UNIVERSITY OF TORONTO
Faculty of Arts and Science

STA2311H: Advanced Computational Methods for Statistics I

Midterm

October 17, 2023

2 hours and 45 minutes

Name: _____

Student Number: _____

- Do not open this test until you are told to begin.
- This midterm is closed-book; a one-sided handwritten cheat sheet is allowed.
- There are four questions (worth a total of 100 points) on the midterm. Take a quick scan through the questions first and prioritize your time accordingly.
- Show all of your work for full marks, and ensure your notation is legible, correct, and consistent with that used in the course.
- If you need to use a result from lecture or the practice problems, briefly describe it.
- If you run out of space, use the back of the page.

Good luck!

1. (25 points) Let $d \geq 1$ and let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be a symmetric positive definite matrix. Let $\mathbf{a} \in \mathbb{R}^d$ and consider the function $f : \mathbb{R}^d \to \mathbb{R}$ defined by

$$f(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{a})^\top \mathbf{A}(\mathbf{x} - \mathbf{a}).$$

   (a) (10 points) Show that f has a global minimizer.[1]

   (b) (15 points) Derive the update rule for the gradient descent algorithm with the optimal step size.

[1] It might help to recall the following identity: if $\mathbf{B} \in \mathbb{R}^{n \times n}$, then $\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\left(\mathbf{x}^\top \mathbf{B}\mathbf{x}\right) = (\mathbf{B} + \mathbf{B}^\top)\mathbf{x}$.

2. (25 points) Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex,[2] differentiable, and L-Lipschitz.[3] Let $\mathbf{x}_1, \mathbf{x}^* \in \mathbb{R}^d$ be arbitrary and let $\epsilon > 0$. Define the following quantities:

$$T = \frac{L^2 \cdot ||\mathbf{x}_1 - \mathbf{x}^*||^2}{\epsilon^2}, \quad \eta = \frac{||\mathbf{x}_1 - \mathbf{x}^*||}{L\sqrt{T}}, \quad \phi_t = \frac{||\mathbf{x}_t - \mathbf{x}^*||^2}{2\eta}, \quad \text{and} \quad \hat{\mathbf{x}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t,$$

where $\mathbf{x}_t$ is defined by the update rule $\mathbf{x}_t = \mathbf{x}_{t-1} - \eta \cdot \nabla f(\mathbf{x}_{t-1})$ for $t \geq 2$.

(a) (5 points) Show that $\phi_{t+1} - \phi_t = \frac{1}{2\eta}\big(2\langle \mathbf{x}_{t+1} - \mathbf{x}_t, \mathbf{x}_t - \mathbf{x}^* \rangle + ||\mathbf{x}_{t+1} - \mathbf{x}_t||^2\big)$.

(b) (5 points) Show that $f(\mathbf{x}_t) + (\phi_{t+1} - \phi_t) \leq f(\mathbf{x}^*) + \frac{1}{2}\eta L^2$.

[2] Recall that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if $f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

[3] Recall that if a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz, then $||\nabla f(\mathbf{x})|| \leq L$ for all $\mathbf{x} \in \mathbb{R}^d$.

(c) (5 points) Show that $\sum_{t=1}^{T} f(\mathbf{x}_t) \leq T \cdot f(\mathbf{x}^*) + \frac{1}{2}\eta T L^2 + \frac{1}{2\eta}||\mathbf{x}_1 - \mathbf{x}^*||^2$.

(d) (10 points) Show that $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

3. (25 points) Suppose you receive a bag containing two (potentially) biased but otherwise identical coins: the first has probability of heads p_1, the second has probability of heads p_2. In order to estimate these parameters, you independently repeat the following experiment n times: pick a coin uniformly at random from the bag, toss it and record the outcome (i.e., heads or tails), and return the coin to the bag. Based on the data you obtain, devise an EM algorithm for estimating $\mathbf{p} = (p_1, p_2)$. It might help to introduce appropriate notation for quantities that appear repeatedly in your derivation.

4. (25 points) Let $d \geq 1$. Suppose that $p(\mathbf{x})$ is some fixed d-dimensional distribution which we wish to approximate with some normal distribution $q(\mathbf{x}) = \mathcal{N}_d(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ such that $\mathrm{KL}(p \,||\, q)$ is minimized. By differentiating $\mathrm{KL}(p \,||\, q)$ with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, show that $\boldsymbol{\mu}$ is given by $\mathbb{E}_p[\mathbf{x}]$ and $\boldsymbol{\Sigma}$ is given by $\mathrm{Var}_p(\mathbf{x})$.[4] It is not necessary to perform any second derivative tests. (If you cannot do this for general d, for partial marks do it for $d = 1$).

[4] It might help to recall that if $\mathbf{X} \in \mathbb{R}^{d \times d}$ is invertible and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, then $\frac{\mathrm{d}}{\mathrm{d}\mathbf{X}} \mathbf{a}^\top \mathbf{X}^{-1} \mathbf{b} = -\mathbf{X}^{-\top} \mathbf{a}\mathbf{b}^\top \mathbf{X}^{-\top}$ and $\frac{\mathrm{d}}{\mathrm{d}\mathbf{X}} \det(\mathbf{X}) = \det(\mathbf{X}) \cdot \mathbf{X}^{-\top}$, where $\mathbf{X}^{-\top} = (\mathbf{X}^\top)^{-1}$