CMPT 409/981: Optimization for Machine Learning

Lecture 5

Sharan Vaswani

September 26, 2022

Recap

For L-smooth, convex functions, GD with $\eta = {}^1\!/_L$ requires $T \geq \frac{2L\|w_0 - w^*\|^2}{\epsilon}$ iterations to obtain point w_T that is ϵ-suboptimal in the sense that $f(w_T) \leq f(w^*) + \epsilon$.

For L-smooth, convex functions, the rate can improved to $\Theta\left({}^1\!/_{\sqrt{\epsilon}}\right)$ using Nesterov acceleration.

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

For L-smooth, μ-strongly convex functions, the rate can improved to $\Theta\left(\sqrt{\kappa}\log\left(\frac{1}{\epsilon}\right)\right)$ using Nesterov acceleration.

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 f is a convex function and $\mathcal{C}$ is a convex set.

Projected GD

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

where, $\Pi_{\mathcal{C}}[x] = \arg\min_{w \in \mathcal{C}} \frac{1}{2} \|w - x\|^2$ 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?

Ans: (i) Yes, since we are minimizing a strongly-convex function over a convex set. (ii) Not necessarily, for example, when the set is the boundary of a circle and we are projecting the centre.

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

Ans: We need to solve $y = \min_{\|w\|_2^2 \leq 1} \frac{1}{2} \|w - x\|_2^2$. If $\|x\|_2^2 \leq 1$, $x \in \mathcal{B}(0,1)$, and $\Pi_{\mathcal{B}(0,1)}[x] = x$. If $\|x\|_2^2 > 1$, then the projection will result in a point on the boundary of $\mathcal{B}$ and have unit length. Consider the set of candidate points of unit length: $\hat{Y} = \{\hat{y} \mid \|\hat{y}\|_2^2 = 1\}$. For $y = \frac{x}{\|x\|_2^2} \in \hat{Y}$ and any other $\hat{y} \in \hat{Y}$,

$$y = \arg\min_{\hat{y} \in \hat{Y}} \frac{1}{2} \|\hat{y} - x\|_2^2 = \frac{1 + \|x\|^2}{2} - \langle \hat{y}, x \rangle$$

Hence, if $\|x\|_2^2 > 1$, then $\Pi_{\mathcal{B}}[x] = \frac{x}{\|x\|_2^2}$. Putting both cases together, $\Pi_{\mathcal{B}}[x] = \frac{x}{\max\{1, \|x\|_2^2\}}$. Can and should be formally done using Lagrange multipliers.

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 f is convex and $w^* \in \arg\min_{w \in \mathcal{C}} f(w)$, 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 $\Pi_{\mathcal{C}}[w^*]$?

Ans: w^* since $w^* \in \mathcal{C}$

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 Lecture 4, 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?

Nesterov Acceleration

Gradient Descent: $w_{k+1} = \mathrm{GD}(w_k)$ where GD is a function such that $\mathrm{GD}(w) := w - \eta \nabla f(w)$.

Nesterov Acceleration: $w_{k+1} = \mathrm{GD}(w_k + \beta_k(w_k - w_{k-1}))$ for $\beta_k \geq 0$ to be determined. Hence,

$$w_{k+1} = [w_k + \beta_k(w_k - w_{k-1})] - \eta \nabla f(w_k + \beta_k(w_k - w_{k-1}))$$

i.e. Nesterov acceleration can be interpreted as doing GD on "extrapolated" points where β_k can be interpreted as the "momentum" in the previous direction $(w_k - w_{k-1})$.

If we define sequence $v_k := w_k + \beta_k(w_k - w_{k-1})$, and initialize $w_0 = v_0$, then,

$$v_k = w_k + \beta_k(w_k - w_{k-1}) \quad ; \quad w_{k+1} = v_k - \eta \nabla f(v_k) \tag{1}$$

Rewriting the above expression only in terms of v_k,

$$v_{k+1} = v_k - \eta_k \nabla f(v_k) + \beta_{k+1}[v_k - v_{k-1}] - \eta\, \beta_{k+1}[\nabla f(v_k) - \nabla f(v_{k-1})]$$

i.e. Nesterov acceleration can be interpreted as moving along a combination of three directions – the gradient direction $\nabla f(v_k)$, the momentum direction for the iterates $[v_k - v_{k-1}]$ and the momentum direction for the gradients $[\nabla f(v_k) - \nabla f(v_{k-1})]$.

Nesterov Acceleration for Smooth, Convex Functions

In order to analyze the convergence of Nesterov acceleration for smooth, convex functions, define $d_k := \beta_k(w_k - w_{k-1})$, set $\eta = \frac{1}{L}$ and define $g_k := -\frac{1}{L}\nabla f(w_k + d_k)$. For $k \geq 1$ (for simplicity, set $w_1 = w_0$),

$$\begin{aligned} w_{k+1} &= [w_k + \beta_k(w_k - w_{k-1})] - \eta \nabla f(w_k + \beta_k(w_k - w_{k-1})) \\ \implies w_{k+1} &= w_k + d_k - \frac{1}{L}\nabla f(w_k + d_k) = w_k + d_k + g_k. \end{aligned}$$

In order to set the momentum parameter β_k, we define a sequence $\{\lambda_k\}_{k=1}^{T}$ such that,

$$\lambda_0 = 0 \quad ; \quad \lambda_k = \frac{1 + \sqrt{1 + 4\lambda_{k-1}^2}}{2} \quad ; \quad \beta_{k+1} = \frac{\lambda_k - 1}{\lambda_{k+1}} \tag{2}$$

Claim: For L-smooth, μ-strongly convex functions, Nesterov acceleration with $\eta = \frac{1}{L}$, β_k set according to Eq. (2) and $T \geq \frac{\sqrt{2L}\,\|w_1 - w^*\|}{\sqrt{\epsilon}}$ iterations to obtain point w_{T+1} that is ϵ-suboptimal in the sense that $f(w_{T+1}) \leq f(w^*) + \epsilon$.

Hence, Nesterov acceleration is optimal for minimizing the class of smooth, convex functions.

Nesterov Acceleration for Smooth, Convex Functions

In order to prove the claim, we will need the following lemma:

Lemma: When using Nesterov acceleration with $\eta = \frac{1}{L}$, for any vector y,
$f(w_{k+1}) - f(y) \leq \langle \nabla f(w_k + d_k), w_k + d_k - y \rangle - \frac{1}{2L} \|\nabla f(w_k + d_k)\|^2$.

Proof: Using L-smoothness, since Nesterov acceleration is equivalent to GD on $w_k + d_k$,

$$\begin{aligned} f(w_{k+1}) - f(w_k + d_k) &\leq \langle \nabla f(w_k + d_k), w_{k+1} - w_k - d_k \rangle + \frac{L}{2} \|w_{k+1} - w_k - d_k\|^2 \\ &= -\frac{1}{L} \langle \nabla f(w_k + d_k), \nabla f(w_k + d_k) \rangle + \frac{1}{2L} \|\nabla f(w_k + d_k)\|^2 \end{aligned}$$

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

$$\implies f(w_{k+1}) - f(y) \leq f(w_k + d_k) - f(y) - \frac{1}{2L} \|\nabla f(w_k + d_k)\|^2$$

Using convexity: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ with $x = w_k + d_k$ and $y = y$

$$\implies f(w_{k+1}) - f(y) \leq \langle \nabla f(w_k + d_k), w_k + d_k - y \rangle - \frac{1}{2L} \|\nabla f(w_k + d_k)\|^2 \qquad (3)$$

Nesterov Acceleration for Smooth, Convex Functions

Using the lemma with $y = w^*$, with $f^* := f(w^*)$ and define $\Delta_k := f(w_k) - f^*$,

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

$$\leq -\frac{L}{2} \left[2 \left\langle \frac{-\nabla f(w_k + d_k)}{L}, (w_k - w^*) + d_k \right\rangle + \frac{1}{L^2} \|\nabla f(w_k + d_k)\|^2 \right]$$

$$\implies \Delta_{k+1} \leq -\frac{L}{2} \left[2\langle g_k, w_k - w^* + d_k \rangle + \|g_k\|^2 \right] \tag{4}$$

Using the lemma with $y = w_k$,

$$[f(w_{k+1}) - f^*] - [f(w_k) - f^*] \leq \langle \nabla f(w_k + d_k), d_k \rangle - \frac{1}{2L} \|\nabla f(w_k + d_k)\|^2$$

$$\implies \Delta_{k+1} - \Delta_k \leq -\frac{L}{2} \left[2 \left\langle \frac{-\nabla f(w_k + d_k)}{L}, d_k \right\rangle + \frac{1}{L^2} \|\nabla f(w_k + d_k)\|^2 \right]$$

$$\implies \Delta_{k+1} - \Delta_k \leq -\frac{L}{2} \left[2\langle g_k, d_k \rangle + \|g_k\|^2 \right] \tag{5}$$

Nesterov Acceleration for Smooth, Convex Functions

For $\lambda_k > 1$,

$$(\lambda_k - 1)\,\text{Eq. (5)} + \text{Eq. (4)} \leq -\frac{L}{2}\left[(\lambda_k - 1)\left[2\langle g_k, d_k\rangle + \|g_k\|^2\right] + \left[2\langle g_k, w_k - w^* + d_k\rangle + \|g_k\|^2\right]\right]$$

Let us first simplify the RHS,

$$\begin{aligned}
&\left[(\lambda_k - 1)\left[2\langle g_k, d_k\rangle + \|g_k\|^2\right] + \left[2\langle g_k, w_k - w^* + d_k\rangle + \|g_k\|^2\right]\right] \\
&= \lambda_k\left[2\langle g_k, d_k\rangle + \|g_k\|^2\right] - \left[2\langle g_k, d_k\rangle + \|g_k\|^2 - 2\langle g_k, w_k - w^* + d_k\rangle - \|g_k\|^2\right] \\
&= \frac{1}{\lambda_k}\left[\lambda_k^2\left(2\langle g_k, d_k\rangle + \|g_k\|^2\right) + 2\lambda_k\langle g_k, w_k - w^*\rangle\right] \\
&= \frac{1}{\lambda_k}\left[\|w_k - w^* + \lambda_k d_k + \lambda_k g_k\|^2 - \|w_k - w^* + \lambda_k d_k\|^2\right]
\end{aligned}$$

Putting everything together,

$$\lambda_k\left[(\lambda_k - 1)\,\text{Eq. (5)} + \text{Eq. (4)}\right] \leq \frac{L}{2}\left[\|w_k - w^* + \lambda_k d_k\|^2 - \|w_k - w^* + \lambda_k d_k + \lambda_k g_k\|^2\right] \quad (6)$$

Nesterov Acceleration for Smooth, Convex Functions

Now let us simplify the LHS of Eq. (6),

$$\lambda_k \left[(\lambda_k - 1)\,\text{Eq. (5)} + \text{Eq. (4)}\right] = \lambda_k \left[(\lambda_k - 1)\,(\Delta_{k+1} - \Delta_k) + \Delta_{k+1}\right] = \lambda_k^2\,\Delta_{k+1} - (\lambda_k^2 - \lambda_k)\,\Delta_k$$

Putting everything together,

$$\lambda_k^2\,\Delta_{k+1} - (\lambda_k^2 - \lambda_k)\,\Delta_k \leq \frac{L}{2}\left[\left\|w_k - w^* + \lambda_k d_k\right\|^2 - \left\|w_k - w^* + \lambda_k d_k + \lambda_k g_k\right\|^2\right]$$

We wish to sum from $k = 1$ to T, and telescope the terms. For the RHS, we want that,

$$w_k - w^* + \lambda_k d_k + \lambda_k g_k = w_{k+1} - w^* + \lambda_{k+1} d_{k+1} = w_k + d_k + g_k - w^* + \lambda_{k+1} d_{k+1}$$

$$= w_k + d_k + g_k - w^* + \lambda_{k+1}\beta_{k+1}[w_{k+1} - w_k]$$

$$= w_k + d_k + g_k - w^* + \lambda_{k+1}\,\beta_{k+1}[w_k + d_k + g_k - w_k]$$

$$\implies \text{We want that: } w_k - w^* + \lambda_k(d_k + g_k) = w_k - w^* + (1 + \lambda_{k+1}\,\beta_{k+1})\,[d_k + g_k]$$

This can be achieved if $\beta_{k+1} = \frac{\lambda_k - 1}{\lambda_{k+1}}$.

Recall that: $\lambda_k^2 \, \Delta_{k+1} - (\lambda_k^2 - \lambda_k) \, \Delta_k \leq \frac{L}{2} \left[\|w_k - w^* + \lambda_k d_k\|^2 - \|w_k - w^* + \lambda_k d_k + \lambda_k g_k\|^2 \right]$.
In order to telescope the LHS, we want that,

$$\lambda_{k-1}^2 = \lambda_k^2 - \lambda_k \implies \lambda_k = \frac{1 + \sqrt{1 + 4\lambda_{k-1}^2}}{2}$$

By using the sequence $\lambda_k = \frac{1+\sqrt{1+4\lambda_{k-1}^2}}{2}$ and setting $\beta_{k+1} = \frac{\lambda_k - 1}{\lambda_{k+1}}$,

$$\lambda_k^2 \, \Delta_{k+1} - \lambda_{k-1}^2 \, \Delta_k \leq \frac{L}{2} \left[\|w_k - w^* + \lambda_k d_k\|^2 - \|w_{k+1} - w^* + \lambda_{k+1} d_{k+1}\|^2 \right]$$

Summing from $k = 1$ to T, since $\lambda_0 = 0$

$$\begin{aligned}
\lambda_T^2 \Delta_{T+1} &\leq \frac{L}{2} \left[\|w_1 - w^* + \lambda_1 d_1\|^2 - \|w_{T+1} - w^* + \lambda_{T+1} d_{T+1}\|^2 \right] \\
&\leq \frac{L}{2} \|w_1 - w^*\|^2 \quad (\text{Since } w_0 = w_1 \implies d_1 = \beta_1 (w_1 - w_0) = 0)
\end{aligned}$$

$$\implies \Delta_{T+1} = f(w_{T+1}) - f^* \leq \frac{L}{2\lambda_T^2} \|w_1 - w^*\|^2 \tag{7}$$

Nesterov Acceleration for Smooth, Convex Functions

Recall that $f(w_{T+1}) - f^* \leq \frac{L}{2\lambda_T^2} \left\|w_1 - w^*\right\|^2$. Let us prove that $\lambda_k \geq \frac{k}{2}$ by induction.

Base case: $k = 1$, $\lambda_1 = \frac{1+\sqrt{1+4\lambda_0^2}}{2} = 1 \geq \frac{1}{2}$.
Inductive step: Assuming the statement is true for $k-1$ i.e. $\lambda_{k-1} \geq \frac{k-1}{2}$,
$\lambda_k = \frac{1+\sqrt{1+4\lambda_{k-1}^2}}{2} = \frac{1+\sqrt{1+(k-1)^2}}{2} \geq \frac{k}{2}$.

Hence, $\lambda_k \geq \frac{k}{2}$ and $\lambda_T \geq \frac{T}{2}$. Hence,

$$f(w_{T+1}) - f^* \leq \frac{2L \left\|w_1 - w^*\right\|^2}{T^2}$$

Hence, Nesterov acceleration with $\eta = \frac{1}{L}$ and a carefully engineered β_k sequence can obtain the accelerated $O\left(\frac{1}{T^2}\right)$ rate for smooth, convex functions.

Nesterov Acceleration for Smooth, Strongly-Convex Functions

Nesterov acceleration also results in the accelerated $O\left(\sqrt{\kappa}\log(1/\epsilon)\right)$ rate for smooth, strongly-convex functions.

In order to obtain this rate, the algorithm requires the following parameter settings: $\eta = \frac{1}{L}$ and,

$$\beta_k = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}$$

Refer to Bubeck, 3.7.1 for the analysis.

Compared to the smooth, convex setting for which β_k decreases, the strongly-convex setting requires a constant β_k in order to attain the accelerated rate.

Compared to GD, for smooth, strongly-convex functions, Nesterov acceleration requires knowledge of κ (and hence μ) in order to set β_k.

Unlike estimating L, estimating μ is difficult, and misestimating it can result in bad empirical performance. Common trick that results in decent performance is to use the convex parameters (with the decreasing β_k) with restarts.

Questions?