CSE 422 Autumn 26 HW 7

Note: For all problems, you will be expected to explain your reasoning. Providing an expression without reasoning will not receive any partial credit.

Problem 1: Better analysis of gradient descent

In class, the professor got lazy and didn't provide a tight analysis of the convergence of gradient descent for L-smooth and ℓ-strongly convex functions. Recall that a function f is L-smooth if for all $x, y \in \mathbb{R}^d$, we have $\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2$, and it is ℓ-strongly convex if for all $x, y \in \mathbb{R}^d$, we have that

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{\ell}{2} \|x - y\|_2^2 \leq f(y) \ .$$

Let x^* be the (unique) global minimizer of f.

(a) [5 points] First, show something slightly stronger than we showed in class: namely, that if f is L-smooth, for all $x, y \in \mathbb{R}^d$, we have the upper bound

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|x - y\|_2^2 \ .$$

(We showed this bound without the factor of $1/2$ in the last term.)

(b) [5 points] Show that for all $x \in \mathbb{R}^d$, we have that

$$\frac{1}{2L} \|\nabla f(x)\|_2^2 \leq f(x) - f(x^*) \ .$$

Hint: consider what happens if you take a gradient step from x.

(c) [5 points] Using the previous bound, show that for all $x \in \mathbb{R}^d$, we have that

$$\frac{1}{L} \|\nabla f(x)\|_2^2 \leq 2\langle \nabla f(x), x - x^* \rangle - \ell \|x - x^*\|_2^2 \ .$$

(d) [5 points] Now, put it all together: let x_0 be an arbitrary initial point, and for $t = 0, \ldots, T-1$, let $x_{t+1} = x_t - \frac{1}{L} \nabla f(x_t)$, so gradient descent with step-size $\eta_t = 1/L$ for all t. Show that

$$\|x_T - x^*\|_2^2 \leq \left(1 - \frac{\ell}{L}\right)^T \|x_0 - x^*\|_2^2 \ .$$

Problem 2: The implicit regularization of gradient descent

Recall from week 2 that for underspecified problems, it is often useful to add a regularization term to encourage simple solutions, such as small norm solutions. It turns out that gradient descent will, in some cases, do this for you *implicitly*, even if you don't have any regularization terms in the optimization objective!

Let $A \in \mathbb{R}^{n \times d}$, where $n < d$, and assume there exists $x \in \mathbb{R}^d$ such that $Ax = b$.

Consider gradient descent on $f(x) = \frac{1}{2}\|Ax - b\|_2^2$, initialized at $x_0 = 0$, with some stepsize schedule η_t. Let x_t, for $t = 0, \ldots, T$ be the iterates of this gradient descent procedure, so that $x_{t+1} = x_t - \eta_t \nabla f(x_t)$, for $t = 0, \ldots, T-1$.

(a) [5 points] Show that x_t is orthogonal to $\ker(A)$ for every t.

(b) [5 points] Show that among all solutions to $Ax = b$, there is exactly one solution orthogonal to $\ker(A)$, and that this solution has minimum Euclidean norm.

(c) [5 points] Suppose that the smallest non-zero singular value of A is $\sqrt{\ell}$, and that the largest singular value of A is $\sqrt{L}$. Using what we showed in class (or indeed, in Problem 1) conclude that gradient descent converges to the minimum-norm solution of $Ax = b$ in a small number of steps. What is the rate of convergence you can show? Warning: f is not strongly convex!

(d) [5 points] Indeed, show the fact that gradient descent favors low-norm solutions is true in the following, strong sense. Suppose that $AA^\top = I$, and run gradient descent starting at $x_0 = 0$ with a fixed stepsize $\eta \in (0, 1)$. Show that for all t, x_t achieves the smallest possible value of $f(x)$ among *all* vectors of norm at most $\|x_t\|_2$.

Problem 3: The randomized Kaczmarz method

There are many subtle variations of SGD, each one having its own advantages and properties. Here, we will investigate the utility of one such variation, known as the *Kaczmarz* method for solving linear systems. Let $A \in \mathbb{R}^{n \times d}$, for $n \geq d$, and suppose it has full column rank. Suppose further that $b \in \mathbb{R}^n$ satisfies that $Ax = b$ has a unique solution x^*. Let $a_1^\top, \ldots, a_n^\top$ denote the rows of A, and assume that all rows are nonzero. The randomized Kaczmarz method maintains an estimate x_t, and in each iteration, performs the following update:

- Sample a row $i \sim [n]$ independently according to some probability distribution p over $[n]$.
- Project x_t onto the hyperplane $\langle a_i, x \rangle = b_i$ to obtain x_{t+1}.

(a) [5 points] Write down the form of the update on x_t. It is, in some sense, a stochastic gradient descent step. Explain why this is the case.

(b) [5 points] Let $e_t = x_t - x^*$ denote the quantity we wish to minimize. Write down an explicit expression for

$$\mathbb{E}[\|e_{t+1}\|_2^2 \,|e_t] \,,$$

for any choice of p.

(c) [5 points] Using this expression, show that there is some choice of p so that

$$\mathbb{E}[\|e_{t+1}\|_2^2 \,|e_t] \leq \left(1 - \frac{\sigma_{\min}(A)^2}{\|A\|_F^2}\right) \|e_t\|_2^2 \,,$$

where $\sigma_{\min}(A)$ is the smallest singular value of A.

(d) [5 points] Finally, use this expression to show that if the condition number $\kappa(A)$ of A is constant, then Kaczmarz initialized at a point x_0 satisfying $\|e_0\|_2 \leq R$ converges to a solution x_T with $\mathbb{E}[\|e_T\|_2^2] \leq \varepsilon^2$ in $O(d \log(R/\epsilon))$ iterations. Here, the condition number of a matrix $\kappa(A)$ is the ratio of its largest to its smallest singular value.