Convex Functions and Gradient Descent

Recall that a function $f$ over $\mathbb{R}^n$ is convex if for any two inputs $x, y \in \mathbb{R}^n$ and any $\lambda \in [0, 1]$ we have $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$. In other words, the line segment from $(x, f(x))$ to $(y, f(y))$ stays above the function.

1. A nice feature about convex functions is that any local minimum is a global minimum. Indeed, show that if $x$ is not a global minimum, there is some direction in which the slope is negative.

   **Solution:** Suppose $x$ is not a global minimum and instead the global minimum is $y$. Then the line segment from $(x, f(x))$ to $(y, f(y))$ has negative slope. Since $f$ stays below this line, it too must have negative slope in this direction at $x$.

   This motivates finding a direction of negative slope and moving in that direction. The gradient of $f$ at $x$, denoted by $\nabla f(x)$ gives the direction of the greatest positive slope, and hence you want to move in the direction of $-\nabla f(x)$.

2. We showed that gradient descent (for both the unconstrained and constrained cases) produced point $\hat{x}$ such that $f(\hat{x}) \leq f(x^*) + \varepsilon$ if you set $\eta = \frac{D}{G\sqrt{T}}$ and run for $T = \left(\frac{DG}{\varepsilon}\right)^2$ steps. This needs knowing $D$ (an upper bound on the distance $\|x_0 - x^*\|$), and $G$ (an upper bound on the gradient), which may not be reasonable for the general problem. But things are better in the constrained case. Suppose you know the function $f(x) = \sum_i c_i x_i$ for some $c = (c_1, \ldots, c_n) \in [0, M]^n$ (i.e., $f$ is linear) and the convex body $K$ is contained within the unit cube: i.e., $K \subseteq \{x \mid 0 \leq x_i \leq 1 \forall i \in \{1, 2, \ldots, n\}\}$.

   (a) What is the diameter of $K$? (The diameter is the maximum Euclidean distance between two points in $K$.)

      **Solution:** The maximum distance is bounded by the max-distance between $(0, 0, \ldots, 0)$ and $(1, 1, \ldots, 1)$, which is $\sqrt{1^2 + 1^2 + \ldots + 1^2} = \sqrt{n}$.

   (b) If you start with some $x_0 \in K$, give an upper bound on $\|x_0 - x^*\|$.

      **Solution:** $\|x_0 - x^*\|$ is at most the diameter of the cube, so setting $D = \sqrt{n}$ suffices.

   (c) What is the maximum value of $\|\nabla f(x)\|$ at any point $x \in K$?

      **Solution:** $\nabla f(x) = \nabla (c_1 x_1 + \ldots + c_n x_n) = c$, so $\|\nabla f(x)\| = \|c\| \leq M\sqrt{n}$. Hence you can set $G = M\sqrt{n}$.
(d) Plugging these values in, what expressions do you get for $T, \eta$?

**Solution:** Recall $T = \left(\frac{DG}{\varepsilon}\right)^2 = \left(\frac{\sqrt{n} M \sqrt{n}}{\varepsilon}\right)^2 = \left(\frac{M n}{\varepsilon}\right)^2$. Substituting, $\eta = \frac{D}{G \sqrt{T}} = \frac{\varepsilon}{M^2 n}$.

3. Suppose you now want to maximize the quadratic function $g(x) = \sum_i c_i x_i^2$ for each $c_i \in [0, M]$, over the same set $K$.

(a) Show the function $g$ is convex. (Prove this in as many ways as you can.)

**Solution:** One way to see it: the univariate function $g_i(x_i) := c_i x_i^2$ is convex. But $g(x) = \sum_i g_i(x_i)$ and the sum of convex functions is convex. (Prove this!) Or use the definitions of convexity directly: either the one above, or that $g(y) \geq g(x) + \langle \nabla g(x), y - x \rangle$. Or the Hessian $\nabla^2 f$ is a non-negative diagonal matrix, and hence positive semi-definite. (Try these, make sure you know how to prove these things.)

(b) What is the maximum value of $\|\nabla g(x)\|$ at any point $x \in K$?

**Solution:** Now $\nabla g(x) = 2(c_1 x_1, \ldots, c_n x_n)$. So the $\|\nabla g(x)\| \leq 2 \max_{x \in K} \sqrt{\sum_i c_i^2 x_i^2} \leq 2M \sqrt{n}$.

4. Suppose $f(x) = \frac{1}{2} x^T A x + b x$ for $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$. Compute the gradient $\nabla f(x)$ and the Hessian $\nabla^2 f(x)$. When is this function convex?

**Solution:** $\nabla f(x) = A x + b$, and $\nabla^2 f(x) = A$. So this function is convex when $A$ is positive semidefinite, when all its eigenvalues are non-negative.

**Multiplicative Weights**

4. In lecture we saw that the simple procedure that multiplied the weight of each expert by $\frac{1}{2}$ whenever the expert made a mistake, resulted in

$$m = \text{#mistakes of algorithm} \leq 2.41(M + \log_2 n),$$

where $M = \text{#mistakes made by the best expert}$ and $n = \# \text{ of experts}$. If we change the weight by $2/3$ at each time, how does this analysis change?

**Solution:** Again, potential is total weight. Every time we make a mistake, total weight goes down by $5/6$. So final weight is $n(5/6)^m$. And every time best expert makes a mistake its weight drops by $2/3$. So $(2/3)^m \geq n(5/6)^m$, and hence

$$m \leq \frac{1}{\log_2(6/5)} (M \log_2 (3/2) + \log_2 n).$$
5. In the lecture: in order to get a better mistake bound of $(1 + \epsilon)M + O\left(\frac{\ln n}{\epsilon}\right)$, we used randomization. Let us now show that you cannot get better than a factor of 2 if you don’t use randomness.

There are two experts. One always predicts 0. The other always predicts 1. Fix any deterministic algorithm $A$ for prediction. Here is one sequence of days: each day, the actual outcome is the opposite of what the algorithm predicts.

After $T$ days, the algorithm would have made $T$ mistakes. Show that the better of the two experts makes at most $T/2$ mistakes. Hence infer that $m \geq 2M$.

**Solution**: Each day exactly one of the two experts is correct. So by the pigeonhole principle, one of them makes $\leq T/2$ mistakes.