

Lecture 5

Last class

▷ Lower bounds

Today

▷ Momentum.

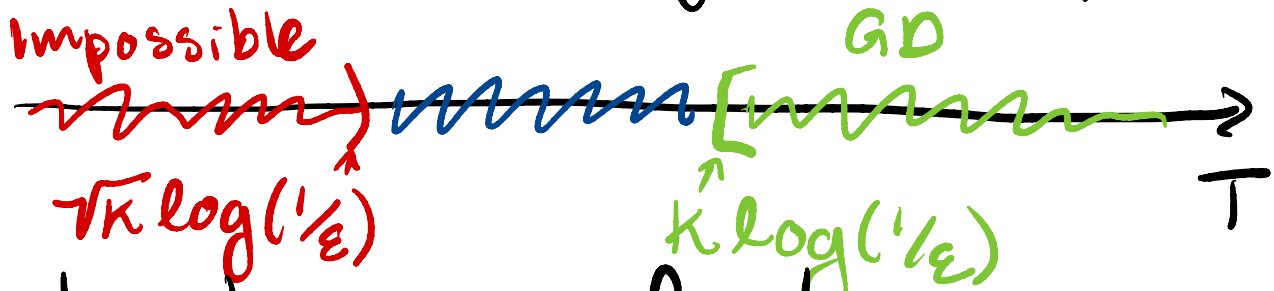
▷ A three sequence strategy

▷ Accelerated gradient descent.

We have a TA: Chandramauli Chakraborty
Email: chandramaulic@uchicago.edu
OH: Th 1:00-2:00 pm, Harper Memorial 141.

Momentum

Last class we saw a gap of the form



for strongly convex functions.

Today we will close this gap. The first idea is to look what happens on quadratics. For badly conditioned quadratics we get slow convergence with $1/L$ in the μ component.

Natural idea due to B. Polyak: Add momentum. That is,

$$(v) \quad w_{k+1} \leftarrow w_k - \eta_k \nabla f(w_k) + \underbrace{\beta (w_k - w_{k-1})}_{\text{Move in our previous direction.}}$$

This works optimally for quadratics (Polyak '64). Yet for general smooth functions, convex functions it is sub-optimal (Lessard, Recht, Packard 2016).

Nonetheless, it is interesting to see what it can teach us. If we unroll (v) to get

$$w_{k+1} - w_k = -\eta \sum_{j=0}^k \beta^{k-j} \nabla f(w_j).$$

Thus, we are using all previous gradient information; albeit, reweighed.

Algorithmic design principle

When possible use previous first-order information.

A three sequence strategy

We will use this principle in tandem with the principle from Lecture 2, to

construct "richer" models. Instead of tackling the μ -SC + L-smooth case, we will tackle convex + L-smooth (in HW 2 you'll use this result to get μ -SC + L-smooth).

Recall that convexity gives

$$(\heartsuit) \quad \ell_y(z) := f(y) + \langle \nabla f(y), z - y \rangle \leq f(z) \quad \forall z.$$

GD uses one gradient to build the model it minimizes at each iteration. Let's use all of them. Define

$$\gamma_k(z) = \sum_{j=1}^k a_j \ell_{y_{j-1}}(z) + \frac{1}{2} \|z - w_0\|^2$$

↑
 First iterate

with $a_j \geq 0$ and $y_j \in \mathbb{R}^d$.

To be picked ↗

Let $A_0 = 0$ and $A_k = a_1 + \dots + a_k$. Then, adding the inequalities $(\heartsuit)$ up yields

$$\gamma_k(z) \leq A_k f(z) + \frac{1}{2} \|z - w_0\|^2 \quad \forall z.$$

A simple first order optimality reveals

$$z_k = w_0 - \sum_{j=1}^k a_j \nabla f(y_{j-1}), \quad \text{or}$$

$$z_{k+1} \leftarrow z_k - a_{k+1} \nabla f(y_k).$$

Since $\psi_k(\cdot)$ is a quadratic with $\nabla^2 \psi_k(z) = I$,

$$\begin{aligned} \psi_k(z) &= \psi_k(z_k) + \langle \nabla f(z_k), z - z_k \rangle + \frac{1}{2} \|z - z_k\|_2^2 \\ (\ddot{v}) \quad &= \psi_k(z_k) + \frac{1}{2} \|z - z_k\|_2^2. \end{aligned}$$

The idea to derive a method is to have two sequences $\{w_k\}$ and $\{y_k\}$ such that

$$(I_k) \quad A_k f(w_k) \leq \min_z \psi_k(z) = \psi(z_k). \quad \text{Third sequence}$$

By evaluating at $z = w^*$ we get

$$A_k f(w_k) \leq \psi_k(z_k) \leq \psi_k(w^*) \leq A_k f^* + \frac{1}{2} \|w_0 - w^*\|^2$$

$$(\ddot{u}) \quad \Rightarrow \quad f(w_k) - f^* \leq \frac{\|w_0 - w^*\|^2}{2A_k}.$$

Before we pick the sequences $\{y_k\}$. Let's assume we always update

$$\begin{aligned} (oo) \quad z_{k+1} &\leftarrow z_k - a \nabla f(y_k), \\ w_{k+1} &\leftarrow y_k - \frac{1}{L} \nabla f(y_k). \end{aligned} \quad \begin{array}{l} \text{To be set.} \\ \text{Conservative stepsize.} \end{array}$$

Question: Suppose that (I_k) holds,

how do we set a and y to ensure (I_{k+1}) ?

Lemma (One step): Suppose (I_k) holds. Then,
$$\Psi_{k+1}(z_{k+1}) \geq A_{k+1} f(w_{k+1}) + \langle \nabla f(y), A_k(w_k - y) + a(z_k - y) \rangle + \frac{1}{2} \left(\frac{A_{k+1}}{L} - a^2 \right) \|\nabla f(y)\|^2.$$

In particular, (I_{k+1}) holds if

(i) $y = \frac{A_k w_k + a z_k}{A_{k+1}},$ and

(ii) $a^2 \leq \frac{A_{k+1}}{L}.$

Proof: By (i) and (I_k) , then $\forall z$,

$$\begin{aligned} \Psi_{k+1}(z) &= \Psi_k(z) + a \ell_y(z) \\ &\geq A_k f(w_k) + \frac{1}{2} \|z - z_k\|^2 \end{aligned}$$

Minimized at $z = z_{k+1} = z_k - a \nabla f(y)$

$$+ a(f(y) + \langle \nabla f(y), z - y \rangle)$$

$$\begin{aligned} &\geq A_k f(w_k) + a f(y) + a \langle \nabla f(y), z_k - y \rangle \\ &\quad - \frac{a^2}{2} \|\nabla f(y)\|^2 \end{aligned}$$

Convexity
(ii)

$$\geq (A_k + a) f(y)$$

$$+ \langle \nabla f(y), A_k(w_k - y) + a(z_k - y) \rangle \\ - \frac{a^2}{2} \|\nabla f(y)\|^2$$

Descent
Lemma

$$\begin{aligned} & \overset{z}{\uparrow} A_{k+1} f(w_{k+1}) \\ & + \langle \nabla f(y), A_k(w_k - y) + a(z_k - y) \rangle \\ & + \left(A_{k+1} - \frac{a^2}{2} \right) \|\nabla f(y)\|^2 \end{aligned}$$

□

A direct consequence of this is the analysis of vanilla GD. Take

$$a_j = 1/L, \quad \text{and} \quad z_0 = w_0.$$

Then, it is too hard to see that

$$y_k = z_k = w_k \quad \forall k.$$

Theorem: Suppose $f: \mathbb{R}^d \rightarrow \mathbb{R}$ is convex and L -smooth. Then GD with $\eta = 1/L$ yields iterates satisfying

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

+

Accelerated Gradient Descent

Taking $a_k = 1/L$ satisfies (ii), but it leaves a lot of slack. Instead we can take $a_k = \sqrt{\frac{A_k}{L}}$. Noting

$$\sqrt{A_{k+1}} - \sqrt{A_k} = \frac{a_{k+1}}{\sqrt{A_{k+1}} - \sqrt{A_k}} \geq \frac{a_{k+1}}{2\sqrt{A_{k+1}}} = \frac{1}{2\sqrt{L}},$$

implies that

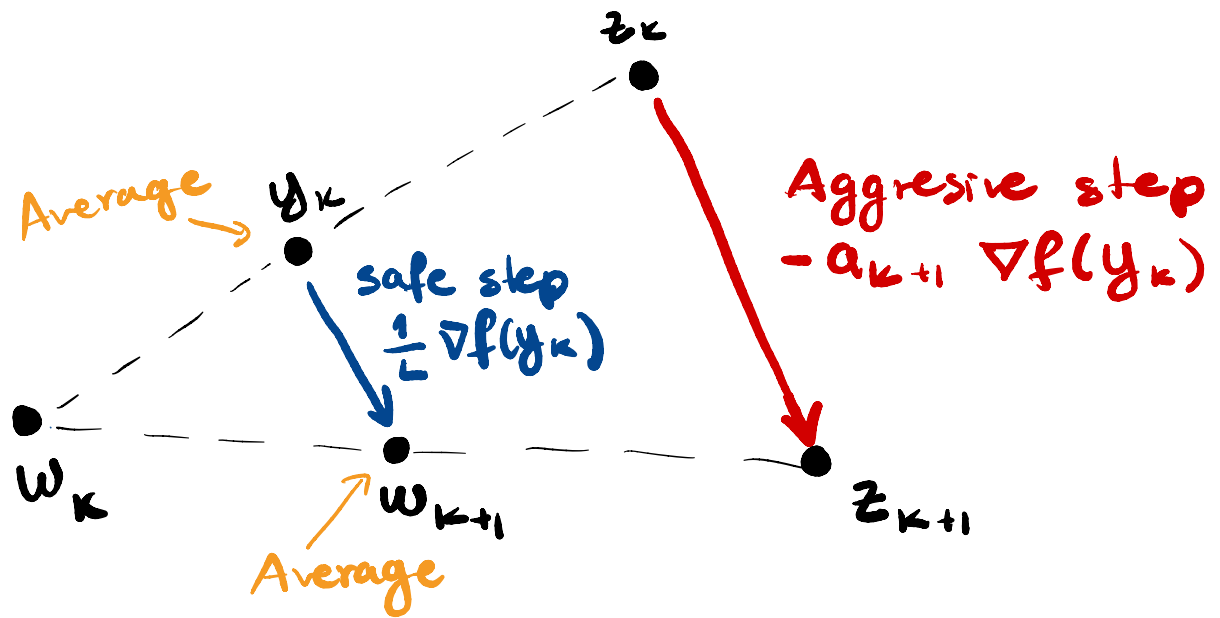
$$(\uparrow) \sqrt{A_{k+1}} - \sqrt{A_0} \geq \frac{k}{2\sqrt{L}} \Rightarrow A_{k+1} \geq \frac{k^2}{2L}.$$

$a_{k+1} \geq \frac{k}{\sqrt{2}L}$
↑

The price to pay is that z_k might diverge if we don't couple it with a much more conservative w_k . To see why this is more conservative, note

$$\begin{aligned} w_{k+1} &= y_k - \frac{1}{L} \nabla f(y_k) \\ &= \frac{A_k w_k + a_{k+1} z_k}{A_{k+1}} - \frac{a_{k+1}^2}{A_{k+1}} \nabla f(y_k) \\ &= \frac{A_k w_k + a_k z_{k+1}}{A_{k+1}}. \end{aligned}$$

Thus,



This exactly the method that Nesterov proposed in '83. Set a_{k+1} recursively via

$$La_{k+1}^2 = A_k + a_{k+1}$$

and iteratively update

$$y_k \leftarrow \frac{A_k w_k + a_{k+1} z_k}{A_{k+1}},$$

(AGD) $w_{k+1} \leftarrow y_k - \frac{1}{L} \nabla f(y_k),$ and

$$z_{k+1} \leftarrow z_k - a_{k+1} \nabla f(y_k).$$

Theorem: Suppose that $f: \mathbb{R}^d \rightarrow \mathbb{R}$ is convex, L -smooth function. Then, the iterates of (AGD) satisfy

$$f(w_k) - f^* \leq \frac{2L \|w_0 - w^*\|^2}{k^2} \quad \forall k.$$

Proof: We prove that (I_k) holds for all k . The first inequality (I_0) holds since $A_0 = 0 = \min \psi_0$. Then, lemma (one step) proves $(I_k) \forall k$. Then $(\uparrow)$ and (ii) gives the result. $\square$

Remarkably, we can write this method as a modified version of Polyak's heavy ball. Define

$$\lambda_0 = 0 \quad \text{and} \quad \lambda_{k+1} = \frac{1 + \sqrt{1 + 4\lambda_k^2}}{2},$$

further,

$$w_{k+1} \leftarrow y_k - \frac{1}{L} \nabla f(y_k), \quad y_k \leftarrow w_k + \beta_k (w_{k+1} - w_k)$$

$$\text{with} \quad \beta_k = \frac{\lambda_{k+1} - 1}{\lambda_{k+2}}.$$

Exercise: Show that the iterates generated by this update are the same as those generated by (AGD). $\dashv$