Optimization for ML: Exercise Session 2

Exercise 1 (tightness and statistical optimality)**.** The goal of this exercise is to explore whether the bounds proven in the lectures are tight and statistically optimal.

Let ξ be a random variable with some distribution $\mathcal{Q}$ with a finite second moment and define $f : \mathbb{R} \to \mathbb{R}$,

$$f(\theta) = \frac{1}{2}\mathbb{E}\,(\xi - \theta)^2 \ .$$

1. Compute the unique minimizer θ_* of f and $f(\theta_*)$.

2. Express a stochastic gradient descent on f that does not have a direct access to $\mathcal{Q}$ but only to i.i.d. samples $\xi_1, \xi_2, \dots \sim \mathcal{Q}$.

3. We consider the stochastic gradient descent with constant stepsize γ.

(a) Using a theorem of the lectures, bound $\mathbb{E}(\theta_k - \theta_*)^2$.

(b) Compute $\mathbb{E}(\theta_k - \theta_*)^2$ exactly and compare with the bound obtained in the previous question.

4. We consider the stochastic gradient descent with variable stepsize $\gamma_k = \beta/(k_0 + k)$.

(a) Using a theorem of the lectures, bound $\mathbb{E}(\theta_k - \theta_*)^2$.

(b) Assume $\beta = 1$ and $k_0 = 1$. Express θ_k as a function of $\xi_1, \dots, \xi_k$. Compute $\mathbb{E}(\theta_k - \theta_*)^2$.

The empirical average is an optimal (minimax) estimator of the mean; stochastic gradient descent is said to be statistically optimal on this problem as its performance differs only by a multiplicative constant.

This exercise motivates the use of decaying stepsizes $\gamma = \Theta(1/k)$ and that it is hopeless to obtain a better rate than $\Theta(1/k)$ in our general setting.

Exercise 2 (importance sampling for coordinate gradient descent)**.** Let $F : \mathbb{R}^p \to \mathbb{R}$ be a continuously differentiable, μ-strongly fonction for some $\mu > 0$. As usual, we denote θ_* the global minimizer of F. We assume that F is $(L_1, \dots, L_p)$-smooth, in the sense that

$$\forall \theta, \theta' \in \mathbb{R}^p, \quad F(\theta') \leqslant F(\theta) + \langle \nabla F(\theta), \theta' - \theta \rangle + \frac{1}{2}\sum_{j=1}^{p} L_j(\theta'(j) - \theta(j))^2 \,.$$

The notion of $(L_1, \dots, L_p)$-smoothness refines the notion of L-smoothness by allowing different curvatures along the different coordinates.

1. In this question, we consider the coordinate gradient descent algorithm: choose $\theta_0 \in \mathbb{R}^p$ and for all $k \in \mathbb{N}$, sample $j_{k+1} \sim \mathrm{Unif}(\{1, \dots, p\})$ independently of the past and compute θ_{k+1} such that

$$\theta_{k+1}(j_{k+1}) = \theta_k(j_{k+1}) - \gamma \partial_{j_{k+1}} F(\theta_k) \,, \tag{1}$$

$$\theta_{k+1}(j) = \theta_k(j) \,, \qquad j \neq j_{k+1} \,, \tag{2}$$

where we choose the stepsize $\gamma = \frac{1}{2 \max_j L_j}$.

(a) Using a result from the lectures, show that the iteration complexity to obtain $\mathbb{E}\|\theta_k - \theta_*\|^2 \leqslant \varepsilon$ is $k = O\left(\left(\log \frac{\|\theta_0-\theta_*\|^2}{\varepsilon}\right) p \max_j \frac{L_j}{\mu}\right)$.

We now show that this upper bound on the iteration complexity is tight. Consider $p \geqslant 2$ and the function $H(\theta) = \frac{1}{2}\sum_{j=1}^{p} L_j \theta(j)^2$.

(b) Show that H is $(L_1, \ldots, L_p)$-smooth and μ-strongly convex with $\mu = \min_j L_j$.

(c) Denote $j_{\min} = \operatorname{argmin}_j L_j$. When $\theta_0 = e_{j_{\min}}$ is the $j_{\min}$-th element of the canonical basis, show that the coordinate gradient descent (1)–(2) on $F = H$ satisfies

$$\mathbb{E}\|\theta_k - \theta_*\|^2 \geqslant \left(1 - \frac{\mu}{p \max_j L_j}\right)^k \|\theta_0 - \theta_*\|^2 .$$

(d) Conclude that in this case, the iteration complexity to obtain $\mathbb{E}\|\theta_k - \theta_*\|^2 \leqslant \varepsilon$ is $k = \Omega\left(\left(\log \frac{\|\theta_0-\theta_*\|^2}{\varepsilon}\right) p \max_j \frac{L_j}{\mu}\right)$.

2. The goal of this question is to show that the iteration complexity of stochastic gradient descent can be improved by an appropriate weighted sampling of the coordinates.

We consider the following weighted generalization of the coordinate gradient descent method. Let $\pi = (\pi_1, \ldots, \pi_p)$ denote a probability distribution on $\{1, \ldots, p\}$. Choose $\theta_0 \in \mathbb{R}^p$ and for all $k \in \mathbb{N}$, sample $j_{k+1} \sim \pi$ independently of the past and compute θ_{k+1} such that

$$\begin{aligned} \theta_{k+1}(j_{k+1}) &= \theta_k(j_{k+1}) - \gamma_{j_{k+1}} \partial_{j_{k+1}} F(\theta_k) , \\ \theta_{k+1}(j) &= \theta_k(j) , \qquad j \neq j_{k+1} , \end{aligned}$$

where $\gamma_1, \ldots, \gamma_p$ are now coordinate-dependent stepsizes.

(a) Prove that, if $\gamma_j \propto \pi_j^{-1}$, the weighted coordinate gradient descent is a stochastic gradient descent in the sense of the lectures.

(b) Show that for all $\theta, \theta' \in \mathbb{R}^p$,

$$\sum_{j=1}^{p} \frac{1}{L_j} \left(\partial_j F(\theta) - \partial_j F(\theta')\right)^2 \leqslant \langle \theta - \theta', \nabla F(\theta) - \nabla F(\theta') \rangle .$$

(c) Consider a weighted coordinate gradient descent with weights $\pi_j = \frac{L_j}{\sum_{j'} L_{j'}}$. Show that, for some appropriate choice of the stepsizes $\gamma_1, \ldots, \gamma_p$ to be determined, the iteration complexity to obtain $\mathbb{E}\|\theta_k - \theta_*\|^2 \leqslant \varepsilon$ is

$$k = O\left(\left(\log \frac{\|\theta_0 - \theta_*\|^2}{\varepsilon}\right) \sum_j \frac{L_j}{\mu}\right) .$$

Importance sampling improved the dependence in the worst of the condition numbers $\max_j \frac{L_j}{\mu}$ to the average of the condition numbers $\frac{1}{p}\sum_j \frac{L_j}{\mu}$.

Exercise 1 (tightness and statistical optimality). The goal of this exercise is to explore whether the bounds proven in the lectures are tight and statistically optimal.

Let ξ be a random variable with some distribution Q with a finite second moment and define $f : \mathbb{R} \to \mathbb{R}$,

$$f(\theta) = \frac{1}{2}\mathbb{E}\left(\xi - \theta\right)^2 .$$

1. Compute the unique minimizer θ_* of f and $f(\theta_*)$.

(1) $f(\theta) = \frac{1}{2} E[\xi^2] + \frac{1}{2}\theta^2 - \theta \cdot E[\xi]$

par la règle de Fermat, $\nabla f(\theta^*) = \theta^* - E[\xi] = 0$

alors $\theta^* = E[\xi]$

$f(\theta^*) = \frac{1}{2} E[(\xi - E[\xi])^2] = \frac{1}{2} \mathrm{Var}[\xi]$

2. Express a stochastic gradient descent on f that does not have a direct access to Q but only to i.i.d. samples $\xi_1, \xi_2, \cdots \sim Q$.

(2) On est dans un cas d'approximation stochastique,

$E[f(\theta,\xi)] = f(\theta) = E[\frac{1}{2}(\xi-\theta)^2]$

$f(\theta,\xi) = \frac{1}{2}(\xi-\theta)^2$

$g(\theta,\xi) = \frac{\partial}{\partial\theta} f(\theta,\xi) = \theta - \xi$

Descent de gradient stochastique

$\theta_0 \in \mathbb{R}$, pour tout $k \in \mathbb{N}$, on prend $\xi_{k+1} \sim Q$, $\gamma_k \in \mathbb{R}$

$\theta_{k+1} = \theta_k - \gamma_k \cdot g(\theta_k, \xi_{k+1})$

$= \theta_k - \gamma_k \cdot (\theta_k - \xi_{k+1})$

$= (1-\gamma_k)\cdot\theta_k + \gamma_k \cdot \xi_{k+1}$

3. We consider the stochastic gradient descent with constant stepsize γ.

(a) Using a theorem of the lectures, bound $\mathbb{E}(\theta_k - \theta_*)^2$.

(3) (a) $\frac{\partial^2}{\partial\theta^2} f(\theta, \xi) = 1 \leq 1$

donc $f(\theta, \xi)$ est M-lisse, avec $M = 1$

$\nabla^2 f(\theta) = 1 \geq 1$

donc $f(\theta)$ est μ-fortement convexe, avec $\mu = 1$

et

$$\sigma^2 = E|g(\theta_*, \xi)|^2 = E[(\xi - E[\xi])^2] = Var[\xi]$$

D'après le Thm. 4.17, Si $\gamma_k = \gamma < \frac{1}{2M} = \frac{1}{2}$, alors

$$E(\theta_k - \theta_*)^2 \leq (1 - \gamma\mu)^k \cdot |\theta_0 - \theta_*|^2 + \frac{2\gamma\sigma^2}{\mu}$$
$$= (1-\gamma)^k \cdot |\theta_0 - \theta_*|^2 + 2\gamma \cdot Var[\xi]$$

(b) Compute $\mathbb{E}(\theta_k - \theta_*)^2$ exactly and compare with the bound obtained in the previous question.

(b) θ_k est un estimateur stochastique de θ_*

$$E(\theta_k - \theta_*)^2 = b(\theta_k)^2 + Var[\theta_k]$$

· $b(\theta_k)^2$

$$b(\theta_k) = E[\theta_k] - \theta_*$$

$$b(\theta_{k+1}) = E[\theta_{k+1}] - \theta_* = E[(1-\gamma)\theta_k + \gamma \cdot \xi_{k+1}] - \theta_*$$
$$= (1-\gamma) \cdot E[\theta_k] + \gamma \cdot \theta_* - \theta_*$$
$$= (1-\gamma) \cdot (E[\theta_k] - \theta_*)$$
$$= (1-\gamma) \cdot b(\theta_k)$$

日期: /

alors $b(\theta_k)^2 = (1-\gamma)^{2k} \cdot b(\theta_0)^2 = (1-\gamma)^{2k} \cdot |\theta_0 - \theta_*|^2$

- $Var[\theta_k]$

$$Var[\theta_{k+1}] = Var[(1-\gamma)\cdot\theta_k + \gamma\cdot\xi_{k+1}]$$
$$= (1-\gamma)^2 \cdot Var[\theta_k] + \gamma^2 \cdot Var[\xi]$$

soit $$V = (1-\gamma)^2 \cdot V + \gamma^2 \cdot Var[\xi]$$
$$V = \frac{\gamma^2 \cdot Var[\xi]}{1-(1-\gamma)^2} = \frac{\gamma \cdot Var[\xi]}{2-\gamma}$$

et
$$Var[\theta_{k+1}] - V = (1-\gamma)^2 \cdot (Var[\theta_k] - V)$$
alors
$$Var[\theta_k] - V = (1-\gamma)^{2k} \cdot (\underbrace{Var[\theta_0]}_{=0} - V)$$
$$Var[\theta_k] = V - (1-\gamma)^{2k} \cdot V$$
$$= (1-(1-\gamma)^{2k}) \cdot \frac{\gamma \cdot Var[\xi]}{2-\gamma}$$

Donc
$$E(\theta_k - \theta_*)^2 = (1-\gamma)^{2k} \cdot |\theta_0 - \theta_*|^2 + (1-(1-\gamma)^{2k}) \cdot \frac{\gamma \cdot Var[\xi]}{2-\gamma}$$

4. We consider the stochastic gradient descent with variable stepsize $\gamma_k = \beta/(k_0 + k)$.

(a) Using a theorem of the lectures, bound $\mathbb{E}(\theta_k - \theta_*)^2$.

(4) (a) D'après (3)(a)

(i) $f(\theta,\xi)$ est M-lisse, $M = 1$

(ii) $f(\theta)$ est μ-fortement convexe, $\mu = 1$

(iii) $\sigma^2 = E|g(\theta_*, \xi)|^2 = Var[\xi]$

Par la théorème 4.19, si $\beta > \frac{1}{\mu}$ et $\gamma_0 = \frac{\beta}{k_0} \leq \frac{1}{2M}$, alors

$$E(\theta_k - \theta_*)^2 \leq \frac{\nu}{k_0 + k}$$

avec $\nu = \max\left\{ k_0 \|\theta_0 - \theta_*\|^2, \frac{2\sigma^2\beta^2}{\beta\mu - 1} \right\}$

(b) Assume $\beta = 1$ and $k_0 = 1$. Express θ_k as a function of $\xi_1, \ldots, \xi_k$. Compute $\mathbb{E}(\theta_k - \theta_*)^2$.

The empirical average is an optimal (minimax) estimator of the mean; stochastic gradient descent is said to be statistically optimal on this problem as its performance differs only by a multiplicative constant.

This exercise motivates the use of decaying stepsizes $\gamma = \Theta(1/k)$ and that it is hopeless to obtain a better rate than $\Theta(1/k)$ in our general setting.

(b) $\beta = 1$, $k_0 = 1 \implies \gamma_k = \frac{1}{k+1}$

$$\theta_{k+1} = \theta_k - \gamma_k \cdot g(\theta_k, \xi_k) = \theta_k - \frac{\theta_k - \xi_k}{k+1} = \frac{k}{k+1} \cdot \theta_k + \frac{1}{k+1} \cdot \xi_k$$

Pour montrer $\theta_k = \frac{\xi_1 + \cdots + \xi_k}{k}$

Récurrence

- $\theta_1 = 0 \cdot \theta_0 + \xi_1 = \xi_1$

日期:　　　/

- soit $\theta_k = \frac{\xi_1 + \cdots + \xi_k}{k}$

$$\theta_{k+1} = \frac{k}{k+1} \cdot \frac{\xi_1 + \cdots + \xi_k}{k} + \frac{\xi_{k+1}}{k+1} = \frac{\xi_1 + \cdots + \xi_{k+1}}{k+1}$$

Donc $\theta_k = \frac{\xi_1 + \cdots + \xi_k}{k}$

$$\begin{aligned} E(\theta_k - \theta_*)^2 &= E\left[\left(\frac{\xi_1 + \cdots + \xi_k}{k} - E[\xi]\right)^2\right] \\ &= E\left[\left(\frac{\xi_1 + \cdots + \xi_k}{k} - E\left[\frac{\xi_1 + \cdots + \xi_k}{k}\right]\right)^2\right] \\ &= \mathrm{Var}\left[\frac{\xi_1 + \cdots + \xi_k}{k}\right] \\ &= \frac{\sigma^2}{k} \end{aligned}$$

Exercise 2 (importance sampling for coordinate gradient descent). Let $F : \mathbb{R}^p \to \mathbb{R}$ be a continuously differentiable, μ-strongly fonction for some $\mu > 0$. As usual, we denote θ_* the global minimizer of F. We assume that F is $(L_1, \ldots, L_p)$-smooth, in the sense that

$$\forall \theta, \theta' \in \mathbb{R}^p, \quad F(\theta') \leqslant F(\theta) + \langle \nabla F(\theta), \theta' - \theta \rangle + \frac{1}{2} \sum_{j=1}^{p} L_j (\theta'(j) - \theta(j))^2 .$$

The notion of $(L_1, \ldots, L_p)$-smoothness refines the notion of L-smoothness by allowing different curvatures along the different coordinates.

1. In this question, we consider the coordinate gradient descent algorithm: choose $\theta_0 \in \mathbb{R}^p$ and for all $k \in \mathbb{N}$, sample $j_{k+1} \sim \mathrm{Unif}(\{1, \ldots, p\})$ independently of the past and compute θ_{k+1} such that

$$\theta_{k+1}(j_{k+1}) = \theta_k(j_{k+1}) - \gamma \partial_{j_{k+1}} F(\theta_k) , \qquad (1)$$

$$\theta_{k+1}(j) = \theta_k(j) , \qquad j \neq j_{k+1} , \qquad (2)$$

where we choose the stepsize $\gamma = \frac{1}{2 \max_j L_j}$.

(a) Using a result from the lectures, show that the iteration complexity to obtain $\mathbb{E}\|\theta_k - \theta_*\|^2 \leqslant \varepsilon$ is $k = O\left(\left(\log \frac{\|\theta_0 - \theta_*\|^2}{\varepsilon}\right) p \max_j \frac{L_j}{\mu}\right)$.

1) (a) Par le corollaire, pour $\mathbb{E}\|\theta_k - \theta_*\|^2 \leq \varepsilon$

si $\quad k \geq 2 \cdot \left(\log \frac{2\|\theta_0 - \theta_*\|^2}{\varepsilon} \right) \cdot \frac{Lp}{\mu}$

$\quad \geq 2 \cdot \left(\log \frac{2\|\theta_0 - \theta_*\|^2}{\varepsilon} \right) \cdot p \cdot \max_j \frac{L_j}{\mu}$

donc

$$k = O\left(\left(\log \frac{\|\theta_0 - \theta_*\|^2}{\varepsilon} \right) \cdot p \cdot \max_j \frac{L_j}{\mu} \right)$$

We now show that this upper bound on the iteration complexity is tight. Consider $p \geqslant 2$ and the function $H(\theta) = \frac{1}{2} \sum_{j=1}^{p} L_j \theta(j)^2$.

(b) Show that H is $(L_1, \ldots, L_p)$-smooth and μ-strongly convex with $\mu = \min_j L_j$.

(b) $\forall j = 1, \ldots, p$, soit $H_j(\theta) = \frac{1}{2} L_j \theta(j)^2$

$\forall \theta, \theta' \in \mathbb{R}^p$,

$$|\partial H_j(\theta) - \partial H_j(\theta')| = |L_j \theta(j) - L_j \theta'(j)| \leq L_j \cdot |\theta(j) - \theta'(j)|$$

donc H_j est L_j — lisse

Alors on a

$$H_j(\theta') \leqslant H_j(\theta) + \partial H_j(\theta)\cdot(\theta'_{(j)} - \theta_{(j)}) + \frac{L_j}{2}(\theta'_{(j)} - \theta_{(j)})^2$$

donc

$$\sum_{j=1}^{p} H_j(\theta') \leqslant \sum_{j=1}^{p} H_j(\theta) + \sum_{j=1}^{p} \partial H_j(\theta)\cdot(\theta'_{(j)} - \theta_{(j)}) + \frac{1}{2}\sum_{j=1}^{p} L_j\cdot(\theta'_{(j)} - \theta_{(j)})^2$$

$$H(\theta') \leqslant H(\theta) + \langle \nabla H(\theta), \theta' - \theta \rangle + \frac{1}{2}\sum_{j=1}^{p} L_j\cdot(\theta'_{(j)} - \theta_{(j)})^2$$

alors H est $(L_1, \dots, L_p)$ — lisse

D'après la définition, $H_j(\theta)$ est fortement convexe de module — L_j, donc on a

$$H_j(\theta) \geqslant H_j(\theta') + \partial H_j(\theta)\cdot(\theta'_{(j)} - \theta_{(j)}) + \frac{L_j}{2}\cdot(\theta'_{(j)} - \theta_{(j)})^2$$

alors

$$\sum_{j=1}^{p} H_j(\theta) \geqslant \sum_{j=1}^{p} H_j(\theta') + \sum_{j=1}^{p} \partial H_j(\theta)\cdot(\theta'_{(j)} - \theta_{(j)}) + \sum_{j=1}^{p} \frac{L_j}{2}(\theta'_{(j)} - \theta_{(j)})^2$$

$$H(\theta) \geqslant H(\theta') + \langle \nabla H(\theta), \theta' - \theta \rangle + \frac{1}{2}\sum_{j=1}^{p} L_j\cdot(\theta'_{(j)} - \theta_{(j)})^2$$

$$H(\theta) \geqslant H(\theta') + \langle \nabla H(\theta), \theta' - \theta \rangle + \frac{1}{2}\cdot \min_j L_j \sum_{j=1}^{p}(\theta'_{(j)} - \theta_{(j)})^2$$

donc H est μ—fortement convexe, avec $\mu = \min_j L_j$

(c) Denote $j_{\min} = \operatorname{argmin}_j L_j$. When $\theta_0 = e_{j_{\min}}$ is the $j_{\min}$-th element of the canonical basis, show that the coordinate gradient descent (1)–(2) on $F = H$ satisfies

$$\mathbb{E}\|\theta_k - \theta_*\|^2 \geqslant \left(1 - \frac{\mu}{p \max_j L_j}\right)^k \|\theta_0 - \theta_*\|^2 .$$

(c) Soit $\theta_0 = e_{j_{min}}$, $H(\theta_0) = \frac{1}{2} L_{j_{min}} = \frac{1}{2}\min_j L_j$

pour $j_{k+1} \sim \mathrm{Unif}(\{1, \dots, p\})$

$$\|\theta_{k+1}-\theta^*\|^2 = \|\theta_k - \gamma\cdot\partial_{j_{k+1}}H(\theta_k)\cdot e_{j_{k+1}} - \theta^*\|^2$$
$$= \|\theta_k - \theta^* - \gamma\cdot\partial_{j_{k+1}}H(\theta_k)\cdot e_{j_{k+1}}\|^2$$
$$= \|\theta_k-\theta^*\|^2 - 2\langle\theta_k-\theta^*, \gamma\cdot L_{j_{k+1}}\cdot\theta_k(j_{k+1})\cdot e_{j_{k+1}}\rangle + \gamma^2\cdot L_{j_{k+1}}^2\cdot\theta^2(j_{k+1})$$
$$\geqslant \|\theta_k-\theta^*\|^2 - \langle\theta_k-\theta^*, 2\cdot\gamma\cdot L_{j_{k+1}}\cdot\theta_k(j_{k+1})\cdot e_{j_{k+1}}\rangle$$

donc

$$E\left[\|\theta_{k+1}-\theta^*\|^2\right] \geqslant E\left[\|\theta_k-\theta^*\|^2 - \langle\theta_k-\theta^*, 2\cdot\gamma\cdot L_{j_{k+1}}\cdot\theta_k(j_{k+1})\cdot e_{j_{k+1}}\rangle\right]$$
$$= E\left[\|\theta_k-\theta^*\|^2\right] - \frac{2\gamma}{p}\cdot E\left[\langle\theta_k-\theta^*, \nabla H(\theta_k)-\nabla H(\theta^*)\rangle\right]$$
$$\geqslant E\left[\|\theta_k-\theta^*\|^2\right] - \frac{2\gamma}{p}\cdot\mu\cdot E\left[\|\theta_k-\theta^*\|^2\right]$$
$$= \left(1-\frac{2\cdot\mu}{p\cdot\max_j L_j}\right)\cdot E\left[\|\theta_k-\theta^*\|^2\right]$$

alors

$$E\|\theta_k-\theta^*\|^2 \geqslant \left(1-\frac{2\cdot\mu}{p\cdot\max_j L_j}\right)^k\cdot\|\theta_0-\theta^*\|^2$$

(d) Conclude that in this case, the iteration complexity to obtain $\mathbb{E}\|\theta_k-\theta_*\|^2 \leqslant \varepsilon$ is $k = \Omega\left(\left(\log\frac{\|\theta_0-\theta_*\|^2}{\varepsilon}\right)p\max_j\frac{L_j}{\mu}\right)$.

(d) Pour avoir $E\|\theta_k-\theta^*\|^2 \leqslant \varepsilon$, d'après (c), on a

$$\left(1-\frac{2\cdot\mu}{p\cdot\max_j L_j}\right)^k\cdot\|\theta_0-\theta^*\|^2 \leqslant \varepsilon$$
$$\left(1-\frac{2\cdot\mu}{p\cdot\max_j L_j}\right)^k \leqslant \frac{\varepsilon}{\|\theta_0-\theta^*\|^2}$$
$$k\cdot\log\left(1-\frac{2\cdot\mu}{p\cdot\max_j L_j}\right) \leqslant \log\frac{\varepsilon}{\|\theta_0-\theta^*\|^2}$$
$$k \geqslant \log\left(\frac{\varepsilon}{\|\theta_0-\theta^*\|^2}\right)\cdot\frac{1}{\log\left(1-\frac{2\cdot\mu}{p\cdot\max_j L_j}\right)}$$
$$k \geqslant -\log\left(\frac{\varepsilon}{\|\theta_0-\theta^*\|^2}\right)\cdot\frac{p\cdot\max_j L_j}{2\mu}$$
$$k \geqslant \log\left(\frac{\|\theta_0-\theta^*\|^2}{\varepsilon}\right)\cdot\frac{p}{2}\cdot\max_j\frac{L_j}{\mu}$$

2. The goal of this question is to show that the iteration complexity of stochastic gradient descent can be improved by an appropriate weighted sampling of the coordinates.

We consider the following weighted generalization of the coordinate gradient descent method. Let $\pi = (\pi_1, \dots, \pi_p)$ denote a probability distribution on $\{1, \dots, p\}$. Choose $\theta_0 \in \mathbb{R}^p$ and for all $k \in \mathbb{N}$, sample $j_{k+1} \sim \pi$ independently of the past and compute θ_{k+1} such that

$$\theta_{k+1}(j_{k+1}) = \theta_k(j_{k+1}) - \gamma_{j_{k+1}} \partial_{j_{k+1}} F(\theta_k)\,,$$
$$\theta_{k+1}(j) = \theta_k(j)\,, \qquad j \neq j_{k+1}\,,$$

where $\gamma_1, \dots, \gamma_p$ are now coordinate-dependent stepsizes.

(a) Prove that, if $\gamma_j \propto \pi_j^{-1}$, the weighted coordinate gradient descent is a stochastic gradient descent in the sense of the lectures.

(a) Pour $F : \mathbb{R}^p \longrightarrow \mathbb{R}$, $\xi = j \sim D(\pi)$

$g(\theta, \xi) = g(\theta, j) = \frac{1}{\pi_j} \cdot \partial_j F(\theta) \cdot e_j$

$E[g(\theta, \xi)] = \nabla F(\theta)$

· Descent de gradient stochastique

$\theta_0 \in \mathbb{R}^p$ et $\forall k \in \mathbb{N}$, on prend $\xi_{k+1} \sim D(\pi)$

soit $\gamma_k = \gamma = c$,

$$\begin{aligned}\theta_{k+1} &= \theta_k - \gamma \cdot g(\theta_k, \xi_{k+1}) \\ &= \theta_k - c \cdot \frac{1}{\pi_{j_{k+1}}} \cdot \partial_{j_{k+1}} F(\theta) \cdot e_{j_{k+1}} \\ &= \theta_k - \gamma_{k+1} \cdot \partial_{j_{k+1}} F(\theta) \cdot e_{j_{k+1}}\end{aligned}$$

(b) Show that for all $\theta, \theta' \in \mathbb{R}^p$,

$$\sum_{j=1}^{p} \frac{1}{L_j} \left(\partial_j F(\theta) - \partial_j F(\theta')\right)^2 \leqslant \langle \theta - \theta', \nabla F(\theta) - \nabla F(\theta') \rangle\,.$$

日期: /

(b) $\forall \theta, \theta', \theta'' \in \mathbb{R}^p$,

$$F(\theta+\theta'') \leqslant F(\theta) + \langle \nabla F(\theta), \theta'' \rangle + \frac{1}{2}\sum_{j=1}^{p} L_j \cdot \theta''(j)^2$$

$$F(\theta+\theta'') \geqslant F(\theta') + \langle \nabla F(\theta'), \theta+\theta''-\theta' \rangle$$

alors

$$F(\theta') + \langle \nabla F(\theta'), \theta+\theta''-\theta' \rangle \leqslant F(\theta) + \langle \nabla F(\theta), \theta'' \rangle + \frac{1}{2}\sum_{j=1}^{p} L_j \cdot \theta''(j)^2$$

$$\langle \nabla F(\theta') - \nabla F(\theta), \theta'' \rangle - \frac{1}{2}\sum_{j=1}^{p} L_j \cdot \theta''(j)^2 \leqslant F(\theta) - F(\theta') - \langle \nabla F(\theta'), \theta-\theta' \rangle$$

par la règle de Fermat, on optim en $\theta''(j)$

$$\forall j=1,\dots,p, \qquad \theta''(j) = \frac{1}{L_j}(\partial_j F(\theta') - \partial_j F(\theta))$$

donc

$$\frac{1}{2}\sum_{j=1}^{p} \frac{1}{L_j} \cdot (\partial_j F(\theta') - \partial_j F(\theta))^2 \leqslant F(\theta) - F(\theta') - \langle \nabla F(\theta'), \theta-\theta' \rangle$$

on inverse θ et θ'

$$\frac{1}{2}\sum_{j=1}^{p} \frac{1}{L_j} \cdot (\partial_j F(\theta) - \partial_j F(\theta'))^2 \leqslant F(\theta') - F(\theta) - \langle \nabla F(\theta), \theta'-\theta \rangle$$

et on somme

$$\sum_{j=1}^{p} \frac{1}{L_j} \cdot (\partial_j F(\theta') - \partial_j F(\theta))^2 \leqslant \langle \nabla F(\theta') - \nabla F(\theta), \theta'-\theta \rangle$$

(c) Consider a weighted coordinate gradient descent with weights $\pi_j = \frac{L_j}{\sum_{j'} L_{j'}}$. Show that, for some appropriate choice of the stepsizes $\gamma_1, \dots, \gamma_p$ to be determined, the iteration complexity to obtain $\mathbb{E}\|\theta_k - \theta_*\|^2 \leqslant \varepsilon$ is

$$k = O\left(\left(\log \frac{\|\theta_0 - \theta_*\|^2}{\varepsilon}\right) \sum_j \frac{L_j}{\mu}\right).$$

Importance sampling improved the dependence in the worst of the condition numbers $\max_j \frac{L_j}{\mu}$ to the average of the condition numbers $\frac{1}{p}\sum_j \frac{L_j}{\mu}$.

日期: /

(c) (i) F est μ-fortement convexe.

(ii) $g(\theta,\xi)$ est M-lisse

$$E\|g(\theta,j)-g(\theta',j)\|^2 = \sum_{j=1}^{p} \lambda_j \cdot \left\| \frac{1}{\lambda_j}\partial_j F(\theta)\cdot e_j - \frac{1}{\lambda_j}\partial_j F(\theta')\cdot e_j \right\|^2$$

$$= \sum_{j=1}^{p} \frac{1}{\lambda_j}\cdot\left(\partial_j F(\theta)-\partial_j F(\theta')\right)^2$$

$$= \left(\sum_{j'=1}^{p} L_{j'}\right)\cdot \sum_{j=1}^{p}\frac{1}{L_j}\left(\partial_j F(\theta)-\partial_j F(\theta')\right)^2$$

$$\leq \left(\sum_{j'=1}^{p} L_{j'}\right)\cdot\langle \theta-\theta', \nabla F(\theta)-\nabla F(\theta')\rangle$$

donc $M = \sum_{j=1}^{p} L_j$

(iii) $\sigma^2 = E\|g(\theta_*,\xi)\|^2 = \sum_{j=1}^{p}\lambda_j\cdot\left\|\frac{1}{\lambda_j}\cdot\partial_j F(\theta_*)\cdot e_j\right\|^2 = 0$

D'après le corollaire du cours, il existe un choix de $\gamma = c$ tel que pour $k \geq O\left(\left(\log\frac{\|\theta_0-\theta_*\|^2}{\varepsilon}\right)\cdot\frac{M}{\mu}\right)$ alors $E\|\theta_k-\theta_*\|^2 \leq \varepsilon$

donc

$$k = O\left(\left(\log\frac{\|\theta_0-\theta_*\|^2}{\varepsilon}\right)\cdot\sum_{j=1}^{p}\frac{L_j}{\mu}\right)$$

日期:　　　/