Last time:

- Grad Descent
- Convergence in a simple crx setting

Today: More structure

- Convergence Rates of GD for:
  - strongly convex fns
  - smooth
  - nonconvex smooth
- What do these mean for practical setups.

Reminder:

str. convexity: $f(x)$ is λ str-crx if $f(x) - \frac{\lambda}{2}\|x\|^2$ is crx

Lm: If $f(\,)$ is λ-str convex and β-smooth. Then:

$$(\nabla f(x) - \nabla f(y))^T (x-y) \geq \frac{\lambda\beta}{\lambda+\beta}\|x-y\|^2 + \frac{1}{\beta+\lambda}\|\nabla f(x) - \nabla f(y)\|^2$$

Cor: $\langle \nabla f(x), x - x^* \rangle \geq c\|x-x^*\|^2 + c'\|\nabla f(x)\|^2$

(strong correlation towards opt)
"attraction"

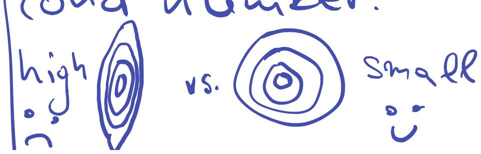
Theorem:

Let f be β-smooth and λ-strongly convex. The Grad. Descent with $\gamma = \frac{2}{\lambda+\beta}$ obtains

$$\|x_t - x^*\|^2 \leq e^{-2t/\kappa}\|x_0 - x^*\|^2$$

where $\kappa = \beta/\lambda$

cond number:



| "stretch of function"

Proof:

$$\underbrace{\|x_{k+1}-x^*\|^2}_{\Delta_{t+1}} = \|x_k - \gamma \nabla f(x_k) - x^*\|^2$$

$$= \Delta_t - 2\gamma \underbrace{\langle \nabla f(x_k), x_k - x^* \rangle}_{\text{apply co-coercivity}} + \gamma^2 \|\nabla f(x_k)\|^2$$

$$\leq \Delta_t - 2\gamma \left(\frac{\lambda\beta}{\lambda+\beta} \|x_t - x^*\|^2 + \frac{1}{\beta+\lambda} \|\nabla f(x_t)\|^2 \right) + \gamma^2 \|\nabla f(x_t)\|^2$$

$\gamma = \frac{2}{\lambda+\beta}$

$$= \|x_t - x^*\| - \frac{4\beta}{(\lambda+\beta)^2} \|x_t - x^*\|^2 + \left[\cancel{\frac{4}{(\lambda+\beta)^2} - \frac{4}{(\lambda+\beta)^2}} \right] \|\nabla f(x_t)\|^2$$

$$= \left(1 - \frac{4\beta}{(\lambda+\beta)^2}\right) \|x_t - x^*\|^2$$

$$= \left(\frac{\lambda^2 + 2\beta + \beta^2 - 4\beta}{(\lambda+\beta)^2} \right) \|x_t - x^*\|^2$$

$$= \left(\frac{\lambda-\beta}{\lambda+\beta}\right)^2 \|x_t - x^*\|^2 = \left(\frac{1-\beta/\lambda}{1+\beta/\lambda}\right)^2 \|x_t-x^*\|^2$$

$$= \left(\frac{\kappa-1}{\kappa+1}\right)^2 \|x_t-x^*\|^2 = \left(\frac{\kappa-1}{\kappa+1}\right)^{2\cdot t}\|x_0-x^*\|^2$$

$$= e^{2t\log\left(1-\frac{1}{\kappa}\right)}\|x_0-x^*\|^2 \le e^{-2t/\kappa}\|x_0-x^*\|^2$$

Comparison of Conv. Rates:

L-Lip:	$\frac{R\cdot L}{\sqrt{T}}$
λ-str.cvx: + β-smooth	$\beta R e^{-2\frac{\lambda}{\beta}}$
β-smooth:	$\frac{\beta\cdot R^2}{T}$
Str_cvx + L-Lip	: $\frac{L^2}{\lambda\cdot T}$

Structure helps us analyze convergence!

Remark:
Not always tight.

Q: What do thes "constants" look like for some real problems?

Computational Complexity of GD:

- Let one ∇f eval be our unit of comp. cost.

- Goal: Measure the total cost wrt $\nabla f(x)$ evals.

- Total cost: $O\left([\#\nabla f \text{ evals}] \cdot [\#\text{iter}(\varepsilon)]\right)$

Let's see an example:

Log. reg. + regularization

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1+e^{-y_i\langle x_i, w\rangle}\right) + \frac{\lambda}{2}\cdot\|w\|^2$$

Reminder:

- $\log(1+e^{x})$ is 1-Lip, 1/4-smooth
- $g_1(g_2(x))$ is $L_{g_1} L_{g_2}$-Lip
- $x^T w + b$ is $\|x\|$-Lip
- $g(x^T w + b)$ is $\beta\|x\|^2$-smooth

$$\frac{1}{n}\sum_i \log\left(1+e^{-y_i\langle x_i,w\rangle}\right)$$

is

- $\frac{1}{n}\sum_i \|x_i\|$ - Lip (L)
- $\frac{1}{n}\sum_i \frac{1}{4}\|x_i\|^2$ - smooth (β_1)

$\frac{\lambda}{2}\|w\|^2$

- is λ str. cvx
- not Lip!
- λ- smooth (β_2)

- Asm Let $\|x_i\|_2 = O(d)$, $\|w^*\| = O(d)$

Then, $R = \|w_0 - w^*\| = O(\sqrt{d})$, $L = O(\sqrt{d})$

$\beta = O(d)$, $\lambda = O(1)$

Rates:

- If $\lambda = 0$. $\frac{\beta R^2}{T} = \frac{d^2}{T}$
- If $\lambda \neq 0$ $\beta R^2 e^{-\frac{T}{2+\frac{\beta}{\lambda}}} = d^2 e^{-c\frac{T}{d}}$

For ε - error $\Rightarrow$ $T = O\left(\frac{d^2}{\varepsilon}\right)$ $[\lambda = 0]$

or if $\lambda \neq 0$ $T = O\left(d \log\left(\frac{d}{\varepsilon}\right)\right)$

Q: Overall complexity?

Time to compute $\nabla f(x)$ is proportional to computing $\langle x_i, w \rangle$ $\forall i$, i.e., $O(n \cdot d)$.

<u>Cor.</u> For $\lambda = O(1)$ and $B = O(d)$ GD takes $O(nd^2 \log d/\varepsilon)$ on logistic regression.

Cost / iteration: $O(nd)$ it requires 1 pass over the data!

Too expensive! Can we make this $O(d)$?

Answer: SGD. Next time!