Relax Quiz Midterm

CS-GY 9223 I: Lecture 7

Smooth functions

~~Preconditioning, acceleration, coordinate decent, etc.~~

Strongly functions

$$\frac{\partial f}{\partial x} = \lim_{t \to 0} \frac{f(x) - f(x+t)}{t}$$

NYU Tandon School of Engineering, Prof. Christopher Musco

$$2A^T(Ax - b) = f(x)$$

$$\lim_{z \to 0} \frac{\left[2A^TAx - A^Tb\right] - \left[2A^TA(x+z) - A^Tb\right]}{z} = 2A^TAz$$

LOGISTICS

- Self-proctored, 2-hour midterm to be taken anytime next week.
- No Collaboration allowed at all. Or outside resources. Just use your own notes and material from the class.
- Sample problems are available on course website. We can review during office hours tomorrow or next week.
- You should have received an invite to Gradescope. Hopefully tonight/tomorrow I can upload a "practice test" to make sure their system works.
- Desmos + Calculator + Coding is allowed + Wolfram Alpha

Conditions:

- **Convexity:** f is a convex function, $\mathcal{S}$ is a convex set.
- **Bounded initial distant:**

$$\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2 \leq R$$

- **Bounded gradients (Lipschitz function):**

$$\|\nabla f(\mathbf{x})\|_2 \leq G \text{ for all } \mathbf{x} \in \mathcal{S}.$$

Theorem

GD Convergence Bound] (Projected) Gradient Descent returns $\hat{\mathbf{x}}$ *with* $f(\hat{\mathbf{x}}) \leq \min_{\mathbf{x}\in\mathcal{S}} f(\mathbf{x}) + \epsilon$ *after*

$$T = \frac{R^2 G^2}{\epsilon^2} \text{ iterations.}$$

ONLINE GRADIENT DESCENT

$\mathbf{x}^* = \min_{\mathbf{x}} \sum_{i=1}^{T} f_i(\mathbf{x}^*)$ (the offline optimum)

Conditions:

- $f_1, \ldots, f_T$ are all convex.
- Each is G-Lipschitz: for all $\mathbf{x}$, i, $\|\nabla f_i(\mathbf{x})\|_2 \leq G$.
- Starting radius: $\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_2 \leq R$.

Theorem (OGD Regret Bound)

After T steps, $\left[\sum_{i=1}^{T} f_i(\mathbf{x}^{(i)})\right] - \left[\sum_{i=1}^{T} f_i(\mathbf{x}^*)\right] \leq RG\sqrt{T}$. I.e. the average regret $\frac{1}{T}\left[\sum_{i=1}^{T} f_i(\mathbf{x}^{(i)})\right]$ is $\leq \epsilon$ after:

$$T = \frac{R^2 G^2}{\epsilon^2} \text{ iterations.}$$

Conditions:

- Finite sum structure: $f(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x})$, with $f_1, \ldots, f_n$ all convex.
- Lipschitz functions: for all $\mathbf{x}$, j, $\|\nabla f_j(\mathbf{x})\|_2 \leq \frac{G'}{n}$.
- Starting radius: $\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_2 \leq R$.

Theorem (SGD Regret Bound)

Stochastic Gradient Descent returns $\hat{\mathbf{x}}$ *with* $\mathbb{E}[f(\hat{\mathbf{x}})] \leq \min_{\mathbf{x} \in \mathcal{S}} f(\mathbf{x}) + \epsilon$ *after*

$$T = \frac{R^2 G'^2}{\epsilon^2} \text{ iterations.}$$

We always have that $G' > G$, *but iterations are typically cheaper by a factor of* n.

BEYOND THE BASIC BOUNDS

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

Goals:

- Improve ϵ dependence below $1/\epsilon^2$.
  - Ideally $1/\epsilon$ or $\log(1/\epsilon)$.
- Reduce or eliminate dependence on G and R.
- Further take advantage of structure in the data (e.g. repetition in features in addition to data points).

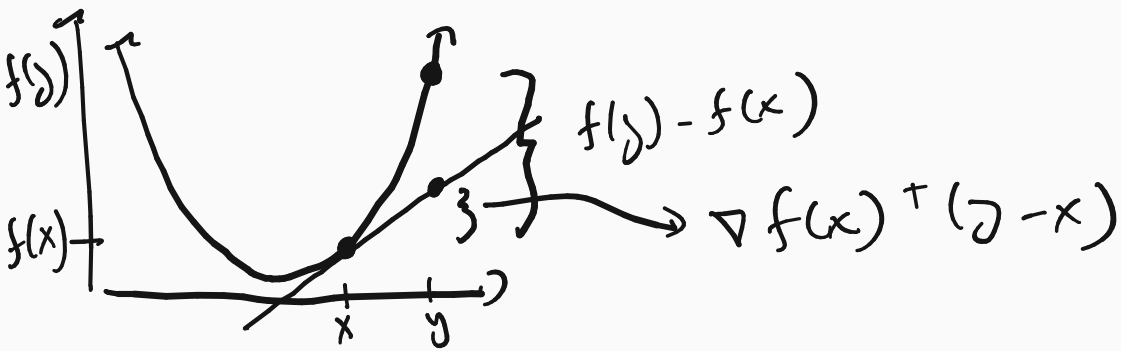
Definition (β-smoothness)

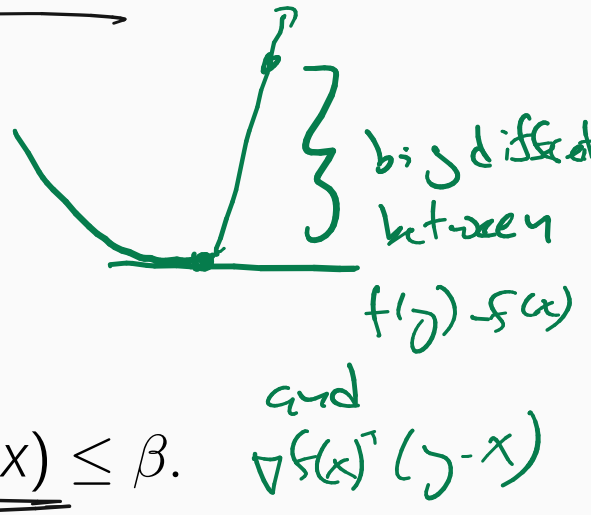
A function f is β smooth if, for all $\mathbf{x}, \mathbf{y}$

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

After some calculus (see Lem 3.4 in Bubeck's book), this implies:

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





For a scalar valued function f, equivalent to $f''(x) \leq \beta$.

Recall from definition of convexity that:

$$f(y) - f(x) \geq \nabla f(x)^T (y - x)$$

So now we have an upper and lower bound.

$$0 \leq [f(y) - f(x)] - \nabla f(x)^T (y - x) \leq \frac{\beta}{2} \|x - y\|_2^2$$

GUARANTEED PROGRESS

Previously learning rate/step size $\eta = 1/\beta$ 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: → definition of β smooth

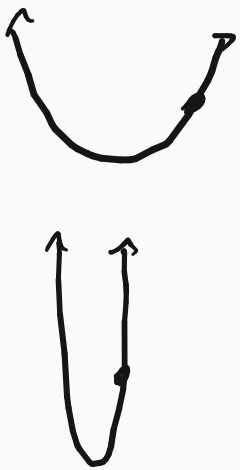
$$\left[f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^{(t)})\right] - \nabla f(\mathbf{x}^{(t)})^T(\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}) \leq \frac{\beta}{2}\|\mathbf{x}^{(t)} - \mathbf{x}^{(t+1)}\|_2^2$$

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

$$\left[f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^{(t)})\right] + \frac{1}{\beta}\|\nabla f(\mathbf{x}^{(t)})\|_2^2 \leq \frac{\beta}{2}\|\frac{1}{\beta}\nabla f(\mathbf{x}^{(t)})\|_2^2$$

$\frac{1}{2\beta}$

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^{(t+1)}) \geq \frac{1}{2\beta}\|\nabla f(\mathbf{x}^{(t)})\|_2^2$$



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:

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

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

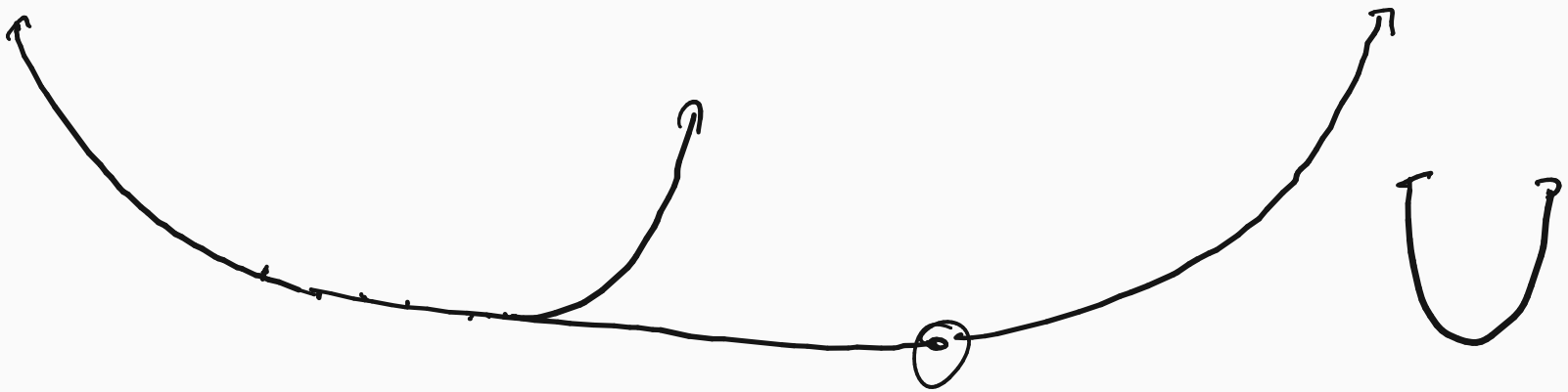
$O\left(\frac{G^2 B^2}{\epsilon^2}\right)$ $\quad B = 2$

Definition (α-strongly convex)

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

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

α is a parameter that will depend on our function.



For a twice-differentiable scalar valued function f, equivalent to $f''(x) \geq \alpha$.

GD FOR STRONGLY CONVEX FUNCTION

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 $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}^{(i)}} f(\mathbf{x}^{(i)})$.

CONVERGENCE GUARANTEE

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?

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

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

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:

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

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

Is it better if κ is large or small?

$T = O\left(\frac{\beta}{\alpha}\log(1/\epsilon)\right)$ $\quad e^{-\log(1/\epsilon)} = e^{\log(\epsilon)} = \epsilon$

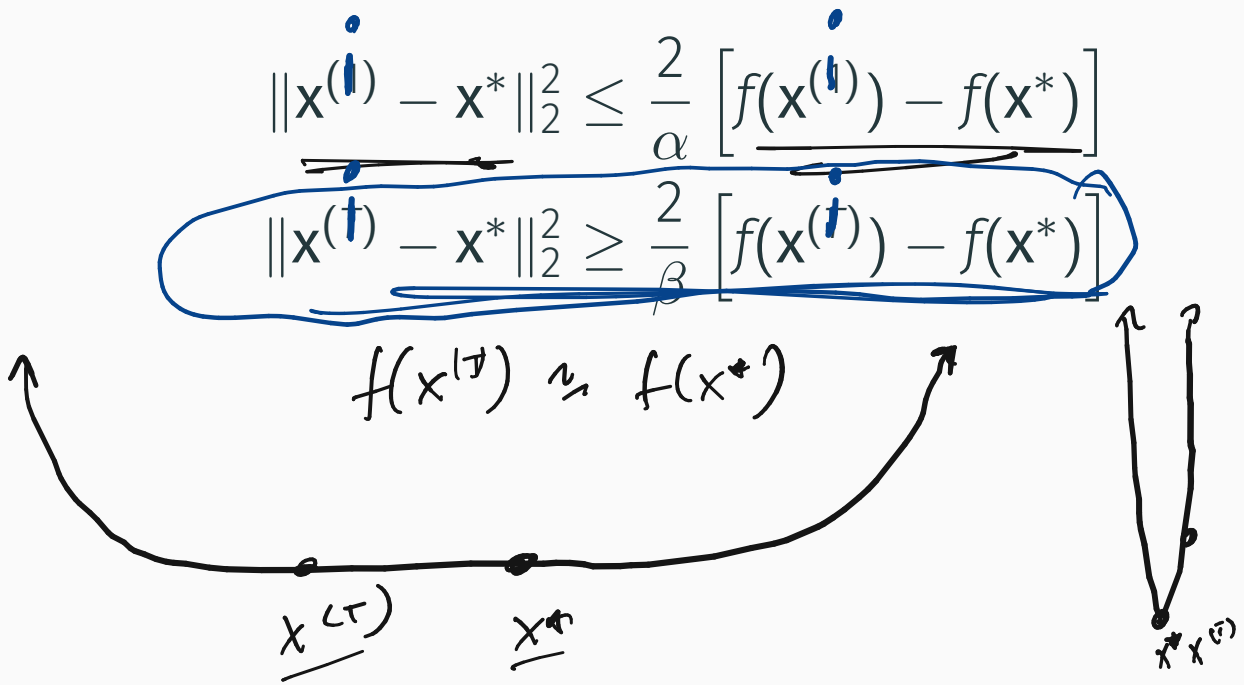
Converting to more familiar form: Using that fact the $\nabla f(\mathbf{x}^*) = \mathbf{0}$ along with

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

we have:

$$\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 \leq \frac{2}{\alpha}\left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)\right]$$

$$\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 \geq \frac{2}{\beta}\left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)\right]$$



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:

$$f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \leq \frac{\beta}{\alpha} e^{-(T-1)\frac{\alpha}{\beta}} \left[f(\mathbf{x}^{(1)}) - f(\mathbf{x}^*) \right]$$

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

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

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

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

Let f be a twice differentiable function from $\mathbb{R}^d \to \mathbb{R}$. Let the Hessian $\mathbf{H} = \nabla^2 f(\mathbf{x})$ contain all of its second derivatives at a point $\mathbf{x}$. So $\mathbf{H} \in \mathbb{R}^{d \times d}$. We have:

$$\mathbf{H}_{i,j} = \left[\nabla^2 f(\mathbf{x})\right]_{i,j} = \frac{\partial^2 f}{\partial x_i x_j}.$$

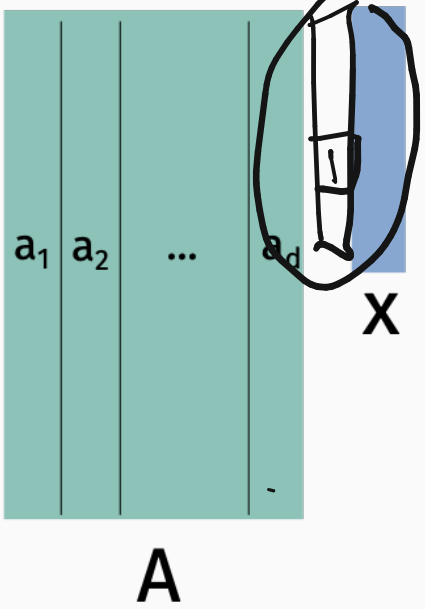
For vector $\mathbf{x}, \mathbf{y}$:

$$\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) \approx \left[\nabla^2 f(\mathbf{x})\right](\mathbf{x} - \mathbf{y}).$$

Let f be a twice differentiable function from $\mathbb{R}^d \to \mathbb{R}$. Let the Hessian $\mathbf{H} = \nabla^2 f(\mathbf{x})$ contain all of its second derivatives at a point $\mathbf{x}$. So $\mathbf{H} \in \mathbb{R}^{d \times d}$. We have:

$$\mathbf{H}_{i,j} = \left[\nabla^2 f(\mathbf{x})\right]_{i,j} = \frac{\partial^2 f}{\partial x_i x_j}.$$

Example: Let $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$. Recall that $\nabla f(\mathbf{x}) = 2\mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{b})$.



$$\nabla^2 f(x) = 2A^TA$$

$$[\nabla f(x)]_i = \frac{\partial f}{\partial x_i} = 2a_i^T(Ax - b)$$

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \lim_{t \to 0} \frac{2a_i^T(Ax-b) - 2a_i^T(A(x+te_j)-b)}{t}$$

$$\lim_{t \to 0} \frac{2a_i^T A t e_j}{t} \qquad 2a_i^T \cdot e_j$$

Claim: If f is twice differentiable, then it is convex if and only if the matrix $\mathbf{H} = \nabla^2 f(\mathbf{x})$ is positive semidefinite for all $\mathbf{x}$.

Definition (Positive Semidefinite (PSD))

A square, symmetric matrix $\mathbf{H} \in \mathbb{R}^{d \times d}$ is positive semidefinite (PSD) for any vector $\mathbf{y} \in \mathbb{R}^d$, $\mathbf{y}^T \mathbf{H} \mathbf{y} \geq 0$.

$y^T I y = y^T y = \|y\|_2^2 \geq 0$

This is a natural notion of "positivity" for symmetric matrices. To denote that $\mathbf{H}$ is PSD we will typically use "Loewner order" notation (`\succeq` in LaTex):

$$\mathbf{H} \succeq 0.$$

We write $\mathbf{B} \succeq \mathbf{A}$ or equivalently $\mathbf{A} \preceq \mathbf{B}$ to denote that $(\mathbf{B} - \mathbf{A})$ is positive semidefinite. This gives a partial ordering on matrices.

$B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ $\quad A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$

Claim: If f is twice differentiable, then it is convex if and only if the matrix $\mathbf{H} = \nabla^2 f(\mathbf{x})$ is positive semidefinite for all $\mathbf{x}$.

Definition (Positive Semidefinite (PSD))

A square, symmetric matrix $\mathbf{H} \in \mathbb{R}^{d \times d}$ is positive semidefinite (PSD) for any vector $\mathbf{y} \in \mathbb{R}^d$, $\mathbf{y}^T \mathbf{H} \mathbf{y} \geq 0$.

For the least squares regression loss function:
$f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, $\mathbf{H} = \nabla^2 f(\mathbf{x}) = 2\mathbf{A}^T\mathbf{A}$ for all $\mathbf{x}$. Is $\mathbf{H}$ PSD?

$$\mathbf{y}^T 2\mathbf{A}^T\mathbf{A}\mathbf{y} = 2\mathbf{y}^T\mathbf{A}^T\mathbf{A}\mathbf{y} = 2(\mathbf{A}\mathbf{y})^T\mathbf{A}\mathbf{y}$$
$$= 2\|\mathbf{A}\mathbf{y}\|_2^2$$
$$\geq 0$$

If f is β-smooth and α-strongly convex then at any point $\mathbf{x}$, $\mathbf{H} = \nabla^2 f(\mathbf{x})$ satisfies:

$$\alpha \mathbf{I}_{d\times d} \preceq \mathbf{H} \preceq \beta \mathbf{I}_{d\times d},$$

where $\mathbf{I}_{d\times d}$ is a $d \times d$ identity matrix.

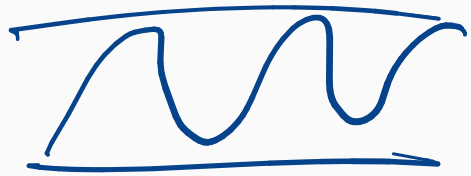
This is the natural matrix generalization of the statement for scalar valued functions:

$$\alpha \leq f''(x) \leq \beta.$$

SMOOTH AND STRONGLY CONVEX HESSIAN

$\sin(x)$ $\sin'(x) = \cos(x)$ $\sin''(x) = -\sin(x)$

$$\alpha \mathsf{I}_{d\times d} \preceq \mathsf{H} \preceq \beta \mathsf{I}_{d\times d}.$$



Equivalently for any z,

$$\alpha\|\mathsf{z}\|_2^2 \leq \mathsf{z}^T\mathsf{H}\mathsf{z} \leq \beta\|\mathsf{z}\|_2^2.$$

Exercise: Show that for $f(\mathsf{x}) = \|\mathsf{A}\mathsf{x} - \mathsf{b}\|_2^2$,

$$[f(\mathsf{y}) - f(\mathsf{x})] - \nabla f(\mathsf{x})^T(\mathsf{y} - \mathsf{x}) = (\mathsf{x} - \mathsf{y})^T \left[2\mathsf{A}^T\mathsf{A}\right](\mathsf{x} - \mathsf{y}).$$

This would imply:

$z = x - y$

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

SIMPLE EXAMPLE

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 are α, β for this problem?

$2\mathbf{D}^T\mathbf{D}$

$$\alpha\|\mathbf{z}\|_2^2 \leq \mathbf{z}^T\mathbf{H}\mathbf{z} \leq \beta\|\mathbf{z}\|_2^2$$

$\frac{\alpha}{2} = \min(d_1^2, d_2^2)$

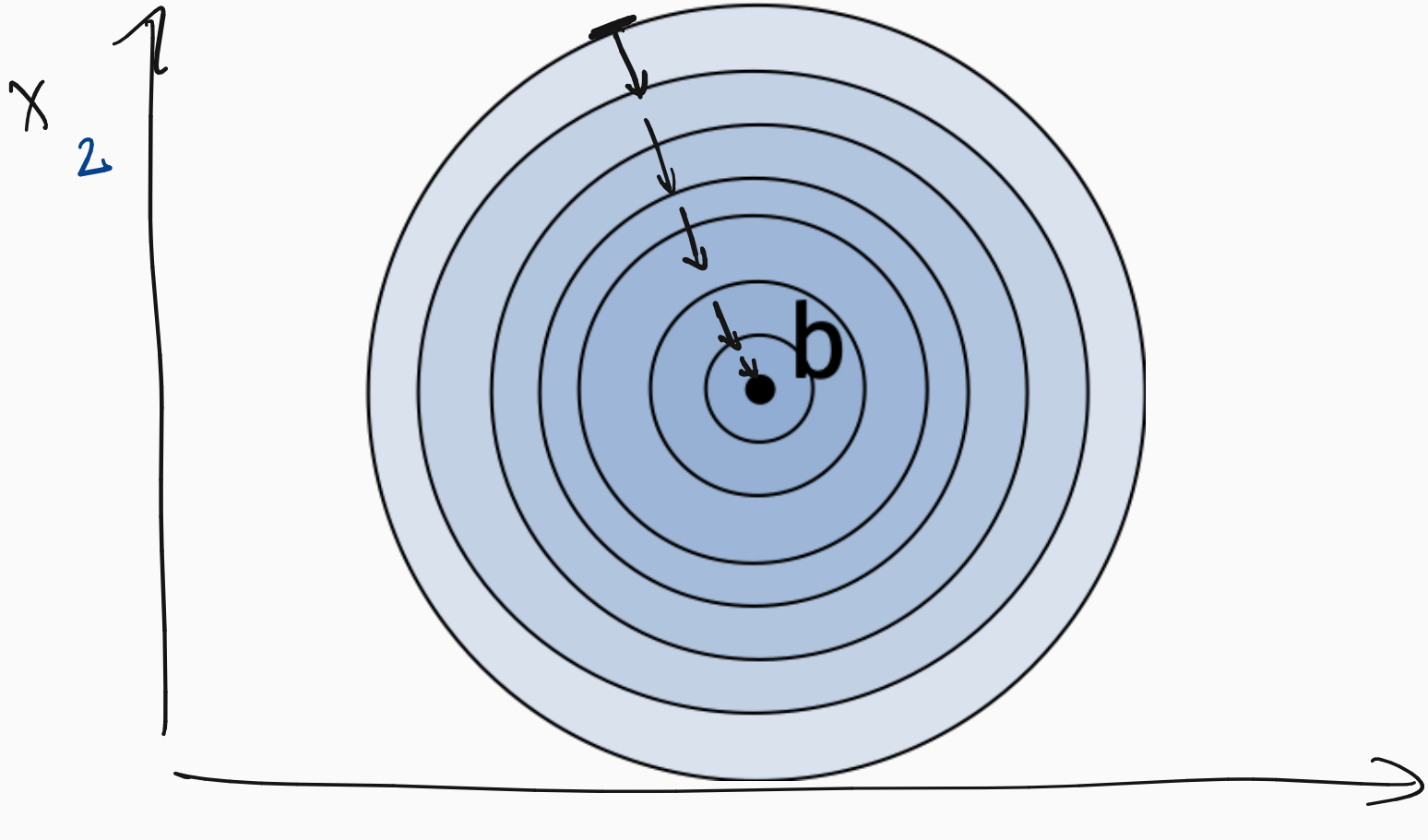
$\frac{\beta}{2} = \max(d_1^2, d_2^2)$

$z = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$z = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $\quad z^T(2\mathbf{D}^T\mathbf{D})z = 2d_2^2\|z\|_2^2$

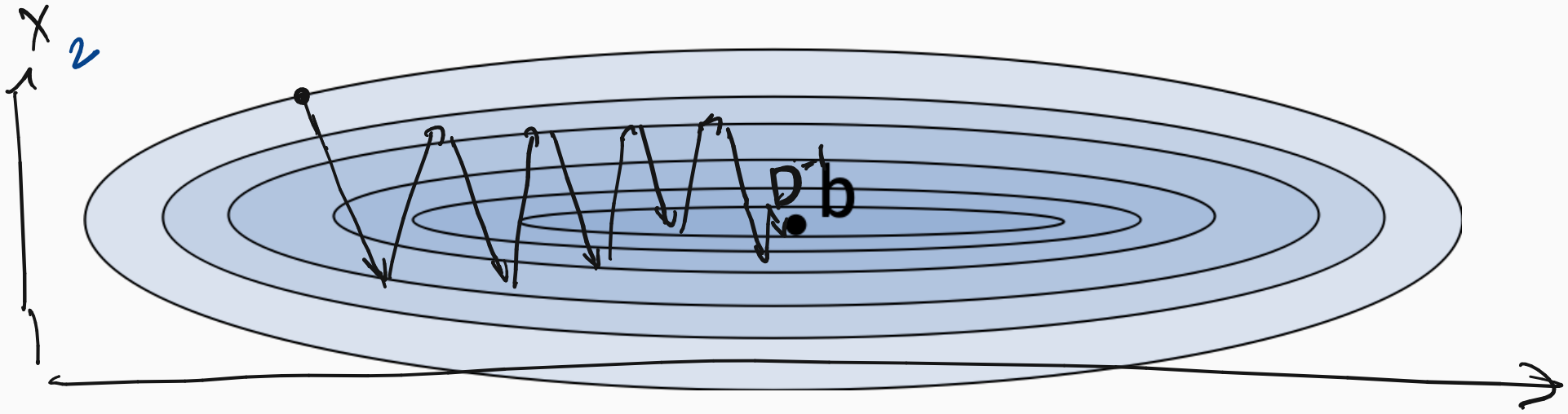
$z = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $\quad z^T(2\mathbf{D}^T\mathbf{D})z = 2d_1^2\|z\|_2^2$

GEOMETRIC VIEW



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

GEOMETRIC VIEW

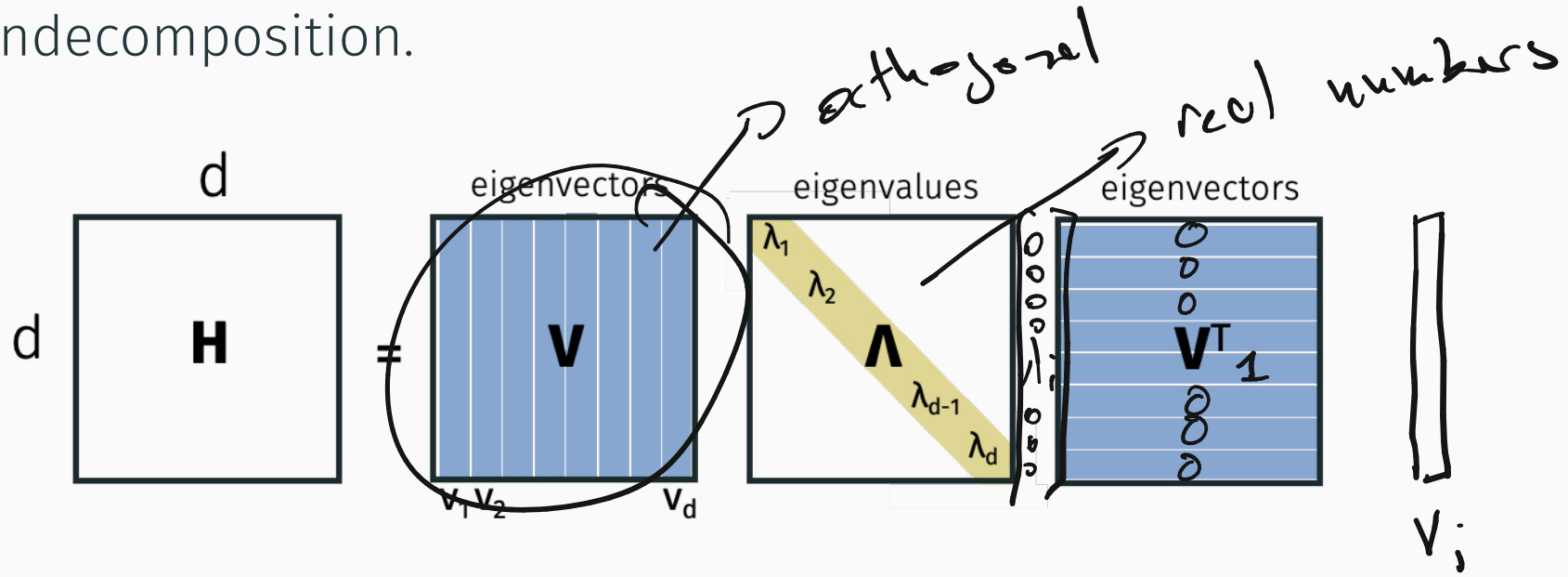


Level sets of $\|\mathbf{Dx} - \mathbf{b}\|_2^2$ when $d_1^2 = \frac{1}{3}, d_2^2 = 2$.

$\alpha = 1/3 \quad \beta = 2 \quad \kappa = \frac{2}{1/3} = 6$

EIGENDECOMPOSITION VIEW

Any symmetric matrix $\mathbf{H}$ has an orthogonal, real valued eigendecomposition.



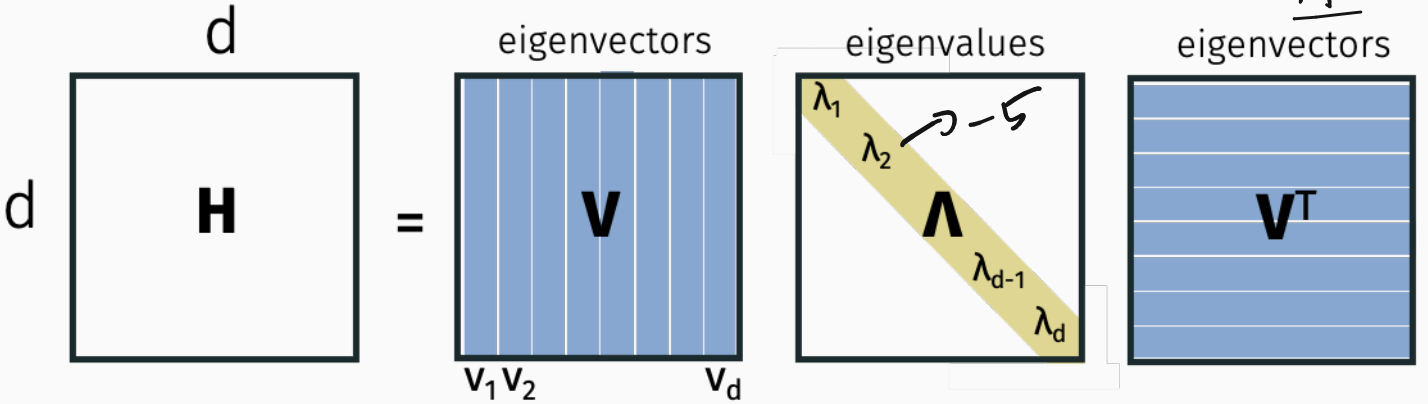
Here $\mathbf{V}$ is square and orthogonal, so $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$. And for each $\mathbf{v}_i$, we have:

$$\mathbf{H}\mathbf{v}_i = \lambda_i \mathbf{v}_i.$$

That's what makes $\mathbf{v}_1, \ldots, \mathbf{v}_d$ eigenvectors.

EIGENDECOMPOSITION VIEW

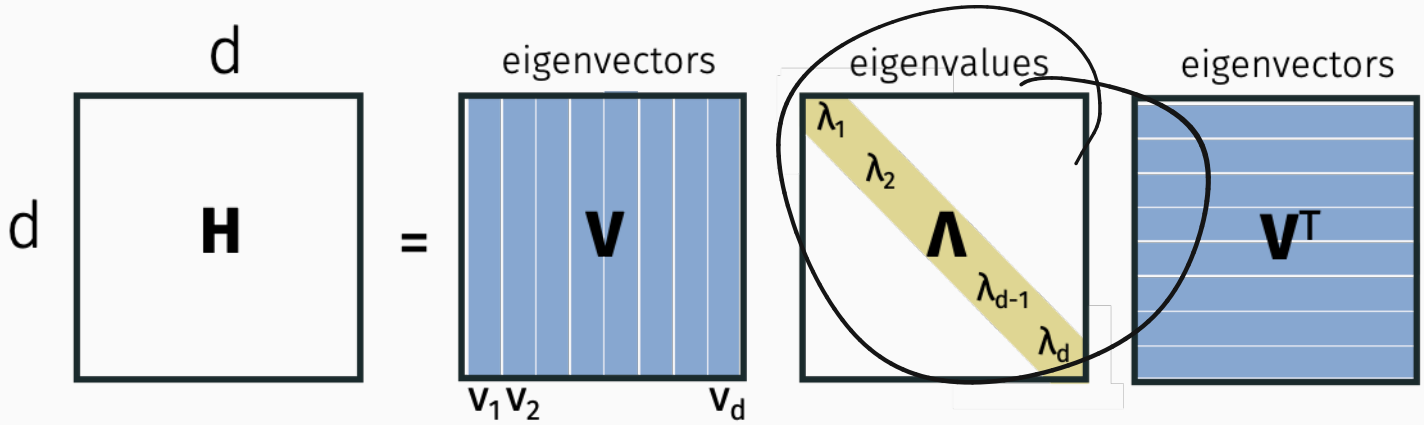
Recall $VV^T = V^TV = I$.



Claim: $H \Leftrightarrow \lambda_1, \ldots, \lambda_d \geq 0$.

EIGENDECOMPOSITION VIEW

Recall $VV^T = V^TV = I$.



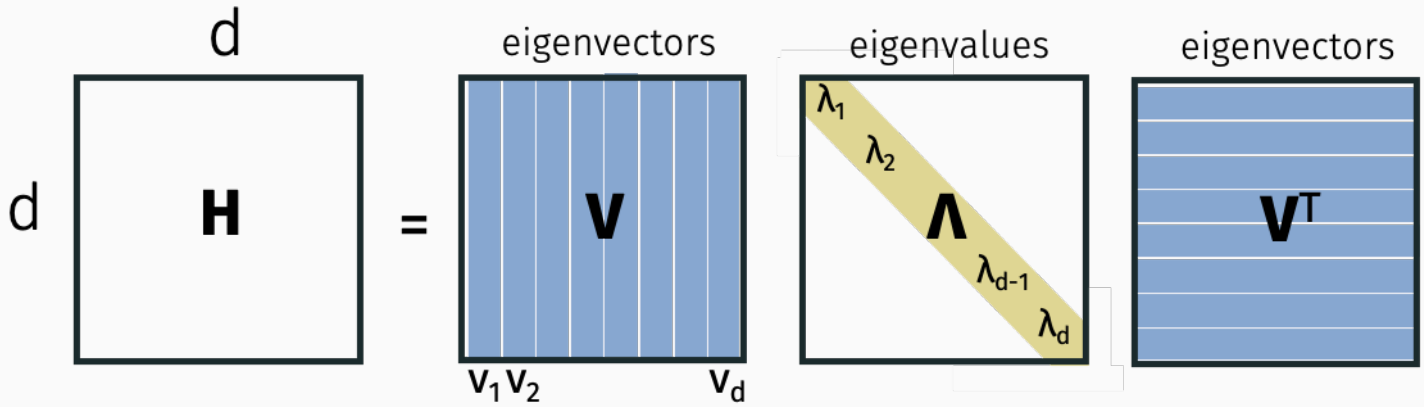
Claim: $\alpha I \preceq H \preceq \beta I \Leftrightarrow \alpha \leq \lambda_1, \ldots, \lambda_d \leq \beta$.

$H \succeq \alpha I \qquad H - \alpha I \succeq 0$

$= V \Lambda V^T - \alpha V V^T = V(\Lambda - \alpha I) V^T$

EIGENDECOMPOSITION VIEW

Recall $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}$.



In other words, if we let $\lambda_{\max}(\mathbf{H})$ and $\lambda_{\min}(\mathbf{H})$ be the smallest and largest eigenvalues of $\mathbf{H}$, then for all $\mathbf{z}$ we have:

$$\mathbf{z}^T\mathbf{H}\mathbf{z} \leq \lambda_{\max}(\mathbf{H}) \cdot \|\mathbf{z}\|^2$$

$$\mathbf{z}^T\mathbf{H}\mathbf{z} \geq \lambda_{\min}(\mathbf{H}) \cdot \|\mathbf{z}\|^2$$

EIGENDECOMPOSITION VIEW

If $f(\mathbf{x})$ is β-smooth and α-strongly convex, then for any $\mathbf{x}$ we have the the maximum eigenvalue of $\mathbf{H} = \nabla^2 f(\mathbf{x}) = \beta$ and the minimum eigenvalue of $\mathbf{H} = \nabla^2 f(\mathbf{x}) = \alpha$.

$$\lambda_{\max}(\mathbf{H}) = \beta$$
$$\lambda_{\min}(\mathbf{H}) = \alpha$$

$H = 2H^T A$

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:

$$\|x^{(T)} - x^*\|_2 \le e^{-T/\kappa} \|x^{(1)} - x^*\|_2$$

$\frac{\beta}{2} = \lambda_{max}(A^TA)$

$\frac{1}{2\lambda_{max}(A^TA)} = \frac{1}{\beta}$

$\beta = \lambda_{max}(2A^TA)$

Goal: Prove for $f(x) = \|Ax - b\|_2^2$.

$\frac{\beta}{2} = \lambda_{max}(A^TA)$

$$\left(I - \frac{1}{\lambda_{max}} A^TA\right)$$

$$VV^T - \frac{1}{\lambda_{max}} V \Lambda V^T = V\left(I - \frac{1}{\lambda_{max}}\Lambda\right)V^T$$

ALTERNATIVE VIEW OF GRADIENT DESCENT

Richardson Iteration view:

$$(x^{(T+1)} - x^*) = \left(I - \frac{1}{\lambda_{\max}} A^T A\right)(x^{(T)} - x^*)$$

$$(x^{(T+1)} - x^*) = \left(x^{(T)} - \frac{1}{2\lambda_{\max}} \cdot \nabla f(x^{(T)}) - x^*\right)$$

$$= \left(x^{(T)} - \frac{1}{\lambda_{\max}} A^T A x^{(T)} - \frac{1}{\lambda_{\max}} A^T b - x^*\right)$$

$A^T b = A^T A x^*$ $\quad \nabla(A^T A x^* - A^T b) = 0$

What is the maximum eigenvalue of the symmetric matrix $\left(I - \frac{1}{\lambda_{\max}} A^T A\right)$ in terms of the eigenvalues $\lambda_{\max} = \lambda_1 \geq \ldots \geq \lambda_d = \lambda_{\min}$ of $A^T A$?

$1 - \frac{\lambda_1}{\lambda_{\max}} = 0$ $\quad 1 - \frac{\lambda_{\min}}{\lambda_{\max}}$ $\quad 1 - \frac{\lambda_i}{\lambda_{\max}}$

UNROLLED GRADIENT DESCENT

$$(\mathbf{x}^{(T+1)} - \mathbf{x}^*) = \left(\mathbf{I} - \frac{1}{\lambda_{\max}}\mathbf{A}^T\mathbf{A}\right)^T (\mathbf{x}^{(1)} - \mathbf{x}^*)$$

number of iterations

$k = \frac{\lambda_{\max}}{\lambda_{\min}}$

What is the maximum eigenvalue of the symmetric matrix $\left(\mathbf{I} - \frac{1}{\lambda_{\max}}\mathbf{A}^T\mathbf{A}\right)^T$?

$\left(1 - \frac{\lambda_i}{\lambda_{\max}}\right)^T$ $\left(1 - \frac{\lambda_{\min}}{\lambda_{\max}}\right)^T$

$(V \Lambda V^T)(V \Lambda V^T) \cdots (V \Lambda V^T) = V \Lambda^T V^T$

$\left((1 - (1/k))^k\right)^{T/k}$

$1/e$

So we have $\|\mathbf{x}^{(T)} - \mathbf{x}^*\|_2 \leq$

$(1/e)^{T/k} = e^{-T/k}$

IMPROVING GRADIENT DESCENT

We now have a really good understanding of gradient descent.

Number of iterations for ϵ error:

	G-Lipschitz	β-smooth
R bounded start	$O\left(\frac{G^2R^2}{\epsilon^2}\right)$	$O\left(\frac{\beta R^2}{\epsilon}\right)$
α-strong convex	$O\left(\frac{G^2}{\alpha\epsilon}\right)$	$O\left(\frac{\beta}{\alpha}\log(1/\epsilon)\right)$

How do we use this understanding to design faster algorithms?

Often it doesn't make sense to sample i uniformly at random:

$$\mathsf{A} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -.5 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & 0 \end{bmatrix} \qquad \mathsf{b} = \begin{bmatrix} 10 \\ 42 \\ -11 \\ -51 \\ 34 \\ -22 \end{bmatrix}$$

Select indices i proportional to $\|\mathsf{a}_i\|_2^2$:

$$\Pr[\text{select index } i \text{ to update}] = \frac{\|\mathsf{a}_i\|_2^2}{\sum_{j=1}^{d} \|\mathsf{a}_j\|_2^2} = \frac{\|\mathsf{a}_i\|_2^2}{\|\mathsf{A}\|_2^2}$$