# EXAM: Optimization for Data Science

Robert M. Gower and Alexandre Gramfort

December 29, 2018

This is an open book exam, meaning you can consult any written or printed material. Electronic devices are prohibited.

The exam consists of four question of increasing difficulty and also of increasing value. The last question is worth over half the total points. To get the maximum number of points, you must justify all of your answers. Good luck!

**Ex. 1 —** (2pt) Prove that $f(x) = x\log(x)$ is convex over its domain. On what subsets of the domain is $f(x)$ strongly convex and smooth?

**Answer (Ex. 1) —** Differentiating gives

$$f'(x) = \log(x) + 1,$$

and again

$$f''(x) = \frac{1}{x} > 0, \quad \forall x > 0.$$

Thus $f(x)$ is convex. If we restrict the domain of $f$ to any interval $[a,\, b]$, where $a > 0$, we have that

$$\frac{1}{b} \leq f''(x) \leq \frac{1}{a}$$

That is, $f$ is $\frac{1}{b}$–strongly convex and $\frac{1}{a}$–smooth over the interval $[a,\, b]$.

**Ex. 2 —** (2pt) Let $f : \mathbb{R} \rightarrow \mathbb{R}$ and $g : \mathbb{R} \rightarrow \mathbb{R}$ be two convex functions. Furthermore, let $g$ be a monotonically increasing function, that is,

$$x \leq y \quad \Rightarrow \quad g(x) \leq g(y), \quad \forall x, y \in \mathbb{R}.$$

Prove that the composition $g(f(x))$ is a convex function.

**Answer (Ex. 2) —** Let $x, y \in \mathbb{R}^d$ and $\lambda \in [0,\, 1]$. By convexity of $f$ we have that

$$f(\lambda x + (1-\lambda)y) \quad \leq \quad \lambda f(x) + (1-\lambda)f(y).$$

Applying $g$ to both sides and using the monotonicity of $g$ followed by the convexity of $g$ we have that

$$\begin{aligned} g(f(\lambda x + (1-\lambda)y)) &\leq g(\lambda f(x) + (1-\lambda)f(y)) \\ &\leq \lambda g(f(x)) + (1-\lambda)g(f(y)). \end{aligned} \tag{1}$$

**Ex. 3** — Let $f(x) = \frac{1}{n}\sum_{i=1}^n \ell(a_i^\top x) + \lambda\|x\|_2^2$, where $\ell : \mathbb{R} \to \mathbb{R}$ is a differentiable real valued function, $x, a_i \in \mathbb{R}^d$ and $\lambda > 0$ is a given regularization parameter.

(1pt) *Part I*

Re-write $f(x)$ as an average of functions $\frac{1}{n}\sum_{i=1}^n f_i(x)$. What is $f_i(x)$ in this case?

(3pt) *Part II*

Using the previous question we can now apply a step of the SGD (stochastic gradient descent) algorithm

$$x^{t+1} = x^t - \alpha_t \nabla f_i(x^t),$$

where $i \in \{1, \ldots, n\}$ is sampled uniformly at random at each step. Assume that $a_i$ has only $s \in \mathbb{N}$ nonzeros elements for $i = 1, \ldots, n$. That is, $a_i$ is $s$–sparse. From the following options, which best describes the computational cost of a single step of the SGD algorithm:<br>
*a)* $O(s)$<br>
*b)* $O(n)$<br>
*c)* $O(d)$<br>
*d)* $O(sd)$<br>
Justify your answer.

(6pt) *Part III*

You will now reduce the computational cost of a SGD step to $O(s)$ using the following change of variables. First you will represent each iterate as $x^t = \beta_t y^t$ where $\beta_t \in \mathbb{R}$ is a real number and $y^t \in \mathbb{R}^d$ a vector. Re-write the SGD update only in terms of the new variables $\beta_t$ and $y_t$, so that updating $\beta_t$ and $y_t$ costs $O(s)$ and $x^{t+1} = \beta_{t+1}y^{t+1}$.

**Answer (Ex. I)** — First note that we can re-write

$$\frac{1}{n}\sum_{i=1}^n \ell(\langle a_i, x\rangle) + \lambda\|x\|_2^2 = \frac{1}{n}\sum_{i=1}^n \underbrace{\left(\ell(\langle a_i, x\rangle) + \lambda\|x\|_2^2\right)}_{f_i(x)} = \frac{1}{n}\sum_{i=1}^n f_i(x).$$

**Answer (Ex. II)** —

$$\begin{aligned} x^{t+1} &= x^t - \alpha_t \nabla f_i(x^t) \\ &= x^t - \alpha_t \left(a_i \ell'(\langle a_i, x\rangle) + \lambda x^t\right) \\ &= (1-\alpha_t\lambda)x^t - \alpha_t \ell'(\left\langle a_i, x^t\right\rangle) a_i. \end{aligned} \tag{2}$$

Thus there is a dense re-scaling and a $s$–sparse addition in the update. The total cost is thus $O(d+s) = O(d)$.

**Answer (Ex. III)** — Let $x^t = \beta_t y^t$. From (2) we have that

$$\begin{aligned} \beta_{t+1} y^{t+1} &= (1-\alpha_t\lambda)\beta_t y^t - \alpha_t \ell'(\langle a_i, x\rangle) a_i. \\ &= \beta_t (1-\alpha_t\lambda)\left(y^t - \frac{\alpha_t \ell'(\langle a_i, x\rangle)}{(1-\alpha_t\lambda)\beta_t} a_i\right). \end{aligned} \tag{3}$$

Thus if we update $\beta_t$ and $y^t$ according to

$$\beta_{t+1} = \beta_t(1-\alpha_t\lambda), \tag{4}$$

$$y^{t+1} = y^t - \frac{\alpha_t \ell'(\beta_t \left\langle a_i, y^t\right\rangle)}{\beta_{t+1}} a_i, \tag{5}$$

then it holds that $x^{t+1} = \beta_{t+1} y^{t+1}$. Furthermore, this costs only $O(s)$ to update $y^t$ and $O(1)$ to update $\beta_t$.

**Ex. 4** — (16pt) Consider the minimization problem

$$x^* = \arg\min_{x\in\mathbb{R}^n} \left( f(x) \overset{\text{def}}{=} \frac{1}{2} x^\top A x - x^\top b \right), \tag{6}$$

where $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, and $x, b \in \mathbb{R}^n$. Consider a step of the stochastic coordinate descent method

$$x^{k+1} = x^k - \alpha_i \frac{\partial f(x^k)}{\partial x_i} e_i, \tag{7}$$

where $e_i \in \mathbb{R}^n$ is the $i$th unit coordinate vector, $\alpha_i = \dfrac{1}{A_{ii}}$, and $i \in \{1,\ldots,n\}$ is chosen uniformly at random at each step with probability $p_i = \dfrac{A_{ii}}{\mathrm{Tr}\,(A)}$. Let $\|x\|_A^2 \overset{\text{def}}{=} x^\top A x$. Prove that

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_A^2\right] \;\leq\; \left(1 - \frac{\lambda_{\min}(A)}{\mathrm{Tr}\,(A)}\right) \mathbb{E}\left[\|x^k - x^*\|_A^2\right] \tag{8}$$

thus (7) converges to the solution.
**Hint:** Since $A$ is symmetric positive definite you can use that

$$\lambda_{\min}(A) = \inf_{x \in \mathbb{R}^n} \frac{x^\top A x}{\|x\|_2^2}.$$

You will need to use that $x^\top Ax \geq \lambda_{\min}(A)\|x\|_2^2$ at some point.

**Answer (Ex. 4)** — First we calculate the partial derivative

$$\frac{\partial f(x)}{\partial x_i} = A_{i:}x - b_i,$$

where $A_{i:}$ is the $i$th row of $A$. Furthermore note that $x^* = A^{-1}b$ thus

$$\frac{\partial f(x)}{\partial x_i} = e_i^\top (Ax - b) = e_i^\top A(x - x^*). \tag{9}$$

Subtracting $x^*$ from both sides of (7) gives

$$\begin{aligned} x^{k+1} - x^* &\overset{(9)+(7)}{=} x^k - x^* - \alpha_i e_i^\top A(x^k - x^*)e_i \\ &= \left(I - \frac{e_i e_i^\top A}{A_{ii}}\right)(x^k - x^*). \end{aligned} \tag{10}$$

Let $P_i = \frac{e_i e_i^\top A}{A_{ii}}$. Taking the squared norm $\|\cdot\|_A$ on both sides of (10) gives

$$\begin{aligned} \|x^{k+1} - x^*\|_A^2 &= \left\langle A(I - P_i)(x^k - x^*), (I - P_i)(x^k - x^*)\right\rangle \\ &= \left\langle (I - P_i^\top)A(I - P_i)(x^k - x^*), x^k - x^* \right\rangle. \end{aligned} \tag{11}$$

Let $r^k = A^{1/2}(x^k - x^*)$ and note that

$$(I - P_i^\top)A(I - P_i) = A - 2AP_i + P_i^\top AP_i = A - \frac{Ae_i e_i^\top A}{A_{ii}}.$$

Using this we have from (11) that

$$\begin{aligned} \|r^{k+1}\|_2^2 &= \left\langle \left(A - \frac{Ae_i e_i^\top A}{A_{ii}}\right)(x^k - x^*), x^k - x^* \right\rangle \\ &= \|r^k\|_2^2 - \left\langle \frac{Ae_i e_i^\top A}{A_{ii}}(x^k - x^*), x^k - x^* \right\rangle \\ &= \|r^k\|_2^2 - \left\langle \frac{A^{1/2} e_i e_i^\top A^{1/2}}{A_{ii}} r^k, r^k \right\rangle. \end{aligned} \tag{12}$$

Taking expectation conditioned on $r^k$ over the second term in the above gives

$$\begin{aligned}
\mathbb{E}\left[\left\langle \frac{A^{1/2}e_ie_i^\top A^{1/2}}{A_{ii}} r^k, r^k \right\rangle \mid r^k\right] &= \sum_{j=1}^n \frac{A_{jj}}{\mathrm{Tr}\,(A)} \left\langle \frac{A^{1/2}e_je_j^\top A^{1/2}}{A_{jj}} r^k, r^k \right\rangle \\
&= \frac{1}{\mathrm{Tr}\,(A)} \left\langle A^{1/2} \sum_{j=1}^n e_je_j^\top A^{1/2} r^k, r^k \right\rangle \\
&= \frac{1}{\mathrm{Tr}\,(A)} \left\langle A r^k, r^k \right\rangle \\
&\geq \frac{\lambda_{\min}(A)}{\mathrm{Tr}\,(A)} \|r^k\|_2^2. \qquad (13)
\end{aligned}$$

Consequently taking expectation conditioned on $r^k$ in (12) gives

$$\mathbb{E}\left[\|r^{k+1}\|_2^2 \mid r^k\right] \leq \left(1 - \frac{\lambda_{\min}(A)}{\mathrm{Tr}\,(A)}\right) \|r^k\|_2^2. \qquad (14)$$

It now remains to take expectation and re-write $\|r^k\|_2^2 = \|x^k - x^*\|_A^2$. ∎