CME 307 Optimization
Assignment 2

May 1, 2023

Due: May 16, 2023 at 5:00PM

Problem 1. Stochastic gradient descent with a biased gradient oracle
In class we discussed stochastic gradient descent in the case when we have an unbiased stochastic gradient oracle for $f(x)$, that is $\mathbb{E}[g(x)] = \nabla f(x)$. In this problem you will explore what happens when $\mathbb{E}[g(x)] \neq \nabla f(x)$. In particular, we will work in the following setup.

Definition 1 (Biased Stochastic Gradient Oracle)**.** We say a map $g(x, \omega) : \mathbb{R}^n \times \Omega$ is a *biased stochastic gradient oracle* for f if

$$g(x, \omega) = \nabla f(x) + b(x) + N(x, \omega),$$

for a bias $b : \mathbb{R}^n \mapsto \mathbb{R}^n$ and a zero mean noise $N : \mathbb{R}^n \times \Omega \to \mathbb{R}^n$, that is $\mathbb{E}_\omega[N(x, \omega)] = 0$, $\forall x \in \mathbb{R}^n$.

The noise and bias are assumed to satisfy the following additional regularity conditions:

Assumption 1. *There exists constants $M, \sigma^2 \geq 0$ such that $\forall x \in \mathbb{R}^d$,*

$$\mathbb{E}_\omega[\|N(x, \omega)\|^2] \leq M\|\nabla f(x) + b(x)\|^2 + \sigma^2.$$

Assumption 2. *There exists constants $m < 1$ and $\zeta \geq 0$ such that $\forall \boldsymbol{x} \in \mathbb{R}^d$,*

$$\|b(x)\|^2 \leq m\|\nabla f(x)\|^2 + \zeta^2.$$

Algorithm 1 Biased SGD

Require: initialization x_0
repeat
  Query oracle at x_k to obtain $g(x_k, \omega_k)$
  $x_{k+1} = x_k - \eta g(x_k, \omega_k)$
until convergence

The focus of this question is to analyze the convergence properties of algorithm 1, when applied to an L-smooth and μ-strongly convex function f.

1. Show that if $\zeta = 0$, then

$$\mathbb{E}_\omega[\langle \nabla f(x), g(x, \omega)\rangle] \geq \frac{(1-m)\|\nabla f(x)\|^2}{2}.$$

The preceding display shows that in expectation, $-g(x, \omega)$ still yields a descent direction. Hence if $\zeta = 0$, we should still expect "convergence", albeit at a slower rate.

2. Show that if we run Algorithm 1 with stepsize $\eta \leq \frac{1}{L(M+1)}$, then for any $k \geq 1$

$$\mathbb{E}_\omega[f(x_k)|x_{k-1}] - f(x_{k-1}) \leq \frac{\eta(m-1)}{2}\|\nabla f(x_{k-1})\|^2 + \frac{\eta\zeta^2}{2} + \frac{\eta^2 L\sigma^2}{2}.$$

3. Under the hypotheses of part 2., establish that

$$\mathbb{E}[f(x_k)] - f(x_\star) \le (1-\eta(1-m)\mu)^k\,[f(x_0)-f(x_\star)] + \frac{\zeta^2}{2(1-m)\mu} + \frac{\eta L\sigma^2}{2(1-m)\mu}.$$

Using the preceding display, show that if we set $\eta = \min\left\{\frac{\epsilon(1-m)\mu}{L\sigma^2}, \frac{1}{L(M+1)}\right\}$, then

$$\mathbb{E}[f(x_k)] - f(x_\star) \le \epsilon + \frac{\zeta^2}{2(1-m)\mu}$$

after $k \ge L\max\left\{\frac{M+1}{(1-m)\mu}, \frac{\sigma^2}{(1-m)^2\mu^2\epsilon}\right\}\log\left(\frac{2(f(x_0)-f(x_\star))}{\epsilon}\right)$ iterations.

How does this result compare to the convergence result for SGD discussed in class? What happens when $\zeta = 0$?

Problem 2. Relative smoothness and convexity
Let $\mathcal{D} \subset \mathbb{R}^n$ be closed and convex and let $f : \mathcal{D} \mapsto \mathbb{R}^n$ be a twice differentiable function. Then the *relative smoothness* constant is given by:

$$\hat{L} = \sup_{x,y\in\mathcal{D}} \int_0^1 2(1-t)\frac{\|y-x\|^2_{H(x+t(y-x))}}{\|y-x\|^2_{H(x)}}dt.$$

Similarly, the *relative convexity* constant is defined as follows:

$$\hat{\mu} = \inf_{x,y\in\mathcal{D}} \int_0^1 2(1-t)\frac{\|y-x\|^2_{H(x+t(y-x))}}{\|y-x\|^2_{H(x)}}dt.$$

Relative smoothness and relative convexity measure the regularity of f with respect to the Hessian norm, rather than the usual 2-norm. Observe $\hat{L}$ and $\hat{\mu}$ satisfy,

$$0 \leq \hat{\mu} \leq \hat{L}.$$

A function f is said to be relatively smooth and relatively convex if $\hat{L} < \infty$ and $\hat{\mu} > 0$. As you will see in the next problem, $\hat{L}$ and $\hat{\mu}$ are natural quantities to consider when analyzing Newton-type methods.

1. Show that if f is L-smooth and μ-strongly convex over $\mathcal{D}$, then it is $\hat{L}$-*relatively* smooth and $\hat{\mu}$ *relatively*-strongly convex, where
$$\frac{\mu}{L} \leq \hat{\mu} \leq \hat{L} \leq \frac{L}{\mu}.$$

2. Show that for any $x, y \in \mathcal{D}$, that
$$f(y) \leq f(x) + \langle \nabla f(x), y-x\rangle + \frac{\hat{L}}{2}\|y-x\|^2_{H(x)},$$
$$f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle + \frac{\hat{\mu}}{2}\|y-x\|^2_{H(x)}.$$
Hint: The fundamental theorem of calculus will be helpful here.

3. A function f is said to be a generalized linear model if it has the form
$$f(x) = \frac{1}{m}\sum_{i=1}^m \phi_i(a_i^T x) + \frac{\nu}{2}\|x\|^2, \tag{1}$$
where $\phi_i : \mathbb{R} \to \mathbb{R}$ is smooth and convex for all $i \in \{1, \ldots, m\}$, $\nu \geq 0$ is the ℓ_2-regularization parameter, and a_i^T is the ith row of the matrix $A \in \mathbb{R}^{m\times n}$. Least-squares and logistic regression are both special cases of GLMs, in which case A corresponds to the data matrix of observations.
Define
$$u := \sup_{1\leq i\leq m}\left(\sup_{x\in\mathcal{D}} \phi_i''(a_i^T x)\right), \quad l := \inf_{1\leq i\leq m}\left(\inf_{x\in\mathcal{D}} \phi_i''(a_i^T x)\right).$$
Show that when f is a GLM (i.e. f is as in (1)), the following inequality holds:
$$\frac{l\sigma_1^2(A) + m\nu}{u\sigma_1^2(A) + m\nu} \leq \hat{\mu} \leq \hat{L} \leq \frac{u\sigma_1^2(A) + m\nu}{l\sigma_1^2(A) + m\nu},$$
were $\sigma_1(A)$ denotes the largest singular value of A.
Hint: The following fact may be useful to you.
The function
$$h(x) = \frac{ax+c}{bx+c}, \quad \text{where } a \geq b \geq 0, c \geq 0$$
is increasing for $x \geq 0$.

Problem 3. Approximate Newton methods

Algorithm 2 Approximate Newton algorithm

Require: initialization x_0, ζ-approximate Hessian oracle $\mathcal{O}_\zeta(x)$
repeat
Query oracle at x_k to obtain $\hat{H}_k$
$x_{k+1} = x_k - \frac{1}{(1+\zeta)\hat{L}} \hat{H}_k^{-1} \nabla f(x_k)$
until convergence

Let f be a smooth strongly convex function, which we wish to minimize. We shall suppose access to an oracle, such that when queried at a point x, produces an approximation $\hat{H}$ satisfying

$$(1-\zeta)\hat{H} \preceq H(x) \preceq (1+\zeta)\hat{H},$$

where $\zeta \in (0,1)$ and $H(x)$ denotes the Hessian evaluated at x of f. We refer to this oracle as a ζ-approximate Hessian oracle, and denote it by $\mathcal{O}_\zeta(x)$. The goal of this problem is to analyze the convergence of the approximate Newton method presented in algorithm 2.

1. Show that if $\eta = \frac{1}{(1+\zeta)\hat{L}}$, then

$$f(x_k) \leq f(x_{k-1}) - \frac{\|\hat{H}_k^{-1/2} \nabla f(x_{k-1})\|^2}{2(1+\zeta)\hat{L}}.$$

2. Show the following identity,

$$(1-\zeta)\hat{\mu}\left(f(x_k) - f(x_\star)\right) \leq \frac{\|\hat{H}_k^{-1/2} \nabla f(x_{k-1})\|^2}{2}.$$

   Hint: The lower-bound you derived in part 2 of problem 2 will be useful here.

3. Using item 2, conclude that

$$f(x_k) - f(x_\star) \leq \left(1 - \left(\frac{1+\zeta}{1-\zeta}\right)^{-1} \frac{\hat{\mu}}{\hat{L}}\right)\left(f(x_{k-1}) - f(x_\star)\right).$$

   Hence deduce that

$$f(x_k) - f(x_\star) \leq \epsilon,$$

   after $k \geq \frac{1+\zeta}{1-\zeta} \frac{\hat{L}}{\hat{\mu}} \log\left(\frac{(f(x_0) - f(x_\star))}{\epsilon}\right)$ iterations.