CMPT 409/981: Optimization for Machine Learning

Lecture 4

Sharan Vaswani

September 17, 2024

Recap

- For L-smooth functions lower-bounded by f^*, GD with backtracking Armijo line-search returns an ϵ stationary-point in $O\left(\frac{1}{\epsilon}\right)$ iterations without requiring the knowledge of L.
- **Convex sets**: Set $\mathcal{C}$ is convex iff $\forall x, y \in \mathcal{C}$, the convex combination $z_\theta := \theta x + (1-\theta) y$ for $\theta \in [0,1]$ is also in $\mathcal{C}$.
    - *Examples*: Half-space: $\{x | Ax \leq b\}$, Norm-ball: $\{x | \, \|x\|_p \leq r\}$.
- **Convex functions**: A function f is convex iff its domain $\mathcal{D}$ is a convex set, and for all $x, y \in \mathcal{D}$ and $\theta \in [0,1]$,$f(\theta x + (1-\theta) y) \leq \theta f(x) + (1-\theta) f(y)$.
    - *First-order definition*: If f is differentiable, it is convex iff its domain $\mathcal{D}$ is a convex set and for all $x, y \in \mathcal{D}$, $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.
    - *Second-order definition*: If f is twice differentiable, it is convex iff its domain $\mathcal{D}$ is a convex set and for all $x \in \mathcal{D}$, $\nabla^2 f(x) \succeq 0$.
    - *Examples*: All norms $\|x\|_p$, Negative entropy: $f(x) = x \log(x)$, Logistic regression: $\sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \langle X_i, w \rangle\right)\right)$, Ridge regression: $\frac{1}{2} \|Xw - y\|^2 + \frac{\lambda}{2} \|w\|^2$.

Jensen's Inequality

- Recall the zero-order definition of convexity: $\forall x, y \in \mathcal{D}$ and $\theta \in [0,1]$, $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$.
  This can be generalized to n points $\{x_1, x_2, \ldots, x_n\}$, i.e. for $p_i \geq 0$ and $\sum_i p_i = 1$,
  $$f(p_1\, x_1 + p_2\, x_2 + \ldots + p_n\, x_n) \leq p_1\, f(x_1) + p_2\, f(x_2) + \ldots + p_n\, f(x_n) \implies f\left(\sum_{i=1}^{n} p_i x_i\right) \leq \sum_{i=1}^{n} p_i f(x_i)$$
- If X is a discrete r.v. that can take value x_i with probability p_i, and f is convex, then,
  $$f\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left[f(X)\right]. \qquad \text{(Jensen's inequality)}$$
- Jensen's inequality can be used to prove inequalities like the AM-GM inequality: $\sqrt{ab} \leq \frac{a+b}{2}$.
- *Proof*: Choose $f(x) = -\log(x)$ as the convex function, and consider two points a and b with $\theta = 1/2$. By Jensen's inequality,
  $$-\log\left(\frac{a+b}{2}\right) \leq \frac{-\log(a) - \log(b)}{2} \implies \log\left(\frac{a+b}{2}\right) \geq \log(\sqrt{ab}) \implies \frac{a+b}{2} \geq \sqrt{ab}.$$

Holder's Inequality

Q: Prove Holder's inequality, for $p, q \geq 1$ s.t. $\frac{1}{p} + \frac{1}{q} = 1$ and $x, y \in R^n$, $|\langle x, y\rangle| \leq \|x\|_p \, \|y\|_q$.

Proof: By repeating the AM-GM proof, but for a general $\theta \in [0, 1]$, for $a, b \geq 0$, we can prove

$$a^\theta b^{1-\theta} \leq \theta a + (1-\theta) b$$

Use $a = \frac{|x_i|^p}{\sum_{j=1}^n |x_j|^p}$, $b = \frac{|y_i|^q}{\sum_{j=1}^n |y_j|^q}$, $\theta = 1/p$, and using the fact that $1 - \theta = 1 - 1/p = 1/q$

$$\left(\frac{|x_i|^p}{\sum_{j=1}^n |x_j|^p}\right)^{1/p} \left(\frac{|y_i|^q}{\sum_{j=1}^n |y_j|^q}\right)^{1/q} \leq \frac{1}{p} \frac{|x_i|^p}{\sum_{j=1}^n |x_j|^p} + \frac{1}{q} \frac{|y_i|^p}{\sum_{j=1}^n |y_j|^p}$$

Summing both sides from $i = 1$ to n and using the fact that $\frac{1}{p} + \frac{1}{q} = 1$

$$\sum_{i=1}^n \frac{|x_i|}{\left(\sum_{j=1}^n |x_j|^p\right)^{1/p}} \frac{|y_i|}{\left(\sum_{j=1}^n |y_j|^q\right)^{1/q}} \leq 1 \implies \sum_i |x_i y_i| \leq \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \left(\sum_{i=1}^n |y_i|^q\right)^{1/q}$$

$$\implies |\langle x, y\rangle| \leq \|x\|_p \, \|y\|_q \qquad \text{(Triangle inequality)}$$

GD for Smooth, Convex Functions

Recall that for convex functions, minimizing the gradient norm results in finding the minimizer. Let us analyze the convergence of GD for smooth, convex problems: $\min_{w \in \mathbb{R}^d} f(w)$.

Claim: For L-smooth, convex functions s.t. for any $w^* \in \arg\min f(w)$, GD with $\eta = \frac{1}{L}$ requires $T \geq \frac{2L\,\|w_0 - w^*\|^2}{\epsilon}$ iterations to obtain point w_T that is ϵ-suboptimal meaning that $f(w_T) \leq f(w^*) + \epsilon$.

Proof: For L-smooth functions, $\forall x, y \in \mathcal{D}$, $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$. Similar to Lecture 2, using GD: $w_{k+1} = w_k - \frac{1}{L}\nabla f(w_k)$ yields

$$f(w_{k+1}) - f(w^*) \leq f(w_k) - f(w^*) - \frac{1}{2L}\|\nabla f(w_k)\|^2 \tag{1}$$

Using $y = w^*$, $x = w_k$ in the first-order condition for convexity: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$,

$$f(w_k) - f(w^*) \leq \langle \nabla f(w_k), w_k - w^* \rangle \leq \|\nabla f(w_k)\|\,\|w_k - w^*\| \qquad \text{(Cauchy Schwarz)}$$

$$\implies \|\nabla f(w_k)\| \geq \frac{f(w_k) - f(w^*)}{\|w_k - w^*\|} \tag{2}$$

GD for Smooth, Convex Functions

In addition to descent on the function, when minimizing smooth, convex functions, GD decreases the distance to a minimizer w^*.

Claim: For GD with $\eta = \frac{1}{L}$, $\|w_{k+1} - w^*\|^2 \leq \|w_k - w^*\|^2 \leq \|w_0 - w^*\|^2$.
Proof:

$$\|w_{k+1} - w^*\|^2 = \|w_k - \eta \nabla f(w_k) - w^*\|^2 = \|w_k - w^*\|^2 - 2\eta\langle \nabla f(w_k), w_k - w^*\rangle + \eta^2 \|\nabla f(w_k)\|^2$$

Using $y = w^*$, $x = w_k$ in the first-order condition for convexity: $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$,

$$\|w_{k+1} - w^*\|^2 \leq \|w_k - w^*\|^2 - 2\eta[f(w_k) - f(w^*)] + \eta^2 \|\nabla f(w_k)\|^2$$

For convex functions, L-smoothness is equivalent to $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2$. Using $x = w^*$, $y = w_k$ in this equation,

$$\leq \|w_k - w^*\|^2 - 2\eta[f(w_k) - f(w^*)] + 2L\eta^2[f(w_k) - f(w^*)]$$

$$\implies \|w_{k+1} - w^*\|^2 \leq \|w_k - w^*\|^2 \qquad \text{(By setting } \eta = \tfrac{1}{L}\text{)}$$

GD for Smooth, Convex Functions

Combining Eq. 2 with the result of the previous claim,

$$\|\nabla f(w_k)\| \geq \frac{f(w_k) - f(w^*)}{\|w_k - w^*\|} \geq \frac{f(w_k) - f(w^*)}{\|w_0 - w^*\|}$$

Combining the above inequality with Eq. 1,

$$f(w_{k+1}) - f(w^*) \leq f(w_k) - f(w^*) - \frac{1}{2L}\|\nabla f(w_k)\|^2 \leq f(w_k) - f(w^*) - \frac{1}{2L}\frac{[f(w_k) - f(w^*)]^2}{\|w_0 - w^*\|^2}$$

Dividing by $[f(w_k) - f(w^*)]\,[f(w_{k+1}) - f(w^*)]$

$$\frac{1}{f(w_k) - f(w^*)} \leq \frac{1}{f(w_{k+1}) - f(w^*)} - \frac{1}{2L}\frac{f(w_k) - f(w^*)}{\|w_0 - w^*\|^2}\frac{1}{f(w_{k+1}) - f(w^*)}$$

$$\implies \frac{1}{2L\,\|w_0 - w^*\|^2}\underbrace{\frac{f(w_k) - f(w^*)}{f(w_{k+1}) - f(w^*)}}_{\geq 1} \leq \left[\frac{1}{f(w_{k+1}) - f(w^*)} - \frac{1}{f(w_k) - f(w^*)}\right] \qquad (3)$$

GD for Smooth, Convex Functions

Summing Eq. 3 from $k = 0$ to $T - 1$,

$$\sum_{k=0}^{T-1} \left[\frac{1}{2L \left\| w_0 - w^* \right\|^2} \right] \leq \sum_{k=0}^{T-1} \left[\frac{1}{f(w_{k+1}) - f(w^*)} - \frac{1}{f(w_k) - f(w^*)} \right]$$

$$\frac{T}{2L \left\| w_0 - w^* \right\|^2} \leq \frac{1}{f(w_T) - f(w^*)} - \frac{1}{f(w_0) - f(w^*)} \leq \frac{1}{f(w_T) - f(w^*)}$$

$$\implies f(w_T) - f(w^*) \leq \frac{2L \left\| w_0 - w^* \right\|^2}{T}$$

The suboptimality $f(w_T) - f(w^*)$ decreases at an $O\left(\frac{1}{T}\right)$ rate, i.e. the function value at iterate w_T approaches the minimum function value $f(w^*)$.

In order to obtain a function value at least ϵ-close to the optimal function value, GD requires $T \geq \frac{2L \left\| w_0 - w^* \right\|^2}{\epsilon}$ iterations.

Minimizing Smooth, Convex Functions

Recall that GD was optimal (amongst first-order methods with no dependence on the dimension) when minimizing smooth (possibly non-convex) functions.

Is GD also optimal when minimizing smooth, convex functions, or can we do better?

Lower Bound: For any initialization, there exists a smooth, convex function such that any first-order method requires $\Omega\left(\frac{1}{\sqrt{\epsilon}}\right)$ iterations.

Possible reasons for the discrepancy between the $O(1/\epsilon)$ upper-bound for GD, and the $\Omega(1/\sqrt{\epsilon})$ lower-bound:

(1) Our upper-bound analysis of GD is loose, and GD actually matches the lower-bound.
(2) The lower-bound is loose, and there is a function that requires $\Omega(1/\epsilon)$ iterations to optimize.
(3) Both the upper and lower-bounds are tight, and GD is sub-optimal. There exists another algorithm that has an $O(1/\sqrt{\epsilon})$ upper-bound and is hence optimal.

Option (3) is correct – GD is sub-optimal for minimizing smooth, convex functions. Using Nesterov acceleration is optimal and requires $\Theta(1/\sqrt{\epsilon})$ iterations.

Questions?