CS-GY 9223 I: Lecture 6
Smoothness, Strong convexity, and more.

NYU Tandon School of Engineering, Prof. Christopher Musco

Assume:

- f is convex.
- Lipschitz function: for all $\mathbf{x}$, $\|\nabla f(\mathbf{x})\|_2 \leq G$.
- Starting radius: $\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_2 \leq R$.

Gradient descent:

- Choose number of steps T.
- $\eta = \frac{R}{G\sqrt{T}}$
- For $i = 1, \ldots, T$:
    - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$
- Return $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}^{(i)}} f(\mathbf{x}^{(i)})$.

Theorem (GD Convergence Bound)

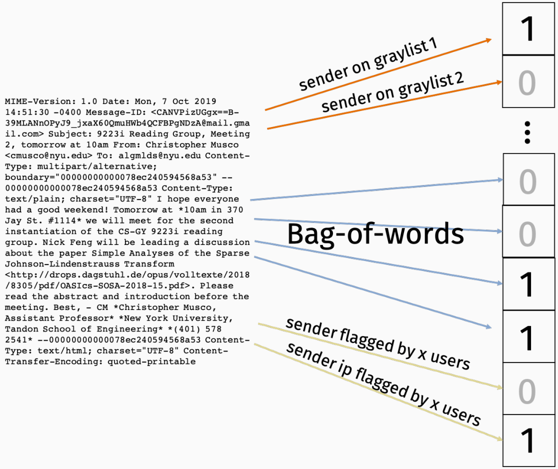
If $T \geq \frac{R^2 G^2}{\epsilon^2}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Instead of a single function f to minimize, assume we have an unknown and changing set of objective functions:

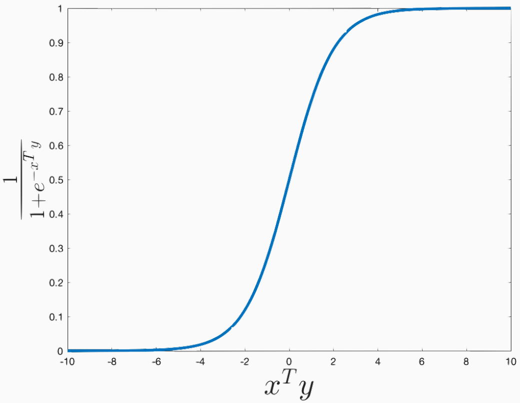
$$f_1, \ldots, f_T.$$

- At each time step, choose $\mathbf{x}^{(i)}$.
- f_i is revealed and we pay cost $f_i(\mathbf{x}^{(i)})$
- **Goal**: Minimize $\sum_{i=1}^{T} f_i(\mathbf{x}^{(i)})$.

Email spam filtering:



- $M_{\mathbf{x}}(\mathbf{y}) = \frac{1}{1+e^{-\mathbf{x}^T\mathbf{y}}}$



Predict $\mathbf{y}$ as spam if $M_{\mathbf{x}}(\mathbf{y}) \geq \frac{1}{2}$.

Logistic loss:

Given label $b \in \{0, 1\}$,

$$L(b, M_{\mathbf{x}}(\mathbf{y})) = -b \log\left(M_{\mathbf{x}}(\mathbf{y})\right) + (1-b) \log\left(1 - M_{\mathbf{x}}(\mathbf{y})\right)$$

Total cost of over time:

$$\sum_{i=1}^{T} L(b^{(i)}, M_{\mathbf{x}^{(i)}}(\mathbf{y}^{(i)})))$$

where $\mathbf{y}^{(i)}$ is the i^{th} email and $b^{(i)}$ is the i^{th} label.

How should we measure how well we did?

For some small value Δ, can we achieve:

$$\sum_{i=1}^{T} f_i(\mathbf{x}^{(i)}) \leq \left[\min_{\mathbf{x}} \sum_{i=1}^{T} f_i(\mathbf{x})\right] + \Delta.$$

I.e. can we compete with the best <u>fixed</u> solution in hindsight.

Δ = "regret"

Assume:

- Lipschitz functions: for all $\mathbf{x}$, i, $\|\nabla f_i(\mathbf{x})\|_2 \leq G$.
- Starting radius: $\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_2 \leq R$.

Online Gradient descent:

- Choose number of steps T.
- $\eta = \frac{D}{G\sqrt{T}}$
- For $i = 1, \ldots, T$:
  - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f_i(\mathbf{x}^{(i)})$
- Play $\mathbf{x}^{(i+1)}$.

Claim (OGD Regret Bound)

After T steps, $\Delta = \left[\sum_{i=1}^{T} f_i(\mathbf{x}^{(i)})\right] - \left[\sum_{i=1}^{T} f_i(\mathbf{x}^*)\right] \leq RG\sqrt{T}$

Recall the machine learning setup. In empirical risk minimization, we can typically write:

$$f(\mathbf{x}) = \sum_{j=1}^{n} f_j(\mathbf{x})$$

where f_i is the loss function for a particular data point.

Linear regression:

$$f(\mathbf{x}) = \sum_{j=1}^{n} (\mathbf{x}^T \mathbf{y}^{(j)} - b^{(j)})^2$$

Pick random $j \in 1, \ldots, n$:

$$\mathbb{E}\left[\nabla f_j(\mathbf{x})\right] = \nabla f(\mathbf{x}).$$

But $\nabla f_j(\mathbf{x})$ can often be computed in a $1/n$ fraction of the time!

Main idea: Use random approximate gradient in place of actual gradient.

Trade slower convergence for cheaper iterations.

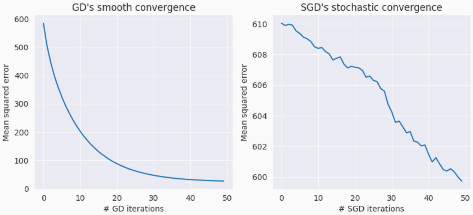
Assume:

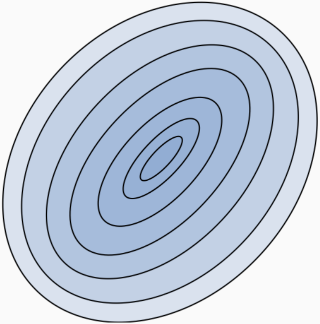
- Lipschitz functions: for all $\mathbf{x}$, j, $\|\nabla f_j(\mathbf{x})\|_2 \leq \frac{G'}{n}$.
- Starting radius: $\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_2 \leq R$.

Stochastic Gradient descent:

- Choose number of steps T.
- $\eta = \frac{D}{G'\sqrt{T}}$
- For $i = 1, \ldots, T$:
    - Pick random $j_i \in 1, \ldots, n$.
    - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f_{j_i}(\mathbf{x}^{(i)})$
- Return $\hat{\mathbf{x}} = \frac{1}{T}\sum_{i=1}^{T} \mathbf{x}^{(i)}$

VISUALIZING SGD





Claim (SGD Convergence)

After $T = \frac{R^2 G'^2}{\epsilon^2}$ iteration:

$$\mathbb{E}\left[f(\hat{\mathbf{x}}) - f(\mathbf{x}^*)\right] \leq \epsilon.$$

Claim (SGD Convergence)

After $T = \frac{R^2 G'^2}{\epsilon^2}$ *iteration:*

$$\mathbb{E}\left[f(\hat{\mathbf{x}}) - f(\mathbf{x}^*)\right] \leq \epsilon.$$

Number of iterations for error ϵ:

- **Gradient Descent**: $T = \frac{R^2 G^2}{\epsilon^2}$.
- **Stochastic Gradient Descent**: $T = \frac{R^2 G'^2}{\epsilon^2}$.

Always have $G \leq G'$:

$$\|\nabla f(x)\|_2 \leq \|\nabla f_1(x)\|_2 + \ldots + \|\nabla f_n(x)\|_2 \leq n \cdot \frac{G'}{n} = G'.$$

Fair comparison:

- SGD cost = (# of iterations) $\cdot$ $O(1)$
- GD cost = (# of iterations) $\cdot$ $O(n)$

Stochastic vs. Full Batch Gradient Descent:

Can the convergence bounds be tightened for certain functions? Can they guide us towards faster algorithms?

Goals:

- Improve ϵ dependence below $1/\epsilon^2$.
- Reduce or eliminate dependence on G and R.
- Etc.

Definition (β-smoothness)

A function f is β smooth if, for all $\mathbf{x}, \mathbf{y}$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq \beta \|\mathbf{x} - \mathbf{y}\|_2$$

β is a parameter that will depend on our function.

Recall from definition of convexity that:

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \nabla f(\mathbf{x})^T(\mathbf{x} - \mathbf{y})$$

How much smaller can left hand side be?

$$\nabla f(\mathbf{x})^T(\mathbf{x} - \mathbf{y}) - [f(\mathbf{x}) - f(\mathbf{y})] \leq \frac{\beta}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$$

Previously learning rate/step size η depended on G. Now choose it based on β:

$$\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} - \frac{1}{\beta}\nabla f(\mathbf{x}^{(t)})$$

Progress per step of gradient descent:

Theorem (GD convergence for β-smooth functions.)

Let f be a β smooth convex function and assume we have $\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_2 \leq R$. If we run GD for T steps with $\eta = \frac{1}{\beta}$ we have:

$$f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \leq \frac{2\beta R^2}{T-1}$$

Corollary: If $T = O\left(\frac{\beta R^2}{\epsilon}\right)$ we have $f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \leq \epsilon$.

Definition (α-strongly convex)

A convex function f is α-strongly convex if, for all $\mathbf{x}, \mathbf{y}$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$$

α is a parameter that will depend on our function.

Completing the picture: If f is α strongly convex and β smooth,

$$\frac{\alpha}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \nabla f(\mathbf{x})^T(\mathbf{x}-\mathbf{y}) - [f(\mathbf{x}) - f(\mathbf{y})] \leq \frac{\beta}{2}\|\mathbf{x}-\mathbf{y}\|_2^2.$$

Gradient descent for strongly convex functions:

- Choose number of steps T.
- For $i = 1, \ldots, T$:
    - $\eta = \frac{2}{\alpha \cdot (i+1)}$
    - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$
- Return $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}^{(i)}} f(\mathbf{x}^{(i)})$.
- Alternatively, return $\hat{\mathbf{x}} = \sum_{i=1}^{T} \frac{2i}{T(T+1)} \mathbf{x}^{(i)}$.

Theorem (GD convergence for α-strongly convex functions.)

Let f be an α-strongly convex function and assume we have that, for all $\mathbf{x}$, $\|\nabla f(\mathbf{x})\|_2 \leq G$. If we run GD for T steps (with adaptive step sizes) we have:

$$f(\hat{\mathbf{x}}) - f(\mathbf{x}^*) \leq \frac{2G^2}{\alpha(T-1)}$$

Corollary: If $T = O\left(\frac{G^2}{\alpha\epsilon}\right)$ we have $f(\hat{\mathbf{x}}) - f(\mathbf{x}^*) \leq \epsilon$

What if f is both β-smooth and α-strongly convex?

$$\frac{\alpha}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \nabla f(\mathbf{x})^T(\mathbf{x}-\mathbf{y}) - [f(\mathbf{x}) - f(\mathbf{y})] \leq \frac{\beta}{2}\|\mathbf{x}-\mathbf{y}\|_2^2.$$

What if $\alpha = \beta$:

What if f is both β-smooth and α-strongly convex?

$$\frac{\alpha}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \nabla f(\mathbf{x})^T(\mathbf{x}-\mathbf{y}) - [f(\mathbf{x}) - f(\mathbf{y})] \leq \frac{\beta}{2}\|\mathbf{x}-\mathbf{y}\|_2^2.$$

What if $\alpha = \beta$:

Theorem (GD for β-smooth, α-strongly convex.)

Let f be a β-smooth and α-strongly convex function. If we run GD for T steps (with step size $\eta = \frac{1}{\beta}$) we have:

$$\|\mathbf{x}^{(t)} - \mathbf{x}^*\|_2^2 \leq e^{-(t-1)\frac{\alpha}{\beta}} \|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2^2$$

$\kappa = \frac{\beta}{\alpha}$ is called the "condition number" of f.

Is it better if κ is large or small?

Converting to more familiar form:

$$\frac{\alpha}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \nabla f(\mathbf{x})^T(\mathbf{x}-\mathbf{y}) - [f(\mathbf{x}) - f(\mathbf{y})] \leq \frac{\beta}{2}\|\mathbf{x}-\mathbf{y}\|_2^2.$$

Corollary (GD for β-smooth, α-strongly convex.)

Let f be a β-smooth and α-strongly convex function. If we run GD for T steps (with step size $\eta = \frac{1}{\beta}$) we have:

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq \frac{\beta}{2} e^{-(t-1)\frac{\alpha}{\beta}} R$$

Corollary: If $T = O\left(\frac{\beta}{\alpha} \log(\beta R/\epsilon)\right)$ we have:

$$f(\hat{\mathbf{x}}) - f(\mathbf{x}^*) \leq \epsilon.$$

Alternative: If $T = O\left(\frac{\beta}{\alpha} \log(\beta/\alpha\epsilon)\right)$ we have:

$$f(\hat{\mathbf{x}}) - f(\mathbf{x}^*) \leq \epsilon \left[f(\mathbf{x}^{(1)}) - f(\mathbf{x}^*)\right]$$

Let $f(\mathbf{x}) = \|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2$ where $\mathbf{D}$ is a diagaonl matrix. For now imagine we're in two dimensions: $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $\mathbf{D} = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}$.

What is β for $f(\mathbf{x}) = \|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2$?

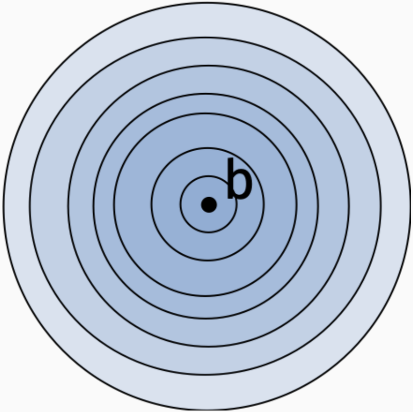
In other words: What is smallest β so that for all $\mathbf{x}, \mathbf{y}$,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq \beta\|\mathbf{x} - \mathbf{y}\|_2$$

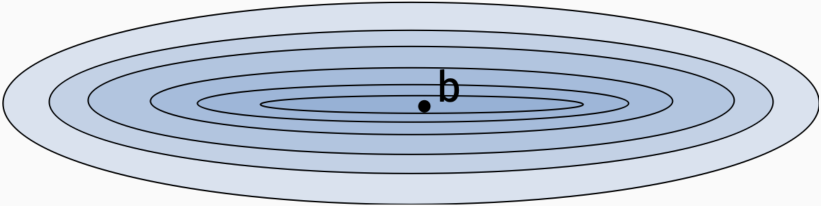
What is α for $f(\mathbf{x}) = \|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2$?

In other words: What is largest α so that for all $\mathbf{x}, \mathbf{y}$,

$$\frac{\alpha}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \leq \nabla f(\mathbf{x})^T(\mathbf{x} - \mathbf{y}) - [f(\mathbf{x}) - f(\mathbf{y})]$$



Level sets of $\|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2$ when $d_1 = 1, d_2 = 1$.



Level sets of $\|\mathbf{Dx} - \mathbf{b}\|_2^2$ when $d_1 = \frac{1}{3}, d_2 = 2$.

Steps to convergence $\approx O\left(\kappa \log(1/\epsilon)\right) = O\left(\frac{\max(\mathbf{D}^2)}{\min(\mathbf{D}^2)} \log(1/\epsilon)\right)$.

For general regression problems $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$,

$$\beta = \lambda_{max}(\mathbf{A}^T\mathbf{A})$$
$$\alpha = \lambda_{min}(\mathbf{A}^T\mathbf{A})$$

Theorem (GD for β-smooth, α-strongly convex.)

Let f be a β-smooth and α-strongly convex function. If we run GD for T steps (with step size $\eta = \frac{1}{\beta}$) we have:

$$\|\mathbf{x}^{(t)} - \mathbf{x}^*\|_2^2 \leq e^{-(t-1)\frac{\alpha}{\beta}} \|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2^2$$

Prove for $f(\mathbf{x}) = \|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2$.

IN-CLASS EXERCISE

IN-CLASS EXERCISE