Optimization for ML: Exercise / Practical Session 3

1 Stochastic gradient descent for finite sums

This section introduces the concept of importance sampling for the optimization of finite sums. Let $f : \mathbb{R}^p \to \mathbb{R}$ be decomposed as $f(\theta) = \frac{1}{n}\sum_{i=1}^n f_i(\theta)$. We assume that f is μ-strongly convex and that each f_i is continuously differentiable and L_i-smooth. As usual, we denote θ_* the global minimizer of f and $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\theta_*)\|^2$.

3. We consider stochastic gradient descent: choose $\theta_0 \in \mathbb{R}^p$ and for all $k \in \mathbb{N}$, sample $i_{k+1} \sim \mathrm{Unif}(\{1, \dots, n\})$ independently of the past and compute

$$\theta_{k+1} = \theta_k - \gamma \nabla f_{i_{k+1}}(\theta_k)\,.$$

Show that, for some appropriate choice of the stepsize γ to be determined, the iteration complexity to obtain $\mathbb{E}\|\theta_k - \theta_*\|^2 \leqslant \varepsilon$ is $k = O\left(\left(\log \frac{\|\theta_0-\theta_*\|^2}{\varepsilon}\right)\left(\max_j \frac{L_j}{\mu} + \frac{\sigma^2}{\mu^2\varepsilon}\right)\right)$.

Again, this dependence in $\max_j \frac{L_j}{\mu}$ is tight. However, one can improve this dependence by using importance sampling. Let $\pi = (\pi_1, \dots, \pi_n)$ denote a probability distribution on $\{1, \dots, n\}$. Choose $\theta_0 \in \mathbb{R}^p$ and for all $k \in \mathbb{N}$, sample $i_{k+1} \sim \pi$ independently of the past and compute

$$\theta_{k+1} = \theta_k - \frac{\gamma}{\pi_{i_{k+1}}} \nabla f_{i_{k+1}}(\theta_k)\,.$$

Finally, denote $\overline{L} = \frac{1}{n}\sum_{i=1}^n L_i$.

4. In this question, we take $\pi_i = \frac{L_i}{n\overline{L}}$. Show that, for some appropriate choice of the stepsize γ to be determined, the iteration complexity to obtain $\mathbb{E}\|\theta_k - \theta_*\|^2 \leqslant \varepsilon$ is $k = O\left(\left(\log \frac{\|\theta_0-\theta_*\|^2}{\varepsilon}\right)\left(\frac{\overline{L}}{\mu} + \frac{\overline{L}}{\min_i L_i}\frac{\sigma^2}{\mu^2\varepsilon}\right)\right)$.

In the iteration complexity, we have improved the dependence of the first term from the worst condition number $\max_i \frac{L_i}{\mu}$ to the average condtion number $\frac{\overline{L}}{\mu}$. This can bring a potentially large improvement, especially when ε is large. However, the second term was worsened by a factor $\frac{\overline{L}}{\min_i L_i}$. This can be harmful when ε is small or σ^2 is large.

5. In this question, we take $\pi_i = \frac{1}{2n}\left(1 + \frac{L_i}{\overline{L}}\right)$. Show that, for some appropriate choice of the stepsize γ to be determined, the iteration complexity to obtain $\mathbb{E}\|\theta_k - \theta_*\|^2 \leqslant \varepsilon$ is $k = O\left(\left(\log \frac{\|\theta_0-\theta_*\|^2}{\varepsilon}\right)\left(\frac{\overline{L}}{\mu} + \frac{\sigma^2}{\mu^2\varepsilon}\right)\right)$.

Partially biasing the sampling allows to enjoy the best of both worlds.

2 Simulations: the least-squares case

Consider the minimization of a least-squares function of the form

$$f(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left(\langle x_i, \theta\rangle - y_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)\,, \qquad f_i(\theta) = \frac{1}{2}\left(\langle x_i, \theta\rangle - y_i\right)^2\,,$$

where $(x_1, y_1), \ldots, (x_n, y_n)$ are given input-output pairs.

6. We denote $X \in \mathbb{R}^{n\times p}$ the design matrix whose rows are $x_1, \ldots, x_n$. Under which condition on X is f strongly convex? If this condition holds, what is the associated strong convexity parameter?

7. Give the minimal value L_i such that f_i is L_i-smooth.

8. We now run simulations with $n = 10^3$ and $p = 10$, in the two following cases:

- **(a)** $x_1, \ldots, x_n \sim_{\text{i.i.d.}} \mathcal{N}(0, I_p)$, $\theta_0 \sim \mathcal{N}(0, I_p)$, $\varepsilon_1, \ldots, \varepsilon_n \sim_{i.i.d.} \mathcal{N}(0, 0.1^2)$ are all independent, and $y_i = \langle x_i, \theta_0\rangle + \varepsilon_i$,
- **(b)** $x_1, \ldots, x_{n-1} \sim_{\text{i.i.d.}} \mathcal{N}(0, I_p)$, $x_n \sim_{\text{i.i.d.}} \mathcal{N}(0, 10^2 I_p)$, $\theta_0 \sim \mathcal{N}(0, I_p)$, $\varepsilon_1, \ldots, \varepsilon_n \sim_{i.i.d.} \mathcal{N}(0, 0.1^2)$ are all independent, and $y_i = \langle x_i, \theta_0\rangle + \varepsilon_i$,

For each one of these cases, generate a function f according to the specified distribution and compare the performance of plain, weighted and partially weighted stochastic gradient descent by plotting the logarithm of the distance to optimum as a function of k. For each algorithm, choose γ either (1) as large as possible, so that the algorithm remains stable or (2) so that it is the same for all algorithms. (This gives a total of $2 \times 2 = 4$ plots with three algorithms on each plot).

1 Stochastic gradient descent for finite sums

This section introduces the concept of importance sampling for the optimization of finite sums. Let $f : \mathbb{R}^p \to \mathbb{R}$ be decomposed as $f(\theta) = \frac{1}{n}\sum_{i=1}^n f_i(\theta)$. We assume that f is μ-strongly convex and that each f_i is continuously differentiable and L_i-smooth. As usual, we denote θ_* the global minimizer of f and $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\theta_*)\|^2$.

3. We consider stochastic gradient descent: choose $\theta_0 \in \mathbb{R}^p$ and for all $k \in \mathbb{N}$, sample $i_{k+1} \sim \mathrm{Unif}(\{1, \ldots, n\})$ independently of the past and compute

$$\theta_{k+1} = \theta_k - \gamma \nabla f_{i_{k+1}}(\theta_k)\,.$$

Show that, for some appropriate choice of the stepsize γ to be determined, the iteration complexity to obtain $\mathbb{E}\|\theta_k - \theta_*\|^2 \leqslant \varepsilon$ is $k = O\left(\left(\log \frac{\|\theta_0-\theta_*\|^2}{\varepsilon}\right)\left(\max_j \frac{L_j}{\mu} + \frac{\sigma^2}{\mu^2\varepsilon}\right)\right)$.

---

(3) soit $g(\theta, i) = \nabla f_i(\theta)$, $i \sim \mathrm{Unif}(\{1, \ldots, n\})$

$$E[g(\theta,i)] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\theta) = \nabla f(\theta)$$

(i) f est μ-fortement convexe

(ii) f est M-lisse

$$\begin{aligned}
E\|g(\theta,i) - g(\theta',i)\|^2 &= E\|\nabla f_i(\theta) - \nabla f_i(\theta')\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{p} \|\nabla f_i(\theta) - \nabla f_i(\theta')\|^2 \\
&\leqslant \frac{1}{n}\sum_{i=1}^{p} \langle \nabla f_i(\theta) - \nabla f_i(\theta'), L_i \cdot (\theta - \theta')\rangle \\
&\leqslant \max_{i=1,\ldots,p} L_i \cdot \left\langle \frac{1}{n}\sum_{i=1}^{p} \nabla f_i(\theta) - \frac{1}{n}\sum_{i=1}^{p}\nabla f_i(\theta'), \theta - \theta'\right\rangle \\
&= \max_{i=1,\ldots,p} L_i \cdot \langle \nabla f(\theta) - \nabla f(\theta'), \theta - \theta'\rangle
\end{aligned}$$

donc $M = \max_{i=1,\ldots,p} L_i$

(iii) $E\|g(\theta_*, i)\|^2 = E\|\nabla f_i(\theta_*)\|^2 = \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(\theta_*)\|^2 = \sigma^2$

Par le corollaire du cours, on choisit $\gamma = \dfrac{1}{2M + \frac{4\sigma^2}{\varepsilon\mu}}$

日期: /

alors on obtient $E\|\theta_k-\theta_*\|^2 \leqslant \varepsilon$

avec $k \geqslant 2\cdot\left(\log\frac{2\|\theta_0-\theta_*\|^2}{\varepsilon}\right)\cdot\left(\frac{M}{\mu}+\frac{2\sigma^2}{\mu^2\varepsilon}\right)$

donc $k = O\left(\left(\log\frac{\|\theta_0-\theta_*\|^2}{\varepsilon}\right)\cdot\left(\max_{j=1,\dots,p}\frac{L_j}{\mu}+\frac{\sigma^2}{\mu^2\varepsilon}\right)\right)$

Again, this dependence in $\max_j \frac{L_j}{\mu}$ is tight. However, one can improve this dependence by using importance sampling. Let $\pi = (\pi_1, \dots, \pi_n)$ denote a probability distribution on $\{1, \dots, n\}$. Choose $\theta_0 \in \mathbb{R}^p$ and for all $k \in \mathbb{N}$, sample $i_{k+1} \sim \pi$ independently of the past and compute

$$\theta_{k+1} = \theta_k - \frac{\gamma}{\pi_{i_{k+1}}} \nabla f_{i_{k+1}}(\theta_k)\,.$$

Finally, denote $\overline{L} = \frac{1}{n}\sum_{i=1}^n L_i$.

4. In this question, we take $\pi_i = \frac{L_i}{n\overline{L}}$. Show that, for some appropriate choice of the stepsize γ to be determined, the iteration complexity to obtain $\mathbb{E}\|\theta_k - \theta_*\|^2 \leqslant \varepsilon$ is $k = O\left(\left(\log \frac{\|\theta_0-\theta_*\|^2}{\varepsilon}\right)\left(\frac{\overline{L}}{\mu} + \frac{\overline{L}}{\min_i L_i}\frac{\sigma^2}{\mu^2\varepsilon}\right)\right)$.

(4) soit $g(\theta,i) = \frac{1}{n\pi_i}\cdot\nabla f_i(\theta)$, $i=1,\dots,p$

$E[g(\theta,i)] = \sum_{i=1}^n \pi_i\cdot\frac{1}{n\pi_i}\cdot\nabla f_i(\theta) = \frac{1}{n}\sum_{i=1}^n\nabla f_i(\theta) = \nabla f(\theta)$

(i) f est μ-fortement convexe

(ii) f est M-lisse

$$\begin{aligned}
E\|g(\theta,i)-g(\theta',i)\|^2 &= \sum_{i=1}^n \pi_i\cdot\left\|\frac{1}{n\pi_i}\nabla f_i(\theta) - \frac{1}{n\pi_i}\cdot\nabla f_i(\theta')\right\|^2 \\
&= \frac{1}{n^2}\sum_{i=1}^n\frac{1}{\pi_i}\cdot\|\nabla f_i(\theta)-\nabla f_i(\theta')\|^2 \\
&= \frac{\overline{L}}{n}\sum_{i=1}^n\frac{1}{L_i}\|\nabla f_i(\theta)-\nabla f_i(\theta')\|^2 \\
&\leqslant \frac{\overline{L}}{n}\sum_{i=1}^n\langle\nabla f_i(\theta)-\nabla f_i(\theta'), \theta-\theta'\rangle \\
&= \overline{L}\cdot\langle\nabla f(\theta)-\nabla f(\theta'), \theta-\theta'\rangle
\end{aligned}$$

日期: /

(iii) $\sigma^2_{biais} = E\|g(\theta_*, i)\|^2 = \sum_{i=1}^{n} \lambda_i \cdot \left\|\frac{1}{n\lambda_i} \cdot \nabla f_i(\theta_*)\right\|^2$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{\lambda_i} \|\nabla f_i(\theta_*)\|^2 = \frac{\bar{L}}{n} \cdot \sum_{i=1}^{n} \frac{1}{L_i} \|\nabla f_i(\theta_*)\|^2$$

$$\leq \frac{\bar{L}}{\min_i L_i} \cdot \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(\theta_*)\|^2$$

$$= \frac{\bar{L}}{\min_i L_i} \cdot \sigma^2$$

Par le corollaire, on choisit $\gamma' = \dfrac{1}{2\bar{L} + \frac{\bar{L}}{\min_i L_i} \cdot \frac{4\sigma^2}{\varepsilon\mu^2}}$

Descent de gradient stochastique à pas constant $\gamma = \frac{\gamma'}{n}$.

$$\theta_{k+1} = \theta_k - \frac{\gamma}{\lambda_{i_{k+1}}} \cdot \nabla f_{i_{k+1}}(\theta_k)$$

$$= \theta_k - \frac{n \cdot \gamma}{n \cdot \lambda_{i_{k+1}}} \cdot \nabla f_{i_{k+1}}(\theta_k)$$

$$= \theta_k - \gamma' \cdot g(\theta_k, i_{k+1})$$

on obtient $E\|\theta_k - \theta_*\|^2 \leq \varepsilon$

quand $k \geq 2 \cdot \log\left(\frac{2 \cdot \|\theta_0 - \theta_*\|^2}{\varepsilon}\right) \cdot \left(\frac{\bar{L}}{\mu} + \frac{\bar{L}}{\min_i L_i} \cdot \frac{2\sigma^2}{\mu^2 \varepsilon}\right)$

donc $k = O\left(\log\left(\frac{\|\theta_0 - \theta_*\|^2}{\varepsilon}\right) \cdot \left(\frac{\bar{L}}{\mu} + \frac{\bar{L}}{\min_i L_i} \cdot \frac{\sigma^2}{\mu^2 \varepsilon}\right)\right)$

In the iteration complexity, we have improved the dependence of the first term from the worst condition number $\max_i \frac{L_i}{\mu}$ to the average condtion number $\frac{\overline{L}}{\mu}$. This can bring a potentially large improvement, especially when ε is large. However, the second term was worsened by a factor $\frac{\overline{L}}{\min_i L_i}$. This can be harmful when ε is small or σ^2 is large.

5. In this question, we take $\pi_i = \frac{1}{2n}\left(1 + \frac{L_i}{\overline{L}}\right)$. Show that, for some appropriate choice of the stepsize γ to be determined, the iteration complexity to obtain $\mathbb{E}\|\theta_k - \theta_*\|^2 \leqslant \varepsilon$ is $k = O\left(\left(\log \frac{\|\theta_0-\theta_*\|^2}{\varepsilon}\right)\left(\frac{\overline{L}}{\mu} + \frac{\sigma^2}{\mu^2\varepsilon}\right)\right)$.

Partially biasing the sampling allows to enjoy the best of both worlds.

(5) Soit $g(\theta,i) = \frac{1}{n\lambda_i}\cdot\nabla f_i(\theta)$, $i \sim D(\lambda)$

$$E[g(\theta,i)] = \sum_{i=1}^{n}\lambda_i\cdot\frac{1}{n\lambda_i}\cdot\nabla f_i(\theta) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\theta) = \nabla f(\theta)$$

(i) f est μ-fortement convexe

(ii) f est M-lisse

$$E\|g(\theta,i)-g(\theta',i)\|^2 = \sum_{i=1}^{n}\lambda_i\cdot\frac{1}{n^2\cdot\lambda_i^2}\cdot\|\nabla f_i(\theta)-\nabla f_i(\theta')\|^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{\lambda_i}\cdot\|\nabla f_i(\theta)-\nabla f_i(\theta')\|^2$$

$\lambda_i = \frac{1}{2n}(1+\frac{L_i}{\overline{L}})$

$$\leq \frac{1}{n^2}\sum_{i=1}^{n}\frac{L_i}{\lambda_i}\cdot\langle\nabla f_i(\theta)-\nabla f_i(\theta'),\theta-\theta'\rangle$$

$\lambda_i \geq \frac{1}{2n}\cdot\frac{L_i}{\overline{L}}$

$$\leq \frac{1}{n^2}\sum_{i=1}^{n}2n\overline{L}\cdot\langle\nabla f_i(\theta)-\nabla f_i(\theta'),\theta-\theta'\rangle$$

$2n\overline{L} \geq \frac{L_i}{\lambda_i}$

$$= 2\overline{L}\cdot\langle\nabla f(\theta)-\nabla f(\theta'),\theta-\theta'\rangle$$

donc $M = 2\overline{L}$

(iii) $$\sigma^2_{biais} = E\|g(\theta_*,i)\|^2 = \sum_{i=1}^{n}\lambda_i\cdot\frac{1}{n^2\lambda_i^2}\cdot\|\nabla f_i(\theta_*)\|^2$$

$$\leq \frac{1}{n^2}\cdot 2n\overline{L}\cdot\sum_{i=1}^{n}\frac{1}{L_i}\cdot\|\nabla f_i(\theta_*)\|^2$$

$$\leq \frac{2\overline{L}}{\min_i L_i}\cdot\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\theta_*)\|^2$$

$$= \frac{2\overline{L}}{\min_i L_i}\cdot\sigma^2$$

$$\leq 2\sigma^2$$

日期: /

Par le corollaire du cours, on choisit $\gamma' = \dfrac{1}{4\bar{L} + \frac{8\sigma^2}{\varepsilon\mu^2}}$

Descent de gradient stochastique à pas constant $\gamma = \frac{\gamma'}{n}$

$$\begin{aligned}\theta_{k+1} &= \theta_k - \frac{\gamma}{\lambda_{i_{k+1}}} \cdot \nabla f_{i_{k+1}}(\theta_k) \\ &= \theta_k - \frac{n \cdot \gamma}{n \cdot \lambda_{i_{k+1}}} \cdot \nabla f_{i_{k+1}}(\theta_k) \\ &= \theta_k - \gamma' \cdot g(\theta_k, i_{k+1})\end{aligned}$$

on obtient $\mathbb{E}\|\theta_k - \theta_*\|^2 \leq \varepsilon$

quand $k \geq 2 \cdot \left(\log \frac{2 \cdot \|\theta_0 - \theta_*\|^2}{\varepsilon}\right) \cdot \left(\frac{2\bar{L}}{\mu} + \frac{4\sigma^2}{\varepsilon\mu^2}\right)$

donc $k = O\left(\left(\log \frac{\|\theta_0 - \theta_*\|^2}{\varepsilon}\right) \cdot \left(\frac{\bar{L}}{\mu} + \frac{\sigma^2}{\varepsilon\mu^2}\right)\right)$

2 Simulations: the least-squares case

Consider the minimization of a least-squares function of the form

$$f(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(\langle x_i, \theta\rangle - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta), \qquad f_i(\theta) = \frac{1}{2}(\langle x_i, \theta\rangle - y_i)^2,$$

where $(x_1, y_1), \ldots, (x_n, y_n)$ are given input-output pairs.

6. We denote $X \in \mathbb{R}^{n\times p}$ the design matrix whose rows are $x_1, \ldots, x_n$. Under which condition on X is f strongly convex? If this condition holds, what is the associated strong convexity parameter?

(6) $f(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$ $\qquad f_i(\theta) = \frac{1}{2}(\langle x_i, \theta\rangle - y_i)^2$

$\nabla^2 f(\theta) = \frac{1}{n}\sum_{i=1}^{n}\nabla^2 f_i(\theta)$

$\mathbb{R}^p \ni \nabla f_i(\theta) = (\langle\theta, x_i\rangle - y_i)\, x_i$

$\mathbb{R}^{p\times p} \ni \nabla^2 f_i(\theta) = x_i^T x_i$

$X^T X \in \mathbb{R}^{p\times p} \qquad X \in \mathbb{R}^{n\times p}$

$$X^T X = [x_1 | \cdots | x_n]\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} x_i^T x_i$$

$$\nabla^2 f(\theta) = \frac{1}{n}\sum_{i=1}^{n}\nabla^2 f_i(\theta) = \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i = \frac{1}{n}X^T X$$

f est fortement convexe

$\iff \exists \mu > 0$, f est μ-fortement convexe

$\iff \exists \mu > 0,\ \forall \theta \in \mathbb{R}^p,\ \nabla^2 f(\theta) \succcurlyeq \mu I_p$

$\iff \exists \mu > 0,\ \frac{1}{n}X^T X \succcurlyeq \mu I_p$

$\iff X^T X \succ 0$

$\iff \text{rang}(X) = p$

f est μ-fortement convexe

$\Longleftrightarrow \frac{1}{n} X^T X \succcurlyeq \mu I_p$

$\Longleftrightarrow \frac{1}{n} \lambda_{min}(X^T X) \geqslant \mu$

donc f est $\frac{1}{n} \lambda_{min}(X^T X)$ — fortement convexe.

7. Give the minimal value L_i such that f_i is L_i-smooth.

(7) $\nabla^2 f_i(\theta) = x_i^T x_i$, soit $L_i > 0$

f_i est L_i — lisse $\Longleftrightarrow \forall \theta \in \mathbb{R}^p, \nabla^2 f_i(\theta) \preccurlyeq L_i \cdot I_p$

$\Longleftrightarrow x_i^T x_i \preccurlyeq L_i \cdot I_p$

$\Longleftrightarrow \|x_i\|^2 \leqslant L_i$

donc la valeur minimale de L_i telle que f_i soit L_i — lisse est $\|x_i\|^2$.