CMPT 409/981: Optimization for Machine Learning

Lecture 5

Sharan Vaswani

September 19, 2024

Recap

- For L-smooth, convex functions, GD with $\eta = {}^1/_L$ requires $T = O\left(\frac{1}{\epsilon}\right)$ iterations to return a point w_T that is ϵ-suboptimal meaning that $f(w_T) \leq f(w^*) + \epsilon$.
- **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.

Nesterov Acceleration

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

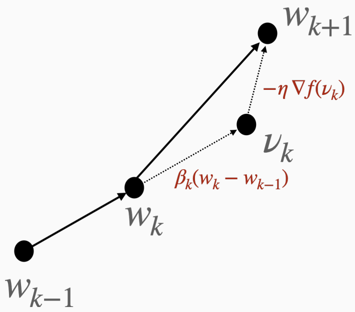
Nesterov Acceleration: $w_{k+1} = \mathrm{GD}(w_k + \beta_k(w_k - w_{k-1}))$ for $\beta_k \geq 0$ to be determined. Hence,

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

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, for $k \geq 1$,

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



Nesterov Acceleration

By eliminating w_k from the equation on the previous slide,

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

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 does not result in monotonic descent in the function values.

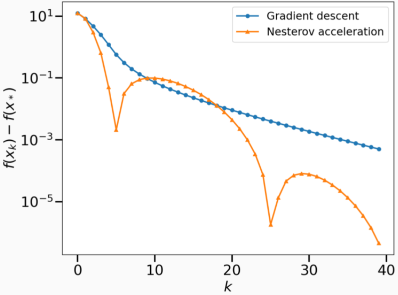


Figure 1: `https://francisbach.com/continuized-acceleration/`

Nesterov Acceleration for Smooth, Convex Functions

Analysis: 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 simplicity, set $w_1 = w_0$. For $k \geq 1$,

$$w_{k+1} = [w_k + \beta_k(w_k - w_{k-1})] - \eta \nabla f(w_k + \beta_k(w_k - w_{k-1}))$$
$$\implies w_{k+1} = w_k + d_k - \frac{1}{L}\nabla f(w_k + d_k) = w_k + d_k + g_k = \text{GD}(w_k + d_k)$$

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, 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 meaning 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 $y = y$

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

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

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

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

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

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

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

- We want to combine equations eq. (4) and eq. (5) in order to get a handle on Δ_T. For $\lambda_k > 1$, let us calculate $(\lambda_k - 1)$ eq. (5) + eq. (4) and also multiply both sides by λ_k,

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

- Let us first simplify the LHS,

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

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

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

Nesterov Acceleration for Smooth, Convex Functions

Simplifying the RHS: $-\frac{L\lambda_k}{2} \underbrace{\left[(\lambda_k - 1)\left[2\langle g_k, d_k\rangle + \|g_k\|^2\right] + \left[2\langle g_k, w_k - w^* + d_k\rangle + \|g_k\|^2\right]\right]}_{(*)}$.

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

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 $\beta_{k+1} = \frac{\lambda_k - 1}{\lambda_{k+1}}$.

Nesterov Acceleration for Smooth, Convex Functions

Recall that:

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

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

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

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

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

$$\implies \Delta_{T+1} = f(w_{T+1}) - f^* \le \frac{L}{2\lambda_T^2}\|w_1 - w^*\|^2 \tag{6}$$

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. $\lambda_{k-1} \geq \frac{k-1}{2}$,

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

This completes the induction. Hence, $\lambda_k \geq \frac{k}{2}$ and $\lambda_T \geq \frac{T}{2}$.

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

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.

Questions?

Strongly convex functions

First-order definition: 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 definition: 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 $f(x) = x^\intercal A x + bx + c$ 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?

Ans: Yes, because for all $y \in \mathcal{D}$, $f(y) > f(w^*)$ and hence w^* is a unique minimizer.

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.

Ans: Recall that $\nabla^2 f(w) = X^\intercal X + \lambda I_d$. Since $\nabla^2 f(w) \succeq (\lambda_{\min}[X^\intercal X] + \lambda)\, I_d$, ridge regression is μ-strongly convex with $\mu = \lambda_{\min}[X^\intercal X] + \lambda$.

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

Ans: Not necessarily, because $\nabla^2 f(w) = X^\intercal X$ might be low-rank, and have $\lambda_{\min}[X^\intercal X] = 0$.

Strongly-convex functions

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

Ans: Yes. $f''(x) = 1/x > 0$ for all $x \in (0, 1)$.

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?

Ans: For logistic regression, $\nabla^2 f(w) = X^\intercal D X$. Here, D is a diagonal matrix such that $D_{i,i} = p_i\,(1 - p_i)$ where $p_i = \sigma\left(\langle X_i, w \rangle\right)$ equal to $\Pr[\hat{y}_i = 1]$ (probability of prediction that point i has label equal to 1) and $\sigma(z) = \frac{1}{1+\exp(-z)}$ is the sigmoid function.
If $X^\intercal X$ is full-rank and $p_i \in (0, 1)$ (the probability of prediction is bounded away from 0 or 1) then $\nabla^2 f(w) \succeq \mu I_d$ for $\mu = \lambda_{\min}[X^\intercal D X]$.
This implies that if $X^\intercal X$ is full-rank, and the parameters are bounded (lie in a compact set) for example, for some finite $C \geq 0$, $\|w\| \leq C$, then, logistic regression is 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 $x = w^*$, $y = w_k$,

$$\implies \|w_{k+1} - w^*\|^2 \leq \|w_k - w^*\|^2 - 2\eta \langle \nabla f(w_k), w_k - w^* \rangle + 2L\eta^2 [f(w_k) - f(w^*)] \qquad (7)$$

GD for Smooth, Strongly-Convex Functions

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

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

$$\implies \langle \nabla f(w_k), w_k - w^* \rangle \geq f(w_k) - f(w^*) + \frac{\mu}{2} \|w_k - w^*\|^2 \tag{8}$$

Combining Eq. 7 and 8,

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

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

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

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? Ans: 1 since $L \geq \mu$.

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

Ans: Recall that $\nabla^2 f(w) = X^\top X + \lambda I_d$. Hence $\kappa = \frac{\lambda_{\max}[X^\top X] + \lambda}{\lambda_{\min}[X^\top X] + \lambda}$

GD for Smooth, Strongly-Convex Functions

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

Ans: Since f is smooth, $f(w_T) - f(w^*) \leq \frac{L}{2} \left\|w_T - w^*\right\|^2$. Hence, if $\left\|w_T - w^*\right\|^2 \leq \frac{2\epsilon}{L}$, this will guarantee that $f(w_T) - f(w^*) \leq \epsilon$. This requires $T \geq \frac{L}{\mu} \log\left(\frac{L\left\|w_0 - w^*\right\|^2}{2\epsilon}\right)$ iterations. We can also directly bound $f(w_T) - f(w^*)$ in terms of $f(w_0) - f(w^*)$ and obtain the same rate as for the iterates (In Assignment 2!).

- 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!).