamma_k \mathbb{E}\big[F(x_k) - F(x^\star)\big] \leq e_0 - e_{T+1} + C^2 \sum_{k=1}^T \gamma_k^2 \leq e_0 + C^2 \sum_{k=1}^T \gamma_k^2.$$
- Devide the above inequality by $\sum_k \gamma_k$ and let $w_k = \frac{\gamma_k}{\sum_k \gamma_k}$. Then
$$\mathbb{E}\Big[\textstyle\sum_{k=1}^T \gamma_k F(x_k) - F(x^\star)\Big] \leq \dfrac{e_0 + C^2 \sum_{k=1}^T \gamma_k^2}{2 \sum_k \gamma_k}.$$

- Very similar to the case of subgradient descent with vanishing step-size.
- Denote the averaged point
$$\bar{x}_T = \sum_{k=1}^{T} w_k x_k.$$
- Owing to convexity, we have
$$F(\bar{x}_T) \leq \sum_{k=1}^{T} w_k F(x_k).$$
- As a result
$$\mathbb{E}\left[F(\bar{x}_T) - F(x^\star)\right] \leq \frac{e_0 + C^2 \sum_{k=1}^{T} \gamma_k^2}{2 \sum_k \gamma_k}.$$
- Choice of γ_k
$$\gamma_k = \frac{c}{k^\delta},\ \delta \in]1/2, 1].$$

Problem with SGD



Trajectory of SGD

- Non-descent method.
- Vanishing step-size.
- Slow rate of convergence.

Finite sum with regularisation

$$\min_{x\in\mathbb{R}^n} \Phi(x) \stackrel{\text{def}}{=} R(x) + \frac{1}{m}\sum_i f_i(x).$$

Proximal SGD: sample i_k randomly from $\{1, \ldots, m\}$,

$$x_{k+1} = \text{prox}_{\gamma_k R}\big(x_k - \gamma_k \nabla f_{i_k}(x_k)\big).$$

- The proximal mapping of R is fully computed.

Same convergence behaviour as SGD:

$$\begin{aligned}\|x_{k+1} - x^\star\|^2 &= \|\text{prox}_{\gamma_k R}(x_k - \gamma_k g_k) - \text{prox}_{\gamma_k R}(x^\star - \gamma_k \nabla F(x^\star))\|^2 \\ &\leq \|x_k - x^\star - \gamma_k g_k\|^2 \\ &= \Delta_k + \gamma_k^2\|g_k\|^2 - 2\gamma_k\langle g_k,\, x_k - x^\star\rangle.\end{aligned}$$

Outline

Perturbation error of SGD:

$$\mathbb{E}[\epsilon_k] = 0 \quad \text{and} \quad \mathbb{E}[\|\epsilon_k\|^2] \leq C(x_k).$$

Vanishing step-size

$$\begin{aligned}\|x_{k+1} - x^*\| &= \|x_k - \gamma_k \nabla F(x_k) - x^* + \gamma_k \nabla F(x^*) + \gamma_k \epsilon_k\| \\ &\leq \|x_k - x^*\| + \gamma_k \|\epsilon_k\|.\end{aligned}$$

Goal: $\gamma_k \|\epsilon_k\| \to 0$

- SGD: $\gamma_k \to 0$.
- Variance reduction $\|\epsilon_k\| \to 0$.

How to: naively, randomise the Incremental Averaged Gradient...

Randomised gradient history

- Initialisation

$$H_0 = \left[\nabla f_1(x_0), \cdots, \nabla f_m(x_0)\right].$$

- Update

$$i_k \text{ randomly from } \{1, \ldots, m\} \text{ and let } g_k = \nabla f_{i_k}(x_k),$$

$$H_k(j) = \begin{cases} g_k : \text{if } j = i_k, \\ H_{k-1}(j) : \text{o.w.} \end{cases}$$

Stochastic averaged gradient (SAG)

$$i_k \text{ randomly from } \{1, \dots, m\} \text{ and let } g_k = \nabla f_{i_k}(x_k),$$

$$x_{k+1} = \text{prox}_{\gamma_k R}\big(x_k - \gamma_k \frac{g_k - H_{k-1}(i_k)}{m} - \gamma_k \frac{1}{m} \sum_{j=1}^{m} H_k(j)\big),$$

$$H_k(j) = \begin{cases} g_k : \text{if } j = i_k, \\ H_{k-1}(j) : \text{o.w.} \end{cases}$$

- Same as IAG, allows constant step-size.
- Biased stochastic gradient estimation

$$\mathbb{E}[\epsilon_k] = \frac{1}{m}\mathbb{E}[g_k] - \nabla F(x_k) = \frac{m-1}{m}\nabla F(x_k).$$

- Convergence proof extremely complicated, due to the biased estimation.

SAGA

$$i_k \text{ randomly from } \{1, \ldots, m\} \text{ and let } g_k = \nabla f_{i_k}(x_k),$$
$$x_{k+1} = \text{prox}_{\gamma_k R}\big(x_k - \gamma_k(g_k - H_{k-1}(i_k)) - \gamma_k \frac{1}{m}\sum_{j=1}^{m} H_k(j)\big),$$
$$H_k(j) = \begin{cases} g_k : \text{if } j = i_k, \\ H_{k-1}(j) : \text{o.w.} \end{cases}$$

- Allows constant step-size.
- Un-biased stochastic gradient estimation

$$\mathbb{E}[\epsilon_k] = \mathbb{E}[g_k] - \nabla F(x_k) = 0.$$

- Convergence rate: let $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$ and $\gamma_k \equiv \frac{1}{3L}$
  - Convex case: $\mathbb{E}[\Phi(\bar{x}_k) - \Phi(x^\star)] = O(1/k)$.
  - α-strongly convex case: $\mathbb{E}[\Phi(\bar{x}_k) - \Phi(x^\star)] = O(\eta^k)$ with $\eta = 1 - \min\left\{\frac{1}{4m}, \frac{\alpha}{3L}\right\}$.

Let P be a positive integer, for $\ell = 0, 1, 2, \cdots$

$$
\begin{aligned}
&\tilde{g}_\ell = \tfrac{1}{m}\textstyle\sum_{i=1}^{m} \nabla f_i(\tilde{x}_\ell), x_{\ell,0} = \tilde{x}_\ell,\\
&\text{For } p = 1, \cdots, P\\
&\quad \text{sample } i_p \text{ uniformly from } \{1, \cdots, m\},\\
&\quad w_k = x_{\ell,p-1} - \gamma_k(\nabla f_{i_p}(x_{\ell,p-1}) - \nabla f_{i_p}(\tilde{x}_\ell) + \tilde{g}_\ell),\\
&\quad x_{\ell,p} = \operatorname{prox}_{\gamma_k R}(w_k).\\
&\text{Option I} : \tilde{x}_{\ell+1} = x_{\ell,P},\\
&\text{Option II} : \tilde{x}_{\ell+1} = \tfrac{1}{P}\textstyle\sum_{p=1}^{P} x_{\ell,p}.
\end{aligned}
$$

- In practice, $P = m, 2m, \ldots$.
- Given $x_{\ell,p}$, denote $k = \ell P + p$ such that $x_k = x_{\ell,p}$. Then

$$\epsilon_k = \nabla f_{i_p}(x_k) - \nabla f_{i_p}(\tilde{x}_\ell) + \tilde{g}_\ell - \nabla F(x_k),$$

  which is unbiased.

Comparison between SGD, SAG/SAGA and SVRG

- Convergence rate:
  - SGD: $O(1/\sqrt{k})$.
  - Variance reduction: $O(1/k)$.
- SAG/SAGA requires extra memory to store the gradient history.
- SVRG does not need extra memory cost.
- Gradient evaluation at each step
  - SGD, SAG/SAGA: 1.
  - SVRG: 3 if choosing $P = m$.

Outline

Let $(h_i, l_i) \in \mathbb{R}^n \times \{\pm 1\}$, $i = 1, \cdots, m$ be the training set, where $l_i \in \mathbb{R}^n$ is the feature vector of each data sample, and l_i is the binary label.

LASSO The formulation of ℓ_1-regularised LSE

$$\min_{x \in \mathbb{R}^n} \frac{\mu}{2}\|x\|_1 + \frac{1}{m}\sum_{i=1}^m \|h_i^T x - l_i\|^2,$$

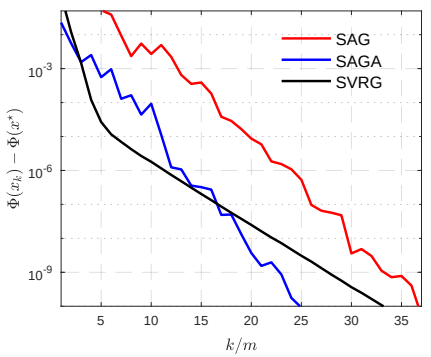
where $\mu > 0$ is a trade-off parameter.

Logistic regression The formulation of ℓ_2-regularised LR

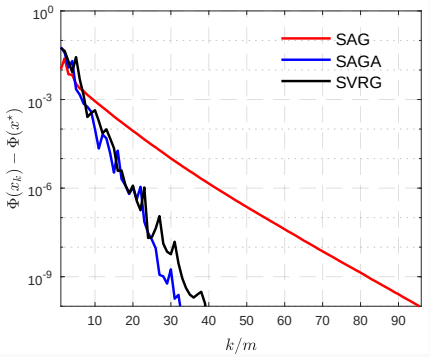
$$\min_{x \in \mathbb{R}^n} \frac{\mu}{2}\|x\|_2^2 + \frac{1}{m}\sum_{i=1}^m \log\left(1 + e^{-l_i h_i^T x}\right),$$

where $\mu > 0$ is a trade-off parameter.

Numerical result



LASSO



LR

Reference

- D. Blatt, A. O. Hero, and H. Gauchman. “A convergent incremental gradient method with a constant step-size”. SIAM Journal on Optimization, 18(1):29?51, 2007.
- M. Schmidt, N. Le Roux, and F. Bach. “Minimizing finite sums with the stochastic average gradient”. Mathematical Programming, 162(1-2):83?112, 2017.
- A. Defazio, F. Bach, and S. Lacoste-Julien. “SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives”. In Advances in Neural Information Processing Systems, pages 1646?1654, 2014.
- L. Xiao and T. Zhang. “A proximal stochastic gradient method with progressive variance reduction”. SIAM Journal on Optimization, 24(4):2057?2075, 2014.