CMPT 409/981: Optimization for Machine Learning

Lecture 6

Sharan Vaswani

October 3, 2022

Recap

Gradient Descent: $w_{k+1} = w_k - \eta \nabla f(w_k)$.

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

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

Function class	L-smooth	L-smooth + convex	L-smooth + μ-strongly convex
Gradient Descent	$\Theta\left(1/\epsilon\right)$	$O\left(1/\epsilon\right)$	$O\left(\kappa \log\left(1/\epsilon\right)\right)$
Nesterov Acceleration	-	$\Theta\left(1/\sqrt{\epsilon}\right)$	$\Theta\left(\sqrt{\kappa} \log\left(1/\epsilon\right)\right)$

Table 1: Optimization Zoo

For all cases, $\eta = \frac{1}{L}$ for both GD and Nesterov acceleration, and we can use Armijo line-search to estimate L and set the step-size.

Gradient Descent is adaptive to strong-convexity, however, Nesterov acceleration requires knowledge of μ to set β_k.

Heavy-Ball Momentum

Heavy-Ball/Polyak Momentum: $w_{k+1} = w_k - \eta \nabla f(w_k) + \beta_k (w_k - w_{k-1})$.

Nesterov Acceleration: $v_k = w_k + \beta_k (w_k - w_{k-1})\,;\, w_{k+1} = v_k - \eta \nabla f(v_k)$ i.e. extrapolate and compute the gradient at the extrapolated point v_k.

Polyak Momentum: $v_k = w_k + \beta_k (w_k - w_{k-1})\,;\, w_{k+1} = v_k - \eta \nabla f(w_k)$ i.e. compute the gradient at w_k and then extrapolate.

Unlike GD, Nesterov acceleration and Polyak momentum are not "descent" methods i.e. it is not guaranteed that $f(w_{k+1}) \leq f(w_k)$ for all k.

In order to minimize quadratics: $f(w) = \frac{1}{2} w^\mathsf{T} A w - bw + c$ where A is symmetric, positive semi-definite, or equivalently solve linear systems of the form: $Aw = b$, using Polyak momentum with *optimal* values of (η, β) is equivalent to Conjugate Gradient.

Heavy-Ball Momentum

Brief History: For L-smooth + μ-strongly convex functions,

- *Quadratics*: HB momentum with a specific (η, β) can achieve the accelerated rate and obtain a dependence on $\sqrt{\kappa}$ (only an asymptotic rate). [Polyak, 1964]
- *General smooth, SC functions*: Using Polyak's (η, β) parameters can result in cycling and HB momentum is not guaranteed to converge. [Lessard et al, 2014]
- *General smooth, SC functions*: Using a different (η, β), HB momentum can converge and match the GD rate (no acceleration). [Ghadimi et al, 2014]
- *General smooth, SC functions + Lipschitz-continuity of Hessian*: Using a different (η, β), HB momentum matches the GD rate at the beginning, but achieves the accelerated rate after $O(\kappa)$ iterations. [Wang et al, 2022]

Heavy-Ball Momentum

Let us focus on minimizing quadratics: $f(w) = \frac{1}{2}w^\intercal Aw - bw + c$, where A is a symmetric positive definite matrix.

Claim: For L-smooth, μ-strongly convex quadratics, HB momentum with $\eta = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ achieves the following convergence rate:

$$\|w_T - w^*\| \leq \sqrt{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} + \epsilon_T\right)^T \|w_0 - w^*\|$$

where $\epsilon_T \geq 0$ and $\lim_{T\to\infty} \epsilon_T = 0$.

HB momentum can also achieve a slightly-worse, but still accelerated non-asymptotic rate [Wang et al, 2021].

$$\|w_T - w^*\| \leq 4\sqrt{\kappa}\left(1 - \frac{1}{2\sqrt{\kappa}}\right)^T \|w_0 - w^*\|$$

Questions?

Minimizing strongly-convex quadratics with GD

As a warm-up, let us first prove the optimal GD rate for smooth, strongly-convex quadratics.

Claim: For L-smooth, μ-strongly convex quadratics, GD with $\eta = \frac{2}{\mu+L}$ achieves the following convergence rate:

$$\|w_T - w^*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^T \|w_0 - w^*\|$$

Proof: For quadratics, $\nabla f(w) = Aw - b$,

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

$$\implies \|w_{k+1} - w^*\| = \|w_k - w^* - \eta [A w_k - b]\|$$

$$= \|w_k - w^* - \eta [A w_k - A w^*]\| \qquad (\text{Since } \nabla f(w^*) = 0 \implies Aw^* = b)$$

$$\implies \|w_{k+1} - w^*\| = \|(I_d - \eta A)(w_k - w^*)\| \leq \|I_d - \eta A\|_2 \|w_k - w^*\|$$

(By definition of the matrix norm: for matrix B, $\|B\|_2 = \max\left\{\frac{\|Bv\|_2}{\|v\|_2}\right\}$ for all vectors $v \neq 0$, and)

We have thus reduced the problem to bounding $\|I_d - \eta A\|_2$.

Minimizing strongly-convex quadratics with GD

Recall that $\|w_{k+1} - w^*\| = \|I_d - \eta A\|_2 \, \|w_k - w^*\|$. Since f is L-smooth and μ-strongly convex, $\mu I_d \preceq \nabla^2 f(w) = A \preceq L I_d$.

If $A = U \Lambda U^\intercal$ is the eigen-decomposition of A, and $\lambda_1, \lambda_2, \ldots, \lambda_d$ are the eigenvalues of A, then, $I_d - \eta A = U S U^\intercal$ where $S_{i,i} = 1 - \eta \lambda_i$.

Since U is an orthonormal matrix, $\|I_d - \eta A\| = \|S\|$. By definition of the matrix norm, for symmetric matrices,

$$\|B\|_2 = \rho(B) := \max\{|\lambda_1[B]|, |\lambda_2[B]|, \ldots, |\lambda_d[B]|\}$$

where $\rho(B)$ is the spectral radius of B.

Hence,

$$\|I_d - \eta A\| = \|S\| = \rho(S) = \max\{|\lambda_1[S]|, |\lambda_2[S]|, \ldots, |\lambda_d[S]|\} = \max_{\lambda \in [\mu, L]} \{|1 - \eta\lambda|\}$$

$$\|I_d - \eta A\| = \max\{|1 - \eta\mu|, |1 - \eta L|\} \qquad \text{(Since } 1 - \eta\lambda \text{ is linear in } \lambda\text{)}$$

Minimizing strongly-convex quadratics with GD

Recall that $\|w_{k+1} - w^*\| = \|I_d - \eta A\| \, \|w_k - w^*\|$ and $\|I_d - \eta A\| = \max\{|1 - \eta\mu|, |1 - \eta L|\}$.

Let us choose a step-size $\eta \in \left[\frac{1}{L}, \frac{1}{\mu}\right]$. Hence,

$$\|I_d - \eta A\| \leq \max\{1 - \eta\mu, \eta L - 1\} = \frac{L - \mu}{L + \mu}$$

(By setting $\eta = \frac{2}{\mu + L}$, we minimize $\max\{1 - \eta\mu, \eta L - 1\}$)

Putting everything together,

$$\|w_{k+1} - w^*\| \leq \frac{L - \mu}{L + \mu} \|w_k - w^*\| = \frac{\kappa - 1}{\kappa + 1} \|w_k - w^*\|$$

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

$$\|w_T - w^*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^T \|w_0 - w^*\|.$$

Questions?

Minimizing strongly-convex quadratics with HB momentum

Update: $w_{k+1} = w_k - \eta \nabla f(w_k) + \beta(w_k - w_{k-1})$

Claim: For L-smooth, μ-strongly convex quadratics, HB momentum with $\eta = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ achieves the following convergence rate:

$\|w_T - w^*\| \leq \sqrt{2} \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} + \epsilon_T \right)^T \|w_0 - w^*\|$, where, $\lim_{T \to \infty} \epsilon_T \to 0$.

Proof:

$$\begin{bmatrix} w_{k+1} - w^* \\ w_k - w^* \end{bmatrix} = \begin{bmatrix} w_k - w^* - \eta \nabla f(w_k) + \beta(w_k - w_{k-1}) \\ w_k - w^* \end{bmatrix}$$

$$= \begin{bmatrix} w_k - w^* - \eta A(w_k - w^*) + \beta(w_k - w^*) - \beta(w_{k-1} - w^*) \\ w_k - w^* \end{bmatrix}$$

(Since $\nabla f(w) = Aw$, $Aw^* = b$)

$$\implies \begin{bmatrix} w_{k+1} - w^* \\ w_k - w^* \end{bmatrix} = \begin{bmatrix} (1+\beta)I_d - \eta A & -\beta I_d \\ I_d & 0 \end{bmatrix} \begin{bmatrix} w_k - w^* \\ w_{k-1} - w^* \end{bmatrix}$$

If $\beta = 0$, we can recover the same equation as GD.

Minimizing strongly-convex quadratics with HB momentum

$$\underbrace{\begin{bmatrix} w_{k+1} - w^* \\ w_k - w^* \end{bmatrix}}_{:=\Delta_{k+1} \in \mathbb{R}^{2d}} = \underbrace{\begin{bmatrix} (1+\beta)I_d - \eta A & -\beta I_d \\ I_d & 0 \end{bmatrix}}_{:=\mathcal{H} \in \mathbb{R}^{2d \times 2d}} \underbrace{\begin{bmatrix} w_k - w^* \\ w_{k-1} - w^* \end{bmatrix}}_{:=\Delta_k \in \mathbb{R}^{2d}} \implies \Delta_{k+1} = \mathcal{H}\,\Delta_k$$

Recursing from $k = 0$ to $T - 1$, and taking norm,

$$\|\Delta_T\| = \left\|\mathcal{H}^T \Delta_0\right\| \leq \left\|\mathcal{H}^T\right\| \left\| \begin{bmatrix} w_0 - w^* \\ w_{-1} - w^* \end{bmatrix} \right\| \qquad \text{(By definition of the matrix norm)}$$

Define $w_{-1} = w_0$ and lower-bounding the LHS,

$$\|w_T - w^*\| \leq \sqrt{2}\,\left\|\mathcal{H}^T\right\| \|w_0 - w^*\|$$

Hence, we have reduced the problem to bounding $\left\|\mathcal{H}^T\right\|$.

Minimizing strongly-convex quadratics with HB momentum

Recall that for symmetric matrices, $\|B\|_2 = \rho(B)$. Unfortunately, this relation is not true for general asymmetric matrices, and $\|B\| \geq \rho(B)$.

Gelfand's Formula: For a matrix $B \in \mathbb{R}^{d \times d}$ such that $\rho(B) := \max_{i \in [d]} |\lambda_i|$, then there exists a sequence $\epsilon_k \geq 0$ such that $\lim_{k \to \infty} \epsilon_k = 0$ and,

$$\left\| B^k \right\| \leq (\rho(B) + \epsilon_k)^k.$$

Using this formula with our bound,

$$\|w_T - w^*\| \leq (\rho(\mathcal{H}) + \epsilon_T)^T \|w_0 - w^*\|$$

Hence, we have reduced the problem to bounding $\rho(\mathcal{H})$.

Minimizing strongly-convex quadratics with HB momentum

Similar to the GD case, let $A = U\Lambda U^\intercal$ be the eigen-decomposition of A, then, $(1+\beta)\, I_d - \eta A = USU^\intercal$ where $S_{i,i} = 1 + \beta - \eta\lambda_i$. Hence,

$$\mathcal{H} = \begin{bmatrix} U^\intercal & 0 \\ 0 & U^\intercal \end{bmatrix} \underbrace{\begin{bmatrix} (1+\beta)I_d - \eta\Lambda & -\beta I_d \\ I_d & 0 \end{bmatrix}}_{:=H} \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}$$

Since U is orthonormal, $\rho(\mathcal{H}) = \rho(H)$. Hence we have reduced the problem to bounding $\rho(H)$.

Minimizing strongly-convex quadratics with HB momentum

Let P be a permutation matrix such that:

$$P_{i,j} = \begin{cases} 1 & i \text{ is odd}, \ j = i \\ 1 & i \text{ is even}, \ j = 2d + i \\ 0 & \text{otherwise} \end{cases} \qquad B = P\,H\,P^{\intercal} = \begin{bmatrix} H_1 & 0 & \dots & 0 \\ 0 & H_2 & \dots & 0 \\ \vdots & & \ddots & \\ 0 & & 0 & H_d \end{bmatrix}$$

where,

$$H_i = \begin{bmatrix} (1+\beta) - \eta\lambda_i & -\beta \\ 1 & 0 \end{bmatrix}$$

Note that $\rho(H) = \rho(B)$ (a permutation matrix does not change the eigenvalues). Since B is a block diagonal matrix, $\rho(B) = \max_i [\rho(H_i)]$. Hence we have reduced the problem to bounding $\rho(H_i)$.

Minimizing strongly-convex quadratics with HB momentum

For a fixed $i \in [2d]$, let us compute the eigenvalues of $H_i \in \mathbb{R}^{2\times 2}$ by solving the characteristic polynomial: $\det(H_i - uI_2) = 0$ w.r.t u.

$$u^2 - (1+\beta-\eta\lambda_i)u + \beta = 0 \implies u = \frac{1}{2}\left[(1+\beta-\eta\lambda_i) \pm \sqrt{(1+\beta-\eta\lambda_i)^2 - 4\beta}\right]$$

Let us set β such that, $(1+\beta-\eta\lambda_i)^2 \leq 4\beta$. This ensures that the roots to the above equation are complex conjugates. Hence,

$$1+\beta-\eta\lambda_i \geq -2\sqrt{\beta} \implies (\sqrt{\beta}+1) \geq \sqrt{\eta\lambda_i} \implies \beta \geq (1-\sqrt{\eta\lambda_i})^2$$

If we ensure that $\beta \geq (1-\sqrt{\eta\lambda_i})^2$

$$u = \frac{1}{2}\left[(1+\beta-\eta\lambda_i) \pm i\sqrt{4\beta - (1+\beta-\eta\lambda_i)^2}\right]$$

$$\implies |u|^2 = \frac{1}{4}\left[(1+\beta-\eta\lambda_i)^2 + 4\beta - (1+\beta-\eta\lambda_i)^2\right] = \beta \implies |u| = \sqrt{\beta}.$$

Hence, if $\beta \geq (1-\sqrt{\eta\lambda_i})^2$, $\rho(H_i) = \sqrt{\beta}$ and $\rho(B) = \max_i\left[\rho(H_i)\right] = \sqrt{\beta}$.

Minimizing strongly-convex quadratics with HB momentum

Using the result from the previous slide, if we ensure that for all i, $\beta \geq (1-\sqrt{\eta\lambda_i})^2$, then, $\rho(B) = \sqrt{\beta}$. Hence, we want that,

$$\beta = \max_i \{(1-\sqrt{\eta\lambda_i})^2\} = \max_{\lambda\in[\mu,L]} \{(1-\sqrt{\eta\lambda})^2\} = \max\{(1-\sqrt{\eta\mu})^2, (1-\sqrt{\eta L})^2\}$$

Similar to GD, we equate the two terms in the max,

$$1+\eta\mu - 2\sqrt{\eta\mu} = 1+\eta L - 2\sqrt{\eta L} \implies \eta = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}.$$

With this value of η, $\rho(\mathcal{H}) = \rho(H) = \rho(B) \leq \sqrt{\beta} = \sqrt{\left(1 - \frac{2\sqrt{\mu}}{(\sqrt{L}+\sqrt{\mu})}\right)^2} = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$.
Putting everything together,

$$\|w_T - w^*\| \leq \sqrt{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} + \epsilon_T\right)^T \|w_0 - w^*\|$$

Questions?