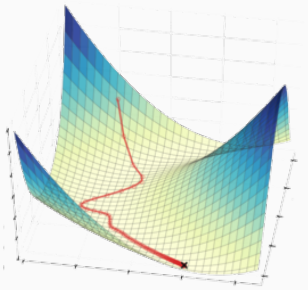
CS-GY 6763: LECTURE 6
GRADIENT DESCENT AND PROJECTED GRADIENT DESCENT

NYU Tandon School of Engineering, Prof. Rajesh Jayaram

GRADIENT DESCENT

Gradient descent: A greedy algorithm for minimizing functions of multiple variables that often works amazingly well.



(and sometimes we can prove it works)

GRADIENT DESCENT ANALYSIS

Assume:

- f is convex.
- Lipschitz function: for all $\mathbf{x}$, $\|\nabla f(\mathbf{x})\|_2 \leq G$.
- Starting radius: $\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_2 \leq R$.

Gradient descent:

- Choose number of steps T.
- Starting point $\mathbf{x}^{(0)}$. E.g. $\mathbf{x}^{(0)} = \vec{0}$.
- $\eta = \frac{R}{G\sqrt{T}}$
- For $i = 0, \ldots, T$:
  - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$
- Return $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}^{(i)}} f(\mathbf{x}^{(i)})$.

Theorem (GD Convergence Bound)

If we run gradient descent for at least $T \geq \frac{R^2 G^2}{\epsilon^2}$ *iterations, with step size* $\eta = \frac{R}{G\sqrt{T}}$, *then*

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

BEYOND THE BASIC BOUND

Can our convergence bound be tightened for certain functions?
Can it 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.

SMOOTHNESS

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

CONVERGENCE GUARANTEE

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

Complete proof in Theorem 3.5 of **Bubeck's book**

STRONG CONVEXITY

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 $f''(x) \geq \alpha$.

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

SMOOTH AND STRONGLY CONVEX

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

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

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

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

Is it better if κ is large or small?

THE LINEAR ALGEBRA OF CONDITIONING

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:

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

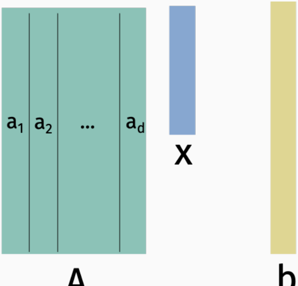
For vector $\mathbf{x}, \mathbf{v}$:

$$\nabla f(\mathbf{x} + t\mathbf{v}) \approx \nabla f(\mathbf{x}) + t\left[\nabla^2 f(\mathbf{x})\right]\mathbf{v}.$$

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 $\nabla f(\mathbf{x}) = 2\mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{b})$.



HESSIAN MATRICES AND POSITIVE SEMIDEFINITENESS

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 $\mathbf{y} \in \mathbb{R}^d$, $\mathbf{y}^T \mathbf{H} \mathbf{y} \geq 0$.

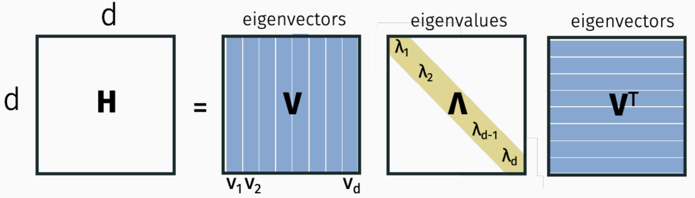
This is a natural notion of "positivity" for symmetric matrices. To denote that $\mathbf{H}$ is PSD we will typically use "Loewner order" notation (`\succeq` in LaTex):

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

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.

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:

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

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

FACTS ABOUT PSD MATRICES

Theorem

Let $\mathbf{H} \in \mathbb{R}^{n \times n}$ *be a symmetric matrix. Then* $\mathbf{H}$ *is PSD if and only if* $\lambda_i(\mathbf{H}) \geq 0$ *for all its eigenvalues* $\lambda_i(\mathbf{H})$ *with* $i = 1, 2, \ldots, n$.

FACTS ABOUT PSD MATRICES

Theorem

Let $\mathbf{H} \in \mathbb{R}^{n \times n}$ *be a symmetric matrix. Then* $\mathbf{H}$ *is PSD if and only if* $\mathbf{H} = \mathbf{V}^T\mathbf{V}$ *for some matrix* $\mathbf{V} \in \mathbb{R}^{n \times n}$.

HESSIAN MATRICES AND POSITIVE SEMIDEFINITENESS

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 $\mathbf{x}$.

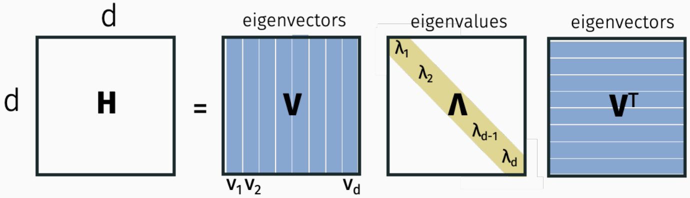
Definition (Positive Semidefinite (PSD))

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?

EIGENDECOMPOSITION VIEW

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



Claim: $\alpha\mathbf{I} \preceq \mathbf{H} \preceq \beta\mathbf{I} \Leftrightarrow \alpha \leq \lambda_1, ..., \lambda_d \leq \beta$.

THE LINEAR ALGEBRA OF CONDITIONING

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

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

Equivalently for any $\mathbf{z}$,

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

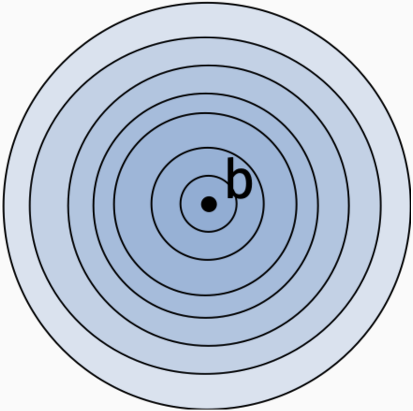
SIMPLE EXAMPLE

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?

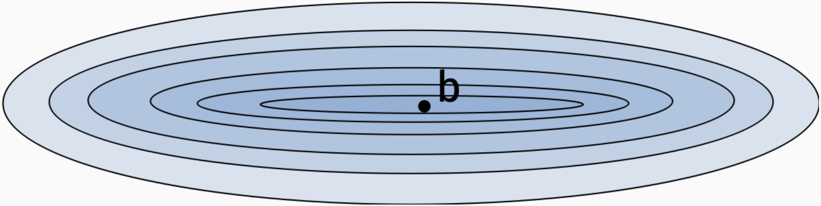
$$\alpha\|\mathbf{z}\|_2^2 \leq \mathbf{z}^T\mathbf{H}\mathbf{z} \leq \beta\|\mathbf{z}\|_2^2$$

GEOMETRIC VIEW



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

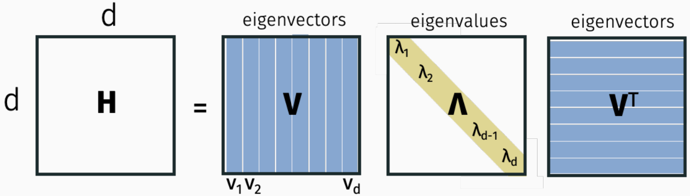
GEOMETRIC VIEW



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

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

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

PRECONDITIONING FOR LEAST-SQUARES REGRESSION

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

Let $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$*, where* $\alpha\mathbf{I} \leq 2\mathbf{A}\mathbf{A}^T \leq \beta\mathbf{I}$*. Then* f *is* β*-smooth and* α*-strongly , and if we run GD for* $T = O\left(\frac{\beta}{\alpha}\log(\frac{R\beta}{\epsilon})\right)$ *steps (with step size* $\eta = \frac{1}{\beta}$*) we have:*

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

PRECONDITIONING FOR LEAST-SQUARES REGRESSION

Given $\mathbf{A} \in \mathbb{R}^{n \times d}$, let $\mathbf{V} \in \mathbb{R}^{n \times d}$ be any matrix with the same *column span* as $\mathbf{A}$. Then

$$\min_x \|\mathbf{A}x - b\|_2 = \min_x \|\mathbf{V}x - b\|_2$$

SINGULAR VALUE DECOMPOSITION

Quick reminder of the SVD:

Theorem (SVD)

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be a rank $r \leq \min\{n, d\}$ matrix. Then $\mathbf{A}$ can be decomposed into $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where the columns of $\mathbf{U} \in \mathbb{R}^{n \times r}$ are the left singular vectors of $\mathbf{A}$, the rows of $\mathbf{V}^T$ are the right singular vectors of $\mathbf{A}$, and $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ is a diagonal matrix with $\mathbf{\Sigma}_{i,i} = \sigma_i$ is the i-th singular value of $\mathbf{\Sigma}$

- Recall, the singular values $\{\sigma_i\}_{i \in [r]}$ are the square roots of the eigenvalues of $\mathbf{D} = \mathbf{A}^T\mathbf{A}^T$, i.e. $\{\lambda_i(\mathbf{D})\}_{i \in [r]}$.
- Note that $\mathbf{U}^T\mathbf{U} = \mathbf{I}_r = \mathbf{V}\mathbf{V}^T$, since $\mathbf{U}$ and $\mathbf{V}$ have orthogonal columns.

SINGULAR VALUE DECOMPOSITION

Quick reminder of the SVD:

Theorem (SVD)

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be a rank $r \leq \min\{n, d\}$ matrix. Then $\mathbf{A}$ can be decomposed into $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where the columns of $\mathbf{U} \in \mathbb{R}^{n \times r}$ are the left singular vectors of $\mathbf{A}$, the rows of $\mathbf{V}^T$ are the right singular vectors of $\mathbf{A}$, and $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ is a diagonal matrix with $\mathbf{\Sigma}_{i,i} = \sigma_i$ is the i-th singular value of $\mathbf{\Sigma}$

- Recall, the singular values $\{\sigma_i\}_{i \in [r]}$ are the square roots of the eigenvalues of $\mathbf{D} = \mathbf{A}^T\mathbf{A}^T$, i.e. $\{\lambda_i(\mathbf{D})\}_{i \in [r]}$.
- Note that $\mathbf{U}^T\mathbf{U} = \mathbf{I}_r = \mathbf{V}\mathbf{V}^T$, since $\mathbf{U}$ and $\mathbf{V}$ have orthogonal columns.
- Can compute the SVD in $O(\min\{nd^2, dn^2\})$ time.

PRECONDITIONING FOR LEAST-SQUARES REGRESSION

Given $\mathbf{A} \in \mathbb{R}^{n \times d}$, let $\mathbf{V} \in \mathbb{R}^{n \times d}$ be any matrix with the same *column span* as $\mathbf{A}$. Then

$$\min_x \|\mathbf{A}x - b\|_2 = \min_x \|\mathbf{V}x - b\|_2$$

Can choose $\mathbf{V}$ to be a *well-conditioned matrix* which spans the columns of $\mathbf{A}$

- Can choose $\mathbf{V} \in \mathbb{R}^{n \times d}$ to be the left singular vectors of $\mathbf{A}$.
- Singular vectors are orthogonal, so $\mathbf{V}^T\mathbf{V} = \mathbf{I}_d$.
- $\nabla^2 f(x) = 2\mathbf{V}^T\mathbf{V} = 2\mathbf{I}_d$, thus $\kappa = 1$!