COMS 4995-004: Optimization for Machine Learning
Homework 3

HW3 is due Tuesday, Nov 14 by 1:00 pm. No late assignments will be accepted[1]. Please refer to `https://www.satyenkale.com/optml-f19/` for instructions on how to submit homework assignments.

In class we studied several algorithms to minimize convex functions. Minimizing *nonconvex* functions $f : \mathbb{R}^d \to \mathbb{R}$ is significantly harder (it is NP-hard in the worst case), so we can only give weak guarantees for first order methods like gradient descent. Typically, the objective here is to show *first order* convergence: i.e. given any $\epsilon > 0$, show that the method yields a point x such that $\|\nabla f(x)\|^2 \leq \epsilon$ after some number of iterations which depends on ϵ (in the case of stochastic optimization, we require x such that $\mathbb{E}[\|\nabla f(x)\|^2] \leq \epsilon$, where the expectation is over the randomness in the stochastic gradients and the algorithm.)

In this homework we will derive such guarantees. Assume that f is a β-smooth *nonconvex* function, and that $f(x) \geq 0$ for all $x \in \mathbb{R}^d$.

Question 1. (9 points) Consider running gradient descent on f with a step-size η: start with an arbitrary point $x_0 \in \mathbb{R}^d$, and iterate $x_{t+1} = x_t - \eta \nabla f(x_t)$ for T steps. Then show there is a choice of the step-size η such that

$$\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \leq 2\beta f(x_0).$$

From this bound, determine how large T needs to be (in terms of $\epsilon, \beta, f(x_0)$) to guarantee that there is an iterate x_t such that $\|\nabla f(x_t)\|^2 \leq \epsilon$.

Question 2. (16 points) Now suppose $f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[g(x, \xi)]$ where $g(\cdot, \xi)$ is differentiable for all ξ and the distribution $\mathcal{D}$ is unknown. Thus it is not possible to evaluate $f(x)$ or $\nabla f(x)$ at any given point x. Assume that that variance of the stochastic gradients is bounded by σ^2, i.e. for any $x \in \mathbb{R}^d$, we have $\mathbb{E}_{\xi \sim \mathcal{D}}[\|\nabla g(x, \xi) - \nabla f(x)\|^2] \leq \sigma^2$. Suppose now that we run *stochastic* gradient descent as follows: start with an arbitrary point $x_0 \in \mathbb{R}^d$, and iterate $x_{t+1} = x_t - \eta \nabla g(x_t, \xi_t)$ where ξ_t is sampled from $\mathcal{D}$, for T steps. Then show that if $\eta \leq \frac{1}{\beta}$, we have

$$\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(x_t)\|^2] \leq \frac{2}{\eta} f(x_0) + \beta \eta \sigma^2 T.$$

[1] Unless you have an emergency; in that case please write to Satyen as soon as possible.

Using this bound, compute a value of η which ensures that

$$\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(x_t)\|^2] \leq O(\beta f(x_0) + \sqrt{\beta f(x_0)\sigma^2 T}).$$

Suppose we output a *random* iterate, i.e. choose $R \in \{0, 1, 2, \ldots, T-1\}$ uniformly at random, and then output x_R. Then conclude that

$$\mathbb{E}[\|\nabla f(x_R)\|^2] \leq O\left(\frac{\beta f(x_0) + \sqrt{\beta f(x_0)\sigma^2 T}}{T}\right),$$

where the expectation is over the choice of R as well as $\xi_0, \xi_1, \ldots, x_{T-1}$. Using this bound, determine how large T needs to be (in terms of $\epsilon, \beta, f(x_0), \sigma$) to guarantee that $\mathbb{E}[\|\nabla f(x_R)\|^2] \leq \epsilon$ (it is fine to use the $\Omega(\cdot)$ notation in your lower bound on T to suppress numerical constants).

Solution: question 1.
By the β-smoothness of f, we have

$$f(x_{t+1}) \leq f(x_t) + \nabla f(x_t) \cdot (x_{t+1} - x_t) + \frac{\beta}{2}\|x_{t+1} - x_t\|^2 = f(x_t) - \eta\|\nabla f(x_t)\|^2 + \frac{\beta\eta^2}{2}\|\nabla f(x_t)\|^2.$$

Setting $\eta = \frac{1}{\beta}$, we get $f(x_{t+1}) \leq f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2$, and so $\|\nabla f(x_t)\|^2 \leq 2\beta(f(x_t) - f(x_{t+1}))$. Summing up this bound from $t = 0$ to $T - 1$, and noticing that the RHS telescopes, we get

$$\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \leq 2\beta(f(x_0) - f(x_T)) \leq 2\beta f(x_0),$$

since $f(x_T) \geq 0$. Thus $\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \leq \frac{2\beta f(x_0)}{T}$, which implies that there exists some iterate x_t for $t \in \{0, 1, \ldots, T-1\}$ such that $\|\nabla f(x_t)\|^2 \leq \frac{2\beta f(x_0)}{T}$. The RHS becomes smaller than ϵ when $T \geq \frac{2\beta f(x_0)}{\epsilon}$.

Solution: question 2.
By the β-smoothness of f, we have

$$f(x_{t+1}) \leq f(x_t) + \nabla f(x_t) \cdot (x_{t+1} - x_t) + \frac{\beta}{2}\|x_{t+1} - x_t\|^2 = f(x_t) - \eta\nabla g(x_t, \xi_t) + \frac{\beta\eta^2}{2}\|\nabla g(x_t, \xi_t)\|^2.$$

Taking expectation on both sides of the inequality above conditioned on x_t, and using the facts that $\mathbb{E}[\nabla g(x_t, \xi_t)|x_t] = \nabla f(x_t)$ and $\mathbb{E}[\|\nabla g(x_t, \xi_t)\|^2|x_t] \leq \|\nabla f(x_t)\|^2 + \sigma^2$, we get

$$\mathbb{E}[f(x_{t+1})|x_t] = f(x_t) - \eta\|\nabla f(x_t)\|^2 + \frac{\beta\eta^2}{2}(\|\nabla f(x_t)\|^2 + \sigma^2) \leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|^2 + \frac{\beta\eta^2}{2}\sigma^2,$$

if we choose $\eta \leq \frac{1}{\beta}$. Taking expectation on both sides of the inequality to remove the conditioning on x_t, we get

$$\mathbb{E}[f(x_{t+1})] = \mathbb{E}[f(x_t)] - \frac{\eta}{2}\mathbb{E}[\|\nabla f(x_t)\|^2] + \frac{\beta\eta^2}{2}\sigma^2 \quad \Rightarrow \quad \mathbb{E}[\|\nabla f(x_t)\|^2] \leq \frac{2}{\eta}(\mathbb{E}[f(x_t)] - \mathbb{E}[f(x_{t+1})]) + \beta\eta\sigma^2.$$

Summing up the inequality from $t = 0$ to $T - 1$, and noticing that the RHS telescopes, we get

$$\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(x_t)\|^2] \leq \frac{2}{\eta}(f(x_0) - \mathbb{E}[f(x_{t+1})]) + \beta\eta\sigma^2 T \leq \frac{2}{\eta} f(x_0) + \beta\eta\sigma^2 T.$$

The above bound uses the fact that $\mathbb{E}[f(x_0)] = f(x_0)$ since x_0 is not random, and that $\mathbb{E}[f(x_{t+1})] \geq 0$. Now suppose we set $\eta = \min\{\frac{1}{\beta}, \sqrt{\frac{2f(x_0)}{\beta\sigma^2 T}}\}$ so that the condition that $\eta \leq \frac{1}{\beta}$ is satisfied, we have

$$\frac{2}{\eta} f(x_0) + \beta\eta\sigma^2 T \leq \max\left\{\beta, \sqrt{\tfrac{\beta\sigma^2 T}{2f(x_0)}}\right\} \cdot 2f(x_0) + \min\left\{\frac{1}{\beta}, \sqrt{\tfrac{2f(x_0)}{\beta\sigma^2 T}}\right\} \cdot \beta\sigma^2 T = O(\beta f(x_0) + \sqrt{\beta f(x_0)\sigma^2 T}).$$

Now if we choose an index $R \in \{0, 1, 2, \ldots, T-1\}$, then we have

$$\mathbb{E}[\|\nabla f(x_R)\|^2] = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(x_t)\|^2] \leq O\left(\frac{\beta f(x_0) + \sqrt{\beta f(x_0)\sigma^2 T}}{T}\right),$$

where the expectation on the LHS is over the choice of R as well as $\xi_0, \xi_1, \ldots, \xi_{T-1}$. In order to make the RHS above smaller ϵ, we need to choose

$$T \geq \Omega\left(\tfrac{\beta f(x_0)}{\epsilon} + \tfrac{\beta f(x_0)\sigma^2}{\epsilon^2}\right).$$