Stat 6341.001: Numerical Linear Algebra and Statistical Computing, Fall 2026 **Topic 1**

Numerical Linear Algebra and Statistical Computing

Lecturer: Yun Wei *Scribe: Students in Fall 2026*

Disclaimer: ***These notes only contains the material of the lectures that use white board instead of slides.*** *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Contents

1 Multi-Armed Bandit [2] **5**

1.1 Introduction . . . 5

1.1.1 Multi-Armed Bandit Protocol . . . 5

1.1.2 Simple Example: Always Pure Exploitation . . . 6

1.1.3 Estimating f^* via Empirical Means . . . 6

1.1.4 The Greedy Algorithm and Exploration . . . 6

1.1.5 Failure of the Greedy Algorithm . . . 7

1.2 ϵ-Greedy Algorithm . . . 7

1.2.1 Sub-Gaussian Concentration Bounds . . . 7

1.3 Upper Confidence Bound (UCB) Algorithm . . . 9

2 Unconstrained optimization [1] **15**

2.1 Newton's method for unconstrained minimization (Section 9.5) . . . 15

2.1.1 The Newton step is a descent direction . . . 15

2.1.2 One-dimensional intuition . . . 15

2.1.3 First interpretation: minimizing a quadratic approximation . . . 16

2.1.4 Second interpretation: linearizing the optimality condition . . . 16

2.2 Newton decrement and stopping criterion . . . 16

2.2.1 Backtracking line search . . . 17

2.2.2 Newton's method algorithm . . . 17

2.3 Convergence discussion . . . 17

3 Equality constrained optimization [1] **21**

3.1 Equality constrained minimization . . . 21

3.1.1 Introduction . . . 21

3.1.2 Equality constrained Convex Optimization . . . 22

3.1.3 Equality constrained convex quadratic minimization . . . 23

3.2 Eliminating equality constraints . . . 26

3.2.1 Review: equality-constrained minimization . . . 26

3.2.2 Review: Quadratic objective . . . 26

3.2.3 Parameterizing the feasible set (Section 10.1.2) . . . 26

3.2.4 Optimal allocation with a resource constraint (Example 10.1) . . . 26

3.3 Newton's method with equality constraints . . . 27

3.3.1 Algorithm for Newton's method with equality constraints . . . 28

3.3.2 Feasible descent directions . . . 28

4 PCA **29**
4.1 Introduction . . . 29
4.2 Sample Covariance Matrix . . . 29
4.2.1 Understand Σ_n (or C_n) . . . 30
4.3 PCA Algorithm . . . 30
4.4 In What Sense PCA Reorganizes the Data . . . 31
4.4.1 Recall Affine Space . . . 31
4.4.2 Least Square Metric . . . 32
4.4.3 PCA Minimizes the Least Square Metric . . . 32
4.5 Complexity of the PCA Algorithm . . . 33
4.6 Singular Value Decomposition (SVD) . . . 33

5 Vector spaces [3] **35**
5.1 Introduction . . . 35
5.2 Vector Spaces, Bases, and Dimension . . . 35
5.2.1 Linear independence and span . . . 35
5.3 Norms and Inner Products . . . 36
5.3.1 Norms and balls . . . 36
5.3.2 Inner products and their induced norms . . . 37
5.3.3 The p-norms on $\mathbb{R}^n$ . . . 38
5.4 Norms on Matrix Spaces . . . 38
5.4.1 The Frobenius inner product and norm . . . 38
5.4.2 Induced matrix norms . . . 38
5.4.3 Spectral norm . . . 39
5.4.4 Maximum absolute column sum norm . . . 39
5.4.5 Maximum absolute row sum norm . . . 40
5.5 The Dual Space . . . 40

6 Closed functions and subgradient [3] **43**
6.1 Extended Real-Valued Functions and Closedness . . . 43
6.2 Subgradient . . . 47

Chapter 1

Multi-Armed Bandit [2]

1.1 Introduction

The multi-armed bandit problem is a sequential decision-making formulation where an agent chooses between A distinct actions (arms) at each time step to maximize cumulative reward (or minimize cumulative regret):

- A available decisions/actions.
- Select an action at each round.
- Receive a stochastic, noisy reward.

1.1.1 Multi-Armed Bandit Protocol

For time steps $t = 1, 2, \ldots, T$:

1. Select a decision $\pi^t \in \Pi \coloneqq \{1, 2, \ldots, A\} = [A]$.
2. Observe a random reward $r^t \in \mathbb{R}$.

Here, t denotes the current time step, T is the total number of time steps, and A is the total number of available actions.

History The interaction history H_{t-1} contains all action-reward pairs observed up to time $t-1$:

$$H_{t-1} \coloneqq \{(a_1, r_1), (a_2, r_2), \ldots, (a_{t-1}, r_{t-1})\}$$

Assumption: Stochastic Reward The reward r_t is a random variable drawn from an underlying distribution conditioned on the interaction history:

$$r_t \sim M^*(\cdot \mid H_{t-1})$$

We briefly recall the multi-armed bandit setup and the ε-Greedy algorithm from the previous lecture. Let $f^*(\pi)$ denote the mean reward of action π, and let $\pi^* = \arg\max_{\pi \in \Pi} f^*(\pi)$ be an optimal action.

Reward Function The true expected reward function f^* for selecting action π is defined as:

$$f^*(\pi) \coloneqq \mathbb{E}[r \mid \pi]$$

Optimal Action The optimal decision π^* yields the highest expected reward:

$$\pi^* \coloneqq \arg\max_{\pi \in \Pi} f^*(\pi)$$

Regret The cumulative regret $\mathrm{Reg}(T)$ measures performance loss relative to the optimal action choice:

$$\mathrm{Reg} \coloneqq \sum_{t=1}^{T} f^*(\pi^*) - \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}[f^*(\pi^t)]$$

The first term is the total mean reward we could obtain by always choosing the optimal action π^*, whereas the second term is the expected total mean reward obtained by the algorithm. Thus, the goal is to minimize the regret.

Objective The goal is to design an algorithm achieving sub-linear regret, meaning the average regret vanishes asymptotically:

$$\lim_{T \to \infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{T} = 0$$

1.1.2 Simple Example: Always Pure Exploitation

Suppose we always select the first decision, $\pi^t = 1$, for all $t = 1, 2, \ldots, T$. The total regret simplifies to:

$$\begin{aligned} \mathrm{Reg}(T) &\coloneqq \sum_{t=1}^{T} f^*(\pi^*) - \sum_{t=1}^{T} f^*(1) \\ &= T\big(f^*(\pi^*) - f^*(1)\big) = \mathcal{O}(T) \end{aligned}$$

This yields linear regret $\mathcal{O}(T)$, failing our sub-linear objective whenever action 1 is suboptimal.

1.1.3 Estimating f^* via Empirical Means

Since the true reward distribution f^* is unknown, we estimate it using the empirical average of past rewards:

$$\hat{f}^t(\pi) \coloneqq \frac{1}{n^t(\pi)} \sum_{s<t} r^s \mathbf{1}_{\{\pi^s = \pi\}}$$

where $n^t(\pi) \coloneqq \sum_{s<t} \mathbf{1}_{\{\pi^s = \pi\}}$ denotes the total number of times decision π was selected up to time t. The indicator function is defined as: $\mathbf{1}_{\{\pi^s = \pi\}} = \begin{cases} 1, & \pi^s = \pi, \\ 0, & \pi^s \neq \pi. \end{cases}$

1.1.4 The Greedy Algorithm and Exploration

A naive approach is the **Greedy Algorithm**, which chooses the decision maximizing the empirical reward estimate at step t:

$$\hat{\pi}^t = \arg\max_{\pi \in \Pi} \hat{f}^t(\pi)$$

While greedy selection exploits current knowledge, it lacks a mechanism for exploration. Because feedback is partial (rewards are observed only for chosen actions), the algorithm can prematurely abandon optimal actions due to early noisy estimates. Once an optimal action's empirical reward drops below another arm's estimate, it is never selected again, leading to linear regret $\mathcal{O}(T)$.

1.1.5 Failure of the Greedy Algorithm

Consider a two-armed bandit setting with $\Pi = \{1, 2\}$:

- Decision 1 yields a deterministic reward: $f^*(1) = \frac{1}{2}$.
- Decision 2 yields a stochastic reward: $r \sim \text{Ber}\left(\frac{3}{4}\right)$, so $f^*(2) = \frac{3}{4}$.

Consider the following trajectory:

1. At $t = 1$, choose Decision 1: receive $r^1 = \frac{1}{2}$. Thus, $\hat{f}^2(1) = \frac{1}{2}$.
2. At $t = 2$, choose Decision 2: receive $r^2 = 0$ (occurring with probability $\frac{1}{4}$). Thus, $\hat{f}^3(2) = 0$.
3. For all $t \geq 3$, the greedy strategy continuously chooses Decision 1 because $\hat{f}^t(1) = \frac{1}{2} > 0 = \hat{f}^t(2)$.

Under this event, the expected regret grows linearly:

$$\text{Reg}(T) = \left(\frac{3}{4} - \frac{1}{2}\right) + \left(\frac{3}{4} - \frac{3}{4}\right) + (T-2)\left(\frac{3}{4} - \frac{1}{2}\right) = \frac{T-1}{4} = \mathcal{O}(T)$$

1.2 ϵ-Greedy Algorithm

To prevent premature convergence to suboptimal choices, the ϵ-greedy algorithm forces exploration:

Exploitation With probability $1-\epsilon$, exploit current knowledge by picking the empirically optimal action:

$$\hat{\pi}^t = \arg\max_{\pi \in \Pi} \hat{f}^t(\pi)$$

Exploration With probability ϵ, explore uniformly at random across all available actions:

$$\hat{\pi}^t \sim \text{Unif}(\{1, 2, \ldots, A\})$$

Thus, with probability $1-\varepsilon$ the algorithm uses the currently best estimated action, while with probability ε it explores.

1.2.1 Sub-Gaussian Concentration Bounds

Definition 1.2.1 (σ-sub-Gaussian Variable)**.** A zero-mean random variable Z is σ-sub-Gaussian if its Moment Generating Function (MGF) satisfies:

$$\mathbb{E}\left[e^{\eta Z}\right] \leq \exp\left(\frac{\sigma^2 \eta^2}{2}\right), \quad \forall \eta \in \mathbb{R}$$

Lemma 1.2.2 (Sub-Gaussian Hoeffding Bound)**.** *Let $Z_1, \ldots, Z_T$ be independent, zero-mean σ-sub-Gaussian random variables. Then for any $u > 0$:*

$$\mathbb{P}\left(\frac{1}{T}\sum_{i=1}^{T} Z_i \geq u\right) \leq \exp\left(-\frac{Tu^2}{2\sigma^2}\right)$$

Proposition 1.2.3. *Assume that $f^*(\pi) \in [0,1]$ and the reward r^t is 1-sub-Gaussian (always think of it as Gaussian with variance 1). Then, for any finite T, by choosing ε appropriately, the ε-greedy algorithm ensures that with probability at least $1-\delta$,*

$$\mathrm{Reg} \lesssim A^{1/3}T^{2/3}\log^{1/3}\left(\frac{AT}{\delta}\right)$$

Consequently,

$$E\left[\mathrm{Reg}\right] \lesssim A^{1/3}T^{2/3}\log^{1/3}\left(AT\right)$$

To prove the proposition, let us consider the following lemma.

Lemma 1.2.4. *With probability at least $1-\delta$,*

$$\max_{\pi}\left|\widehat{f}^t(\pi) - f^*(\pi)\right| \lesssim \sqrt{\frac{A\log(AT/\delta)}{\varepsilon t}}$$

Comment: The left-hand side measures the largest estimation error over all actions. As t becomes larger, more observations are collected, so the estimation error becomes smaller. For a proof of the lemma, see Section 2.2 of [2].

Proof of Proposition 2.1

We begin with

$$\mathrm{Reg} = \sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim p^t}\left[f^*(\pi^*) - f^*(\pi^t)\right].$$

Using the definition of the ε-greedy algorithm, this can be decomposed as

$$\mathrm{Reg} = \sum_{t=1}^{T}(1-\varepsilon)\left[f^*(\pi^*) - f^*(\widehat{\pi}^t)\right] + \varepsilon\sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim\mathrm{unif}([A])}\left[f^*(\pi^*) - f^*(\pi^t)\right].$$

Since $f^*(\pi)\in[0,1]$,$f^*(\pi^*) - f^*(\pi^t) \le 1$. Therefore,

$$\mathrm{Reg} \le \sum_{t=1}^{T}\left[f^*(\pi^*) - f^*(\widehat{\pi}^t)\right] + \varepsilon T.$$

Now decompose

$$f^*(\pi^*) - f^*(\widehat{\pi}^t) = \left[f^*(\pi^*) - \widehat{f}^t(\pi^*)\right] + \left[\widehat{f}^t(\pi^*) - \widehat{f}^t(\widehat{\pi}^t)\right] + \left[\widehat{f}^t(\widehat{\pi}^t) - f^*(\widehat{\pi}^t)\right].$$

Because $\widehat{\pi}^t = \arg\max_\pi \widehat{f}^t(\pi)$, we have $\widehat{f}^t(\pi^*) - \widehat{f}^t(\widehat{\pi}^t) \le 0$. Hence,

$$f^*(\pi^*) - f^*(\widehat{\pi}^t) \le 2\max_{\pi}\left|f^*(\pi) - \widehat{f}^t(\pi)\right|.$$

Therefore,

$$\mathrm{Reg} \le 2\sum_{t=1}^{T}\max_{\pi}\left|f^*(\pi) - \widehat{f}^t(\pi)\right| + \varepsilon T.$$

Using the lemma 2.2, with probability at least $1-\delta$

$$\text{Reg} \lesssim 2\sum_{t=1}^{T}\sqrt{\frac{A\log(AT/\delta)}{\varepsilon t}} + \varepsilon T.$$

Using the fact $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 2\sqrt{T}-1$, we obtain

$$\text{Reg} \lesssim 2\sqrt{\frac{A\log(AT/\delta)}{\varepsilon}}\sqrt{T} + \varepsilon T.$$

Comment: The first term is associated with the estimation error, while the second term comes from random exploration.

To balance these two terms, choose $\varepsilon \propto \left(\frac{A\log(AT/\delta)}{T}\right)^{1/3}$.

Plugging this choice into the previous bound gives

$$\text{Reg} \lesssim A^{1/3}T^{2/3}\log^{1/3}\left(\frac{AT}{\delta}\right).$$

This leads to the following question: can we improve the algorithm? The ε-greedy algorithm explores with probability ε, even for those well-estimated bad arms/actions. Can we choose actions in a more adaptive way instead of exploring uniformly?

1.3 Upper Confidence Bound (UCB) Algorithm

Suppose that at each time t we can construct a lower and upper confidence bound

$$\underline{f}^t, \overline{f}^t : \Pi \to \mathbb{R}$$

such that, with probability at least $1-\delta$,

$$f^*(\pi) \in \left[\underline{f}^t(\pi), \overline{f}^t(\pi)\right]$$

for every $t \in [T]$ and every $\pi \in [A]$.

The UCB algorithm chooses $\pi^t = \arg\max_{\pi\in\Pi}\overline{f}^t(\pi)$. Thus, at time t, the chosen action is the one having the largest upper confidence bound.

Lemma 1.3.1 (Instantaneous regret of UCB)**.** *For t, suppose $f^*(\pi) \in \left[\underline{f}^t(\pi), \overline{f}^t(\pi)\right]$*
Then $\pi^t = \arg\max_\pi \overline{f}^t(\pi)$ has the following property:

$$f^*(\pi^*) - f^*(\pi^t) \leq \overline{f}^t(\pi^t) - f^*(\pi^t) \leq \overline{f}^t(\pi^t) - \underline{f}^t(\pi^t).$$

Proof. Because π^t maximizes the upper confidence bound, $\overline{f}^t(\pi^*) \leq \overline{f}^t(\pi^t)$. Also, because the confidence interval contains the true mean reward, $f^*(\pi^*) \leq \overline{f}^t(\pi^*)$. Combining these gives $f^*(\pi^*) \leq \overline{f}^t(\pi^t)$. Therefore,

$$f^*(\pi^*) - f^*(\pi^t) \leq \overline{f}^t(\pi^t) - f^*(\pi^t).$$

Finally, $f^*(\pi^t) \geq \underline{f}^t(\pi^t)$, so $\overline{f}^t(\pi^t) - f^*(\pi^t) \leq \overline{f}^t(\pi^t) - \underline{f}^t(\pi^t)$.

□

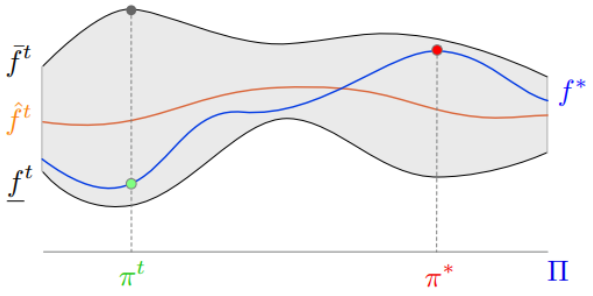


Figure 1.1: Illustration of the UCB algorithm.

Lemma 1.3.2 (Construction of confidence bounds)**.** *With probability 1-δ,*

$$\max \left| \widehat{f}^t(\pi) - f^*(\pi) \right| \leq \sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi)}}.$$

We skip the proof. Then by the lemma,

$$\underline{f}^t(\pi) := \widehat{f}^t(\pi) - \sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi)}} \leq f^*(\pi) \leq \widehat{f}^t(\pi) + \sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi)}} =: \overline{f}^t(\pi)$$

UCB:

Define

$$\widehat{f}^{\,t}(\pi) = \frac{1}{n^t(\pi)} \sum_{s<t} r^s \mathbf{1}(\pi^s = \pi)\,.$$

The upper confidence bound is

$$\overline{f}^{\,t}(\pi) = \widehat{f}^{\,t}(\pi) + \sqrt{\frac{2\log\left(\frac{2\pi^2 t^2 A}{\delta}\right)}{n^t(\pi)}}.$$

Similarly, the lower confidence bound is

$$\underline{f}^{\,t}(\pi) = \widehat{f}^{\,t}(\pi) - \sqrt{\frac{2\log\left(\frac{2\pi^2 t^2 A}{\delta}\right)}{n^t(\pi)}}.$$

With probability at least $1-\delta$,

$$\underline{f}^{\,t}(\pi) \leq f^\star(\pi) \leq \overline{f}^{\,t}(\pi).$$

The UCB action selection rule is

$$\pi^t = \arg\max_{\pi} \overline{f}^{\,t}(\pi).$$

Initialization:

At $t = 1$,

$$n^t(\pi) = 0, \qquad \forall \pi \in [A],$$

and thus

$$\overline{f}^{\,1}(\pi) = \infty, \qquad \forall \pi \in [A].$$

In the first step, any arm may be selected since it is all infinity. Once a certain arm is selected, then the upper confidence bound $\overline{f}$ for that arm is finite. So at $t = 2$, any remaining arm maybe selected. In the first A steps, each arm will be selected once.

Proposition 1.3.3 (UCB regret)**.** *Assume $f^*(\pi) \in [0,1]$ and r^t is 1-sub-Gaussian. Then, with probability at least $1-\delta$,*

$$\mathrm{Reg} \lesssim \sqrt{AT \log(AT/\delta)}.$$

Comment: This gives a $T^{1/2}$ dependence, compared with the $T^{2/3}$ dependence of the ε-greedy bound. A smaller power of T is better because the regret grows more slowly. Thus, the UCB algorithm has a better regret bound than the ε-greedy algorithm.

Proof of Proposition 1.3.3. Using lemma 1.3.1 and our construction of the confidence bounds, the regret can be bounded by the sum of the confidence widths: with probability 1-δ,

$$f^*(\pi^*) - f^*(\pi^t) \le \overline{f}^t(\pi^t) - \underline{f}^t(\pi^t) \le 2\sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi^t)}}$$

On the other hand,

$$f^*(\pi^*) - f^*(\pi^t) \le 1.$$

Thus, with probability 1-δ,

$$f^*(\pi^*) - f^*(\pi^t) \le 2\sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi^t)}} \wedge 1$$

where $a \wedge b = \min\{a, b\}$.

Thus with probability 1-δ,

$$\begin{aligned}\mathrm{Reg} &= \sum_{t=1}^{T} \left[f^*(\pi^*) - f^*(\pi^t)\right] \\ &\lesssim \sum_{t=1}^{T} \left(2\sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi^t)}} \wedge 1\right).\end{aligned}$$

here, the "$\wedge 1$" term appears because we can write off the regret for early rounds where $n^t(\pi^t) = 0$ as 1.

To be continued. □

Lemma 1.3.4 (Confidence width potential lemma)**.** *We have*

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n^t(\pi^t)}} \wedge 1 \lesssim \sqrt{AT}.$$

Proof:

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n^t(\pi^t)}} \wedge 1 = \sum_{\pi} \sum_{t=1}^{T} \frac{\mathbf{1}(\pi^t = \pi)}{\sqrt{n^t(\pi)}} \wedge 1.$$

Since

$$n^t(\pi^t) = \sum_{\pi} n^t(\pi)\mathbf{1}(\pi^t = \pi),$$

we can write

$$\frac{1}{\sqrt{n^t(\pi^t)}} = \sum_{\pi} \frac{\mathbf{1}(\pi^t = \pi)}{\sqrt{n^t(\pi)}}.$$

Therefore,

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n^t(\pi^t)}} \wedge 1 = \sum_{\pi} \sum_{t=1}^{T} \frac{\mathbf{1}(\pi^t = \pi)}{\sqrt{n^t(\pi)}} \wedge 1.$$

Re-indexing according to the number of times arm π is selected,

$$= \sum_{\pi} \sum_{t=1}^{n^{T+1}(\pi)} \frac{1}{\sqrt{t-1}} \wedge 1.$$

Using

$$\sum_{t=1}^{n} \frac{1}{\sqrt{t-1}} \wedge 1 \leq 1 + 2\sqrt{n},$$

we obtain

$$\leq \sum_{\pi} \left(1 + 2\sqrt{n^{T+1}(\pi)}\right).$$

Thus,

$$= A + 2 \sum_{\pi} \sqrt{n^{T+1}(\pi)}.$$

By Jensen's inequality, since $F(x) = \sqrt{x}$ is concave,

$$\frac{1}{A} \sum_{\pi} \sqrt{n^{T+1}(\pi)} \leq \sqrt{\frac{1}{A} \sum_{\pi} n^{T+1}(\pi)}.$$

Hence,

$$A + 2 \sum_{\pi} \sqrt{n^{T+1}(\pi)} = A + 2A \left(\frac{1}{A} \sum_{\pi} \sqrt{n^{T+1}(\pi)} \right)$$

$$\leq A + 2A \sqrt{\frac{1}{A} \sum_{\pi} n^{T+1}(\pi)}.$$

Since

$$\sum_{\pi} n^{T+1}(\pi) = T,$$

we get

$$\leq A + 2A \sqrt{\frac{T}{A}} = A + 2\sqrt{AT}.$$

Finally, consider two cases:

1. If $A \leq \sqrt{AT}$ (equivalently, $A \leq T$),

then

$$A + 2\sqrt{AT} \leq 3\sqrt{AT}.$$

2. If $A > T$,

then trivially

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n^t(\pi^t)}} \wedge 1 \leq T \leq \sqrt{AT}.$$

Therefore,

$$\boxed{\sum_{t=1}^{T} \frac{1}{\sqrt{n^t(\pi^t)}} \wedge 1 \leq 3\sqrt{AT}}$$

Chapter 2

Unconstrained optimization [1]

2.1 Newton's method for unconstrained minimization (Section 9.5)

Now consider minimizing $f(x)$ without equality constraints. At a current point $x \in \mathrm{dom}(f)$, the Newton step is

$$\Delta x_{\mathrm{nt}} = -[\nabla^2 f(x)]^{-1} \nabla f(x). \tag{2.1}$$

Equivalently, it solves the linear system $\nabla^2 f(x) \Delta x_{\mathrm{nt}} = -\nabla f(x)$.

2.1.1 The Newton step is a descent direction

For a direction v, the directional derivative at x is

$$\left. \frac{d}{dt} f(x + tv) \right|_{t=0} = \langle \nabla f(x), v \rangle = \nabla f(x)^T v.$$

A negative value means that v is a descent direction. Substituting (2.1) yields

$$\left. \frac{d}{dt} f(x + t\Delta x_{\mathrm{nt}}) \right|_{t=0} = -\nabla f(x)^T [\nabla^2 f(x)]^{-1} \nabla f(x) \leq 0.$$

2.1.2 One-dimensional intuition

In one dimension, we seek a zero of $f'(x)$. Approximating f' by its tangent line at x_k gives

$$f'(x_k) + f''(x_k)(x_{k+1} - x_k) = 0, \qquad x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}.$$

The next iterate is the horizontal-axis intercept of that tangent line. Repeating the construction gives successive Newton iterates.
Geometrically, Newton's method uses a tangent line to approximate $f'(x)$ and find where $f'(x) = 0$. The tangent line has the form

$$g'(x) = ax + b.$$

This means that we are approximating $f(x)$ using a quadratic function of the form

$$g(x) = \frac{1}{2} a x^2 + bx + c.$$

The derivative of this quadratic is $g'(x) = ax + b$, so finding the zero of the tangent-line approximation of $f'(x)$ is equivalent to finding the minimum of the quadratic approximation of $f(x)$ when $a > 0$.

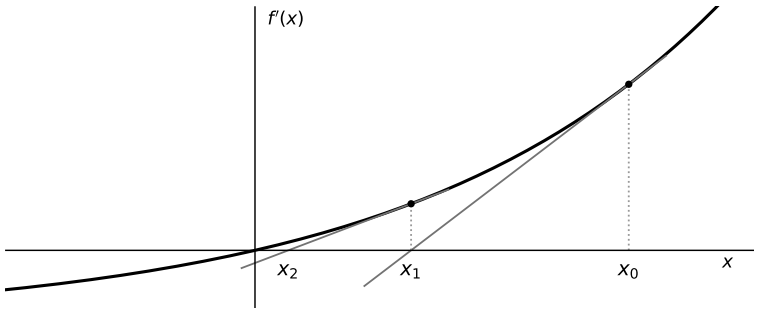


Figure 2.1: Illustration of the tangent-line construction applied to a general f'..

2.1.3 First interpretation: minimizing a quadratic approximation

The second-order Taylor approximation about x is

$$\widehat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v. \tag{2.2}$$

Treat the right-hand side as a function $g(v)$. Its gradient is

$$\nabla g(v) = \nabla f(x) + \nabla^2 f(x) v.$$

Because the Hessian of g is positive definite, the condition $\nabla g(v) = 0$ is necessary and sufficient for its unique minimum. Solving gives

$$v = -[\nabla^2 f(x)]^{-1} \nabla f(x) = \Delta x_{\mathrm{nt}}.$$

Thus the Newton step minimizes the local quadratic model.

2.1.4 Second interpretation: linearizing the optimality condition

For unconstrained convex minimization, the first-order optimality condition is $\nabla f(x+v) = 0$. The first order Taylor expansion gives

$$\nabla f(x+v) \approx \nabla f(x) + \nabla^2 f(x) v.$$

Setting this approximation equal to zero again gives $v = \Delta x_{\mathrm{nt}}$. Unlike the negative gradient direction, the Newton direction uses curvature information through the Hessian. If $\nabla^2 f(x) = I$, then $\Delta x_{\mathrm{nt}} = -\nabla f(x)$, simplifying to the standard gradient descent direction.

2.2 Newton decrement and stopping criterion

Definition 2.2.1. The Newton decrement at x is

$$\lambda(x) = \left(\nabla f(x)^T [\nabla^2 f(x)]^{-1} \nabla f(x) \right)^{1/2}. \tag{2.3}$$

It measures the improvement predicted by the quadratic approximation. Since (2.2) is minimized at the

Newton step,

$$\begin{aligned} f(x) - p^* = f(x) - \inf_y \widehat{f}(y) &= f(x) - \widehat{f}(x + \Delta x_{\text{nt}}) \\ &= -\nabla f(x)^T \Delta x_{\text{nt}} - \frac{1}{2} \Delta x_{\text{nt}}^T \nabla^2 f(x) \Delta x_{\text{nt}} \\ &= \nabla f(x)^T [\nabla^2 f(x)]^{-1} \nabla f(x) - \frac{1}{2} \nabla f(x)^T [\nabla^2 f(x)]^{-1} \nabla f(x) \\ &= \frac{1}{2} \lambda(x)^2. \end{aligned}$$

In simplifying the quadratic term, we use symmetry of the Hessian and its inverse.

2.2.1 Backtracking line search

Suppose Δx is a descent direction for f at $x \in \operatorname{dom} f$ $\alpha \in (0, \frac{1}{2})$ and $\beta \in (0, 1)$.

1. Initialize $t = 1$.
2. While
$$f(x + t\Delta x) \leq f(x) + \alpha t \nabla f(x)^\top \Delta x$$
is not satisfied, reduce the step size by setting $t \leftarrow \beta t$.
3. Return the first step size satisfying the condition.

The right-hand side is smaller than $f(x)$ because $\nabla f(x)^\top \Delta x < 0$. Thus, the condition requires a sufficient decrease in the objective value.

2.2.2 Newton's method algorithm

Given a starting point $x \in \operatorname{dom}(f)$ and tolerance $\epsilon > 0$, repeat:

1. Compute the Newton step and squared decrement:
$$\Delta x_{\text{nt}} = -[\nabla^2 f(x)]^{-1} \nabla f(x), \qquad \lambda(x)^2 = \nabla f(x)^T [\nabla^2 f(x)]^{-1} \nabla f(x).$$
2. Stop if $\lambda(x)^2/2 \leq \epsilon$.
3. Choose a step size t by backtracking line search.
4. Update $x \leftarrow x + t\Delta x_{\text{nt}}$.

2.3 Convergence discussion

Newton's method can converge much faster than gradient descent because it uses curvature information in addition to the gradient.

The convergence argument uses the following assumptions:

1. f is twice continuously differentiable.
2. f is strongly convex with parameter $m > 0$; in particular,
$$\nabla^2 f(x) \succeq mI, \qquad x \in \mathcal{S}.$$
$$\nabla^2 f(x) \preceq MI, \qquad x \in \mathcal{S},$$

3. The Hessian is bounded above and is Lipschitz continuous. For constants $M > 0$ and $L > 0$,

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_2 \leq L \left\|x - y\right\|_2, \qquad x, y \in \mathcal{S}.$$

Here, $\|A\|_2$ denotes the spectral norm, which is the largest singular value of A.

$$\text{We will show } \exists \ \eta, \gamma \text{ s.t.} \quad 0 < \eta \leq \frac{m^2}{L}, \quad \gamma > 0,$$

$$\text{s.t. the following hold:}$$

- If $\|\nabla f(x^{(k)})\|_2 \geq \eta$, then
$$f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma.$$
- If $\|\nabla f(x^{(k)})\|_2 < \eta$, then the backtracking line search selects $t^{(k)} = 1$ and
$$\frac{L}{2m^2}\|\nabla f(x^{(k+1)})\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2\right)^2.$$

Suppose that for some iteration k, $\left\|\nabla f(x^{(k)})\right\|_2 < \eta$, then $\left\|\nabla f(x^{(k)})\right\|_2 \leq m^2/L$

$$\frac{L}{2m^2}\left\|\nabla f(x^{(k)})\right\|_2 \leq \frac{1}{2} < 1.$$

$$\frac{L}{2m^2}\left\|\nabla f(x^{(k+1)})\right\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2\right)^2 \leq \frac{L}{2m^2}\left\|\nabla f(x^{(k)})\right\|_2$$

$$\left\|\nabla f(x^{(k+1)})\right\|_2 \leq \left\|\nabla f(x^{(k)})\right\|_2 < \eta$$

$$\boxed{\text{So } \left\|\nabla f(x^{(s)})\right\|_2 < \eta, \forall s \geq k.}$$

Note

$$\frac{L}{2m^2}\left\|\nabla f(x^{(k+1)})\right\|_2 \leq (\frac{1}{2})^2 < 1$$

for all $\ell \geq k$,

$$\begin{aligned}
\frac{L}{2m^2}\left\|\nabla f(x^{(\ell)})\right\|_2 &\leq \left(\frac{L}{2m^2}\left\|\nabla f(x^{(\ell-1)})\right\|_2\right)^2 \\
&\leq \left(\frac{L}{2m^2}\left\|\nabla f(x^{(\ell-2)})\right\|_2\right)^{2^2} \\
&\leq \cdots \\
&\leq \left(\frac{L}{2m^2}\left\|\nabla f(x^{(k)})\right\|_2\right)^{2^{\ell-k}} \\
&\leq \left(\frac{1}{2}\right)^{2^{\ell-k}}.
\end{aligned}$$

Recall the strongly convex objective-gap bound

$$f(x) - p^\star \leq \frac{1}{2m}\left\|\nabla f(x)\right\|_2^2,$$

where $p^\star = \min_x f(x)$. Therefore,

$$\begin{aligned} f(x^{(\ell)}) - p^\star &\le \frac{1}{2m}\|\nabla f(x^{(\ell)})\|_2^2 \\ &\le \frac{1}{2m}\left(\frac{2m^2}{L}\left(\frac{1}{2}\right)^{2^{\ell-k}}\right)^2 \\ &= \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^{\ell-k+1}}. \end{aligned}$$

Let $x^{(0)}$ is the initial point then the number of iterations in Stage 1 is at most

$$\frac{f(x^{(0)}) - p^\star}{\gamma}.$$

For Stage 2, we want

$$f(x^{(\ell)}) - p^\star \le \epsilon.$$

To achieve an objective gap no larger than ε, it is sufficient that

$$\frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^{\ell-k+1}} \le \epsilon.$$

$$\left(\frac{1}{2}\right)^{2^{\ell-k+1}} \le \frac{\epsilon}{\epsilon_0}, \qquad \text{where } \epsilon_0 = \frac{2m^3}{L^2}.$$

$$2^{-2^{\ell-k+1}} \le \frac{\epsilon}{\epsilon_0}.$$

Taking logarithms base 2,

$$-2^{\ell-k+1} \le \log_2\left(\frac{\epsilon}{\epsilon_0}\right).$$

Therefore,

$$2^{\ell-k+1} \ge \log_2\left(\frac{\epsilon_0}{\epsilon}\right).$$

Taking logarithms base 2 again,

$$\boxed{\ell - k + 1 \ge \log_2 \log_2\left(\frac{\epsilon_0}{\epsilon}\right)}$$

Combining the two stages gives the below iteration-count estimate:

$$\frac{f(x^{(0)}) - p^\star}{\gamma} + \log_2\left[\log_2\left(\frac{\varepsilon_0}{\varepsilon}\right)\right],$$

The number of iterations in the second stage is bounded by a term of order $\log_2 \log_2(\varepsilon_0/\varepsilon)$, which is very small (almost constant) compared with the first-stage bound. The main contribution is coming from the first stage.

Chapter 3

Equality constrained optimization [1]

3.1 Equality constrained minimization

3.1.1 Introduction

In this lecture, we begin the discussion of the methods for solving a convex optimization problem with equality constraints. We want to minimize $f(x)$,

$$\begin{aligned} &\min f(x) \\ &\quad \text{s.t. } Ax = b, \end{aligned}$$

where $A \in \mathbb{R}^{p \times n}$.

To compare this problem with the general framework discussed in the previous lectures, we can rewrite the problem as follows:

$$\begin{aligned} &\min f(x) \\ &\quad \text{s.t. } Ax - b = 0, \end{aligned}$$

where $A = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_p^T \end{bmatrix}$ and a_i, $i \in \{1, 2, \ldots, p\}$ are column vectors.

Then

$$Ax - b = \begin{bmatrix} a_1^T x - b_1 \\ a_2^T x - b_2 \\ \vdots \\ a_p^T x - b_p \end{bmatrix} = 0. \tag{3.1}$$

From Eqn. (3.1) we can determine the equality constraint h_j:

$$h_j(x) = a_j^T x - b_j,$$

for $j \in \{1, 2, \ldots, p\}$

3.1.2 Equality constrained Convex Optimization

We are working with the convex optimization problem. To find a solution, make the following assumptions:

1. $f(x) : \mathbb{R}^n \to \mathbb{R}$ is convex and twice continuously differentiable
2. $A \in \mathbb{R}^{p \times n}$ has full row rank, i.e. it has linearly independent rows and $\text{rank}(A) = p < n$
3. Optimal solution $x^* \in \text{dom}\, f$ exists and $p^* = f(x^*)$ will produce the smallest value of $f(x)$.

To solve this problem, try using the KKT conditions.
The Lagrangian without the inequality constraint λ is:

$$\begin{aligned} L(x, \nu) &= f(x) + \sum_{j=1}^{p} \nu_j h_j(x) \\ &= f(x) + \sum_{j=1}^{p} \nu_j (a_j^T x - b_j) \\ &= f(x) + \nu^T (Ax - b) \\ &= f(x) + ((A^T \nu)^T x - \nu^T b), \end{aligned} \tag{3.2}$$

where $\nu \in \mathbb{R}^p$ is associated with the equality constraints.
The KKT conditions are as follows:

1. $\nabla_x L(x, \nu) = \nabla f(x) + A^T \nu = 0$, from (3.2)
2. NA since there are no inequality constraints,
3. $Ax = b$, from $h_j(x) = 0$, where $h_j = Ax - b$
4. NA since there are no inequality constraints,
5. NA since there are no inequality constraints.

Then (x^*, ν^*) satisfies

$$\begin{cases} \nabla f(x^*) + A^T \nu^* = 0 \\ Ax^* = b. \end{cases} \tag{3.3}$$

Since $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{p \times n}$,

$$\nabla f(x^*) + A^T \nu^* = 0$$

has n equations, and

$$Ax^* = b$$

has p equations.
Therefore, there are $n + p$ equations in the $n + p$ variables $x^*,\ \nu^*$.

3.1.3 Equality constrained convex quadratic minimization

Consider a simple example to build intuition.

10.1.1. Equality constrained convex quadratic minimization.

Let $f(x) = \frac{1}{2}x^T Px + q^T x + r$ be a quadratic function. We want to minimize $f(x)$,

$$\begin{aligned} &\min f(x) \\ &\text{s.t. } Ax = b, \end{aligned} \tag{3.4}$$

where $P \in \mathbb{S}_+^n$ and $A \in \mathbb{R}^{p\times n}$. Note that $\mathbb{S}_+^n$ are $n \times n$ PSD matrices with all eigenvalues ≥ 0.

Then

$$\nabla f(x) = Px + q$$

and

$$\nabla^2 f(x) = P,$$

hence, $\nabla^2 f(x)$ is PSD, which implies that $f(x)$ is convex.

Applying the derived KKT conditions in Eqn. (3.3) to this special case yields the following optimality conditions:

$$\begin{cases} Px^* + q + A^T\nu^* = 0 \\ Ax^* = b, \end{cases}$$

which implies

$$\begin{cases} Px^* + A^T\nu^* = -q \\ Ax^* + 0 \times \nu^* = b \end{cases} \tag{3.5}$$

The linear system of equations in Eqn. (3.5) can be rewritten in the matrix notation:

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \nu^* \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}, \tag{3.6}$$

where

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix}$$

is called the KKT matrix.

Solving this linear system of equations produces the desired solution to the equality constrained quadratic minimization problem.

There is a unique optimal primal-dual pair $(x^*,\ \nu^*)$ if the KKT matrix is non-singular. If the KKT matrix is singular and the system is solvable, any solution produces an optimal pair $(x^*,\ \nu^*)$. If the system is unsolvable, the quadratic optimization problem is unbounded below.

Lemma 3.1.1. *If (3.6) has no solution, then (3.4) is unbounded below.*

To prove this lemma, we need another lemma:

Lemma 3.1.2. *If a matrix $B \in \mathbb{R}^{d\times d}$ is symmetric and $y \notin \text{colspan}(B)$, then $\exists z\ Bz = 0$ and $\langle z,\ y\rangle > 0$.*

Proof. Recall: $B = \begin{bmatrix} b_1 & b_2 & \dots & b_d \end{bmatrix}$, where b_j for $j \in \{1, 2, \dots, d\}$ are column vectors. Then

$$\begin{aligned} \text{colspan}(B) &= \{\sum_{j=1}^{d} \lambda_j b_j \mid \lambda_j \in \mathbb{R}\} \\ &= \{B\lambda \mid \lambda \in \mathbb{R}^d\}. \end{aligned} \tag{3.7}$$

From Eqn. (3.7),

$$y \notin \text{colspan}(B) \Leftrightarrow B\lambda = y \text{ has no solution.}$$

Since $y \notin \text{colspan}(B)$, $\exists\, y_1, y_2$ $y = y_1 + y_2$, where $y_1 \in \text{colspan}(B)$ and $y_2 \perp \text{colspan}(B)$.

In other words, we can decompose the vector y into components parallel and perpendicular to colspan(B) as in Fig. (3.1).

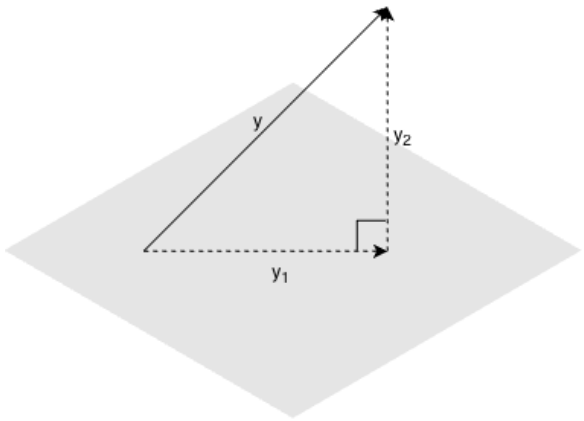


Figure 3.1: Vector y is decomposed into components y_1 (parallel to the column span of B) and y_2 (perpendicular to the column span of B).

Assume $y_2 = 0$. Then there is no component of y perpendicular to colspan(B), which means $y = y_1 \in$ colspan(B). It contradicts the assumption that $y \notin$ colspan(B). Therefore, $y_2 \neq 0$.

Since $y_2 \perp$ colspan(B), $\langle y_2, b_i \rangle = 0$ for $i \in \{1, 2, \dots, d\}$.

The equation above can be rewritten as

$$\begin{bmatrix} b_1^T y_2 \\ b_2^T y_2 \\ \vdots \\ b_d^T y_2 \end{bmatrix} = 0 \Leftrightarrow B^T y_2 = 0 \Leftrightarrow B y_2 = 0$$

since B is symmetric.

Now, check that $\langle y_2, y \rangle > 0$.

$$\begin{aligned} \langle y_2, y \rangle = \langle y_2, y_1 + y_2 \rangle &= \langle y_2, y_1 \rangle + \langle y_2, y_2 \rangle \\ &= 0 + \langle y_2, y_2 \rangle \\ &= \|y_2\|^2 > 0, \end{aligned}$$

since $y_2 \neq 0$ and $y_2 \perp y_1 \implies \langle y_2, y_1 \rangle = 0$.

□

Using Lemma 4.2, we can prove Lemma 4.1.

Proof. If (3.6) has no solution, then

$$\begin{bmatrix} -q \\ b \end{bmatrix} \notin \text{colspan} \begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix}.$$

By Lemma 4.2, $\exists z = \begin{bmatrix} v \\ w \end{bmatrix}$ such that

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and

$$\langle \begin{bmatrix} v \\ w \end{bmatrix}, \begin{bmatrix} -q \\ b \end{bmatrix} \rangle > 0 \Leftrightarrow -q^T v + b^T w > 0 \tag{3.8}$$

Let $\hat{x}$ be feasible point of (3.4), i.e. $\hat{x}$ satisfies the constraint $A\hat{x} = b$.
Then $\hat{x} + tv$ is also feasible for all t, where v is a particular direction and t is the step size.
Indeed,

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies \begin{cases} Pv + A^T w = 0 \\ Av = 0 \end{cases} \implies \begin{cases} Pv = -A^T w \\ Av = 0 \end{cases} \tag{3.9}$$

and

$$A(\hat{x} + tv) = A(\hat{x}) + t(Av) = b + t \times 0 = b,$$

which implies that $\hat{x} + tv$ is feasible.
Then

$$\begin{aligned} f(\hat{x} + tv) &= f(\hat{x}) + t(v^T P\hat{x} + q^T v) + \frac{1}{2}t^2 v^T P v \\ &= f(\hat{x}) + t(\hat{x}^T P v + q^T v) + \frac{1}{2}t^2 v^T P v \\ &= f(\hat{x}) + t(\hat{x}^T(-A^T w) + q^T v) + \frac{1}{2}t^2 v^T(-A^T w) && \text{from (3.9)} \\ &= f(\hat{x}) + t(-\hat{x}^T A^T w + q^T v) - \frac{1}{2}t^2 w^T A v \\ &= f(\hat{x}) + t(-b^T w + q^T v) - 0 && \text{using } A\hat{x} = b \text{ and } Av = 0 \end{aligned}$$

Since $(-q^T v + b^T w) > 0$ by Eqn. (3.8), we have $(-b^T w + q^T v) < 0$. Then

$$\lim_{t \to \infty} f(\hat{x} + tv) = -\infty.$$

Therefore, the quadratic optimization problem is unbounded below. □

3.2 Eliminating equality constraints

3.2.1 Review: equality-constrained minimization

Consider

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b, \end{array} \tag{3.10}$$

where f is convex and twice continuously differentiable, $A \in \mathbb{R}^{p\times n}$, and $\operatorname{rank}(A) = p$. Thus the equality constraints have no redundant rows. At a primal optimum x^*, the KKT conditions are

$$\nabla f(x^*) + A^T\nu^* = 0, \qquad Ax^* = b, \tag{3.11}$$

where $\nu^* \in \mathbb{R}^p$ is dual optimal.

3.2.2 Review: Quadratic objective

For the quadratic objective

$$f(x) = \frac{1}{2}x^T Px + q^T x + r,$$

the gradient is $Px + q$. Consequently, (3.11) becomes the linear system

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \nu^* \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}. \tag{3.12}$$

The block matrix has size $(n+p)\times(n+p)$. If it is invertible, the system has a unique solution. If the feasible convex quadratic problem has no solution to (3.12), its objective is unbounded below.

3.2.3 Parameterizing the feasible set (Section 10.1.2)

Choose any particular solution $\widehat{x}$ of $Ax = b$, and let $F \in \mathbb{R}^{n\times(n-p)}$ have columns forming a basis of the null space of A. In particular,

$$AF = 0, \qquad \operatorname{range}(F) = \operatorname{null}(A).$$

Every point can be written as

$$\{x : Ax = b\} = \{Fz + \widehat{x} : z \in \mathbb{R}^{n-p}\}. \tag{3.13}$$

To verify, take $x = Fz + \widehat{x}$. Then

$$Ax = A(Fz + \widehat{x}) = AFz + A\widehat{x} = 0 + b = b.$$

Define $\widetilde{f}(z) = f(Fz + \widehat{x})$. Problem (3.10) is equivalent to the unconstrained problem

$$\underset{z\in\mathbb{R}^{n-p}}{\text{minimize}} \quad \widetilde{f}(z).$$

After finding a minimizer z^*, we recover a minimizer of the original problem as $x^* = Fz^* + \widehat{x}$.

3.2.4 Optimal allocation with a resource constraint (Example 10.1)

Consider the separable objective and resource constraint

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{i=1}^n f_i(x_i) \\ \text{subject to} & \displaystyle\sum_{i=1}^n x_i = b. \end{array}$$

Here $A = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}$ has rank one. Eliminating the last variable gives

$$x_n = b - \sum_{i=1}^{n-1} x_i,$$

and hence the unconstrained function in $n-1$ variables is

$$\sum_{i=1}^{n-1} f_i(x_i) + f_n\left(b - \sum_{i=1}^{n-1} x_i\right).$$

Now the constraint is removed.
This substitution can be written in the format of (3.13). Let e_n be the last canonical basis vector in $\mathbb{R}^n$, and choose

$$\widehat{x} = be_n = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ b \end{bmatrix}, \qquad F = \begin{bmatrix} I_{n-1} \\ -\mathbf{1}^T \end{bmatrix} \in \mathbb{R}^{n \times (n-1)},$$

where $\mathbf{1} \in \mathbb{R}^{n-1}$ is the vector of ones. We have $A\widehat{x} = b$, $AF = 0$, and the columns of F are linearly independent, so they form a basis of $\text{null}(A)$. Therefore

$$Fz + \widehat{x} = \begin{bmatrix} z \\ b - \sum_{i=1}^{n-1} z_i \end{bmatrix}, \qquad z \in \mathbb{R}^{n-1},$$

which recovers precisely the elimination formula above.

3.3 Newton's method with equality constraints

Now consider the equality-constrained problem

$$\min_x f(x) \qquad \text{subject to} \qquad Ax = b,$$

where $A \in \mathbb{R}^{p \times n}$. Suppose the current iterate x is feasible, meaning $Ax = b$.

For a step v, the second-order approximation to the objective is

$$f(x+v) \approx f(x) + \nabla f(x)^\top v + \frac{1}{2} v^\top \nabla^2 f(x) v := \hat{f}(x+v).$$

Feasibility of the new point requires

$$A(x+v) = b.$$

Since $Ax = b$, the step must satisfy $Av = 0$. Therefore, the local quadratic problem is

$$\begin{aligned} &\text{minimize}_v \quad \nabla f(x)^\top v + \frac{1}{2} v^\top \nabla^2 f(x) v \\ &\text{subject to} \quad Av = 0. \end{aligned}$$

The constant term $f(x)$ can be omitted because it does not affect the minimizer.
Recall:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} x^T P x + q^T x + r \\ \text{subject to} \quad & Ax = b \end{aligned}$$

KKT system:

$$Ax^\star = b$$
$$Px^\star + q + A^T\nu^\star = 0$$

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^\star \\ \nu^\star \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}$$

Optimize over Δx_{nt}:

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\text{nt}} \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}$$

Newton decrement

$$\lambda(x) = \left(\Delta x_{\text{nt}}^T \nabla^2 f(x) \Delta x_{\text{nt}}\right)^{1/2}$$

$$f(x) - p^\star \approx f(x) - \inf_{\substack{v \\ A(x+v)=b}} \left\{\hat{f}(x+v)\right\} = \frac{\lambda(x)^2}{2},$$

3.3.1 Algorithm for Newton's method with equality constraints

Given a starting point $x \in \operatorname{dom} f$ satisfying $Ax = b$ and a tolerance $\varepsilon > 0$, repeat the following steps.

1. Compute the constrained Newton step Δx_{nt} by solving the KKT system. Compute the squared Newton decrement
$$\lambda(x)^2 = \Delta x_{\text{nt}}^\top \nabla^2 f(x) \Delta x_{\text{nt}}.$$
2. **Stopping criterion:** stop if
$$\frac{\lambda(x)^2}{2} \le \varepsilon.$$
3. **Line search:** choose a step size t using backtracking line search
4. **Update:** set
$$x \leftarrow x + t\Delta x_{\text{nt}}.$$

Since the constrained Newton direction satisfies $A\Delta x_{\text{nt}} = 0$, each update preserves $Ax = b$.

3.3.2 Feasible descent directions

A vector $v \in \mathbb{R}^n$ is a feasible direction for the equality constraint if

$$Av = 0.$$

It is a feasible descent direction at x if

$$Av = 0 \qquad \text{and} \qquad f(x+tv) < f(x) \quad \text{for all sufficiently small } t > 0.$$

Chapter 4

PCA

4.1 Introduction

We have a high dim data set. It distributes around a low dim set.
Dimension Reduction: Reorganize the data in a low dim space. We want to preserve some properties of the data set in the low dim representation.
Principal Component Analysis (PCA).

1. Algorithm.
2. In what sense it reorganizes the data.
3. Efficient algorithm through SVD.

PCA is a linear dimension reduction method.
Data set:

$$\{X_1, \dots, X_n\} \subset \mathbb{R}^p, \qquad p \text{ is large.}$$

$$X_i = \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}, \qquad p \text{ vector.}$$

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} X_i \in \mathbb{R}^p.$$

4.2 Sample Covariance Matrix

$$C_n = \sum_{i=1}^{n} \underbrace{(X_i - \mu_n)}_{\mathbb{R}^{p\times 1}} \underbrace{(X_i - \mu_n)^T}_{\mathbb{R}^{1\times p}} \in \mathbb{R}^{p\times p}. \tag{*}$$

$$\Sigma_n = \frac{1}{n-1} C_n \qquad \text{Sample covariance matrix.}$$

Properties of C_n.

1. C_n is symmetric.

2.

$$C_n = U\Lambda U^T, \qquad U \in O(p).$$

U is an orthogonal $p \times p$ matrix, i.e.

$$UU^T = U^T U = I_{p\times p}.$$

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}, \qquad U = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_p \\ | & & | \end{bmatrix}.$$

$$\lambda_1 \geq \cdots \geq \lambda_p \geq 0.$$

$\{u_1, \ldots, u_p\}$ are orthonormal eigenvectors, i.e. C_n is positive semi-definite. This follows from the structure of C_n in $(*)$ (e.g. eigenvalues ≥ 0).

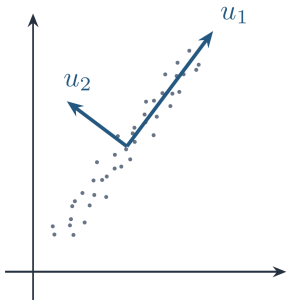
$$(\lambda_i, u_i)_{i=1}^p \text{ are the orthonormal eigenpairs of } C_n.$$

4.2.1 Understand Σ_n (or C_n)

$$u_i: \text{ direction of the } i\text{th largest variance of the data } \{X_1, \ldots, X_n\}.$$

$$\underbrace{\frac{\lambda_i}{n-1}}_{\text{eigenvalue of } \Sigma_n} : \quad i\text{th largest variance.}$$

$$p = 2, \qquad \{X_1, \ldots, X_n\} \subset \mathbb{R}^2$$



4.3 PCA Algorithm

Input: $\{X_1, \ldots, X_n\} \subset \mathbb{R}^p$.
# of principal components: k.
Step 1. Construct C_n using $\{X_1, \ldots, X_n\}$.
Step 2. Find the eigendecomp of C_n:

$$C_n = U\Lambda U^T.$$

$$U = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_p \\ | & & | \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}, \qquad \lambda_1 \geq \cdots \geq \lambda_p \geq 0.$$

Step 3. Project $X_i - \mu_n$ onto the first k eigenvectors.

$$\widetilde{U} = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_k \\ | & & | \end{bmatrix}.$$

Projection:

$$\widetilde{U}^T(X_i - \mu_n) = \widetilde{X}_i \in \mathbb{R}^k.$$

$$X_i \xrightarrow{\text{PCA}} \widetilde{X}_i = \begin{bmatrix} u_1^T(X_i - \mu_n) \\ \vdots \\ u_k^T(X_i - \mu_k) \end{bmatrix} \in \mathbb{R}^k.$$

4.4 In What Sense PCA Reorganizes the Data

PCA reduce the dimension of the data from $\mathbb{R}^p$ to $\mathbb{R}^k$.

$$\sum_{i=1}^{n} \widetilde{X}_i = \widetilde{U}^T \left(\underbrace{\sum_{i=1}^{n} (X_i - \mu_n)}_{=0} \right) = 0.$$

PCA reorganize the data in the following sense.

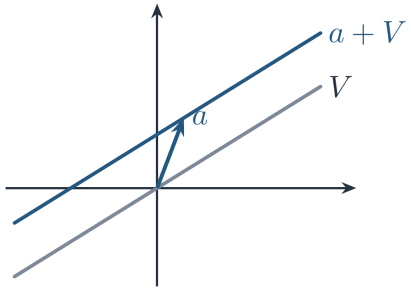
1. **Projection with maximum variability.** Project $\{X_1 - \mu_n, \ldots, X_n - \mu_n\}$ onto the principal components corresponding to the directions with k largest variances of the data.
2. It finds a k-dim affine subspace of $\mathbb{R}^p$ that best fits the data in the least square sense.

4.4.1 Recall Affine Space

V is a vector subspace of $\mathbb{R}^p$, $a \in \mathbb{R}^p$.

$$\text{affine subspace} \qquad a + V = \{a + v \mid v \in V\}.$$

$$\dim(a + V) = \dim(V).$$



$$\dim(V) = k, \qquad V \text{ is a subspace of } \mathbb{R}^p.$$

$\{\widetilde{u}_1, \ldots, \widetilde{u}_k\}$ form an orthonormal basis of V.

$$\beta_i \in \mathbb{R}^k, \qquad i = 1, \ldots, n.$$

$$\beta_i = \begin{bmatrix} \beta_{i,1} \\ \vdots \\ \beta_{i,k} \end{bmatrix} \in \mathbb{R}^k, \qquad \widetilde{U} = \begin{bmatrix} | & & | \\ \widetilde{u}_1 & \cdots & \widetilde{u}_k \\ | & & | \end{bmatrix}.$$

$$y_i = a + \widetilde{U}\beta_i = a + \sum_{j=1}^{k} \beta_{i,j}\widetilde{u}_j \in a + V.$$

4.4.2 Least Square Metric

Given $\{X_1, \ldots, X_n\} \subset \mathbb{R}^p$ and an affine subspace $a + V$ of $\mathbb{R}^p$,

$$\{y_1, \ldots, y_n\} \subset a + V.$$

$$\widetilde{U} = \begin{bmatrix} | & & | \\ \widetilde{u}_1 & \cdots & \widetilde{u}_k \\ | & & | \end{bmatrix}.$$

$\{\widetilde{u}_i\}$ are orthonormal basis of V. We can find $\beta_1, \ldots, \beta_n \in \mathbb{R}^k$ s.t.

$$y_i = a + \widetilde{U}\beta_i.$$

$$I\Big(\underbrace{a, \widetilde{U}}_{\substack{\text{determines the affine} \\ \text{subspace } a+V}}, \underbrace{\beta_1, \ldots, \beta_n}_{\text{determines } y_i \in a+V} \Big) = \sum_{i=1}^{n} \|X_i - y_i\|_{\mathbb{R}^p}^2$$

$$= \boxed{\sum_{i=1}^{n} \|X_i - a - \widetilde{U}\beta_i\|_{\mathbb{R}^p}^2}.$$

4.4.3 PCA Minimizes the Least Square Metric

Theorem 4.4.1. *Given $\{X_1, \ldots, X_n\} \subset \mathbb{R}^p$, let*

$$a = \mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i \in \mathbb{R}^p, \qquad \widetilde{U} = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_k \\ | & & | \end{bmatrix} \in \mathbb{R}^{p \times k}.$$

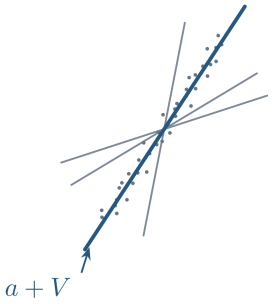
$u_1, \ldots, u_k$ are the orthonormal eigenvectors of C_n corresponding to the decreasing order of the eigenvalues.

$$\beta_i = \widetilde{X}_i \qquad \textit{(output of PCA).}$$

Then $I(a, \widetilde{U}, \beta_1, \ldots, \beta_n)$ is minimized.

I.e. PCA finds $y_1, \ldots, y_n$ in the affine space $\mu_n + V$ with coordinates $\beta_1, \ldots, \beta_n$ and V generated by $u_1, \ldots, u_k$ so that

$$\sum_{i=1}^{n} \|X_i - y_i\|^2 \text{ is minimized.}$$



4.5 Complexity of the PCA Algorithm

Step 1.

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad C_n = \sum_{i=1}^{n}(X_i - \mu_n)(X_i - \mu_n)^T.$$

Recall the Big O notation. Complexity (# of operations, e.g. addition and multiplication).

Complexity in step 1:

$$\underbrace{n}_{\mu_n} + \underbrace{p \cdot p \cdot n}_{C_n} \leq 2p^2 n = O(p^2 n).$$

Step 2. Eigendecomposition. Complexity $O(p^3)$.

Step 3. Projection. Complexity is dominated by the complexity in the first 2 steps.

Total complexity is

$$O(p^2 n + p^3).$$

4.6 Singular Value Decomposition (SVD)

$$M \in \mathbb{R}^{p \times n}.$$

$$\text{SVD:} \qquad M = USV^T.$$

$$\begin{aligned} U: &\quad p \times p \text{ orthogonal matrix,} \\ V: &\quad n \times n \text{ orthogonal matrix,} \\ &S \in \mathbb{R}^{p \times n}. \end{aligned}$$

$$S = \underbrace{\begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_{\min(p,n)} \\ & 0 & \end{bmatrix}}_{p \geq n} \quad \text{or} \quad \underbrace{\begin{bmatrix} \sigma_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & \sigma_{\min(p,n)} & \end{bmatrix}}_{p \leq n}.$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(p,n)} \geq 0.$$

They are singular values of M. Column vectors of U (or V) are the left (right) singular vectors.

$$Y = \begin{bmatrix} | & & | \\ X_1 - \mu_n & \cdots & X_n - \mu_n \\ | & & | \end{bmatrix} \in \mathbb{R}^{p \times n}.$$

Observation: $\boxed{C_n = YY^T}$.

SVD of Y:

$$Y = \widetilde{U}\Sigma\widetilde{V}^T.$$

$$\begin{aligned} C_n &= \widetilde{U}\Sigma\underbrace{\widetilde{V}^T\widetilde{V}}_{I_{n\times n}}\Sigma^T U^T \\ &= \widetilde{U}\Sigma\Sigma^T\widetilde{U}^T \\ &= U\Lambda U^T \qquad \text{eigendecomposition.} \end{aligned}$$

$$\implies \underbrace{\lambda_i}_{\text{eigenvalues of } C_n} = \sigma_i^2 \geq 0, \qquad \Lambda = \Sigma\Sigma^T, \qquad U = \widetilde{U}.$$

First k singular vectors of Y can be calculated in $O(pnk)$.

$$pnk \leq p^2 n + p^3.$$

Use SVD of Y to replace the first two steps of PCA.

Chapter 5

Vector spaces [3]

5.1 Introduction

The lecture begins by recalling principal component analysis (PCA) and Newton's method for equality-constrained minimization. We then review vector spaces and their bases, norms and inner products, matrix norms, and the dual space, following Chapter 1 of Beck's *First-Order Methods in Optimization* [3, Sections 1.1–1.4, 1.6–1.8, and 1.11]. Throughout these notes, scalars are real. Statements about a finite basis or the representation of every linear functional are made in the finite-dimensional setting.

5.2 Vector Spaces, Bases, and Dimension

Definition 5.2.1 (Real vector space)**.** A real vector space E is a nonempty set equipped with addition $E \times E \to E$ and scalar multiplication $\mathbb{R} \times E \to E$ satisfying the following properties. For all $x, y, z \in E$ and $\alpha, \beta \in \mathbb{R}$:

(i) Addition is commutative: $x + y = y + x$.

(ii) Addition is associative: $x + (y + z) = (x + y) + z$.

(iii) There is a zero vector $0 \in E$ such that $x + 0 = x$.

(iv) Each $x \in E$ has an additive inverse $-x \in E$ such that $x + (-x) = 0$.

(v) Scalar multiplication is associative: $\alpha(\beta x) = (\alpha\beta)x$.

(vi) Multiplication by one leaves a vector unchanged: $1x = x$.

(vii) Scalar multiplication distributes over vector addition: $\alpha(x + y) = \alpha x + \alpha y$.

(viii) Scalar multiplication distributes over scalar addition: $(\alpha + \beta)x = \alpha x + \beta x$.

The zero vector and the additive inverse of each vector are unique. For example, if 0 and $0'$ are both zero vectors, then $0 = 0 + 0' = 0'$. The familiar spaces $\mathbb{R}^n$ and $\mathbb{R}^{m \times n}$ are vector spaces under entrywise addition and scalar multiplication. In particular, a "vector" in an abstract vector space may itself be a matrix.

5.2.1 Linear independence and span

Definition 5.2.2 (Linear independence)**.** A finite family $v_1, \dots, v_k \in E$ is *linearly independent* if

$$\sum_{i=1}^{k} \alpha_i v_i = 0 \quad \Longrightarrow \quad \alpha_1 = \cdots = \alpha_k = 0.$$

Thus, the only linear combination of these vectors that equals zero is the trivial combination. Otherwise the family is linearly dependent.

Definition 5.2.3 (Span and basis)**.** The span of $v_1, \dots, v_k$ is

$$\operatorname{span}\{v_1, \dots, v_k\} = \left\{ \sum_{i=1}^{k} \beta_i v_i : \beta_1, \dots, \beta_k \in \mathbb{R} \right\}.$$

The vectors *span* E if every $x \in E$ can be written as $x = \sum_{i=1}^{k} \beta_i v_i$. A *basis* of E is a linearly independent family that spans E.

Spanning guarantees that a coordinate representation exists, while linear independence guarantees that it is unique. Indeed, if $\sum_i \alpha_i v_i = \sum_i \beta_i v_i$, then $\sum_i (\alpha_i - \beta_i) v_i = 0$, so $\alpha_i = \beta_i$ for every i. All bases of a finite-dimensional vector space have the same number of vectors. This number is the *dimension* of E, denoted $\dim(E)$.

Example 5.2.4 (Standard basis of $\mathbb{R}^n$)**.** Let e_i be the vector with a one in coordinate i and zeros elsewhere. Every $x = (x_1, \dots, x_n)^{\mathsf{T}}$ has the representation

$$x = \sum_{i=1}^{n} x_i e_i.$$

Moreover, $\sum_i \alpha_i e_i = 0$ implies $\alpha_i = 0$ by examining coordinate i. Hence $e_1, \dots, e_n$ form a basis and $\dim(\mathbb{R}^n) = n$.

Example 5.2.5 (Standard basis of $\mathbb{R}^{m \times n}$)**.** For $1 \le i \le m$ and $1 \le j \le n$, let E_{ij} denote the $m \times n$ matrix whose (i, j) entry is one and whose other entries are zero. Then

$$A = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} E_{ij}, \qquad A \in \mathbb{R}^{m \times n}.$$

If $\sum_{i,j} \alpha_{ij} E_{ij} = 0$, its (i, j) entry is α_{ij}, so every coefficient vanishes. Thus the mn matrices E_{ij} form a basis and

$$\dim\left(\mathbb{R}^{m \times n}\right) = mn.$$

5.3 Norms and Inner Products

5.3.1 Norms and balls

Definition 5.3.1 (Norm)**.** A norm on a real vector space E is a function $\|\cdot\| : E \to \mathbb{R}$ satisfying, for all $x, y \in E$ and $\lambda \in \mathbb{R}$:

(i) *Nonnegativity and definiteness:* $\|x\| \ge 0$, with $\|x\| = 0$ if and only if $x = 0$.

(ii) *Absolute homogeneity:* $\|\lambda x\| = |\lambda| \, \|x\|$.

(iii) *Triangle inequality:* $\|x + y\| \le \|x\| + \|y\|$.

A norm measures the size of a vector, and $\|x - y\|$ measures the distance between two vectors. For example, the Euclidean norm on $\mathbb{R}^n$ is

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}.$$

Given a center $c \in E$ and a radius $r > 0$, the open and closed balls in the chosen norm are, respectively,

$$B(c, r) = \{x \in E : \|x - c\| < r\}, \qquad B[c, r] = \{x \in E : \|x - c\| \le r\}.$$

Their shapes depend on the norm, so a ball need not look like a Euclidean sphere.

5.3.2 Inner products and their induced norms

Definition 5.3.2 (Real inner product)**.** An inner product on E is a function $\langle \cdot, \cdot \rangle : E \times E \to \mathbb{R}$ with the following properties:

(i) *Symmetry:* $\langle x, y \rangle = \langle y, x \rangle$.

(ii) *Linearity in the first argument:*

$$\langle \alpha_1 x_1 + \alpha_2 x_2, y \rangle = \alpha_1 \langle x_1, y \rangle + \alpha_2 \langle x_2, y \rangle.$$

(iii) *Positive definiteness:* $\langle x, x \rangle \geq 0$, with $\langle x, x \rangle = 0$ if and only if $x = 0$.

Here $x, y, x_1, x_2 \in E$ and $\alpha_1, \alpha_2 \in \mathbb{R}$. Symmetry and linearity in the first argument also give linearity in the second argument.

Example 5.3.3 (Euclidean inner product)**.** On $\mathbb{R}^n$, the usual dot product is

$$\langle x, y \rangle = x^{\mathsf{T}} y = \sum_{i=1}^{n} x_i y_i.$$

Its induced norm $\sqrt{\langle x, x \rangle}$ is the Euclidean norm $\|x\|_2$.

Proposition 5.3.4 (Norm induced by an inner product)**.** *Every real inner product induces a norm by*

$$\|x\| = \sqrt{\langle x, x \rangle}. \tag{5.1}$$

Proof. Nonnegativity and definiteness follow from the inner product axioms, and bilinearity gives $\|\lambda x\|^2 = \lambda^2 \|x\|^2$. To verify the triangle inequality, first note the Cauchy–Schwarz inequality

$$|\langle x, y \rangle| \leq \|x\| \, \|y\|.$$

For $y \neq 0$, it follows by expanding $0 \leq \|x - ty\|^2$ at $t = \langle x, y \rangle / \|y\|^2$; the case $y = 0$ is immediate. Consequently,

$$\|x + y\|^2 = \|x\|^2 + 2\langle x, y \rangle + \|y\|^2 \leq (\|x\| + \|y\|)^2.$$

Taking square roots proves the claim. □

Example 5.3.5 (Q-inner product and Q-norm)**.** Let $Q \in \mathbb{S}^n_{++}$, where $\mathbb{S}^n_{++}$ denotes the set of real symmetric positive definite $n \times n$ matrices. Thus $Q^{\mathsf{T}} = Q$ and $x^{\mathsf{T}} Q x > 0$ for every $x \neq 0$. Define

$$\langle x, y \rangle_Q = x^{\mathsf{T}} Q y, \qquad \|x\|_Q = \sqrt{x^{\mathsf{T}} Q x}.$$

Verifying that this is an inner product was an exercise in the lecture. Linearity follows from matrix multiplication, and symmetry follows from

$$x^{\mathsf{T}} Q y = (x^{\mathsf{T}} Q y)^{\mathsf{T}} = y^{\mathsf{T}} Q^{\mathsf{T}} x = y^{\mathsf{T}} Q x.$$

Positive definiteness follows from the assumption on Q. Equation (5.1) then gives the stated Q-norm. For $Q = I_n$, these reduce to the ordinary dot product and Euclidean norm.

5.3.3 The p-norms on $\mathbb{R}^n$

For $1 \le p < \infty$, the p-norm of $x \in \mathbb{R}^n$ is

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

In particular, $\|x\|_1 = \sum_i |x_i|$, and $p = 2$ gives the Euclidean norm. The restriction $p \ge 1$ is necessary for this formula to define a norm: when $0 < p < 1$, the triangle inequality fails in general. For $p \ge 1$, its triangle inequality is Minkowski's inequality. The infinity norm is

$$\|x\|_\infty = \max_{1 \le i \le n} |x_i|.$$

It is the limiting case of the p-norms:

$$\lim_{p \to \infty} \|x\|_p = \|x\|_\infty. \tag{5.2}$$

To see this, put $M = \max_i |x_i|$. At least one term in the sum is M^p, and all n terms are at most M^p. Therefore

$$M \le \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \le n^{1/p} M.$$

Since $n^{1/p} \to 1$, the squeeze theorem proves (5.2), including the case $x = 0$.

5.4 Norms on Matrix Spaces

5.4.1 The Frobenius inner product and norm

Because $\mathbb{R}^{m \times n}$ is a vector space, its elements can also be equipped with an inner product. For $A, B \in \mathbb{R}^{m \times n}$, define

$$\langle A, B \rangle_F = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij} = \operatorname{tr}(A^\mathsf{T} B). \tag{5.3}$$

Here $A^\mathsf{T} B$ is $n \times n$, and the trace is the sum of its diagonal entries. Indeed, $(A^\mathsf{T} B)_{jj} = \sum_i A_{ij} B_{ij}$, which proves the last equality. This is the ordinary Euclidean inner product applied to all mn matrix entries.

The associated norm is the *Frobenius norm*:

$$\|A\|_F = \sqrt{\langle A, A \rangle_F} = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2 \right)^{1/2} = \sqrt{\operatorname{tr}(A^\mathsf{T} A)}. \tag{5.4}$$

Thus $\operatorname{tr}(A^\mathsf{T} A) = \|A\|_F^2$; the square root is essential when writing the norm itself.

5.4.2 Induced matrix norms

An $m \times n$ matrix can also be viewed as a linear map $A : \mathbb{R}^n \to \mathbb{R}^m$, sending x to Ax. Suppose the domain has a norm $\|\cdot\|_a$ and the codomain has a norm $\|\cdot\|_b$.

Definition 5.4.1 (Induced matrix norm)**.** The matrix norm induced by these two vector norms is

$$\|A\|_{a,b} = \max_{\|x\|_a \le 1} \|Ax\|_b = \max_{\|x\|_a = 1} \|Ax\|_b = \sup_{x \ne 0} \frac{\|Ax\|_b}{\|x\|_a}. \tag{5.5}$$

The first subscript specifies the input norm, and the second specifies the output norm. In finite dimensions the closed unit ball is compact and $x \mapsto \|Ax\|_b$ is continuous, so the maximum exists. The equivalent forms in (5.5) follow by scaling any nonzero vector to have unit norm.

In particular, the definition gives

$$\|Ax\|_b \le \|A\|_{a,b}\,\|x\|_a \qquad \text{for every } x \in \mathbb{R}^n. \tag{5.6}$$

For $x \neq 0$, apply the definition to $x/\|x\|_a$ and multiply by $\|x\|_a$; for $x = 0$, both sides vanish. Thus the induced norm measures the largest possible amplification of a vector's size by the linear map.

It is also a norm on the vector space of matrices. Absolute homogeneity is immediate. If $A \neq 0$, some Ax is nonzero, so $\|A\|_{a,b} > 0$. Finally, the vector triangle inequality gives

$$\|(A+B)x\|_b \le \|Ax\|_b + \|Bx\|_b \le (\|A\|_{a,b} + \|B\|_{a,b})\|x\|_a,$$

and maximizing over the unit ball proves the matrix triangle inequality.

5.4.3 Spectral norm

When both vector norms are Euclidean, we obtain the *spectral norm*:

$$\|A\|_2 := \|A\|_{2,2} = \max_{\|x\|_2 \le 1} \|Ax\|_2 = \sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^{\mathsf{T}}A)}. \tag{5.7}$$

Here $\sigma_{\max}(A)$ is the largest singular value of A, and $\lambda_{\max}(A^{\mathsf{T}}A)$ is the largest eigenvalue of the symmetric positive semidefinite matrix $A^{\mathsf{T}}A$.

Verification of the spectral-norm formula (lecture exercise). For a unit vector x, $\|Ax\|_2^2 = x^{\mathsf{T}}A^{\mathsf{T}}Ax$. Write $A^{\mathsf{T}}A = U\Lambda U^{\mathsf{T}}$, with U orthogonal and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, where $\lambda_i \ge 0$. If $y = U^{\mathsf{T}}x$, then $\|y\|_2 = 1$ and

$$x^{\mathsf{T}}A^{\mathsf{T}}Ax = \sum_{i=1}^{n} \lambda_i y_i^2 \le \lambda_{\max} \sum_{i=1}^{n} y_i^2 = \lambda_{\max}.$$

Equality is attained at a unit eigenvector for $\lambda_{\max}$. Taking square roots gives (5.7). □

5.4.4 Maximum absolute column sum norm

For $a = b = 1$, the induced matrix norm is

$$\|A\|_1 := \|A\|_{1,1} = \max_{\|x\|_1 \le 1} \|Ax\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{m} |A_{ij}|. \tag{5.8}$$

Thus the matrix 1-norm is the largest absolute column sum.

Verification of the column-sum formula (lecture exercise). Let $c = \max_j \sum_i |A_{ij}|$. For every x,

$$\|Ax\|_1 = \sum_{i=1}^{m} \left| \sum_{j=1}^{n} A_{ij}x_j \right| \le \sum_{j=1}^{n} |x_j| \sum_{i=1}^{m} |A_{ij}| \le c\|x\|_1.$$

Hence $\|A\|_1 \le c$. Choose a column j_* attaining c and let $x = e_{j_*}$. Then $\|x\|_1 = 1$, and Ax is that column, so $\|Ax\|_1 = c$. This proves equality. □

5.4.5 Maximum absolute row sum norm

For $a = b = \infty$, the induced norm is

$$\|A\|_\infty := \|A\|_{\infty,\infty} = \max_{\|x\|_\infty \le 1} \|Ax\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |A_{ij}|. \tag{5.9}$$

Thus the matrix infinity norm is the largest absolute row sum.

Verification of the row-sum formula (lecture exercise). Let $r = \max_i \sum_j |A_{ij}|$. Then

$$\|Ax\|_\infty = \max_i \left| \sum_j A_{ij} x_j \right| \le \max_i \sum_j |A_{ij}|\,|x_j| \le r\|x\|_\infty.$$

For equality, choose a row i_* attaining r, and set $x_j = 1$ if $A_{i_* j} \ge 0$ and $x_j = -1$ otherwise. This choice has $\|x\|_\infty = 1$ and

$$(Ax)_{i_*} = \sum_{j=1}^{n} A_{i_* j} x_j = \sum_{j=1}^{n} |A_{i_* j}| = r.$$

Therefore $\|A\|_\infty = r$. □

Remark 5.4.2 (Distinguishing matrix norms)**.** The Frobenius norm measures the Euclidean size of all matrix entries, whereas the spectral norm measures the largest Euclidean amplification of a vector. For example, $\|I_n\|_F = \sqrt{n}$ but $\|I_n\|_2 = 1$. Also, the induced matrix 1-norm is a maximum column sum, not the sum of the absolute values of all entries.

5.5 The Dual Space

Definition 5.5.1 (Linear functional and dual space)**.** A linear functional on E is a linear map $f : E \to \mathbb{R}$, meaning

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y) \qquad (x, y \in E,\ \alpha, \beta \in \mathbb{R}).$$

The dual space is the set of all such functionals:

$$E^* = \{f : E \to \mathbb{R} : f \text{ is linear}\}.$$

With pointwise operations $(f+g)(x) = f(x) + g(x)$ and $(\alpha f)(x) = \alpha f(x)$, the set E^* is itself a vector space.

Lemma 5.5.2 (Representation of linear functionals)**.** *Let E be a finite-dimensional real inner product space. For every $f \in E^*$, there is a unique vector $v \in E$ such that*

$$f(x) = \langle v, x\rangle \qquad \textit{for every } x \in E. \tag{5.10}$$

Proof. Choose an orthonormal basis $u_1, \dots, u_d$ of E, where $d = \dim(E)$, and set

$$v = \sum_{i=1}^{d} f(u_i) u_i.$$

For $x = \sum_i x_i u_i$, linearity and orthonormality imply

$$f(x) = \sum_{i=1}^{d} x_i f(u_i) = \left\langle \sum_{i=1}^{d} f(u_i) u_i, \sum_{j=1}^{d} x_j u_j \right\rangle = \langle v, x\rangle.$$

This proves existence. If v and w represent the same functional, then $\langle v - w, x\rangle = 0$ for every x. Setting $x = v - w$ gives $\|v - w\|^2 = 0$, so $v = w$. □

For each $v \in E$, the map $J(v) : x \mapsto \langle v, x\rangle$ is a linear functional. The lemma shows that $J : E \to E^*$ is a bijection, and bilinearity shows that J is linear. Therefore

$$E \cong E^*, \qquad \dim(E^*) = \dim(E).$$

The board notation $E^* = E$ expresses this identification: vectors and linear functionals are different objects, but the chosen inner product provides a one-to-one linear correspondence between them. Finite dimensionality is part of the statement here; the same claim about all linear functionals does not hold for arbitrary infinite-dimensional inner product spaces.

Example 5.5.3 (Coordinate form of the representation)**.** On $\mathbb{R}^n$ with the standard dot product, every functional is $f(x) = a^{\mathsf{T}}x$, where $a = (f(e_1), \dots, f(e_n))^{\mathsf{T}}$. If instead we use the Q-inner product from Example 5.3.5, the representing vector is $v = Q^{-1}a$, since

$$\langle Q^{-1}a, x\rangle_Q = (Q^{-1}a)^{\mathsf{T}}Qx = a^{\mathsf{T}}x.$$

Thus the functional remains the same, but its representing vector depends on the chosen inner product.

Definition 5.5.4 (Dual norm)**.** Given a norm $\|\cdot\|$ on E, the dual norm on E^* is

$$\|\mathbf{y}\|_* := \underbrace{\sup_{\|\mathbf{x}\|\le 1} \langle \mathbf{y}, \mathbf{x}\rangle}_{(\text{all } \mathbf{x} \text{ in unit ball})} = \underbrace{\sup_{\|\mathbf{x}\|=1} \langle \mathbf{y}, \mathbf{x}\rangle}_{(\text{unit sphere})}$$

The two suprema will be the same.

Lemma 5.5.5. *Consider $p, q \ge 1$, satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Then on $\mathbb{R}^n$, the dual norm of the p-norm is the q-norm.*

Example 5.5.6.

$$\|\cdot\|_1 \overset{\text{dual norm}}{\longleftrightarrow} \|\cdot\|_\infty, \qquad \|\cdot\|_2 \overset{\text{dual norm}}{\longleftrightarrow} \|\cdot\|_2.$$

The dual norm of the Euclidean norm is itself: if E has $\|\cdot\|_2$, then E^* has $\|\cdot\|_2$.

Lemma 5.5.7 (Generalized Cauchy-Schwarz inequality)**.**

$$|\langle \mathbf{y}, \mathbf{x}\rangle| \le \|\mathbf{y}\|_* \cdot \|\mathbf{x}\|, \qquad \forall \mathbf{y} \in E^*,\ \mathbf{x} \in E.$$

Remark 5.5.8. If $\|\cdot\| = \|\cdot\|_2$, this is the Cauchy-Schwarz inequality

$$|\langle \mathbf{y}, \mathbf{x}\rangle| \le \|\mathbf{y}\|_2 \cdot \|\mathbf{x}\|_2.$$

Chapter 6

Closed functions and subgradient [3]

6.1 Extended Real-Valued Functions and Closedness

Our goal is to minimize some function, where E is the domain. We consider extended real-valued functions

$$f : E \to [-\infty, \infty],$$

where we allow the value ∞ in the range.

Definition 6.1.1 (Effective domain)**.**

$$\operatorname{dom}(f) := \{\mathbf{x} \in E : f(\mathbf{x}) < \infty\}.$$

Since we want to minimize, we do not allow the function to reach ∞ on its domain.

Example 6.1.2 (Indicator function)**.** Consider $C \subseteq E$. Then the indicator function of C is

$$\delta_C(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in C \\ \infty, & \mathbf{x} \notin C, \end{cases} \qquad \operatorname{dom}(\delta_C) = C.$$

This turns a constrained problem into an unconstrained one:

$$\begin{array}{l} \min\ f(\mathbf{x}) \\ \text{s.t. } \mathbf{x} \in C \end{array} \iff \min_{\mathbf{x} \in E}\ f(\mathbf{x}) + \delta_C(\mathbf{x}),$$

where the right-hand side is an unconstrained minimization problem. Both find the smallest value of $f(\mathbf{x})$ over $\mathbf{x} \in C$.

Definition 6.1.3 (Epigraph)**.**

$$\operatorname{epi}(f) = \{(\mathbf{x}, y)\ :\ f(\mathbf{x}) \le y,\ \mathbf{x} \in E,\ y \in \mathbb{R}\}.$$

The epigraph is the region that lies on or above the graph of $f(\mathbf{x})$.

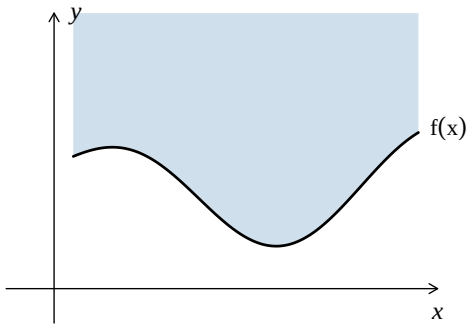


Figure 6.1: The epigraph of f is the region on or above the graph of f.

Definition 6.1.4 (Proper function)**.** f is called *proper* if there exists at least one $\mathbf{x} \in E$ such that $f(\mathbf{x}) < \infty$, i.e., $\operatorname{dom}(f) \neq \emptyset$. As long as the function is not $+\infty$ for every value, it is called proper.

Definition 6.1.5 (Closed function)**.** A function $f : E \to [-\infty, \infty]$ is *closed* if its epigraph is closed.

Example 6.1.6.

$$f(x) = \begin{cases} 0, & 1 < x < 2 \\ \infty, & \text{otherwise} \end{cases} = \delta_{(1,2)}(x).$$

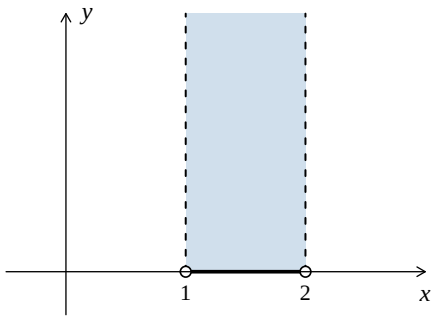


Figure 6.2: The epigraph of $\delta_{(1,2)}$. The dashed boundary lines are not included, so the epigraph is not closed.

Proposition 6.1.7. *The indicator function δ_C is closed if and only if C is a closed set.*

Proof.

$$\operatorname{epi}(\delta_C) = \{(\mathbf{x}, y) \ : \ \delta_C(\mathbf{x}) \leq y, \ \mathbf{x} \in E, \ y \in \mathbb{R}\} = C \times \mathbb{R}_+,$$

since $\mathbf{x}$ can take any value in C and y any nonnegative real value. Hence $\operatorname{epi}(\delta_C)$ is closed if and only if C is closed. □

Example 6.1.8. Let $f : \mathbb{R} \to [-\infty, \infty]$ be given by

$$f(x) = \begin{cases} \frac{1}{x}, & x > 0 \\ \infty, & \text{else} \end{cases}.$$

Then $\operatorname{epi}(f) = \left\{(x, y) \ : \ x > 0, \ \frac{1}{x} \leq y\right\}$ is closed, so f is a closed function. Note that $\operatorname{dom}(f) = (0, \infty)$ is open, so a closed function need not have a closed domain.

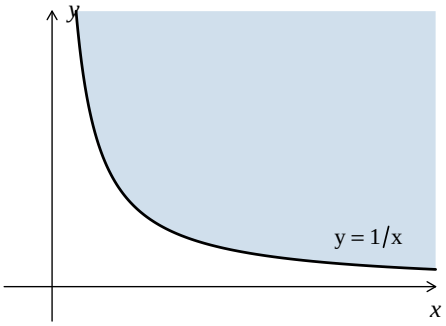


Figure 6.3: The epigraph of $f(x) = 1/x$.

Definition 6.1.9 (Lower semicontinuity)**.** A function $f : E \to [-\infty, \infty]$ is called *lower semicontinuous at* $\mathbf{x} \in E$ if

$$f(\mathbf{x}) \leq \liminf_{n\to\infty} f(\mathbf{x}_n)$$

for any sequence $\{\mathbf{x}_n\} \subset E$ with $\mathbf{x}_n \to \mathbf{x}$. It is *lower semicontinuous* if it is lower semicontinuous at every $\mathbf{x} \in E$.

Definition 6.1.10 (Upper semicontinuity)**.** A function $f : E \to [-\infty, \infty]$ is called *upper semicontinuous at* $\mathbf{x} \in E$ if

$$f(\mathbf{x}) \geq \liminf_{n\to\infty} f(\mathbf{x}_n)$$

for any sequence $\{\mathbf{x}_n\} \subset E$ with $\mathbf{x}_n \to \mathbf{x}$. It is *upper semicontinuous* if it is upper semicontinuous at every $\mathbf{x} \in E$.

Remark 6.1.11. A function is continuous if and only if it is both lower and upper semicontinuous.

Definition 6.1.12 (α-level set)**.**

$$\mathrm{Lev}(f, \alpha) = \{\mathbf{x} \in E \ : \ f(\mathbf{x}) \leq \alpha\}.$$

Theorem 6.1.13. *Let $f : E \to [-\infty, \infty]$. Then the following three statements are equivalent:*

(i) f is lower semicontinuous.

(ii) f is closed.

(iii) For any $\alpha \in \mathbb{R}$, $\mathrm{Lev}(f, \alpha)$ is closed.

Remark 6.1.14. Any continuous function is lower semicontinuous, so it is closed.

Theorem 6.1.15 (Operations preserving closedness)**.**

(a) Let $\{f_i\}_{i=1}^m$, $f_i : E \to (-\infty, \infty]$, be closed, and consider $\{\alpha_i\}_{i=1}^m \subset \mathbb{R}_+$. Then $f = \sum_{i=1}^m \alpha_i f_i$ is closed.

(b) Let $\{f_i\}_{i\in I}$, $f_i : E \to (-\infty, \infty]$, be closed. Then $f(\mathbf{x}) = \max_{i\in I} f_i(\mathbf{x})$ is closed.

(c) Let f be closed. Then $g(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b})$ is closed.

Remark 6.1.16. If f_1, f_2 are closed, then $\mathrm{epi}(\max\{f_1, f_2\}) = \mathrm{epi}(f_1) \cap \mathrm{epi}(f_2)$ is an intersection of closed sets, so it is closed.

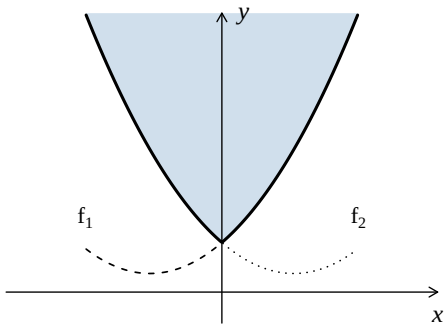


Figure 6.4: The epigraph of $\max\{f_1, f_2\}$.

6.2 Subgradient

Definition 6.2.1 (Subgradient)**.** Let $f : E \to (-\infty, \infty]$ be proper, and $\mathbf{x} \in \text{dom}(f)$. A vector $\mathbf{g} \in E^*$ is called a *subgradient* of f at $\mathbf{x}$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle, \qquad \forall \mathbf{y} \in E.$$

Remark 6.2.2. Recall: if f is convex and differentiable, then

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \qquad \forall \mathbf{y}.$$

Example 6.2.3. Let $f(x) = |x|$, $x \in \mathbb{R}$. Then f is not differentiable at $x = 0$. At $x = 0$, g is a subgradient if $\forall y$,

$$\begin{aligned} f(y) &\geq f(0) + g \cdot (y - 0) \\ &= 0 + g \cdot y \end{aligned}$$

so

$$|y| \geq g \cdot y, \qquad \forall y.$$

So we want to find some g such that this is true.

For example: $g = 0$, $g = 1$, and in general when $g \in [-1, 1]$.

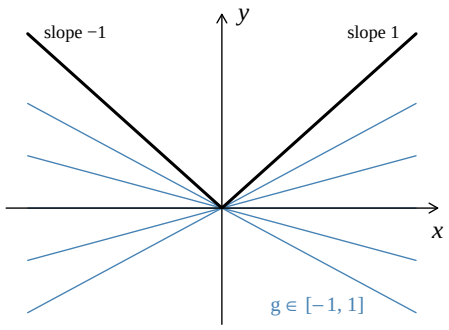


Figure 6.5: For all the possible slopes $g \in [-1, 1]$, the lines stay below $|x|$, so g is a subgradient of $|x|$ at 0.

Bibliography

[1] S. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[2] D. J. Foster and A. Rakhlin, *Foundations of reinforcement learning and interactive decision making*, arXiv preprint arXiv:2312.16730, 2023. Sections 2.2 and 2.3.

[3] A. Beck, *First-Order Methods in Optimization*, MOS-SIAM Series on Optimization, vol. 25, Society for Industrial and Applied Mathematics and Mathematical Optimization Society, Philadelphia, PA, 2017. doi: 10.1137/1.9781611974997.