+0/1/60+

Exam Optimization for Machine Learning – CS-439
Prof. Martin Jaggi

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

20 June 2019 - from 08h15 to 11h15 in PO01

ID

STUDENT NAME

SCIPER : **SCIPER** Signature :

Wait for the start of the exam before turning to the next page. This document is printed double sided, 16 pages.

- This is a closed book exam. No electronic devices of any kind.
- Place on your desk: your student ID, writing utensils, one double-sided A4 page cheat sheet (handwritten or 11pt min font size) if you have one; place all other personal items below your desk or on the side.
- You each have a different exam.
- For technical reasons, **do use black or blue pens for the MCQ part, no pencils!** Use white corrector if necessary.

Respectez les consignes suivantes | Observe this guidelines | Beachten Sie bitte die unten stehenden Richtlinien		
choisir une réponse | select an answer Antwort auswählen	ne PAS choisir une réponse | NOT select an answer NICHT Antwort auswählen	Corriger une réponse | Correct an answer Antwort korrigieren
ce qu'il ne faut **PAS** faire | what should **NOT** be done | was man **NICHT** tun sollte		

First part, multiple choice

There is **exactly one** correct answer per question.

Lasso Coordinate Descent

The optimization problem for sparse least squares linear regression (also known as the Lasso) is given by

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|A\mathbf{x}-\mathbf{b}\|^2 + \lambda\|\mathbf{x}\|_1$$

for some regularization parameter $\lambda > 0$.

We write A_{-i} for the $(d-1)\times n$ matrix obtained by removing the i-th column A_i from A, and same for the vector $\mathbf{x}_{-i}$ with one entry removed accordingly. The soft thresholding operator S is defined as

$$S_a(b) := \begin{cases} 0, & |b| \le a, \\ b-a & b > a, \\ b+a & b < -a \end{cases}.$$

Question 1 The solution to exact coordinate minimization for the Lasso problem above, for the i-th coordinate, is

- ☐ $x_i^\star = S_{\frac{\lambda}{\|A_i\|^2}}\left(A_i^\top(\mathbf{b}-A_{-i}\mathbf{x}_{-i})/\|A_i\|^2\right)$
- ☐ $x_i^\star = S_{\frac{\lambda}{\|A_i\|^2}}\left(A_i^\top(\mathbf{b}-A\mathbf{x})/\|A_i\|^2\right)$
- ☐ $x_i^\star = S_{\frac{\lambda/2}{\|A_i\|^2}}\left(2A_i^\top(\mathbf{b}-A\mathbf{x})/\|A_i\|^2\right)$
- ☐ $x_i^\star = S_{\frac{\lambda/2}{\|A_i\|^2}}\left(2A_i^\top(\mathbf{b}-A_{-i}\mathbf{x}_{-i})/\|A_i\|^2\right)$
- ■ $x_i^\star = S_{\frac{\lambda/2}{\|A_i\|^2}}\left(A_i^\top(\mathbf{b}-A_{-i}\mathbf{x}_{-i})/\|A_i\|^2\right)$

Hint: If you don't recall the precise expression, verify a concrete example with a toy matrix A and a large value of λ.

Stochastic Gradient Descent

In this section we are interested in finding the minimum of a *strongly convex* function $f\colon \mathbb{R}^n \to \mathbb{R}$,

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}),$$

with iterative schemes of the form

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \mathbf{g}(\mathbf{x}_t),$$

for *gradient oracles* $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^n$.

Question 2 Given access to a gradient oracle $\mathbf{g}_G\colon \mathbb{R}^n \to \mathbb{R}^n$, with $\mathbf{g}_G(\mathbf{x}) := \nabla f(\mathbf{x})$, $\forall \mathbf{x}\in\mathbb{R}^n$, we can implement gradient descent (with constant stepsize $\gamma_t \equiv \gamma$). What is the convergence rate of gradient descent (with optimal stepsize), i.e. how many iterations T does it take to reach suboptimality $f(\mathbf{x}_T) - f^\star \le \varepsilon$?

- ■ no answer is correct
- ☐ $T = \mathcal{O}(\log \frac{1}{\varepsilon})$
- ☐ $T = \mathcal{O}(\log\log \frac{1}{\varepsilon})$
- ☐ $T = \mathcal{O}(e^{\varepsilon})$

Question 3 Given access to a stochastic gradient oracle $\mathbf{g}_{\mathrm{SG}} \colon \mathbb{R}^n \to \mathbb{R}^n$ we can implement stochastic gradient descent on f. Assume the stochastic oracle outputs

$$\mathbf{g}_{\mathrm{SG}}(\mathbf{x}) := \mathbf{g}_{\mathrm{G}} + \boldsymbol{\xi}$$

for every call, where $\boldsymbol{\xi} \in \mathbb{R}^n$ is a random variable with $\mathbb{E}\,\boldsymbol{\xi} = \mathbf{0}$, and $\mathbb{E}\,\|\boldsymbol{\xi}\|^2 \leq \sigma^2$ (and $\sigma^2 > 0$). What is the convergence rate of stochastic gradient descent (with optimal constant stepsize $\gamma_t \equiv \gamma$), for the last iterate (not the average iterate), i.e. how many iterations T does it take to reach suboptimality $\mathbb{E}\, f(\mathbf{x}_T) - f^\star \leq \varepsilon$?

- [x] no answer is correct
- [ ] $T = \mathcal{O}(\frac{1}{\varepsilon})$
- [ ] $T = \mathcal{O}(e^{\varepsilon})$
- [ ] $T = \mathcal{O}(\log \frac{1}{\varepsilon})$

Consider the following two stochastic oracles:

$$\mathbf{g}_{\mathrm{A}}(\mathbf{x}) := \begin{cases} 2\mathbf{g}_{\mathrm{G}}(\mathbf{x}), & \text{w. prob. } \frac{1}{2} \\ \mathbf{0}, & \text{w. prob. } \frac{1}{2} \end{cases} \qquad \mathbf{g}_{\mathrm{B}}(\mathbf{x}) := \begin{cases} \mathbf{g}_{\mathrm{G}}(\mathbf{x}), & \text{w. prob. } \frac{1}{2} \\ \mathbf{g}_{\mathrm{SG}}(\mathbf{x}), & \text{w. prob. } \frac{1}{2} \end{cases}$$

Question 4 Which statement is true? (Here biased means not having the correct expectation)

- [ ] Oracle A and B are both biased.
- [ ] Oracle A is unbiased, oracle B is biased.
- [ ] Oracle A is biased, oracle B is unbiased.
- [x] Oracle A and B are both unbiased.

Question 5 Which statement is true?

- [x] no answer is correct
- [ ] The variance of oracle B is smaller than the variance of oracle A.
- [ ] The variance of oracle A is smaller than the variance of oracle B.

Question 6 Consider two new oracles, $\mathbf{g}_{\mathrm{C}}$ and $\mathbf{g}_{\mathrm{D}}$. Suppose stochastic gradient descent (with constant stepsize γ) converges as:

$$\text{oracle C:} \quad \mathbb{E}\, f(\mathbf{x}_t) - f^\star \leq \left(1 - \frac{a}{100}\right)^t (f(\mathbf{x}_0) - f^\star)$$
$$\text{oracle D:} \quad \mathbb{E}\, f(\mathbf{x}_t) - f^\star \leq (1-a)^t (f(\mathbf{x}_0) - f^\star) + b$$

where here $a \in (0,1)$ and $b > 0$ are two parameters. Which algorithm do you prefer, to reach accuracy ε (in terms of function suboptimality, $\mathbb{E}\, f(\mathbf{x}_t) - f^\star \leq \varepsilon$) as fast as possible? (Assume $f(\mathbf{x}_0) - f^\star \geq 100b$).

- [ ] Both algorithms converge equally fast.
- [x] Oracle D over C if $\varepsilon > 10b$.
- [ ] Oracle D over C if $\varepsilon \leq 10b$.
- [ ] no answer is correct

Convexity and Smoothness

For each of the functions below, verify whether they are (1) convex, (2) strictly convex, (3) strongly convex, and (4) smooth:

A. $f(x) = x,\ x \in \mathbb{R}$

B. $f(x) = \sin(x),\ x \in \mathbb{R}$

C. $f(x) = \mathrm{ReLu}(ax + b),\ x \in \mathbb{R}$

D. $f(\mathbf{x}) = \mathrm{ReLu}(a_2 x_2 (a_1 x_1 + b_1) + b_2),\ \mathbf{x} \in \mathbb{R}^2$

E. $f(x) = e^{-x},\ x \in \mathbb{R}$

F. $f(\mathbf{x}) = \exp(-\mathbf{a}^\top \mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|_2^2,\ \mathbf{x} \in \mathbb{R}^2$

G. $f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x},\ \mathbf{x} \in \mathbb{R}^2,$

where

$$A := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \mathrm{ReLu}(x) := \begin{cases} 0, & x < 0 \\ x, & \text{otherwise} \end{cases}, \qquad a, b, a_i, b_i \in \mathbb{R}, \qquad \mathbf{a}, \mathbf{b} \in \mathbb{R}^2.$$

Question 7 Given the function **A.** above, which are all of its properties?

- [x] convex + smooth
- [ ] convex
- [ ] convex + strictly convex + strongly convex
- [ ] convex + strictly convex
- [ ] convex + strictly convex + smooth
- [ ] smooth
- [ ] convex + strictly convex + strongly convex + smooth
- [ ] none of these properties

Question 8 Given the function **B.** above, which are all of its properties?

- [ ] convex
- [ ] convex + strictly convex
- [ ] convex + strictly convex + strongly convex
- [ ] convex + smooth
- [x] smooth
- [ ] convex + strictly convex + strongly convex + smooth
- [ ] convex + strictly convex + smooth
- [ ] none of these properties

Question 9 Given the function **C.** above, which are all of its properties?

- [ ] convex + strictly convex + smooth
- [x] convex
- [ ] convex + smooth
- [ ] convex + strictly convex + strongly convex
- [ ] convex + strictly convex + strongly convex + smooth
- [ ] smooth
- [ ] convex + strictly convex
- [ ] none of these properties

Question 10 Given the function **D.** above, which are all of its properties?

- [ ] convex + strictly convex
- [ ] convex + strictly convex + strongly convex
- [ ] smooth
- [ ] convex + strictly convex + smooth
- [ ] convex
- [ ] convex + smooth
- [ ] convex + strictly convex + strongly convex + smooth
- [x] none of these properties

Question 11 Given the function **E.** above, which are all of its properties?

- [ ] convex + strictly convex + strongly convex
- [ ] smooth
- [ ] convex + smooth
- [ ] convex
- [ ] convex + strictly convex + strongly convex + smooth
- [x] convex + strictly convex
- [ ] convex + strictly convex + smooth
- [ ] none of these properties

Question 12 Given the function **F.** above, which are all of its properties?

- [ ] convex + strictly convex + strongly convex + smooth
- [ ] smooth
- [x] convex + strictly convex + strongly convex
- [ ] convex + smooth
- [ ] convex + strictly convex + smooth
- [ ] convex
- [ ] convex + strictly convex
- [ ] none of these properties

Question 13 Given the function **G.** above, which are all of its properties?

- [ ] convex + strictly convex + strongly convex
- [ ] convex + strictly convex
- [ ] convex + strictly convex + strongly convex + smooth
- [ ] convex
- [ ] convex + strictly convex + smooth
- [ ] convex + smooth
- [x] smooth
- [ ] none of these properties

Smoothness and Strong Convexity

Consider an iterative optimization procedure.

Question 14 Which one of the following three inequalities is valid for a *smooth* convex function f for some $L \in \mathbb{R}$:

- ☐ $f(\mathbf{x}^\star) - f(\mathbf{x}_t) \leq \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^\star) + \frac{L}{2} \|\mathbf{x}^\star - \mathbf{x}_t\|^2$
- ■ $f(\mathbf{x}^\star) - f(\mathbf{x}_t) \leq \nabla f(\mathbf{x}_t)^\top (\mathbf{x}^\star - \mathbf{x}_t) + \frac{L}{2} \|\mathbf{x}^\star - \mathbf{x}_t\|^2$
- ☐ $f(\mathbf{x}^\star) - f(\mathbf{x}_t) \leq \nabla f(\mathbf{x}_t)^\top (\mathbf{x}^\star - \mathbf{x}_t) - \frac{L}{2} \|\mathbf{x}^\star - \mathbf{x}_t\|^2$

Question 15 Which one of the following three inequalities is valid for a *strongly convex* function f for some $\mu \in \mathbb{R}$:

- ☐ $f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) \geq \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) + \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2$
- ■ $f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) \leq \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) - \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2$
- ☐ $f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) \leq \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) + \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2$

Second part, true/false questions

Question 16 (Linear Minimization Oracle) The LMO used in the Frank-Wolfe algorithm is given as $\mathrm{LMO}_X(\mathbf{g}) := \operatorname*{argmin}_{\mathbf{s}\in X} \langle \mathbf{s}, \mathbf{g}\rangle$. For $X := conv(\mathcal{A})$ being the convex hull of any bounded set $\mathcal{A} \subset \mathbb{R}^d$, we have that

$$\mathrm{LMO}_X(\mathbf{g}) = \mathrm{LMO}_{\mathcal{A}}(\mathbf{g}) \ .$$

■ TRUE ☐ FALSE

Question 17 (Hearn Gap in Frank-Wolfe) The duality gap for constrained optimization problems $\min_{\mathbf{x}\in X} f(\mathbf{x})$ as resulting from the Frank-Wolfe algorithm is

$$g(\mathbf{x}) := \langle \mathbf{s} - \mathbf{x}, \nabla f(\mathbf{x})\rangle \geq f(\mathbf{x}) - f(\mathbf{x}^\star) \ .$$

where $\mathbf{s} = \mathrm{LMO}_X(\nabla f(\mathbf{x}))$ is the output of the Linear Minimization Oracle.

☐ TRUE ■ FALSE

Question 18 (Accelerated Gradient Descent) Accelerated Gradient Descent on an L-smooth and $(\mu > 0)$-strongly convex function f converges as $\mathcal{O}(1/\sqrt{\varepsilon})$.

■ TRUE ☐ FALSE

Question 19 (Accelerated Gradient Descent) Accelerated Gradient Descent on an L-smooth and convex function f converges as $\mathcal{O}(1/\sqrt{\varepsilon})$.

■ TRUE ☐ FALSE

Question 20 (Convexity) A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if its *graph* is a convex set.

☐ TRUE ■ FALSE

Question 21 (Random search) Consider derivative-free random search as discussed in the lecture. For L-smooth convex functions, random search, with step-size $1/L$, converges as $\mathcal{O}(dL/\varepsilon)$

☐ TRUE ■ FALSE

Third part, open questions

Answer in the space provided! Your answer must be justified with all steps. Do not cross any checkboxes, they are reserved for correction.

Importance Sampling for SGD

Consider a smooth sum-structured objective function:

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})\,.$$

The SGD algorithm samples $i \in [n]$ uniformly and sets $\nabla f_i(\mathbf{x}_t)$ to be the stochastic gradient. Sometimes it is possible to speed up SGD by performing *importance sampling.*

Question 22: *2 points.* Consider any probability distribution $\mathbf{p} = (p_1, \ldots, p_n)$ with $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$. We sample i according to distribution $\mathbf{p}$ and define $\mathbf{g}_t$ as:

$$\mathbf{g}_t := \frac{1}{p_i n}\nabla f_i(\mathbf{x}_t)\,. \tag{IS}$$

Then show that $\mathbf{g}_t$ is an unbiased gradient estimator i.e. $\mathbb{E}[\mathbf{g}_t|\mathbf{x}_t] = \nabla f(\mathbf{x}_t)$.

☐0 ☐1 ■2

Solution:

$$\mathbb{E}[\mathbf{g}_t|\mathbf{x}_t] = \sum_{i=1}^{n} p_i \frac{1}{p_i n}\nabla f_i(\mathbf{x}_t) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_t) = \nabla f(\mathbf{x}_t)\,.$$

Question 23: *3 points.* In the same setting as the previous page, recall that the standard simplex is defined as $\Delta_n := \{\mathbf{y} \in \mathbb{R}^n : \sum_{i=1}^n y_i = 1, y_i \geq 0 \ \forall i\}$. For some fixed positive constants $c_i \in \mathbb{R}$ for $i \in [n]$, let $\mathbf{y}^\star$ be the optimum of

$$\mathbf{y}^\star = \operatorname*{argmin}_{\mathbf{y} \in \Delta_n} \left\{ g(\mathbf{y}) := \sum_{i=1}^n \frac{c_i^2}{y_i} \right\}.$$

Using the *first-order optimality condition*, prove that

$$y_i^\star = \frac{|c_i|}{\sum_{j=1}^n |c_j|}, \forall i \in [n].$$

☐ 0 ☐ 1 ☐ 2 ■ 3

Solution: The first-order optimality condition states that if $\mathbf{y}^\star$ is an optimum, then for all $\mathbf{y} \in \Delta_n$,

$$\nabla g(\mathbf{y}^\star)^\top (\mathbf{y} - \mathbf{y}^\star) > 0.$$

The ith coordinate of the gradient at the claimed optimum point is:

$$\nabla_i g(\mathbf{y}^\star) = -\frac{c_i^2}{(y_i^\star)^2} = -\frac{c_i^2}{c_i^2}\left(\sum_{j=1}^n |c_j|\right)^2 = -\left(\sum_{j=1}^n |c_j|\right)^2$$

Substituting the above gradient and $\mathbf{y}^\star$, the optimality conditions becomes:

$$\sum_{i=1}^n -\left(\sum_{j=1}^n |c_j|\right)^2 \left(y_i - \frac{|c_i|}{\sum_{j=1}^n |c_j|}\right) = -\left(\sum_{j=1}^n |c_j|\right)^2 (\|\mathbf{y}\|_1 - 1) = 0.$$

Question 24: *3 points.* Using the previous result, compute the optimum sampling probability $\mathbf{p}^\star$ to minimize the *variance* $\mathbb{E}[\|\mathbf{g}_t - \nabla f(\mathbf{x}_t)\|^2]$ of our estimator $\mathbf{g}_t$ defined in (IS).

☐ 0 ☐ 1 ☐ 2 ■ 3

Solution:

$$\begin{aligned}\mathbb{E}[\|\mathbf{g}_t - \nabla f(\mathbf{x}_t)\|^2] &= \mathbb{E}[\|\mathbf{g}_t\|^2] - \|\nabla f(\mathbf{x}_t)\|^2 \\ &= \sum_{i=1}^n p_i \frac{1}{p_i^2 n^2} \|\nabla f_i(\mathbf{x}_t)\|^2 - \|\nabla f(\mathbf{x}_t)\|^2.\end{aligned}$$

Thus, the optimal sampling distribution to minimize variance is

$$\mathbf{p}^\star = \operatorname*{argmin}_{\mathbf{p} \in \Delta_n} \sum_{i=1}^n \frac{1}{p_i} \|\nabla f_i(\mathbf{x}_t)\|.$$

By the result from the previous question, we know that:

$$\mathbf{p}_i^\star = \frac{\|\nabla f_i(\mathbf{x}_t)\|}{\sum_{j=1}^n \|\nabla f_j(\mathbf{x}_t)\|}.$$

Convergence of Signed Gradient Descent

Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is an L-smooth function. Let us look at an algorithm which only uses the coordinate-wise signs of the gradient, with step-size $\gamma > 0$:

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma sign(\nabla f(\mathbf{x}_t))\,. \tag{sgnGD}$$

Question 25: *3 points.* What is the best step-size γ to use in (sgnGD)?

Hint: plug in the update (sgnGD) *into the smoothness condition and maximize the function decrease.*

☐ 0 ☐ 1 ☐ 2 ■ 3

Solution: Using the smoothness condition,

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\le f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 \\ &= f(\mathbf{x}_t) - \gamma \nabla f(\mathbf{x}_t)^\top sign(\nabla f(\mathbf{x}_t)) + \frac{L\gamma^2}{2}\|sign(\nabla f(\mathbf{x}_t))\|_2^2 \\ &= f(\mathbf{x}_t) - \gamma\|\nabla f(\mathbf{x}_t)\|_1 + \frac{Ld\gamma^2}{2}\,. \end{aligned}$$

The above expression is a quadratic in γ and we can compute the value at which it attains its minimum to be

$$\gamma = \frac{\|\nabla f(\mathbf{x}_t)\|_1}{Ld}\,.$$

Question 26: *3 points.* Suppose that function f has an optimum value $f^\star$ and satisfies the following PL-condition for a constant $\mu_\infty > 0$:

$$\tfrac{1}{2}\|\nabla f(\mathbf{x})\|_1^2 \ge \mu_\infty(f(\mathbf{x}) - f^\star)\ \forall \mathbf{x}\,.$$

Then prove that (sgnGD) with the best step-size γ from the previous question gives the following rate:

$$f(\mathbf{x}_t) - f^\star \le \left(1 - \frac{\mu_\infty}{dL}\right)^t (f(\mathbf{x}_0) - f^\star)\,.$$

☐ 0 ☐ 1 ☐ 2 ■ 3

Solution: Using the above computed $\gamma = \frac{\|\nabla f(\mathbf{x}_t)\|_1}{Ld}$, we get that

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\le f(\mathbf{x}_t) - \gamma\|\nabla f(\mathbf{x}_t)\|_1 + \frac{Ld\gamma^2}{2} \\ &= f(\mathbf{x}_t) - \frac{1}{2Ld}\|\nabla f(\mathbf{x}_t)\|_1^2 \\ &\le f(\mathbf{x}_t) - \frac{\mu_\infty}{Ld}(f(\mathbf{x}_t) - f^\star)\,. \end{aligned}$$

In the last step we use PL inequality. Subtracting $f^\star$ from both sides and rearranging gives the required rate:

$$f(\mathbf{x}_{t+1}) - f^\star \le f(\mathbf{x}_t) - f^\star - \frac{\mu_\infty}{Ld}(f(\mathbf{x}_t) - f^\star) = \left(1 - \frac{\mu_\infty}{dL}\right)(f(\mathbf{x}_t) - f^\star)\,.$$

Coordinate descent vs. Gradient descent

Recall that for a function f, L_c coordinate-wise smoothness is defined as

$$f(\mathbf{x} + \gamma\mathbf{e}_i) \le f(\mathbf{x}) + \gamma\nabla_i f(\mathbf{x}) + \frac{L_c}{2}\gamma^2,\ \forall \mathbf{x} \in \mathbb{R}^d, \forall\gamma \in \mathbb{R}, \forall i \in [d]\,.$$

In contrast, standard (full gradient) smoothness is defined as

$$f(\mathbf{x} + \mathbf{y}) \le f(\mathbf{x}) + \nabla f(\mathbf{x})^\top\mathbf{y} + \frac{L_g}{2}\|\mathbf{y}\|_2^2,\ \forall \mathbf{x} \in \mathbb{R}^d, \forall \mathbf{y} \in \mathbb{R}^d\,.$$

Question 27: *3 points.* Assume that

(a) L_g is the smallest constant such that f is L_g smooth,

(b) L_c is the smallest constant such that f is L_c coordinate-wise smooth,

(c) f is convex.

Prove the following two relations:

$$L_c \le L_g \le dL_c\,.$$

☐ 0 ☐ 1 ☐ 2 ■ 3

Solution: For the first inequality, clearly substituting $\mathbf{y} = \gamma \mathbf{e}_i$ in the full gradient smoothness condition shows that f is also L_g coordinate-wise smooth.
For the second inequality note that

$$\begin{aligned}
f(\mathbf{x}+\mathbf{y}) &= f\left(\tfrac{1}{d}\sum_{i=1}^{d}(\mathbf{x}+dy_i\mathbf{e}_i)\right)\\
&\le \frac{1}{d}\sum_{i=1}^{d} f(\mathbf{x}+dy_i\mathbf{e}_i)\\
&\le \frac{1}{d}\sum_{i=1}^{d}\left\{f(\mathbf{x})+\nabla_i f(\mathbf{x})(dy_i)+\frac{L_c d^2 y_i^2}{2}\right\}\\
&= f(\mathbf{x})+\nabla f(\mathbf{x})^\top \mathbf{y}+\frac{L_c d}{2}\left\|\mathbf{y}\right\|_2^2\,.
\end{aligned}$$

In the first step we used convexity of f.

Question 28: *3 points.* Define the symmetric matrix $A \in \mathbb{R}^{d\times d}$ to be $A := \varepsilon I_d + \mathbf{1}_d \mathbf{1}_d^\top$ where I_d is the identity matrix and $\mathbf{1}_d$ is a vector of all 1s. For some $\mathbf{b} \in \mathbb{R}^d$, consider the quadratic function

$$f(\mathbf{x}) := \tfrac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{b}^\top \mathbf{x}\,. \tag{FQ}$$

Compute the L_c and L_g smoothness constants for f.

☐0 ☐1 ☐2 ■3

Solution: L_g is an upper bound on the the spectral norm of the Hessian. Here the Hessian is A and has a spectral norm of $\varepsilon + d$ i.e. $L_g = \varepsilon + d$.
For L_c note that

$$\begin{aligned}
f(\mathbf{x} + \gamma\mathbf{e}_i) &= \tfrac{1}{2}(\mathbf{x} + \gamma\mathbf{e}_i)^\top A(\mathbf{x} + \gamma\mathbf{e}_i) - \mathbf{b}^\top(\mathbf{x} + \gamma\mathbf{e}_i)\\
&= \tfrac{1}{2}\mathbf{x}^\top A\mathbf{x} + \gamma(A\mathbf{x})_i + \frac{A_{i,i}\gamma^2}{2} - \mathbf{b}^\top\mathbf{x} - \gamma\mathbf{b}_i\\
&= \tfrac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{b}^\top\mathbf{x} + \gamma(A\mathbf{x} - \mathbf{b})_i + \frac{A_{i,i}\gamma^2}{2}\\
&= f(\mathbf{x}) + \gamma\nabla_i f(\mathbf{x}) + \frac{A_{i,i}\gamma^2}{2}\,.
\end{aligned}$$

Thus $L_c = \max_i A_{i,i}$ which in this case is $\varepsilon + 1$.

Question 29: *2 points.* Suppose that performing 1 step of gradient descent on (FQ) requires the same time as performing d steps of coordinate descent. Which algorithm would you expect to converge faster? How would the rates of the two algorithms compare for $\varepsilon \to 0$?

☐0 ☐1 ■2

Solution: Coordinate descent (CD) would be d times faster than gradient descent (GD). This is because GD has a rate proportional to L_g whereas the rate of CD is proportional to dL_c. Since in our example $L_g = dL_c$ for $\varepsilon \to 0$, GD and CD require the same number of iterations. However each iteration of CD is d times faster and so it is overall d times faster.

Smooth non-convex functions

Question 30: *3 points.* Suppose that f is a possibly non-convex, twice differentiable function such that the Hessian is bounded in spectral norm

$$\left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L, \ \forall \mathbf{x}.$$

Show that the function f_L as defined below is convex:

$$f_L(\mathbf{x}) := f(\mathbf{x}) + \frac{L}{2}\|\mathbf{x}\|_2^2 .$$

☐ 0 ☐ 1 ☐ 2 ■ 3

Solution: A very short proof:
the Hessian of f_L is $\nabla^2 f(\mathbf{x}) + LI$. Since the eigenvalues of $\nabla^2 f(\mathbf{x})$ lie in the interval $[-L, L]$, the eigenspectrum of $\nabla^2 f(\mathbf{x}) + LI$ lies in $[0, 2L]$. Thus f_L is convex and $2L$-smooth.
Longer proof:
Since f is possibly non-convex, the eigenvalues of the Hessian may be either positive or negative. The bounded Hessian condition in this case implies that for any $\mathbf{x}, \mathbf{y}, \mathbf{z}$:

$$-L\|\mathbf{x}-\mathbf{y}\|^2 \leq (\mathbf{x}-\mathbf{y})^\top \nabla^2 f(\mathbf{z})(\mathbf{x}-\mathbf{y}) \leq L\|\mathbf{x}-\mathbf{y}\|^2 .$$

Using the mean-value form of the remainder term in Taylor's Theorem, we know that for any $\mathbf{x}, \mathbf{y}$ there exists $\mathbf{z}$ such that

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^\top \nabla^2 f(\mathbf{z})(\mathbf{y}-\mathbf{x}) .$$

Combining the above two together we have:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) - \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2 .$$

Simply expanding the Euclidean norm gives:

$$\begin{aligned}
\frac{L}{2}\|\mathbf{y}\|^2 &= \frac{L}{2}\|\mathbf{x} + (\mathbf{y}-\mathbf{x})\|^2 \\
&= \frac{L}{2}\|\mathbf{x}\|^2 + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2 + L\mathbf{x}^\top(\mathbf{y}-\mathbf{x}) \\
&= \frac{L}{2}\|\mathbf{x}\|^2 + (\nabla \tfrac{L}{2}\|\mathbf{x}\|^2)^\top(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2 .
\end{aligned}$$

Thus we have proved that

$$\begin{aligned}
f_L(\mathbf{y}) &= f(\mathbf{y}) + \frac{L}{2}\|\mathbf{y}\|^2 \\
&\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) - \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2 + \frac{L}{2}\|\mathbf{y}\|^2 \\
&= f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + (\nabla \tfrac{L}{2}\|\mathbf{x}\|^2)^\top(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{x}\|^2 \\
&= f_L(\mathbf{x}) + (\nabla f_L(\mathbf{x}))^\top(\mathbf{y}-\mathbf{x}) .
\end{aligned}$$

In the third step we used the expansion of $\|\mathbf{y}\|^2$. This proves that f_L is convex.

Over-parameterized problems

Suppose that f satisfies the sum structure:

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}),$$

where each of the function f_i is L-smooth. In this question assume we are in the over-parameterized setting which means:

$$\text{there exists } \mathbf{x}^\star \text{ such that } \nabla f_i(\mathbf{x}^\star) = 0 \ \forall i \in [n].$$

We will run standard SGD on this problem by picking i uniformly and updating with some step-size $\gamma > 0$:

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla f_i(\mathbf{x}_t).$$

Question 31: *4 points.* Given that f is over-parameterized, show that

$$\mathbb{E}\left[\, \|\nabla f_i(\mathbf{x}_t)\|^2 \,\middle|\, \mathbf{x}_t\right] \le 2L(f(\mathbf{x}_t) - f(\mathbf{x}^\star)).$$

Hint: use the fact that the gradient of f_i is L-Lipschitz and that it is $\mathbf{0}$ *at* $\mathbf{x}^\star$.

☐ 0 ☐ 1 ☐ 2 ☐ 3 ■ 4

Solution: Since f_i is L-smooth, the following holds for all $\mathbf{y}$:

$$f_i(\mathbf{y}) \le f_i(\mathbf{x}_t) + \nabla f_i(\mathbf{x}_t)^\top(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2.$$

The inequality holds even if we minimize both sides of the above equation giving that

$$\begin{aligned}\min_{\mathbf{y}} f_i(\mathbf{y}) &\le f_i(\mathbf{x}_t) + \min_{\mathbf{y}}\left\{\nabla f_i(\mathbf{x}_t)^\top(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2\right\}\\ &= f_i(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f_i(\mathbf{x}_t)\|^2.\end{aligned}$$

Further if f is convex, $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ implies that $f(\mathbf{x}^\star) = \min_{\mathbf{y}} f(\mathbf{y})$. Substituting this and rearranging the terms in the above equation we get:

$$\frac{1}{2L}\|\nabla f_i(\mathbf{x}_t)\|^2 \le f_i(\mathbf{x}_t) - f_i(\mathbf{x}^\star).$$

Now taking conditional expectation on both sides gives us the desired result.

Question 32: *2 points.* Using the result in the previous question prove that

$$\mathbb{E}[f(\mathbf{x}_{t+1})|\mathbf{x}_t] \leq f(\mathbf{x}_t) - \gamma \|\nabla f(\mathbf{x}_t)\|^2 + \gamma^2 L^2 (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) . \tag{OPS}$$

Hint: Plug in the SGD update into the smoothness bound on f.

☐ 0 ☐ 1 ■ 2

Solution: Since each f_i is L-smooth, this implies that f is also L-smooth. Then we can write that

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &= f(\mathbf{x}_t) - \gamma \nabla f(\mathbf{x}_t)^\top \nabla f_i(\mathbf{x}_t) + \frac{L\gamma^2}{2} \|\nabla f_i(\mathbf{x}_t)\|^2 . \end{aligned}$$

Taking expectation on both sides and using the result in question 31 gives

$$\mathbb{E}[f(\mathbf{x}_{t+1})|\mathbf{x}_t] \leq f(\mathbf{x}_t) - \gamma \|\nabla f(\mathbf{x}_t)\|^2 + \gamma^2 L^2 (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) .$$

Question 33: *4 points.* Now suppose that f is μ-strongly convex. By picking an appropriate step-size γ, prove using (OPS) that SGD converges at a linear rate, i.e.,

$$\mathbb{E}[f(\mathbf{x}_t)] - f(\mathbf{x}^\star) \leq \left(1 - \frac{\mu^2}{L^2}\right)^t (f(\mathbf{x}_0) - f(\mathbf{x}^\star)) .$$

Hint: The best step-size is not $\frac{1}{L}$ and depends on μ.

☐ 0 ☐ 1 ☐ 2 ☐ 3 ■ 4

Solution: Since f is s.c., it satisfies the PL-condition

$$\|\nabla f(\mathbf{x}_t)\|^2 \geq 2\mu (f(\mathbf{x}_t) - f^\star) ,$$

where $f^\star = f(\mathbf{x}^\star)$. Replacing this in the result of Question 32 gives

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}_{t+1})|\mathbf{x}_t] &\leq f(\mathbf{x}_t) - 2\mu\gamma (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) + \gamma^2 L^2 (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \\ &= f(\mathbf{x}_t) - (2\mu\gamma - \gamma^2 L^2)(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) . \end{aligned}$$

Let us pick $\gamma = \frac{\mu}{L^2}$ to maximize the dependent term above. Then, subtracting $f(\mathbf{x}^\star)$ from both sides gives

$$\mathbb{E}[f(\mathbf{x}_{t+1})|\mathbf{x}_t] - f(\mathbf{x}^\star) \leq \left(1 - \frac{\mu^2}{L^2}\right) .$$

Unrolling the above while taking expectations gives the desired result.