Notes on Learning Theory

Dan Edelstein

September 2025

About this text

Each chapter or subchapter refers to one from Francis Bach's "Learning Theory from First Principles" [1], and proceeds as follows:
Preliminary Definitions: Necessary definitions and theorems for the content in the following chapter, presented without derivation as definitional.
Derivations and Further Definitions: Key results from the given chapter, presented as an analytic narrative, proceeding from premises or preliminary definitions to a relevant conclusion.
Proofs: Occasional standalone proofs that support or otherwise engage with the content of the chapter, topics chosen in large part by guidance from the author's supervisor[1]
Discussion: Brief summary of the chapter's content, builds a textual narrative through the text on the significance of analytic results.
Proofs and derivations are a combination of material from Bach's text and secondary sources, and while the majority of strategies for derivations and proofs are not created de novo by the author, the author has worked through each step-by-step and endeavored to carefully present all analytic arguments as complete and well-reasoned without excessive jumps in logic or unstated assumptions. The author has attempted to demonstrate a consistent body of knowledge on several key results in Learning Theory.

Ch. 1 - Mathematical Preliminaries

Preliminary Definitions

Convexity: A differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ is convex iff:

$$F(\eta) \geq F(\theta) + F'(\theta)^\top(\eta - \theta), \quad \forall \eta, \theta \in \mathbb{R}^d \tag{1}$$

Which is equivalent to a function being at or above its tangent at all points. We can also use a discrete definition with $\varphi : I \to \mathbb{R}$, where I is an interval in $\mathbb{R}$, $\forall x \in I$ and any $\lambda \in [0, 1]$:

$$\varphi\left(\lambda x_1 + (1-\lambda)x_2\right) \leq \lambda\varphi\left(x_1\right) + (1-\lambda)\varphi\left(x_2\right) \tag{2}$$

Almost Surely: Let $(\Omega, \mathcal{F}, P)$ be a probability space. An event $E \in \mathcal{F}$ happens 'almost surely' if $P(E) = 1$.

σ-Algebra: Given a sample space Ω of all possible outcomes, a σ-Algebra $\mathcal{F}$ is the collection of subsets of Ω that satisfies:

$$\begin{gathered} \Omega \in \mathcal{F} \\ A \in \mathcal{F} \Rightarrow \Omega \setminus A \in \mathcal{F} \\ A_1, A_2, \cdots \in \mathcal{F} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{F} \end{gathered} \tag{3}$$

Filtration: $\mathbb{F} := (\mathcal{F}_i)_{i \in I}$ is a sequence of σ-algebras $(\mathcal{F}_i)_{i \geq 0}$ that is non-decreasing, s.t.:

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n \subseteq \mathcal{F} \tag{4}$$

Martingale: a model of a fair game where knowledge of past events never helps predict future changes.

[1] Prof. Michael Riis Andersen

A discrete-time stochastic process $X = (X_n)_{n \geq 0}$ is a martingale w/r/t a filtration $(\mathcal{F}_n)_{n \geq 0}$ if:

$$\begin{aligned} &\mathbb{E}[|X_n|] < \infty \quad \forall n \\ &\quad X_n \text{ is } \mathcal{F}_n\text{-measurable (value of } X \text{ at time } n \text{ is known at time } n) \\ &\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_n \text{ almost surely} \end{aligned} \tag{5}$$

Doob Martingale: For an integrable function $f(Z_1, \ldots, Z_n)$, a Doob martingale sequence $(Y_i)_{i=0}^n$ is constructed by taking successive conditional expectations w/r/t a filtration $\mathcal{F}_i = \sigma(Z_1, \ldots, Z_i)$:

$$Y_i = E[f(Z_1, \ldots, Z_n)|\mathcal{F}_i] \tag{6}$$

This gives us the boundary conditions: $Y_0 = E[f(Z)]$ (unconditional expectation). $Y_n = f(Z)$ (the function itself).
The martingale difference sequence V_i consists of increments of this process:

$$V_i = Y_i - Y_{i-1} = E[f(Z)|\mathcal{F}_i] - E[f(Z)|\mathcal{F}_{i-1}] \tag{7}$$

Markov's Inequality: Let X be a random variable that is almost surely non-negative. Then, for every constant $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a} \tag{8}$$

Union bound: Given events indexed by $f \in \mathcal{F}$, we have

$$\mathbb{P}\left(\bigcup_{f \in \mathcal{F}} A_f\right) \leq \sum_{f \in \mathcal{F}} \mathbb{P}(A_f) \tag{9}$$

which we can use in upper-bounding the tail probability of the supremum of random variables:

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} Z_f > t\right) = \mathbb{P}\left(\bigcup_{f \in \mathcal{F}} \{Z_f > t\}\right) \leq \sum_{f \in \mathcal{F}} \mathbb{P}(Z_f > t). \tag{10}$$

Chernoff Bound is also referred to as the exponential Markov bound. For any random variable X and parameter $t > 0$: $\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta})$. Applying Markov's Inequality (where our non-negative variable is e^{tX} and our constant is e^{ta}): $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}}$. Since this holds for all $t > 0$, we get

$$\mathbb{P}(X \geq a) \leq \inf_{t > 0} \frac{\mathbb{E}[e^{tX}]}{e^{ta}} \tag{11}$$

Some concentration inequalities:
Jensen's Inequality on $\mathbb{R}$: (see Eq. 12)
If $F : \mathbb{R} \to \mathbb{R}$ is convex and X is a real-valued random variable, then

$$F(\mathbb{E}[X]) \leq \mathbb{E}[F(X)] \tag{12}$$

Hoeffding's Inequality:
If $Z_1, \ldots, Z_n$ are independent random variables s.t. $Z_i \in [0, 1]$ almost surely, then, for any $t \geq 0$:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i] \geq t\right) \leq \exp(-2nt^2) \tag{13}$$

and the corollary:
Two-sided Hoeffding's Inequality:
With the same premise as Eq. 13:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i]\right| \geq t\right) \leq 2\exp(-2nt^2) \tag{14}$$

Which can be extended to the assumption that $Z_i \in [a, b]$ almost surely, leading to

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Z_i]\right| > t\right) \leq 2\exp\left(-\frac{2nt^2}{(a-b)^2}\right). \tag{15}$$

and which is often used "in reverse" starting from the probability and deriving t from it as follows: For any $\delta \in (0, 1)$, with probability greater than $1 - \delta$, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Z_i]\right| < \frac{|a-b|}{\sqrt{2n}}\sqrt{\log\left(\frac{2}{\delta}\right)}. \tag{16}$$

McDiarmid's Inequality: Let $Z_1, \ldots, Z_n$ be independent random variables (in measurable space $\mathcal{Z}_i$), and $f : \mathcal{Z}^n \to \mathbb{R}$ a function of bounded variation s.t. $\forall i \in \{1, \ldots, n\}$, and all $z_1, \ldots, z_n, z_i' \in \mathcal{Z}$, we have

$$|f(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \leq c.$$

Then

$$\mathbb{P}\left(|f(Z_1, \ldots, Z_n) - \mathbb{E}[f(Z_1, \ldots, Z_n)]| > t\right) \leq 2\exp\left(-\frac{2t^2}{(nc^2)}\right). \tag{17}$$

Derivations and Further Definitions

Let X be a random variable with finite mean $\mu = \mathbb{E}[X]$ and finite, non-zero variance $\sigma^2 = Var(X)$ and consider random variable $Y = (X - \mu)^2$, which is almost surely non-negative. If we apply Eq. 8 on Y with constant $a = k^2$, we can see:

$$\mathbb{P}(Y \geq a) \leq \frac{\mathbb{E}[Y]}{a}$$
$$\mathbb{P}((X-\mu)^2 \geq k^2) \leq \frac{\mathbb{E}[(X-\mu)^2]}{k^2}$$
$$\sigma^2 = \mathbb{E}[(X-\mu)^2]$$

We note that $(X - \mu)^2 \geq k^2 \iff |X - \mu| \geq k$. This gives us **Chebyshev's Inequality:**

$$\mathbb{P}(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2} \tag{18}$$

Take as before n i.i.d. variables $Z_1, \ldots, Z_n$ with sample mean $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Linearity of expectation and independence gives us $\mathbb{E}[\bar{Z}_n] = \mathbb{E}[Z]$ and $Var(\bar{Z}_n) = \frac{\sigma^2}{n}$. Substituting these values into Chebyshev's Inequality with threshold $k = \varepsilon$ yields:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z]\right| \geq \varepsilon\right) \leq \frac{Var(\bar{Z}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$

As $n \to \infty$, the upper bound approaches 0, implying that, for a large enough sample size, the probability of sample mean deviation from the true mean by any ε becomes negligible. In the context of machine learning, if we want to be $1 - \delta$ confident that our sample mean is within ε of the true mean, we set the bound $\frac{\sigma^2}{n\varepsilon^2} \leq \delta$ and solve for n:

$$n \geq \frac{\sigma^2}{\delta\varepsilon^2}$$

Proofs

For Jensen's Inequality, we begin with a discrete case, for $x \in I$, $\lambda \in [0, 1]$ with $\sum_{i=1}^{n} \lambda_i = 1$ with I as an open interval in $\mathbb{R}, \varphi : I \to \mathbb{R}$

We will assume that the relation $\sum_{i=1}^{n} \lambda_i \varphi(x_i) \geq \varphi(\sum_{i=1}^{n} \lambda_i x_i)$ holds

$$\text{for } n = 2, \text{ we have}$$

$$\lambda_1 \varphi(x_1) + \lambda_2 \varphi(x_2) \geq \varphi(\lambda_1 x_1 + \lambda_2 x_2)$$

$$\text{and if we take } \lambda_2 = 1 - \lambda_1$$

$$\lambda_1 \varphi(x_1) + (1 - \lambda_1)\varphi(x_2) \geq \varphi(\lambda_1 x_1 + (1 - \lambda_1) x_2)$$

which we note to be equiv. to Eq. 2

$$\text{for any } x_{n+1} \in I, \ \lambda_{n+1} \in [0, 1]$$

$$q_i = \frac{\lambda_i}{1 - \lambda_{n+1}}, \quad \sum_{i=1}^{n+1} q_i = 1$$

$$\sum_{i=1}^{n+1} q_i \varphi(x_i) = \sum_{i=1}^{n} q_i \varphi(x_i) + q_{n+1} \varphi(x_{n+1})$$

$$= \frac{1}{1 + \lambda_{n+1}} \sum_{i=1}^{n} \lambda_i \varphi(x_i) + \frac{\lambda_{n+1}}{1 + \lambda_{n+1}} \varphi(x_{n+1})$$

and as per or assumption, we have :

$$\geq \frac{1}{1 + \lambda_{n+1}} \varphi\left(\sum_{i=1}^{n} \lambda_i x_i \right) + \frac{\lambda_{n+1}}{1 + \lambda_{n+1}} \varphi(x_{n+1})$$

and we can look to our $n = 2$ case to arrive at :

$$\geq \varphi\left(\frac{1}{1 + \lambda_{n+1}} \sum_{i=1}^{n} \lambda_i x_i + \frac{\lambda_{n+1}}{1 + \lambda_{n+1}} x_{n+1} \right)$$

which we note to be equivalent to :

$$\geq \varphi\left(\sum_{i=1}^{n+1} q_i x_i \right) \square$$

The generalization to the continuous case is not included.

Figure 1: Our predecessor in Danish study of convex functions, Johan Jensen

We now move to the Hoeffding's inequality. We start with

$$\begin{aligned}
\text{Let } \varphi(s) &= \log(\mathbb{E}[\exp(s(Z-\mathbb{E}[Z]))]) \\
\varphi'(s) &= \frac{1}{\mathbb{E}[\exp(s(Z-\mathbb{E}[Z]))]} \cdot \frac{d}{ds}\mathbb{E}[\exp(s(Z-\mathbb{E}[Z]))] \\
&= \frac{\mathbb{E}[(Z-\mathbb{E}[Z]) \cdot \exp(s(Z-\mathbb{E}[Z]))]}{\mathbb{E}[\exp(s(Z-\mathbb{E}[Z]))]} \\
\varphi''(s) &= \frac{\mathbb{E}[(Z-\mathbb{E}[Z])^2 \cdot \exp(s(Z-\mathbb{E}[Z]))]}{\mathbb{E}[\exp(s(Z-\mathbb{E}[Z]))]} - \left[\frac{\mathbb{E}[(Z-\mathbb{E}[Z]) \cdot \exp(s(Z-\mathbb{E}[Z]))]}{\mathbb{E}[\exp(s(Z-\mathbb{E}[Z]))]}\right]^2
\end{aligned}$$

These expressions can be parsed effectively with the technique of **exponential tilting**. For a given value s, we define a "tilted" expectation $\mathbb{E}_s$ as relates to the original density:

$$f_s(z) = \frac{\exp(sz)f(z)}{\mathbb{E}[\exp(sZ)]} \tag{19}$$

With this formulation, typically described as a "tilted measure", we find the second derivative of the log-moment generating function:

$$\begin{aligned}
\varphi''(s) &= \mathbb{E}_s[(Z-\mathbb{E}[Z])^2] - (\mathbb{E}_s[Z-\mathbb{E}[Z]])^2 \\
&= \mathrm{Var}_s(Z-\mathbb{E}[Z]) = \mathrm{Var}_s(Z)
\end{aligned}$$

As $Z \in [0,1]$, the variance of Z under any probability measure is explicitly bounded. For any random variable $\tilde{Z} \in [0,1]$:

$$\begin{aligned}
\mathrm{Var}(\tilde{Z}) &= \inf_{\nu \in [0,1]} \mathbb{E}[(\tilde{Z}-\nu)^2] \leq \mathbb{E}[(\tilde{Z}-1/2)^2] \\
&= \frac{1}{4}\mathbb{E}[(2\tilde{Z}-1)^2] \leq \frac{1}{4}
\end{aligned}$$

The last inequality holds as $2\tilde{Z}-1 \in [-1,1]$ almost surely, which gives us $(2\tilde{Z}-1)^2 \leq 1$. Therefore, $\forall s \geq 0$:

$$\varphi''(s) = \mathrm{Var}_s(Z) \leq \frac{1}{4}$$

By Taylor's theorem with integral remainder, for $s \geq 0$:

$$\varphi(s) \leq \varphi(0) + s\varphi'(0) + \frac{1}{2}s^2 \sup_{u \in [0,s]} \varphi''(u) \leq \frac{s^2}{8}$$

with $\varphi(0) = 0$ and $\varphi'(0) = 0$, we obtain the bound:

$$\varphi(s) \leq \frac{s^2}{8} \tag{20}$$

Hoeffding's Inequality for the Sum
Let $Z_1, \ldots, Z_n$ be independent random variables, assume $Z_i \in [0,1]$ almost surely for all $i = 1, \ldots, n$, and use $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$. For any $t \geq 0$, we bound the tail probability:

$$\begin{aligned}
\mathbb{P}(\bar{Z} - \mathbb{E}[\bar{Z}] \geq t) &= \mathbb{P}\left(\exp(s(\bar{Z}-\mathbb{E}[\bar{Z}])) \geq \exp(st)\right) \quad \text{(for any } s > 0) \\
&\leq \frac{\mathbb{E}[\exp(s(\bar{Z}-\mathbb{E}[\bar{Z}]))]}{\exp(st)} \quad \text{(by Eq. 11)}
\end{aligned}$$

Now we compute the numerator using independence. As per Eq. 20, for any variable $X \in [a,b]$ with

mean zero, $\mathbb{E}[\exp(\lambda X)] \leq \exp(\frac{\lambda^2(b-a)^2}{8})$. Applying this with $\lambda = s/n$ and interval size 1:

$$\begin{aligned}
\mathbb{E}[\exp(s(\bar{Z} - \mathbb{E}[\bar{Z}]))] &= \mathbb{E}\left[\exp\left(s \cdot \frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])\right)\right] \\
&= \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\frac{s}{n}(Z_i - \mathbb{E}[Z_i])\right)\right] \quad \text{(by independence)} \\
&\leq \prod_{i=1}^{n} \exp\left(\frac{(s/n)^2 \cdot 1^2}{8}\right) \quad \text{(by 13)} \\
&= \exp\left(\sum_{i=1}^{n} \frac{s^2}{8n^2}\right) = \exp\left(\frac{ns^2}{8n^2}\right) = \exp\left(\frac{s^2}{8n}\right)
\end{aligned}$$

Therefore:

$$\mathbb{P}(\bar{Z} - \mathbb{E}[\bar{Z}] \geq t) \leq \exp\left(-st + \frac{s^2}{8n}\right)$$

If we optimize over s, the right-hand side is minimized when:

$$\frac{d}{ds}\left(-st + \frac{s^2}{8n}\right) = -t + \frac{s}{4n} = 0 \implies s^* = 4nt$$

Substituting back:

$$\begin{aligned}
\exp\left(-s^t + \frac{(s^*)^2}{8n}\right) &= \exp\left(-4nt^2 + \frac{16n^2t^2}{8n}\right) \\
&= \exp(-2nt^2)
\end{aligned}$$

We conclude:

$$\mathbb{P}\left(\bar{Z} - \mathbb{E}[\bar{Z}] \geq t\right) \leq \exp(-2nt^2) \quad \forall t \geq 0 \quad \square$$

By symmetry on $-Z_i$, we also have:

$$\mathbb{P}\left(\bar{Z} - \mathbb{E}[\bar{Z}] \leq -t\right) \leq \exp(-2nt^2)$$

Combining both tails:

$$\mathbb{P}\left(|\bar{Z} - \mathbb{E}[\bar{Z}]| \geq t\right) \leq 2\exp(-2nt^2) \quad \square$$

McDiarmid's Inequality
Let $Z_1, \ldots, Z_n$ be independent random variables and $f : \mathcal{Z}^n \to \mathbb{R}$ satisfy the bounded variation property s.t. changing the i-th coordinate changes the function value by at most c.
We construct a Doob martingale (6) $(Y_i)_{i=0}^n$ w/r/t the filtration (4) $\mathcal{F}_i = \sigma(Z_1, \ldots, Z_i)$:

$$Y_i = E[f(Z_1, \ldots, Z_n)|\mathcal{F}_i].$$

This satisfies the boundary conditions $Y_0 = E[f(Z)]$ and $Y_n = f(Z)$.
We define the martingale difference sequence V_i (7) as increments of this process:

$$V_i = Y_i - Y_{i-1} = E[f(Z)|\mathcal{F}_i] - E[f(Z)|\mathcal{F}_{i-1}].$$

We observe $f(Z) - E[f(Z)] = \sum_{i=1}^n V_i$, which allows us to analyze the deviation of the function from its mean as a sum of martingale differences.
Conditioned on $\mathcal{F}_{i-1}$, V_i depends only on the randomness of Z_i. Due to bounded variation of f, the difference between the maximum and minimum possible values of V_i is bounded almost surely by c. Since V_i is a martingale difference, $E[V_i|\mathcal{F}_{i-1}] = 0$ (see Eq. 7).
We apply the logic of Hoeffding's Lemma (20), which bounds the moment generating function of a bounded random variable with zero mean. For any $s > 0$ and a random variable bounded in an interval of size c, the expectation of the exponential is bounded by $\exp(s^2c^2/8)$. Applying this conditionally to the martingale differences we get:

$$\mathbb{E}[\exp(sV_i)|\mathcal{F}_{i-1}] \leq \exp\left(\frac{s^2c^2}{8}\right)$$

.
Using the Chernoff Bound (11) we look at $E[\exp(s(f(Z) - E[f(Z)]))]$:

$$E\left[\exp\left(s\sum_{i=1}^{n} V_i\right)\right].$$

Using the law of total expectation and conditional independence of martingale differences:

$$E\left[\exp\left(s\sum_{i=1}^{n} V_i\right)\right] = E\left[E\left[\exp(sV_n)\prod_{i=1}^{n-1}\exp(sV_i)\middle|\mathcal{F}_{n-1}\right]\right]$$
$$= E\left[\left(\prod_{i=1}^{n-1}\exp(sV_i)\right)E[\exp(sV_n)|\mathcal{F}_{n-1}]\right]$$

Substituting the bound from Hoeffding's Lemma:

$$\leq E\left[\prod_{i=1}^{n-1}\exp(sV_i)\right]\exp\left(\frac{s^2c^2}{8}\right).$$

Iterating this process for $i = n-1$ down to 1, we obtain:

$$E[\exp(s(f(Z) - E[f(Z)]))] \leq \exp\left(\frac{ns^2c^2}{8}\right).$$

Applying the Chernoff Bound, for any $t > 0$:

$$P(f(Z) - E[f(Z)] \geq t) \leq \exp(-st)E[\exp(s(f(Z) - E[f(Z)]))]$$
$$\leq \exp\left(-st + \frac{ns^2c^2}{8}\right).$$

We optimize the bound w/r/t s. Taking the derivative and setting to 0 yields $-t + \frac{nc^2s}{4} = 0$, which implies $s = \frac{4t}{nc^2}$. Substituting this optimal s back into the exponent:

$$-\left(\frac{4t}{nc^2}\right)t + \frac{nc^2}{8}\left(\frac{16t^2}{n^2c^4}\right) = -\frac{4t^2}{nc^2} + \frac{2t^2}{nc^2} = -\frac{2t^2}{nc^2}.$$

Thus, the one-sided tail bound is $\exp\left(\frac{-2t^2}{nc^2}\right)$. By symmetry on $-f$, we obtain the two-sided McDiarmid's Inequality,:

$$P(|f(Z) - E[f(Z)]| \geq t) \leq 2\exp\left(\frac{-2t^2}{nc^2}\right) \quad \square$$

Ch. 2 - Introduction to Supervised Learning

Preliminary Definitions

Law of total expectation: With X and Y as random variables,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = \int_Y \mathbb{E}[X \mid Y = y]\, dp(y). \tag{21}$$

Conditional risk: given any $x' \in X$ (i.e., $y|x = x'$), we can define the conditional risk for any $z \in Y$ (as a deterministic function of z and x'):

$$r(z|x') = \mathbb{E}[\ell(y, z)|x = x'] \tag{22}$$

Expected Risk: Given a prediction function $f : X \to \mathcal{Y}$, a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and a probability distribution p on $X \times \mathcal{Y}$, the expected risk of f is defined as:

$$\mathcal{R}(f) = \mathbb{E}[\ell(y, f(x))] = \int_{X\times\mathcal{Y}} \ell(y, f(x))\, dp(x, y). \tag{23}$$

Empirical Risk: Given a prediction function $f : X \to \mathcal{Y}$, a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and data $(x_i, y_i) \in X \times \mathcal{Y}$, $i = 1, \ldots, n$, the empirical risk of f is defined as:

$$\hat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i)). \tag{24}$$

Derivations and Further Definitions

We can derive from the above:
$\mathcal{R}(f) = \mathbb{E}[\mathbb{E}[\ell(y, f(x)) \mid x]] = \int_X \mathbb{E}[\ell(y, f(x')) \mid x = x']\, dp(x')$ via Eq.'s 21, 23.
For a fixed $x' \in X$, we write: $\mathbb{E}_{x' \sim p}\,[\mathbb{E}[\ell(y, f(x')) \mid x = x']] = \int_{\mathcal{X}} \mathbb{E}[\ell(y, f(x')) \mid x = x']\, dp(x')$
As $f(x')$ is deterministic given x', from Eq. 22 we have: $\mathcal{R}(f) = \int_{\mathcal{X}} r(f(x')|x')dp(x')$
To minimize $\mathcal{R}(f)$ over all measurable functions $f : X \to \mathcal{Y}$, we can minimize over each $x' \in X$ independently (if we assume X is finite).

We arrive at the **Bayes predictor:**

$$f_*(x') \in \arg\min_{z \in \mathcal{Y}} \mathbb{E}[\ell(y, z) \mid x = x'] = \arg\min_{z \in \mathcal{Y}} r(z \mid x') \tag{25}$$

and subsequently the **Bayes risk**, which is the risk of all Bayes predictors:

$$\mathcal{R}^* = \mathbb{E}_{x' \sim p}\left[\inf_{z \in \mathcal{Y}} \mathbb{E}[\ell(y, z) \mid x = x']\right] \tag{26}$$

We move from this to **Excess Risk** of a function $f : X \to Y$, from the proposition that the expected risk is minimized at a Bayes predictor $f^* : X \to Y, \quad \forall x' \in X$:
We have

$$\begin{aligned}
\mathcal{R}(f) - \mathcal{R}^* &= \mathcal{R}(f) - \mathcal{R}(f_*) \\
&= \int_{\mathcal{X}} r(f(x')|x')\, dp(x') - \int_{\mathcal{X}} r(f_*(x')|x')\, dp(x') \\
&= \int_{\mathcal{X}} r(f(x')|x')\, dp(x') - \int_{\mathcal{X}} \min_{z \in \mathcal{Y}} r(z|x')\, dp(x') \text{ from Eq.'s 25 and 26} \\
&= \int_{\mathcal{X}} \left[r(f(x')|x') - \min_{z \in \mathcal{Y}} r(z|x')\right] dp(x')
\end{aligned}$$

$$\text{Excess Risk } = \mathcal{R}(f) - \mathcal{R}^* \tag{27}$$

Any particular value of $r(f(x')|x')$ must be at least the value of the minimum over all possible choices, and will be equal to the minimum iff the predictor is a Bayes predictor.

Next we describe **empirical risk minimization** with parametrized functions:
Consider a parameterized family of prediction functions $f_\theta : X \to \mathcal{Y}$ for $\theta \in \Theta$. This class of learning methods aims at minimizing the empirical risk with respect to a $\theta \in \Theta$:

$$\hat{\mathcal{R}}(f_\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)). \tag{28}$$

We define an estimator $\hat{\theta} \in \arg\min_{\theta \in \Theta} \hat{\mathcal{R}}(f_\theta)$ (using Eq. 28) and a prediction function $f_{\hat{\theta}} : X \to \mathcal{Y}$
We can decompose parameterized excess risk as follows:
given any $\theta \in \Theta$, we can write the excess risk of $f_{\hat{\theta}}$ (using Eq.'s 23 and 26) as

$$\mathcal{R}(f_{\hat{\theta}}) - \mathcal{R}^* = \underbrace{\left\{\mathcal{R}(f_{\hat{\theta}}) - \inf_{\theta \in \Theta} \mathcal{R}(f_\theta)\right\}}_{\text{estimation error}} + \underbrace{\left\{\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \mathcal{R}^*\right\}}_{\text{approximation error}}. \tag{29}$$

We make use of the infimum $\inf_{\theta \in \Theta} \mathcal{R}(f_\theta)$ to indicate is the smallest possible expected risk achievable by any function in the parametric family.
The approximation error is always nonnegative, does not depend on the chosen $f_{\hat{\theta}}$, and depends only on the class of functions parameterized by $\theta \in \Theta$. The estimation error is also always nonnegative and is typically random because the function $f_{\hat{\theta}}$ is random. It typically decreases in n and increases when Θ grows.

We continue our abstraction by considering the class of machine learning algorithms $\mathcal{A}$ as functions that go from a dataset, i.e., an element of $\mathcal{D}_n(p) = (\mathcal{X} \times \mathcal{Y})^n \quad \forall n$, to a function from $\mathcal{X} \to \mathcal{Y}$. We use $\mathcal{D}_n(p)$ to

indicate the assumption that the data points in $\mathcal{D}$ are obtained as i.i.d. observations from some unknown distribution p from some family $\mathcal{P}$.
We consider an expression of excess risk dependent on p: $\mathcal{R}_p(\mathcal{A}(\mathcal{D}_n(p))) - \mathcal{R}_p^*$
Given that $\mathcal{D}_n(p)$ is random, and that we may not know p, the risk is random. We must nonetheless endeavor to minimize the excess risk via $\mathcal{A}$ as the ultimate goal of Machine Learning.
We measure **Expected Error** as:

$$\mathbb{E}\left[\mathcal{R}_p(\mathcal{A}(\mathcal{D}_n(p)))\right] \tag{30}$$

and describe an Algorithm $\mathcal{A}$ as **Consistent in Expectation** ("Consistency") for p if the Excess Risk measured using the Expected Error tends to zero as n goes to infinity.
We can relax our conditions to make use of **Probably approximately correct (PAC)** learning, by choosing some $\delta \in (0, 1)$ and $\varepsilon > 0$ s.t.

$$\mathbb{P}(\mathcal{R}_p(\mathcal{A}(\mathcal{D}_n(p))) - \mathcal{R}_p^* \leq \varepsilon) \geq 1 - \delta \tag{31}$$

While it may not be possible to achieve optimal risk performance, we hope to find bounds on Excess Risk for a given set of information.
We can define **Asymptotic analysis** as results on sample size $n \to \infty$, such as that for Consistency in Expectation, which requires excess risk to tend to zero as n grows. While these results establish convergence, they do not specify when the behavior begins to apply nor do they guarantee performance for any specific finite dataset size.
In contrast, **non-asymptotic analysis** provides explicit upper bounds on error that are valid for any fixed n, such as that from Hoeffding's (Eq. 13) inequality, guaranteeing that deviation between empirical and expected risk exceeds an ϵ with probability at most δ.

Discussion

Here we move from general probability spaces to the specific goal of risk minimization. The decomposition of Excess Risk into Approximation Error and Estimation Error (Eq. 29) is a central focus for machine learning. It reveals the fundamental trade-off wherein a more complex function class Θ reduces approximation error (getting closer to the Bayes Predictor f_*) but increases estimation error (harder to find the best function). Specifically, we established that the Estimation Error is a random variable dependent on the training data, while Approximation Error is a fixed geometric distance determined by our choice of class Θ. By defining the Bayes Risk $\mathcal{R}^*$ (Eq. 26) as the irreducible error floor, we set the target for consistency: our algorithm $\mathcal{A}$ must drive the excess risk to zero as $n \to \infty$.

Ch. 3 - Linear Least-Squares Regression

I will omit some details/derivations of OLS as it is familiar.

Preliminary Definitions

From the text [1] we have:
Let $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ be the vector of outputs (sometimes called the *response vector*), and $\Phi \in \mathbb{R}^{n \times d}$ the matrix of inputs, whose rows are $\varphi(x_i)^\top$. It is called the *design* matrix or *data* matrix. In this notation, the empirical **ordinary least squares risk** is

$$\hat{\mathcal{R}}(\theta) = \frac{1}{n}\|y - \Phi\theta\|_2^2 \tag{32}$$

We get

$$\hat{\mathcal{R}}(\theta) = \frac{1}{n}\left(\|y\|_2^2 - 2\theta^\top\Phi^\top y + \theta^\top\Phi^\top\Phi\theta\right) \quad \text{and} \quad \hat{\mathcal{R}}'(\theta) = \frac{2}{n}\left(\Phi^\top\Phi\theta - \Phi^\top y\right).$$

$\hat{\mathcal{R}}'(\theta) = 0$ gives the so-called **normal equation**:

$$\Phi^\top\Phi\hat{\theta} = \Phi^\top y. \tag{33}$$

Assuming Φ has full column rank, the multidimensional linear normal equations have a unique solution: $\hat{\theta} = (\Phi^\top\Phi)^{-1}\Phi^\top y$

If we assume that the input data $(x_1, \ldots, x_n)$ are not random (but the output data $(y_1, \ldots, y_n)$ are themselves random), we can say that we are in the fixed design setting, and our goal is to minimize the **Fixed Design Risk:**

$$\mathcal{R}(\theta) = \mathbb{E}_y \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \varphi(x_i)^\top \theta)^2 \right] = \mathbb{E}_y \left[\frac{1}{n} \|y - \Phi\theta\|_2^2 \right] \tag{34}$$

With the homoscedastic noise assumption
$\big(\forall \varepsilon_i, i \in \{1, \ldots, n\}$, are independent, with expectation $\mathbb{E}[\varepsilon_i] = 0$ and variance $\mathbb{E}[\varepsilon_i^2] = \sigma^2\big)$
we assume there is a vector $\theta_* \in \mathbb{R}^d$ s.t. the relationship between input and output is, for $i \in \{1, \ldots, n\}$ $y_i = \varphi(x_i)^\top \theta_* + \varepsilon_i$
We take $\hat{\Sigma} = \frac{1}{n}\Phi^\top \Phi$ as the sample covariance matrix. With a linear model assumption in this setting we have $\mathcal{R}^* = \sigma^2$ (see Eq. 26) and

$$\mathcal{R}(\theta) - \mathcal{R}^* = \frac{1}{n}\|\Phi(\theta - \theta_*)\|_2^2 = (\theta - \theta_*)^\top \hat{\Sigma}(\theta - \theta_*) = \|\theta - \theta_*\|_{\hat{\Sigma}}^2$$

We can further **decompose the OLS expected risk** in this context as per Eq. 29 into:

$$\mathbb{E}[\mathcal{R}(\hat{\theta})] - \mathcal{R}^* = \underbrace{\|\mathbb{E}[\hat{\theta}] - \theta_*\|_{\hat{\Sigma}}^2}_{\text{Bias}} + \underbrace{\mathbb{E}\left[\|\hat{\theta} - \mathbb{E}[\hat{\theta}]\|_{\hat{\Sigma}}^2\right]}_{\text{Variance}} = \frac{\sigma^2 d}{n} \tag{35}$$

We know that $\mathbb{E}[\hat{\theta}] = \theta_*$, $\text{var}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta_*)(\hat{\theta} - \theta_*)^\top\right] = \frac{\sigma^2}{n}\hat{\Sigma}^{-1}$
We return to the regular random design setting with x and y considered random, and each pair (x_i, y_i) assumed i.i.d. from a p on $\mathcal{X} \times \mathbb{R}$. We maintain assumptions of linearity and homoscedasticity from our fixed setting.
We then arrive at this expression for **expected risk:**

$$\mathcal{R}(\theta) - \mathcal{R}^* = \|\theta - \theta_*\|_{\Sigma}^2 = \frac{\sigma^2}{n}\mathbb{E}\left[\text{tr}\left(\Sigma\hat{\Sigma}^{-1}\right)\right] \tag{36}$$

Discussion

In contrast to the abstract function classes from Ch. 2, OLS offers closed-form solutions and exact risk decompositions. The analysis of Fixed Design versus Random Design settings highlights a significant distinction: whether we treat our inputs as constants or stochastic variables changes our risk guarantees. By explicitly calculating the Bias-Variance decomposition (Eq. 35) for Fixed Design OLS, we replace the general approximation-estimation trade-off with concrete algebraic terms. We derived the exact excess risk $\mathbb{E}[\mathcal{R}(\hat{\theta})] - \mathcal{R}^* = \frac{\sigma^2 d}{n}$. This result demonstrates that for unregularized linear models, the error scales linearly with the dimension d and inversely with sample size n, providing a clear benchmark for model complexity.

Ch. 4 - Empirical Risk Minimization

Preliminary Definitions

Lipschitz continuity in real-valued functions: A real-valued function $f : \mathbb{R} \to \mathbb{R}$ is called Lipschitz continuous if $\exists K \in \mathbb{R}^+$ s.t. $\forall (x_1, x_2) \in \mathbb{R}$:

$$|f(x_1) - f(x_2)| \leq K|x_1 - x_2| \tag{37}$$

Hölder's inequality: Let S, Σ, μ be a measure space and let $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then for all measurable real- or complex-valued functions f and g on S:

$$\|fg\|_1 \leq \|f\|_p \|g\|_q \tag{38}$$

We take $\mathcal{Y} = \{-1, 1\}$, $g : \mathcal{X} \to \mathbb{R}$, $f(x) = \text{sign}(g(x))$ with a random sampling from the uniform distribution over $\{-1, 1\}$ when $g(x) = 0$. We use $\Phi : \mathbb{R} \to \mathbb{R}$ to indicate a loss function mapping the margin $u = yg(x)$ (distance from prediction to decision boundary) to a loss value.

We then get the **Φ-Risk:**

$$\mathcal{R}_\Phi(g) = \mathbb{E}[\Phi(yg(x))] \tag{39}$$

There are several **Convex Surrogates** for Φ that avoid our non-continuous (and as such non-convex) 0–1 risk for $f = \text{sign} \circ g$ (where we randomly sample at $g(x) = 0$).
We take the following:

- **Quadratic Square Loss:**

$$\Phi(u) = (u-1)^2 \tag{40}$$

$$\begin{aligned} y^2 &= 1 \\ \Phi(u) &= (yg(x)-1)^2 \\ &= (y(g(x)-y))^2 \\ &= (g(x)-y)^2 \end{aligned}$$

- **Logistic loss:**

$$\Phi(u) = \log(1+e^{-u}) \tag{41}$$

$$\begin{aligned} \Phi(u) &= \log(1+e^{-yg(x)}) \\ &= -\log\left(\frac{1}{e^{-yg(x)}}\right) \\ &= -\log\left(\sigma(yg(x))\right) \text{ for } \sigma \text{ as the sigmoid fn} \end{aligned}$$

- **Hinge loss:**

$$\Phi(u) = \max(1-u, 0) \tag{42}$$

- **Squared Hinge loss:**

$$\Phi(u) = \max(1-u, 0)^2 \tag{43}$$

- **Exponential loss:**

$$\Phi(u) = \exp(-u) \tag{44}$$

Derivations and Further Definitions

We take as a proposition (Bartlett et al., 2006): Let $\Phi : \mathbb{R} \to \mathbb{R}$ be a convex function. The surrogate function Φ is classification-calibrated if and only if Φ is differentiable at 0 and $\Phi'(0) < 0$.
⟨ This is an interesting proof, i would like to work through it ⟩
We reiterate (i think i said it already) that calibration is a property of the surrogate loss function Φ which ensures that minimizing the (convex) Φ-risk leads to optimal predictions for the original, non-convex 0–1 loss.
We return to Eq. 29 in the context of our surrogate functions with a family of prediction functions $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$ for $\Theta \subset \mathbb{R}^d$ and investigate the two elements of the risk in this context.
We start with Approximation Error $\inf_{f \in \mathcal{F}} \mathcal{R}(f) - \mathcal{R}^*$, which, when we assume that θ^* does not belong to Θ, **we can further decompose the approximation error** into:

$$\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \mathcal{R}^* = \left\{ \inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \inf_{\theta' \in \mathbb{R}^d} \mathcal{R}(f_{\theta'}) \right\} + \left\{ \inf_{\theta' \in \mathbb{R}^d} \mathcal{R}(f_{\theta'}) - \mathcal{R}^* \right\} \tag{45}$$

with the right-hand element as the incompressible approximation error due to our model choice.
We finally get into some meat of the learning theory and state that the fn $\theta \mapsto \mathcal{R}(f_\theta) - \inf_{\theta' \in \mathbb{R}^d} \mathcal{R}(f_{\theta'})$ is nonnegative on $\mathbb{R}^d$ and can be typically upper-bounded by a specific norm (or its square) $\Omega(\theta - \theta_*)$.
We can see our left-hand term as a notion of "distance" between θ_* and Θ.
If we assume our loss function $\ell(y, \cdot)$ is Lipschitz-continuous w.r.t. real-valued $f_\theta(x), f_{\theta'}(x)$ such as is the case in our convex conjugates:

$$|\ell(y, f_\theta(x)) - \ell(y, f_{\theta'}(x))| \leq G|f_\theta(x) - f_{\theta'}(x)|$$

we arrive at a **bounded excess risk:**

$$\mathcal{R}(f_\theta) - \mathcal{R}(f_{\theta'}) = \mathbb{E}\left[\ell(y, f_\theta(x)) - \ell(y, f_{\theta'}(x))\right] \leq G\mathbb{E}\left[|f_\theta(x) - f_{\theta'}(x)|\right] \tag{46}$$

We can also look into **further decompositions of the estimation error**, remaining within linear models that are bounded by D in ℓ_2 norm with G-Lipschitz-losses, which we indicate as $\mathcal{F}$.
Using $g_\mathcal{F} \in \arg\min_{g\in\mathcal{F}} \mathcal{R}(g)$ as the minimizer of the expected risk and $\hat{f} \in \arg\min_{f\in\mathcal{F}} \hat{\mathcal{R}}(f)$ as the minimizer of the empirical risk:

$$\begin{aligned}
\mathcal{R}(\hat{f}) - \inf_{f\in\mathcal{F}} \mathcal{R}(f) &= \mathcal{R}(\hat{f}) - \mathcal{R}(g_\mathcal{F}) \\
&= \left\{\mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f})\right\} + \left\{\hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(g_\mathcal{F})\right\} + \left\{\hat{\mathcal{R}}(g_\mathcal{F}) - \mathcal{R}(g_\mathcal{F})\right\}
\end{aligned} \tag{47}$$

We decompose the excess risk into three terms: The deviation of the true risk from the empirical risk at the empirical minimizer $\hat{f}$ - $\mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f})$, The empirical optimization difference - $\hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(g_\mathcal{F})$, and the deviation of the empirical risk from the true risk at the optimal function $g_\mathcal{F}$ - $\hat{\mathcal{R}}(g_\mathcal{F}) - \mathcal{R}(g_\mathcal{F})$.

$$\begin{aligned}
&\leq \sup_{f\in\mathcal{F}} \left\{\mathcal{R}(f) - \hat{\mathcal{R}}(f)\right\} + \left\{\hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(g_\mathcal{F})\right\} + \sup_{f\in\mathcal{F}} \left\{\hat{\mathcal{R}}(f) - \mathcal{R}(f)\right\} \\
&\leq \sup_{f\in\mathcal{F}} \left\{\mathcal{R}(f) - \hat{\mathcal{R}}(f)\right\} + 0 + \sup_{f\in\mathcal{F}} \left\{\hat{\mathcal{R}}(f) - \mathcal{R}(f)\right\}
\end{aligned} \tag{48}$$

We replace two of these terms with the supremum of their maximal difference within $\mathcal{F}$ to establish bounds and note that by definition of $\hat{f}$ the empirical optimization difference is 0. We can simplify this if desired to $\leq 2\sup_{f\in\mathcal{F}} |\mathcal{R}(f) - \hat{\mathcal{R}}(f)|$.
We can re-express the difference between the true risk and the empirical risk as the difference between expectation and empirical average of the loss function

$$\mathcal{R}(f) - \hat{\mathcal{R}}(f) = E[\ell(y, f(x))] - \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$$

We can assume that the training data (x_i, y_i) are i.i.d., which ensures the individual losses $\ell(y_i, f(x_i))$ are independent RV's, and that these loss values are almost surely bounded within $[0, \ell_\infty]$. We can apply Hoeffding's Inequality (Eq.'s 15 - 16) to yield the expression

$$P\left(\mathcal{R}(f) - \hat{\mathcal{R}}(f) \geq t\right) = P\left(\frac{1}{n}\sum E[\ell(y_i, f(x_i))] - \frac{1}{n}\sum \ell(y_i, f(x_i)) \geq t\right) \leq \exp\left(\frac{-2nt^2}{\ell_\infty^2}\right)$$

for some t. We can set this probability upper bound t to $\delta \in (0, 1)$ s.t. with probability at least $1 - \delta$:

$$\begin{aligned}
\delta &= \exp\left(\frac{-2nt^2}{\ell_\infty^2}\right) \\
\log\delta &= \frac{-2nt^2}{\ell_\infty^2} \\
t^2 &= \frac{\ell_\infty^2}{-2n}\log\delta \\
&= \frac{\ell_\infty^2}{2n}\log\frac{1}{\delta} \\
t &= \frac{\ell_\infty}{\sqrt{2n}}\sqrt{\log\frac{1}{\delta}}
\end{aligned}$$

$$\mathcal{R}(f) - \hat{\mathcal{R}}(f) \leq \frac{\ell_\infty}{\sqrt{2n}}\sqrt{\log\frac{1}{\delta}}$$

We can look at **applications of the decomposed estimation error when the class of prediction functions $\mathcal{F}$ is finite**: Our goal is to bound the uniform deviation of the empirical risk from the true risk over $\mathcal{F}$: $\sup_{f\in\mathcal{F}} \left|\hat{\mathcal{R}}(f) - \mathcal{R}(f)\right|$. We retain the assumption that $\ell(y, f(x))$ is bounded almost surely by ℓ_∞.

We can take Hoeffding's inequality with the union bound (Eq.'s 10, 14) which states that the probability that any event in a finite collection occurs is less than or equal to the sum of the probabilities of the individual events. Here, the "event" is that the risk deviation is large for a specific $f \in \mathcal{F}$:

$$P\left(\sup_{f\in\mathcal{F}}\left|\hat{\mathcal{R}}(f)-\mathcal{R}(f)\right|\geq t\right)\leq\sum_{f\in\mathcal{F}}P\left(\left|\hat{\mathcal{R}}(f)-\mathcal{R}(f)\right|\geq t\right)$$

Since there are $|\mathcal{F}|$ terms, and each is bounded by the same Hoeffding bound:

$$P\left(\sup_{f\in\mathcal{F}}\left|\hat{\mathcal{R}}(f)-\mathcal{R}(f)\right|\geq t\right)\leq 2|\mathcal{F}|\exp\left(\frac{-2nt^2}{\ell_\infty^2}\right)$$

We can set this probability upper bound t to $\delta \in (0,1)$ s.t. with probability at least $1-\delta$:

$$\begin{aligned}\delta &= 2|\mathcal{F}|\exp\left(\frac{-2nt^2}{\ell_\infty^2}\right)\\ t &= \ell_\infty\sqrt{\frac{1}{2n}}\sqrt{\log\left(\frac{2|\mathcal{F}|}{\delta}\right)}\\ &= \ell_\infty\sqrt{\frac{\log(2|\mathcal{F}|)+\log(1/\delta)}{2n}}\end{aligned}$$

Which we can separate into two terms:

$$\sup_{f\in\mathcal{F}}\left|\hat{\mathcal{R}}(f)-\mathcal{R}(f)\right|\leq\ell_\infty\sqrt{\frac{\log(2|\mathcal{F}|)}{2n}}+\ell_\infty\sqrt{\frac{\log(1/\delta)}{2n}}$$

where the left-hand term depends on $|\mathcal{F}|$ and the right-hand term depends on δ. As such we can bound this supremum of the risk with one term that quantifies model complexity and one term that quantifies statistical error rate.

Proofs

Exercise 4.5: Show that for $\Theta = \{\theta \in \mathbb{R}^d, \|\theta\|_1 \leq D\}$ (ℓ_1-norm instead of the ℓ_2-norm), we have

$$\inf_{\theta\in\Theta}\mathcal{R}(f_\theta)-\inf_{\theta\in\mathbb{R}^d}\mathcal{R}(f_\theta)\leq G\mathbb{E}\left[\|\varphi(x)\|_\infty\right](\|\theta_*\|_1-D)_+. \tag{49}$$

Generalize to all norms.

We take as unstated assumptions built into the exercise that we are using a generalized linear model $f_\theta(x) = \theta^\top\varphi(x)$ with risk $\mathcal{R}(f_\theta) = \mathbb{E}[\ell(y, f_\theta(x))]$ where ℓ is G-Lipschitz and convex.

We attempt to find $\theta_* = \arg\inf_{\theta\in\mathbb{R}^d}\mathcal{R}(f_\theta)$ within $\mathbb{R}^D$ as well as an optimal θ_Θ that satisfies the above while constrained to Θ within the ℓ_1 ball of size D. Whenever the minimizer θ_* exists, since θ_* is the global minimum, we must have $\mathcal{R}(f_{\theta_\Theta}) \geq \mathcal{R}(f_{\theta_*})$.
We consider two cases:
If $\theta_* \in \Theta$, we have $\|\theta_*\|_1 \leq D$ and $\inf_{\theta\in\Theta}\mathcal{R}(f_\theta) - \inf_{\theta\in\mathbb{R}^d}\mathcal{R}(f_\theta) = 0$.
We note that the RHS uses $(z)_+ = \max(0, z)$ and as $\|\theta_*\|_1 \leq D$, $(\|\theta_*\|_1 - D) \leq 0$ and $(\|\theta_*\|_1 - D)_+ = 0$, so we arrive at a trivial case with $0 \leq 0$.
If $\theta_* \notin \Theta$, $\|\theta_*\|_1 > D$ and $(\|\theta_*\|_1 - D)_+ = \|\theta_*\|_1 - D$. We can rewrite our formula as

$$\inf_{\theta\in\Theta}\mathcal{R}(f_\theta)-\inf_{\theta\in\mathbb{R}^d}\mathcal{R}(f_\theta)\leq G\mathbb{E}\left[\|\varphi(x)\|_\infty\right](\|\theta_*\|_1-D)$$

To arrive at a feasible upper bound, we scale θ_* to the ℓ_1 ball: $\tilde{\theta} = D\frac{\theta_*}{\|\theta_*\|_1}$. We verify that $\tilde{\theta}$ lies on the surface of the ball as $\|\tilde{\theta}\|_1 = \|D\frac{\theta_*}{\|\theta_*\|_1}\|_1 = D\frac{\|\theta_*\|_1}{\|\theta_*\|_1} = D$ through the homogeneity of the ℓ_1 norm.
As θ_Θ is $\arg\inf_{\theta\in\mathbb{R}^d}\mathcal{R}(f_\theta)$ within Θ and $\tilde{\theta} \in \Theta$:

$$\inf_{\theta\in\Theta}\mathcal{R}(f_\theta)=\mathcal{R}(f_{\theta_\Theta})\leq\mathcal{R}(f_{\tilde{\theta}})$$

If we look at $\mathcal{R}(f_\theta) - \mathcal{R}(f_{\theta'}) = \mathbb{E}[\ell(y, \theta^\top \varphi(x)) - \ell(y, \theta'^\top \varphi(x))] \leq G\mathbb{E}\left[|\theta^\top \varphi(x) - \theta'^\top \varphi(x)|\right]$ w.r.t. Eq. 46 and our GLM assumption, we can simplify the RHS to $\leq G\mathbb{E}\left[|(\theta - \theta')^\top \varphi(x)|\right]$. By Hölder's inequality (Eq. 38) we have $|u^\top v| \leq \|u\|_1 \|v\|_\infty$ and therefore $|(\theta - \theta')^\top \varphi(x)| \leq \|\theta - \theta'\|_1 \cdot |\varphi(x)\|_\infty$. We note that $\|\theta - \theta'\|_1$ is deterministic and move it outside of our original term to get $G\mathbb{E}\left[|\theta^\top \varphi(x) - \theta'^\top \varphi(x)|\right] \leq G\mathbb{E}\left[\|\varphi(x)\|_\infty\right] \|\theta - \theta'\|_1$.
Take $\theta = \tilde{\theta}$ and $\theta' = \theta_*$, we have $\mathcal{R}(f_{\tilde{\theta}}) - \mathcal{R}(f_{\theta_*}) \leq G\mathbb{E}\left[\|\varphi(x)\|_\infty\right] \|\tilde{\theta} - \theta_*\|_1$
We simplify $\|\tilde{\theta} - \theta_*\|_1 = \|D\frac{\theta_*}{\|\theta_*\|_1} - \theta_*\|_1 = \|(\frac{D}{\|\theta_*\|_1} - 1)\theta_*\|_1 = \left|\frac{D}{\|\theta_*\|_1} - 1\right| \cdot \|\theta_*\|_1 = \frac{\|\theta_*\|_1 - D}{\|\theta_*\|_1} \cdot \|\theta_*\|_1 = \|\theta_*\|_1 - D$ (with $\|\theta_*\|_1 > D$ and $\frac{D}{\|\theta_*\|_1} < 1$).
So we have

$$\mathcal{R}(f_{\tilde{\theta}}) - \mathcal{R}(f_{\theta_*}) \leq G\mathbb{E}\left[\|\varphi(x)\|_\infty\right] (\|\theta_*\|_1 - D)$$

With $\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) \leq \mathcal{R}(f_{\tilde{\theta}})$ and the $(z)_+$ operator to cover the trivial case, we get

$$\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta) \leq G\mathbb{E}\left[\|\varphi(x)\|_\infty\right] (\|\theta_*\|_1 - D)_+$$

To extend this beyond the ℓ_1 norm/ball: $\Theta = \{\theta : \|\theta\| \leq D\}$ for any $\|\cdot\|$, we observe that our method of scaling, the Hölder inequality, and our distance measure holds for general norms. □

Discussion

Moving beyond linear equality, we encounter classification wherein the ideal 0-1 loss is intractable. This leads us to the introduction of Convex Surrogates (Hinge, Logistic, Exponential). These surrogates upper-bound the 0-1 loss, allowing us to use efficient optimization algorithms while still guaranteeing classification success. However, using these richer classes creates a new problem: we can no longer simply count parameters (d) to guarantee generalization. We addressed this for finite hypothesis classes $|\mathcal{F}|$ using Hoeffding's inequality, deriving a specific generalization bound that scales with $\sqrt{\frac{\log|\mathcal{F}|}{n}}$. This connects the cardinality of the function class directly to the probability of large error.

Ch. 4.5 - Rademacher Complexity

Preliminary Definitions

We consider n i.i.d. random variables $z_1, \ldots, z_n \in \mathcal{Z}$, and a class $\mathcal{H}$ of functions from $\mathcal{Z}$ to $\mathbb{R}$, s.t. $z = (x, y)$, $\mathcal{H} = \{(x, y) \mapsto \ell(y, f(x)), f \in \mathcal{F}\}$, with data $\mathcal{D} = \{z_1, z_2 \ldots\}$ We endeavor to provide an upper bound on $\sup_{f \in \mathcal{F}}\{\mathcal{R}(f) - \widehat{\mathcal{R}}(f)\}$ using a vector of Rademacher random variables ϵ with universal distribution over $\{-1, 1\}$ independent of $\mathcal{D}$ represented as $\varepsilon \in \mathbb{R}^n$.

Rademacher complexity of the class $\mathcal{H}$:

$$R_n(\mathcal{H}) = \mathbb{E}_{\epsilon, D}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h(z_i)\right] \tag{50}$$

Triangle inequality for the supremum: we start from the standard triangle inequality: $|x + y| \leq |x| + |y|$, and then if A and B are functions in $\mathcal{H}$ yielding values in $\mathbb{R}$:

$$\sup_x[f(x) + g(x)] \leq \sup_x[f(x)] + \sup_x[g(x)] \tag{51}$$

Dual norm: Take a vector u, a norm Ω, and any vector θ that is constrained to have a norm less than or equal to 1 under Ω. The dual norm Ω is defined as

$$\Omega^*(u) = \sup_{\Omega(\theta) \leq 1} u^\top \theta \tag{52}$$

When the original norm is the ℓ_p-norm, the dual norm Ω^* is the ℓ_q-norm, where p and q are conjugate exponents satisfying $\frac{1}{p} + \frac{1}{q} = 1$ (with $p, q \in [1, \infty]$).

Derivations and Further Definitions

We can extend the Rademacher Complexity into an empirical context. If we fix the expectation to be only over the Rademacher variables, we have **Empirical Rademacher complexity:**

$$\hat{R}_n(\mathcal{H}) = \mathbb{E}_\varepsilon \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(z_i) \right] \tag{53}$$

which is computable as it is not dependent on an unknown data distribution. We can use McDiarmid's inequality (Eq. 17) on empirical risk, where if $h(z) \in [0, \ell_\infty]$ for all $h \in \mathcal{H}$, then with probability greater than $1 - \delta$, $\forall h \in \mathcal{H}$

$$\mathbb{E}[h(z)] \leq \frac{1}{n} \sum_{i=1}^{n} h(z_i) + 2\hat{R}_n(\mathcal{H}) + 3\frac{\ell_\infty}{\sqrt{2n}} \sqrt{\log \frac{2}{\delta}} \tag{54}$$

The factor of $3 = 1+2$ comes from applying McDiarmid's inequality twice, once for $\sup_{h \in \mathcal{H}} \{\mathcal{R}(h) - \hat{\mathcal{R}}(h)\}$ and once for $\hat{R}_n(\mathcal{H})$ (with an extra factor of 2 since it appears $2R_n(\mathcal{H})$). The derivation of the above is omitted.

We extend our analysis of Rademacher complexity with insights from Lipschitz continuity (Eq. 46), with the proposition that for any functions $b, a_i : \Theta \to \mathbb{R}$, and any 1-Lipschitz functions $\phi_i : \mathbb{R} \to \mathbb{R}$, the **Contraction Principle** holds:

$$\mathbb{E}_\epsilon \left[\sup_{\theta \in \Theta} \left\{ b(\theta) + \sum_{i=1}^{n} \epsilon_i \phi_i(a_i(\theta)) \right\} \right] \leq \mathbb{E}_\epsilon \left[\sup_{\theta \in \Theta} \left\{ b(\theta) + \sum_{i=1}^{n} \epsilon_i a_i(\theta) \right\} \right] \tag{55}$$

This is proved by induction for n from $[0, \infty]$. We omit $n = 0$ as this is trivially valid by equality and move to a comparison of n, $n+1$, $n > 0$. We assume that the proposition holds for n terms and examine $n + 1$. We analyze expectation over $\epsilon_1, \ldots, \epsilon_{n+1}$ by factoring out the expectation w.r.t. ϵ_{n+1}, using the law of total expectation (Eq. 21) $\mathbb{E}_{\epsilon_1,\ldots,\epsilon_{n+1}}[X] = \mathbb{E}_{\epsilon_1,\ldots,\epsilon_n}\left[\mathbb{E}_{\epsilon_{n+1}}\left[X \mid \epsilon_1, \ldots, \epsilon_n\right]\right]$. Let $A(\theta) = b(\theta) + \sum_{i=1}^{n} \epsilon_i \phi_i(a_i(\theta))$

$$LHS_{n+1} = \mathbb{E}_{\epsilon_1,\ldots,\epsilon_n} \left[\mathbb{E}_{\epsilon_{n+1}} \left[\sup_{\theta \in \Theta} \{A(\theta) + \epsilon_{n+1}\phi_{n+1}(a_{n+1}(\theta))\} \right] \right]$$

As ϕ_{n+1} is assumed to be 1-Lipschitz, the expected supremum involving $\phi_{n+1}(a_{n+1}(\theta))$ is bounded by the expected supremum with $a_{n+1}(\theta)$:

$$\mathbb{E}_{\epsilon_{n+1}} \left[\sup_{\theta \in \Theta} \{A(\theta) + \epsilon_{n+1}\phi_{n+1}(a_{n+1}(\theta))\} \right] \leq \mathbb{E}_{\epsilon_{n+1}} \left[\sup_{\theta \in \Theta} \{A(\theta) + \epsilon_{n+1}a_{n+1}(\theta)\} \right]$$

We substitute this into the full expectation

$$LHS_{n+1} \leq E_{\epsilon_1,\ldots,\epsilon_{n+1}} \left[\sup_{\theta \in \Theta} \left\{ b(\theta) + \sum_{i=1}^{n} \epsilon_i \phi_i(a_i(\theta)) + \epsilon_{n+1}a_{n+1}(\theta) \right\} \right]$$

The expression contains n terms involving the 1-Lipschitz functions ϕ_i and one term involving $a_{n+1}(\theta)$. We apply the inductive hypothesis to the inner n terms.
using $C(\theta) = b(\theta) + \epsilon_{n+1}a_{n+1}(\theta)$, we restate:

$$LHS_{n+1} \leq \mathbb{E}_{\epsilon_1,\ldots,\epsilon_{n+1}} \left[\sup_{\theta \in \Theta} \left\{ C(\theta) + \sum_{i=1}^{n} \epsilon_i \phi_i(a_i(\theta)) \right\} \right]$$

The inductive hypothesis applies directly: for any baseline $C(\theta)$ and n Lipschitz-contracted terms, the supremum is bounded by the version with uncontracted arguments. Thus:

$$\mathbb{E}_{\epsilon_1,\ldots,\epsilon_n} \left[\sup_{\theta \in \Theta} \left\{ C(\theta) + \sum_{i=1}^{n} \epsilon_i \phi_i(a_i(\theta)) \right\} \right] \leq \mathbb{E}_{\epsilon_1,\ldots,\epsilon_n} \left[\sup_{\theta \in \Theta} \left\{ C(\theta) + \sum_{i=1}^{n} \epsilon_i a_i(\theta) \right\} \right]$$

Taking expectations over ϵ_{n+1} completes the step.

$$\mathbb{E}_{\epsilon_{n+1}}\left[\mathbb{E}_{\epsilon_1,\ldots,\epsilon_n}\left[\sup_{\theta\in\Theta}\left\{C(\theta)+\sum_{i=1}^{n}\epsilon_i a_i(\theta)\right\}\right]\right]=\mathbb{E}_{\epsilon_1,\ldots,\epsilon_{n+1}}\left[\sup_{\theta\in\Theta}\left\{b(\theta)+\epsilon_{n+1}a_{n+1}(\theta)+\sum_{i=1}^{n}\epsilon_i a_i(\theta)\right\}\right]$$

We obtain the RHS:

$$LHS_{n+1}\leq\mathbb{E}_{\epsilon_1,\ldots,\epsilon_{n+1}}\left[\sup_{\theta\in\Theta}\left\{b(\theta)+\sum_{i=1}^{n+1}\epsilon_i a_i(\theta)\right\}\right]=RHS_{n+1}$$

Let $a_i(f)=f(x_i)$ be the output of the prediction function $f\in\mathcal{F}$ on the i-th input. If the loss $\ell(y,u)$ is G-Lipschitz, then the function $\phi_i(u)=\frac{1}{G}\ell(y_i,u)$ is 1-Lipschitz (conditional on y_i). Applying the Contraction Principle conditionally on $\mathcal{D}$ relates the empirical Rademacher complexity of the loss functions $\hat{R}_n(\mathcal{H})$ to the empirical Rademacher complexity of the prediction functions $\hat{R}_n(\mathcal{F})$:

$$E_\epsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ell(y_i,f(x_i))\mid\mathcal{D}\right]\leq G\cdot E_\epsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x_i)\mid\mathcal{D}\right]$$

Taking the expectation over the data $\mathcal{D}$ yields the **Relation between complexities** of loss and prediction classes:

$$R_n(\mathcal{H})\leq G\cdot R_n(\mathcal{F}) \tag{56}$$

This tells us that the complexity of the entire hypothesis class expressed through the loss function ($\mathcal{H}$) is bounded by the complexity of the core prediction function class ($\mathcal{F}$), scaled by the Lipschitz constant G of the loss.

We extend Rademacher complexities into ball-constrained norms, as in Exercise 4.5. We assume that $\mathcal{F}=\{f_\theta(x)=\theta^\top\varphi(x),\Omega(\theta)\leq D\}$, where Ω is a norm on $\mathbb{R}^d$. We find (proof omitted) using the dual norm that

$$R_n(\mathcal{F})=\frac{D}{n}\mathbb{E}\left[\Omega^*(\Phi^\top\varepsilon)\right] \tag{57}$$

When $\Omega=\|\cdot\|_2$, we get

$$R_n(\mathcal{F})=\frac{D}{\sqrt{n}}\sqrt{\mathbb{E}\left[\|\phi(x)\|_2^2)\right]} \tag{58}$$

This is convenient in that the bound depends only on the constraint size D, the sample size n, and the expected squared ℓ_2 norm of the feature vector. It does not explicitly depend on the dimension d of the feature vector $\phi(x)\in R^d$.

Proofs

Given the Rademacher complexity of $\mathcal{H}$ defined in Eq. 50, prove that

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}h(z_i)-\mathbb{E}[h(z)]\right\}\right]\leq 2R_n(\mathcal{H}),\quad \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\mathbb{E}[h(z)]-\frac{1}{n}\sum_{i=1}^{n}h(z_i)\right\}\right]\leq 2R_n(\mathcal{H}) \tag{59}$$

We begin with the original dataset $\mathcal{D}$ and an independent copy $\mathcal{D}'$. As z_i' is an independent sample from the same distribution as z, $E[h(z_i')|\mathcal{D}]=E[h(z)]$. We assume $\mathcal{H}$, with $\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}h(z_i)-\mathbb{E}[h(z)]\right\}$, is a measurable random variable for each finite sample. We consider $\sup_{f\in\mathcal{F}}\{\mathcal{R}(f)-\widehat{\mathcal{R}}(f)\}$ as a loss

expression in the above context and express the expectation

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{E[h(z)]-\frac{1}{n}\sum_{i=1}^{n}h(z_i)\right\}\right]=\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}E[h(z_i')\mid D]-\frac{1}{n}\sum_{i=1}^{n}h(z_i)\right\}\right]$$

We can combine terms in the sum

$$=\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}E\left[h(z_i')-h(z_i)\mid\mathcal{D}\right]\right\}\right]$$

Considering that the supremum of a family of convex functions is convex, we have, using Jensen's Inequality Eq. 12:

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}E\left[h(z_i')-h(z_i)\mid\mathcal{D}\right]\right\}\right]\leq\mathbb{E}\left[\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}[h(z_i')-h(z_i)]\right\}\mid\mathcal{D}\right]\right]$$

and with the law of total expectation Eq. 21:

$$\leq\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}[h(z_i')-h(z_i)]\right\}\right]$$

We introduce a sequence $\epsilon=(\epsilon_1,\ldots,\epsilon_n)$ of independent Rademacher random variables ($\epsilon_i\in\{-1,1\}$ with $P(\epsilon_i=1)=P(\epsilon_i=-1)=1/2$), independent of D and D'. As z_i,z_i' are i.i.d., (z_i,z_i') has the same distribution as (z_i',z_i). Therefore $h(z_i')-h(z_i)$ is symmetric around zero. Multiplying this difference by ϵ_i does not change the distribution, so the expectation remains equal:

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}[h(z_i')-h(z_i)]\right\}\right]=\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i[h(z_i')-h(z_i)]\right\}\right]$$

$$=\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}\epsilon_ih(z_i')-\frac{1}{n}\sum_{i=1}^{n}\epsilon_ih(z_i)\right\}\right]$$

Using the triangle inequality Eq. 51 we can decompose

$$\leq\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}\epsilon_ih(z_i')\right\}\right]+\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{-\frac{1}{n}\sum_{i=1}^{n}\epsilon_ih(z_i)\right\}\right]$$

and using the definition of the Rademacher complexity Eq. 50

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}\epsilon_ih(z_i')\right\}\right]=R_n(\mathcal{H})$$

we note the symmetrical relationship:

$$\sup_{h\in\mathcal{H}}-\sum\epsilon_ih(z_i)=\sup_{h\in\mathcal{H}}\sum(-\epsilon_i)h(z_i)$$

as $-\epsilon_i$ has the same distribution as ϵ_i

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{-\frac{1}{n}\sum_{i=1}^{n}\epsilon_ih(z_i)\right\}\right]=R_n(\mathcal{H})$$

so now we have

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}h(z_i)-\mathbb{E}[h(z)]\right\}\right]\leq R_n(\mathcal{H})+R_n(\mathcal{H})=2R_n(\mathcal{H})\quad\square$$

Discussion

To manage infinite function classes, we introduced Rademacher Complexity ($R_n(\mathcal{H})$) as a data-dependent measure of how well a class can fit random noise. The central element of this is the Symmetrization Lemma (Eq. 59), which bounds the expected uniform deviation of empirical risk from true risk by $2R_n(\mathcal{H})$. Furthermore, the Contraction Principle (Eq. 56) allows us to simplify the problem to the complexity of the prediction functions themselves: $R_n(\mathcal{H})\leq GR_n(\mathcal{F})$. These results allow us to bound

generalization error based purely on the function class's correlation with noise, rather than its geometric parameters.

Ch. 4.5.4 - Linear Predictions with Rademacher Techniques

Preliminary Definitions

Cauchy-Schwarz inequality: For all vectors a and b of an inner product space,

$$|a^\top b| \leq |a|_2 |b|_2 \tag{60}$$

Derivations and Further Definitions

As in Eq. 58, we study prediction functions of the form $f_\theta(x) = \theta^\top \phi(x)$, with the assumption that parameter vector θ is constrained within a ball of radius D defined by some norm Ω
s.t. $\mathcal{F} = f_\theta(x) = \theta^\top \phi(x), \Omega(\theta) \leq D$. We further assume that the expected squared ℓ_2-norm of the feature vector is bounded as follows: $\mathbb{E}[\|\phi(x)\|_2^2] \leq R^2$. We first attempt to bound the estimation error, with $f_{\hat{\theta}}$ as the empirical risk minimizer over $\mathcal{F}$ and $\inf_{f\in\mathcal{F}} R(f)$ as the minimum true risk achievable within $\mathcal{F}$:

$$\mathbb{E}\left[R(f_{\hat{\theta}}) - \inf_{f\in\mathcal{F}} R(f)\right]$$

We observe as per Eq. 48

$$R(f_{\hat{\theta}}) - R(g_{\mathcal{F}}) \leq 2 \sup_{f\in\mathcal{F}} |R(f) - \hat{R}(f)|$$

$$\mathbb{E}\left[R(f_{\hat{\theta}}) - \inf_{f\in\mathcal{F}} R(f)\right] \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} |R(f) - \hat{R}(f)|\right]$$

We can make use of Eq. 59

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} |R(h) - \hat{R}(h)|\right] \leq 2R_n(\mathcal{H})$$

$$\therefore \quad \mathbb{E}\left[R(f_{\hat{\theta}}) - \inf_{f\in\mathcal{F}} R(f)\right] \leq 4R_n(\mathcal{H})$$

We assume $\ell(y, \cdot)$ is G-Lipschitz-continuous and use Eq. 56

$$R_n(\mathcal{H}) \leq G \cdot R_n(\mathcal{F})$$

$$\mathbb{E}\left[R(f_{\hat{\theta}}) - \inf_{f\in\mathcal{F}} R(f)\right] \leq 4GR_n(\mathcal{F})$$

and we take the dimension-independent bound in Eq. 58

$$R_n(\mathcal{F}) = \frac{D}{\sqrt{n}} \sqrt{\mathbb{E}\left[\|\phi(x)\|_2^2)\right]}$$

with the assumption that

$$\mathbb{E}[\|\phi(x)\|_2^2] \leq R^2$$

$$\mathbb{E}\left[R(f_{\hat{\theta}}) - \inf_{f\in\mathcal{F}} R(f)\right] \leq 4G \cdot \frac{DR}{\sqrt{n}}$$

We derive from this a bounded expression for **Estimation error linear predictions**:

$$\mathbb{E}[R(f_{\hat{\theta}})] \leq \inf_{\|\theta\|_2 \leq D} R(f_\theta) + \frac{4GRD}{\sqrt{n}} \tag{61}$$

With similar constraints and assumptions, and the assumption that there is a minimizer θ^* of $R(f\theta)$ over $\mathbb{R}^d$ we explore the approximation error

$$\inf_{\|\theta\|_2 \leq D} R(f_\theta) - R(f_{\theta^*})$$

we use bounded excess risk from Eq. 46

$$R(f_\theta) - R(f_{\theta'}) = E[\ell(y, f_\theta(x)) - \ell(y, f_{\theta'}(x))] \leq GE[|f_\theta(x) - f_{\theta'}(x)|]$$

and set $f_{\theta'} = f_{\theta^*}$

$$R(f_\theta) - R(f_{\theta^*}) \leq GE[|f_\theta(x) - f_{\theta^*}(x)|]$$

and as per the linear structure of the models

$$|f_\theta(x) - f_{\theta^*}(x)| = |\theta^\top \phi(x) - (\theta^*)^\top \phi(x)| = |(\theta - \theta^*)^\top \phi(x)|$$

$$R(f_\theta) - R(f_{\theta^*}) \leq G\mathbb{E}[|(\theta - \theta^*)^\top \phi(x)|]$$

making use of the Cauchy-Schwartz inequality (Eq. 60) with

$$a = \theta - \theta^*, \quad b = \phi(x)$$

$$|(\theta - \theta^*)^\top \phi(x)| \leq \|\theta - \theta^*\|_2 \cdot \|\phi(x)\|_2$$

$$R(f_\theta) - R(f_{\theta^*}) \leq G\mathbb{E}[\|\theta - \theta^*\|_2 \cdot \|\phi(x)\|_2]$$

we can move $\|\theta - \theta^*\|_2$ outside of the expectation

$$R(f_\theta) - R(f_{\theta^*}) \leq G\|\theta - \theta^*\|_2 \cdot \mathbb{E}[\|\phi(x)\|_2]$$

using Jensen's inequality (Eq. 12) as $\|\phi(x)\|_2$ is a non-negative function

$$F(\mathbb{E}[X]) \leq \mathbb{E}[F(X)] \quad \text{or} \quad (\mathbb{E}[Y])^2 \leq \mathbb{E}[Y^2], \quad \mathbb{E}[Y] \leq \sqrt{\mathbb{E}[Y^2]}, Y = \|\phi(x)\|_2$$

$$\mathbb{E}[\|\phi(x)\|_2] \leq \sqrt{\mathbb{E}[\|\phi(x)\|_2^2]}$$

with the assumption that

$$\mathbb{E}[\|\phi(x)\|_2^2] \leq R^2$$

$$R(f_\theta) - R(f_{\theta^*}) \leq GR\|\theta - \theta^*\|_2$$

now we take the infimum

$$\inf_{\|\theta\|_2 \leq D} R(f_\theta) - R(f_{\theta^*}) \leq \inf_{\|\theta\|_2 \leq D} GR\|\theta - \theta^*\|_2$$

$$\leq GR \inf_{\|\theta\|_2 \leq D} \|\theta - \theta^*\|_2$$

and similar to the proof for Eq. 49

$$\inf_{\|\theta\|_2 \leq D} \|\theta - \theta^*\|_2 = (\|\theta^*\|_2 - D)_+$$

We arrive at a bounded expression for **Approximation error linear predictions**

$$\inf_{\|\theta\|_2 \leq D} R(f_\theta) - R(f_{\theta^*}) \leq GR(\|\theta^*\|_2 - D)_+ \tag{62}$$

We can combine approximation and estimation error bounds into a single **Linear Generalization Bound:**

$$E[R(f_{\hat{\theta}})] - R(f_{\theta^*}) \leq \underbrace{GR(\|\theta^*\|_2 - D)_+}_{\text{Approximation Error}} + \underbrace{\frac{4GRD}{\sqrt{n}}}_{\text{Estimation Error}} \tag{63}$$

We arrive at a bounded expression dependent on the Lipschitz constant G of the loss function (sensitivity of the function to changes in prediction output), the radius R of the expected feature norm of $\phi(x)$, the radius D of the parameter norm, the sample size n, and the optimal unconstrained parameter vector θ^*.

Discussion

This section combines our work on linear models with complexity theory. By applying Rademacher techniques to linear predictors constrained by ℓ_2 norms, we derived the explicit Linear Generalization Bound (Eq. 63): Excess Risk $\leq GR(\|\theta^*\|_2 - D)_+ + \frac{4GRD}{\sqrt{n}}$. This bound depends on the parameter norm D and feature norm R rather the input dimension d. This proves that we can learn effectively even in infinite-dimensional spaces, provided that our data and parameters have small norms.

Ch. 5 - Optimization for Machine Learning

Preliminary Definitions

A differentiable function F is said to be μ **strongly-convex** with $\mu > 0$, iff

$$F(\eta) \geq F(\theta) + F'(\theta)^\top(\eta - \theta) + \frac{\mu}{2}\|\theta - \eta\|_2^2, \quad \forall \eta, \theta \in \mathbb{R}^d \tag{64}$$

A differentiable function F is said to be L**-smooth** iff

$$|F(\eta) - F(\theta) - F'(\theta)^\top(\eta - \theta)| \leq \frac{L}{2}\|\theta - \eta\|_2^2, \quad \forall \theta, \eta \in \mathbb{R}^d \tag{65}$$

A twice-differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ is **convex** $\theta \in \mathbb{R}^d$ iff its Hessian is positive semidefinite:

$$x^\top F''(\theta)x \geq 0, \quad \forall x \in \mathbb{R}^d \tag{66}$$

A twice-differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ is μ**-strongly convex** iff its Hessian satisfies the positive semidefinite condition with $\mu > 0$:

$$x^\top F''(\theta)\mathbf{x} \geq \mu|x|_2^2, \quad \forall x \in \mathbb{R}^d \tag{67}$$

A twice-differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ is **L-smooth** iff its Hessian satisfies the positive semidefinite condition:

$$x^\top F''(\theta)\mathbf{x} \leq L|x|_2^2, \quad \forall x \in \mathbb{R}^d \tag{68}$$

For a parameterization $\{f_\theta\}_{\theta \in \mathbb{R}^d}$ and a regularizer $\Omega : \mathbb{R}^d \to \mathbb{R}$, the function $F : \mathbb{R}^d \to \mathbb{R}$ is called the **Objective Function**:

$$F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) + \Omega(\theta) \tag{69}$$

We note a convenient consequence of convexity:

$$\text{If } F : \mathbb{R}^d \to \mathbb{R} \text{ is convex and } A : \mathbb{R}^{d'} \to \mathbb{R}^d \text{ is affine, then } F \circ A : \mathbb{R}^{d'} \to \mathbb{R} \text{ is convex} \tag{70}$$

Derivations and Further Definitions

We can restate excess risk in estimation error as in Eq. 47 with the central term restated as **optimization error:**

$$\left\{\hat{\mathcal{R}}(f_{\hat{\theta}}) - \inf_{\theta \in \mathbb{R}^d} \hat{\mathcal{R}}(f_\theta)\right\} \tag{71}$$

which will equal zero if $\hat{\theta}$ is the minimizer of $\hat{\mathcal{R}}$.
If we attempt to minimize our objective function $F : \mathbb{R}^d \to \mathbb{R}$, $\min_{\theta \in \mathbb{R}^d} F(\theta)$, we take $\theta_0 \in \mathbb{R}^d$ and make use of the **Gradient Descent** algorithm for $t \geq 1$:

$$\theta_t = \theta_{t-1} - \gamma_t F'(\theta_{t-1}) \tag{72}$$

where $\gamma_t > 0$ is the step size. When Gradient Descent is used to minimize the empirical risk $\hat{R}(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$, computing the full gradient $F'(\theta_{t-1})$ requires accessing the full dataset to compute all individual loss gradients and take the mean.
We can parse some critical elements of gradient descent through its applications to OLS, where our objective function takes the form $F(\theta) = \frac{1}{2n}|\Phi\theta - y|_2^2$ with $\Phi \in \mathbb{R}^{n \times d}$ as the design matrix. We can find the first derivative

$$F'(\theta) = \frac{1}{n}\Phi^\top(\Phi\theta - y) = \frac{1}{n}\Phi^\top\Phi\theta - \frac{1}{n}\Phi^\top y$$

from which we derive the Hessian

$$H = F''(\theta) = \frac{\partial}{\partial\theta}\left[F'(\theta)\right] = \frac{\partial}{\partial\theta}\left[\frac{1}{n}\Phi^\top\Phi\theta\right] - \frac{\partial}{\partial\theta}\left[\frac{1}{n}\Phi^\top y\right] = \frac{1}{n}\Phi^\top\Phi \in \mathbb{R}^{d \times d}$$

We are looking for the optimal solution η^*, characterized by the zero gradient condition: $F'(\eta^*) = 0$, or

$$\frac{1}{n}\Phi^\top\Phi\eta^* = \frac{1}{n}\Phi^\top y$$

.
We have $F'(\eta^*) = H\eta^* - \frac{1}{n}\Phi^\top y$, and as $F'(\eta^*) = 0$, $H\eta^* = \frac{1}{n}\Phi^\top y$, so we restate

$$F'(\theta) = H\theta - H\eta^* = H(\theta - \eta^*)$$

We note that as per Eq. 66, as H is independent of θ, F is convex. We further assume that the Hessian is bounded in $[\mu, L]$ as per Eq.'s 67, 68 and define the **condition number**:

$$\kappa = \frac{L}{\mu} \tag{73}$$

where we note that μ and L are the smallest and largest eigenvector of the Hessian, respectively.
We return to gradient descent from Eq. 72 on OLS, wherein we substitute the Hessian and optimal solution

$$\begin{aligned}
\theta_t &= \theta_{t-1} - \gamma H(\theta_{t-1} - \eta^*) \\
&\quad \text{from which we subtract the optimal solution} \\
\theta_t &= \theta_{t-1} - \gamma H\theta_{t-1} + \gamma H\eta^* \\
\theta_t - \eta^* &= \theta_{t-1} - \gamma H\theta_{t-1} + \gamma H\eta^* - \eta^* \\
&= (I - \gamma H)\theta_{t-1} - (I - \gamma H)\eta^* \\
&= (I - \gamma H)(\theta_{t-1} - \eta^*) \\
&\quad \text{which after } t \text{ steps becomes} \\
\theta_t - \eta^* &= (I - \gamma H)^t(\theta_0 - \eta^*)
\end{aligned}$$

This shows that the deviation from the optimal solution is multiplied by the matrix $(I - \gamma H)$ at every step.
As we assume that the eigenvalues of H are in $[\mu, L]$, the eigenvalues of $(I - \gamma H)$ are in $[1 - \gamma L, 1 - \gamma\mu]$. Therefore the squared distance to the optimum, $|\theta_t - \eta^*|_2^2$, is bounded by analyzing the magnitude of the largest eigenvalue of $(I - \gamma H)^t$:

$$|\theta_t - \eta^*|_2^2 \leq \left(\max_{\lambda \in [\mu, L]} |1 - \gamma\lambda|\right)^{2t} |\theta_0 - \eta^*|_2^2$$

From this emerges that $\max_{\lambda \in [\mu, L]} |1 - \gamma\lambda|$ is minimized when the step size is set to the **Optimal Step Size:**

$$\gamma = \frac{2}{\mu + L} \tag{74}$$

and as such the convergence rate is $\propto \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2t}$.

Discussion

We established the fundamental definitions of the objective landscape, Strong Convexity (μ) and L-Smoothness, which act as lower and upper quadratic bounds on the function, respectively.
By analyzing Gradient Descent on the concrete case of OLS, we derived the Optimal Step Size $\gamma = \frac{2}{\mu + L}$ (Eq. 74). This derivation is significant because it proves that the convergence speed is determined by the eigenvalues of the Hessian $(I - \gamma H)$. This establishes the Condition Number $\kappa = L/\mu$ as the critical factor: if the Hessian's eigenvalues are spread out ($\kappa \gg 1$), the optimal step size must be small to accommodate the sharpest direction, slowing progress in the flattest direction.

Ch.s 5.2.2-3 - Convex Functions and Their Properties, Analysis of Gradient Descent for Strongly Convex and Smooth Functions

Preliminary Definitions

Łojasiewicz's inequality states that if F is differentiable and μ-strongly convex, with unique minimizer η^* characterized by $F'(\eta^*) = 0$, we have

$$\|F'(\theta)\|_2^2 \geq 2\mu(F(\theta) - F(\eta_*)), \quad \forall\theta \in \mathbb{R}^d \tag{75}$$

For $F : \mathbb{R}^d \to \mathbb{R}$ convex and μ is a probability measure on $\mathbb{R}^d$, we define **Jensen's Inequality in the Continuous Case**:

$$F\left(\int_{\mathbb{R}^d} \theta \, d\mu(\theta)\right) \leq \int_{\mathbb{R}^d} F(\theta) \, d\mu(\theta) \tag{76}$$

Derivations and Further Definitions

It is worth noticing that when a function is both smooth and strongly convex, we can observe the condition number (Eq. 73) $\kappa = \frac{L}{\mu} \geq 1$

We can take Eq. 1 in the context of η^* and rearrange to arrive at **convexity with a global minimizer**:

$$F(\theta) - F(\eta^*) \leq F'(\theta)^\top (\theta - \eta^*) \tag{77}$$

This form gives us an upper bound on the distance to the optimum, $F(\theta) - F(\eta^*)$, bounded by the inner product of the gradient $F'(\theta)$ and the vector from the optimum in the direction of the current point, $(\theta - \eta^*)$. In addition, any stationary point can be the global minimizer.
We note from Ex. 5.5 that when an objective function F is convex, $F + \frac{\mu}{2}\|\cdot\|_2^2$is μ-strongly convex.
We tie the previous elements of Ch. 5 together in a proposition on **Convergence of GD for smooth strongly convex functions:** Assume that F is L-smooth and μ-strongly convex. Choosing $\gamma_t = 1/L$, the iterates $(\theta_t)_{t\geq 0}$ of GD on F satisfy

$$F(\theta_t) - F(\eta_*) \leq \left(1 - \frac{\mu}{L}\right)^t (F(\theta_0) - F(\eta_*)) \leq \exp(-\frac{t}{\kappa})(F(\theta_0) - F(\eta_*)) \tag{78}$$

We can arrive here by applying the L-smoothness inequality to Eq.'s 65 and 72 at θ_{t-1} for the next iterate $\theta_t = \theta_{t-1} - \frac{1}{L}F'(\theta_{t-1})$:

$$\begin{aligned}
F(\theta_t) &= F(\theta_{t-1} - F'(\theta_{t-1})/L) \\
&\leq F(\theta_{t-1}) + F'(\theta_{t-1})^\top(-F'(\theta_{t-1})/L) + \frac{L}{2}\|-F'(\theta_{t-1})/L\|_2^2 \\
&= F(\theta_{t-1}) - \frac{1}{L}\|F'(\theta_{t-1})\|_2^2 + \frac{1}{2L}\|F'(\theta_{t-1})\|_2^2 \\
F(\theta_t) &\leq F(\theta_{t-1}) - \frac{1}{2L}\|F'(\theta_{t-1})\|_2^2 \\
F(\theta_t) - F(\eta^*) &\leq F(\theta_{t-1}) - F(\eta^*) - \frac{1}{2L}\|F'(\theta_{t-1})\|_2^2 \\
\|F'(\theta)\|_2^2 &\geq 2\mu(F(\theta) - F(\eta^*)) \\
F(\theta_t) - F(\eta^*) &\leq F(\theta_{t-1}) - F(\eta^*) - \frac{1}{2L}\left[2\mu(F(\theta_{t-1}) - F(\eta^*))\right] \\
F(\theta_t) - F(\eta^*) &\leq \left(1 - \frac{\mu}{L}\right)(F(\theta_{t-1}) - F(\eta^*)) \\
F(\theta_t) - F(\eta^*) &\leq \left(1 - \frac{1}{\kappa}\right)(F(\theta_{t-1}) - F(\eta^*))
\end{aligned}$$

Recursion on t brings us to Eq. 78.

Proofs

Exercise 5.5: Show that function $F : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly-convex if and only if function $\theta \mapsto F(\theta) - \frac{\mu}{2}\|\theta\|_2^2$ is convex.

$$
\begin{aligned}
&\text{Assume } F \text{ is a differentiable function}\\
\text{Let } G(\theta) &= F(\theta) - \frac{\mu}{2}\|\theta\|_2^2\\
G'(\theta) &= F'(\theta) - \mu\theta\\
F'(\theta) &= G'(\theta) + \mu\theta\\
&\text{Assume } F \text{ is } \mu\text{-strongly convex:}\\
F(\eta) &\geq F(\theta) + F'(\theta)^\top(\eta-\theta) + \frac{\mu}{2}\|\eta-\theta\|_2^2\\
\left(G(\eta) + \frac{\mu}{2}\|\eta\|_2^2\right) &\geq \left(G(\theta) + \frac{\mu}{2}\|\theta\|_2^2\right) + (G'(\theta)+\mu\theta)^\top(\eta-\theta) + \frac{\mu}{2}\|\eta-\theta\|_2^2\\
\frac{\mu}{2}\|\eta-\theta\|_2^2 &= \frac{\mu}{2}\left(\|\eta\|_2^2 - 2\eta^\top\theta + \|\theta\|_2^2\right)\\
\mu\theta^\top(\eta-\theta) &= \mu\eta^\top\theta - \mu\|\theta\|_2^2\\
G(\eta) + \frac{\mu}{2}\|\eta\|_2^2 &\geq G(\theta) + \frac{\mu}{2}\|\theta\|_2^2 + G'(\theta)^\top(\eta-\theta) + (\mu\eta^\top\theta - \mu\|\theta\|_2^2) + \frac{\mu}{2}\|\eta\|_2^2 - \mu\eta^\top\theta + \frac{\mu}{2}\|\theta\|_2^2\\
G(\eta) + \frac{\mu}{2}\|\eta\|_2^2 &\geq G(\theta) + G'(\theta)^\top(\eta-\theta) + \frac{\mu}{2}\|\eta\|_2^2\\
G(\eta) &\geq G(\theta) + G'(\theta)^\top(\eta-\theta)\\
&\therefore G \text{ is convex}\\
&\text{Assume } G(\theta) = F(\theta) - \frac{\mu}{2}\|\theta\|_2^2 \text{ is convex}\\
\left(F(\eta) - \frac{\mu}{2}\|\eta\|_2^2\right) &\geq \left(F(\theta) - \frac{\mu}{2}\|\theta\|_2^2\right) + (F'(\theta)-\mu\theta)^\top(\eta-\theta)\\
F(\eta) &\geq F(\theta) + F'(\theta)^\top(\eta-\theta) - \mu\theta^\top(\eta-\theta) + \frac{\mu}{2}\|\eta\|_2^2 - \frac{\mu}{2}\|\theta\|_2^2\\
-\mu\theta^\top(\eta-\theta) + \frac{\mu}{2}\|\eta\|_2^2 - \frac{\mu}{2}\|\theta\|_2^2 &= (-\mu\eta^\top\theta + \mu\|\theta\|_2^2) + \frac{\mu}{2}\|\eta\|_2^2 - \frac{\mu}{2}\|\theta\|_2^2\\
&= \frac{\mu}{2}\|\eta\|_2^2 - \mu\eta^\top\theta + \left(\mu\|\theta\|_2^2 - \frac{\mu}{2}\|\theta\|_2^2\right)\\
&= \frac{\mu}{2}\left(\|\eta\|_2^2 - 2\eta^\top\theta + \|\theta\|_2^2\right)\\
F(\eta) &\geq F(\theta) + F'(\theta)^\top(\eta-\theta) + \frac{\mu}{2}\left(\|\eta\|_2^2 - 2\eta^\top\theta + \|\theta\|_2^2\right)\\
F(\eta) &\geq F(\theta) + F'(\theta)^\top(\eta-\theta) + \frac{\mu}{2}\|\eta-\theta\|_2^2\\
&\therefore F \text{ is } \mu \text{ strongly convex}
\end{aligned}
$$

$\therefore F$ is μ-strongly convex iff $G(\theta) = F(\theta) - \frac{\mu}{2}\|\theta\|_2^2$ is convex $\square$

Exercise 5.6: Show that if function $F : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly-convex, then it has a unique minimizer.

Assume that a global minimizer exists in $\mathbb{R}^d$. Assume that F possesses two distinct global minimizers,

η^* and θ^*, s.t. $\eta^* \neq \theta^*$. This necessitates: $F(\eta^*) = F(\theta^*)$ As a consequence of μ-strong-convexity:

$$
\begin{aligned}
F'(\eta^*) = 0 \text{ and } F'(\theta^*) = 0 \\
F(\eta) \geq F(\theta) + F'(\theta)^\top(\eta - \theta) + \frac{\mu}{2}\|\eta - \theta\|_2^2 \\
\text{for } \eta = \eta^* \text{ and } \theta = \theta^* \\
F(\eta^*) \geq F(\theta^*) + F'(\theta^*)^\top(\eta^* - \theta^*) + \frac{\mu}{2}\|\eta^* - \theta^*\|_2^2 \\
F'(\theta^*) = 0 \\
\text{we get } F(\eta^*) \geq F(\theta^*) + \frac{\mu}{2}\|\eta^* - \theta^*\|_2^2 \\
\text{for } \eta = \theta^* \text{ and } \theta = \eta^* \\
F(\theta^*) \geq F(\eta^*) + F'(\eta^*)^\top(\theta^* - \eta^*) + \frac{\mu}{2}\|\theta^* - \eta^*\|_2^2 \\
F'(\eta^*) = 0 \\
\text{we get } F(\theta^*) \geq F(\eta^*) + \frac{\mu}{2}\|\theta^* - \eta^*\|_2^2 \\
F(\theta^*) = F(\eta^*) \\
F(\eta^*) \geq F(\eta^*) + \frac{\mu}{2}\|\eta^* - \theta^*\|_2^2 \\
F(\eta^*) - F(\eta^*) \geq F(\eta^*) + \frac{\mu}{2}\|\eta^* - \theta^*\|_2^2 - F(\eta^*) \\
0 \geq \frac{\mu}{2}\|\eta^* - \theta^*\|_2^2 \\
\text{by } \mu\text{-strong convexity we have } \mu > 0 \\
\therefore \|\eta^* - \theta^*\|_2^2 = 0 \\
\eta^* = \theta^* \\
\text{but } \eta^* \neq \theta^* \\
\therefore \ \eta^* \text{ must be unique } \square
\end{aligned}
$$

Discussion

Here we formalized the intuition from the OLS case into a general proof for all L-smooth, μ-strongly convex functions. We derived the explicit Linear Convergence Rate $F(\theta_t) - F(\eta_*) \leq \left(1 - \frac{1}{\kappa}\right)^t (F(\theta_0) - F(\eta_*))$ (Eq. 78). This result confirms that "conditioning" is the speed limit of gradient descent. When $\kappa \approx 1$ (spherical geometry), the error contracts rapidly. As $\kappa \to \infty$ (elongated valleys), the term $(1-1/\kappa)$ approaches 1, and convergence stalls.

Ch. 5.2.5 - Beyond Gradient Descent

Preliminary Definitions

We can extend our convex optimization techniques with that of Nesterov's Accelerated Gradient Descent, which is characterized by a two-sequence iteration with a 'momentum' term. We define the **extrapolation parameter** β_t as a weighting factor that determines the influence of the previous trajectory on the current iterate ('momentum'). In the smooth convex setting $\mu = 0$, this parameter is time-dependent: $\beta_t = \frac{t-1}{t+2}$. In the smooth strongly convex setting $\mu > 0$, the parameter is fixed based on the condition number $\kappa = L/\mu$: $\beta = \frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}$.

Derivations and Further Definitions

The algorithmic framework for **Nesterov AGD** [2] utilizes two coupled sequences, θ_t and η_t, initialized s.t. $\theta_0 = \eta_0$. Proofs are omitted and we will instead here summarize Nesterov's results.
For an L-smooth, μ-strongly convex objective function F, θ_t is computed via a gradient step from the

auxiliary sequence η_{t-1}, which is then followed by the update of the auxiliary point through extrapolation:

$$\theta_t = \eta_{t-1} - \frac{1}{L}F'(\eta_{t-1})$$

$$\eta_t = \theta_t + \frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}(\theta_t - \theta_{t-1})$$

By performing gradient evaluation at the extrapolated point η_{t-1} rather than θ_{t-1}, the method achieves a linear convergence rate:

$$F(\theta_t) - F(\eta^*) \leq \exp(-t\sqrt{\mu/L})(F(\theta_0) - F(\eta^*))$$

This effectively improves the characteristic time of convergence from κ to $\sqrt{\kappa}$.

For L-smooth convex functions lacking strong convexity ($\mu = 0$), the algorithm uses a specific schedule for the extrapolation parameter to ensure accelerated sublinear convergence:

$$\theta_t = \eta_{t-1} - \frac{1}{L}F'(\eta_{t-1})$$

$$\eta_t = \theta_t + \frac{t-1}{t+2}(\theta_t - \theta_{t-1})$$

This sequence establishes that the objective function sub-optimality decays at a rate of $O(1/t^2)$, satisfying the bound $F(\theta_t) - F(\eta^*) \leq \frac{2L|\theta_0 - \eta^*|_2^2}{(t+1)^2}$.

Discussion

We identified that standard Gradient Descent is not information-theoretically optimal; its dependence on κ is too strong. By introducing Nesterov's Accelerated Gradient Descent, we utilized a momentum term via the auxiliary sequence η_t to correct the trajectory. The result is a fundamental improvement in the convergence rate: for strongly convex functions, the dependence improves from $\exp(-t/\kappa)$ to $\exp(-t/\sqrt{\kappa})$. For general convex functions ($\mu = 0$), it improves from $O(1/t)$ to $O(1/t^2)$. This suggests that respecting the local geometry (via momentum) allows us to beat the standard gradient descent barrier.

Ch. 5.3 - Gradient Methods on Non-Smooth Problems

Preliminary Definitions

With $F : \mathbb{R}^d \to \mathbb{R}$ as a convex function, the set of vectors z that serve as the slopes of lower-bounding tangents to the function at the point θ is called the **subdifferential**, denoted:

$$\partial F(\theta) = \{z \in \mathbb{R}^d, \forall \eta \in \mathbb{R}^d, F(\eta) \geq F(\theta) + z^\top(\eta - \theta)\} \tag{79}$$

A **subgradient** is defined as any arbitrary element z selected from this subdifferential set.
We can describe a function $F : \mathbb{R}^d \to \mathbb{R}$ as B-**Lipschitz-continuous** iff the magnitude of the change in function value is linearly bounded by the distance between points, specifically:

$$|F(\eta) - F(\theta)| \leq B|\eta - \theta|_2, \forall \theta, \eta \in \mathbb{R}^d \tag{80}$$

Derivations and Further Definitions

If we consider a function that is B-Lipschitz-continuous but is not differentiable, we can consider non-smooth optimization using a subgradient method. The **subgradient method** proceeds iteratively, utilizing any selected subgradient $z_t \in \partial F(\theta_{t-1})$ as the descent direction, usually defined as $\theta_t = \theta_{t-1} - \gamma_t z_t$, where $\gamma_t > 0$ is the step size.
Expanding the squared distance to the minimizer η^*:

$$\|\theta_t - \eta^*\|_2^2 = \|\theta_{t-1} - \eta^*\|_2^2 - 2\gamma_t z_t^\top(\theta_{t-1} - \eta^*) + \gamma_t^2\|z_t\|_2^2$$

Using the subgradient definition $F(\eta^*) \geq F(\theta_{t-1}) + z_t^\top(\eta^* - \theta_{t-1})$ and the Lipschitz bound $|z_t|_2 \leq B$:

$$\|\theta_t - \eta^*\|_2^2 \leq \|\theta_{t-1} - \eta^*\|_2^2 - 2\gamma_t[F(\theta_{t-1}) - F(\eta^*)] + \gamma_t^2 B^2 \tag{81}$$

Rearranging to isolate the error term and summing over t steps:

$$2\sum_{s=1}^{t} \gamma_s[F(\theta_{s-1}) - F(\eta^*)] \leq \sum_{s=1}^{t}(|\theta_{s-1} - \eta^*|_2^2 - |\theta_s - \eta^*|_2^2) + B^2\sum_{s=1}^{t}\gamma_s^2$$

The first sum telescopes to $|\theta_0 - \eta^*|_2^2 - |\theta_t - \eta^*|_2^2$. Assuming $|\theta_0 - \eta^*|_2 \leq D$ and dropping the negative term:

$$2\sum_{s=1}^{t} \gamma_s[F(\theta_{s-1}) - F(\eta^*)] \leq D^2 + B^2\sum_{s=1}^{t}\gamma_s^2$$

We bound the minimum excess error $\min_{0\leq s<t}(F(\theta_s) - F(\eta^*))$ using the weighted average:

$$\min_{0\leq s<t}(F(\theta_s) - F(\eta^*)) \leq \frac{D^2 + B^2\sum_{s=1}^{t}\gamma_s^2}{2\sum_{s=1}^{t}\gamma_s}$$

Choosing a diminishing step size $\gamma_s = \frac{D}{B\sqrt{s}}$, we bound the series:
1. $\sum_{s=1}^{t}\gamma_s^2 = \frac{D^2}{B^2}\sum_{s=1}^{t}\frac{1}{s} \leq \frac{D^2}{B^2}(1 + \log t)$ 2. $\sum_{s=1}^{t}\gamma_s = \frac{D}{B}\sum_{s=1}^{t}\frac{1}{\sqrt{s}} \geq \frac{D}{B}\sqrt{t}$ (Using loose lower bound $\sum_{s=1}^{t} s^{-1/2} \geq \sqrt{t}$)

Substituting these into the error bound:

$$\min_{0\leq s<t}(F(\theta_s) - F(\eta^*)) \leq \frac{D^2 + D^2(1 + \log t)}{2\frac{D}{B}\sqrt{t}} = \frac{DB(2 + \log t)}{2\sqrt{t}}$$

This demonstrates a convergence rate of $O(\frac{\log t}{\sqrt{t}})$
We can eliminate the logarithmic factor to achieve $O(1/\sqrt{t})$ by optimizing the step size choice. Instead of the decaying schedule $\gamma_s \propto 1/\sqrt{s}$, we select a constant step size γ fixed in advance based on the total number of iterations t.
Starting from the general bound derived above:

$$\min_{0\leq s\leq t-1}(F(\theta_s) - F(\eta^*)) \leq \frac{D^2 + B^2\sum_{s=1}^{t}\gamma_s^2}{2\sum_{s=1}^{t}\gamma_s}$$

If we set a constant step size $\gamma_s = \gamma$ for all $s = 1 \dots t$:

$$\cdots \leq \frac{D^2 + B^2 t\gamma^2}{2t\gamma} = \frac{D^2}{2t\gamma} + \frac{B^2\gamma}{2}$$

Minimizing this expression wrt γ yields the optimal constant step size $\gamma = \frac{D}{B\sqrt{t}}$. Substituting this back into the equation:

$$\cdots \leq \frac{D^2}{2t\left(\frac{D}{B\sqrt{t}}\right)} + \frac{B^2\left(\frac{D}{B\sqrt{t}}\right)}{2} = \frac{DB}{2\sqrt{t}} + \frac{DB}{2\sqrt{t}} = \frac{DB}{\sqrt{t}}$$

Thus, with a tuned constant step size, the convergence rate is $O(1/\sqrt{t})$
We note that if the non-smooth objective function F is also μ-strongly convex, the convergence rate improves significantly. By adopting a step size of $\gamma_t = \frac{2}{\mu(t+1)}$, it can be shown (proof omitted) that the error bounds decay at $O(1/t)$ rather than $O(1/\sqrt{t})$:

$$F(\theta_t) - F(\eta^*) \leq \frac{2B^2}{\mu(t+1)} \tag{82}$$

Discussion

As per Eq. 61, for convex Lipschitz-continuous losses for linear predictions with feature ℓ_2-norms smaller than R and a parameter bounded in the ℓ_2-norm by D, the estimation error for the empirical risk minimizer is upper-bounded by a constant times $GRD/\sqrt{n}$. We now extend this to see that the optimization error after t iterations of the subgradient method is upper-bounded by a constant times $GRD/\sqrt{t}$ since the Lipschitz constant of the objective function is $B \leqslant GR$. The combination of these facts indicates that iterations t need not be larger than the number of observations n.

Ch. 5.4 - Stochastic Gradient Descent

Preliminary Definitions

A **telescoping sum** is a finite series in which subsequent terms cancel each other out, leaving only the initial and final boundary terms. For a sequence of numbers (a_k), the sum telescopes as follows:

$$\sum_{k=1}^{T}(a_{k+1} - a_k) = a_{T+1} - a_1 \tag{83}$$

Derivations and Further Definitions

For gradient descent on $F(\theta) = \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, f_\theta(x_i))+\Omega(\theta)$, we must compute the gradient $F'(\theta_{t-1})$ across the full dataset on each iteration. We can instead use the **unbiased stochastic estimate** $g_t(\theta_{t-1})$ over a subset of the data s.t.

$$\mathbb{E}[g_t(\theta_{t-1})|\theta_{t-1}] = F'(\theta_{t-1}) \tag{84}$$

which we can use in the **Stochastic gradient descent** (SGD) algorithm: take a step-size sequence $(\gamma_t)_{t\geqslant 0}$, pick $\theta_0 \in \mathbb{R}^d$, and for $t \geqslant 1$, let

$$\theta_t = \theta_{t-1} - \gamma_t g_t(\theta_{t-1}), \tag{85}$$

where $g_t(\theta_{t-1})$ satisfies Eq. 84.
We can look at SGD in the contexts of Empirical and Expected Risk Minimization. Empirical ($\hat{R}$): When the objective function $F(\theta)$ is the empirical risk, $\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, f_\theta(x_i))$, the stochasticity is introduced by sampling an index $i(t) \in 1, \ldots, n$ and using the gradient of the loss for that single sample. Expected (R): When the objective function $F(\theta)$ is the expected risk, $R(\theta) = E[\ell(y, f_\theta(x))]$, the stochastic gradient g_t is derived from a new, independent observation (x_t, y_t), which is the gradient of the loss $\ell(y_t, f_\theta(x_t))$.
If we assume the unbiased behavior of Eq. 84 as well the bounded gradient

$$|g_t(\theta^{t-1})|_2^2 \leq B^2 \quad \text{almost surely} \tag{86}$$

both for $\forall t \geq 1$ we can derive some interesting **Convergence of SGD** results: Assume that F is convex, is B-Lipschitz, and admits a minimizer θ_* that satisfies $\|\theta_* - \theta_0\|_2 \leqslant D$, with our previously stated assumptions. Choosing $\gamma_t = (D/B)/\sqrt{t}$, the iterates $(\theta_t)_{t\geqslant 0}$ of SGD on F satisfy

$$\mathbb{E}\left[F(\bar{\theta}_t) - F(\theta_*)\right] \leqslant DB\frac{2+\log(t)}{2\sqrt{t}}, \tag{87}$$

It is noted that Agarwal et al. [3] that the bound in $O(\frac{BD}{\sqrt{t}})$ is optimal for this class of problem (having a better convergence rate is impossible).
We can extend our convergence with a constant step size γ instead of a strictly decreasing one when the optimization horizon T is known. This approach derives from summing the fundamental descent inequality over T steps:

$$\gamma_s\mathbb{E}[F(\theta^{s-1}) - F(\theta^*)] \leq \frac{1}{2}\left(\mathbb{E}\left[|\theta^{s-1} - \theta^*|_2^2\right] - \mathbb{E}\left[|\theta^s - \theta^*|_2^2\right]\right) + \frac{1}{2}\gamma_s^2 B^2$$

For a constant step size $\gamma_s = \gamma$, the summation yields the bound for the uniformly averaged iterate $\bar{\theta}_T = \frac{1}{T}\sum_{s=1}^{T}\theta^{s-1}$:

$$\frac{1}{T}\sum_{s=1}^{T}\mathbb{E}\left[F(\theta^{s-1})\right] - F(\theta^*) \leq \frac{D^2}{2\gamma T} + \frac{\gamma B^2}{2}$$

.

By optimizing this upper bound with respect to γ, the optimal constant step size is found to be $\gamma^* = D/(B\sqrt{T})$. Substituting this optimal step size back into the expression minimizes the right-hand side, resulting in a cleaner convergence rate:

$$\mathbb{E}[F(\bar{\theta}_T) - F(\theta^*)] \leq \frac{DB}{\sqrt{T}}$$

By tuning the constant step size based on the known horizon T, we derive a pure $O(1/\sqrt{T})$ convergence rate.

This convergence rate allows us to compare SGD and GD wrt Empirical Risk Minimization problems defined over a finite dataset of n samples. For convex Lipschitz-continuous loss functions, the generalization error of the precise ERM solution is often bounded by $O(1/\sqrt{n})$, as established by complexity measures such as Rademacher bounds.

If SGD runs for $t = n$ iterations (corresponding to a single pass over the data), the resultant predictor $\bar{\theta}_n$ achieves an expected excess risk bounded by $O(DB/\sqrt{n})$, which matches the statistical generalization capacity of the ERM problem itself.

The initial assumption relies on almost sure boundedness of the stochastic gradient magnitude, $|g_t(\theta^{t-1})|_2^2 \leq B^2$. The analysis remains valid if the assumption is weakened to require the gradient to be bounded only in expected squared norm rather than almost surely:

$$\mathbb{E}[\|g_t(\theta^{t-1})\|_2^2] \leq B^2$$

Proofs

Proof for **Convergence of SGD** (Eq. 87:

We denote $D^2 = |\theta^0 - \theta_*|_2^2$ as the squared distance of the initial iterate to the optimum. Starting with:

$$\begin{aligned}
|\theta^t - \theta_*|_2^2 &= |\theta^{t-1} - \gamma_t g_t(\theta^{t-1}) - \theta_*|_2^2 \\
|a-b|_2^2 &= |a|_2^2 - 2a^\top b + |b|_2^2 \\
|\theta^t - \theta^*|_2^2 &= |\theta^{t-1} - \theta^*|_2^2 - 2\gamma_t g_t(\theta^{t-1})^\top(\theta^{t-1} - \theta^*) + \gamma_t^2 |g_t(\theta^{t-1})|_2^2
\end{aligned}$$

Next, we take the conditional expectation with respect to $\mathcal{F}_{t-1}$ (the history up to iteration $t-1$):

$$E[|\theta^t - \theta^*|_2^2|\mathcal{F}_{t-1}] = |\theta^{t-1} - \theta^*|_2^2 - 2\gamma_t E[g_t(\theta^{t-1})^\top(\theta^{t-1} - \theta^*)|\mathcal{F}_{t-1}] + \gamma_t^2 E[|g_t(\theta^{t-1})|_2^2|\mathcal{F}_{t-1}]$$

Applying $E[g_t(\theta^{t-1})|\mathcal{F}_{t-1}] = F'(\theta^{t-1})$, and $E[|g_t(\theta^{t-1})|_2^2|\mathcal{F}_{t-1}] \leq B^2$:

$$E[|\theta^t - \theta^*|_2^2|\mathcal{F}_{t-1}] \leq |\theta^{t-1} - \theta^*|_2^2 - 2\gamma_t F'(\theta^{t-1})^\top(\theta^{t-1} - \theta^*) + \gamma_t^2 B^2$$

The convexity of F implies:

$$F(\theta^{t-1}) - F(\theta^*) \leq F'(\theta^{t-1})^\top(\theta^{t-1} - \theta^*)$$

Substituting this convexity bound into expected squared distance relation, we isolate expected excess error, take the full expectation $E[\cdot]$ by the law of total expectation (Eq. 21) and again isolate expected excess error:

$$\begin{aligned}
E[|\theta^t - \theta^*|_2^2|\mathcal{F}_{t-1}] &\leq |\theta^{t-1} - \theta^*|_2^2 - 2\gamma_t[F(\theta^{t-1}) - F(\theta^*)] + \gamma_t^2 B^2 \\
E[|\theta^t - \theta^*|_2^2] &\leq E[|\theta^{t-1} - \theta^*|_2^2] - 2\gamma_t E[F(\theta^{t-1}) - F(\theta^*)] + \gamma_t^2 B^2 \\
\gamma_t E[F(\theta^{t-1}) - F(\theta^*)] &\leq \frac{1}{2}\left(E[|\theta^{t-1} - \theta^*|_2^2] - E[|\theta^t - \theta^*|_2^2]\right) + \frac{1}{2}\gamma_t^2 B^2
\end{aligned}$$

We sum this inequality over $s = 1$ to t (replacing t with s in the indices):

$$\sum_{s=1}^{t}\gamma_s E[F(\theta^{s-1}) - F(\theta^*)] \leq \frac{1}{2}\left(E[|\theta^0 - \theta^*|_2^2] - E[|\theta^t - \theta^*|_2^2]\right) + \frac{1}{2}B^2\sum_{s=1}^{t}\gamma_s^2$$

The first term on the right-hand side is a telescoping sum. Since $E[|\theta^t - \theta^*|_2^2] \geq 0$ and $|\theta^0 - \theta^*|_2^2 \leq D^2$:

$$\sum_{s=1}^{t} \gamma_s E[F(\theta^{s-1}) - F(\theta^*)] \leq \frac{1}{2}D^2 + \frac{1}{2}B^2 \sum_{s=1}^{t} \gamma_s^2$$

We utilize Jensen's inequality (Eq. 12)for F applied to the average iterate $\bar{\theta}^t = \frac{1}{\sum_{s=1}^{t} \gamma_s} \sum_{s=1}^{t} \gamma_s \theta^{s-1}$:

$$E[F(\bar{\theta}^t)] - F(\theta^*) \leq \frac{1}{\sum_{s=1}^{t} \gamma_s} \sum_{s=1}^{t} \gamma_s E[F(\theta^{s-1}) - F(\theta^*)]$$

Combining the bounds:

$$E[F(\bar{\theta}^t) - F(\theta^*)] \leq \frac{\frac{1}{2}D^2 + \frac{1}{2}B^2 \sum_{s=1}^{t} \gamma_s^2}{\sum_{s=1}^{t} \gamma_s}$$

To achieve the desired convergence rate, we select the constant step size $\gamma_s = \gamma$ for all s (based on t), following the successful constant step size choice from the subgradient method: Choosing $\gamma_s = \gamma = \frac{D}{B\sqrt{t}}$, we calculate the numerator and denominator:

$$\sum_{s=1}^{t} \gamma_s = t\gamma = \frac{D\sqrt{t}}{B}$$

$$\sum_{s=1}^{t} \gamma_s^2 = t\gamma^2 = t\frac{D^2}{B^2 t} = \frac{D^2}{B^2}$$

Substituting these back into the bound:

$$E[F(\bar{\theta}^t) - F(\theta^*)] \leq \frac{\frac{1}{2}D^2 + \frac{1}{2}B^2 \frac{D^2}{B^2}}{\frac{D\sqrt{t}}{B}} = \frac{D^2}{\frac{D\sqrt{t}}{B}} = \frac{DB}{\sqrt{t}}$$

Discussion

To achieve the $O(1/\sqrt{n})$ rate, Batch GD requires $O(n)$ iterations, with each iteration computing the full gradient over all n samples (complexity $O(nd)$ per iteration), resulting in a total computational complexity of $O(n^2 d)$. Single-pass SGD requires n iterations, with each iteration sampling only one data point (complexity $O(d)$ per iteration), resulting in a total complexity of $O(nd)$. Therefore, SGD achieves the optimal statistical rate with a computational complexity that is $O(n)$ times faster than batch GD.

Ch. 5.4.1 - SGD in Strongly Convex Problems

Derivations and Further Definitions

We assume the objective function $G(\theta) = F(\theta) + \frac{\mu}{2}|\theta|_2^2$, where F is convex, is minimized, and let θ^* be the unique minimizer. We assume the stochastic gradient $g_t(\theta^{t-1})$ satisfies the unbiased condition $E[g_t(\theta^{t-1})|\theta^{t-1}] = F'(\theta^{t-1})$ and the bounded variance condition $E[|g_t(\theta^{t-1})|_2^2] \leq B^2$ (or almost sure boundedness).
The SGD update rule for the strongly convex problem is given by incorporating the gradient of the quadratic regularization term $\frac{\mu}{2}|\theta|_2^2$:

$$\theta_t = \theta_{t-1} - \gamma_t[g_t(\theta^{t-1}) + \mu\theta^{t-1}]$$

We begin with the expected squared distance of the iterate θ_t from the optimum θ_*:

$$E[|\theta_t - \theta_*|_2^2] = E[|\theta_{t-1} - \gamma_t(g_t(\theta^{t-1}) + \mu\theta^{t-1}) - \theta^*|_2^2]$$

Expanding the square and taking the conditional expectation with respect to the history F_{t-1} (using the law of total expectation (Eq. 21) later for the full expectation), and applying the unbiased gradient assumption $E[g_t(\theta^{t-1})|\theta^{t-1}] = F'(\theta^{t-1})$, we arrive at the standard descent inequality structure (here applied to G'):

$$E[|\theta_t - \theta^*|_2^2|F_{t-1}] \leq |\theta_{t-1} - \theta^*|_2^2 - 2\gamma_t G'(\theta^{t-1})^\top(\theta^{t-1} - \theta^*) + \gamma_t^2 E[|g_t(\theta^{t-1}) + \mu\theta^{t-1}|_2^2|F_{t-1}]$$

.
For the strongly convex objective $G(\theta)$, we use strong convexity rewritten in terms of G':

$$G'(\theta^{t-1})^\top(\theta^{t-1}-\theta^*) \geq G(\theta^{t-1}) - G(\theta^*) + \frac{\mu}{2}|\theta^{t-1}-\theta^*|_2^2$$

.
Substituting this and incorporating the bound on $E[|g_t(\theta^{t-1}) + \mu\theta^{t-1}|_2^2|F_{t-1}] \leq 4B^2$ (which holds if g_t is bounded and θ_{t-1} is bounded, w/ overall term bounded by $4B^2$ as the update involves $\mu\theta_{t-1}$), and taking the full expectation, we obtain:

$$E[|\theta_t-\theta^*|_2^2] \leq E[|\theta_{t-1}-\theta^*|_2^2] - 2\gamma_t E[G(\theta^{t-1}) - G(\theta^*)] - \gamma_t\mu E[|\theta^{t-1}-\theta^*|_2^2] + 4\gamma_t^2 B^2$$

.
Rearranging for the expected excess risk term $E[G(\theta^{t-1}) - G(\theta^*)]$, we get:

$$\gamma_t E[G(\theta^{t-1}) - G(\theta^*)] \leq \frac{1}{2}\left((1-\gamma_t\mu)E[|\theta^{t-1}-\theta^*|_2^2] - E[|\theta^t-\theta^*|_2^2]\right) + 2\gamma_t^2 B^2$$

.
We select decaying step size $\gamma_t = 1/(\mu t)$. Substituting this choice:

$$\frac{1}{\mu t}E[G(\theta^{t-1}) - G(\theta^*)] \leq \frac{1}{2}\left(\left(1-\frac{1}{t}\right)E[|\theta^{t-1}-\theta^*|_2^2] - E[|\theta^t-\theta^*|_2^2]\right) + 2\frac{B^2}{\mu^2 t^2}$$

Multiplying by $2\mu t$:

$$E[G(\theta^{t-1}) - G(\theta^*)] \leq \mu t\left(\left(1-\frac{1}{t}\right)E[|\theta^{t-1}-\theta^*|_2^2] - E[|\theta^t-\theta^*|_2^2]\right) + 4\frac{B^2}{\mu t}$$

Take $\delta_t = E[|\theta^t-\theta^*|_2^2]$. Using this to rewrite the first term on the RHS, we get:

$$t\left(\frac{t-1}{t}\delta_{t-1} - \delta_t\right) = (t-1)\delta_{t-1} - t\delta_t$$

Summing the full inequality over $s=1$ to t yields:

$$\sum_{s=1}^{t} E[G(\theta^{s-1}) - G(\theta^*)] \leq \sum_{s=1}^{t}((s-1)\delta_{s-1} - s\delta_s) + \sum_{s=1}^{t}\frac{4B^2}{\mu s}$$

The first sum is a telescoping sum. With the assumptions that $\delta_0 = |\theta^0-\theta^*|_2^2$ is bounded and $t\delta_t$ is non-negative, the sum yields $0\cdot\delta_0 - t\delta_t \leq 0$.

$$\sum_{s=1}^{t} E[G(\theta^{s-1}) - G(\theta^*)] \leq 0 + \frac{4B^2}{\mu}\sum_{s=1}^{t}\frac{1}{s}$$

From the bound for the harmonic series $\sum_{s=1}^{t}\frac{1}{s} \leq 1+\log t$ and Jensen's inequality for $\bar{\theta}_t = \frac{1}{t}\sum_{s=1}^{t}\theta^{s-1}$ we can arrive at:

$$E[G(\bar{\theta}_t) - G(\theta^*)] \leq \frac{4B^2}{\mu t}(1+\log t)$$

This establishes the $O(\frac{\log(t)}{t})$ convergence rate for the uniformly averaged iterate of SGD on strongly convex problems.

Discussion

When we add Strong Convexity (μ) to the SGD setting, we fundamentally change the convergence behavior. With a decaying step size $\gamma_t = 1/(\mu t)$, the error decays at a rate of $O(\frac{\log t}{t})$. This is an improvement over the general convex rate of $O(1/\sqrt{t})$, and demonstrates that the curvature μ helps SGD "lock in" to the minimum more effectively, even in the presence of noise. However, the presence of noise variance B^2 means that we are fundamentally limited by the variance of our gradients and can never achieve the exponential convergence of deterministic gradient descent.

Ch. 5.4.3 - Bias-Variance Trade-offs for Least-Squares

Preliminary Definitions

Define a joint distribution on $(x, y) \in \mathcal{X} \times \mathbb{R}$, the linear model $y = \varphi(x)^\top \theta_* + \epsilon$ with feature map $\varphi : \mathcal{X} \to \mathbb{R}^d$, and optimal parameter vector $\theta_* \in \mathbb{R}^d$.
LMS Recursion: SGD update rule as applied to the quadratic loss

$$\mathcal{L}(y, \theta^\top \varphi(x)) = \frac{1}{2}(y - \theta^\top \varphi(x))^2 \tag{88}$$

Uniformly Averaged Iterate: The mean of all previous iterates

$$\bar{\theta}_t = \frac{1}{t} \sum_{s=1}^{t} \theta_{s-1} \tag{89}$$

Mahalanobis Norm: The squared norm of the inverse covariance matrix $\Sigma = \mathbb{E}[\varphi(x)\varphi(x)^\top]$, used to measure the distance of the initial bias, with $\Delta = (\eta_0 - \theta_*)$

$$\|\Delta\|^2_{\Sigma^{-1}} = \Delta^\top \Sigma^{-1} \Delta \tag{90}$$

Derivations and Further Definitions

Consider LMS recursion (Eq. 88) with constant step size γ: $\theta_t = \theta_{t-1} - \gamma(\theta_{t-1}^\top \varphi(x_t) - y_t)\varphi(x_t)$. If we assume $y_t = \varphi(x_t)^\top \theta_* + \epsilon_t$, we can subsequently decompose the error of the averaged iterate $\bar{\eta}_t$ into separate bias and variance components:
Bias Component: This captures 'forgetting' of initial condition η_0. Through expanding/unrolling LMS recursion with $\gamma \leq 1/R^2$ (where $\|\varphi(x)\|_2^2 \leq R^2$ almost surely), bias decays at an accelerated rate:

$$\|\bar{\eta}_t^{(bias)} - \theta_*\|^2_\Sigma \leq \frac{1}{\gamma^2 t^2} \|\eta_0 - \theta_*\|^2_{\Sigma^{-1}} \tag{91}$$

Variance Component: This captures the impact of σ^2 noise . The variance is bounded by:

$$\mathbb{E}[\|\bar{\eta}_t^{(var)} - \theta_*\|^2_\Sigma] \leq \frac{\sigma^2 d}{t} \tag{92}$$

Dimension-Free Extension: If we wish to extend the above to infinite-dimensional spaces, variance can be re-characterized by $tr[\Sigma] = \mathbb{E}[\|\varphi(x)\|_2^2]$.

Discussion

The analysis of bias-variance trade-offs in the domain of averaged SGD with quadratic losses indicates that the variance term $\frac{\sigma^2 d}{t}$ asymptotically approaches the efficiency of the batch OLS estimator. Unlike standard SGD, which converges to the noise-limited region $O(\gamma)$, use of the averaged iterate with constant step size allows for convergence to the exact global minimizer θ_*. Averaging enables the bias term to decay at $O(1/t^2)$ rate when the initial distance is measured in the Σ^{-1} norm. These results provide a basis for dimension-free guarantees, as the variance can be re-characterized using $tr[\Sigma]$, which depends on the properties of the feature map φ rather than the explicit dimension d.

Ch. 5.4.4 - Variance Reduction

Derivations and Further Definitions

Take an objective function $F(\theta)$ defined over a dataset of size n:

$$F(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\theta)$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ is the loss associated with the i-th data point. We assume that F is μ-strongly convex, each component function f_i is R^2-smooth, meaning its gradient is R^2-Lipschitz continuous, where $R^2 = \max_i L_i$ is the maximal smoothness constant of f_i, and the unique minimizer of $F(\theta)$ is η^*.

The **SAGA algorithm** is a variance reduction technique that employs a table of previously calculated individual gradients to maintain an unbiased yet low-variance stochastic gradient estimate. At iteration t, an index $i(t) \in 1, \ldots, n$ is selected uniformly at random, and the update for the parameter vector θ is given by:

$$\theta_t = \theta_{t-1} - \gamma \left[f'_{i(t)}(\theta_{t-1}) + \frac{1}{n} \sum_{i=1}^{n} z_i^{(t-1)} - z_{i(t)}^{(t-1)} \right] \tag{93}$$

where $z_i^{(t-1)}$ is the last stored gradient for the i-th function, and the table is updated by setting $z_{i(t)}^{(t)} = f'_{i(t)}(\theta_{t-1})$. Under strong convexity, this method achieves linear convergence to the optimum, with a rate that depends on the number of samples n and the condition number $\kappa = L/\mu$:

$$\mathbb{E}[F(\theta_t) - F(\eta^*)] \leq C \left(1 - \min\left\{\frac{1}{n}, \frac{\mu}{L}\right\}\right)^t \approx C \exp\left(-\min\left\{\frac{1}{n}, \frac{\mu}{L}\right\} t\right) \tag{94}$$

This explicitly bounds the number of iterations required for finite sums.
For finite sum problems that are convex but not strongly convex ($\mu = 0$), variance reduction methods (such as SAGA) achieve a convergence rate of $O(n/t)$, which is faster than the standard SGD rate of $O(1/\sqrt{t})$ for high-precision regimes.

Discussion

Here we addressed the previously discussed limitation of standard SGD from Ch. 5.4.1: the gradient noise prevents linear convergence. The SAGA (Eq. 93) algorithm resolves this by introducing a memory term (z_i) to reduce the variance of the stochastic gradient estimate as the algorithm progresses. By ensuring that the variance goes to zero as we approach the optimum, SAGA recovers the fast linear convergence rate (Eq. 93) characteristic of deterministic methods, without incurring the full cost of Batch GD. This represents modern state-of-the-art (SotA) algorithms that are statistically efficient (like SGD) but converge exponentially fast (like GD) by intelligently managing gradient information.

Ch. 5.5 - Conclusions on Convergence Rates

Summary Tables

In Table 1, we categorize convergence rates based on the geometry of the objective function (Convex vs. Strongly Convex), its regularity (Smooth vs. Nonsmooth), and the nature of the algorithm (Deterministic vs. Stochastic). In the tables below, L represents the smoothness constant, μ the strong convexity constant, and B the Lipschitz constant.

Table 1: Convergence Rates (Excess Error after t iterations)

	Convex	**Strongly Convex**
Non-smooth	Deterministic: $1/\sqrt{t}$ (Sec. 5.3) Stochastic: $1/\sqrt{t}$ (Sec. 5.4)	Deterministic: $B^2/(t\mu)$ (Eq. 82) Stochastic: $B^2/(t\mu)$ (Sec. 5.4.1)
Smooth	Deterministic: $1/t^2$ (Sec. 5.2.5) Stochastic: $1/\sqrt{t}$ (Eq. 87)	Deterministic: $\exp(-t\sqrt{\mu/L})$ (Sec. 5.2.5) Stochastic: $L/(t\mu)$ (Sec. 5.4.1)
Finite Sum	n/t (Sec. 5.4.4)	$\exp(-\min\{1/n, \mu/L\}t)$ (Eq. 94)

We can invert these formulas to determine the computational complexity with the number of gradient accesses t required to achieve an excess error of ϵ, shown in Table 2.

Discussion

The convergence rates in Tables 1 and 2 give us useful results on optimization.

The Acceleration of Strong Convexity. Strong convexity ($\mu > 0$) fundamentally changes convergence speeds. In the non-smooth regime, it improves the rate from $O(1/\sqrt{t})$ to $O(1/t)$. In the smooth

Table 2: Computational Complexity (Iterations t to reach error ϵ)

	Convex	**Strongly Convex**
Non-smooth	Deterministic: $1/\epsilon^2$ Stochastic: $1/\epsilon^2$	Deterministic: $B^2/(\epsilon\mu)$ Stochastic: $B^2/(\epsilon\mu)$
Smooth	Deterministic: $1/\sqrt{\epsilon}$ Stochastic: $1/\epsilon^2$	Deterministic: $\sqrt{L/\mu}\log(1/\epsilon)$ Stochastic: $L/(\epsilon\mu)$
Finite Sum	n/ϵ	$\max\{n, L/\mu\}\log(1/\epsilon)$

deterministic regime, it enables the transition from polynomial convergence ($O(1/t^2)$) to linear convergence ($\exp(-t\sqrt{\mu/L})$). This justifies the use of ℓ_2-regularization to induce favorable geometry for exponential optimization speed.

The Speed Limit of Non-smoothness. Non-smoothness imposes a hard ceiling on performance. Regardless of the algorithm, non-smooth convex problems are generally bound to $O(1/\sqrt{t})$ rates (complexity $O(1/\epsilon^2)$). Even with strong convexity, they cannot achieve the exponential convergence characteristic of smooth problems. This quantifies the computational cost of using non-differentiable objectives like Hinge loss or L1 regularization.

Deterministic vs. Stochastic Trade-off. While deterministic methods offer superior theoretical rates (e.g., $1/t^2$ vs. $1/\sqrt{t}$ for smooth convex problems), they incur an $O(n)$ cost per iteration. Stochastic methods, costing only $O(1)$ per iteration, are often computationally superior for moderate precision regimes where $\epsilon \approx 1/\sqrt{n}$.

Variance Reduction. The "Finite Sum" results show how variance reduction techniques (like SAGA) resolve this trade-off. By exploiting the sum structure, they achieve the exponential rates of deterministic methods ($\exp(-t)$) without the full $O(n)$ per-step cost, making them the standard for high-precision finite-sum optimization.

Ch. 14.4 - PAC Bayes

Preliminary Definitions

Shattering: Let $\mathcal{F}$ be a class of binary classifiers that maps input $\mathcal{X}$ to labels in $\{-1, 1\}$. $\mathcal{F}$ shatters $S \subset \mathcal{X}$ if every binary labeling of S is achievable by some $f \in \mathcal{F}$.

Vapnik–Chervonenkis (VC) dimension: The VC-dimension of the hypothesis class $\mathcal{F}$, $\mathbf{VC}(\mathcal{F})$, is the largest integer n s.t. $\exists$ a set of n points in $\mathcal{X}$ that can be shattered by $\mathcal{F}$. If sets of arbitrary size can be shattered, then $VC(\mathcal{F}) = \infty$.

Radon–Nikodym derivative: Let X be a measurable space that consists of set X and σ-algebra $\mathcal{M}_X$, and let μ and ν be measures on X (possibly infinite in $\mathbb{R}$, or finite in $\mathbb{C}$). Take measurable function f (with real or complex values) on X. f is a Radon–Nikodym (RN) derivative of μ w/r/t ν if, given any measurable $A \subset X$:

$$\mu(A) = \int_A f\nu = \int_{x\in A} f(x)d\nu(x) \tag{95}$$

(This exists if μ is absolutely continuous with respect to ν; any event with zero probability under ν also has zero probability under μ)

Kullback–Leibler (KL) divergence: Given measurable space Θ ($\theta \in \Theta$) and probability distributions ρ and q on Θ, if $\exists \frac{d\rho}{dq}$ (Eq. 95), KL-divergence $\mathbf{D}_{KL}$ is defined as:

$$D_{\mathrm{KL}}(\rho\|q) = \int_\Theta \log\left(\frac{d\rho}{dq}(\theta)\right) d\rho(\theta) \tag{96}$$

Donsker-Varadhan Formula: For any measurable function $h : \Theta \to \mathbb{R}$ s.t. $\int e^h dq < \infty$ where $\mathcal{P}(\Theta)$ is the set of all probability measures on Θ, the logarithm of the expectation w/r/t q satisfies:

$$\log \int_\Theta \exp(h(\theta))dq(\theta) = \sup_{\rho\in\mathcal{P}(\Theta)} \left[\int_\Theta h(\theta)d\rho(\theta) - D_{KL}(\rho\|q)\right] \tag{97}$$

Gibbs Posterior Distribution: For parameter $s > 0$, prior q, and empirical risk function $\hat{\mathcal{R}}$, the Gibbs posterior $\hat{\rho}_s$ is defined by:

$$\frac{d\hat{\rho}_s}{dq}(\theta) \propto \exp(-s\hat{\mathcal{R}}(\theta)) \tag{98}$$

Derivations and Further Definitions

Randomized vs. Aggregated Predictors: In the PAC-Bayes framework, we begin with a posterior probability distribution ρ over parameter space Θ. We first consider the **randomized predictor**, wherein for novel input x, we sample $\theta \sim \rho$ to predict $f_\theta(x)$ and take expected risk averaged over ρ:

$$\int_\Theta \mathcal{R}(\theta) d\rho(\theta) \tag{99}$$

For the **aggregated predictor** we make predictions via posterior mean $\bar{f}(x) = \int_\Theta f_\theta(x) d\rho(\theta)$.
Convexity and Jensen's Inequality: If we assume the loss $u \mapsto l(y, u)$ is convex for all y, we can apply Jensen's inequality to relate randomized and aggregated predictor risks. As $\mathcal{R}(f) = \mathbb{E}_{(X,Y)}[l(Y, f(X))]$, the convexity of l implies that $\mathcal{R}(\cdot)$ is convex w/r/t the predictor f. As such if the loss function $l(y, \cdot)$ in our prediction is convex, Jensen's inequality guarantees that the randomized predictor risk serves as a valid conservative bound for the aggregated predictor risk:

$$\mathcal{R}(\bar{f}) = \mathcal{R}\left(\mathbb{E}_{\theta\sim\rho}[f_\theta]\right) \leq \mathbb{E}_{\theta\sim\rho}[\mathcal{R}(f_\theta)] \tag{100}$$

Generalization Bound Derivation: Assume that a loss function is bounded almost surely s.t. $\forall\theta \in \Theta$, $l(y, f_\theta(x)) \in [0, l_\infty]$. For a fixed θ, the difference between expected risk $\mathcal{R}(\theta)$ and empirical risk $\hat{\mathcal{R}}(\theta)$ satisfies Hoeffding's inequality. Recall that for any $s \in \mathbb{R}_+$, the expectation over the selection of training data satisfies:

$$\mathbb{E}\left[\exp\left(s(\mathcal{R}(\theta) - \hat{\mathcal{R}}(\theta))\right)\right] \leq \exp\left(\frac{s^2 l_\infty^2}{8n}\right)$$

Integrating this over the prior q and applying the Donsker-Varadhan formula (Eq. 97) with $h(\theta) = s(\mathcal{R}(\theta) - \hat{\mathcal{R}}(\theta))$ allows us to bound the risk over the entire distribution $\mathcal{P}(\Theta)$:

$$\mathbb{E}\left[\exp\left(\sup_{\rho\in\mathcal{P}(\Theta)}\left(s\int_\Theta (\mathcal{R}(\theta) - \hat{\mathcal{R}}(\theta))d\rho(\theta) - D_{KL}(\rho\|q)\right)\right)\right] \leq \exp\left(\frac{s^2 l_\infty^2}{8n}\right)$$

Applying the Chernoff bound to the above yields the standard **PAC-Bayes generalization bound**. With probability at least $1-\delta$, $\forall\rho \in \mathcal{P}(\Theta)$:

$$\int_\Theta \mathcal{R}(\theta)d\rho(\theta) \leq \int_\Theta \hat{\mathcal{R}}(\theta)d\rho(\theta) + \frac{1}{s}D_{KL}(\rho\|q) + \frac{1}{s}\log\frac{1}{\delta} + \frac{sl_\infty^2}{8n} \tag{101}$$

Minimizing the RHS w/r/t ρ yields the optimal distribution, which is the Gibbs posterior $\hat{\rho}_s$ (Eq. 98). If we substitute $\hat{\rho}_s$ into the bound, we get:

$$\int_\Theta \mathcal{R}(\theta)d\hat{\rho}_s(\theta) \leq \inf_{\rho\in\mathcal{P}(\Theta)}\left\{\int_\Theta \hat{\mathcal{R}}(\theta)d\rho(\theta) + \frac{1}{s}D_{KL}(\rho\|q)\right\} + \frac{1}{s}\log\frac{1}{\delta} + \frac{sl_\infty^2}{8n} \tag{102}$$

If we choose the scaling parameter $s \propto \sqrt{n}$, we recover the standard generalization error of $O(1/\sqrt{n})$. Consider a finite hypothesis class with m predictors and a uniform prior $q(\theta) = 1/m$. If we restrict distributions to Dirac measures concentrated on single parameters, $D_{KL}(\rho\|q)$ becomes $\log(1/q(\theta)) = \log m$. By optimizing s, we recover the standard Empirical Risk Minimization bound:

$$\mathcal{R}(\theta) \leq \hat{\mathcal{R}}(\theta) + l_\infty\sqrt{\frac{\log m}{n}} \tag{103}$$

The above section contains derivations from both Bach [1] and Alquier [4].

Discussion

The PAC-Bayes structure gives a validation for Bayesian averaging while remaining agnostic on the correctness of the prior. The generalization bound (Eq. 101) balances empirical risk against KL-complexity with a confidence term; we minimize this bound to discourage overfitting. The scaling parameter s is fixed to $s = n$ in standard Bayesian inference, but setting $s \propto \sqrt{n}$ with PAC-BAYES recovers the standard $O(1/\sqrt{n})$ convergence rate, generally PAC-Bayes seems to require some optimization on s.
We can consider PAC-Bayes as a data-dependent alternative to Rademacher complexity or VC dimension. Standard learning theory (Rademacher/VC) controls the uniform deviation of the empirical risk from the true risk across the entire hypothesis class $\mathcal{F}$. As derived in Eq. 50, the bound is the supremum:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}(R(f) - \hat{R}(f))\right] \leq 2R_n(\mathcal{F})$$

This bound is a "worst-case" approach: $R_n(\mathcal{F})$ (Rademacher complexity) measures the ability of the full class $\mathcal{F}$ to fit random noise. This must hold even for the predictor that maximizes the error regardless of which predictor the algorithm selects. PAC-Bayes, in contrast, bounds the expected risk of a randomized predictor defined by posterior distribution ρ.

$$\int_\Theta \mathcal{R}(\theta)d\rho(\theta) \leq \int_\Theta \hat{\mathcal{R}}(\theta)d\rho(\theta) + \frac{1}{s}D_{KL}(\rho\|q) + \frac{1}{s}\log\frac{1}{\delta} + \frac{sl_\infty^2}{8n}$$

Here, the bound does not need to hold uniformly for all $\theta \in \Theta$, but only for the specific ρ chosen.
The specific advantage of PAC-Bayes lies in the complexity term. For Rademacher/VC, complexity $R_n(\mathcal{F})$ or $VC(\mathcal{F})$ is a property of the hypothesis class architecture. For PAC-Bayes, the complexity term is the KL-divergence $D_{KL}(\rho\|q)$. This is fully data-dependent because the posterior ρ is optimized after observing data. This allows PAC-Bayes to provide useful bounds in high-dimensional settings such as over-parameterized neural networks, where the VC dimension would be essentially infinite. Even if the class is extremely large, if the learning algorithm finds a "simple" solution (e.g. one close to the prior q), the generalization error remains tightly bounded. Finally, PAC-Bayes integrates optimization directly into the bound. Minimizing PAC-Bayes bounds explicitly yields the Gibbs posterior $\hat{\rho}_s \propto \exp(-s\hat{R}(\theta))$. This creates a theoretical link between the learning algorithm's objective (Empirical Risk Minimization) and the statistical guarantee, whereas Rademacher analysis separates the capacity of the class from the method used to search over it.

Appendix

References: Unless otherwise noted, the majority of proofs and derivations in this book take inspiration or direct guidance from either Bach's text [1] or the Wikipedia article on the relevant subject. That said, all derivations and proofs were worked through step by step by the author with a diligent effort to define all relevant techniques and theorems and to build a coherent mathematical narrative from start to finish.
Uses of Generative A.I. Tools: Anthropic Claude Haiku 4.5 was used to translate some equations into LaTeX as well as general assistance with formatting in this document. Gemini Flash 3.0 was used to check for errors and omissions in refining draft versions of the worksheet.

References

[1] F. Bach. *Learning Theory from First Principles*. Adaptive Computation and Machine Learning series. MIT Press, 2024. ISBN: 9780262381369 (cit. on pp. 1, 9, 34, 35).

[2] Yurii Nesterov. "A method for solving the convex programming problem with convergence rate $O(1/k^2)$". In: *Proceedings of the USSR Academy of Sciences* 269 (1983), pp. 543–547. URL: https://api.semanticscholar.org/CorpusID:145918791 (cit. on p. 24).

[3] Alekh Agarwal et al. *Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization*. 2011. arXiv: 1009.0571 [stat.ML]. URL: https://arxiv.org/abs/1009.0571 (cit. on p. 27).

[4] Pierre Alquier. “User-friendly Introduction to PAC-Bayes Bounds”. In: *Foundations and Trends in Machine Learning* 17.2 (Jan. 2024), pp. 174–303. ISSN: 1935-8245. DOI: 10.1561/2200000100. URL: http://dx.doi.org/10.1561/2200000100 (cit. on p. 34).