Optimization for AI **Name:** ____________________

Fall 2020

Final **ID:** ____________________

This exam contains 2 pages (including this cover page) and 10 questions.

1. (10 points) Consider a concave function $f : \mathbb{R}^d \to \mathbb{R}$. Show that $\{x \in \mathbb{R}^d : f(x) \geq a\}$ is a convex set.

2. (10 points) Let $x \in \mathbb{R}^d$ be a d dimensional vector and x_i be the i-th element of x. Show that $f(x) = \sum_{i=1}^d \sum_{j=1}^d x_i x_j$ is a convex function.

3. (10 points) Consider two convex function f and g. Assume that $\{x : g(x) \leq 0\}$ is not empty. Show that there exists $\lambda \geq 0$ such that
$$f(x^\star) = \min_{x:g(x)\leq 0} f(x) \quad \text{where} \quad x^\star = \arg\min_x f(x) + \lambda g(x).$$

4. (10 points) Show that
$$(1 - \frac{1}{\kappa})^T \leq \exp(-\frac{T}{\kappa}).$$

5. (10 points) In TensorFlow, the Nesterov accelerated gradient method updates parameter θ as follows:
$$\begin{aligned} v_t &= \gamma v_{t-1} + \eta \nabla L(\theta_t - \gamma v_{t-1}) \\ \theta_{t+1} &= \theta_t - v_t, \end{aligned}$$
where L is the α-strongly convex and β-smooth loss function. Let $\kappa = \frac{\beta}{\alpha}$ be the condition number of L. Find the convergence speed of the Nesterov accelerated gradient method with the optimal η and γ for the given κ.

6. (10 points) When $f(x)$ has the same Hessian matrix for all x, the mirror descent can reproduce Newton's method. Find a mirror map that can make the mirror descent equal to Newton's method.

7. (10 points) Show that f is $\sum_{i=1}^d \beta_i$ smooth when f is β_i smooth with respect to the i-th coordinate for all i.

8. (10 points) Consider a twice differentiable function f. Assume that all diagonal elements of $\nabla^2 f(x)$ are 1 for all x. Then, is it possible that f is 2-strongly convex?

9. (10 points) Consider a Lasso regression problem:
$$\min_x \|Ax - b\|^2 \quad s.t. \quad \|x\|_1 \leq 1.$$
We would like to optimize the Lasso regression problem using Frank-Wolfe. Describe the full details of Frank-Wolfe that solves the Lasso regression problem and find the convergence rate (i.e., $f(x_T) - f(x^\star) \leq ?$) with finding the appropriate step size.

10. (10 points) Policy gradients with parameter θ in general leverage a baseline function to reduce the variance of stochastic gradients. Show that

$$\mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta \log(\pi_\theta(a_t|s_t))b(s_t)] = 0,$$

where $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, a_{T-1}, r_{T-1}, s_T)$ is the trajectory played by π_θ policy that define the action distribution for each state.