ORIE 4741/5741
Learning with Big Messy Data

Instructor: Haiyun He
TA:

Discussion #3: More Linear Algebra + Gradient Descent
March 6th, 2023

1 Notations

(a) Capital letters, e.g. A, X: Matrices.
(b) Small letters, x, w: Column vectors.
(c) Small letters with subscripts, e.g. x_i, w_j, a_{ij}: Entries of vectors or matrices.
(d) Letters with superscript "*", e.g. p^*, f^*, x^*: Function or independent variable values at the optimal point.

2 Basic Linear Algebra

In the following, vectors (bold, lower-case) have length d, square matrices have size $d \times d$, and other matrices have arbitrary dimensions as long as the sizes in the multiplications are consistent.

- Inner product: $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v} = \mathbf{v}^T \mathbf{u} = \sum_{i=1}^d u_i v_i$ (here and below, $(\cdot)^T$ means transpose)
  - We say that $\mathbf{u}$ and $\mathbf{v}$ are orthogonal if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$ (Visually: At right angles from each other)
  - More generally, $\langle \mathbf{u}, \mathbf{v} \rangle = \|\mathbf{u}\| \cdot \|\mathbf{v}\| \cdot \cos(\text{angle}(\mathbf{u}, \mathbf{v}))$ (maximized when they point in the same direction, as negative as possible when they point in opposite directions)
  - Cauchy-Schwartz inequality: $|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \cdot \|\mathbf{v}\|$ with equality if $\mathbf{u} = c\mathbf{v}$ for some constant c.
- (Euclidean) norm: $\|\mathbf{u}\| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle} = \sqrt{\sum_{i=1}^d u_i^2}$
  - Triangle inequality: $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$
  - Expansion: $\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + 2\langle \mathbf{u}, \mathbf{v} \rangle + \|\mathbf{v}\|^2$
- Transpose operations:
  - $(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$
  - $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$ (Note: The analogous property certainly doesn't hold in general for the matrix inverse, defined below)
- The identity matrix is a square matrix with 1's on the diagonal, and 0's elsewhere. Then $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$.
- Matrix inverse (of a $d \times d$ square matrix): $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_d$, where $\mathbf{I}_d$ is the $d \times d$ identity matrix
  - The inverse exists only for *invertible* matrices
  - $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
- Trace (of a $d \times d$ square matrix): $\text{Tr}[\mathbf{A}] = \sum_{i=1}^d A_{ii}$ (i.e., sum of diagonal entries)
  - Useful property: $\text{Tr}[\mathbf{AB}] = \text{Tr}[\mathbf{BA}]$
  - Special case (let $\mathbf{A} = \mathbf{u}$ and $\mathbf{B} = \mathbf{u}^T$): $\|\mathbf{u}\|^2 = \text{Tr}[\mathbf{u}\mathbf{u}^T]$
- A $d \times d$ square matrix $\mathbf{A}$ is said to be positive semidefinite if $\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^d$

  - It is positive definite if the only case that equality holds (i.e., $\mathbf{x}^T\mathbf{A}\mathbf{x} = 0$) is the case $\mathbf{x} = \mathbf{0}$.
  - Being positive semidefinite is equivalent to all eigenvalues being non-negative (you will not be assessed on questions involving eigenvalues/eigenvectors, but these concepts are also useful). Being positive definite is equivalent to all eigenvalues being positive.
  - Positive definite matrices are always invertible (but positive semidefinite matrices may not be, e.g., consider the all-zero matrix)
  - The positive semidefinite (respectively, positive definite) property implies that $\det(\mathbf{A}) \geq 0$ (respectively, $\det(\mathbf{A}) > 0$), where $\det(\cdot)$ is the determinant. However, matrices with a positive determinant still might fail to be positive semidefinite (e.g., $\mathbf{A} = -\mathbf{I}$ when $d = 2$).
- The general definition of determinant will not be needed in this course, but you should at least know the formula for the 2×2 case: $\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = ad - bc$.

3 Some Linear Algebra Conclusions

3.1 $(Ax = 0 \Leftrightarrow x = 0) \overset{\text{(a)}}{\Longleftrightarrow} A$ has full column rank $\overset{\text{(b)}}{\Longleftrightarrow} A^TA$ is invertible

This is the prerequisite for pseudoinverse. Namely, if a matrix A has full column rank, then we can explicitly write out its pseudoinverse that contains $(A^TA)^{-1}$.

We will first show the correctness of 2.1(a).

Proof. Let A be an $m \times n$ matrix. Write A as the concatenation of column vectors $(a_1, a_2, \cdots, a_n)$. $Ax = 0$ can then be written as $\sum_{i=1}^{n} a_i x_i = 0$. Thus $(Ax = 0 \Leftrightarrow x = 0)$ is equivalent to the columns of A being linearly independent, i.e. A has full column rank. □

For 2.1(b):

Proof. A^TA is an $n \times n$ symmetric matrix, thus A^TA being invertible is equivalent to A^TA having full column rank. From 2.1(a), we know this is equivalent to $(A^TAx = 0 \Leftrightarrow x = 0)$. From this we can prove 2.1(b) by proving a stronger equivalence $Ax = 0 \overset{\text{(c)}}{\Longleftrightarrow} A^TAx = 0$, i.e. these two equations have the same solution space. Thus we are now going to prove equivalence (c).

It is evident that $Ax = 0 \Rightarrow A^TAx = 0$. As for the other direction, we left multiply x^T on both sides of $A^TAx = 0$ to get $x^TA^TAx = 0$.

Note this is equivalent to $(Ax)^TAx = 0$, namely $||Ax||_2^2 = 0$. From the norm property that $||x|| = 0 \Leftrightarrow x = 0$, we get $Ax = 0$, which completes the proof. □

3.2 $y \in \mathbf{range}(X)$, then $XX^\dagger y = y$

Recall that when an $n \times d$ matrix X has full column rank, then X's pseudoinverse $X^\dagger = (X^TX)^{-1}X^T$. Thus we have its properties

(a) $X^\dagger X = I_d$

(b) $XX^\dagger \neq I_n$ (Here "$\neq$" means not always equal)
Now we are going to prove 2.2.

Proof. Range$(X) \triangleq \{Xv|v \in \mathbb{R}^d\}$. Thus $y \in$ range(X) means $\exists z \in \mathbb{R}^d$, s.t. $Xz = y$. We left multiply $XX^\dagger$ to both sides and get $XX^\dagger Xz = XX^\dagger y$. Using property (a), the left-hand side equals Xz, which equals y, thus we get $XX^\dagger y = y$. □

Now we discuss the intuitive understanding of this equality.

Matrices can be regarded as linear operators, which can map one vector to the other through the matrix-vector multiplication. Property (a) says the effect of first imposing operator X and then $X^\dagger$ on **any** vector is equivalent to the effect of an identity map. While given property (b), we know first imposing $X^\dagger$ and then X on a vector may not get the original vector, which corresponds to this "inverse" being "pseudo"; however, from 2.2, we know we can return to the original vector when this vector is in the range (or column space) of X.

3.3 $\frac{\partial(w^T v)}{\partial w}$ and $\frac{\partial(w^T A w)}{\partial w}$

When performing gradient descent to the least squares problem, we need to calculate $\nabla_w ||y - Xw||^2 = \nabla_w (y^T y - y^T Xw - w^T X^T y + w^T X^T Xw)$. The partial derivatives in the above forms occur in this gradient.

Recall the definition of gradient. Let $f(x_1, x_2, ..., x_n)$ be a multivariate scalar function. The gradient of f, ∇f, is the multivariate generalization of the derivative of f. The gradient is a vector, where each entry corresponds to a partial derivative with respect to a variable of the function.

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

Let $w, v \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$ in the following context.

3.3.1 $\frac{\partial(w^T v)}{\partial w}$

We explicitly write out the summation for calculating $w^T v$: $w^T v = \sum_{i=1}^n w_i v_i$.

Now consider the kth ($k \in \{1, 2, \cdots, n\}$) component in gradient, namely $\frac{\partial(\sum_{i=1}^n w_i v_i)}{\partial w_k}$. Only the term with $i = k$ contributes to this partial derivative. Thus

$$\frac{\partial(\sum_{i=1}^n w_i v_i)}{\partial w_k} = \frac{\partial(w_k v_k)}{\partial w_k} = v_k$$

Then, concatenating the results of the partial derivatives on different k, we get

$$\frac{\partial(w^T v)}{\partial w} = \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_n \end{bmatrix} = v$$

3.3.2 $\frac{\partial(w^T A w)}{\partial w}$

Similarly, we consider the kth component in gradient, namely $\frac{\partial(\sum_{i=1}^n \sum_{j=1}^n w_i a_{ij} w_j)}{\partial w_k}$. Only the terms with $i = k$ or $j = k$ will contribute to this partial derivative. Thus

$$\begin{aligned} \frac{\partial(\sum_{i=1}^n \sum_{j=1}^n w_i a_{ij} w_j)}{\partial w_k} &= \underbrace{\frac{\partial(\sum_{j=1,j\neq k}^n w_k a_{kj} w_j)}{\partial w_k}}_{i=k,j\neq k} + \underbrace{\frac{\partial(\sum_{i=1,i\neq k}^n w_i a_{ik} w_k)}{\partial w_k}}_{j=k,i\neq k} + \underbrace{\frac{\partial a_{kk} w_k^2}{\partial w_k}}_{i=j=k} \\ &= \sum_{j=1,j\neq k}^n a_{kj} w_j + \sum_{i=1,i\neq k}^n w_i a_{ik} + 2a_{kk} w_k \\ &= \sum_{j=1}^n a_{kj} w_j + \sum_{i=1}^n w_i a_{ik} \end{aligned} \tag{1}$$

Denote the ith row and jth column of matrix A as $A_{i:}$ and $A_{:j}$, respectively. Let $B = A^T$. Thus

$$\sum_{j=1}^n a_{kj} w_j = A_{k:} w$$

$$\sum_{i=1}^n w_i a_{ik} = w^T A_{:k} = B_{k:} w$$

Finally, concatenating the results of the partial derivatives on different k, we get

$$\frac{\partial(w^T A w)}{\partial w} = \begin{bmatrix} A_{1:} + B_{1:} \\ A_{2:} + B_{2:} \\ \cdot \\ \cdot \\ \cdot \\ A_{n:} + B_{n:} \end{bmatrix} w = (A + B) w = (A + A^T) w$$

which completes the calculation.

More conclusions regarding derivatives with respect to matrices can be found in *The Matrix Cookbook*[**petersen2008matrix**].

4 Other Basics

- **Basic properties of logarithms and exponentials:**
  - $\log(xy) = \log x + \log y$
  - $\log \frac{1}{x} = -\log x$
  - $\log \frac{y}{x} = \log y - \log x$
  - $\log x^c = c \log x$
  - $\log_a x = \frac{\log_b x}{\log_b a}$
  - $e^{a+b} = e^a e^b$
  - $(e^a)^b = e^{ab}$
  - $e^{\log x} = \log e^x = x$ (for log having base e)
- **Basic calculus:**
  - $\frac{d}{dx} x^c = cx^{c-1}$
  - $\frac{d}{dx} e^{cx} = ce^{cx}$
  - $\frac{d}{dx} \log x = \frac{1}{x}$ (Note: Assume $\log(\cdot)$ has base e except where indicated otherwise)
  - Chain rule: $\frac{d}{dx} f(g(x)) = f'(g(x))\, g'(x)$
  - Product rule: $\frac{d}{dx}(f(x)g(x)) = f'(x)g(x) + f(x)g'(x)$
  - Quotient rule: $\frac{d}{dx}(f(x)/g(x)) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$

5 The Least Squares Problem

5.1 Convexity

For more details on this part, refer to Chapter 3 of *Convex Optimization* by Boyd & Vandenberghe [**boyd2004convex**].

Definition 1. A function $f : \mathbb{R}^n \to \mathbb{R}$ being convex if the domain of f (denoted as $\mathbf{dom}(f)$) is a convex set and $\forall x, y \in \mathbf{dom}(f)$ and $\theta \in [0, 1]$, $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$.

This is also called the Jensen's Inequality, or the zeroth-order convexity condition.

Equivalently, the first-order convexity condition is:

Theorem 1. *(First-order Convexity Condition) Suppose a function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable in $\mathbf{dom}(f)$. Then f is convex if and only if $\mathbf{dom}(f)$ is convex and $\forall x, y \in \mathbf{dom}(f)$, $f(y) \geq f(x) + \nabla f(x)^T (y - x)$.*

And we also have the second-order convexity condition:

Theorem 2. *(Second-order Convexity Condition) Suppose a function $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable in $\mathbf{dom}(f)$. Then f is convex if and only if $\forall x \in \mathbf{dom}(f)$, $\nabla^2 f \succeq 0$ (positive semi-definite).*

5.2 Convergence Rate of Gradient Descent on Smooth Functions [unconstrained]

If we perform gradient descent on a function f which is convex and "smooth" (i.e. its gradient does not change too fast), and make the step size to be not too large, then we can ensure the gap between function value f at step k and the global minimum p^* to be upper bounded by a value

which is inversely proportional to k.

Formally, we make the following assumptions:
(a) $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable with $\mathbf{dom}(f) = \mathbb{R}^n$
(b) f is smooth[1] with parameter $\beta > 0$
(c) The optimal value $p^* = \inf_x f(x)$ is finite and is attained at x^*.
Then if we pick a constant step size t s.t. $0 < t \leq \frac{1}{\beta}$ and perform updates $x^+ = x - t\nabla f(x^{(k)})$ at each step, we have

$$\begin{aligned} f(x^+) &\leq f(x) + \nabla f(x)(x^+ - x) + \frac{\beta}{2}||x^+ - x||^2 \\ &= f(x) - t||\nabla f(x)||^2 + t^2\frac{\beta}{2}||\nabla f(x)||^2 \end{aligned} \tag{2}$$

When $0 < t \leq \frac{1}{\beta}$, we have

$$\begin{aligned} f(x^+) &\leq f(x) - \frac{t}{2}||\nabla f(x)||^2 \\ &\leq f(x^*) + \nabla f(x)^T(x - x^*) - \frac{t}{2}||\nabla f(x)||^2 (\textit{using first-order convexity condition}) \\ &= p^* + \frac{1}{2t}(||x - x^*||^2 - ||x - x^* - t\nabla f(x)||^2) \\ &= p^* + \frac{1}{2t}(||x - x^*||^2 - ||x^+ - x^*||^2) \end{aligned} \tag{3}$$

Take average of the above inequality over iterations $1, 2, \cdots, k$:

$$\begin{aligned} \frac{1}{k}\sum_{i=1}^{k} f(x^{(i)}) - p^* &\leq \frac{1}{k}\sum_{i=1}^{k}\frac{1}{2t}(||x^{(i)} - x^*||^2 - ||x^{(i+1)} - x^*||^2) \\ &\leq \frac{1}{2tk}(||x^{(0)} - x^*||^2 - ||x^{(k+1)} - x^*||^2) \\ &\leq \frac{1}{2tk}||x^{(0)} - x^*||^2 \end{aligned} \tag{4}$$

Since $f(x^{(k)})$ is non-increasing over iterations, we have

$$f(x^{(k)}) - p^* \leq \frac{1}{2tk}||x^{(0)} - x^*||^2 \tag{5}$$

which shows the number of iterations k taken to reach $f(x^{(k)}) - p^* \leq \epsilon$ is $O(\frac{1}{\epsilon})$.

5.3 Properties of the Least Squares Problem

We have $\nabla_w||y - Xw||^2 = 2X^T(Xw - y)$, and then $\nabla_w^2||y - Xw||^2 = 2X^TX$, which is positive semi-definite. Thus from the second-order convexity condition, we can also show the convexity of the least-squares problem, which is a substitute of the use of first-order condition that was discussed in class. From this we can show using the first-order condition that the least squares problem only

[1] A function f is smooth if and only if $\forall x, y \in \mathbf{dom}(f)$, $f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{\beta}{2}||x-y||^2$. This is equivalent to ∇f being Lipschitz continuous with parameter β, i.e. $\forall x, y \in \mathbf{dom}(f)$, $||\nabla f(x) - \nabla f(y)|| \leq \beta||x - y||$.

has one minimum, which is a global minimum.

As for smoothness (or Lipschitz continuity), we have

$$||\nabla_w||y - Xw_1||^2 - \nabla_w||y - Xw_2||^2|| \leq 2||X^T X||||w_1 - w_2||$$

Thus if we limit the step size t to be $0 \leq t \leq \frac{1}{2||X^T X||}$, we can get a convergence rate of $O(\frac{1}{k})$ with respect to the number of steps k.