# Chapter 0
# Prologue and Preliminaries

Ernest K. Ryu
Seoul National University

Mathematical Machine Learning Theory
Spring 2024

# Outline

# Motivating example: Binary classification

Consider the binary classification problem where we have grayscale images $X_1, \dots, X_N \in \mathcal{X} = \mathbb{R}^d$, where $d$ is the number of pixels, and corresponding labels $Y_1, \dots, Y_N \in \{-1, +1\}$. (E.g. labels indicate whether the image contains a cat or dog.)

Goal: Learn a function $f \colon \mathcal{X} \to \{-1, +1\}$ that approximately solves

$$\underset{f}{\text{minimize}} \quad \underbrace{\mathbb{P}_{(X,Y)\sim P}[f(X) \neq Y]}_{=\mathcal{R}[f]}.$$

I.e., minimize $\mathcal{R}[f] = (\text{probability of error of } f)$.

Implicit constraint: We must be able to implement $f$ on a computer.

# Direct empirical risk minimization

We do not have direct access to

$$\mathcal{R}[f] = \mathop{\mathbb{P}}_{(X,Y)\sim P}[f(X) \neq Y] = (\text{Probability of error}),$$

but we can use the approximation

$$\mathop{\mathbb{P}}_{(X,Y)\sim P}[f(X) \neq Y] = \mathop{\mathbb{E}}_{(X,Y)\sim P}\left[\mathbf{1}_{\{f(X)\neq Y\}}\right]$$

$$\overset{?}{\approx} \underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{\{f(X_i)\neq Y_i\}}}_{=\hat{\mathcal{R}}[f]} = (\%\text{ of error on training data}).$$

We call $\mathcal{R}[f]$ the *population risk* or the *true risk* and call $\hat{\mathcal{R}}[f]$ the *empirical risk*. Intuitively, based on the law of large numbers, we expect

$$\hat{\mathcal{R}}[f] \overset{?}{\approx} \mathcal{R}[f]$$

when the *sample size* $N$ is large. The $\overset{?}{\approx}$ will be precisely defined and rigorously justified in this course.

# Direct empirical risk minimization

A seemingly straightforward approach is

$$\underset{f}{\text{minimize}} \quad \underbrace{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{f(X_i)\neq Y_i\}}}_{=\hat{\mathcal{R}}[f]}.$$

I.e., minimize (% of error on training data of $f$). This is used in some setups, such as tree-based learning methods.

However, this combinatorial optimization problem is often difficult (NP-hard) to solve exactly, and it's not the focus of modern machine learning research.[1]

[1] One can argue that tree-based methods such as XGBoost are most common in current machine learning *practice* of the industry. However, methods based on continuous optimization of surrogate losses are more mainstream in current machine learning *research*, and it is the basis of modern deep learning.

# Surrogate loss

Instead, use a *surrogate loss* and solve the resulting continuous optimization problem.

Let $f(x) = \operatorname{sign}(g(x))$, where $g\colon \mathcal{X} \to \mathbb{R}$. (For the sake of concreteness, define $\operatorname{sign}(0) = 0$, although this specific choice does not matter.)

Then (no surrogate yet),

$$\begin{aligned}
\mathcal{R}[g] &= \mathop{\mathbb{P}}_{(X,Y)\sim P}[\operatorname{sign}(g(X)) \neq Y] \\
&= \mathop{\mathbb{P}}_{(X,Y)\sim P}[\operatorname{sign}(Yg(X)) \neq 1] \\
&= \mathop{\mathbb{E}}_{(X,Y)\sim P}[\mathbf{1}_{\{\operatorname{sign}(Yg(X))\neq 1\}}] \\
&= \mathop{\mathbb{E}}_{(X,Y)\sim P}[\Phi^{0\text{-}1}(Yg(X))], \qquad \Phi^{0\text{-}1}(u) = \begin{cases} 1 & \text{for } u < 0 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

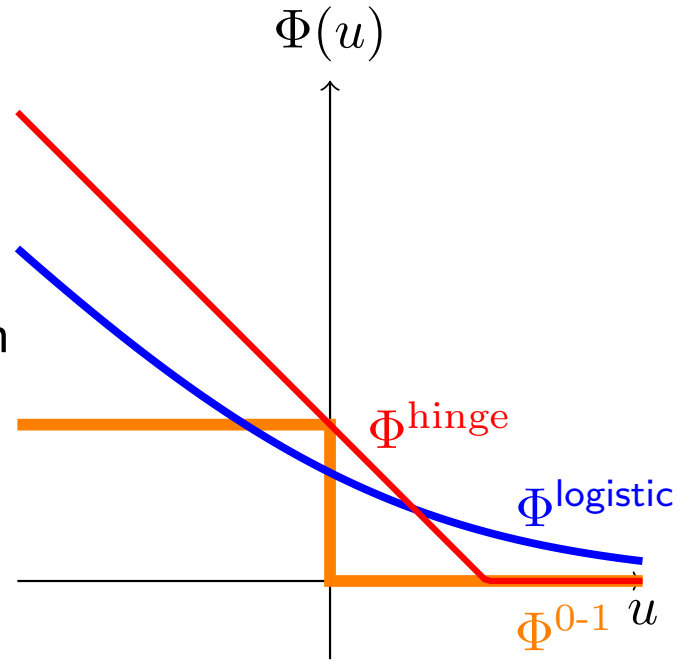
Since $\mathcal{R}[f] = \mathcal{R}[\operatorname{sign} \circ g] = \mathcal{R}[g]$, this is a bit of abuse of notation.

# Surrogate loss

Replace $\Phi^{0\text{-}1}(u)$ with a surrogate loss such as

$$\Phi^{\text{hinge}}(u) = \max\{1-u, 0\}$$
$$\Phi^{\text{logistic}}(u) = \log(1+e^{-u}),$$

which are nice continuous, convex functions.




Finally, solve the continuous optimization problem

$$\underset{g}{\text{minimize}} \quad \underbrace{\mathop{\mathbb{E}}_{(X,Y)\sim P}[\Phi(Yg(X))]}_{=\mathcal{R}_\Phi[g]}$$

or its approximation

$$\underset{g}{\text{minimize}} \quad \underbrace{\frac{1}{N}\sum_{i=1}^{N}\Phi(Y_i g(X_i))}_{=\hat{\mathcal{R}}_\Phi[g]}.$$

# Minimize surrogate loss $\stackrel{?}{\Rightarrow}$ Minimize original loss

Do not forget that we have changed the optimization problem from minimizing $\mathcal{R}$ to $\mathcal{R}_\Phi$ to $\hat{\mathcal{R}}_\Phi$.

Is this a valid approach? We will rigorously establish this later.

The surrogate losses majorize the 0-1 loss, i.e., $\Phi^{\text{hinge}} \geq \Phi^{\text{0-1}}$ or $\frac{1}{\log(2)}\Phi^{\text{logistic}} \geq \Phi^{\text{0-1}}$, and this is an important observation. Therefore, $\mathcal{R}_\Phi[g] \approx 0 \Rightarrow \mathcal{R}[g] \approx 0$, provided that the original and surrogate losses are nonnegative.

However, if

$$\mathcal{R}_\Phi[g_\star^{\text{surr}}] = \inf_g \mathcal{R}_\Phi[g] > 0,$$

then it is not immediately clear whether

$$\mathcal{R}[g_\star^{\text{surr}}] \stackrel{?}{\approx} \inf_g \mathcal{R}[g].$$

In this case, we need further justification for using the surrogate loss.

# Measures of performance

We want to prove that our ML method is good.
How do we quantify "good"?

If we can find a $g$ such that

$$\mathcal{R}[g] - \mathcal{R}^\star < \text{small}$$

where $\mathcal{R}^\star$ is some suitable baseline, then $g$ will be a good function, and the machine learning task has succeeded.

However, our $g$ is itself random, as it depends on that data $X_1, \ldots, X_N, Y_1, \ldots, Y_N$ and the randomness of the training algorithm (perhaps SGD). Therefore, a statement about the goodness of $g$ must be probabilistic sense.

# Measures of performance: Expected error

One approach is to show that the *expected error* is small:

$$\mathop{\mathbb{E}}_{g}\left[\mathcal{R}[g] - \mathcal{R}^{\star}\right] < \text{small}.$$

Interpretation: $g$ is good in expectation.

We will mostly focus on this measure of performance in this course.

# Measures of performance: PAC

Another approach is to show that the *probably approximately correct* (PAC) result:

$$\mathop{\mathbb{P}}_{g}\left(\mathcal{R}[g] - \mathcal{R}^\star < \text{small}\right) > 1 - \text{small}$$

Interpretation: $g$ is good with high probability. Arguably, this better aligns with our intuitive notion of what a "good" ML method is.

PAC is a weaker notation (easier to establish) than expected error:

- A bound on expected error implies PAC by Markov inequality.
- If an algorithm catastrophically fails with a small probability, then the expected error will be large, but PAC may hold.

# Measures of performance: Data, memory, and compute complexities

In either

$$\mathop{\mathbb{E}}_{g}\left[\mathcal{R}[g] - \mathcal{R}^\star\right] < \text{small}$$

or

$$\mathop{\mathbb{P}}_{g}\left(\mathcal{R}[g] - \mathcal{R}^\star < \text{small}\right) > 1 - \text{small},$$

"small" will be formally quantified with $\varepsilon$'s and $\delta$'s.

The "small" will (usually) be a decreasing function of the expended resources: data, memory, and compute.[2]

[2] In modern ML vernacular, "compute" is used as a noun to mean "computation", "computing power", or "flops".

# Outline

# Norm and dual norm

We say $\|\cdot\|$ is a *norm* on $\mathbb{R}^d$ if $\|x\| \in [0, \infty)$ for all $x \in \mathbb{R}^d$ (so the output of $\|x\|$ is a finite non-negative scalar), and

- $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^d$,
- $\|\alpha x\| = |\alpha| \|x\|$ for all $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^d$, and
- $\|x\| = 0$ if and only if $x = 0$.

Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. The *dual norm* $\|\cdot\|_*$ is defined as

$$\|y\|_* = \sup\{y^\mathsf{T} x \,|\, x \in \mathbb{R}^d, \|x\| \leq 1\}$$

for all $y \in \mathbb{R}^d$. The dual norm is a norm on $\mathbb{R}^d$.

If $p, q \in [1, \infty]$ and $1/p + 1/q = 1$, then the $p$-norm and the $q$-norm are duals of each other, i.e.,

$$(\|\cdot\|_p)_* = \|\cdot\|_q.$$

In particular, the Euclidean norm $\|\cdot\|_2$ is self-dual.

# Singular value decomposition (SVD)

Let $A \in \mathbb{R}^{m \times n}$. The singular value decomposition (SVD) of $A$ has the form

$$A = \underbrace{\begin{bmatrix} u_1 & \cdots & u_m \end{bmatrix}}_{\in \mathbb{R}^{m \times m}} \underbrace{\begin{bmatrix} \sigma_1 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \sigma_r & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}}_{\in \mathbb{R}^{m \times n}} \underbrace{\begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix}}_{\in \mathbb{R}^{n \times n}}.$$

where $u_1, \ldots, u_m \in \mathbb{R}^m$ is an orthonormal basis of $\mathbb{R}^m$, $r$ is the rank of $A$, $\sigma_1, \ldots, \sigma_r > 0$, and $v_1, \ldots, v_n \in \mathbb{R}^n$ is an orthonormal basis of $\mathbb{R}^n$. The SVD always exists and can be computed.[3]

[3] In comparison, the eigenvalue decomposition only exists when the matrix is diagonalizable. Also, the Jordan canonical form cannot be computed stably on a computer with finite-precision arithmetic.

# Compact SVD

Let $A \in \mathbb{R}^{m \times n}$. The compact SVD of $A$ has the form

$$A = \underbrace{\begin{bmatrix} u_1 & \cdots & u_r \end{bmatrix}}_{=U \in \mathbb{R}^{m \times r}} \underbrace{\operatorname{diag}(\sigma_1, \ldots, \sigma_r)}_{=\Sigma \in \mathbb{R}^{r \times r}} \underbrace{\begin{bmatrix} v_1^\intercal \\ \vdots \\ v_r^\intercal \end{bmatrix}}_{=V^\intercal \in \mathbb{R}^{r \times n}},$$

where $u_1, \ldots, u_r \in \mathbb{R}^m$ is an orthonormal set of vectors, $r$ is the rank of $A$, $\sigma_1, \ldots, \sigma_r > 0$, and $v_1, \ldots, v_r \in \mathbb{R}^n$ is an orthonormal set of vectors.

Note, $U$ and $V$ are not orthogonal matrices despite containing orthonormal columns. So

$$U^\intercal U = I, \qquad UU^\intercal \neq I, \qquad V^\intercal V = I, \qquad VV^\intercal \neq I.$$

# Moore–Penrose inverse pseudo-inverse

The (Moore–Penrose) pseudo-inverse is as defined as follows:
Let $A \in \mathbb{R}^{m \times n}$ have rank $r$ and let

$$A = \underbrace{\begin{bmatrix} u_1, \dots, u_r \end{bmatrix}}_{=U} \underbrace{\operatorname{diag}(\sigma_1, \dots, \sigma_r)}_{=\Sigma} \underbrace{\begin{bmatrix} v_1^\intercal \\ \vdots \\ v_r^\intercal \end{bmatrix}}_{=V^\intercal}$$

be the compact SVD of $A$. Then, the pseudo-inverse is defined as

$$A^\dagger = \begin{bmatrix} v_1, \dots, v_r \end{bmatrix} \operatorname{diag}(\sigma_1^{-1}, \dots, \sigma_r^{-1}) \begin{bmatrix} u_1^\intercal \\ \vdots \\ u_r^\intercal \end{bmatrix} = V \Sigma^{-1} U^\intercal.$$

# Abstract definition of pseudo-inverse

Alternatively, we can define $A^\dagger$ to be the linear operator such that

$$A^\dagger\big|_{\mathcal{R}(A)^\perp} = 0, \qquad A^\dagger\big|_{\mathcal{R}(A)} = \left(A\big|_{\mathcal{N}(A)^\perp \to \mathcal{R}(A)}\right)^{-1}.$$

To clarify, by restricting $A$'s domain and range as

$$A\big|_{\mathcal{N}(A)^\perp \to \mathcal{R}(A)} : \mathcal{N}(A)^\perp \to \mathcal{R}(A)$$

we have a bijection and, therefore, invertible linear mapping.

# Unitary matrices and unitary invariance of $\|\cdot\|_2$

We say $U \in \mathbb{R}^{n\times n}$ is *unitary* or *orthogonal* if any one of the following equivalent conditions hold:

- $U^\intercal U = I$.
- $UU^\intercal = I$.
- The columns of $U$ form an orthonormal basis of $\mathbb{R}^n$.
- The rows of $U$ form an orthonormal basis of $\mathbb{R}^n$.

The Euclidean norm $\|\cdot\|_2$ is unitarily invariant, i.e., if $U \in \mathbb{R}^{n\times n}$ is unitary, then

$$\|Ux\|_2 = \|x\|_2 \qquad \forall x \in \mathbb{R}^n.$$

# Positive definite matrices

We say a symmetric matrix $A \in \mathbb{R}^{n \times n}$ is positive definite if its eigenvalues are all strictly positive and positive semidefinite if its eigenvalues are all nonnegative. (Note that symmetric real matrices necessarily have real eigenvalues.) (In this class, we will not refer to asymmetric matrices as positive definite or semidefinite.)

For a symmetric $A \in \mathbb{R}^{n \times n}$, write

$$A \succ 0 \quad \text{and} \quad A \succeq 0$$

to respectively denote that $A$ is positive definite and positive semidefinite.

For symmetric $A, B \in \mathbb{R}^{n \times n}$, we use the notation

$$A \succ B \quad \Leftrightarrow \quad (A - B) \succ 0$$

and

$$A \succeq B \quad \Leftrightarrow \quad (A - B) \succeq 0.$$

# Cholesky factorization and matrix square root

Let $A \in \mathbb{R}^{n \times n}$ be symmetric positive semidefinite. Then there is a lower triangular $L \in \mathbb{R}^{n \times n}$ such that

$$LL^\intercal = A.$$

This is the *Cholesky factorization*, and it can be computed efficiently.

Let $A \in \mathbb{R}^{n \times n}$ be symmetric positive semidefinite. If we let $A = U\mathrm{diag}(\lambda_1, \ldots, \lambda_n)U^\intercal$ be the eigenvalue decomposition of $A$, then

$$A^{1/2} = U\mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})U^\intercal$$

is called the matrix square root of $A$. Of course, $A^{1/2}$ is itself a symmetric positive semidefinite matrix, and

$$A^{1/2}(A^{1/2})^\intercal = A^{1/2}A^{1/2} = A.$$

We will encounter several instances where given a symmetric positive semidefinite $A$, we need a $B$ such that $BB^\intercal = A$. Theoretically, one can choose either the Cholesky factorization or the matrix square root. (Or something else, since there are many other choices.) Computationally, the Cholesky is usually the best option.

# Matrix inversion lemma

Let $A \in \mathbb{R}^{n \times n}$ and $D \in \mathbb{R}^{m \times m}$ be invertible. Let $B \in \mathbb{R}^{n \times m}$ and $C \in \mathbb{R}^{m \times n}$. Then, the *Sherman–Morrison–Woodbury* matrix inversion lemma states:

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}.$$

Often applied with $A$ and/or $B$ equal to a scalar multiple of identity.

# Convexity

We say a set $C \subseteq \mathbb{R}^d$ is *convex* if

$$\theta x + (1-\theta)y \in C, \qquad \forall\, x, y \in C,\, \theta \in (0,1).$$

In other words, the line segment connecting $x$ and $y$ is contained in $C$.

We say a function $f : \mathbb{R}^d \to \mathbb{R}$ is *convex* if

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y), \qquad \forall\, x, y \in \mathbb{R},\, \theta \in [0,1].$$

In other words, the line segment connecting $(x, f(x))$ and $(y, f(y))$ is above the graph of $f$. Finite-valued convex functions are continuous. Convex functions are not necessarily differentiable.

# Jensen's inequality

## Lemma (Jensen's inequality)

*Let $X \in \mathbb{R}^d$ be a random variable such that $\mathbb{E}[X] \in \mathbb{R}^d$ is well defined, and let $\varphi \colon \mathbb{R}^d \to \mathbb{R}$ be convex. Then.*

$$\varphi\big(\mathbb{E}[X]\big) \leq \mathbb{E}[\varphi(X)].$$

*(The LHS is assumed to be finite, but the RHS can take value $\infty$.)*

**Proof.** Immediate from the notion of subgradients and the linearity of expectation. We will revisit the proof later. □

# Outline

# Markov inequality

## Theorem

*Let $X \in \mathbb{R}$ be a nonnegative random variable. For any $\varepsilon > 0$,*

$$\mathbb{P}(X \geq \varepsilon) \leq \frac{1}{\varepsilon}\mathbb{E}[X].$$

**Proof.** Using

$$\mathbf{1}_{\{X \geq \varepsilon\}} \leq \frac{1}{\varepsilon}X,$$

we conclude

$$\mathbb{P}(X \geq \varepsilon) = \mathbb{E}[\mathbf{1}_{\{X \geq \varepsilon\}}] \leq \mathbb{E}\left[\frac{1}{\varepsilon}X\right].$$

□

# Chebyshev inequality

## Corollary

*Let $X_1, \ldots, X_N$ be a IID scalar random variables with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma \in \mathbb{R}$. Let*

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i.$$

*For any $\varepsilon > 0$,*

$$\mathbb{P}\big(|\bar{X} - \mu| \geq \varepsilon\big) \leq \frac{\sigma^2}{N\varepsilon^2}.$$

## Proof.

By the Markov inequality,

$$\mathbb{P}\big(|\bar{X} - \mu| \geq \varepsilon\big) = \mathbb{P}\big((\bar{X} - \mu)^2 \geq \varepsilon^2\big) \leq \frac{\sigma^2/N}{\varepsilon^2}.$$

□

# Concentration: $\bar{X} \approx \mu$ when $N$ is large?

Can we bound the deviation of $\bar{X}$ from its mean $\mu$? I.e., can we show that $\bar{X}$ *concentrates* around $\mu$?

Without assumptions, we cannot. Even as $N \to \infty$. (E.g. Cauchy.)

If we assume $\sigma < \infty$, Chebyshev applies and shows a concentration result.

When $X$ is bounded, a stronger condition, Hoeffding applies and provides a stronger concentration result.

# Hoeffding inequality

## Theorem

*Let* $X_1, \ldots, X_N \in [0,1]$ *be independent random variables with means* $\mu_1, \ldots, \mu_N \in \mathbb{R}$. *Let* $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ *and* $\bar{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mu_i$. *For any* $\varepsilon > 0$,

$$\mathbb{P}(\bar{X} - \bar{\mu} \geq \varepsilon) \leq \exp(-2N\varepsilon^2).$$

The exponential dependence on $N$ is much stronger (better) than the $\mathcal{O}(1/N)$ dependence of Chebyshev.

# Two-sided Hoeffding inequality

## Corollary

*Let $X_1, \dots, X_N \in [0,1]$ be independent random variables with means $\mu_1, \dots, \mu_N \in \mathbb{R}$. Let $\bar{X} = \frac{1}{N}\sum_{i=1}^N X_i$ and $\bar{\mu} = \frac{1}{N}\sum_{i=1}^N \mu_i$. For any $\varepsilon > 0$,*

$$\mathbb{P}(|\bar{X} - \bar{\mu}| \geq \varepsilon) \leq 2\exp(-2N\varepsilon^2).$$

**Proof.** Decompose the deviation into two events and use one-sided Hoeffding twice with $X_i, (1 - X_i) \in [0,1]$:

$$\begin{aligned}
\mathbb{P}(|\bar{X} - \bar{\mu}| \geq \varepsilon) &= \mathbb{P}(\bar{X} - \bar{\mu} \geq \varepsilon) + \mathbb{P}(\bar{X} - \bar{\mu} \leq -\varepsilon) \\
&= \mathbb{P}(\bar{X} - \bar{\mu} \geq \varepsilon) + \mathbb{P}(-\bar{X} + \bar{\mu} \geq \varepsilon) \\
&= \mathbb{P}(\bar{X} - \bar{\mu} \geq \varepsilon) + \mathbb{P}((1 - \bar{X}) - (1 - \bar{\mu}) \geq \varepsilon) \\
&\leq \exp(-2N\varepsilon^2) + \exp(-2N\varepsilon^2).
\end{aligned}$$

□

# Light and heavy tails

We say a scalar-valued random variable $X$ is *light-tailed* if

$$\mathbb{P}\big(|X| > \text{large}\big) \leq \text{small}$$

and *heavy-tailed* if

$$\mathbb{P}\big(|X| > \text{large}\big) \geq \text{not small}$$

where "small" and "large" depends on the context.[4] Light-tailed random variables are much easier to control and will allow us to get good rates, so we exclude heavy-tailed random variables with appropriate assumptions.

Cauchy RVs with density

$$f(x) = \frac{1}{\pi(1+x^2)}$$

and tail probability

$$\mathbb{P}(X \geq x) \sim \frac{1}{\pi x} \qquad \text{as } x \to \infty$$

are heavy-tailed.

[4] In probability theory, there is a more quantitative definition of light- and heavy-tailedness. In this course, this informal definition will suffice.

# Light and heavy tails

Gaussian RVs with density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

and tail probability

$$\mathbb{P}(X \geq x) \sim \frac{1}{\sqrt{2\pi}x} e^{-\frac{x^2}{2}} \qquad \text{as } x \to \infty$$

are light-tailed.

Bounded RVs

$$X \in [a, b] \qquad \text{for some } -\infty < a \leq b < \infty$$

have tail probability

$$\mathbb{P}(X \geq x) = 0 \qquad \text{as } x \to \infty,$$

so they are light-tailed.

# Chebyshev vs. Hoeffding

## Theorem (More general Chebyshev)

*Let $X_1, \ldots, X_N$ be IID .... For $p \geq 2$, there is some $C_{\varepsilon,p} > 0$ such that*

$$\mathbb{P}(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{C_{\varepsilon,p}\mathbb{E}[|X_1 - \mu|^p]}{N^{p/2}}.$$

## Theorem (Hoeffding)

*Let $X_1, \ldots, X_N \in [0,1]$ be IID .... Then, there is some $C_\varepsilon > 0$ such that*

$$\mathbb{P}(|\bar{X} - \mu| \geq \varepsilon) \leq \exp(-C_\varepsilon N).$$

Hoeffding is substantially stronger for large $N$.

The main difference is that Chebyshev applies to light- and heavy-tailed RVs, while Hoeffding applies to only light-tailed RVs.

Soon, we generalize Hoeffding to sub-Gaussian random variables.

# Hoeffding lemma

We now proceed with the proof of the Hoeffding inequality.

## Lemma

*Let $Z \in [0,1]$ be a random variable. For any $s \geq 0$,*

$$\mathbb{E}[e^{s(Z-\mathbb{E}[Z])}] \leq e^{s^2/8}.$$

**Proof.** We first consider the cumulant-generating function

$$\varphi(s) = \log \mathbb{E}[e^{s(Z-\mathbb{E}[Z])}].$$

Clearly, $\varphi(0) = 0$. We will show that $\varphi'(0) = 0$ and $\varphi''(s) \leq 1/4$ for all $s \geq 0$. This implies $\varphi'(s) \leq s/4$ and $\varphi(s) \leq s^2/8$, and we conclude the statement.

Let $f$ be the density function[5] of $Z$, and let $\tilde{Z} \in [0,1]$ is a random variable with density

$$\frac{e^{s(z-\mathbb{E}[Z])} f(z)}{\int_0^1 e^{s(w-\mathbb{E}[Z])} f(w)\, dw} \qquad \text{for } z \in [0,1].$$

Then,[6]

$$\begin{aligned}
\varphi'(s) &= \frac{\mathbb{E}[(Z-\mathbb{E}[Z])e^{s(Z-\mathbb{E}[Z])}]}{\mathbb{E}[e^{s(Z-\mathbb{E}[Z])}]} = \mathbb{E}[\tilde{Z}-\mathbb{E}[Z]] \\
\varphi''(s) &= \frac{\mathbb{E}[(Z-\mathbb{E}[Z])^2 e^{s(Z-\mathbb{E}[Z])}]}{\mathbb{E}[e^{s(Z-\mathbb{E}[Z])}]} - \left(\frac{\mathbb{E}[(Z-\mathbb{E}[Z])e^{s(Z-\mathbb{E}[Z])}]}{\mathbb{E}[e^{s(Z-\mathbb{E}[Z])}]}\right)^2 \\
&= \operatorname{Var}(\tilde{Z}-\mathbb{E}[Z]) = \operatorname{Var}(\tilde{Z}).
\end{aligned}$$

We see that $\varphi'(0) = 0$. Since

$$\varphi''(s) = \operatorname{Var}(\tilde{Z}) = \operatorname{Var}(\tilde{Z}-1/2) \le \mathbb{E}[(\tilde{Z}-1/2)^2] \le \frac{1}{4}$$

since $(\tilde{Z}-1/2)^2 \le 1/4$ for all $\tilde{Z} \in [0,1]$. □

[5] If $Z$ is not a continuous RV we can use the Radon–Nikodym derivative.

[6] To be rigorous, one should justify differentiating under the integral, i.e. one should justify $\frac{d}{ds}\,\mathbb{E}_Z[\cdots] = \mathbb{E}_Z[\frac{d}{ds}\cdots]$, with a Lebesgue DCT argument.

# Proof of Hoeffding inequality

Using the Hoeffding Lemma, we now prove the Hoeffding inequality.

**Proof.**

$$\begin{aligned}
\mathbb{P}(\bar{X} - \bar{\mu} \geq \varepsilon) &= \mathbb{P}(e^{s(\bar{X}-\bar{\mu})} \geq e^{s\varepsilon}) \qquad \text{for } s > 0 \quad \text{(monotonicity of exponential)} \\
&\leq e^{-s\varepsilon}\mathbb{E}[e^{s(\bar{X}-\bar{\mu})}] \qquad \text{(Markov ineq.)} \\
&= e^{-s\varepsilon}\prod_{i=1}^{N}\mathbb{E}[e^{\frac{s}{N}(X_i-\mu_i)}] \qquad \text{(independence)} \\
&\leq e^{-s\varepsilon}\prod_{i=1}^{N}e^{\frac{s^2}{8N^2}} \qquad \text{(Hoeffding Lem.)} \\
&= e^{-s\varepsilon+\frac{s^2}{8N}}
\end{aligned}$$

Finally, we plug in $s = 4N\varepsilon$, the minimizer of the bound, to conclude

$$\mathbb{P}(\bar{X} - \bar{\mu} \geq \varepsilon) \leq e^{-2N\varepsilon^2}.$$

□

# Hoeffding inequality with non-uniform bounds

Hoeffding's bound $X_i \in [0,1]$ is arbitrary, and it can be generalized.

## Theorem

*Let $X_1, \dots, X_N \in \mathbb{R}$ be independent random variables such that $|X_i| \leq c_i < \infty$ almost surely and $\mathbb{E}[X_i] = \mu_i \in \mathbb{R}$ for $i = 1, \dots, N$. Let $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ and $\bar{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mu_i$. For any $\varepsilon > 0$,*

$$\mathbb{P}(\bar{X} - \bar{\mu} \geq \varepsilon) \leq \exp\left(-\frac{N^2\varepsilon^2}{2(c_1^2 + \cdots + c_N^2)}\right).$$

**Proof.** Follow the same reasoning as the regular Hoeffding inequality. □

# Hoeffding for martingales (Azuma)

The independence assumption of Hoeffding inequality can be relaxed to a martingale difference assumption.

## Theorem (Azuma's inequality)

*Let $X_1, \ldots, X_N \in \mathbb{R}$ be a martingale difference sequence, i.e.,*

$$\mathbb{E}[X_i \,|\, X_1, \ldots, X_{i-1}] = 0 \qquad \text{for } i = 1, \ldots, N,$$

*such that $|X_i| \leq c_i < \infty$ for $i = 1, \ldots, N$. (The martingale difference condition implies $X_1, \ldots, X_N$ have zero mean.) Let $\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$. For any $\varepsilon > 0$,*

$$\mathbb{P}(\bar{X} \geq \varepsilon) \leq \exp\left(-\frac{N^2 \varepsilon^2}{2(c_1^2 + \cdots + c_N^2)}\right).$$

**Proof.** Follow the same reasoning as the regular Hoeffding inequality. □

---

In probability theory, whenever a result holds for sums of independent random variables, there is a good chance that the result can be generalized to sums of martingale difference sequences, i.e., martingales.

# Sub-Gaussian random variables

A random variable $X \in \mathbb{R}$ is *sub-Gaussian* with constant $\tau \geq 0$ if

$$\mathbb{E}[e^{s(X-\mathbb{E}[X])}] \leq e^{\tau^2 s^2/2} \qquad \forall\, s \in \mathbb{R}.$$

## Fact

*If $X$ is sub-Gaussian with constant $\tau$, then $sX$ is sub-Gaussian with constant $|s|\tau$ for any $s \in \mathbb{R}$.*

**Proof.** Check definition. □

## Fact

*If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $X$ is sub-Gaussian with constant $\tau = \sigma$.*

**Proof.** With direct calculation, show $\mathbb{E}[e^{s(X-\mathbb{E}[X])}] = e^{\sigma^2 s^2/2}$. □

## Fact

*If $X \in [a, b]$, then $X$ is sub-Gaussian with constant $\tau^2 = (b-a)^2/4$.*

**Proof.** Re-do the proof of Hoeffding lemma. □

# Hoeffding for sub-Gaussians

## Fact

*Let $X_1, \dots, X_N$ be independent sub-Gaussian random variables with constant $\tau$. Then $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$ is sub-Gaussian with constant $\tau/\sqrt{N}$.*

**Proof.** Follow the reasoning of Hoeffding inequality. □

## Theorem

*Let $X_1, \dots, X_N$ be independent sub-Gaussian random variables with constant $\tau > 0$. Write $\mu_1, \dots, \mu_N \in \mathbb{R}$ for the means. Let $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$ and $\bar{\mu} = \frac{1}{N} \sum_{i=1}^N \mu_i$. For any $\varepsilon > 0$,*

$$\mathbb{P}(\bar{X} - \bar{\mu} \geq \varepsilon) \leq \exp\left(-\frac{N\varepsilon^2}{2\tau^2}\right).$$

**Proof.** Re-do the proof of Hoeffding inequality. □

# Tails of sub-Gaussians

The Hoeffding inequality with $N = 1$ implies

$$\mathbb{P}(|X - \mu| \geq \varepsilon) \leq 2 \exp\left(-\frac{\varepsilon^2}{2\tau^2}\right),$$

so sub-Gaussians are light-tailed.

In fact, we can equivalently characterize sub-Guassian random variables via this light-tailed property.

## Theorem

*$X$ is a sub-Gaussian with some $\tau > 0$ if and only if there is some $\omega > 0$ such that*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{\omega}\right) \qquad \forall\, \varepsilon > 0.$$

# McDiarmid inequality

## Theorem

*Let $Z_1, \ldots, Z_N \in \mathcal{Z}$ be independent random variables. (No assumption on $\mathcal{Z}$.) Let $f \colon \mathcal{Z}^N \to \mathbb{R}$ be a function satisfying the following "bounded differences property" with $c > 0$:*

$$\left| f(Z_1, \ldots, Z_{i-1}, Z_i, Z_{i+1}, \ldots, Z_N) - f(Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \ldots, Z_N) \right| \leq c$$

*for all $Z_1, \ldots, Z_N, Z_i' \in \mathcal{Z}$. Then, for all $\varepsilon > 0$,*

$$\mathbb{P}\left(f(Z_1, \ldots, Z_N) - \mathbb{E}[f(Z_1, \ldots, Z_N)] \geq \varepsilon\right) \leq \exp(-2\varepsilon^2/(Nc^2)).$$

(Note, McDiarmid generalizes Hoeffding with $f = \frac{1}{N}\sum_{i=1}^N Z_i$ and $c = 1$.)

# McDiarmid inequality

**Proof.** We show

$$\mathbb{P}\left(f(Z_1,\ldots,Z_N) - \mathbb{E}[f(Z_1,\ldots,Z_N)] \geq \varepsilon\right) \leq \exp(-2\varepsilon^2/(Nc^2)).$$

and the two-sided bound from the decomposition argument of p. 30.

For $i = 1,\ldots,N$, define

$$V_i = \mathbb{E}[f(Z_1,\ldots,Z_N) \,|\, Z_1,\ldots,Z_i] - \mathbb{E}[f(Z_1,\ldots,Z_N) \,|\, Z_1,\ldots,Z_{i-1}],$$

where for $i = 1$, we have

$$\mathbb{E}[f(Z_1,\ldots,Z_N) \,|\, \emptyset] = \mathbb{E}[f(Z_1,\ldots,Z_N)].$$

By definition,

$$f(Z_1,\ldots,Z_N) - \mathbb{E}[f(Z_1,\ldots,Z_N)] = \sum_{i=1}^{N} V_i.$$

By the law of iterated expectations,

$$\mathbb{E}[V_i \,|\, Z_1,\ldots,Z_{i-1}] = 0.$$

(So $V_1,\ldots,V_N$ is a martingale difference sequence.)

# McDiarmid inequality

By the bounded differences property,

$$V_i = \mathbb{E}\left[ \mathop{\mathbb{E}}_{Z_i'} \left[ f(Z_1, \ldots, Z_{i-1}, Z_i, Z_{i+1}, \ldots, Z_N) - f(Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \ldots, Z_N) \right] \middle| Z_1, \ldots, Z_i \right]$$

$$|V_i| \leq \mathbb{E}\left[ \mathop{\mathbb{E}}_{Z_i'} \left[ \left| f(Z_1, \ldots, Z_{i-1}, Z_i, Z_{i+1}, \ldots, Z_N) - f(Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \ldots, Z_N) \right| \right] \middle| Z_1, \ldots, Z_i \right]$$

$$\leq \mathbb{E}[\mathop{\mathbb{E}}_{Z_i'}[c] \mid Z_1, \ldots, Z_i]$$

$$= c$$

Finally, we use the Hoeffding inequality with the martingale difference sequence $V_1, \ldots, V_N$ to conclude the statement. □

# Expectation of maximum

## Theorem

*Let $X_1, \dots, X_N$ be (potentially dependent) zero-mean random variables that are sub-Gaussian with constant $\tau > 0$. Then*

$$\mathbb{E}[\max\{X_1, \dots, X_N\}] \leq \sqrt{2\tau^2 \log N}.$$

**Proof.**

$$\begin{aligned}
&\mathbb{E}[\max\{X_1, \dots, X_N\}] \\
&\quad \leq \frac{1}{t} \log \mathbb{E}[e^{t \max\{X_1, \dots, X_N\}}] \qquad \text{for } t > 0 \text{ (Jensen inequality)} \\
&\quad = \frac{1}{t} \log \mathbb{E}[\max\{e^{tX_1}, \dots, e^{tX_N}\}] \\
&\quad \leq \frac{1}{t} \log \mathbb{E}[e^{tX_1} + \cdots + e^{tX_N}] \\
&\quad \leq \frac{1}{t} \log(N e^{\tau^2 t^2/2}).
\end{aligned}$$

Finally, plug in $t = \frac{1}{\tau}\sqrt{2 \log N}$, the minimizer of the bound, to conclude the result. □

# Expectation of maximum

## Corollary

*Let $X_1, \ldots, X_N$ be (potentially dependent) zero-mean random variables that are sub-Gaussian with constant $\tau > 0$. Then*

$$\mathbb{E}[\max\{|X_1|, \ldots, |X_N|\}] \leq \sqrt{2\tau^2 \log(2N)}.$$

**Proof.** Use the previous theorem with $X_1, \ldots, X_N, -X_1, \ldots, -X_N$. □

# Bernstein inequality

When we know $\sigma$, Bernstein's inequality improves upon Hoeffding.

## Theorem

*Let $X_1, \ldots, X_N$ be independent zero-mean random variables such that $|X_i| \leq c$ almost surely for $i = 1, \ldots, N$. Then, for $\varepsilon > 0$,*

$$\mathbb{P}(|\bar{X}| \geq \varepsilon) \leq 2 \exp\Big( - \frac{N\varepsilon^2}{2\sigma^2 + 2c\varepsilon/3} \Big),$$

*where $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$ and $\sigma^2 = \frac{1}{N} \sum_{i=1}^N \mathrm{Var}(X_i)$.*

Bernstein is not uniformly better than Hoeffding, but it is "not worse" than Hoeffding for small $\varepsilon > 0$, where "not worse" is meant in the sense illustrated in the next slide.

# Proof of Bernstein

**Proof.** We start by showing two Lemmas.

Lemma 1) The function

$$h(x) = \frac{1}{x^2}\sum_{k=1}^{\infty}\frac{x^k}{k!} = \begin{cases} \frac{e^x - x - 1}{x^2} & \text{for } x \neq 0 \\ \frac{1}{2} & \text{for } x = 0 \end{cases}$$

is a monotonically increasing function. We skip the proof.

Lemma 2) Let $|u| < 3$. Then,

$$\begin{aligned} e^u - 1 - u &= \sum_{k=0}^{\infty}\frac{u^{k+2}}{(k+2)!} \\ &\leq \sum_{k=0}^{\infty}\frac{u^{k+2}}{2\cdot 3^k} = \frac{u^2}{2}\sum_{k=0}^{\infty}\frac{u^k}{3^k} = \frac{u^2}{2}\frac{1}{1-u/3}. \end{aligned}$$

# Proof of Bernstein

Lemma 3) Let $Z$ be a random variable such that $|Z| \leq c$, $\mathbb{E}[Z] = 0$, and $\mathbb{E}[Z^2] = \sigma^2$. Then,

$$\begin{aligned}
\mathbb{E}[e^{sZ}] &= \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{s^k}{k!} Z^k\right] = \sum_{k=0}^{\infty} \frac{s^k}{k!} \mathbb{E}[Z^k] \\
&= 1 + \mathbb{E}[sZ] + \sum_{k=2}^{\infty} \frac{s^k}{k!} \mathbb{E}[Z^k] = 1 + \mathbb{E}[h(sZ) s^2 Z^2] \\
&\leq 1 + \mathbb{E}[h(sc) s^2 Z^2] = 1 + h(sc) s^2 \mathbb{E}[Z^2] \\
&= 1 + \tfrac{\sigma^2}{c^2}(e^{sc} - 1 - sc) \leq 1 + \tfrac{\sigma^2 s^2}{2} \tfrac{1}{1 - sc/3} \\
&\leq \exp\left(\tfrac{s^2 \sigma^2}{2(1 - sc/3)}\right),
\end{aligned}$$

where we swap the order of the summation and expectation using the boundedness of $Z$ and the Lebesgue dominated convergence theorem.

# Proof of Bernstein

Finally, we proceed with the main proof.

Let $\sigma_i^2 = \mathrm{Var}(X_i)$ for $i = 1, \ldots, N$ and note $N\sigma^2 = \sigma_1 + \cdots + \sigma_N^2$. Then,

$$\begin{aligned}
\mathbb{P}(\bar{X} \geq \varepsilon) = \mathbb{P}(e^{sN\bar{X}} \geq e^{sN\varepsilon}) &\leq e^{-sN\varepsilon}\mathbb{E}[e^{sN\bar{X}}] = e^{-sN\varepsilon}\prod_{i=1}^{N}\mathbb{E}[e^{sX_i}] \\
&\leq e^{-sN\varepsilon}\prod_{i=1}^{N}\exp\left(\tfrac{s^2\sigma_i^2}{2(1-sc/3)}\right) \\
&\leq \exp\left(-sN\varepsilon + \tfrac{Ns^2\sigma^2}{2(1-sc/3)}\right)
\end{aligned}$$

for $s < \frac{3}{c}$. Finally, plugging in

$$s = \frac{\varepsilon}{\sigma^2 + c\varepsilon/3} < \frac{1}{c}$$

and simplifying leads to the stated bound. □

# Hoeffding vs. Bernstein

Let $X_1, \ldots, X_N \in [a, b]$ be IID random variables with mean $\mu \in \mathbb{R}$.
Imagine we want to estimate $\mu$ with $\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$.

Using Hoeffding, one can (and you will) show

$$N = \mathcal{O}\left( \frac{(b-a)^2}{\varepsilon^2} \log(1/\delta) \right) \quad \Rightarrow \quad \mathbb{P}(|\bar{X} - \mu| < \varepsilon) \leq 1 - \delta.$$

Using Bernstein, one can (and you will) show

$$N = \mathcal{O}\left( \left( \frac{\sigma^2}{\varepsilon^2} + \frac{b-a}{\varepsilon} \right) \log(1/\delta) \right) \quad \Rightarrow \quad \mathbb{P}(|\bar{X} - \mu| < \varepsilon) \leq 1 - \delta.$$

Consider small $\varepsilon > 0$ and $\delta > 0$. Note that for small $\varepsilon > 0$, the term $\frac{\sigma^2}{\varepsilon^2}$ dominates the other term $\frac{b-a}{\varepsilon}$.

If $\sigma \ll (b-a)$, Bernstein's sample complexity is better than Hoeffding's.
If $\sigma \approx (b-a)$, then the two sample complexities are roughly the same.

# Matrix functions

Let $f : \mathbb{R} \to \mathbb{R}$. For any $A = A^\intercal \in \mathbb{R}^{d \times d}$, define the *matrix function* $f(A) \in \mathbb{R}^{d \times d}$ by first taking the eigenvalue decomposition

$$A = U \operatorname{diag}(\lambda_1, \ldots, \lambda_d) U^\intercal$$

and then forming

$$f(A) = U \operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_d)) U^\intercal.$$

(One can define matrix functions for asymmetric matrices, but we will not do so in this class.)

# Lieb's theorem

## Theorem

*Let $S = S^\intercal \in \mathbb{R}^{d\times d}$. The function*

$$A \mapsto -\mathrm{Tr}\big(\exp(S + \log A)\big)$$

*is a convex map on the set of $d \times d$ positive definite matrices.*

(Note that the set of positive definite matrices is a convex set.)

The proof is quite involved. We will use this result without proof.

---

E. H. Lieb, Convex trace functions and the Wigner–Yanase–Dyson conjecture, *Adv. Math.*, 1973.

# Matrix Bernstein inequality

## Theorem

*Let $X_1, \dots, X_N \in \mathbb{R}^{d \times d}$ be independent symmetric random matrices such that $\lambda_{\max}(X_i) \leq c$ almost surely and $\mathbb{E}[X_i] = 0$ for $i = 1, \dots, N$. Then, for $\varepsilon > 0$,*

$$\mathbb{P}(\lambda_{\max}(\bar{X}) \geq \varepsilon) \leq d \exp\Big( - \frac{N\varepsilon^2}{2\sigma^2 + 2c\varepsilon/3} \Big),$$

*where $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$ and*

$$\sigma^2 = \lambda_{\max}\Big( \frac{1}{N} \sum_{i=1}^N \mathbb{E}[X_i^2] \Big).$$

# Proof of matrix Bernstein

**Proof.** We start by showing three Lemmas.

Lemma 1) Let $Z = Z^\intercal \in \mathbb{R}^{d\times d}$ be a random matrix such that $\lambda_{\max}(Z) \leq c$, and $\mathbb{E}[Z] = 0$. Consider the same function $h$ as in the scalar matrix Bernstein inequality. Then,

$$\begin{aligned}
\mathbb{E}[e^{sZ}] &= \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{s^k}{k!} Z^k\right] = \sum_{k=0}^{\infty} \frac{s^k}{k!}\mathbb{E}[Z^k] \\
&= I + \mathbb{E}[sZ] + \sum_{k=2}^{\infty} \frac{s^k}{k!}\mathbb{E}[Z^k] = I + \mathbb{E}[h(sZ)s^2Z^2] \\
&\preceq I + \mathbb{E}[h(sc)s^2Z^2] = I + h(sc)s^2\mathbb{E}[Z^2] \\
&= I + \tfrac{1}{c^2}(e^{sc} - 1 - sc)\mathbb{E}[Z^2] \preceq I + \tfrac{s^2}{2(1-sc/3)}\mathbb{E}[Z^2] \\
&\preceq \exp\left(\tfrac{s^2}{2(1-sc/3)}\mathbb{E}[Z^2]\right)
\end{aligned}$$

for $s < c/3$, where we follow similar steps as in the scalar case.

# Proof of matrix Bernstein

Lemma 2) Let $s > 0$ and $M = M^\intercal \in \mathbb{R}^{d \times d}$. Then

$$e^{s\lambda_{\max}(M)} \leq \mathrm{Tr}(e^{sM}).$$

To see why, let $\lambda_1, \dots \lambda_d$ be the eigenvalues of $M$. Then,

$$e^{s \max_{i=1,\dots,d}\{\lambda_i\}} = \max_{i=1,\dots,d} e^{s\lambda_i} \leq \sum_{i=1,\dots,d} e^{s\lambda_i} = \mathrm{Tr}(e^{sM}).$$

Lemma 3) In the matrix case, $\mathbb{E}[\exp(s\sum_{i=1}^N X_i)] \neq \prod_{i=1}^N \mathbb{E}[\exp(sX_i)]$ even though the $X_i$ are independent, so the scalar proof needs modification. We use the following bound

$$\begin{aligned}
\operatorname{Tr}\left(\mathbb{E}\left[\exp\left(s\sum_{i=1}^N X_i\right)\right]\right) &= \mathbb{E}\left[\operatorname{Tr}\exp\left(s\sum_{i=1}^N X_i\right)\right] \\
&= \mathbb{E}\left[\operatorname{Tr}\exp\left(s\sum_{i=1}^{N-1} X_i + \log\exp(sX_N)\right)\right] \\
&= \mathop{\mathbb{E}}_{X_1,\ldots,X_{N-1}}\left[\mathop{\mathbb{E}}_{X_N}\left[\operatorname{Tr}\exp\left(s\sum_{i=1}^{N-1} X_i + \log\exp(sX_N)\right)\right]\right] \\
&\leq \mathop{\mathbb{E}}_{X_1,\ldots,X_{N-1}}\left[\operatorname{Tr}\exp\left(s\sum_{i=1}^{N-1} X_i + \log\mathop{\mathbb{E}}_{X_N}[\exp(sX_N)]\right)\right] \\
&\vdots \\
&\leq \operatorname{Tr}\left(\exp\left(\sum_{i=1}^N \log\mathbb{E}[\exp(sX_i)]\right)\right)
\end{aligned}$$

# Proof of matrix Bernstein

Next, we proceed with the main proof.

Using the bound of the lemma 3, we have

$$\begin{aligned}
\mathbb{P}(\lambda_{\max}(\bar{X}) \geq \varepsilon) =& \mathbb{P}(e^{sN\lambda_{\max}(\bar{X})} \geq e^{sN\varepsilon}) \leq e^{-sN\varepsilon}\mathbb{E}[e^{s\lambda_{\max}(N\bar{X})}] \\
&\leq e^{-sN\varepsilon}\mathrm{Tr}(\mathbb{E}[e^{sN\bar{X}}]) \\
&\leq e^{-sN\varepsilon}\mathrm{Tr}\left(\exp\left(\sum_{i=1}^{N}\log\mathbb{E}[\exp(sX_i)]\right)\right) \\
&\leq \mathrm{Tr}\left(\exp\left(-sN\varepsilon I + \tfrac{s^2}{2(1-sc/3)}\sum_{i=1}^{N}\mathbb{E}[X_i^2]\right)\right) \\
&\leq \mathrm{Tr}\left(\exp\left(-sN\varepsilon I + \tfrac{Ns^2\sigma^2}{2(1-sc/3)}I\right)\right)
\end{aligned}$$

for $s < \frac{3}{c}$. Finally, plugging in

$$s = \frac{\varepsilon}{\sigma^2 + c\varepsilon/3} < \frac{1}{c}$$

and noting that $\mathrm{Tr}(I) = d$ leads to the stated bound. □

# Outline

# Extended-valued convex functions

Let $C \subseteq \mathbb{R}^d$ be convex. We say a function $f : C \to \mathbb{R}$ is convex if

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y), \qquad \forall\, x, y \in C,\, \theta \in (0,1).$$

We say a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is an extended-valued convex function if

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y), \qquad \forall\, x, y \in \mathbb{R}^d,\, \theta \in (0,1).$$

Define the (effective) domain of $f$ as

$$\mathbf{dom}\, f = \{x \in \mathbb{R}^d \mid f(x) < \infty\}.$$

Then, $\mathbf{dom}\, f \subseteq \mathbb{R}^d$ is convex. Also, we can identify an extended-valued convex $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ with a convex $\tilde{f} : \mathbf{dom}\, f \to \mathbb{R}$ such that $\tilde{f}(x) = f(x)$ for all $x \in \mathbf{dom}\, f$.

# Gradient provides global lower bound

Given an extended-valued convex function $f$, we say $f$ is differentiable at $x \in \mathbb{R}^d$ if $f$ is finite on an open neighborhood of $x$ and $f$ is differentiable at $x$ in the usual sense.

## Theorem

*Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be convex. Assume $f$ is differentiable at $x$. Then,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \qquad \forall\, y \in \mathbb{R}^d.$$

**Proof.** By convexity,

$$f(x + \theta(y - x)) \leq (1 - \theta) f(x) + \theta f(y), \qquad \forall\, \theta \in (0, 1).$$

Reorganizing, we get

$$f(y) \geq f(x) + \frac{f(x + \theta(y - x)) - f(x)}{\theta}, \qquad \forall\, \theta \in (0, 1).$$

By taking $\theta \to 0$, we get the desired inequality. □

# Projection onto convex sets are well defined

## Lemma

*Let $A \subseteq \mathbb{R}^d$ be a nonempty closed convex set and let $p \in \mathbb{R}^d$. Then*

$$\operatorname*{argmin}_{x \in A} \|x - p\|_2$$

*uniquely exists.*

**Proof.** Proof in homework. □

# Supporting hyperplane theorem

## Theorem (Supporting hyperplane theorem)

*Let $A \subset \mathbb{R}^d$ be a nonempty closed convex set and let $p \in \partial A$. Then, there is a non-zero $v \in \mathbb{R}^d$ such that*

$$v^\intercal x \leq v^\intercal p, \qquad \forall\, x \in A.$$

**Proof.** For any $\varepsilon > 0$, it must be that $B(p, \varepsilon) \not\subset A$, since $p \in \partial A$. Choose $p_n \in B(p, 1/2^n) \backslash A$ for $n \in \mathbb{N}$, so $p_n \to p$. Also, let

$$x_n = \operatorname*{argmin}_{x \in A} \|x - p_n\|,$$

i.e., $x_n$ is the projection of $p_n$ onto $A$, which uniquely exists for $n \in \mathbb{N}$. Then, $\|x_n - p_n\| > 0$, so let

$$v_n = \frac{p_n - x_n}{\|p_n - x_n\|}.$$

Since $\{v_n\}_{n \in \mathbb{N}}$ is a sequence on the unit ball (which is compact), it has an accumulation point, which we denote $v_\infty$.

We claim

$$v_n^\intercal x \le v_n^\intercal p_n, \qquad \forall\, x \in A.$$

Otherwise, if there is an $x \in A$ such that $v^\intercal x > v^\intercal p_n$, i.e., such that $(p_n - x_n)^\intercal x > (p_n - x_n)^\intercal p_n$, then

$$\begin{aligned}
&\| \underbrace{\lambda x + (1-\lambda) x_n}_{\in A} - p_n \|^2 = \|\lambda(x - p_n) + (1-\lambda)(x_n - p_n)\|^2 \\
&= \lambda^2 \|(x - p_n)\|^2 + (1-\lambda)^2 \|x_n - p_n\|^2 + 2\lambda(1-\lambda)(x - p_n)^\intercal (x_n - p_n) \\
&= (1 - 2\lambda)\|x_n - p_n\|^2 + 2\lambda \underbrace{(x - p_n)^\intercal (x_n - p_n)}_{<0} + \mathcal{O}(\lambda^2) \le \|x_n - p_n\|^2
\end{aligned}$$

for small $\lambda > 0$, which contradicts the assumption that $x_n$ is the projection of $p_n$ into $A$.

Since there is a subsequence $n_j \to \infty$ such that $v_{n_j} \to v_\infty \stackrel{\text{def}}{=} v$ and $p_{n_j} \to p$, we conclude

$$v^\intercal x \le v^\intercal p, \qquad \forall\, x \in A.$$

□

# Subgradients

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be convex (but not necessarily differentiable). We say $g \in \mathbb{R}^d$ is a *subgradient* of $f$ at $x$ if

$$f(y) \geq f(x) + g^\intercal (y - x), \qquad \forall\, y \in \mathbb{R}^d.$$

We write $\partial f(x) \subseteq \mathbb{R}^d$ to denote the set of subgradients at $x$.

We have already established that $\nabla f(x) \in \partial f(x)$ if $f$ is differentiable at $x$, but convex functions can be non-differentiable. Nevertheless, a subgradient always exists on $\operatorname{int} \operatorname{dom} f$.

# Existence of a subgradient

## Theorem

*Let* $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ *be convex. If* $x \in \operatorname{int} \operatorname{dom} f$ *(*$x$ *in interior of* $\operatorname{dom} f$*), then exists a subgradient of* $f$ *at* $x$*, i.e.,* $g \in \partial f(x)$ *exists.*

**Proof.** Consider the *epigraph* of $f$:

$$A = \{(x, t) \,|\, f(x) \leq t,\, x \in \mathbb{R}^d,\, t \in \mathbb{R}\} \subset \mathbb{R}^{d+1}.$$

Convexity of $f$ as a function implies convexity of $A$ as a set.

By construction, $(x, f(x)) \in \partial A$. By the supporting hyperplane theorem, there is a $v = (\tilde{g}, \tau) \in \mathbb{R}^{d+1}$ such that

$$\tilde{g}^\intercal y + \tau s \leq \tilde{g}^\intercal x + \tau f(x), \qquad \forall (y, s) \in A.$$

Given any $y$, we can take $s \to \infty$, so $\tau \leq 0$. If $\tau = 0$, then,

$$\tilde{g}^\intercal (x + \delta) \leq \tilde{g}^\intercal x$$

for sufficiently small $\delta \in \mathbb{R}^d$ such that $x + \delta \in \operatorname{dom} f$. Since $x \in \operatorname{int} \operatorname{dom} f$, this implies that $\tilde{g} = 0$, but this contradicts $(\tilde{g}, \tau) \neq 0$. Therefore, we conclude $\tau < 0$.

# Existence of a subgradient

Finally, we divide both sides of the inequality by $\tau < 0$ and let $g = \tilde{g}/\tau$ to get

$$-g^\intercal y + s \geq -g^\intercal x + f(x), \qquad \forall (y, s) \in A.$$

Plugging in $s = f(y)$, we conclude

$$f(y) \geq f(x) + g^\intercal (y - x), \qquad \forall y \in \mathbb{R}^d.$$

□

# Uniqueness of subgradient implies differentiability

**Theorem**
*Let $f : \mathbb{R}^d \to \mathbb{R}$. Then $\{g\} = \partial f(x)$ if and only if $f$ is differentiable at $x$ and $g = \nabla f(x)$.*

We won't use this result, so we won't prove it.

# Gradient of cvx. $f$ provides a cutting plane for $\operatorname{argmin} f$

## Lemma

*Let $f : \mathbb{R} \to \mathbb{R}$ be convex. Assume $f$ is differentiable at $x = 0$.*

(a) *If $f'(0) < 0$, then $\operatorname{argmin} f \subseteq (0, \infty)$.*

(b) *If $f'(0) > 0$, then $\operatorname{argmin} f \subseteq (-\infty, 0)$.*

**Proof.** We only prove (a) as (b) follows from the same reasoning with the sign flipped. By convexity

$$f(y) \geq f(0) + f'(0) \cdot y, \qquad \forall\, y \in \mathbb{R}.$$

Therefore, for $y < 0$, we have

$$f(y) \geq f(0) + f'(0) \cdot y > f(0).$$

So $\operatorname{argmin} f \subseteq [0, \infty)$.

By standard calculus arguments, for small $y$,

$$f(y) = f(0) + f'(0) \cdot y + \mathcal{O}(y^2)$$

so $\inf f < f(0)$. Thus we conclude $\operatorname{argmin} f \subseteq (0, \infty)$. □

# Jensen's inequality

## Lemma (Jensen's inequality)

*Let $X \in \mathbb{R}^d$ be a random variable such that $\mathbb{E}[X] \in \mathbb{R}^d$ is well defined, and let $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ be convex. Then,*

$$\varphi\big(\mathbb{E}[X]\big) \leq \mathbb{E}[\varphi(X)].$$

**Proof.** Let $g \in \partial\varphi(\mathbb{E}[X])$. Then,

$$\varphi(X) \geq \varphi(\mathbb{E}[X]) + g^\intercal (X - \mathbb{E}[X]).$$

Taking expectations on both sides completes the proof. □

# General Jensen's inequality

## Lemma (General Jensen's inequality)

*Let $C \subseteq \mathbb{R}^d$ be a nonempty convex set and let $\varphi\colon C \to \mathbb{R}$ be convex. Let $X$ be a random variable such that $X \in C$ with probability $1$ and $\mathbb{E}[X] \in \mathbb{R}^d$ is well defined. Then, $\mathbb{E}[X] \in C$ and*

$$\varphi\big(\mathbb{E}[X]\big) \leq \mathbb{E}[\varphi(X)].$$

**Proof.** Step-by-step in homework. □

# Continuity of univariate convex functions

## Theorem

*Let* $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ *be convex. Then,* $f$ *is continuous on* $\operatorname{int} \operatorname{dom} f$*.*

**Proof.** W.L.O.G. consider continuity at $x = 0 \in \operatorname{int} \operatorname{dom} f$. W.L.O.G. assume $0 \in \operatorname{argmin} f$ and $0 = \min f$, since otherwise we can consider

$$\tilde{f}(x) = f(x) - f(0) - gx,$$

where $g \in \partial f(x)$ and noting that continuity of $f$ and $\tilde{f}$ are equivalent.

For any $x \neq 0$, the convexity inequality with $y = 0$ imlpies

$$f(\varepsilon x) \leq \varepsilon f(x)$$

for all $\varepsilon \in [0, 1]$. Also note that $0 \leq f(\varepsilon x)$. Therefore, by taking $x = \pm\delta$ for sufficiently small $\delta > 0$ and $\varepsilon \to 0$, we conclude

$$\lim_{z \to 0} f(z) = 0.$$

□

# Lemma: Convex fn. are maximized at extreme points

For any $\varepsilon > 0$, let

$$K^\varepsilon = \{(\pm\varepsilon, \ldots, \pm\varepsilon) \in \mathbb{R}^d\}, \qquad \text{(So } |K| = 2^d\text{)}$$
$$C^\varepsilon = \{x \in \mathbb{R}^d \mid \|x\|_\infty \leq \varepsilon\}.$$

## Lemma

*Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. Then,*

$$\sup_{x \in C^\varepsilon} f(x) = \max_{x \in K^\varepsilon} f(x).$$

# Lemma: Convex fn. are maximized at extreme points

**Proof.** Let $x^\star \in \operatorname{argmax}_{x \in K^\varepsilon} f(x)$. Assume for contradiction that there is an $x^\circ \in C^\varepsilon \backslash K^\varepsilon$ such that $f(x^\circ) > f(x^\star)$. Then, there exists an index $i \in \{1, \ldots, d\}$ such that $x_i^\circ \in (-\varepsilon, +\varepsilon)$. Let $g \in \partial f(x^\circ)$. Then the subgradient inequality tells us that

$$f(x^\circ + \delta e_i) \geq f(x^\circ) + \delta g_i$$

for all $\delta \in \mathbb{R}$. Therefore, by taking $\delta > 0$ if $g_i \geq 0$ and $\delta < 0$ if $g_i < 0$, we can find a $\delta$ such that

$$x_i^\circ + \delta = \pm\varepsilon, \qquad f(x^\circ + \delta e_i) \geq f(x^\circ).$$

Therfore, $x^\circ + \delta e_i$ has one fewer coordinate in $(-\varepsilon, +\varepsilon)$ and it has a function value that is not smaller.

Repeating this process at most $d$ times, we get a point in $K^\varepsilon$ with function value not not smaller than $f(x^\circ) > f(x^\star)$. This contradicts the optimality of $x^\star$, and we are forced to conclude that

$$\sup_{x \in C^\varepsilon} f(x) \leq f(x^\star).$$

# Continuity of multivariate convex functions

## Theorem

*Let* $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ *be convex. Then,* $f$ *is continuous on* $\operatorname{int} \operatorname{dom} f$.

**Proof.** W.L.O.G. consider continuity at $x = 0 \in \operatorname{int} \operatorname{dom} f$. W.L.O.G. assume $0 \in \operatorname{argmin} f$ and $0 = \min f$. Consider $K^\varepsilon$ and $C^\varepsilon$ as previously defined.

Let $\{x^{(1)}, \dots x^{(2^d)}\} = K^\varepsilon$. Then,

$$0 = f(0) \leq f(x) \leq \max_{j=1,\dots,2^d} f(x^{(j)}), \qquad \forall\, x \in C^\varepsilon.$$

Since univariate convex functions are continuous,

$$\lim_{\varepsilon \to 0} \max_{j=1,\dots,2^d} f(x^{(j)}) = \max_{j=1,\dots,2^d} \lim_{\varepsilon \to 0} f(x^{(j)}) = f(0) = 0.$$

Therefore,

$$0 \leq \inf_{x \in C^\varepsilon} f(x) \leq \sup_{x \in C^\varepsilon} f(x) = \max_{x \in K^\varepsilon} f(x) \to 0$$

as $\varepsilon \to 0$, and we conclude continuity. □

# CCP functions

$f$ is CCP if closed, convex, and proper:

- $f$ is proper if $f(x) < \infty$ somewhere.
- Proper $f$ is closed if epigraph of $f$

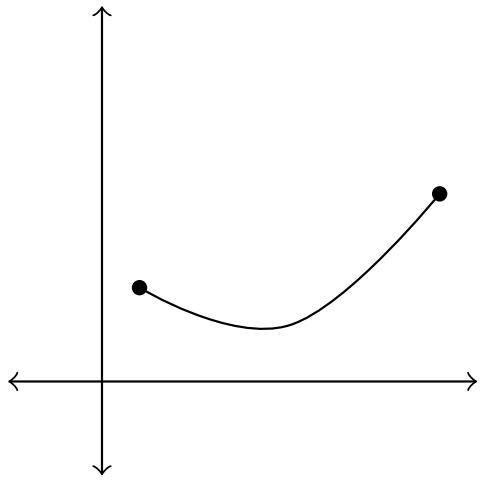
$$\mathbf{epi}\, f = \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \,|\, f(x) \leq \alpha\}$$

  is closed.

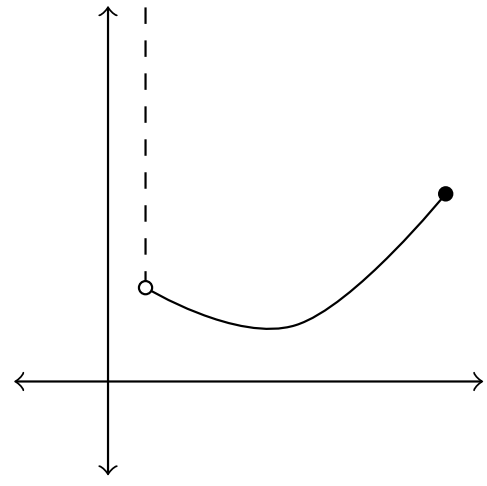
Properties:

- Most convex functions of interest are closed and proper.
- [$f$ is convex] $\Leftrightarrow$ [$\mathbf{epi}\, f$ is convex]
- For proper $f$, [$f$ closed] $\Leftrightarrow$ [$f$ is lower semi-continuous]
- [$f$ CCP] $\Leftrightarrow$ [$\mathbf{epi}\, f$ nonempty closed convex without a vertical line] (vertical line $= \{x_0\} \times \mathbb{R}$.)
- If $f$ is convex and $f(x) < \infty$ for all $x$, then $f$ is CCP

# CCP function example



Closed convex function

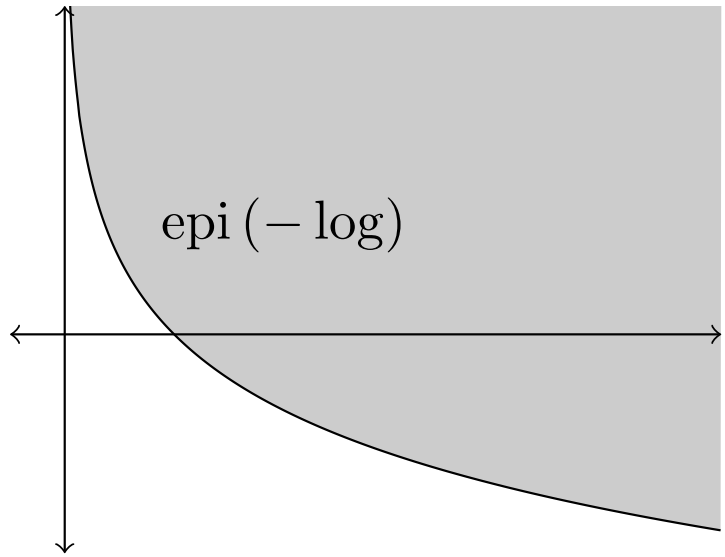


Convex but not closed

The dashed line denotes the function value of $\infty$.

# CCP function example

Epigraph of the CCP $-\log$ is a nonempty closed convex set.



# Conjugate function

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. Conjugate function of $f$:

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \{\langle y, x \rangle - f(x)\}$$

Properties: when $f$ is CCP

- $f^*$ is CCP and $f^{**} = f$
- $(\nabla f)^{-1} = \nabla f^*$ when $f$ and $f^*$ are differentiable
- $(\partial f)^{-1} = \partial f^*$ in general

# Strong convexity

With $\mu > 0$, CCP $f$ is $\mu$-strongly convex if:

- $f(x) - (\mu/2)\|x\|^2$ is convex.
- $f(y) \geq f(x) + \langle g, y - x\rangle + \frac{\mu}{2}\|x - y\|^2$ for all $x, y$ and $g \in \partial f(x)$.
- $\nabla^2 f(x) \succeq \mu I$ for all $x$ if $f$ is twice continuously differentiable.

These conditions are equivalent.

If $f$ is $\mu$-strongly convex and $g$ is convex, then $f + g$ is $\mu$-strongly convex. To clarify, strong convexity does not imply differentiability.

# $L$-smooth function

With $L > 0$, CCP $f$ is $L$-smooth if:

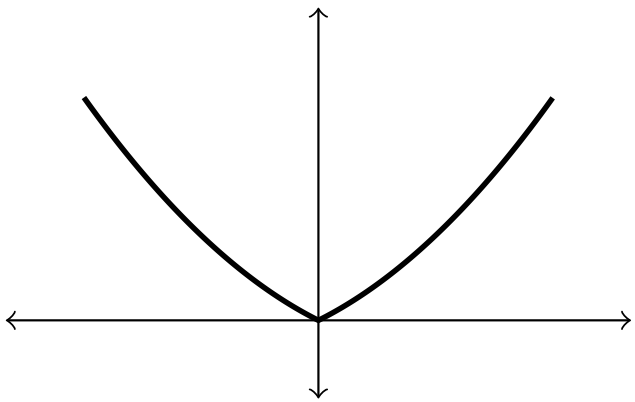
- $f(x) - (L/2)\|x\|^2$ is concave.
- $f$ is differentiable and $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2$ for all $x, y$.
- $f$ is differentiable and $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2$ for all $x, y$.
- $f$ is differentiable and $\nabla f$ is $L$-Lipschitz.
- $\nabla^2 f(x) \preceq LI$ for all $x$ if $f$ is twice continuously differentiable.

These conditions are equivalent.

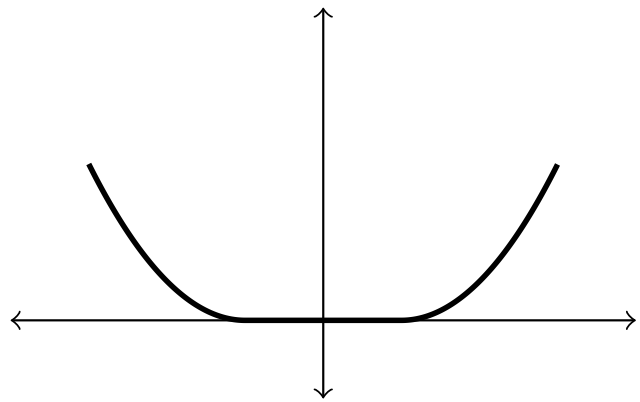
"$L$-smoothness", which implies once-continuous differentiability, is somewhat non-standard; "smoothness" often means infinite differentiability in other fields of mathematics.

# Strong convexity and smoothness

Informally speaking, $\mu$-strongly convex functions have upward curvature of at least $\mu$ and $L$-smooth convex functions have upward curvature of no more than $L$. We can think of nondifferentiable points to be points with infinite curvature.



Strongly convex but not smooth



Smooth but not strongly convex.

# Strong convexity and smoothness

If $f$ is $\mu$-strongly convex and $L$-smooth, then $\mu \leq L$.

Strong convexity and smoothness are dual properties:
if $f$ CCP, [$f$ is $\mu$-strongly convex] $\Leftrightarrow$ [$f^*$ is $(1/\mu)$-smooth]

# Fenchel–Young inequality

## Theorem

*Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be convex and proper. Then,*

$$f(x) + f^*(y) \geq x^\intercal y, \qquad \forall\, x, y \in \mathbb{R}^d.$$

**Proof.** By definition of the conjugate,

$$f^*(y) = \sup_{z \in \mathbb{R}^d} \{z^\intercal y - f(z)\} \geq x^\intercal y - f(x).$$

□

Example consequences of Fenchel–Young:

$$a^\intercal b \leq \frac{\varepsilon}{2}\|a\|_2^2 + \frac{1}{2\varepsilon}\|b\|_2^2, \qquad \forall\, a, b \in \mathbb{R}^d,\ \varepsilon > 0$$

$$a^\intercal b \leq \frac{1}{p}\|a\|_p^p + \frac{1}{q}\|b\|_q^q, \qquad \forall\, a, b \in \mathbb{R}^d,\ p, q \in (1, \infty),\ \tfrac{1}{p} + \tfrac{1}{q} = 1.$$

# Fenchel–Young for smooth functions

## Theorem

*Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and $L$-smooth. Then,*

$$f(x) + f^*(y) \geq x^\intercal y + \frac{1}{2L}\|y - \nabla f(x)\|^2, \qquad \forall\, x, y \in \mathbb{R}^d.$$

**Proof.** If $f^*(y) = \infty$, then the inequality holds vacuously. Assume $y \in \operatorname{int}\operatorname{dom} f^*$. Then, there exists $z \in \partial f^*(y)$, and $\nabla f(z) = y$. By smoothness of $f$, we have

$$\begin{aligned} f(x) &\geq f(z) + \nabla f(z)^\intercal (x - z) + \frac{1}{2L}\|\nabla f(x) - \nabla f(z)\|^2 \\ &= f(z) + y^\intercal (x - z) + \frac{1}{2L}\|\nabla f(x) - y\|^2. \end{aligned}$$

Plugging this in, we get

$$f^*(y) = \sup_{u \in \mathbb{R}^d} \{u^\intercal y - f(u)\} \geq z^\intercal y - f(z) \geq -f(x) + y^\intercal x + \frac{1}{2L}\|\nabla f(x) - u\|^2.$$

Finally, the case $y \in \partial(\operatorname{dom} f^*)$ follows from a lower semi-continuity argument that we won't consider for now. □

# Affine set

Affine set: $A \subseteq \mathbb{R}^d$ is affine if

$$(1-\theta)x + \theta y \in A, \qquad \forall\, x, y \in A,\, \theta \in \mathbb{R}.$$

(An empty set is defined to be an affine set.)

## Lemma

*A nonempty affine set $A \subseteq \mathbb{R}^d$ can be written as*

$$A = x_0 + \mathcal{V} = \{x_0 + v \,|\, V \in \mathcal{V}\},$$

*where $\mathcal{V} \subseteq \mathbb{R}^d$ is a subspace and $x_0 \in \mathcal{V}^\perp$.*

**Proof.** Let $x_0' \in A$. Then

$$A - x_0' = \{a - x_0' \,|\, a \in A\} \subseteq \mathbb{R}^d$$

is a subspace: Clearly, $0 = x_0' - x_0' \in A$,

$$x, y \in A - x_0' \quad \Rightarrow \quad x + x_0', y + x_0' \in A \quad \Rightarrow \quad \frac{1}{2}(x + x_0') + \frac{1}{2}(y + x_0') \in A$$
$$\Rightarrow \quad (x + x_0') + (y + x_0') - x_0' \in A \quad \Rightarrow \quad x + y \in A - x_0',$$

and

$$x \in A - x_0' \quad \Rightarrow \quad x + x_0' \in A \quad \Rightarrow \quad \alpha(x + x_0') + (1 - \alpha)x_0' \in A$$
$$\Rightarrow \quad \alpha x + x_0' \in A \quad \Rightarrow \quad \alpha x \in A - x_0'.$$

Finally, we let $\mathcal{V} = A - x_0'$, and $x_0 = \mathrm{Proj}_A(0)$. □

# Affine hull

Affine hull of $C \subseteq \mathbb{R}^d$:

$$\operatorname{aff} C = \{\theta_1 x_1 + \cdots + \theta_k x_k \,|\, x_1, \ldots, x_k \in C,\, \theta_1 + \cdots + \theta_k = 1,\, k \geq 1\}.$$

## Lemma

*Let $C \subseteq \mathbb{R}^d$ be nonempty. If $x_0 \in C$, then*

$$\operatorname{aff} C = x_0 + \operatorname{aff}(C - x_0) = x_0 + \operatorname{span}(C - x_0).$$

**Proof.** Proof in homework. □

# Interior

Closed ball of radius $r$ centered at $x$:

$$B(x, r) = \{y \in \mathbb{R}^d \mid \|y - x\| \leq r\}$$

$C \subseteq \mathbb{R}^d$ is open: For all $x \in C$, there is an $r > 0$ such that

$$B(x, r) \subseteq C$$

Interior of $C \subseteq \mathbb{R}^d$:

$$\operatorname{int} C = \{x \in C \mid B(x, r) \subseteq C \text{ for some } r > 0\}$$

Closure of $C \subseteq \mathbb{R}^d$: $\operatorname{cl} C$

Boundary of $C \subseteq \mathbb{R}^d$: $\operatorname{cl} C \backslash \operatorname{int} C$

# Subspace topology

Let $S \subseteq \mathbb{R}^d$.

$C \subseteq S$ is open relative to $S$: For all $x \in C$, there is an $r > 0$ such that

$$B(x, r) \cap S = \{s \in S \mid \|s - x\| \leq r\} \subseteq C.$$

(This definition is for any $S \subseteq \mathbb{R}^d$, but we will only consider the case where $S$ is affine.)

In this class, if we say $C$ is relatively open without specifying $S$, then we mean $S = \operatorname{aff} C$.

# Relative interior

Relative interior of $C \subset \mathbb{R}^d$:

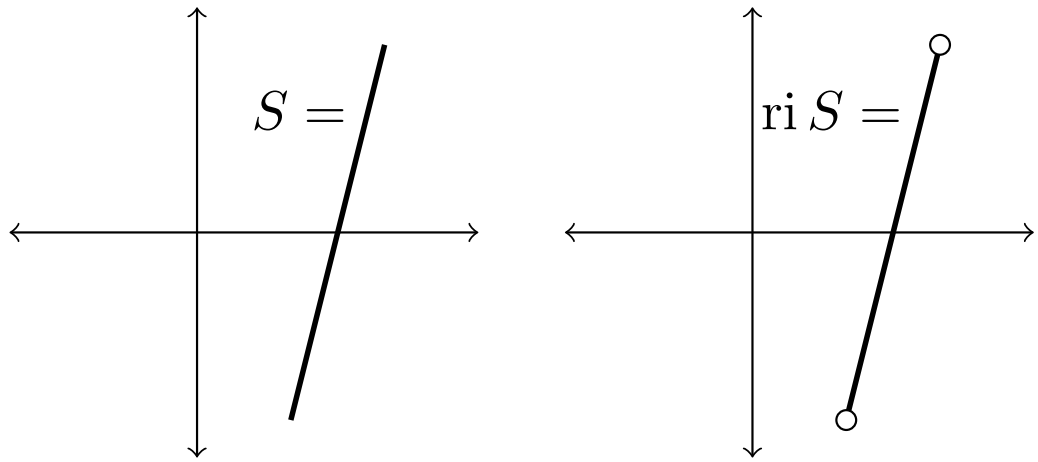
$$\operatorname{ri} C = \{x \in C \mid B(x, r) \cap \operatorname{aff} C \subseteq C \text{ for some } r > 0\}$$

If $C = \{x\} \subset \mathbb{R}^d$, then the definition implies $\operatorname{ri}\{x\} = \{x\}$.

Relative boundary of $C \subseteq \mathbb{R}^d$: $\operatorname{cl} C \backslash \operatorname{ri} C$

# Relative interior example

$$S = \left\{ (x, y) \in \mathbb{R}^2 \,|\, x \in [0.5, 1],\, y = 4x - 3 \right\}.$$



# Relative interior is nonempty

## Lemma

*Let $C \subseteq \mathbb{R}^d$ be a nonempty convex set. Then, $\operatorname{ri} C \neq \emptyset$.*

**Proof.** Let $x_0 \in C$. If $C = \{x_0\}$ is a singleton, then $\operatorname{ri}\{x_0\} = \{x_0\} \neq \emptyset$. Now assume $C$ has at least two elements.

Since $C - x_0$ has a nonzero element, it has a basis (a maximal linearly independent subset) $\{b_1, \ldots, b_k\}$. Write $M = \operatorname{span}\{b_1, \ldots, b_k\} \subseteq \mathbb{R}^d$. Then, $\varphi\colon \mathbb{R}^k \to M$ defined by

$$\varphi(\alpha_1, \ldots, \alpha_k) = \sum_{i=1}^{k} \alpha_i b_i$$

is a one-to-one linear isomorphism. Since $C - x_0$ is convex and $0 \in C - x_0$, we have that

$$\varphi\big(\{(\alpha_1, \ldots, \alpha_k) \,|\, \alpha_1 + \cdots + \alpha_k < 1,\ \alpha_i > 0 \text{ for } i = 1, \ldots, k\}\big) \subset C - x_0.$$

Since $\varphi$ maps an open set to a (relatively) open set, we conclude that $C - x_0$ has nonempty relative interior. □

# Relative interior through explicit parameterization

## Lemma

*Let $C \subseteq \mathbb{R}^d$ be a nonempty convex set, and let*

$$\operatorname{aff}(C) = x_0 + \mathcal{V}$$

*such that $x_0 \in \mathcal{V}^\perp$. Let the columns of $U \in \mathbb{R}^{d \times r}$ form an orthonormal basis of $\mathcal{V}$, then,*

$$C = \{x_0 + Uy \,|\, y \in U^\intercal C\},$$

*and*

$$\operatorname{ri} C = \{x_0 + Uy \,|\, y \in \operatorname{int}(U^\intercal C)\}.$$

**Proof.** Note that $U^\intercal U = I$ by orthonormality and $UU^\intercal = \mathrm{Proj}_{\mathcal{V}}$.

If $x \in C \subseteq \mathrm{aff}\,(C) = x_0 + \mathcal{V}$, then $x = x_0 + v$ with $x_0 \in \mathcal{V}^\perp$ and $v \in \mathcal{V}$, and

$$x = x_0 + v = x_0 + UU^\intercal v = x_0 + UU^\intercal x \in \{x_0 + Uy \,|\, y \in U^\intercal C\}.$$

Therefore,

$$C \subseteq \{x_0 + Uy \,|\, y \in U^\intercal C\}.$$

On the other hand, let $x' \in \{x_0 + Uy \,|\, y \in U^\intercal C\}$, i.e., $x' = x_0 + UU^\intercal x$ for some $x \in C$. Since $x \in C \subseteq \mathrm{aff}\,(C) = x_0 + \mathcal{V}$, we have $x = x_0 + v$ with $x_0 \in \mathcal{V}^\perp$ and $v \in \mathcal{V}$. Therefore,

$$x' = x_0 + UU^\intercal x = x_0 + v = x \in C,$$

and

$$C \supseteq \{x_0 + Uy \,|\, y \in U^\intercal C\}.$$

We have established

$$C = \{x_0 + Uy \,|\, y \in U^\intercal C\},$$

which is equivalent to

$$C - x_0 = \{Uy \,|\, y \in U^\intercal C\} = U\big(U^\intercal C\big).$$

Next, note that $U : \mathbb{R}^k \to \mathcal{V}$ is a one-to-one linear isomorphism that maps open sets to (relatively) open sets. Therefore, $U$ maps the interior $U^\intercal C$ to the interior of $C - x_0$ relative to $\mathcal{V}$, i.e.,

$$\operatorname{ri}(C - x_0) = \{Uy \,|\, y \in \operatorname{int}(U^\intercal C)\},$$

and we conclude the statement after translation by $x_0$. □

# Existence of a subgradient

## Theorem

*Let $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be proper and convex. If $x \in \operatorname{ri}\operatorname{dom} f$ ($x$ in relative interior of $\operatorname{dom} f$), then exists a subgradient of $f$ at $x$.*

**Proof.** Let

$$\operatorname{aff}(\operatorname{dom} f) = x_0 + \mathcal{V}$$

where $x_0 \in \mathcal{V}^\perp$. Let the columns of $U \in \mathbb{R}^{d\times r}$ form an orthonormal basis of $\mathcal{V}$. Then, by the previous lemma,

$$\operatorname{ri}\operatorname{dom} f = \{x_0 + Uy \,|\, y \in \operatorname{int}(U^\intercal \operatorname{dom} f)\}.$$

Define $\tilde{f}\colon \mathbb{R}^r \to \mathbb{R} \cup \{\infty\}$ as

$$\tilde{f}(\tilde{x}) = f(x_0 + U\tilde{x}), \qquad \forall\, \tilde{x} \in \mathbb{R}^r.$$

Then, $\tilde{f}$ is proper and convex. (Recall that $f(x) = \infty$ outside of $x_0 + \mathcal{V}$.) Since $U^\intercal U = I$ and $U^\intercal(x_0 + U\tilde{x}) = \tilde{x}$, we have

$$U^\intercal \operatorname{dom} f = \operatorname{dom} \tilde{f}$$

$$U^\intercal \operatorname{ri}\operatorname{dom} f = \operatorname{int}(U^\intercal \operatorname{dom} f) = \operatorname{int}\operatorname{dom} \tilde{f}.$$

Let $x \in \operatorname{ri}\operatorname{dom} f$ and $\tilde{x} = U^\intercal x \in \operatorname{int}\operatorname{dom}\tilde{f}$. Then, by a previous theorem, $\tilde{f}$ has a subgradient $\tilde{g}$ at $\tilde{x}$, and

$$\tilde{f}(\tilde{y}) \geq \tilde{f}(\tilde{x}) + \langle \tilde{g}, \tilde{y} - \tilde{x} \rangle, \qquad \forall\, \tilde{y} \in \mathbb{R}^r.$$

Let $g = U\tilde{g}$. Then, using $U^\intercal U = I$,

$$\begin{aligned} f(x_0 + U\tilde{y}) &\geq f(x_0 + U\tilde{x}) + \langle \tilde{g}, U^\intercal U(\tilde{y} - \tilde{x}) \rangle \\ &= f(x_0 + U\tilde{x}) + \langle g, U(\tilde{y} - \tilde{x}) \rangle \\ &= f(x_0 + U\tilde{x}) + \langle g, x_0 + U\tilde{y} - (x_0 + U\tilde{x}) \rangle \end{aligned}$$

for all $\tilde{y} \in \mathbb{R}^r$, i.e., and

$$f(y) \geq f(x) + \langle g, y - x \rangle$$

for all $y \in x_0 + \mathcal{V}$. Since the inequality vacuously holds for all $y \notin x_0 + \mathcal{V}$ (LHS $= \infty$), we conclude $g \in \partial f(x)$. □