MATH562 - Theory of Machine Learning

Based on lectures from Winter 2026 by Prof. Courtney Paquette.
Notes by Louis Meunier

Contents

§1 Introduction

§1.1 Some Linear Algebra

1.1.1 Inverting Block Matrices

Let

$$M := \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in \mathbb{R}^{(p+q)\times(p+q)},$$

i.e. $A \in \mathbb{R}^{p\times p}, B \in \mathbb{R}^{p\times q}, C \in \mathbb{R}^{q\times p}$ and $D \in \mathbb{R}^{q\times q}$ (where we use the convention that if $A \in \mathbb{R}^{m\times n}$, then A has m rows and n columns, so in particular maps $\mathbb{R}^n \to \mathbb{R}^m$). If A is invertible, let

$$M \setminus A := D - CA^{-1}B =: \text{Schur Complement (of } A \text{ with respect to } M).$$

Then,

$$M^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(M \setminus A)^{-1}CA^{-1} & -A^{-1}B(M \setminus A)^{-1} \\ -(M \setminus A)^{-1}CA^{-1} & (M \setminus A)^{-1} \end{pmatrix}.$$

Similarly, if D invertible and $M \setminus D := A - BD^{-1}C$, then

$$M^{-1} = \begin{pmatrix} (M \setminus D)^{-1} & -(M \setminus D)^{-1}BD^{-1} \\ -D^{-1}C(M \setminus D)^{-1} & D^{-1} + D^{-1}C(M \setminus D)^{-1}BD^{-1} \end{pmatrix}.$$

1.1.2 Eigenvalues and Singular Values

Given $A \in \mathbb{R}^{n\times n}$ symmetric, there exists $U \in \mathbb{R}^{n\times n}$ orthogonal (i.e., $U^T = U^{-1}$) such that

$$A = U \operatorname{diag}(\lambda) U^T,$$

where $\lambda = (\lambda_1, ..., \lambda_n)$ for λ_i's the eigenvectors of A. In particular, if $U^{(i)}$ enumerate the columns of U, we have

$$AU^{(i)} = \lambda_i U^{(i)},$$

i.e. the $U^{(i)}$'s are the eigenvectors of A.

Given $X \in \mathbb{R}^{n\times d}, n \geq d$, then there exists an orthogonal matrix $V \in \mathbb{R}^{d\times d}$ and $U \in \mathbb{R}^{n\times d}$ with orthogonal columns, and a matrix of *singular values* $s \in \mathbb{R}^d_+ = \left\{(v_1, ..., v_d) \in \mathbb{R}^d \mid v_i \geq 0 \forall i = 1, ..., d\right\}$ such that

$$X = U \operatorname{Diag}(s) V^T.$$

Remark 1.1:

1. if $u_i \in \mathbb{R}^n, v_j \in \mathbb{R}^d$ are the columns of U, V resp., then $X = \sum_{i=1}^d s_i u_i v_i^T$
2. if s_i a singular value of X, then s_i^2 an eigenvalue of XX^T and X^TX.

§1.2 Concentration Inequalities

Here we study the question of how the magnitude of the average of n independent, mean 0 random variables behaves compared to their average magnitude, specifically with respect to n.

We know that the central-limit theorem states that for z_i iid with variance σ^2, $\sqrt{n}\left(\frac{1}{n}\sum z_i - \mathbb{E}[z]\right)$ converges in distribution to a $\mathcal{N}(0, \sigma^2)$; this is an asymptotic result, which gives no

information about the rate of this converge with respect to n, which is what we care about in our study.

↪**Proposition 1.1** (Markov's): Let Y be a nonnegative r.v. with finite mean. Then,

$$\mathbb{P}(Y \geq \varepsilon) \leq \frac{1}{\varepsilon}\mathbb{E}[Y], \quad \forall \varepsilon > 0.$$

↪**Proposition 1.2** (Chebyshev's): Let X be a r.v. with finite mean and variance, then

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\mathrm{Var}[X]}{\varepsilon^2}, \quad \forall \varepsilon > 0.$$

↪**Corollary 1.1**: If $z_i, i = 1, ..., n$ are iid with variance σ^2, then

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} z_i - \mathbb{E}[z]\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

↪**Proposition 1.3** (Union Bound, Max/Tail Bound):
1. $\mathbb{P}\left(\bigcup_{f\in\mathcal{F}} A_f\right) \leq \sum_{f\in\mathcal{F}} \mathbb{P}\left(A_f\right)$
2. $\mathbb{P}\left(\sup_{f\in\mathcal{F}} Z_f \geq t\right) \leq \sum_{f\in\mathcal{F}} \mathbb{P}\left(Z_f \geq t\right)$

↪**Proposition 1.4** (Jensen's Inequality): If $F : \mathbb{R} \to \mathbb{R}$ convex and X an r.v. with finite mean,

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

1.2.1 Hoeffding Inequality

↪**Proposition 1.5** (Hoeffding Inequality): Let $z_1, ..., z_n$ be independent r.v.s with $z_i \in [0, 1]$ a.s.. Then, for any $t \geq 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} z_i - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[z_i] \geq t\right) \leq \exp(-2nt^2).$$

Remark 1.2: Read this result as a *fast* (exponential) convergence of the empirical mean to the true mean as the sample size n grows.

Proof. First we claim that

$$z \in [0, 1] \text{ a.s.} \Rightarrow \mathbb{E}[\exp(s(z - \mathbb{E}[z]))] \leq \exp\left(\frac{s^2}{8}\right). \quad (\dagger)$$

We'll assume z is centered for the sake of notation. Let $\varphi(s) := \log(\mathbb{E}[\exp(sz)])$. Remark that

$$\varphi(0) = 0$$

$$\varphi'(s) = \frac{\mathbb{E}[z\exp(sz)]}{\mathbb{E}[\exp(sz)]}$$

$$\varphi''(s) = \frac{\mathbb{E}\big[z^2\exp(sz)\big]}{\mathbb{E}[\exp(sz)]} - \left(\frac{\mathbb{E}[z\exp(sz)]}{\mathbb{E}[\exp(sz)]}\right)^2.$$

In particular, if we define a new probability density

$$\tilde{z} \mapsto \frac{e^{s\tilde{z}}}{\mathbb{E}[e^{sz}]}$$

with respect to that of z, and let $\tilde{z}$ be distributed with respect to this distibution, then

$$\mathrm{Var}(\tilde{z}) = \varphi''(s).$$

Note that $\tilde{z} \in [0,1]$ a.s.. In addition, we have that

$$\begin{aligned}\mathrm{Var}(\tilde{z}) &= \inf_{v\in[0,1]}\mathbb{E}\big[(\tilde{z}-v)^2\big] \\ &\leq \mathbb{E}\left[\left(\tilde{z}-\frac{1}{2}\right)^2\right] = \frac{1}{4}\mathbb{E}\left[\left(\underbrace{2\tilde{z}-1}_{\leq 1 \text{ a.s.}}\right)^2\right] \leq \frac{1}{4},\end{aligned}$$

so that $\varphi''(s) \leq \frac{1}{4}$ for all s. Thus, by Taylor expanding φ, we find

$$\varphi(s) \leq \varphi(0) + \varphi'(0)s + \frac{s^2}{2}\frac{1}{4} = \frac{s^2}{8},$$

using the bound above and the fact $\varphi'(0) = 0$ (checking the above formula). Thus,

$$\varphi(s) = \log(\mathbb{E}[\exp(sz)]) \leq \frac{s^2}{8},$$

from which the claim $(\dagger)$ follows by taking exp of both sides.

Next, let $t \geq 0$ and put $\bar{z} = \frac{1}{n}\sum z_i$. Then,

$$\begin{aligned}\mathbb{P}(\bar{z} - \mathbb{E}[\bar{z}] \geq t) &= \mathbb{P}(\exp(s(\bar{z} - \mathbb{E}[\bar{z}])) \geq \exp(st)) \\ \text{(Markov's)} \quad &\leq e^{-st}\mathbb{E}[\exp(s(\bar{z} - \mathbb{E}[\bar{z}]))] \\ \text{(Indep.)} \quad &= e^{-st}\prod_{i=1}^{n}\mathbb{E}\left[\exp\left(\frac{s}{n}(z_i - \mathbb{E}[z_i])\right)\right] \\ (\dagger) \quad &\leq e^{-st}\prod_{i=1}^{n}\exp\left(\frac{s^2}{8n^2}\right) = \exp\left(-st + \frac{s^2}{8n}\right).\end{aligned}$$

This bound held for all s, so in particular holds at $\bar{s} = \mathrm{argmin}\left\{-st + \frac{s^2}{8n}\right\} = 4nt$. Plugging in this value for s gives the result. ■

↪**Corollary 1.2** (2-sided Hoeffding): With the same hypotheses as the previous proposition, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum z_i - \frac{1}{n}\sum \mathbb{E}[z_i]\right| \geq t\right) \leq 2\exp(-2nt^2), \forall t \geq 0.$$

If instead $z_i \in [a,b]$ a.s., we can replace the rhs with

$$\leq 2\exp\left(\frac{-2nt^2}{(a-b)^2}\right).$$

Remark 1.3:

1. How is Hoeffding used? Start with a probability, then derive the necessary t (usually, as a function of n) to achieve that bound. e.g., if $z_i \in [a,b] a.s.$ and for any $\delta \in (0,1)$, then with probability $1-\delta$,

$$\left|\frac{1}{n}\sum z_i - \frac{1}{n}\sum \mathbb{E}[z_i]\right| < \frac{|a-b|}{\sqrt{2n}}\sqrt{\log\left(\frac{2}{\delta}\right)}$$

2. An extension exists for martingales. If $Z_i, i = 1, \ldots, n$ martingales with respect to a filtration $\{\mathcal{F}_i\}$ and $|Z_i| \leq c_i$ a.s., then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \geq t\right) \leq \exp\left(-\frac{n^2t^2}{2\|c\|^2}\right), \quad c := (c_1, \ldots, c_n).$$

↪**Definition 1.1** (Sub-Gaussian): We say an r.v. X is *sub-Gaussian* if there exists $\tau \in \mathbb{R}_+$ such that

$$\mathbb{E}[\exp(s(X - \mathbb{E}[X]))] \leq \exp\left(\frac{\tau^2}{2}s^2\right), \quad \forall s \in \mathbb{R}.$$

We define the *sub-Gaussian norm* by

$$\|X\|_{\psi_2} := \inf\left\{k \geq 0 : \mathbb{E}\left[\exp\left(\frac{X^2}{k^2}\right)\right] \leq 2\right\},$$

i.e. the "best" sub-Gaussian parameter for X.

Remark 1.4: Interpretation: X has tails decaying as fast (or faster) than a Gaussian.

Remark 1.5: *Different texts may define this differently, i.e. with/without a 2 factor under the τ^2. The notational advantage of this definition is that a Gaussian random variable with variance σ^2 has sub-Gaussian parameter σ.*

↪**Proposition 1.6**: X is sub-Gaussian iff there exists a $k \in \mathbb{R}_+$ such that

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq 2\exp\left(-\frac{t^2}{k^2}\right), \quad \forall t \in \mathbb{R}.$$

↪**Definition 1.2** (Sub-Exponential): We say X *sub-exponential* if

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq 2\exp\left(-\frac{t}{k}\right),$$

for some k and for all $t \geq 0$. We define the *sub-Gaussian norm* by

$$\|X\|_{\psi_1} := \inf\left\{k \geq 0 : \mathbb{E}\left[\exp\left(\frac{|X|}{k^2}\right)\right] \leq 2\right\},$$

i.e. the "best" sub-Gaussian parameter for X.

Remark 1.6: This is a similar, but slower, tail bound than sub-Gaussian.

1.2.2 McDiamarid's Inequality

For a measure space Z and nonnegative integer n, we say $f : Z^n \to \mathbb{R}$ is a function of bounded variation with constant c if for all $i \in [n] := \{1, ..., n\}$ and points $z_1, ..., z_n, z_i' \in Z$, then

$$|f(z_1, ..., z_i, ..., z_n) - f(z_1, ..., z_i', ..., z_n)| \leq c.$$

↪**Proposition 1.7** (McDiamarid's Inequality): Let $z_1, ..., z_n$ be independent r.v.s on some measure space Z and $f : Z^n \to \mathbb{R}$ be a function of bounded variation with constant c. Then,

$$\mathbb{P}(|f(z_1, ..., z_n) - \mathbb{E}[f(z_1, ..., z_n)]| \geq t) \leq 2\exp\left(-\frac{2t^2}{nc^2}\right), \quad \forall t \geq 0.$$

Remark 1.7: We can extend this to $z_i \in \mathbb{R}^d$; if $\|z_i\|_2 \leq c$ a.s., then $\left\|\frac{1}{n}\sum z_i\right\|_2 \leq \frac{c}{\sqrt{n}}\left(1 + \sqrt{2\log\left(\frac{1}{\delta}\right)}\right)$ with probability $\geq 1 - \delta$.

1.2.3 Bernstein's Inequality

↪**Proposition 1.8** (Bernstein's): Let $z_i, i = 1, ..., n$ be independent with $|z_i| \leq c$ a.s. and mean zero. Then for all $t \geq 0$,

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_i z_i\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2\sigma^2 + 2ct/3}\right), \quad \sigma^2 := \frac{1}{n}\sum_i \mathrm{Var}(z_i).$$

In particular, for $\delta \in (0, 1)$, then with probability $\geq 1 - \delta$,

$$\left|\frac{1}{n}\sum z_i\right| \leq \sqrt{\frac{2\sigma^2 \log\left(\frac{2}{\delta}\right)}{n}} + \frac{2c\log\left(\frac{2}{\delta}\right)}{3n}.$$

↪**Proposition 1.9** (Extension of Bernstein's, sub-exponential): Let $x_1, ..., x_n$ be mean zero, independent, sub-exponential r.v.s with constants k_i, and let $a \in \mathbb{R}^n$. Then, for all $t \geq 0$,

$$\mathbb{P}\left(\left|\sum a_i x_i\right| \geq t\right) \leq 2\exp\left(-c\min\left\{\frac{t^2}{k^2\|a\|_2^2}, \frac{1}{k\|a\|_\infty}\right\}\right).$$

↪**Proposition 1.10** (Extension of Bernstein's, non-zero means): With the same hypothesis as Bernstein's but without the zero means, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum z_i - \frac{1}{n}\sum \mathbb{E}[z_i]\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2\sigma^2 + 2c\frac{t}{3}}\right).$$

1.2.4 Expectation of the Maximum

↪**Proposition 1.11**: Let z_i be (possible dependent) mean-zero, $\mathbb{R}$-values r.v.s which are all sub-Gaussian with constant τ^2. Then,

$$\mathbb{E}[\max\{z_1, ..., z_n\}] \leq \sqrt{2\tau^2 \log(n)}.$$

Proof. For all $t > 0$,

$$\begin{aligned}
\mathbb{E}[\max\{z_1, ..., z_n\}] &\leq \frac{1}{t}\log)\mathbb{E}[\exp(t\max(Z_i))] \quad \text{(Jensen's)} \\
&= \frac{1}{t}\log(\mathbb{E}[\max\{\exp(tZ_i)\}]) \quad \text{(exp increasing)} \\
&\leq \frac{1}{t}\log\Big(\mathbb{E}\Big[\sum \exp(tZ_i)\Big]\Big) \quad \text{(max leq sum)} \\
&\leq \frac{1}{t}\log\left(n\exp\left(\tau^2\frac{t^2}{2}\right)\right) \quad \text{(sub-Gaussian)} \\
&= \frac{\log(n)}{t} + \frac{\tau^2 t}{2}.
\end{aligned}$$

The proof follows by taking $t := \tau^{-1}\sqrt{2\log(n)}$. ■

§2 Introduction to Supervised Learning

§2.1 Training Data Predictions

The goal of supervised learning is to take a series of observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i \in [n]$ (called *training data*) and to predict a new $y \in \mathcal{Y}$ given a (previously unseen) $x \in \mathcal{X}$ (*testing data*).

We write

- $\mathcal{X}$ for our space of *inputs*, typically embedded in $\mathbb{R}^d$ (where d tends to be very large; think images encoded as large matrices of pixels, text, videos, etc)
- $\mathcal{Y}$ for our space of *outputs* or *labels* for the data

The challenges in supervised learning are twofold:

1. $y \in \mathcal{Y}$ may not be a deterministic function of $x \in \mathcal{X}$
2. inputs may live a high-dimensional space, hence it is computationally expensive to work with them

We make two primary blanket assumptions of our problem:

1. we aim to maximize the expectation of some measure of performance with respect to some testing distribution we put on our data

2. we assume (x_i, y_i) are iid, with the training data having the same distribution as the testing data

↪**Definition 2.1** (Machine Learning (ML) Algorithm): An *ML algorithm* is a function from the data set, $\left(\mathcal{X} \times \mathcal{Y}\right)^n$ to a function $\mathcal{X} \to \mathcal{Y}$.

§2.2 Decision Theory

The question we aim to answer here is, what is the *optimal* performance of an algorithm, regardless of the finiteness of the data? I.e., if we havd perfect knowledge of the underlying probability distribution of our data, how should we design our algorithm?

We assume for now a fixed (testing) distribution $P_{x,y}$ on $\mathcal{X} \times \mathcal{Y}$ with P_x marginal distribution on $\mathcal{X}$.

2.2.1 Supervised Learning and Loss Functions

↪**Definition 2.2** (Loss Function): A *loss function* is a mapping $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ where $\ell(y, z)$ some measure of how close a true label y is to a predicted label z.

⊛ **Example 2.1**:

1. (Binary classification) Let $\mathcal{Y} = \{0, 1\}$, or even $\mathcal{Y} = \{0, ..., 9\}$. A typical loss on such labels is the "0-1 loss", $\ell(y, z) := \mathbb{1}_{\{y \neq z\}}$.
2. (Regression) Let $\mathcal{Y} = \mathbb{R}$, then two typical loss functions are the *mean-square loss*

$$\ell(y, z) := (y - z)^2$$

or *absolute loss*

$$\ell(y, z) := |y - z|.$$

2.2.2 Risks

↪**Definition 2.3** (Expected Risk): Given a prediction function $f : \mathcal{X} \to \mathcal{Y}$, a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and a probability distribution P on $\mathcal{X} \times \mathcal{Y}$, the *expected risk* of f is defined by

$$\mathcal{R}(f) := \mathbb{E}_{x,y}[\ell(y, f(x))] = \int_{\mathcal{X} \times \mathcal{Y}} \ell(y, f(x)) \, \mathrm{d}P(x, y).$$

↪**Definition 2.4** (Empirical Risk): Given a prediction function $f : \mathcal{X} \to \mathcal{Y}$, a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and $(x_i, y_i)_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$, the *empirical risk* is given by

$$\begin{aligned} \widehat{\mathcal{R}}(f) &:= \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) \\ &= \int_{\mathcal{X} \times \mathcal{Y}} \ell(y, f(x)) \, \mathrm{d}\mu(x, y), \quad \mu(x, y) := \frac{1}{n} \sum_{i=1}^{n} \delta_{\{(x_i, y_i)\}}. \end{aligned}$$

Remark 2.1: Heuristically, $\widehat{\mathcal{R}}(f)$ should approach $\mathcal{R}(f)$ as $n \to \infty$.

⊛ **Example 2.2**:

1. If $\mathcal{Y} = \{0,1\}, \ell(y,z) = \mathbb{1}_{\{y \neq z\}}$, then

   $$\mathcal{R}(f) = \mathbb{E}\left[\mathbb{1}_{\{y \neq f(z)\}}\right] = \mathbb{P}(y \neq f(x)) = \text{probability of missclassifying}$$

2. $\mathcal{Y} = \mathbb{R}, \ell(y,z) = (y-z)^2$,

   $$\mathcal{R}(f) = \mathbb{E}\left[(y - f(x))^2\right], \quad \text{mean-square error (MSE)}$$

2.2.3 Baye's Risk, Predictor

Here, we answer the question: what's the best predictor f we could find, assuming we knew everything about the underlying distibution on $\mathcal{X} \times \mathcal{Y}$?

We can write, by law of total expectation,

$$\begin{aligned}
\mathcal{R}(f) &= \mathbb{E}[\ell(y, f(x))] \\
&= \mathbb{E}[\mathbb{E}[\ell(y, f(x))|x]] \\
&= \mathbb{E}_{x' \sim p}[\mathbb{E}[\ell(y, f(x')) \mid x = x']] \\
&= \int_{\mathcal{X}} \mathbb{E}[\ell(y, f(x')) \mid x = x'] \, \mathrm{d}p(x').
\end{aligned}$$

Define the *conditional risk* given $x' \in \mathcal{X}$ by

$$r(z|x') := \mathbb{E}_y[\ell(y,z) \mid x = x'],$$

so that we can write

$$\mathcal{R}(f) = \int_{\mathcal{X}} r(f(x')|x) \, \mathrm{d}p(x') \overset{\mathcal{X} \text{ finite}}{=} \sum_{x' \in \mathcal{X}} r(f(x')|x') \mathbb{P}(x = x').$$

In particular, in the finite case, we can see that to minimize the risk $\mathcal{R}(f)$, we can minimize the individual conditional risks $r(f(x')|x')$ for each $x' \in \mathcal{X}$. The so-called *Baye's predictor* is a function f_* which for each x' minimizes $r(f(x')|x')$. Formally,

↪**Proposition 2.1** (Baye's Predictor/Risk): The expected risk is minimized at a *Baye's predictor* $f_* : \mathcal{X} \to \mathcal{Y}$ such that, for all $x' \in \mathcal{X}$,

$$f_*(x') \in \operatorname{argmin}_{z \in \mathcal{Y}} \mathbb{E}[\ell(y,z) \mid x = x'].$$

The *Baye's risk* is the risk of a (any) Baye's predictor, written

$$\mathcal{R}^* := \mathbb{E}_{x' \sim p}\left[\inf_{z \in \mathcal{Y}} \mathbb{E}[\ell(y,z) \mid x = x']\right] = \mathbb{E}[\ell(y, f_*(x'))].$$

Remark 2.2:

1. Finding an f_* is an impossible task in practice. Instead, we'll usually assume f takes some parametrized form, and optimize these parameters.
2. Baye's predictor may not be unique, but all Baye's predcitors have the same risk
3. Baye's risk is usually nonzero, unless the dependency between x and y is deterministic.

↪**Definition 2.5** (Excess Risk): The *excess risk* of a predictor f is the value $\mathcal{R}(f) - \mathcal{R}(f_*) \geq 0$.

Remark 2.3: Thus, if we knew the conditional distribution $(y|x)$ for each x, the optimal predictor is known. ML can be succinctly be described as dealing with the general case in which we do not know $(y|x)$ for all x, and can only work with given samples of data.

⊛ **Example 2.3**:

1. (Binary classification) With $\mathcal{Y} := \{-1, 1\}$ and $\ell(y, z) = \mathbb{1}_{\{y \neq z\}}$ the 0-1 loss, we can see that

$$\begin{aligned} f_*(x') &\in \operatorname{argmin}_{z' \in \{-1,1\}} P(y \neq z \mid x = x') \\ &= \operatorname{argmax}_{z \in \{-1,1\}} \mathbb{P}(y = z \mid x = x') \\ &= \begin{cases} 1 & \mathbb{P}(y = 1 \mid x = x') > \frac{1}{2} \\ -1 & \mathbb{P}(y = 1 \mid x = x') < \frac{1}{2} \\ \text{anything} & \mathbb{P}(y = 1 \mid x = x') = \frac{1}{2} \end{cases}. \end{aligned}$$

   Putting $\mathcal{L}(x') := \mathbb{P}(y = 1 \mid x = x')$, this implies

$$\mathcal{R}^* = \mathbb{E}[\min\{\mathcal{L}(x), 1 - \mathcal{L}(x)\}].$$

2. (Regression) With $\mathcal{Y} = \mathbb{R}$, $\ell(y, z) = (y - z)^2$, we see that

$$\begin{aligned} \operatorname{argmin}_{z \in \mathbb{R}} \mathbb{E}\big[(y - z)^2 \mid x = x'\big] = \operatorname{argmin}_{z \in \mathbb{R}} \Bigg\{ & \underbrace{\mathbb{E}\Big[(y - \mathbb{E}[y \mid x = x'])^2 \mid x = x'\Big]}_{\text{independent of } z} \\ & + \underbrace{(z - \mathbb{E}[y \mid x = x'])^2}_{\text{minimize this}} \Bigg\} \\ = \mathbb{E}[y \mid x = x']. \end{aligned}$$

   Hence, $f_*(x') = \mathbb{E}[y|x = x']$, and so

$$\mathcal{R}^* = \mathbb{E}_{x' \sim p}\Big[\inf_{z \in \mathbb{R}} \mathbb{E}\big[(y - z)^2 \mid x = x'\big]\Big] = \mathbb{E}_{x'}\Big[(y - \mathbb{E}[y|x = x'])^2\Big] \quad \text{(conditional variance)}$$

§2.3 Empirical Risk Minimization

We don't know the underlying distributions we work with (of course, otherwise we'd be done), and we need to work with samples, and need to simplify what kind of prediction functions we consider (since we don't know the underlying distribution, thus can't find the Baye's predictor in general).

We'll assume a parametrized family of predictor functions (called our *model*),

$$f_\theta : \mathcal{X} \to \mathcal{Y}, \quad \theta \in \Theta,$$

where $\Theta \subset \mathbb{R}^d$ typically. Heuristically, as d increases, if we could find the best f_θ predictor for $\theta \in \Theta$, that predictor will approach the Baye's predicotr.

↪**Definition 2.6** (Empirical risk with respect to a parameter): The *empirical risk w.r.t* $\theta \in \Theta$ is

$$\hat{R}(f_\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)).$$

We consider the optimal parameter minimizing this empirical risk as

$$\hat{\theta} \in \operatorname{argmin}_{\theta\in\Theta} \hat{R}(f_\theta),$$

and so our "optimal" prediction function with respect to Θ is $f_{\hat{\theta}}$.

⊛ **Example 2.4**: A typical linear least-squares problem takes this form,

$$\min_{\theta\in\Theta=\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \theta^T \varphi(x_i)\right)^2,$$

so that here, $f_\theta(x) = \theta^T \varphi(x_i)$ and our loss function is the square loss.

2.3.1 Risk Decomposition

Given a $\hat{\theta} \in \Theta$ (not necessarily optimal w.r.t Θ), we would like to break down the excess risk of the predictor $f_{\hat{\theta}}$ w.r.t the Baye's predictor to see the difference in error coming from our choice of model (we call this *approximation error,* i.e. how far our model is from approoximating our true predictor function) versus from the choice of $f_{\hat{\theta}}$ over the "true" best predictor with respect to Θ (as defined in the previous section. This is called the *estimation error*, and should be thought of as how well any underlying optimization algorithm used to find $\hat{\theta}$ performed compared to the theoretical best).

Mathematically, we can write

$$\underbrace{\mathcal{R}(f_{\hat{\theta}}) - \mathcal{R}^*}_{\substack{\text{Excess Risk}\\ \text{how good our estimator is}\\ \text{from the best possible}}} = \underbrace{\left\{\mathcal{R}(f_{\hat{\theta}}) - \inf_{\theta\in\Theta} \mathcal{R}(f_\theta)\right\}}_{\substack{\text{Estimation Error}\\ \text{how good our estimator is compared}\\ \text{to the best the model can do}}} + \underbrace{\left\{\inf_{\theta\in\Theta} \mathcal{R}(f_\theta) - \mathcal{R}^*\right\}}_{\substack{\text{Approximation Error}\\ \text{how good our model (theoretically) is}\\ \text{compared to the best possible}}}.$$

Note that the approximation error is due to the modelling choice, and is independent of the specific $f_{\hat{\theta}}$. Vaguely, "as Θ grows, the approximation error should shrink".

The estimation error can further be broken down into three parts; let $\theta' \in \Theta$ be the minimizer of $\theta \mapsto \mathcal{R}(f_{\theta'})$ (e.g., $\mathcal{R}(f_{\theta'}) = \inf_{\theta\in\Theta} \mathcal{R}(f_\theta)$), then

$$\underbrace{\mathcal{R}(f_{\hat{\theta}}) - \mathcal{R}(f_{\theta'})}_{\text{Estimation Error}} = \left\{\mathcal{R}(f_{\hat{\theta}}) - \widehat{\mathcal{R}}(f_{\hat{\theta}})\right\} \leftarrow \begin{array}{l}\text{How good the model risk is}\\ \text{on data vs true risk of model}\end{array}$$

$$\begin{array}{l}\underline{\text{Empirical Optimization Error}}\\ \text{How bad our choice of predictor is}\\ \text{compared true best in terms}\\ \text{of performance on the data}\\ (\text{for } \hat{\theta}\,)\end{array} \rightarrow \quad + \left\{\hat{R}(f_{\hat{\theta}}) - \hat{R}(f_{\theta'})\right\}$$

$$+ \left\{\hat{R}(f_{\theta'}) - \mathcal{R}(f_{\theta'})\right\} \leftarrow \begin{array}{l}\text{How good the model risk is}\\ \text{on data vs true risk of model}\\ (\text{for } \theta'\,)\end{array}$$

$$\leq \underbrace{2\sup_{\theta\in\Theta}|\mathcal{R}(f_\theta) - \widehat{\mathcal{R}}(f_\theta)|}_{\text{should} \downarrow \text{ as } n\uparrow} + \underbrace{\left\{\hat{R}(f_{\hat{\theta}}) - \hat{R}(f_{\theta'})\right\}}_{\begin{array}{c}\uparrow \text{ as } \Theta\uparrow\\ (\Theta \text{ gets too large to optimize over})\end{array}} .$$

In brief, we expect that as the parameter space Θ *grows*, the esimation error *increases*, but the approximation error *decreases*. But as n (number of samples) increases, we expect the estimation error to decrease (and there is no effect on the approximation error). Thus, there is a subtle interplay between $d := \dim(\Theta)$ and n.

§2.4 Statistical Learning Theory

"Statistical learning theory" asks how to provide guarantees of performance of an algorithm on previously unseen data.

We assume we have data

$$D_n(p) := \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

which are assumed to be iid from some unknown distribution p which is part of some family P of distributions.

An algorithm then is a mapping A from $D_n(p)$ to a prediction function $A(D_n(p)) = f : \mathcal{X} \to \mathcal{Y}$. Our goal is to find an algorithm such that the excess risk of the prediction function given by A is "small", in a sense we'll define in the next section.

2.4.1 Measures of Performance

↪**Definition 2.7** (Expected Risk): The *expected risk* of an algorithm A on sample size n and probability distribution p is the quantity

$$\mathbb{E}\left[\mathcal{R}_p\left(A\left(D_{n(p)}\right)\right)\right],$$

where the expected value is taken over all possible n-size subsets of the training data. We say that an algorithm is *consistent in expectation* if the above quantity converges, with p fixed, to $\mathcal{R}^*$ as $n \to \infty$.

↪**Definition 2.8** (Probability Approximately Correct $*$): We say an algorithm A is Probability Approximately Correct (PAC) if for any given $\delta \in (0, 1)$ and $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n \geq N$,

$$\mathbb{P}\left(\mathcal{R}_p(A(D_n(p))) - \mathcal{R}^* \leq \varepsilon\right) \geq 1 - \delta.$$

2.4.2 Notions of Consistency over Classes of Probability distributions

Remark that our definition of consistency in expectation gave no guarantee over rate of convergence, especially not with respect to the specific distribution.

↪**Definition 2.9**: An algorithm is *uniformly consistent* if for all probability distributions on (x, y), the algorithm is consistent.

↪**Definition 2.10** (Minimax risk): The minimax risk is defined to be, given $\mathcal{X} \times \mathcal{Y}$,

$$\inf_{A\text{: algorithm}} \sup_{\substack{p \in P\text{:} \\ \text{class of dists.}}} \left\{ \mathbb{E}\Big[\mathcal{R}_p(A(D_n(p))) \Big] - \mathcal{R}^* \right\}.$$

Remark 2.4: This is hard to evaluate in general, but is easy to upper bound (just fix any A and evaluate the inner supremum, i.e., look at the worst-case performance of the algorithm). Lower bounds are much harder to compute, since they need to hold for any possible algorithm.

§2.5 "No Free Lunch"

No, this section is not about SSMU shutting down Midnight Kitchen…

Here, we make clearer the remarks of the previous section in terms of performance of algorithms for arbitrarily distributed data. Namely, we show that, for a specific loss function and input/output space, for any size of data n, we can construct a distribution on our data such that any algorithm we can come up with will perform "poorly" (i.e. it's excess risk is bounded away from 0). Hence, there is no "free lunch", i.e no "easy algorithm" that will work without further assumptions on what our possible probability distributions could be

↪**Proposition 2.2** (No Free Lunch): Consider a binary classification with $0-1$ loss and with $\mathcal{X}$ infinite. Let $\mathcal{P}$ be the class of all probability distributions on $\mathcal{X} \times \{0, 1\}$. Then, for all n and for all algorithms A,

$$\sup_{p \in \mathcal{P}} \left\{ \mathbb{E}\Big[\mathcal{R}_p(A(D_n(p))) \Big] - \mathcal{R}^* \right\} \geq \frac{1}{2}.$$

Remark 2.5: As we'll see in the proof, the bounds we obtain will not give any rate in n, asymptotic or not. Indeed, the probability distribution for each n will crucially depend on a certain parameter n being much larger than n. Indeed, we can state (but will not prove) the much stronger statement as follows.

↪**Theorem 2.1** (Devroye, '96): Consider the same setup as Prop. 2.2. Then, for any decreasing sequence $a_n \to 0$ with $a_1 \leq \frac{1}{16}$, then for any algorithm A, there exists a $p \in \mathcal{P}$ such that for all $n \geq 1$,

$$\mathbb{E}\Big[\mathcal{R}_p(A(D_n(p))) \Big] - \mathcal{R}^* \geq a_n.$$

I.e., the supremum over $\mathcal{P}$ has excess risk going to zero *arbitrarily slowly*.

Proof. (of Prop. 2.2) Fix $n \in \mathbb{N}$ and assume wlog $\mathbb{N} \subset \mathcal{X}$ (by relabelling otherwise). Let $k \in \mathbb{Z}_+$, and, given a k-length vector $r = (r_1, ..., r_k) \in \{0,1\}^k$, define a joint probability distribution p on (x, y) such that

$$\mathbb{P}\big(x = j, y = r_j\big) = \frac{1}{k}, \quad j = 1, ..., k.$$

In particular, in this case, y is a deterministic function of x; given $x = j$, $y = r_j$. In particular, this means $\mathcal{R}^* = 0$.

Denote $\hat{f}_{D_n} := A(D_n(p))$ as the classifier under p given by algorithm A, and write $S(r) := \mathbb{E}\Big[\mathcal{R}_p\big(\hat{f}_{D_n}\big)\Big]$ as the expectation of the expected risk under this given probability distribution of the classifier given by the algorithm A for the given vector $r \in \{0,1\}^k$. We aim to pick r such that we maximize this quantity; if we can pick r such that this quantity is larger than $\frac{1}{2}$, we'll be done (why?).

This is hard to do directly, so we'll instead lower bound the max probabilistically; given any distribution q on $\{0,1\}^k$, we certainly have

$$\max_{r \in \{0,1\}^k} S(r) \geq \mathbb{E}_{r \sim q}[S(r)].$$

Thus, we'll design some q so that this quantity on the right is large. Specifically, let q be uniform on $\{0,1\}^k$, i.e. each $r = (r_1, ..., r_k)$ a vector of r_j's each independent, unbiased, Bernoulli r.v.'s. Then,

$$\mathbb{E}_{r \sim q}[S(r)] = \mathbb{P}\big(\hat{f}_{D_n}(x) \neq y\big) = \mathbb{P}\big(\hat{f}_{D_n}(x) \neq r_x\big),$$

which follows from the fact that we can write

$$\begin{aligned}
\mathbb{E}[S(r)] &= \mathbb{E}_r\Big[\mathbb{E}\Big[\mathcal{R}_p\big(\hat{f}_{D_n}\big)\Big]\Big] \\
&= \mathbb{E}_r\Big[\mathbb{E}_p\Big[\mathbb{E}_{x,y}\Big[\ell\big(x, \hat{f}_{D_n}(y)\big)\Big]\Big]\Big] \\
&= \mathbb{E}_{r,p,(x,y)}\Big[\mathbb{1}_{y \neq \hat{f}_{D_n}(x)}\Big] = \mathbb{P}\big(y \neq \hat{f}_{D_n}(x)\big),
\end{aligned}$$

just by unpacking all of the definitions. Continuing from above, we can write then

$$\begin{aligned}
\mathbb{E}_{r \sim q}[S(r)] &= \mathbb{E}_p\Big[\mathbb{P}\big(\hat{f}_{D_n}(x) \neq r_x\big) \mid x_1, ..., x_n; r_{x_1}, ..., r_{x_n}\Big] \quad \text{(total expectation)} \\
&\geq \mathbb{E}\Big[\mathbb{P}\big(\hat{f}_{D_n}(x) \neq r_x, x \neq x_1, ..., x_n\big) \mid x_1, ...\Big].
\end{aligned}$$

By Baye's rule,

$$\mathbb{P}\big(\hat{f}_{D_n}(x) \neq r_x, x \neq x_1, ..., x_n \mid x_1, ..., x_n\big) = \underbrace{\mathbb{P}\big(\hat{f}_{D_n}(x) \neq r_x \mid x \neq x_1, ..., x_n; x_1, ..., x_n\big)}_{=\frac{1}{2}} \cdot \mathbb{P}(x \neq x_1, ..., x_n \mid x_1, ..., x_n),$$

since, supposing we didn't observe x, x has equal probability of being labeled $0, 1$. Thus, all together,

$$\begin{aligned}
\mathbb{E}[S(r) &\geq \frac{1}{2}\mathbb{E}[\mathbb{P}(x \notin \{x_1, ..., x_n\} \mid x_1, ..., x_n)] \\
&= \frac{1}{2}\mathbb{P}(x \notin \{x_1, ..., x_n\}) \\
&= \frac{1}{2}\mathbb{E}[\mathbb{P}(x \notin \{x_1, ..., x_n\} \mid x)] \\
&= \frac{1}{2}\mathbb{E}\left[\prod_{i=1}^{n} \mathbb{P}(x \neq x_i \mid x)\right] \quad \text{(independence)} \\
&= \frac{1}{2}\left(1 - \frac{1}{k}\right)^n.
\end{aligned}$$

We have n fixed; as $k \to \infty$, this quantity $\to \frac{1}{2}$, proving the result. ■

§3 Linear Least Squares

§3.1 Framework

Goal: Consider $f_\theta : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}$, for some parameter $\theta \in \Theta \subset \mathbb{R}^d$, and minimize the empirical risk

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2.$$

Specifically, we'll study when f_θ is linear in θ, but not necessarily x, i.e.

$$f_\theta(x) \coloneqq \sum_{i=1}^{d} a_i(x)\theta_i = \varphi(x)^T\theta,$$

where $\varphi(x) = (a_1, ..., a_d)^T(x) \in \mathbb{R}^d$. Our goal then will be to compute

$$\min_{\theta \in \mathbb{R}^d} \left\{ \widehat{\mathcal{R}}(\theta) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \varphi(x_i)^T\theta\right)^2 \right\}.$$

or equivalently, writing $y = (y_1, ..., y_n)^T$ and

$$\Phi(x) = \begin{pmatrix} \vdots \\ -\varphi(x_i)^T- \\ \vdots \end{pmatrix} \in \mathbb{R}^{n \times d},$$

then

$$\hat{R}(\theta) = \frac{1}{n}\|y - \Phi(x)\theta\|^2.$$

§3.2 Ordinary Least Squares

Assume Φ from above has full column rank, i.e. $d \leq n$ (we say the problem then is "overdetermined/underparametrized"). This implies $\Phi^T\Phi \in \mathbb{R}^{d \times d}$ is invertible.

↪**Proposition 3.1** (OLS): When Φ has rank d, the minimizer of what we now call the *ordinary least squares problem* (OLS) is unique, and given by

$$\hat{\theta} = \left(\Phi^T \Phi\right)^{-1} \Phi^T y.$$

In particular, we call the relation

$$\Phi^T \Phi \hat{\theta} = \Phi^T y$$

the *normal equations*.

Proof. By homework (this is just a quadratic). ■

§3.3 Statistical Analysis of OLS

There are two main assumptions on OLS we will study:

1. *Random design setting:* assume both the inputs and outputs are random
2. *Fixed design setting:* assume the inputs are fixed, but the outputs are random. In this case, φ is deterministic, and thus our goal is to minimize

$$\mathcal{R}(\theta) = \mathbb{E}_y\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \varphi(x_i)^T\theta\right)^2\right].$$

§3.4 Fixed Design

We assume the following:

1. Φ is deterministic, and $\hat{\Sigma} := \frac{1}{n}\Phi^T\Phi$ is invertible
2. $\exists \theta_* \in \mathbb{R}^d$ such that $y_i = \varphi(x_i)^T \varphi_* + \varepsilon_i$
3. ε_i's are independent with mean zero and variance σ^2. We define $\varepsilon := (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$.

↪**Proposition 3.2** (Risk Decomposition of OLS - Fixed Design): Under the linear model and the assumptions above, for $\theta \in \mathbb{R}^d$, $\mathcal{R}^* = \sigma^2$ and the excess risk is given by

$$\mathcal{R}(\theta) - \mathcal{R}^* = \|\theta - \theta_*\|_{\hat{\theta}}^2 = (\theta - \theta_*)^T \hat{\Sigma}(\theta - \theta_*).$$

If $\hat{\theta}$ a random variable, then

$$\mathbb{E}_{\hat{\theta}}\left[\mathcal{R}\left(\hat{\theta}\right) - \mathcal{R}^*\right] = \underbrace{\left\|\mathbb{E}\left[\hat{\theta}\right] - \theta_*\right\|_{\hat{\Sigma}}^2}_{\text{bias}} + \underbrace{\mathbb{E}\left[\left\|\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right\|_{\hat{\Sigma}}^2\right]}_{\text{variance}}.$$

Remark 3.1: Since y has some noise, it makes sense to assume $\hat{\theta}$ could be random in its own right.

Proof. Using $y = \Phi\theta_* + \varepsilon$, we readily see

$$
\begin{aligned}
\mathcal{R}(\theta) &= \mathbb{E}_y\left[\frac{1}{n}\|y - \Phi\theta\|_2^2\right] \\
&= \mathbb{E}_\varepsilon\left[\frac{1}{n}\|\Phi(\theta_* - \theta) + \varepsilon\|^2\right] \\
&= \frac{1}{n}\mathbb{E}_\varepsilon\left[\underbrace{\|\Phi(\theta_* - \theta)\|_2^2}_{\perp\varepsilon} + \underbrace{\|\varepsilon\|_2^2}_{\rightsquigarrow n\sigma^2} + \underbrace{2(\Phi(\theta_* - \theta))^T\varepsilon}_{\text{mean zero}}\right] \\
&= \frac{1}{n}\|\Phi(\theta_* - \theta)\|_2^2 + \sigma^2 \\
&= \sigma^2 + (\theta_* - \theta)^T\underbrace{\frac{\Phi^T\Phi}{n}}_{=\hat{\Sigma}}(\theta_* - \theta) \\
&= \sigma^2 + \|\theta_* - \theta\|_{\hat{\Sigma}}^2.
\end{aligned}
$$

It's clear that this is minimized at $\theta = \theta_*$ (uniquely, since $\hat{\Sigma}$ invertible), which thus has risk $\mathcal{R}_* = \sigma^2$.

Suppose now $\hat{\theta}$ random. Then,

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{R}\left(\hat{\theta}\right) - \mathcal{R}^*\right] &= \mathbb{E}\left[\left\|\hat{\theta} - \theta_*\right\|_{\hat{\Sigma}}^2 \pm \mathbb{E}\left[\hat{\theta}\right]\right] \\
&= \mathbb{E}\left[\left\|\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right\|_{\hat{\Sigma}}^2\right] + \underbrace{2\mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)^T\hat{\Sigma}\left(\mathbb{E}\left[\hat{\theta}\right] - \theta_*\right)\right]}_{=0} + \mathbb{E}\left[\left\|\mathbb{E}\left[\hat{\theta}\right] - \theta_*\right\|_{\hat{\Sigma}}^2\right].
\end{aligned}
$$

■

3.4.1 Statistical Properties of the OLS Estimator

Recall the OLS estimator,

$$
\hat{\theta} = \left(\Phi^T\Phi\right)^{-1}\Phi^T y = \hat{\Sigma}^{-1}\left(\frac{1}{n}\Phi^T y\right), \quad y = \Phi\theta_* + \varepsilon,
$$

where the only randomness comes from ε.

↪**Proposition 3.3** (Estimation Properties of OLS): The OLS estimator $\hat{\theta}$ has the following properties:

1. $\mathbb{E}\left[\hat{\theta}\right] = \theta_*$, i.e. $\hat{\theta}$ is unbiased
2. $\mathrm{Var}\left(\hat{\theta}\right) = \mathbb{E}\left[\left(\hat{\theta} - \theta_*\right)\left(\hat{\theta} - \theta_*\right)^T\right] = (\sigma^2/n)\hat{\Sigma}^{-1}$, where $\hat{\Sigma}^{-1}$ is often called the *precision matrix*.

Proof. Since $\mathbb{E}_\varepsilon[y] = \mathbb{E}[\Phi\theta_* + \varepsilon] = \Phi\theta_*$, we find

$$
\mathbb{E}\left[\left(\Phi^T\Phi\right)^{-1}\Phi^T y\right] = \left(\Phi^T\Phi\right)^{-1}\Phi^T\Phi\theta_* = \theta_*,
$$

since Φ is deterministic.

Next, for the variance, compute

$$\hat{\theta} - \theta_* = \left(\Phi^T\Phi\right)^{-1}\Phi^T\varepsilon,$$

so that

$$\begin{aligned}
\mathrm{Var}\left(\hat{\theta}\right) &= \mathbb{E}\left[\left(\Phi^T\Phi\right)^{-1}\Phi^T\varepsilon\left(\left(\Phi^T\Phi\right)^{-1}\Phi^T\varepsilon\right)^T\right] \\
&= \mathbb{E}\left[\left(\Phi^T\Phi\right)^{-1}\Phi^T\varepsilon\varepsilon^T\Phi\left(\Phi^T\Phi\right)^{-1}\right] \\
&= \sigma^2\left(\Phi^T\Phi\right)^{-1}\Phi^T\Phi\left(\Phi^T\Phi\right)^{-1} \\
&= \sigma^2\left(\Phi^T\Phi\right)^{-1} = \frac{\sigma^2}{n}\hat{\Sigma}^{-1},
\end{aligned}$$

as claimed. ■

↪**Proposition 3.4** (Risk of OLS): The excess risk of the OLS estimator is

$$\mathbb{E}\left[\mathcal{R}\left(\hat{\theta}\right)\right] - \mathcal{R}^* = \sigma^2\frac{d}{n}.$$

Proof. We plug in the previous result to Prop. 3.2. We find

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}\left(\hat{\theta}\right)\right] - \mathcal{R}^* &= \mathbb{E}\left[\left\|\hat{\theta} - \theta_*\right\|_{\hat{\Sigma}}^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta} - \theta_*\right)^T\hat{\Sigma}\left(\hat{\theta} - \theta^*\right)\right] \\
&= \mathbb{E}\left[\mathrm{tr}\left(\left(\hat{\theta} - \theta_*\right)^T\hat{\Sigma}\left(\hat{\theta} - \theta^*\right)\right)\right] \\
&= \mathrm{tr}\left(\mathbb{E}\left[\hat{\Sigma}\left(\hat{\theta} - \theta_*\right)\left(\hat{\theta} - \theta_*\right)^T\right]\right) \\
&= \mathrm{tr}\left(\mathbb{E}\left[\hat{\Sigma}\frac{\sigma^2}{n}\hat{\Sigma}^{-1}\right]\right) \\
&= \frac{\sigma^2}{n}\,\mathrm{tr}(I_d) = \sigma^2\frac{d}{n},
\end{aligned}$$

where we use the linearity of the trace, and the fact that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. ■

Observations:

- In fixed design setting, OLS thus leads to unbiased estimation with excess risk of $\sigma^2\frac{d}{n}$.
- For excess risk being small compared to σ^2, need $\frac{d}{n}$ to be small. This seems to exclude *high-dimensional problems* where d is close to n (let alone $d > n$ or $d \gg n$). Regularization (ridge) can help.

§3.5 Ridge Least-Squares Regression

When $d > n$, $\Phi^T\Phi$ is no longer invertible, and so the normal equations admit a whole subspace of solutions. We have two main solutions to this:

1. *Dimension reduction:* aims to replace the feature vector $\varphi(x)$ with another feature vector of lower dimension

2. *Regularization:* adds a term to the least-squares objective function, (i.e. ℓ^1-penalty, which leads to *lasso*, or ℓ^2-penalty, which leads to *ridge*)

↪**Definition 3.1** (Ridge Least Squares Regression Estimator): For a regularization parameter $\lambda > 0$, we define the *ridge least squares estimators* $\hat{\theta}_\lambda$ as the minimizer of

$$\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \|y - \Phi\theta\|_2^2 + \lambda \|\theta\|_2^2 \right\} \quad \text{(ridge regularization)}.$$

We can express the ridge regression estimator in closed form; we don't even need $\Phi^T\Phi$ to be invertible as in the OLS case.

↪**Proposition 3.5**: With, as usual, $\hat{\Sigma} = \frac{1}{n}\Phi^T\Phi \in \mathbb{R}^{d\times d}$, then

$$\hat{\theta}_\lambda = \frac{1}{n}\left(\hat{\Sigma} + \lambda I\right)^{-1} \Phi^T y.$$

Remark 3.2: In particular, when $\lambda = 0$, we recover the OLS estimator assuming $\hat{\Sigma}$ invertible.

Proof. This is essentially the same as the proof for the OLS; one recognizes that we have a quadratic in θ. The invertibility of $\hat{\Sigma} + \lambda I$ follows from the fact that $\hat{\Sigma}$ positive semidefinite, and thus has nonnegative eigenvalues, and thus $\hat{\Sigma} + \lambda I$ has strictly positive eigenvalues and is thus invertible. ■

3.5.1 Statistical Properties of Ridge Least Squares Estimator

↪**Proposition 3.6**: The ridge least squares estimator, under the fixed-design assumptions, is equal to $\hat{\theta}_\lambda = \frac{1}{n}\left(\hat{\Sigma} + \lambda I\right)^{-1}\Phi^T y$ has excess risk

$$\mathbb{E}\left[\mathcal{R}\left(\hat{\theta}_\lambda\right)\right] - \mathcal{R}^* = \underbrace{\lambda^2 \theta_*^T \left(\hat{\Sigma} + \lambda I\right)^{-2} \hat{\Sigma} \theta_*}_{\text{bias}} + \underbrace{\frac{\sigma^2}{n} \operatorname{tr}\left(\hat{\Sigma}^2 \left(\hat{\Sigma} + \lambda I\right)^{-2}\right)}_{\text{variance}}.$$

Observations:

- The result above gives a bias/variance decomposition of the excess risk; this is related to the approximation/estimation error decomposition of the risk.

The bias term is part of the approximation error - it has the main influences of λ, which only effects θ and is thus really a modelling assumption. The variance term captures the "noise" (σ^2 only appears here), and is really about estimation error.

- *Bias:* as $\lambda \uparrow$, bias $\uparrow$ and is equal to zero if $\lambda = 0$ (and $\hat{\Sigma}$ is invertible, of course). In particular, the bias grows like λ^2, and as $\lambda \to \infty$, the bias approximates $\lambda^2 \cdot \theta_*^T \hat{\Sigma}^{-1} \theta_*$ (which is independent of n).
- *Variance* as $\lambda \uparrow$, the variance $\downarrow$ and when $\lambda = 0$, the variance is $\frac{\sigma^2}{n}$; this depends on n.

In particular, since the excess risk is the sum of these two, we expect a kind of parabolic relationship between excess risk and λ, implying the existence of some optimal λ.

- The quantity $\operatorname{tr}\left(\hat{\Sigma}^2\left(\hat{\Sigma}+\lambda I\right)^{-2}\right)$ is called the *degrees of freedom*, and is considered as an "implicit number of parameters".
- As $\lambda \to 0$, $\hat{\theta}_\lambda$ converges to the OLS estimator
- $\lambda = 0$ is not usually the optimal choice (i.e. yielding the best excess risk); we want to bias our estimator in general

3.5.2 Choice of λ

Can we tune λ to achieve a better bound than our OLS?

↪**Proposition 3.7** (Choice of regularization parameter): Let

$$\lambda_* = \frac{\sigma \cdot \operatorname{tr}\left(\hat{\Sigma}\right)^{1/2}}{\|\theta_*\|_2 \sqrt{n}}.$$

Then,

$$\mathbb{E}\left[\mathcal{R}\left(\hat{\theta}_{\lambda_*}\right)\right] - \mathcal{R}^* = \frac{\sigma \cdot \operatorname{tr}\left(\hat{\Sigma}\right)^{1/2} \|\theta_*\|_2}{\sqrt{n}}.$$

Proof. Recall that the eigenvalues of $\lambda I + \hat{\Sigma}$ are of the form $\lambda + \mu$ for μ an eigenvalue of $\hat{\Sigma}$. In addition, we will need to use the fact that $\operatorname{tr}(AB) \le \lambda_{\max}(A) \operatorname{tr}(B)$ (special case of Holder's inequality).

We need first a bound on the eigenvalues of the matrix $\left(\hat{\Sigma}+\lambda I\right)^{-2}\lambda\hat{\Sigma}$. Let μ be an eigenvalue of $\hat{\Sigma}$, so $\mu+\lambda$ an eigenvalue of $\hat{\Sigma}+\lambda I$. We know $(\mu+\lambda)^2 \ge 2\lambda\mu$ (AM-GM), and hence $\lambda\mu(\mu+\lambda)^{-2} \le \frac{1}{2}$ and so the eigenvalues of $\left(\hat{\Sigma}+\lambda I\right)^{-2}\lambda\hat{\Sigma}$ are always $\le \frac{1}{2}$, i.e.

$$\lambda_{\max}\left\{\left(\hat{\Sigma}+\lambda I\right)^{-2}\hat{\Sigma}\right\} \le \frac{1}{2}.$$

Therefore, we can bound the bias

$$\text{bias} = \lambda\theta_*^T\left(\hat{\Sigma}+\lambda I\right)^{-2}\lambda\hat{\Sigma}\theta_* \le \lambda\lambda_{\max}\left\{\left(\hat{\Sigma}+\lambda I\right)^{-2}\lambda\hat{\Sigma}\right\}\|\theta_*\|_2^2 \le \frac{\lambda}{2}\|\theta_*\|_2^2.$$

Similarly, we can bound the variance,

$$\begin{aligned}
\text{variance} = \frac{\sigma^2}{n}\operatorname{tr}\left(\hat{\Sigma}^2\left(\hat{\Sigma}+\lambda I\right)^{-2}\right) &= \frac{\sigma^2}{n\lambda}\operatorname{tr}\left(\hat{\Sigma}\left[\lambda\hat{\Sigma}\left(\hat{\Sigma}+\lambda I\right)^{-2}\right]\right) \\
&\le \frac{\sigma^2}{n\lambda}\operatorname{tr}\left(\hat{\Sigma}\right)\lambda_{\max}\left\{\lambda\hat{\Sigma}\left(\hat{\Sigma}+\lambda I\right)^{-2}\right\} \\
&\le \frac{\sigma^2}{2n\lambda}\operatorname{tr}\left(\hat{\Sigma}\right).
\end{aligned}$$

Together, these imply that, for any $\lambda > 0$,

$$\mathbb{E}\left[\mathcal{R}\left(\hat{\theta}_\lambda\right)\right] - \mathcal{R}_* = \text{bias} + \text{variance} \le \frac{\lambda}{2}\|\theta_*\|_2^2 + \frac{\sigma^2}{2n\lambda}\operatorname{tr}\left(\hat{\Sigma}\right).$$

We optimize the right-hand side, which is of the form $f(\lambda) = a\lambda + \frac{b}{\lambda}$. One verifies that the minimum is $\lambda = \sqrt{\frac{b}{a}}$, which has value $f\left(\sqrt{\frac{b}{a}}\right) = 2\sqrt{ab}$. Since $a = \frac{\|\theta_*\|_2^2}{2}$ and $b = \frac{\sigma^2}{2n}\operatorname{tr}\left(\hat{\Sigma}\right)$, this implies

$$\lambda_* = \frac{\sigma \operatorname{tr}\left(\hat{\Sigma}\right)^{1/2}}{\sqrt{n}\|\theta_*\|_2},$$

as claimed, and similarly, we get the actual excess risk of $\hat{\theta}_{\lambda_*}$ upon plugging in the appropriate values. ■

Remark 3.3:

1. Let $R = \max_{i\in[n]} \|\varphi(x_i)\|_2$. Then

$$\operatorname{tr}\left(\hat{\Sigma}\right) = \sum_{j=1}^{d} \hat{\Sigma}_{jj} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\left(\varphi(x_i)_j\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\|\varphi(x_i)\|_2^2 \leq R^2.$$

   Namely, the dimension d of the parameter space plays no role in the excess risk, and thus d could be infinite, provided R and $\|\theta_*\|_2^2$ are finite (normally, R grows with d, so need these extra assumptions).

2. We can compare this to the excess risk of the OLS estimator, which we found to be $\sigma^2\frac{d}{n}$. We see that

- With λ_*, our excess risk of ridge goes to zero as $n \to \infty$ like $\frac{1}{\sqrt{n}}$, which is slower than that of OLS (which goes like $\frac{1}{n}$)
- However, the ridge estimator has risk proportional to σ while the OLS is proportional to σ^2, i.e. OLS has a higher dependency on the noise variance

3. λ_* involves constants we don't know, i.e. σ, $\|\theta_*\|_2$ - in practice, we have to find λ_* by trial and error.
4. λ_* is not necessarily the best choice in the sense of minimizing the excess risk, since we found it by optimizing an upper bound of the excess risk.
5. λ is an example of a "hyperparameter" - something a user must choose. It should not be taken as an *absolute* - rather, it should be consider as a "guide" as to how to pick λ.

§3.6 Random Design Analysis

Here, we assume the following:

1. Both x and y are random and each pair (x_i, y_i) from a probability distribution p on $\mathcal{X} \times \mathbb{R}$
2. $\exists\theta_* \in \mathbb{R}^d$ s.t. $y_i = \varphi(x_i)^T\theta_* + \varepsilon_i$
3. ε_i's are independent from x_j's and each other, and are mean zero, variance σ^2.

In particular, remark that under these assumptions, $\mathbb{E}[y_i \mid x_i] = \varphi(x_i)^T\theta_*$.

↪**Proposition 3.8**: Under the above assumptions, for any $\theta \in \mathbb{R}^d$, we have

$$\mathcal{R}(\theta) - \mathcal{R}^* = \|\theta - \theta_*\|_\Sigma^2, \quad \Sigma := \mathbb{E}\left[\varphi(x)\varphi(x)^T\right],$$

and $\mathcal{R}^* = \sigma^2$.

Proof.

$$\begin{aligned}
\mathcal{R}(\theta) &= \mathbb{E}_{x,y,\varepsilon}\left[\left(y - \theta^T\varphi(x)\right)^2\right] \\
&= \mathbb{E}\left[\left(\varphi(x)^T\theta_* + \varepsilon - \theta^T\varphi(x)\right)^2\right] \\
&= \mathbb{E}\left[\left(\varphi(x)^T(\theta_* - \theta) - \varepsilon\right)^2\right] \\
&= \mathbb{E}\left[(\theta - \theta_*)^T\varphi(x)\varphi(x)^T(\theta - \theta_*) + \varepsilon^2 - 2\varepsilon\varphi(x_0)^T(\theta - \theta_*)\right] \\
&= \mathbb{E}_x\mathbb{E}_\varepsilon[[\cdots \mid x]] \\
\text{(independence)} \quad &= \mathbb{E}\left[\varepsilon^2\right] + \mathbb{E}\left[\left(\varphi(x)^T(\theta - \theta_*)\right)^2\right] \\
&= \sigma^2 + (\theta - \theta_*)^T\Sigma(\theta - \theta_*).
\end{aligned}$$

■

↪**Proposition 3.9**: Under the above assumptions, assume $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \varphi(x_i)\varphi(x_i)^T = \frac{1}{n}\Phi^T\Phi$ is invertible. Then, the expected excess risk of the OLS estimator is given by

$$\frac{\sigma^2}{n}\mathbb{E}\left[\operatorname{tr}\left(\Sigma\hat{\Sigma}^{-1}\right)\right].$$

Remark 3.4: We see Σ as the "expected covariance matrix" and $\hat{\Sigma}$ as the "empirical covariance matrix"; in particular, these two were equal in the fixed-design case (as our inputs were deterministic) and thus the above quantity becomes

$$\frac{\sigma^2}{n}\mathbb{E}[\operatorname{tr}(I_d)] = \frac{d}{n}\sigma^2.$$

Proof. Recall that $\hat{\sigma} = \frac{1}{n}\hat{\Sigma}^{-1}\Phi^T y$, which, by assumption on the y's, becomes

$$\hat{\theta} = \frac{1}{n}\hat{\Sigma}^{-1}\Phi^T(\Phi\theta_* + \varepsilon) = \theta_* + \frac{1}{n}\hat{\Sigma}^{-1}\Phi^T\varepsilon.$$

By the previous proposition, then, it follows that

$$\begin{aligned}
\mathbb{E}\Big[\mathcal{R}\big(\hat{\theta}\big) - \mathcal{R}^*\Big] &= \mathbb{E}\left[\left(\frac{1}{n}\hat{\Sigma}^{-1}\Phi^T\varepsilon\right)^T \Sigma\left(\frac{1}{n}\hat{\Sigma}^{-1}\Phi^T\varepsilon\right)\right] \\
&= \frac{1}{n^2}\mathbb{E}\Big[\operatorname{tr}\Big(\Sigma\Big(\hat{\Sigma}^{-1}\Phi^T\varepsilon\varepsilon^T\Phi\hat{\Sigma}^{-1}\Big)\Big)\Big] \quad \text{(i)} \\
&= \frac{\sigma^2}{n^2}\mathbb{E}\left[\operatorname{tr}\left(\Sigma\hat{\Sigma}^{-1}\underbrace{\Phi^T\Phi}_{=nI}\hat{\Sigma}^{-1}\right)\right] \quad \text{(ii)} \\
&= \frac{\sigma^2}{n}\mathbb{E}\Big[\operatorname{tr}\Big(\Sigma\hat{\Sigma}^{-1}\Big)\Big].
\end{aligned}$$

In (i) we use the fact that for any real matrices A, B, $\operatorname{tr}(AB) = \operatorname{tr}(BA)$; in particular, here this case, $AB \in \mathbb{R}$ so that $AB = \operatorname{tr}(AB)$ (where A, B are the appropriate matrices above). In (ii) we use the linearity of the trace, as well as the fact that, by conditioning on x first and using independence of ε, x, we can factor out $\mathbb{E}\big[\varepsilon\varepsilon^T\big] = n\sigma^2$. ■

3.6.1 Gaussian Design

Here, we briefly study what more we can say in the case that $\varphi(x) \sim \mathcal{N}(0, \Sigma)$ for some $\Sigma \in \mathbb{R}^{d\times d}$. We can write

$$\varphi(x) = \Sigma^{1/2}\hat{Z}, \quad \hat{Z} \sim \mathcal{N}(0, I_d).$$

Generating n (independent) $\hat{Z} \in \mathbb{R}^{n\times d}$, we then form the random matrix

$$Z \coloneqq \begin{pmatrix} -\hat{Z}_1^T- \\ \vdots \\ -\hat{Z}_n^T- \end{pmatrix} \in \mathbb{R}^{n\times d}.$$

This gives

$$Z\Sigma^{1/2} = \Phi \in \mathbb{R}^{n\times d} \Rightarrow \hat{\Sigma} = \frac{1}{n}\Phi^T\Phi = \frac{1}{n}\Sigma^{1/2}Z^TZ\Sigma^{1/2}.$$

Thus, apply the "trace trick" from the previous proposition again, we find that

$$\mathbb{E}\Big[\operatorname{tr}\Big(\Sigma\hat{\Sigma}^{-1}\Big)\Big] = n\mathbb{E}\left[\operatorname{tr}\left(\Sigma\left(\Sigma^{-\frac{1}{2}}\big(Z^TZ\big)^{-1}\Sigma^{-1/2}\right)\right)\right] = n\mathbb{E}\left[\operatorname{tr}\left(\big(Z^TZ\big)^{-1}\right)\right] = \frac{nd}{n-d-1},$$

which implies the excess risk $\approx \frac{d}{n}\sigma^2$ (this equality above uses the fact that much is known about the spectral properties of the $\big(Z^TZ\big)^{-1}$, which we won't discuss here).