CME 307/MS&E 311/OIT 676
Final

December 9, 2024

Name: _____

SUNetID: _____

(Your email address is yourSUNetID@stanford.edu.)

Instructions

This is a 90-minute exam that consists of three questions. Please answer them to the best of your ability.

Problem 1. Transforming optimization problems using duality
Let $c \in \mathbf{R}^n$, $a \in \mathbf{R}^n$, $P \in \mathbf{R}^{n \times m}$, $z \in \mathbf{R}^m$, $b \in \mathbf{R}$, and $x \in \mathbf{R}^n$.

(a) Consider the following optimization problem (x is the optimization variable):

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & a^T x + \max_{z:\|z\|_2 \leq \rho} (P^T x)^T z \leq b. \end{array} \tag{1}$$

Prove that the optimization problem (1) is equivalent to

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & a^T x + \rho \|P^T x\|_2 \leq b. \end{array} \tag{2}$$

What kind of optimization problem is (2) among the following classes: LP, QP, SOCP, SDP? Please give the narrowest possible designation (e.g., if the problem is an LP, do not say it is a QP).

(b) Consider the following optimization problem (x is the optimization variable):

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & a^T x + \max_{z:\|z\|_\infty \leq \rho} (P^T x)^T z \leq b. \end{array} \tag{3}$$

Transform the optimization problem (3) into a LP. The number of variables and constraints in your LP must be linear in n and/or m.

Solution.

(a) Using either duality or Cauchy-Schwarz, we see that $\max_{z:\|z\|_2 \leq \rho} (P^T x)^T z = \rho \|P^T x\|_2$. The result follows.

The optimization problem (2) is a SOCP.

(b) Using either duality or Hölder's inequality, we see that $\max_{z:\|z\|_\infty \leq \rho} (P^T x)^T z = \rho \|P^T x\|_1$.

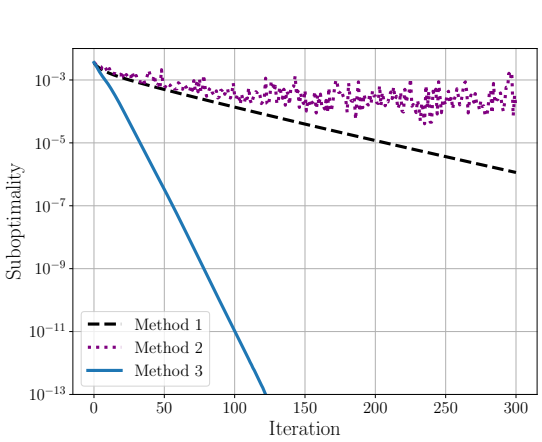
The resulting constraint is $a^T x + \rho \|P^T x\|_1 \leq b$, but this constraint is not linear. To make the constraint linear, we introduce the variable $y \in \mathbf{R}^m$ and add the constraint $-y \leq P^T x \leq y$. Then the constraint $a^T x + \rho \|P^T x\|_1 \leq b$ can be transformed into $a^T x + \rho \mathbf{1}_m^T y \leq b$.

The resulting LP is

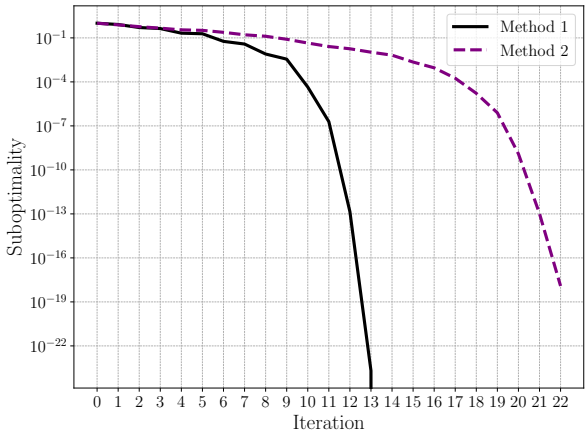
$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & a^T x + \rho \mathbf{1}_m^T y \leq b \\ & y \geq -P^T x \\ & y \geq P^T x. \end{array} \tag{4}$$

The number of variables in the LP (4) is $m+n$ and the number of constraints is $2m+1$. Both of these quantities are linear in m and n.

Problem 2. Optimization algorithms



(a) Suboptimality of three 1st-order methods



(b) Suboptimality of a quasi-Newton and a 2nd-order method

Figure 1: Suboptimality plots for several optimizers.

(a) Fig. 1a plots the suboptimality of three different first-order methods on an objective function that is L-smooth and μ-strongly convex. Based on the suboptimality trajectories, match each of Method 1, Method 2, and Method 3 to *one* of the following optimizers: Accelerated Gradient Descent, BFGS, Gradient Descent, and Stochastic Gradient Descent. Provide a 2-3 sentence answer justifying your selections.

(b) Fig. 1b plots the suboptimality of a quasi-Newton method and a second-order method on an L-smooth and μ-strongly convex function. Based on the suboptimality trajectories, please select two algorithms we discussed in class and assign them to Method 1 and Method 2, respectively. Provide a 2-3 sentence answer justifying your assignments.

(c) You are a research scientist at the social media company Atem. You have been tasked with optimizing ad placement to maximize the number of clicks and keep advertisers happy. Atem has built a powerful neural network for click prediction to aid in this task. Making sure the predictive model is consistent with the latest data is essential for good performance. However, retraining the model from scratch as new data comes is too expensive. It turns out, the 1048576 features learned by Atem's model remain stable over time as new data comes in. Thus, only the last binary classification layer must be updated, leading to a logistic regression problem with a variable of dimension 1048576. Note, this logistic regression problem involves dense data, as the feature vectors output by the network are dense.

Consider the following two scenarios for updating the weights of the classification layer:

(i) The model is updated monthly based on a dataset of 20000000 samples formed from all past click data.

(ii) The model is updated as soon as a new batch of click data is processed, and the data batch is discarded.

For scenarios (i) and (ii), please identify an algorithm we discussed in class that you think best fits that scenario. For each scenario, provide a 2-3 sentence answer justifying your choice of algorithm.

Hint: How does scenario (ii) differ from scenario (i)?

Solution.

(a) We can eliminate BFGS as it is a quasi-Newton Method. We are told that each method is run with a fixed step size. From this, it is clear Method 2 must be SGD, as it is the only method that does not converge to the optimum with a fixed step size. Both AGD and GD converge linearly in this setting, but AGD converges at a faster rate. Therefore Method 1 is GD, and Method 2 is AGD.

(b) Method 1 is Newton's method, and Method 2 could be BFGS or L-BFGS. Flipping these labelings is also perfectly acceptable, as Newton's method is not always guaranteed to do better than BFGS or L-BFGS.

(c) (i) For this item, the only reasonable answers are SGD or SVRG. Full-gradient methods should be eliminated purely from an efficiency standpoint. Newton's method and friends make no sense at all, given the problem's dimensionality.

(ii) SGD is the only sensible method, as in the online setting, there is no notion of a "full gradient". So, none of the other methods we discussed in the course make sense.

Problem 3. Discrete optimization
Consider a finite set of elements $N := \{1, 2, \ldots, n\}$, a vector $r \in \mathbb{R}^n$ satisfying $r > 0$ and another vector $c \in \mathbb{R}^n$, and the following function $f : 2^N \to \mathbb{R}$:

$$f(S) = \log\left(1 + \sum_{i \in S} r_i\right) - \sum_{i \in S} c_i, \, \forall \, S \subseteq N.$$

Moreover, define the following polyhedral set $P(f)$:

$$P(f) = \Big\{x \in \mathbb{R}^n \ : \ \sum_{i \in S} x_i \leq f(S), \, \forall \, S \subseteq N\Big\}.$$

Consider a vector $w \in \mathbb{R}^n$ with $w \geq 0$.

(a) Show that f is a submodular function.

(b) What is an optimal solution and the optimal value of the problem $\max_{x \in P(f)} w^T x$? Provide a proof for your answers.

(c) Suppose we sort the elements of w in decreasing order, i.e., $w_{j_1} \geq w_{j_2} \geq \cdots \geq w_{j_n}$, where $j_1, \ldots, j_n$ is a permutation of $1, \ldots, n$. Define the function $\hat{f} : \mathbb{R}^n \to \mathbb{R}$:

$$\hat{f}(w) = \sum_{i=1}^{n} w_{j_k} \cdot \big[\big(f(\{j_1, \ldots, j_k\}) - f(\{j_1, \ldots, j_{k-1}\})\big]$$

Use the result in part (b) to prove that the function $\hat{f}(w)$ is convex.

Solution.

(a) Notice that $1 + \sum_{i \in S} r_i$ is linear, and the weights r_i are all positive. Since $\log(\cdot)$ is a concave function, it follows that $f_1(S) \coloneqq \log(1 + \sum_{i \in S} r_i)$ is submodular. Moreover, $f_2(S) \coloneqq -\sum_{i \in S} c_i$ is linear, so it is submodular. Since sums of submodular functions are submodular, we conclude that $f_1(S) + f_2(S) = f(S)$ is submodular.

(b) The solution is similar to the example given in the lecture notes and problem 2 on HW3. For brevity, we do not present the full proof here. The optimal solution is

$$x_i^\star = f(S^i) - f(S^{i-1}) \quad \forall i \in \{1, 2, \ldots, n\},$$

where $S^0 = \emptyset$ and $S^j \coloneqq \{1, 2, \ldots, j\}$ for $1 \leq j \leq n$. The optimal value is (assuming w is sorted in decreasing order)

$$w^T x^\star = \sum_{j=1}^{n} w_j (f(S^j) - f(S^{j-1})).$$

(c) From part (b), we have $\hat{f}(w) = \max_{x \in P(f)} w^T x$. For a fixed x, $w^T x$ is a convex function of w. It follows that $\hat{f}(w)$ is a pointwise maximum of convex functions, and is therefore convex.