CMPT 409/981: Optimization for Machine Learning

Lecture 4

Sharan Vaswani

September 22, 2022

Recap

Convex optimization: Minimizing a convex function over a convex set.

Convex sets: Set $\mathcal{C}$ is convex iff $\forall x, y \in \mathcal{C}$, the convex combination $z := \theta x + (1-\theta)y$ 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 condition for convexity: 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 condition for convexity: 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: $\frac{1}{2} \|Xw - y\|^2 + \frac{\lambda}{2} \|w\|^2$.

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, 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 in the sense 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 $[f(w_k) - f(w^*)] [f(w_{k+1}) - f(w^*)]$

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

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

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/oracle calls.

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 actual 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 (Will cover it next week!).

Questions?

Strongly convex functions

First-order condition: If f is differentiable, it is μ-strongly convex iff its domain $\mathcal{D}$ is a convex set and for all $x, y \in \mathcal{D}$ and $\mu > 0$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2$$

i.e. for all y, the function is lower-bounded by the quadratic defined in the RHS.

Second-order condition for convexity: If f is twice differentiable, it is strongly-convex iff its domain $\mathcal{D}$ is a convex set and for all $x \in \mathcal{D}$,

$$\nabla^2 f(x) \succeq \mu I_d$$

i.e. for all x, the eigenvalues of the Hessian are lower-bounded by μ.

Alternative condition: Function $g(x) = f(x) - \frac{\mu}{2} \|x\|^2$ is convex, i.e. if we "remove" a quadratic (curvature) from f, it still remains convex.

Examples: Quadratics $f(x) = x^\intercal Ax + bx + c$ are μ-strongly convex if $A \succeq \mu I_d$. If f is a convex loss function, then $g(x) := f(x) + \frac{\lambda}{2} \|x\|^2$ (the ℓ_2-regularized loss) is λ-strongly convex.

Strongly-convex functions

Strict-convexity: If f is differentiable, it is strictly-convex iff its domain $\mathcal{D}$ is a convex set and for all $x, y \in \mathcal{D}$,

$$f(y) > f(x) + \langle \nabla f(x), y - x \rangle$$

If f is μ strongly-convex, then it is also strictly convex.

Q: For a strictly-convex f, if $\nabla f(w^*) = 0$, then is w^* a unique minimizer of f?

Q: Prove that the ridge regression loss function: $f(w) = \frac{1}{2} \|Xw - y\|^2 + \frac{\lambda}{2} \|w\|^2$ is strongly-convex. Compute μ.

Q: Is $f(w) = \frac{1}{2} \|Xw - y\|^2$ strongly-convex?

Strongly-convex functions

Q: Is negative entropy function $f(x) = x \ln(x)$ strictly-convex on $(0, 1)$?

Q: Is logistic regression: $f(w) = \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \langle X_i, w \rangle\right)\right)$ strongly-convex?

Questions?

GD for Smooth, Strongly-Convex Functions

Recall that for convex functions, minimizing the gradient norm results in finding the minimizer, and for strongly-convex functions, the minimizer w^* is unique.

Let us analyze the convergence of GD for smooth, strongly-convex problems: $\min_{w \in \mathbb{R}^d} f(w)$.

Claim: For L-smooth, μ-strongly convex functions, GD with $\eta = \frac{1}{L}$ requires $T \geq \frac{L}{\mu} \log\left(\frac{\|w_0 - w^*\|^2}{\epsilon}\right)$ iterations to obtain a point w_T that is ϵ-suboptimal in the sense that $\|w_T - w^*\|^2 \leq \epsilon$.

Proof: Bounding the distance of the iterates to w^*,

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

L-smoothness: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2$. Using $x = w^*$, $y = w_k$,

$$\implies \|w_{k+1} - w^*\|^2 \leq \|w_k - w^*\|^2 - 2\eta\langle \nabla f(w_k), w_k - w^* \rangle + 2L\eta^2[f(w_k) - f(w^*)] \qquad (4)$$

GD for Smooth, Strongly-Convex Functions

μ-strongly convexity: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2$. Using $x = w_k$, $y = w^*$,

$$f(w^*) \geq f(w_k) + \langle \nabla f(w_k), w^* - w_k \rangle + \frac{\mu}{2} \|w_k - w^*\|^2$$

$$\implies \langle \nabla f(w_k), w_k - w^* \rangle \geq f(w_k) - f(w^*) + \frac{\mu}{2} \|w_k - w^*\|^2 \tag{5}$$

Combining Eq. 4 and 5,

$$\|w_{k+1} - w^*\|^2 \leq \|w_k - w^*\|^2 - 2\eta \left[f(w_k) - f(w^*) + \frac{\mu}{2} \|w_k - w^*\|^2 \right] + 2L\eta^2 [f(w_k) - f(w^*)]$$

$$= \|w_k - w^*\|^2 (1 - \mu\eta) + [f(w_k) - f(w^*)] \left(-2\eta + 2L\eta^2\right)$$

$$\implies \|w_{k+1} - w^*\|^2 \leq \left(1 - \frac{\mu}{L}\right) \|w_k - w^*\|^2 \qquad \text{(Since } \eta = \tfrac{1}{L},\ \left(-2\eta + 2L\eta^2\right) = 0\text{)}$$

Recursing from $k = 0$ to $T - 1$,

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

(Using $1 - x \leq \exp(-x)$ for all x)

GD for Smooth, Strongly-Convex Functions

The suboptimality $\|w_T - w^*\|^2$ decreases at an $O(\exp(-T))$ rate, i.e. the iterate w_T approaches the unique minimizer w^*. In order to obtain an iterate at least ϵ-close to w^*, we need to make the RHS less than ϵ and quantify the number of required iterations.

$$\exp\left(-\frac{\mu T}{L}\right)\|w_0 - w^*\|^2 \leq \epsilon \implies T \geq \frac{L}{\mu}\log\left(\frac{\|w_0 - w^*\|^2}{\epsilon}\right).$$

Hence, the convergence rate is $O(\log(1/\epsilon))$ which is exponentially faster compared to the convergence rate for smooth, convex functions. This rate of convergence rate is referred to as the **linear rate**.

Condition number: $\kappa := \frac{L}{\mu}$ is a problem-dependent constant that quantifies the hardness of the problem (smaller κ implies that we need fewer iterations of GD).

Q: What κ corresponds to the easiest problem?

Q: What is the condition number for ridge regression: $\frac{1}{2}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|^2$.

GD for Smooth, Strongly-Convex Functions

Q: For L-smooth, μ-strongly convex functions, how many iterations do we need to ensure that $f(w_T) - f(w^*) \leq \epsilon$?

Gradient Descent is "adaptive" to strong-convexity i.e. it does not need to know μ to converge.

The algorithm remains the same (use step-size $\eta = \frac{1}{L}$) regardless of whether we run it on a convex or strongly-convex function.

Since GD only requires knowledge of L, we can use the Back-tracking Armijo line-search to estimate the smoothness, and obtain faster convergence in practice (In Assignment 1!).

Minimizing Smooth, Strongly-Convex Functions

Recall that for smooth, convex functions, GD is sub-optimal (convergence rate of $O(1/\epsilon)$) and can be improved by using Nesterov acceleration (convergence rate of $O(1/\sqrt{\epsilon})$).

For smooth, strongly-convex functions, the convergence rate of GD is $O\left(\kappa \log\left(1/\epsilon\right)\right)$.

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

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

GD is sub-optimal for minimizing smooth, convex functions. Using Nesterov acceleration is optimal and requires $\Theta\left(\sqrt{\kappa} \log\left(1/\epsilon\right)\right)$ iterations

Questions?

Dealing with Constrained Domains

We have characterized the convergence of GD on smooth, (strongly)-convex functions when the domain was $\mathbb{R}^d$ i.e. the optimization was "unconstrained".

In general, convex optimization can be constrained to be over a convex set.

Examples: Linear programming, Optimizing over the probability simplex or a norm-ball.

We can modify GD to solve problems such as $\min_{w \in \mathcal{C}} f(w)$ where $\mathcal{C}$ is a convex set.

Projected GD

$$w_{k+1} = \Pi_{\mathcal{C}}\left[w_k - \eta \nabla f(w_k)\right]$$

where, $\Pi_{\mathcal{C}}[x] = \arg\min_{w \in \mathcal{C}} \frac{1}{2} \left\|w - x\right\|^2$ is the Euclidean projection onto the convex set $\mathcal{C}$.

Dealing with Constrained Domains

Q: (i) Is $\Pi_{\mathcal{C}}[x]$ unique for convex sets? (ii) For non-convex sets?

Q: For $x \in \mathbb{R}^d$, compute the Euclidean projection onto the ℓ_2-ball: $\mathcal{B}(0,1) = \{w | \, \|w\|_2^2 \leq 1\}$?

Dealing with Constrained Domains

For convex optimization over unconstrained domains, we know that the minimizer can be characterized by its gradient norm i.e. if w^* is a minimizer, then, $\nabla f(w^*) = 0$.

Optimality conditions: For constrained convex domains, if $w^* \in \arg\min_{x \in \mathcal{C}} f(x)$, then $\forall w \in \mathcal{C}$,

$$\langle \nabla f(w^*), w - w^* \rangle \geq 0$$

i.e. if we are at the optimal, either the gradient is zero (if w^* is inside $\mathcal{C}$) or moving in the negative direction of the gradient will push us out of $\mathcal{C}$ (if w^* is at the boundary of $\mathcal{C}$).

For the Euclidean projection, if $y := \Pi_{\mathcal{C}}[x] = \arg\min_{w \in \mathcal{C}} \frac{1}{2} \|w - x\|^2$, then, using the optimal conditions above, $\forall w \in \mathcal{C}$,

$$\langle x - y, w - y \rangle \leq 0$$

i.e. the angle between the rays $y \to x$ and $y \to w$ for all $w \in \mathcal{C}$ is greater than 90°.

Q: For convex set $\mathcal{C}$, if $w^* = \arg\min_{w \in \mathcal{C}} f(w)$, what is $\Pi_{\mathcal{C}}[w^*]$?

Dealing with Constrained Domains

Claim: Projections onto a convex set are non-expansive operations i.e. for all x_1, x_2, if $y_1 := \Pi_{\mathcal{C}}[x_1]$ and $y_2 := \Pi_{\mathcal{C}}[x_2]$, then, $\|y_1 - y_2\| \leq \|x_1 - x_2\|$.

Proof: Recall from the last slide, that for the Euclidean projection, $y = \Pi_{\mathcal{C}}[x]$, $\langle x - y, w - y \rangle \leq 0$ for all $w \in \mathcal{C}$. Hence,

$$\langle x_1 - y_1, w - y_1 \rangle \leq 0 \implies \langle x_1 - y_1, y_2 - y_1 \rangle \leq 0 \qquad \text{(Set } w = y_2\text{)}$$

$$\langle x_2 - y_2, w - y_2 \rangle \leq 0 \implies \langle x_2 - y_2, y_1 - y_2 \rangle \leq 0 \qquad \text{(Set } w = y_1\text{)}$$

Adding the two equations,

$$\langle x_2 - y_2, y_1 - y_2 \rangle + \langle x_1 - y_1, y_2 - y_1 \rangle \leq 0 \implies \langle x_2 - x_1 + y_1 - y_2, y_1 - y_2 \rangle \leq 0$$

$$\implies \langle y_1 - y_2, y_1 - y_2 \rangle \leq \langle x_1 - x_2, y_1 - y_2 \rangle \implies \|y_1 - y_2\|^2 \leq \|x_1 - x_2\| \, \|y_1 - y_2\|$$

(Cauchy Schwartz)

$$\implies \|y_1 - y_2\| \leq \|x_1 - x_2\|$$

Projected GD for Smooth, Strongly-Convex Functions

Recall the projected GD update: $w_{k+1} = \Pi_{\mathcal{C}}[w_k - \eta \nabla f(w_k)]$. Since $w^* = \Pi_{\mathcal{C}}[w^*]$, using the non-expansiveness of projections with $x_1 = w^*$, $x_2 = w_k - \eta \nabla f(w_k)$, $y_1 = w^*$, $y_2 = w_{k+1}$,

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

i.e. by projecting onto $\mathcal{C}$, the distance to the minimizer w^* (that lies in $\mathcal{C}$) has not increased.

With this change, the proof proceeds as before. In particular,

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

Using smoothness, strong-convexity similar to Slides 10-11, we can derive the same linear rate.

$$\|w_{k+1} - w^*\|^2 \leq \exp\left(-T/\kappa\right) \|w_0 - w^*\|^2$$

Using non-expansivenss of projections, we can redo the proof for smooth, convex functions and get the same $O\left(1/\epsilon\right)$ convergence rate.

Hence, projected GD is a good option for minimizing convex functions over convex sets when the projection operation is computationally cheap.

Questions?