Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Department of Computer Science
Bernd Gärtner, Niao He, David Steurer

Optimization for Data Science Final Exam (1 September 2021) FS21

Candidate

First name: ....

Last name: ....

Student ID (Legi) Nr.: ....

I attest with my signature that I was able to take the exam without any impediments and that I have read and understood the general remarks below.

Signature: ....

General remarks and instructions

1. Check your exam documents for completeness (pages numbered from 1 to 21).
2. Immediately inform an assistant in case you experience any impediments. Complaints after the exam cannot be accepted.
3. You can solve the exercises in any order. They are not ordered by difficulty. Solutions should be written into the provided spaces. If you need scratch paper and/or extra paper for solutions, please ask an assistant.
4. Pencils are not allowed. Pencil-written solutions will not be graded.
5. No auxiliary material is allowed. All electronic devices must be turned off and are not allowed to be on your desk or carried with you to the toilet.
6. Attempts to cheat will be noted and reported to the examination office who will decide on the appropriate legal measures.
7. Provide only one solution to each exercise. Cross out invalid solutions clearly. If multiple solutions are provided, none of them will be graded.
8. **For full points, the explanations for your solutions must be clear, without any gaps, and mathematically rigorous unless explicitly stated otherwise.**
9. You may use (without proof) any statement that has been proved in the lecture and the exercise sessions, appears in previous subtasks of the same assignment, or as a hint in the assignments, or on the **cheat sheet** (last page of the exam). If you need something *different* from that, you must write a new proof or at least list all necessary changes.

Good luck!

To be filled by corrector

	achieved points (maximum)
1	(15)
2	(15)
3	(15)
4	(20)
5	(20)
Σ	(85)

Assignment 1. *A matrix* $M \in \mathbb{R}^{n\times n}$ *is called* completely positive *if for some natural number* m*, there exists a matrix* $A \in \mathbb{R}^{m\times n}$ *with nonnegative entries such that* $M = A^\top A$*. Let* $\mathrm{POS}_n \subseteq \mathbb{R}^{n\times n}$ *be the set of completely positive matrices.*

(a) Prove that every completely positive matrix is positive semidefinite.

(b) Let $M \in \mathbb{R}^{2\times 2}$ *be a (symmetric) positive definite matrix with nonnegative entries. Prove that* M *is completely positive.*

(c) A (matrix) cone is a subset $C \subseteq \mathbb{R}^{n\times n}$ *such that* $M \in C$ *implies* $\lambda M \in C$ *for all real numbers* $\lambda \geq 0$*. Prove that* POS_n *is a cone.*

(d) Prove that a cone C *is convex if and only if for all* $M, M' \in C$*, we also have* $M + M' \in C$*.*

(e) Prove that POS_n *is a convex cone.*

↓ Space for solution to Assignment 1 ↓

↓ Space for solution to Assignment 1 ↓

↓ Space for solution to Assignment 1 ↓

↓ Space for solution to Assignment 1 ↓

Solution:

(a) $x^\top M x = x^\top A^\top A x = ||Ax||^2 \geq 0$.

(b) *Master Solution*:

We claim that we find a matrix $A \in \mathbb{R}^{2\times 2}$ consisting of two columns a_1, a_2, such that $M = A^\top A$. In order for this to hold, we need to satisfy the following constraints:

$$\begin{aligned} a_1^\top a_1 = ||a_1||^2 &= M_{1,1}, \\ a_2^\top a_2 = ||a_2||^2 &= M_{2,2}, \\ a_1^\top a_2 = a_2^\top a_1 &= M_{1,2} = M_{2,1}. \end{aligned}$$

Since M is positive semidefinite, we also know that $\det(M) = M_{1,1}M_{2,2} - M_{1,2}^2 \geq 0$, so $M_{1,2} \leq \sqrt{M_{1,1}M_{2,2}}$.

In other words, we need two vectors a_1, a_2 of lengths $\sqrt{M_{1,1}}$ and $\sqrt{M_{2,2}}$ and given scalar product $a_1^\top a_2 = M_{1,2} \leq ||a_1||||a_2||$. This holds if we choose two vectors of the required lengths, with an angle α between them that satisfies $M_{1,2} = \cos(\alpha)||a_1||||a_2||$.

Alternative Solution I:

Since M is positive definite, M has an eigendecomposition $UDU^\top$ such that D is a diagonal matrix with only positive diagonal entries and U is an orthonormal matrix. Since all diagonal entries of D are positive, we have $M = UD^{\frac{1}{2}}D^{\frac{1}{2}}U^\top = \left(D^{\frac{1}{2}}U^\top\right)^\top \left(D^{\frac{1}{2}}U^\top\right)$. However, $U^\top$ may contain negative entries, e.g.,

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

To solve this issue, we need to "rotate" $D^{1/2}U^\top$. More precisely, let u_1 and u_2 be the two columns of $U^\top$. Since u_1 and u_2 are orthogonal to each other, we can rotate u_1 and u_2 to be the two positive axes. We use U as such a rotation matrix and will see that this will also work for the columns of $D^{1/2}U^\top$. We can write

$$\begin{aligned} M &= \left(D^{\frac{1}{2}}U^\top\right)^\top \left(D^{\frac{1}{2}}U^\top\right) \\ &= \left(D^{\frac{1}{2}}U^\top\right)^\top U^\top U \left(D^{\frac{1}{2}}U^\top\right) \\ &= \left(UD^{\frac{1}{2}}U^\top\right)^\top \left(UD^{\frac{1}{2}}U^\top\right) \\ &= A^\top A. \end{aligned}$$

Next, we need to prove that all entries of $A = UD^{\frac{1}{2}}U^\top$ are nonnegative. Let

$$D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \quad \text{and} \quad U = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

If $\lambda_1 = \lambda_2$, $A = U(\sqrt{\lambda_1} \cdot I)U^\top = \sqrt{\lambda_1}UU^\top = \sqrt{\lambda_1} \cdot I$, whose entries are trivially nonnegative. Thus, we assume that $\lambda_1 > \lambda_2 > 0$ below. Since U is orthogonal $ac + bd = 0$. By definition,

$$M = UDU^\top = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} a & c \\ b & d \end{pmatrix} = \begin{pmatrix} \lambda_1 a^2 + \lambda_2 b^2 & \lambda_1 ac + \lambda_2 bd \\ \lambda_1 ac + \lambda_2 bd & \lambda_1 c^2 + \lambda_2 d^2 \end{pmatrix}.$$

Since

$$0 \le M_{1,2} = \lambda_1 ac + \lambda_2 bd = (\lambda_1 - \lambda_2)ac + \lambda_2(ac + bd) = (\lambda_1 - \lambda_2)ac$$

and $\lambda_1 > \lambda_2$, we have $ac \ge 0$. Also,

$$A = UD^{\frac{1}{2}}U^\top = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{pmatrix} \begin{pmatrix} a & c \\ b & d \end{pmatrix} = \begin{pmatrix} \sqrt{\lambda_1} a^2 + \sqrt{\lambda_2} b^2 & \sqrt{\lambda_1} ac + \sqrt{\lambda_2} bd \\ \sqrt{\lambda_1} ac + \sqrt{\lambda_2} bd & \sqrt{\lambda_1} c^2 + \sqrt{\lambda_2} d^2 \end{pmatrix}.$$

The diagonal entries of A is trivially nonnegative. The other entries are also nonnegative since

$$\sqrt{\lambda_1} ac + \sqrt{\lambda_2} bd = \underbrace{\left(\sqrt{\lambda_1} - \sqrt{\lambda_2}\right)}_{>0} \underbrace{ac}_{\ge 0} + \sqrt{\lambda_2} \underbrace{(ac + bd)}_{=0} \ge 0.$$

Alternative Solution II:

Since M is symmetric and $M \in \mathbb{R}^{2 \times 2}$, let

$$M := \begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

Let λ_1 and λ_2 be the eigenvalues of M. Since M is positive definite we know that both are positive. Since $ac - b^2 = \det(M) = \lambda_1 \cdot \lambda_2 > 0$ implying that $ac > 0$ we know that both a and c are strictly positive (this uses that $a, c \ge 0$). Therefore, we have

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix} = \underbrace{\begin{pmatrix} \sqrt{\frac{ac-b^2}{c}} & \frac{b}{\sqrt{c}} \\ 0 & \sqrt{c} \end{pmatrix}}_{:=A^\top} \underbrace{\begin{pmatrix} \sqrt{\frac{ac-b^2}{c}} & 0 \\ \frac{b}{\sqrt{c}} & \sqrt{c} \end{pmatrix}}_{:=A}.$$

(c) if $M = A^\top A$, then $\lambda M = (\sqrt{\lambda}A)^\top \sqrt{\lambda}A) \in \text{POS}_n$.

(d) If C is convex and $M, M' \in C$, then $\frac{1}{2}M + \frac{1}{2}M' \in C$ by convexity. Since C is a cone, $M + M' \in C$ follows. For the other direction, let $M, M' \in C$. We need to show that $\lambda M + (1 - \lambda)M' \in C$ for $0 \leq \lambda \leq 1$. Let $X = \lambda M \in C$ by the cone property, and similarly $X' = (1-\lambda)M' \in C$. Then $X+X' \in C$, and this just means $\lambda M + (1 - \lambda)M' \in C$.

(e) If $M = A^\top A$ and $M' = B^\top B$, then $M + M' = C^\top C$, where

$$C = \left(\frac{A}{B} \right).$$

Assignment 2. *We know that for smooth functions, gradient descent has the property that the gradient norms converge to 0, even in the nonconvex case. Concretely, we have the following result.*

Suppose $f(x) : \mathbb{R}^d \to \mathbb{R}$ *is smooth with parameter* $L > 0$ *and has a global minimum* $x^\star$. *Then gradient descent with stepsize* $\gamma = 1/L$ *yields*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \leq \frac{2L}{T}(f(x_0) - f(x^\star)), \quad T > 0.$$

Here we want to establish a similar bound for stochastic *gradient descent. For this, we consider the stochastic optimization problem*

$$\min_{x \in \mathbb{R}^d} f(x) := \mathbb{E}_\xi[F(x, \xi)],$$

where $\xi \sim P$ *is a random vector chosen from some distribution* P. *We assume that the function* $f(x)$ *is well-defined, smooth with parameter* $L > 0$, *and has a global minimum* $x^\star$.

Further, we assume that the stochastic gradient is an unbiased estimator of the true gradient, meaning that

$$\mathbb{E}_\xi[\nabla F(x, \xi)] = \nabla f(x), \quad x \in \mathbb{R}^d.$$

We finally assume that the variance of the stochastic gradient is bounded, namely that for some fixed $\sigma > 0$,

$$\mathbb{E}_\xi[\|\nabla F(x, \xi) - \nabla f(x)\|^2] \leq \sigma^2, \quad x \in \mathbb{R}^d.$$

Recall that stochastic gradient descent (SGD) with fixed stepsize γ *performs the update*

$$x_{t+1} = x_t - \gamma \nabla F(x_t, \xi_t), \quad t = 0, \ldots, T-1,$$

starting from some fixed $x_0 \in \mathbb{R}^d$, *where the* $\{\xi_t\}$ *are independently and identically drawn from* P.

(a) Prove that

$$\mathbb{E}_{\xi_t}[f(x_{t+1}) - f(x_t) | x_t] \leq -\left(\gamma - \frac{L}{2}\gamma^2\right) \|\nabla f(x_t)\|^2 + \frac{L}{2}\sigma^2\gamma^2.$$

(b) You may use without proof that (a) implies

$$\mathbb{E}[f(x_{t+1}) - f(x_t)] \leq -\left(\gamma - \frac{L}{2}\gamma^2\right) \mathbb{E}[\|\nabla f(x_t)\|^2] + \frac{L}{2}\sigma^2\gamma^2,$$

where the expectation is now over all random choices in the algorithm. Prove that for $\gamma \leq 1/L$, *we have*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(x_t)\|^2] \leq \frac{2(f(x_0) - f(x^\star))}{T\gamma} + L\sigma^2\gamma.$$

(c) *For stochastic gradient variance* $\sigma^2 = 0$ *and* $\gamma = 1/L$*, we (unsurprisingly) recover the bound for gradient descent. If* $\sigma^2 > 0$*, which choice of* $\gamma = \gamma(T)$ *yields the best bound in (b), and what is this bound (you may assume that* T *is sufficiently large so that* $\gamma(T) \leq 1/L$*)? Discuss how the resulting bound compares to the one for gradient descent!*

↓ Space for solution to Assignment 2 ↓

↓ Space for solution to Assignment 2 ↓

↓ Space for solution to Assignment 2 ↓

Solution:

(a) By the smoothness of function $f(x)$, we have

$$f(x_{t+1}) - f(x_t) \leq \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{L}{2} \|x_{t+1} - x_t\|^2.$$

Plugging in the definition of x_{t+1}, we have

$$f(x_{t+1}) - f(x_t) \leq -\gamma \nabla f(x_t)^T \nabla F(x_t, \xi_t) + \frac{L}{2} \gamma^2 \|\nabla F(x_t, \xi_t)\|^2.$$

Now we take the conditional expectation over ξ_t, given x_t. From our assumptions, $\mathbb{E}_{\xi_t}[\nabla F(x_t, \xi_t)|x_t] = \nabla f(x_t)$, and $\mathbb{E}_{\xi_t}[\|\nabla F(x_t, \xi_t)\|^2 | x_t] \leq \|\nabla f(x_t)\|^2 + \sigma^2$ (using that for a vector of random variables, $\text{Var}(X) := E(\|X - E(X)\|^2) = E[\|X\|^2] - \|E[X]\|^2$). Hence,

$$\mathbb{E}_{\xi_t}[f(x_{t+1}) - f(x_t)|x_t] \leq -\gamma \|\nabla f(x_t)\|^2 + \frac{L}{2} \gamma^2 \|\nabla f(x_t)\|^2 + \frac{L}{2} \gamma^2 \sigma^2.$$

The desired bound follows.

(b) For $\gamma \leq 1/L$, we have $\gamma - \frac{L}{2}\gamma^2 \geq \gamma/2$, so the bound from (a) yields

$$\mathbb{E}[f(x_{t+1}) - f(x_t)] \leq -\frac{\gamma}{2} \mathbb{E}[\|\nabla f(x_t)\|^2] + \frac{L}{2} \sigma^2 \gamma^2,$$

Rewriting this (and using linearity of expectation) gives

$$\mathbb{E}[\|\nabla f(x_t)\|^2] \leq \frac{2\mathbb{E}[f(x_t) - f(x_{t+1})] + L\sigma^2\gamma^2}{\gamma} = \frac{2\mathbb{E}[f(x_t) - f(x_{t+1})]}{\gamma} + L\sigma^2\gamma.$$

Summing up, we have telescoping:

$$\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(x_t)\|^2] \leq \frac{2\mathbb{E}[f(x_0) - f(x_T)]}{\gamma} + TL\sigma^2\gamma \leq \frac{2(f(x_0) - f(x^\star))}{\gamma} + TL\sigma^2\gamma.$$

Dividing by T, the claimed bound follows.

(c) Let $E := f(x_0) - f(x^\star)$. We need to minimize

$$q(\gamma) := \frac{2E}{T\gamma} + L\sigma^2\gamma$$

which happens when

$$q'(\gamma) = -\frac{2E}{T\gamma^2} + L\sigma^2 = 0.$$

Hence,

$$\gamma = \sqrt{\frac{2E}{TL\sigma^2}}.$$

This yields

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(x_t)\|^2] \leq \frac{2E}{T\gamma} + L\sigma^2\gamma = 2\frac{\sqrt{2EL\sigma^2}}{\sqrt{T}}.$$

Hence, when gradient descent needs S steps to reach error ε, stochastic gradient descent needs $\Theta(S^2)$ steps, considering everything else as constant.

Assignment 3. *Newton's method has known global convergence guarantees only in specific scenarios. In this assignment, we will deal with one such scenario.*

We are a given a twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ *that is strongly convex with parameter* $\mu > 0$ *and smooth with parameter* $L > \mu$*. In particular, the two functions* $g(x) = f(x) - \frac{\mu}{2}x^\top x$ *and* $h(x) = \frac{L}{2}x^\top x - f(x)$ *are convex. For all* $x \in \mathbb{R}^n$ *and with* Id *denoting the identity matrix, the second-order characterization therefore yields that the two matrices*

$$\nabla^2 f(x) - \mu \cdot \mathrm{Id}, \quad L \cdot \mathrm{Id} - \nabla^2 f(x)$$

are positive semidefinite, meaning that all eigenvalues of $\nabla^2 f(x)$ *are between* μ *and* L*. This in turn means that for all natural numbers* k*, the matrices*

$$\left(\nabla^2 f(x)\right)^{-k} - \frac{1}{L^k} \cdot \mathrm{Id}, \quad \frac{1}{\mu^k} \cdot \mathrm{Id} - \left(\nabla^2 f(x)\right)^{-k} \tag{1}$$

are positive semidefinite (all eigenvalues of $\left(\nabla^2 f(x)\right)^{-k}$ *are between* $1/L^k$ *and* $1/\mu^k$*). You may use (1) without proof.*

We consider Newton's method with fixed stepsize $\alpha > 0$*, i.e., we generate a sequence* $x_0, x_1, \ldots$ *of points where* x_0 *is an arbitrary point and for each* $t \geq 0$*,*

$$x_{t+1} = x_t - \alpha \left(\nabla^2 f(x_t)\right)^{-1} \nabla f(x_t).$$

(a) For each $t \geq 0$*, prove that*

$$f(x_{t+1}) \leq f(x_t) - \frac{\alpha}{L}\|\nabla f(x_t)\|^2 + \frac{L\alpha^2}{2\mu^2}\|\nabla f(x_t)\|^2.$$

(b) Let x^* *be the unique global minimizer of* f*. Prove that for some suitable* $\alpha > 0$*, and for every* $t \geq 0$*, we have*

$$f(x_{t+1}) - f(x^*) \leq (f(x_t) - f(x^*))\left(1 - \frac{\mu^3}{L^3}\right).$$

This implies (but you don't have to prove it) that for every constant $\varepsilon > 0$*, we have* $f(x_T) - f(x^*) \leq \varepsilon$ *after* $T = O\left(\log\left(\frac{f(x_0) - f(x^*)}{\varepsilon}\right)\right)$ *many steps.*

You can use part (a) and the inequality

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f(x^*)) \tag{2}$$

without proof.

↓ Space for solution to Assignment 3 ↓

↓ Space for solution to Assignment 3 ↓

↓ Space for solution to Assignment 3 ↓

↓ Space for solution to Assignment 3 ↓

Solution:

(a) From smoothness of f, we have:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{L}{2}||\mathbf{y}-\mathbf{x}||^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

Let $\mathbf{d}_t$ be the Newton direction such that $\mathbf{d}_t = \left(\nabla^2 f(\mathbf{x}_t)\right)^{-1} \nabla f(\mathbf{x}_t)$. Writing the above inequality for $\mathbf{y} = \mathbf{x}_{t+1} = \mathbf{x}_t - \alpha \mathbf{d}_t$ and $\mathbf{x} = \mathbf{x}_t$, we have:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \alpha \nabla f(\mathbf{x}_t)^T \mathbf{d}_t + \frac{L\alpha^2}{2}||\mathbf{d}_t||^2$$

Replacing $\mathbf{d}_t$ with $\left(\nabla^2 f(\mathbf{x}_t)\right)^{-1} \nabla f(\mathbf{x}_t)$, we have:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \alpha \nabla f(\mathbf{x}_t)^T \left(\nabla^2 f(\mathbf{x}_t)\right)^{-1} \nabla f(\mathbf{x}_t) + \frac{L\alpha^2}{2} \nabla f(\mathbf{x}_t)^T \left(\left(\nabla^2 f(\mathbf{x}_t)\right)^{-1}\right)^2 \nabla f(\mathbf{x}_t)$$

Using (1) with $k = 1$ and $k = 2$ and for any $\mathbf{y} \in \mathbb{R}^n$, we can bound the quadratic terms as follows:

$$\frac{||\mathbf{y}||^2}{L} \leq \mathbf{y}^T \left(\nabla^2 f(\mathbf{x}_t)\right)^{-1} \mathbf{y}$$
$$\mathbf{y}^T \left(\left(\nabla^2 f(\mathbf{x}_t)\right)^{-1}\right)^2 \mathbf{y} \leq \frac{||\mathbf{y}||^2}{\mu^2}$$

Hence:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{\alpha}{L}||\nabla f(\mathbf{x}_t)||^2 + \frac{L\alpha^2}{2\mu^2}||\nabla f(\mathbf{x}_t)||^2$$

(b) By setting α to $\frac{\mu^2}{L^2}$, the choice that minimizes the right-hand side in (a), we have

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{\mu^2}{2L^3}||\nabla f(\mathbf{x}_t)||^2.$$

Plugging in the given lower bound (2) for $||\nabla f(\mathbf{x}_t)||^2$, we get

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{\mu^3}{L^3}\left(f(\mathbf{x}_t) - f(\mathbf{x}^*)\right)$$
$$\Rightarrow f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \leq \left(f(\mathbf{x}_t) - f(\mathbf{x}^*)\right)\left(1 - \frac{\mu^3}{L^3}\right).$$

For the sake of completeness, here is the proof of the fact $||\nabla f(\mathbf{x})||^2 \geq 2\mu\left(f(\mathbf{x}) - f(\mathbf{x}^*)\right)$. From strong convexity of f, we have:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}||\mathbf{y}-\mathbf{x}||^2 \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

For fixed x, the RHS is a quadratic function on y. Its gradient is $\nabla f(\mathsf{x}) + \mu\mathsf{y} - \mu\mathsf{x}$. So it is minimized at $\mathsf{y}^* = \mathsf{x} - \frac{\nabla f(\mathsf{x})}{\mu}$. Replacing y^* in the RHS, we get the following for any x and y:

$$f(\mathsf{y}) \geq f(\mathsf{x}) - \frac{1}{2\mu}\|\nabla f(\mathsf{x})\|^2.$$

Replacing y with x^* and rearranging the terms concludes the proof.

Assignment 4. *Consider the linear regression model*

$$y = X\beta^* + w$$

where $x_1, \dots x_n \overset{\text{i.i.d}}{\sim} D$ *are the rows of* X, *for some distribution* D *over* d-*dimensional vectors,* $\beta^* \in \mathbb{R}^d$ *and* $w \sim N(0, \mathrm{Id})$. *Suppose* w *and* X *are independent. You can use the following fact without proof:*

Fact: *For every realization of* X, *the least-squares estimator* $\hat{\beta}$ *achieves the following guarantee:*

$$\mathcal{L}_n := \frac{1}{n}\mathbb{E}\left[\|X\hat{\beta} - X\beta^*\|^2 \;\middle|\; X\right] \leq \frac{d}{n}\,.$$

In the lecture, you have seen the proof of this fact when the realization of X *has rank* d. *The random variable* $\mathcal{L}_n$ *is also referred to as the* in-sample prediction error, *since the error is measured with respect to the observed data points* $x_1, \dots, x_n$. *The aim of this exercise is to prove a bound on the* out-of-sample prediction error *of the least-squares estimator, which is defined as*

$$\mathcal{L} := \mathbb{E}\left[\left(\langle\hat{\beta}, x\rangle - \langle\beta^*, x\rangle\right)^2 \;\middle|\; X\right]$$

where $x \sim D$ *is independent of* X *and* w.

(a) Show that

$$\mathcal{L} = \mathbb{E}\left[\|M^{1/2}\hat{\beta} - M^{1/2}\beta^*\|^2 \;\middle|\; X\right]$$

where $M = \mathbb{E}xx^\top$ *for* $x \sim D$.

(b) Suppose M *is invertible. Define* $\hat{M} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$ *and* $\Delta = \mathrm{Id} - M^{-1/2}\hat{M}M^{-1/2}$. *Show that for* $\|\Delta\| < 1$

$$\mathcal{L} \leq \frac{1}{1 - \|\Delta\|} \cdot \frac{d}{n}$$

Remark: *For* $\|\Delta\| \leq 1/2$, *this implies that* $\mathcal{L} \leq (1 + 2\|\Delta\|) \cdot \frac{d}{n} \leq 2 \cdot \frac{d}{n}$ *(you don't need to prove this). In the exercises we have seen that if the distribution* D *equals* $N(0, \Sigma)$ *and* $n \geq C \cdot d \log d$ *for some absolute constant* C *then* $\|\Delta\| \leq 1/2$ *with high probability.*

↓ Space for solution to Assignment 4 ↓

↓ Space for solution to Assignment 4 ↓

↓ Space for solution to Assignment 4 ↓

↓ Space for solution to Assignment 4 ↓

Solution:

(a) It follows from linear algebra calculations and the independence between x and $\widehat{\beta}, X$.

$$
\begin{aligned}
\mathcal{L} &:= \mathbb{E}\left[\left(\langle\widehat{\beta}, x\rangle - \langle\beta^*, x\rangle\right)^2 \,\middle|\, \mathbf{X}\right] \\
&= \mathbb{E}\left[\langle\widehat{\beta} - \beta^*, x\rangle^2 \,\middle|\, \mathbf{X}\right] \\
&= \mathbb{E}\left[\left(\widehat{\beta} - \beta^*\right)^\top x x^\top \left(\widehat{\beta} - \beta^*\right) \,\middle|\, \mathbf{X}\right] \\
&= \mathbb{E}\left[\left(\widehat{\beta} - \beta^*\right)^\top \mathbb{E} x x^\top \left(\widehat{\beta} - \beta^*\right) \,\middle|\, \mathbf{X}\right] \\
&= \mathbb{E}\left[\left(\widehat{\beta} - \beta^*\right)^\top M \left(\widehat{\beta} - \beta^*\right) \,\middle|\, \mathbf{X}\right] \\
&= \mathbb{E}\left[\left\|M^{1/2}\left(\widehat{\beta} - \beta^*\right)\right\|^2 \,\middle|\, X\right]
\end{aligned}
$$

(b) Since $\widehat{\mathbf{M}} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ and $M - \widehat{\mathbf{M}} = M^{1/2}\mathbf{\Delta}M^{1/2}$, we have

$$
\begin{aligned}
\mathcal{L} - \mathcal{L}_n &= \mathbb{E}\left[\langle\beta^* - \widehat{\beta}, (M - \widehat{\mathbf{M}})(\beta^* - \widehat{\beta})\rangle \,\middle|\, \mathbf{X}\right] \\
&= \mathbb{E}\left[\langle\beta^* - \widehat{\beta}, M^{1/2}\mathbf{\Delta}M^{1/2}(\beta^* - \widehat{\beta})\rangle \,\middle|\, \mathbf{X}\right] \\
&= \mathbb{E}\left[\langle M^{1/2}(\beta^* - \widehat{\beta}), \mathbf{\Delta}M^{1/2}(\beta^* - \widehat{\beta})\rangle \,\middle|\, \mathbf{X}\right] \\
&\le \|\mathbf{\Delta}\| \cdot \mathbb{E}\left[\left\|M^{1/2}(\beta^* - \widehat{\beta})\right\|^2 \,\middle|\, X\right] \\
&= \|\mathbf{\Delta}\| \cdot \mathcal{L}
\end{aligned}
$$

Therefore

$$
(1 - \|\mathbf{\Delta}\|) \cdot \mathcal{L} \le \mathcal{L}_n \,.
$$

The claim then follows.

Assignment 5. *Let* $n \in \mathbb{N}$ *be sufficiently large (e.g.,* $n \geq 10^{10}$*), let* $a, b \in \mathbb{R}$ *with* $0.1 \leq b < a \leq 1$*, and let* $x \in \{-1, 1\}^n$ *be an* n*-dimensional vector. Let* $\mathbf{G}$ *be an undirected random graph with vertex set* $[n] = \{1, 2, \ldots, n\}$ *constructed as follows. For every pair of distinct vertices* $i < j \in [n]$ *independently, add an edge between* i *and* j *in* $\mathbf{G}$ *with probability* c_{ij}*, where*

$$c_{ij} = \begin{cases} a & \textit{if } x_i = x_j \,, \\ b & \textit{if } x_i \neq x_j \,. \end{cases}$$

Let $\mathbf{L}$ *denote the adjacency matrix of* $\mathbf{G}$*, i.e.,* $\mathbf{L}_{i,j} = 1$ *if there is an edge between* i *and* j *in* $\mathbf{G}$ *and* $\mathbf{L}_{i,j} = 0$ *otherwise. Note that the diagonal entries of* $\mathbf{L}$ *are* 0 *and that* $\mathbf{L}$ *is symmetric because* $\mathbf{G}$ *is undirected.*

Our goal is to approximately recover x*, given* $n, a, b, \mathbf{L}$*.*

To this end, we consider the matrix $\mathbf{M} = \mathbf{L} - \frac{a+b}{2} \cdot J + a \cdot I_n$*, where* J *is an* n*-by-*n *matrix with all entries equal to* 1*.*

(a) Show that $\mathbb{E}\mathbf{M} = \lambda \cdot xx^\top$*, where* $\lambda = \frac{a-b}{2}$*.*

(b) Show that $\|\mathbf{M} - \mathbb{E}\mathbf{M}\| \leq C\sqrt{n}$ *with probability at least* 0.99 *for some large enough constant* C *(e.g.,* $C = 40$*). Recall that for matrices, we use* $\|\cdot\|$ *to denote the spectral norm.*

(c) Show that there exists a polynomial-time algorithm that given $\mathbf{L}$ *computes an estimator* $\hat{x}$ *taking values in* $\mathbb{R}^n$ *such that with probability at least* 0.99*,*

$$\left\| \lambda \cdot \hat{x}\hat{x}^\top - \lambda \cdot xx^\top \right\|_F^2 \leq O(n) \,.$$

(d) Show that there is a way to post-process the previous estimator $\hat{x}$ *in polynomial time to obtain an estimator* $\hat{x}'$ *taking values in* $\{-1, 1\}^n$ *that satisfies the following error bound with probability at least* 0.99*,*

$$\left| \left\{ (i, j) \in [n] \times [n] \;\middle|\; \hat{x}'_i \cdot \hat{x}'_j \neq x_i \cdot x_j \right\} \right| \leq O(n/\lambda^2) \,.$$

(e) Consider a naive estimator $\tilde{x}$ *with every entry* $\tilde{x}_i$ *independently drawn uniformly from* $\{-1, 1\}$*. Find the smallest constant* $c_0 \in \mathbb{R}$*, such that for every constant* $c > c_0$ *and* $\lambda = n^c$*, the error bound in the previous part (d) is substantially better than the error of this naive estimator in the following sense,*

$$\lim_{n \to \infty} \frac{n/\lambda^2}{\mathbb{E}\,|\{(i, j) \in [n] \times [n] \mid \tilde{x}_i \cdot \tilde{x}_j \neq x_i \cdot x_j\}|} = 0 \,.$$

For this part, a short justification for your choice of c_0 *is enough. A full proof is not necessary.*

↓ Space for solution to Assignment 5 ↓

↓ Space for solution to Assignment 5 ↓

↓ Space for solution to Assignment 5 ↓

↓ Space for solution to Assignment 5 ↓

Solution:

(a) For $i, j \in [n], i \neq j$, we have $\mathbb{E}L_{ij} = a$ if $x_i = x_j$ and $\mathbb{E}L_{ij} = b$ if $x_i \neq x_j$. Summarizing the two cases, we have $\mathbb{E}L_{ij} = \frac{a+b}{2} + \frac{a-b}{2}x_i x_j$. By definition $L_{ii} = 0$ for $i \in [n]$, it then follows that $\mathbb{E}L = \frac{a+b}{2} \cdot J + \frac{a-b}{2}xx^\top - a\mathrm{Id}_n$. By the definition of M, we have

$$\mathbb{E}M = \mathbb{E}L - \frac{a+b}{2} \cdot J + a\mathrm{Id}_n = \frac{a-b}{2}xx^\top$$

(b) Since $L_{ij} \in \{0, 1\}$, we have $M_{ij} \in [-a, 1 + a]$. Further $\mathbb{E}M_{ij} = \frac{a-b}{2}x_i x_j \in \{\frac{b-a}{2}, \frac{a-b}{2}\}$. Thus $|M_{ij} - \mathbb{E}M_{ij}| \leq 10$. Since M_{ij} are independent for $i, j \in [n]$, and $M - \mathbb{E}M$ has zero mean, we can apply the hint to the random matrix $M - \mathbb{E}M$ and get that $\|M - \mathbb{E}M\| \leq 40\sqrt{n}$ with probability at least 0.99.

(c) We denote $X^0 := \mathbb{E}M = \lambda xx^\top$ and $W := M - X^0$. Then $M = X^0 + W$.

Since M is symmetric, we know that it has an eigendecomposition of the form

$$M = \sum_{i=1}^{n} \lambda_i v_i v_i^\top = \sum_{i=1}^{n} |\lambda_i|(\mathrm{sgn}(\lambda_i) \cdot v_i)v_i^\top$$

where $\lambda_1 \geq \ldots \geq \lambda_n$ and $v_1, \ldots, v_n$ are orthonormal. Because of the last equality, this in fact also gives rise to an SVD of M and can hence be computed in polynomial time. Here $|\lambda_i|$ are the singular values, and $\mathrm{sgn}(\lambda_i)v_i$ and v_i are taken as left and right singular vectors respectively. This means that the largest singular value is given by $\max\{|\lambda_1|, |\lambda_n|\}$. Suppose that it is in fact given by λ_1. Note that this implies $\lambda_1 \geq 0$.[1] Then from lecture and the above we know that $\lambda_1 v_1 v_1^\top$ is a best rank-1 approximation for M. Since $\lambda_1 \geq 0$ we can then choose our estimator as $\hat{x} = \sqrt{\frac{\lambda_1}{\lambda}}v_1$. This implies that

$$\hat{X} := \lambda \cdot \hat{x}\hat{x}^\top \in \arg\min \left\{ \|X - M\|_F^2 \mid \mathrm{rank}(X) \leq 1 \right\}$$

The rest of the proof is then almost the same as the proof of the guarantees of the best rank-1 approximation seen in lecture. First, we will show that $\left\|\hat{X} - X^0\right\|_F \leq 2\sqrt{2} \cdot \|W\|$. To this end, we denote $U := \hat{X} - X^0$. By definition of $\hat{X}$, we have

$$\begin{aligned}
0 &\leq \left\|X^0 - M\right\|_F^2 - \|\hat{X} - M\|_F^2 \\
&= \|W\|_F^2 - \|U - W\|_F^2 \\
&= \|W\|_F^2 - \|U\|_F^2 - \|W\|_F^2 + 2\langle U, W\rangle \\
&= -\|U\|_F^2 + 2\langle U, W\rangle
\end{aligned}$$

[1] We will show this in the very end. Solutions who assumed one or both of these statements without proof were awarded full points if they were otherwise correct.

Thus $\|\mathbf{U}\|_F^2 \leq 2\langle \mathbf{U}, \boldsymbol{W} \rangle$. Since $\mathbf{U} = \hat{X} - X^0$ has rank at most 2, by the relation between nuclear norm and Frobenius norm, we have

$$\|\mathbf{U}\|_* \leq \sqrt{2}\,\|\mathbf{U}\|_F$$

By Holder's inequality, we have

$$\begin{aligned}\|\mathbf{U}\|_F^2 &\leq 2\langle \mathbf{U}, \boldsymbol{W} \rangle \\ &\leq 2\|\mathbf{U}\|_*\|\boldsymbol{W}\| \\ &\leq 2 \cdot \sqrt{2} \cdot \|\mathbf{U}\|_F \cdot \|\boldsymbol{W}\|\end{aligned}$$

Therefore we have $\left\|\hat{X} - X^0\right\|_F = \|\mathbf{U}\|_F \leq 2\sqrt{2} \cdot \|\boldsymbol{W}\|$.

By definition, $X^0 = \lambda \cdot xx^\top$. By part (b), with probability at least 0.99, we have $\|\boldsymbol{W}\| \leq 40\sqrt{n}$. Therefore with probability at least 0.99, we have $\left\|\lambda \cdot \hat{x}\hat{x}^\top - \lambda xx^\top\right\|_F \leq 80\sqrt{2}\sqrt{n}$

In what follows we will prove the assumption we made about the top singular value of $\boldsymbol{M}$. I.e., we give a proof that the top singular value is given by $\boldsymbol{\lambda}_1 \geq 0$. W.l.o.g., we can assume that $\lambda \geq \frac{100}{\sqrt{n}}$. Otherwise we can take $\hat{x} = 0$ and get

$$\left\|\lambda\hat{x}\hat{x}^\top - \lambda xx^\top\right\|_F = \left\|\lambda xx^\top\right\|_F \leq 100\sqrt{n}$$

and we get the desired bound.

We would like to show that $\max_{1\leq i\leq n}|\boldsymbol{\lambda}_i| = \max\{|\boldsymbol{\lambda}_1|, |\boldsymbol{\lambda}_n|\} = \boldsymbol{\lambda}_1$. From b) we know that with probability 0.99 we have $\|\boldsymbol{M} - \mathbb{E}\boldsymbol{M}\| \leq 40\sqrt{n}$ and hence:

$$|\boldsymbol{\lambda}_1| = \|\boldsymbol{M}\| = \|\boldsymbol{M} - \mathbb{E}\boldsymbol{M} + \mathbb{E}\boldsymbol{M}\| \geq \left\|\lambda xx^\top\right\| - \|\boldsymbol{M} - \mathbb{E}\boldsymbol{M}\| \geq \lambda n - 40\sqrt{n} \geq 60\sqrt{n}$$

Further, we have that

$$\begin{aligned}\boldsymbol{\lambda}_n &= \boldsymbol{v}_n^\top \boldsymbol{M} \boldsymbol{v}_n = \lambda \boldsymbol{v}_n^\top xx^\top \boldsymbol{v}_n + \boldsymbol{v}_n^\top (\boldsymbol{M} - \mathbb{E}\boldsymbol{M})\boldsymbol{v}_n \\ &\geq 0 - \|\boldsymbol{M} - \mathbb{E}\boldsymbol{M}\| \geq -40\sqrt{n}.\end{aligned}$$

Thus $\boldsymbol{\lambda}_1 \geq 60\sqrt{n} > |\boldsymbol{\lambda}_n|$ and the top singular value is given by $\boldsymbol{\lambda}_1$.

(d) The estimator is given by

$$\hat{x}'_i = \begin{cases} 1 & \text{if } \hat{x}_i > 0\,, \\ -1 & \text{if } \hat{x}_i \leq 0\,. \end{cases}$$

We note that for any $i, j \in [n]$ such that $\widehat{x'}_i \cdot \widehat{x'}_j \neq x_i \cdot x_j$, we have $|\hat{x}_i \cdot \hat{x}_j - x_i \cdot x_j| \geq 1$. Therefore with probability at least 0.99,

$$\left|\left\{(i,j) \in [n] \times [n] \;\middle|\; \hat{x}'_i \cdot \hat{x}'_j \neq x_i \cdot x_j\right\}\right| \leq \left\|\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top - \boldsymbol{x}\boldsymbol{x}^\top\right\|_F^2 \leq \frac{12800 \cdot n}{\lambda^2}\,.$$

(e) Fix a pair $(i, j) \in [n] \times [n]$. If $i = j$, then $\tilde{x}_i^2 = 1 = x_i^2$. Else, $\tilde{x}_i \cdot \tilde{x}_j$ is uniform over $\{-1, 1\}$ and hence it is equal to $x_i x_j$ with probability $1/2$. Since there are $n^2 - n$ such pairs it follows by linearity of expectation that

$$\mathbb{E}\,|\{(i, j) \in [n] \times [n] \mid \tilde{x}_i \cdot \tilde{x}_j \neq x_i \cdot x_j\}| = \frac{n^2 - n}{2}$$

Therefore, setting $\lambda = n^c$, the expression in the limit becomes $\frac{2 \cdot n^{1-2c}}{n^2 - n} = \Theta\left(\frac{1}{n^{1+2c}}\right)$. Thus, for $c > -1/2$ the limit is 0 whereas it is non-zero for $c_0 = 1/2$.

Cheat sheet

In your solutions, you can use the following facts without proof.

Fact: Let $n \geq 10^6$, Let $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ be a symmetric random matrix with independent upper triangular entries, if for each $i, j \in [n]$ we have $\mathbb{E}[\boldsymbol{W}_{ij}] = 0$ and $|\boldsymbol{W}_{i,j}| \leq 10$, then $\|\boldsymbol{W}\| \leq 40\sqrt{n}$ with probability at least 0.99.

Holder inequality: For any real symmetric matrices $U, W \in \mathbb{R}^{n\times n}$, $|\langle U, W\rangle| \leq \|U\|_* \cdot \|W\|$, where $\|U\|_*$ is the nuclear norm of matrix U and $\|W\|$ the spectral norm of matrix W.

Fact: For any real symmetric matrix U, the nuclear norm $\|U\|_* \leq \sqrt{\mathrm{rank}(U)} \cdot \|U\|_F$.

Cyclicity of trace: For every pair of matrices $A, B \in \mathbb{R}^{m\times n}$, we have $\mathrm{Tr}(A^\top B) = \mathrm{Tr}(B^\top A)$.

Singular value decomposition: without further justification, you can use that singular value decompositions can be computed in polynomial time.

Fact: For $n > 10^6$, suppose $x_1, x_2, \ldots, x_n \in \mathbb{R}$ are independently sampled from $\{-1, 1\}$ with probability 1/2 each, then with probability at least 0.99,

$$\left|\sum_{i=1}^{n} x_i\right| \leq 20\sqrt{n}.$$