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):
2. Student ID(s):

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).
2. Poisson regression: $f(w) = -y^\top Xw + 1^\top v$ (where $v_i = \exp(w^\top x^i)$).
3. Weighted infinity-norm: $f(w) = \max_{j \in \{1,2,\dots,d\}} L_j |w_j|$.
   Hint: Max and aboluste value are not differentiable in general, so you cannot use the Hessian for this question.

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: $f(w) = \|Xw - y\|_p + \lambda\|Aw\|_q$.
5. Support vector regression: $f(w) = \sum_{i=1}^N \max\{0, |w^T x_i - y_i| - \epsilon\} + \frac{\lambda}{2}\|w\|_2^2$.
6. Indicator function for linear constraints: $f(w) = \begin{cases} 0 & \text{if } Aw \le b \\ \infty & \text{otherwise} \end{cases}$.

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 $\|\nabla f(w^k)\|^2 \le \epsilon$ 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 $\|\nabla f(w^k)\|^2 \le 2L(f(w^0) - f^*)/t^{4/3}$. How many iterations of this algorithm would we need to find a w^k with $\|\nabla f(w^k)\|^2 \le \epsilon$?
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

[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.

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 $L' \geq L$.

3. Suppose that, in the previous question, we initialized with $L^k = 2L^{k-1}$. 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

Coming soon...

2 Computation Questions

Coming soon....

3 Very-Short Answer Questions

Coming soon...