Lecture 6

HW 1 was due an hour ago.

Scribe?

Last time

- Subdifferential Calculus
- Gradient Descent
- Descent Lemma
- Stepsizes

Today

- Nonconvex smooth guarantees
- Characterization of L-smooth convex f.
- Better guarantees for convex.

Nonconvex smooth opt guarantees

Consider solving $\min_{x \in \mathbb{R}^d} f(x)$ with L-Lipschitz gradient via

$$x_{k+1} \leftarrow x_k - \alpha_k \nabla f(x_k)$$

with $x_0 \in \mathbb{R}^d$.

Theorem Suppose f is diff with L-Lips grad. Then for $T \geq 0$

$$\frac{1}{T} \sum_{k=0}^{T-1} \|\nabla f(x_k)\|_2^2 \leq \frac{2L\,(f(x_0) - \min f)}{T}$$

when $\alpha_k = 1/L$ or with exact linesearch.

Moreover,

$$\frac{1}{T}\sum_{k=0}^{T-1} \|\nabla f(x_k)\|_2^2 \le \max\left\{\frac{1}{\eta a}, \frac{L}{2\tau\eta(1-\eta)}\right\} \frac{(f(x_0) - \min f)}{T}$$

when we use Armijo backtracking. $+$

Consequence

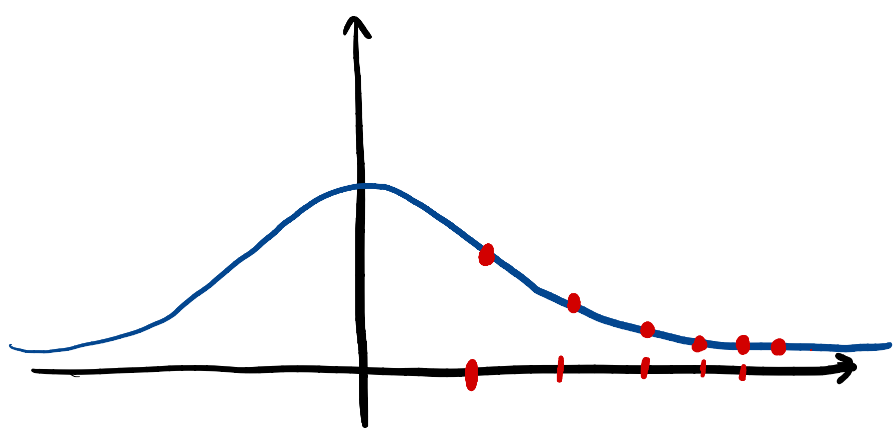
$T \geq c\frac{1}{\varepsilon}$ for $c>0$.

Picking $T = \Omega(\frac{1}{\varepsilon})$ then

$\exists k \le T$ s.t. $\|\nabla f(x_k)\|_2^2 \le \varepsilon$.

Warnings

- x_k might not converge! Consider $f(x) = \exp(-x^2)$



- Even if $x_k \to x^*$, the limit might not be a local min.

Exercise: Think of an example where this happens.

Proof: We prove it for $\alpha_k = \frac{1}{L}$, the rest of the proofs are similar.

By DL, we have $\forall k \geq 0$

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2$$

Summing all of these up to $T-1$

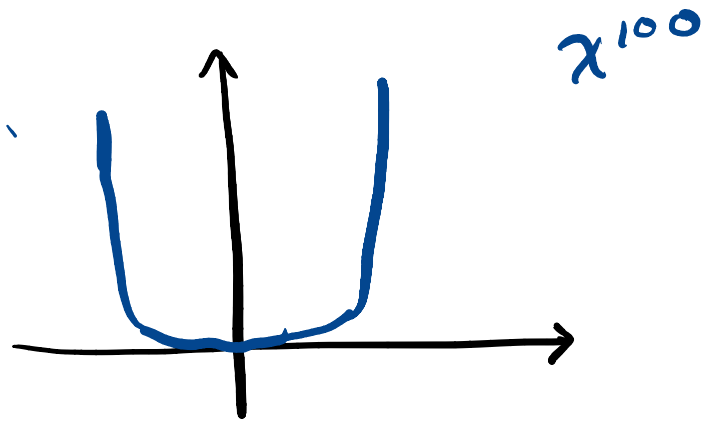
$$f(T) \leq f(x_0) - \frac{1}{2L} \sum_{k=0}^{T-1} \|\nabla f(x_k)\|^2$$

$$\Rightarrow \sum_{k=0}^{T-1} \|\nabla f(x_k)\|^2 \leq 2L\left[f(x_0) - f(x_T)\right]$$

$$\leq 2L\left[f(x_0) - \min f\right].$$

Dividing both sides by T gives the result. $\square$

The reason why we have such slow converges is that our function can grow very slowly



when the gradient is small, you don't move that much.

Theorem. Assume f is twice diff and x^* is a second-order critical point

$$\nabla f(x^*) = 0 \quad \text{and} \quad \nabla^2 f(x^*) \succeq \lambda I$$

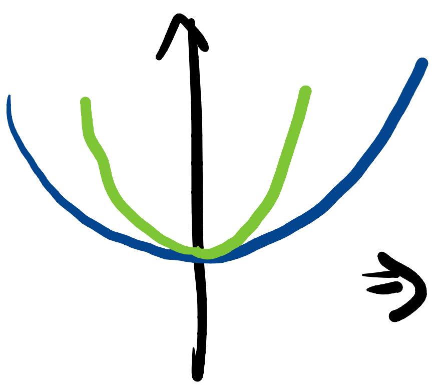
$$\lambda_{min}(\nabla^2 f(x^*)) \geq \lambda$$

There exists $\varepsilon > 0$ s.t if $x_k \in B_\varepsilon(x^*)\ \forall_k$ then

$$f(x_T) - f(x^*) \leq \left(1 - \frac{\lambda^2}{2L^2}\right)\left(f(x_k) - f(x^*)\right).$$

Intuition

For points where 2nd-order approximation grows, we have that if we start close



$$\Rightarrow T = \Omega\left(\left(\frac{\lambda^2}{L^2}\right)^{-1} \log\left(\frac{f(x_0) - f(x^*)}{\varepsilon}\right)\right)$$

suffice for $f(x_T) - f(x_0) \le \varepsilon$.

Proof: Since $\lambda_{min}(\nabla^2 f(x))$ is continuous $\Rightarrow \exists\, \varepsilon > 0$ s.t. $\forall x \in B_\varepsilon(x^*)$

$$\lambda_{min}(\nabla^2 f(x)) \ge \frac{\lambda}{2}.$$

Then, for any $\|\bar{s}\| \le \varepsilon$ we can define

$$\varphi_s(t) = f(x^* + t\bar{s}) \quad \text{and}$$

$$\varphi'(1) = \varphi'(0) + \int_0^1 \varphi''(t)\,dt$$

$$\Rightarrow \nabla f(x^* + \bar{s})^T s = 0 + \int_0^1 \underbrace{s^T \nabla^2 f(x^* + t\bar{s})\, s}_{\ge \frac{\lambda}{2}\|s\|^2}\,dt$$

$$\ge \frac{\lambda}{2}\|s\|^2.$$

$$\Rightarrow \frac{\lambda\|s\|}{2} \le \|\nabla f(x+s)\|. \quad (\because)$$

By Taylor Approximation:

$$\frac{L}{2}\|s\|^2 \geq f(x^*+s) - \left(f(x^*) + 0^T s\right)$$
$$= f(x^*+s) - f(x^*) \qquad (♡)$$

Combining (☺) and (♡)

$$\frac{4}{\lambda^2}\|\nabla f(x+s)\|^2 \geq \frac{2}{L}\left(f(x^*+s) - f(x^*)\right) \qquad (★)$$

Then, using DL

$$f(x_{k+1}) - f(x^*) \leq f(x_k) - f(x^*) - \frac{2}{L}\|\nabla f(x_k)\|^2$$

Follows from (★)

$$\leq \left(1 - \frac{\lambda^2}{2L^2}\right)\left(f(x_k) - f(x^*)\right).$$

$\square$

Better guarantees for convex functions

Lemma (Characterization L-smoothness for convex functions)

Suppose that $f: \mathbb{R}^d \to \mathbb{R}$ is diff and convex.

Then the following are equivalent

1) f has L-Lipschitz gradient

2) $\frac{L}{2}\|\cdot\|_2^2 - f(\cdot)$ is convex.

3) $f(y) \le f(x) + \langle \nabla f(x), y-x \rangle + \frac{L}{2}\|x-y\|^2 \quad \forall x, y$

4) $\langle \nabla f(y) - \nabla f(x), y-x \rangle \ge \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2 \quad \forall x, y.$

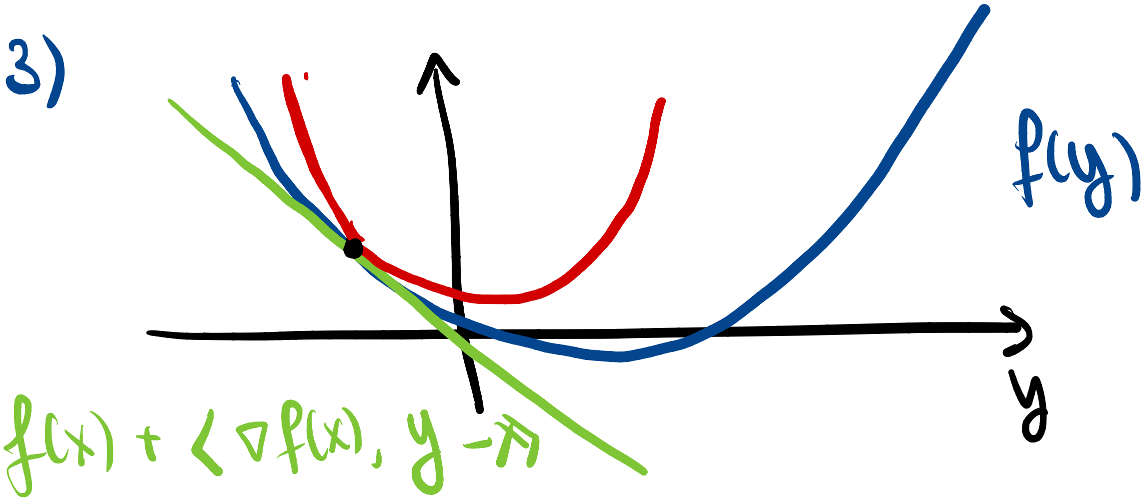
If further f is twice diff the following are also equivalent to the above

5) $\nabla^2 f(x) \preceq LI \quad \forall x \quad (LI - \nabla^2 f(x) \succeq 0)$

Intuition

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

3)



Proof: (2) $\Leftrightarrow$ (5) $h(x) = \frac{L}{2}\|x\|^2 - f(x)$ is convex

$\Leftrightarrow \quad \nabla^2 h(x) \succeq 0$

$\Leftrightarrow \quad LI \succeq \nabla^2 f(x)$

second order characterization

(2) $\Leftrightarrow$ (3) $h(x) = \frac{L}{2}\|x\|^2 - f(x)$ is convex

$\Leftrightarrow \quad h(x) + \langle \nabla h(x), y-x \rangle \leq h(y) \quad \forall y, x$

$\Leftrightarrow \quad \frac{L}{2}\|x\|^2 - f(x) + L\langle x, y-x \rangle - \langle \nabla f(x), y-x \rangle \leq \frac{L}{2}\|y\|^2 - f(y)$

$\Leftrightarrow \quad f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + \frac{L}{2}\|x-y\|^2$

(4) $\Rightarrow$ (1) By Cauchy-Schwarz

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq \|\nabla f(x) - \nabla f(y)\| \, \|x-y\|$$

$$\Updownarrow$$

$$\|\nabla f(x) - f(y)\| \leq L\|x-y\|$$

(1) $\Rightarrow$ (3) Taylor Approximation Theorem.