Homework 2

Convex Optimization 10-725

Due Friday September 27 at 11:59pm

Submit your work as a single PDF on Gradescope. Make sure to prepare your solution to each problem on a separate page. (Gradescope will ask you select the pages which contain the solution to each problem.)

Total: 80 points (+ 10 bonus points)

1 Gradient descent convergence analysis (18 points)

In this problem, we will analyze gradient descent under suitable assumptions. Consider minimizing a differentiable function f with $\mathrm{dom}(f) = \mathbb{R}^n$, whose gradient is L-Lipschitz continuous for a constant $L > 0$, meaning

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2, \quad \text{for all } x, y.$$

We will run gradient descent, starting from $x^{(0)}$, with the updates

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots,$$

where each $t_k \le 1/L$. As usual, we will write a generic update as $x^+ = x - t\nabla f(x)$, where $t \le 1/L$.

1.1 Nonconvex case (8 points)

Here we will assume nothing about convexity of f. We will show that gradient descent reaches an ϵ-substationary point x, such that $\|\nabla f(x)\|_2 \le \epsilon$, in $O(1/\epsilon^2)$ iterations. Important note: you may use here that

$$f(y) \le f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|_2^2, \quad \text{for all } x, y. \tag{1}$$

Recall that you assumed convexity and twice differentiability of f on Homework 1 to show that the above is equivalent to the L-Lipschitz condition on ∇f. But (1) is in fact a consequence of ∇f being L-Lipschitz, and does not actually require convexity or twice differentiability of f.

(a, 2 pts) Plug in $y = x^+ = x - t\nabla f(x)$ to (1) to show that

$$f(x^+) \le f(x) - \Big(1 - \frac{Lt}{2}\Big) t \|\nabla f(x)\|_2^2.$$

(b, 2 pts) Use $t \le 1/L$, and rearrange the previous result, to get

$$\|\nabla f(x)\|_2^2 \le \frac{2}{t}(f(x) - f(x^+)).$$

(c, 2 pts) Sum the previous result over all iterations from $1, \dots, k+1$ to establish

$$\sum_{i=0}^{k} \|\nabla f(x^{(i)})\|_2^2 \leq \frac{2}{t}(f(x^{(0)}) - f^\star).$$

(d, 2 pts) Lower bound the sum in the previous result to get

$$\min_{i=0,\dots,k} \|\nabla f(x^{(i)})\|_2 \leq \sqrt{\frac{2}{t(k+1)}(f(x^{(0)}) - f^\star)},$$

which establishes the desired $O(1/\epsilon^2)$ rate for achieving ϵ-substationarity.

1.2 Convex case (10 points)

Now we will assume that f is convex. We will show that gradient descent reaches an ϵ-suboptimal point x, such that $f(x) - f^\star \leq \epsilon$, in $O(1/\epsilon)$ iterations. Going back to part (b) from the nonconvex case, we can rearrange this to get

$$f(x^+) \leq f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2. \tag{2}$$

Note that, by this property, we see that gradient descent is indeed a descent method for $t \leq 1/L$ (it decreases the criterion at each iteration).

(a, 3 pts) Starting with (2), apply the first-order condition for convexity of f, to show

$$f(x^+) \leq f^\star + \nabla f(x)^T(x - x^\star) - \frac{t}{2}\|\nabla f(x)\|_2^2.$$

(b, 3 pts) From the previous result, show that

$$f(x^+) \leq f^\star + \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x^+ - x^\star\|_2^2\right).$$

(c, 2 pts) Sum the previous result over all iterations $1, \dots, k$ to get

$$\sum_{i=1}^{k}(f(x^{(i)}) - f^\star) \leq \frac{1}{2t}\|x^{(0)} - x^\star\|_2^2.$$

(d, 2 pts) Use the fact that gradient descent is a descent method to lower bound the sum above, and conclude

$$f(x^{(k)}) - f^\star \leq \frac{\|x^{(0)} - x^\star\|_2^2}{2tk},$$

which establishes the desired $O(1/\epsilon)$ rate for achieving ϵ-suboptimality.

2 Properties and examples of subgradients (18 points)

We will inspect various properties and examples of subgradients.

(a, 2 pts) Show that $\partial f(x)$ is a closed and convex set for any function f (not necessarily convex) and any point x in its domain.

(b, 2 pts) Show that $g \in \partial f(x)$ if and only if $(g, -1)$ defines supporting hyperplane to epigraph of f at $(x, f(x))$ (i.e., $(g, -1)$ is the normal vector to this hyperplane).

(c, 2 pts) For a convex function f, show that if $x \in U$ where U is a open neighborhood in its domain, then

$$f(y) \geq f(x) + g^T(y - x), \quad \text{for all } y \in U \quad \Rightarrow \quad g \in \partial f(x).$$

In other words, if the tangent line inequality holds in a local open neighborhood of x, then it holds globally.

(d, 1 pt) For a convex function f and subgradients $g_x \in \partial f(x)$, $g_y \in \partial f(y)$, prove that

$$(g_x - g_y)^T(x - y) \geq 0.$$

This property is called *monotonicity* of the subdifferential ∂f.

(e, 2 pts) For $f(x) = \|x\|_2$, show that all subgradients $g \in \mathbb{R}^n$ at a point $x \in \mathbb{R}^n$ are of the form

$$g \in \begin{cases} \{x/\|x\|_2\} & x \neq 0 \\ \{v : \|v\|_2 \leq 1\} & x = 0. \end{cases}$$

(f, 3 pts) For $f(x) = \max_{s \in S} f_s(x)$, where each f_s is convex, show that

$$\partial f(x) \supseteq \operatorname{conv}\Bigg(\bigcup_{s : f_s(x) = f(x)} \partial f_s(x) \Bigg).$$

Bonus (4 pts): when S is a discrete set, prove the other direction.

(g, 6 pts): For $f(X) = \|X\|_{\text{tr}}$, show that subgradients at $X = U\Sigma V^T$ (this is an SVD of X) satisfy

$$\partial f(X) \supseteq \{UV^T + W : \|W\|_{\text{op}} \leq 1,\ U^T W = 0,\ WV = 0\}.$$

Hint: you may use the fact that $\|\cdot\|_{\text{tr}}$ and $\|\cdot\|_{\text{op}}$ are dual norms, which implies $\langle A, B\rangle \leq \|A\|_{\text{tr}}\|B\|_{\text{op}}$ for any matrices A, B, where recall $\langle A, B\rangle = \text{tr}(A^T B)$. **Bonus (5 pts):** prove the other direction.

3 Properties and examples of proximal operators (22 points)

We will inspect various properties and examples of proximal operators. Unless otherwise specified, take h to be a convex function with domain $\text{dom}(h) = \mathbb{R}^n$, and $t > 0$ be arbitrary, and consider its associated proximal operator

$$\text{prox}_{h,t}(x) = \underset{z}{\text{argmin}}\ \frac{1}{2t}\|x - z\|_2^2 + h(z).$$

(a, 3 pts) Prove that $\text{prox}_{h,t}$ is a well-defined function on $\mathbb{R}^n$, that is, each point $x \in \mathbb{R}^n$ gets mapped to a unique value $\text{prox}_{h,t}(x)$.

(b, 2 pts) Prove that $\text{prox}_{h,t}(x) = u$ if and only if

$$h(y) \geq h(u) + \frac{1}{t}(x-u)^T(y-u), \quad \text{for all } y.$$

Hint: use subgradient optimality.

(c, 6 pts) Prove that $\text{prox}_{h,t}$ is nonexpansive, meaning

$$\|\text{prox}_{h,t}(x) - \text{prox}_{h,t}(y)\|_2 \leq \|x-y\|_2, \quad \text{for all } x, y.$$

Hint: use the previous question, and the monotonicity of subgradients from Q2(d).

(d, 3 pts) The proximal minimization algorithm (a special case of proximal gradient descent) repeats the updates:

$$x^{(k+1)} = \text{prox}_{h,t}(x^{(k)}), \quad k = 1, 2, 3, \dots.$$

Write out these updates when applied to $h(x) = \frac{1}{2}x^T Ax - b^T x$, where $A \in \mathbb{S}^n$. Show that this is equivalent to the *iterative refinement* algorithm for solving the linear system $Ax = b$:

$$x^{(k+1)} = x^{(k)} + (A + \epsilon I)^{-1}(b - Ax^{(k)}), \quad k = 1, 2, 3, \dots,$$

where $\epsilon > 0$ is some constant. **Bonus (1 pt):** assuming that proximal minimization converges to the minimizer of $h(x) = \frac{1}{2}x^T Ax - b^T x$ (which is does, under suitable step sizes), what would the iterations of iterative refinement converge to in the case when A is singular, $Ax = b$, and $x^{(0)} = 0$?

(e, 8 pts) For a matrix-variate function h, we define its proximal operator as

$$\text{prox}_{h,t}(X) = \underset{Z}{\text{argmin}} \ \frac{1}{2t}\|X - Z\|_F^2 + h(Z),$$

For $h(X) = \|X\|_{\text{tr}}$, show that the proximal operator evaluated at $X = U\Sigma V^T$ (this is an SVD of X) is so-called matrix soft-thresholding,

$$\text{prox}_{h,t}(X) = U\Sigma_t V^T, \quad \text{where } \Sigma_t = \text{diag}\Big((\Sigma_{11} - t)_+, \dots, (\Sigma_{nn} - t)_+\Big),$$

and $x_+ = \max\{x, 0\}$ denotes the positive part of x. Hint: start with subgradient optimality as you developed in Q3(b), and use the subgradients of the trace norm from Q2(g).

4 Group lasso logistic regression (22 points)

Suppose we have features $X \in \mathbb{R}^{n \times (p+1)}$ that we divide into J groups:

$$X = \Big[\mathbb{1} \ X_{(1)} \ X_{(2)} \ \cdots \ X_{(J)}\Big],$$

where $\mathbb{1} = (1, \dots, 1) \in \mathbb{R}^n$ and each $X_{(j)} \in \mathbb{R}^{n \times p_j}$. To achieve sparsity over groups of features, rather than individual features, we can use a *group lasso* penalty. Write $\beta = (\beta_0, \beta_{(1)}, \dots, \beta_{(J)}) \in \mathbb{R}^{p+1}$, where β_0 is an intercept term and each $\beta_{(j)} \in \mathbb{R}^{p_j}$. Consider the problem

$$\min_{\beta} \ g(\beta) + \lambda \sum_{j=1}^{J} w_j \|\beta_{(j)}\|_2, \tag{3}$$

where g is a loss function and $\lambda \geq 0$ is a tuning parameter. The penalty $h(\beta) = \lambda \sum_{j=1}^{J} w_j \|\beta_{(j)}\|_2$ is called the group lasso penalty. A common choice for w_j is $\sqrt{p_j}$ to adjust for the group size.

(a, 3 pts) Derive the proximal operator $\text{prox}_{h,t}(\beta)$ for the group lasso penalty defined above.

(b, 2 pts) Let $y \in \{0,1\}^n$ be a binary label, and let g be the logistic loss

$$g(\beta) = -\sum_{i=1}^{n} y_i (X\beta)_i + \sum_{i=1}^{n} \log(1 + \exp\{(X\beta)_i\}),$$

Write out the steps for proximal gradient descent applied to the logistic group lasso problem (3) in explicit detail.

(c, 5 pts) Now we'll use the logistic group lasso to classify a person's age group from his movie ratings. The movie ratings can be categorized into groups according to a movie's genre (e.g., all ratings for action movies can be grouped together). Load the training data in `trainRatings.txt`, `trainLabels.txt`; the features have already been arranged into groups and you can find information about this in `groupTitles.txt`, `groupLabelsPerRating.txt`. Solve the logistic group lasso problem (3) with regularization parameter $\lambda = 5$ by running proximal gradient descent for 1000 iterations with fixed step size $t = 10^{-4}$. Plot $f^{(k)} - f^\star$ versus k, where $f^{(k)}$ denotes the objective value at iteration k, and use as an optimal objective value $f^\star = 336.207$. Make sure the plot is on a semi-log scale (where the y-axis is in log scale).

(d, 5 pts) Now implement Nesterov acceleration for the same problem. You should again run accelerated proximal gradient descent for 1000 iterations with fixed step size $t = 10^{-4}$. As before, produce a plot $f^{(k)} - f^\star$ versus k. Describe any differences you see in the criterion convergence curve.

(e, 5 pts) Lastly, implement backtracking line search (rather than a fixed step size), and rerun proximal gradient for 400 iterations, without acceleration. (Note this means 400 outer iterations; the backtracking loop itself can take several inner iterations.) You should set $\beta = 0.1$ and $\alpha = 0.5$. Produce a plot of $f^{(k)} - f^\star$ versus $i(k)$, where $i(k)$ counts the *total* number of iterations performed at outer iteration k (total, meaning the sum of the iterations in both the inner and outer loops).

Note: since it makes for an easier comparison, you can draw the convergence curves from (c), (d), (e) on the same plot.

(f, 2 pts) Finally, use the solution from accelerated proximal gradient descent in part (d) to make predictions on the test set, available in `testRatings.txt`, `testLabels.txt`. What is the classification error? What movie genre are important for classifying whether a viewer is under 40 years old?