CPSC 540 Assignment 2 (due February 1 at midnight)

The assignment instructions are the same as for the previous assignment, but for this assignment you can work in groups of 1-3. However, please only hand in one assignment for the group.

1. Name(s): Zhen Wang, Hongru Li
2. Student ID(s): 98552169

1 Calculation Questions

1.1 Convexity

Show that the following functions are convex, by only using one of the definitions of convexity (i.e., without using the "operations that preserve convexity" or using convexity results stated in class):[1]

1. L2-regularized weighted least squares: $f(w) = \frac{1}{2}(Xw - y)^\top V(Xw - y) + \frac{\lambda}{2}\|w\|^2$.
   (V is a diagonal matrix with positive values on the diagonal).
   Answer: $\nabla^2 f(w) = X^\top VX + \lambda I > 0$. Thus $f(w)$ is convex.

2. Poisson regression: $f(w) = -y^\top Xw + 1^\top v$ (where $v_i = \exp(w^\top x^i)$).
   Answer: $f(w) = -y^\top Xw + \sum v_i$, $\nabla^2 f(w) = 0 + r$ where r is $d * 1$ matrix and each $d_j = \sum_{i=1}^n (x_j^i)^2 exp(w^\top x^i) > 0$. Thus $\nabla^2 f(w) > 0$. Thus $f(w)$ is convex.

3. Weighted infinity-norm: $f(w) = \max_{j \in \{1,2,\ldots,d\}} L_j |w_j|$.
   Hint: Max and aboluste value are not differentiable in general, so you cannot use the Hessian for this question.
   Answer:

$$
\begin{aligned}
f(\theta w + (1-\theta)v) &= \max_{j \in \{1,2,\ldots,d\}} L_j |\theta w_j + (1-\theta) v_j| \\
&\leq \max_{j \in \{1,2,\ldots,d\}} L_j (\theta |w_j| + (1-\theta)|v_j|) \\
&= \max_{j \in \{1,2,\ldots,d\}} \theta L_j |w_j| + (1-\theta) L_j |v_j| \\
&\leq \max_{j \in \{1,2,\ldots,d\}} \theta L_j |w_j| + \max_{j \in \{1,2,\ldots,d\}} (1-\theta) L_j |v_j| \\
&= \theta f(w) + (1-\theta) f(v)
\end{aligned}
$$

   . Thus it is convex.

Show that the following functions are convex (you can use results from class and operations that preserve convexity if they help):

4. Regularized regression with arbitrary p-norm and weighted q-norm: $f(w) = \|Xw - y\|_p + \lambda \|Aw\|_q$.
   Answer: All norms are convex and sum of convex functions is convex. Thus $f(w)$ is convex.

[1] That C^0 convex functions are below their chords, that C^1 convex functions are above their tangents, or that C^2 convex functions have a positive semidefinite Hessian.

5. Support vector regression: $f(w) = \sum_{i=1}^{N} \max\{0, |w^\top x_i - y_i| - \epsilon\} + \frac{\lambda}{2}\|w\|_2^2$.
Answer: Square of norm is convex thus $\frac{\lambda}{2}\|w\|_2^2$ is convex for $\lambda > 0$. $|w^\top x_i - y_i|$ is convex thus $|w^\top x_i - y_i| - \epsilon$ is convex. As 0 is convex and maximum of convex functions is also convex, thus $\max\{0, |w^\top x_i - y_i| - \epsilon\}$ is convex. Sum of convex functions is convex, thus $f(w)$ is convex.

6. Indicator function for linear constraints: $f(w) = \begin{cases} 0 & \text{if } Aw \leq b \\ \infty & \text{otherwise} \end{cases}$.
Answer: If $Aw \leq b$ and $Av \leq b$, $A(\theta w + (1-\theta)v) \leq A(\max(w, v)) \leq b$, so $f(\theta w + (1-\theta)v) = 0 = \theta f(w) + (1-\theta)f(v)$. If one or both of $Aw > b$ and $Av > b$, $\theta f(w) + (1-\theta)f(v) = \infty$. So $f(\theta w + (1-\theta)v) \leq \theta f(w) + (1-\theta)f(v)$. Thus $f(w)$ is convex.

1.2 Convergence of Gradient Descent

For these questions it will be helpful to use the "convexity inequalities" notes posted on the webpage.

1. In class we showed that if ∇f is L-Lipschitz continuous and f is bounded below then with a step-size of $1/L$ gradient descent is guaranteed to have found a w^k with $\|\nabla f(w^k)\|^2 \leq \epsilon$ after $t = O(1/\epsilon)$ iterations. Suppose that a more-clever algorithm exists which, on iteration t, is guaranteed to have found a w^k satisfying $\|\nabla f(w^k)\|^2 \leq 2L(f(w^0) - f^*)/t^{4/3}$. How many iterations of this algorithm would we need to find a w^k with $\|\nabla f(w^k)\|^2 \leq \epsilon$?
Answer:$\|\nabla f(w^k)\|^2 \leq 2L(f(w^0) - f^*)/t^{4/3} \leq \epsilon$. Thus $t \geq (\epsilon/(2L(f(w^0) - f^*)))^{3/4}$.

2. In practice we typically don't know L. A common strategy in this setting is to start with some small guess L^0 that we know is smaller than the true L (usually we take $L^0 = 1$). On each iteration k, we initialize with $L^k = L^{k-1}$ and we check the inequality

$$f\left(w^k - \frac{1}{L^k}\nabla f(w^k)\right) \leq f(w^k) - \frac{1}{2L^k}\|\nabla f(w^k)\|^2.$$

If this is not satisfied, we double L^k and test it again. This continues until we have an L^k satisfying the inequality, and then we take the step. Show that gradient descent with $\alpha_k = 1/L^k$ defined in this way has a linear convergence rate of

$$f(w^k) - f(w^*) \leq \left(1 - \frac{\mu}{2L}\right)^k [f(w^0) - f(w^*)],$$

if ∇f is L-Lipschitz continuousn and f is μ-strongly convex.
Hint: if a function is L-Lipschitz continuous that it is also L'-Lipschitz continuous for any $L' \geq L$.
Answer:

$$\begin{aligned} f(w^k) - f(w^*) &\leq f(w^{k-1}) - f(w^*) - \frac{1}{2L^{k-1}}(2\mu(f(w^{k-1}) - f(w^*))) \\ &= (1 - \frac{\mu}{L^{k-1}})(f(w^{k-1}) - f(w^*)) \\ &\leq \prod_{i=0}^{k-1}(1 - \frac{\mu}{L^i})(f(w^0) - f(w^*)) \\ &\leq \left(1 - \frac{\mu}{2L}\right)^k [f(w^0) - f(w^*)] \end{aligned}$$

3. Suppose that, in the previous question, we initialized with $L^k = \frac{1}{2}L^{k-1}$. Describe a setting where this could work much better.

4. In class we showed that if ∇f is L-Lipschitz continuous and f is strongly-convex, then with a step-size of $\alpha_k = 1/L$ gradient descent has a convergence rate of

$$f(w^k) - f(w^*) = O(\rho^k).$$

Show that under these assumptions that a convergence rate of $O(\rho^k)$ in terms of the function values implies that the iterations have a convergence rate of

$$\|w^k - w^*\| = O(\rho^{k/2}).$$

1.3 Beyond Gradient Descent

1. We can write the proximal-gradient update as

$$\begin{aligned} w^{k+\frac{1}{2}} &= w^k - \alpha_k \nabla f(w^k) \\ w^{k+1} &= \underset{v\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2}\|v - w^{k+\frac{1}{2}}\|^2 + \alpha_k r(v) \right\}. \end{aligned}$$

Show that this is equivalent to setting

$$w^{k+1} \in \underset{v\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ f(w^k) + \nabla f(w^k)^\top (v - w^k) + \frac{1}{2\alpha_k}\|v - w^k\|^2 + r(v) \right\}.$$

Answer:

$$\begin{aligned} w^{k+1} &\in \underset{v\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ f(w^k) + \nabla f(w^k)^\top (v - w^k) + \frac{1}{2\alpha_k}\|v - w^k\|^2 + r(v) \right\} \\ &\in \underset{v\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \alpha_k f(w^k) + \alpha_k \nabla f(w^k)^\top (v - w^k) + \frac{1}{2}\|v - w^k\|^2 + \alpha_k r(v) \right\} \\ &\in \underset{v\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ 2\alpha_k f(w^k) + 2\alpha_k \nabla f(w^k)^\top (v - w^k) + \|v - w^k\|^2 + 2\alpha_k r(v) \right\} \\ &\in \underset{v\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \alpha_k^2 (\nabla f(w^k))^2 + 2\alpha_k \nabla f(w^k)^\top (v - w^k) + \|v - w^k\|^2 + 2\alpha_k r(v) \right\} \\ &\in \underset{v\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2}(\|v - w^k + \alpha_k \nabla f(w^k)\|^2) + \alpha_k r(v) \right\} \\ &\in \underset{v\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2}(\|v - (w^k - \alpha_k \nabla f(w^k)\|^2)) + \alpha_k r(v) \right\} \end{aligned}$$

This is the same as

$$\begin{aligned} w^{k+\frac{1}{2}} &= w^k - \alpha_k \nabla f(w^k) \\ w^{k+1} &= \underset{v\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2}\|v - w^{k+\frac{1}{2}}\|^2 + \alpha_k r(v) \right\}. \end{aligned}$$

2. The "sum" version of multi-class SVMs uses an objective of the form

$$f(W) = \sum_{i=1}^{n} \sum_{c\neq y^i} [1 - w_{y^i}^\top x^i + w_c^\top x^i]^+ + \frac{\lambda}{2}\|W\|_F^2,$$

where $[\gamma]^+$ sets negative values to zero (and you can use k as the number of classes so the inner loop is over $(k-1)$ elements). Derive the sub-differential of this objetive.

Answer: Let $L_i = \sum_{c \neq y^i}[1 - w_{y^i}^\top x^i + w_c^\top x^i]^+$. The sub-gradient of L_i is

$$f(w) = \begin{cases} 0 & \text{if } Aw \leq b \\ \infty & \text{otherwise} \end{cases}$$

3. In some situations it might be hard to accurately compute the elements of the gradient, but we might have access to the sign of the gradient (this can also be useful in distributed settings where communicating one bit for each element of the gradient is cheaper than communicating a floating point number for each gradient element). Consider an f that is bounded below and where ∇f is Lipschitz continuous in the ∞-norm, meaning that

$$f(v) \leq f(u) + \nabla f(u)^\top (v - u) + \frac{L_\infty}{2}\|v - u\|_\infty^2,$$

for all v and w and some L_∞. For this setting, consider a sign-based gradient descent algorithm of the form

$$w^{k+1} = w^k - \frac{\|\nabla f(w^k)\|_1}{L_\infty}\text{sign}(\nabla f(w^k)),$$

where we define the sign function element-wise as

$$\text{sign}(w_j) = \begin{cases} +1 & w_j > 0 \\ 0 & w_j = 0 \\ -1 & w_j < 0 \end{cases},$$

Show that this sign-based gradient descent algorithm finds a w^k satisfying $\|\nabla f(w^k)\|^2 \leq \epsilon$ after $t = O(1/\epsilon)$ iterations.

2 Computation Questions

Coming soon....

3 Very-Short Answer Questions

Coming soon...