Chapter 1

Supervised Learning Formulations

In this chapter, we will set up the standard theoretical formulation of supervised learning and introduce the *empirical risk minimization* (ERM) paradigm. The setup will apply to almost the entire monograph and the ERM paradigm will be the main focus of Chapter 2, 3, and 4.

1.1 Supervised learning

In supervised learning, we have a dataset where each data point is associated with a label, and we aim to learn from the data a function that maps data points to their labels. The learned function can be used to infer the labels of test data points. More formally, suppose the data points, also called inputs, belong to some input space $\mathcal{X}$ (e.g. images of birds), and labels belong to the output space $\mathcal{Y}$ (e.g. bird species). Suppose we are interested in a specific joint probability distribution P over $\mathcal{X} \times \mathcal{Y}$ (e.g. images of birds in North America), from which we draw a *training set*, i.e we draw a a set of n independent and identically distributed (i.i.d.) data points $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ from P. The goal of supervised learning is to learn a mapping (i.e. a function) from $\mathcal{X}$ to $\mathcal{Y}$ using the training data. Any such function $h : \mathcal{X} \to \mathcal{Y}$ is called a *predictor* (also *hypothesis* or *model*).

Given two predictors, how do we decide which is better? For that, we define a *loss function* over the predictions. There are several ways to define loss functions: for now, define a loss function ℓ as a function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Intuitively, the loss function takes two labels, the prediction made by a model $\hat{y}$ and the true label y, and gives a number that captures how different the two labels are. We assume ℓ is non-negative, i.e $\ell(\hat{y}, y) \geq 0$. Then, the loss of a model h on an example (x, y) is $\ell(h(x), y)$, i.e. the difference (as measured by ℓ) between the prediction made by h and the true label.

With these definitions, we are able to formalize the problem of supervised learning. Precisely, we seek to find a model h that minimizes what we call the expected loss (or population loss or expected risk or population risk):

$$L(h) \triangleq \mathop{\mathbb{E}}_{(x,y)\sim p}[\ell(h(x), y)]. \tag{1.1}$$

Note that L is nonnegative because ℓ is nonnegative. Typically, the loss function is designed so that the best possible loss is zero when $\hat{y}$ matches y exactly. Therefore, the goal is to find h such that $L(h)$ is as close to zero as possible.

Examples: regression and classification problems. Here are two standard types of supervised learning problems based on the properties of the output space:

- In the problem of *regression*, predictions are real numbers ($\mathcal{Y} = \mathbb{R}$). We would like predictions to be as close as possible to the real labels. A classical loss function that captures this is the squared error, $\ell(\hat{y}, y) = (\hat{y} - y)^2$.

- In the problem of *classification*, predictions are in a discrete set of k unordered classes $\mathcal{Y} = [k] = \{1, \cdots, k\}$. One possible classification loss is the $0-1$ loss: $\ell(\hat{y}, y) = \mathbb{1}(\hat{y} \neq y)$, i.e. 0 if the prediction is equal to the true label, and 1 otherwise.

Hypothesis class. So far, we said we would like to find *any function* that minimizes population risk. However, in practice, we do not have a way of optimizing over arbitrary functions. Instead, we work within a more constrained set of functions $\mathcal{H}$, which we call the *hypothesis family* (or *hypothesis class*). Each element of $\mathcal{H}$ is a function $h : \mathcal{X} \to \mathcal{Y}$. Usually, we choose a set $\mathcal{H}$ that we know how to optimize over (e.g. linear functions, or neural networks).

Given one particular function $h \in \mathcal{H}$, we define the *excess risk* of h with respect to $\mathcal{H}$ as the difference between the population risk of h and the best possible population risk inside $\mathcal{H}$:

$$E(h) \triangleq L(h) - \inf_{g \in \mathcal{H}} L(g).$$

Generally we need more assumptions about a specific problem and hypothesis class to bound absolute population risk, hence we focus on bounding the excess risk.

Usually, the family we choose to work with can be parameterized by a vector of parameters $\theta \in \Theta$. In that case, we can refer to an element of $\mathcal{H}$ by h_θ, making that explicit. An example of such a parametrization of the hypothesis class is $\mathcal{H} = \{h : h_\theta(x) = \theta^\top x, \theta \in \mathbb{R}^d\}$.

1.2 Empirical risk minimization

Our ultimate goal is to minimize population risk. However, in practice we do not have access to the entire population: we only have a *training set* of n data points, drawn from the same distribution as the entire population. While we cannot compute population risk, we can compute *empirical risk*, the loss over the training set, and try to minimize that. This is, in short, the paradigm known as *empirical risk minimization* (ERM): we optimize the training set loss, with the hope that this leads us to a model that has low population loss. From now on, with some abuse of notation, we often write $\ell(h_\theta(x), y)$ as $\ell((x, y), \theta)$ and use the two notations interchangeably. Formally, we define the empirical risk of a model h as:

$$\widehat{L}(h_\theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{n} \sum_{i=1}^{n} \ell((x^{(i)}, y^{(i)}), \theta). \tag{1.2}$$

Empirical risk minimization is the method of finding the minimizer of $\widehat{L}$, which we call $\hat{\theta}$:

$$\hat{\theta} \triangleq \operatorname*{argmin}_{\theta \in \Theta} \widehat{L}(h_\theta). \tag{1.3}$$

Since we are assuming that our training examples are drawn from the same distribution as the whole population, we know that empirical risk and population risk are equal *in expectation* (over the randomness of the training dataset):

$$\begin{aligned}
\mathop{\mathbb{E}}_{(x^{(i)}, y^{(i)}) \overset{\text{iid}}{\sim} P} \widehat{L}(h_\theta) &= \mathop{\mathbb{E}}_{(x^{(i)}, y^{(i)}) \overset{\text{iid}}{\sim} P} \frac{1}{n} \sum_{i=1}^{n} \ell(h_\theta(x^{(i)}), y^{(i)}) && (1.4) \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{(x^{(i)}, y^{(i)}) \overset{\text{iid}}{\sim} P} \ell(h_\theta(x^{(i)}), y^{(i)}) && (1.5) \\
&= \frac{1}{n} \cdot n \cdot \mathop{\mathbb{E}}_{(x^{(i)}, y^{(i)}) \overset{\text{iid}}{\sim} P} \ell(h_\theta(x^{(i)}), y^{(i)}) && (1.6) \\
&= L(h_\theta). && (1.7)
\end{aligned}$$

This is one reason why it makes sense to use empirical risk: it is an unbiased estimator of the population risk.

The key question that we seek to answer in the first part of this course is: **what guarantees do we have on the excess risk for the parameters learned by ERM?** The hope with ERM is that minimizing the training error will lead to small testing error. One way to make this rigorous is by showing that the ERM minimizer's excess risk is bounded.

Chapter 2

Asymptotic Analysis

In this chapter, we use an asymptotic approach (i.e. assuming number of training samples $n \to \infty$) to achieve a bound on the ERM. We then instantiate these results to the case where the loss function is the maximum likelihood and discuss the limitations of asymptotics. (In future chapters we will assume finite n and provide a non-asymptotic analysis.)

2.1 Asymptotics of empirical risk minimization

For the asymptotic analysis of ERM, we would like to prove that excess risk is bounded as shown below:

$$L(\hat{\theta}) - \inf_{\theta \in \Theta} L(\theta) \leq \frac{c}{n} + o\left(\frac{1}{n}\right). \tag{2.1}$$

Here c is a problem dependent constant that does not depend on n, and $o(1/n)$ hides all dependencies except n. The equation above shows that as we have more training data (i.e. as n increases) the excess risk of ERM decreases at the rate of $\frac{1}{n}$.

Let $\{(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})\}$ be the training data and let $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}^p\}$ be the parameterized family of hypothesis functions. Let the ERM minimizer be $\hat{\theta}$ as defined in Equation (1.3). Let θ^* be the minimizer of the population risk L, i.e. $\theta^* = \operatorname{argmin}_\theta L(\theta)$. The theorem below quantifies the excess risk $L(\hat{\theta}) - L(\theta^*)$:

Theorem 2.1 (Informally stated)**.** *Suppose that (a) $\hat{\theta} \xrightarrow{p} \theta^*$ as $n \to \infty$ (i.e. consistency of $\hat{\theta}$), (b) $\nabla^2 L(\theta^*)$ is full rank, and (c) other appropriate regularity conditions hold.*[1] *Then,*

1. $\sqrt{n}(\hat{\theta} - \theta^*) = O_P(1)$, *i.e. for every $\epsilon > 0$, there is an M such that* $\sup_n \mathbb{P}(\|\sqrt{n}(\hat{\theta} - \theta^*)\|_2 > M) < \epsilon$. *(This means that the sequence $\{\sqrt{n}(\hat{\theta} - \theta^*)\}$ is "bounded in probability".)*

2. $\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, (\nabla^2 L(\theta^*))^{-1} \operatorname{Cov}(\nabla \ell((x, y), \theta^*))(\nabla^2 L(\theta^*))^{-1}\right)$.

3. $n(L(\hat{\theta}) - L(\theta^*)) = O_P(1)$.

4. $n(L(\hat{\theta}) - L(\theta^*)) \xrightarrow{d} \frac{1}{2}||S||_2^2$ *where* $S \sim \mathcal{N}\left(0, (\nabla^2 L(\theta^*))^{-1/2} \operatorname{Cov}(\nabla \ell((x, y), \theta^*))(\nabla^2 L(\theta^*))^{-1/2}\right)$.

5. $\lim_{n \to \infty} \mathbb{E}\left[n(L(\hat{\theta}) - L(\theta^*))\right] = \frac{1}{2} \operatorname{tr}\left(\nabla^2 L(\theta^*)^{-1} \operatorname{Cov}(\nabla \ell((x, y), \theta^*))\right)$.

[1] $X_n \xrightarrow{p} X$ implies that for all $\epsilon > 0$, $\mathbb{P}\left(\|X_n - X\| > \epsilon\right) \to 0$, while $X_n \xrightarrow{d} X$ implies that $\mathbb{P}(X_n \leq t) \to \mathbb{P}(X \leq t)$ at all points t for which $\mathbb{P}(X \leq t)$ is continuous. These two notions of convergence are known as convergence in probability and convergence in distribution, respectively. These concepts are not essential to this course, but additional information can be found by reading the Wikipedia article on convergence of random variables.

Remark: In the theorem above, Parts 1 and 3 only show the rate or order of convergence, while Parts 2 and 4 define the limiting distribution for the random variables.

Theorem 2.1 is a powerful conclusion because once we know that $\sqrt{n}(\hat{\theta}-\theta^*)$ is (asymptotically) Gaussian, we can easily work out the distribution of the excess risk. If we believe in our assumptions and n is large enough such that we can assume $n \to \infty$, this allows us to analytically determine quantities of interest in almost any scenario (for example, if our test distribution changes). The key takeaway is that our parameter error $\hat{\theta}-\theta^*$ decreases in order $1/\sqrt{n}$ and the excess risk decreases in order $1/n$. While we will not discuss the regularity assumptions in Theorem 2.1 in great detail, we note that the assumption that L is twice differentiable is crucial.

2.1.1 Key ideas of proofs

We will prove the theorem above by applying the following main ideas:

1. Obtain an expression for the excess risk by Taylor expansion of the derivative of the empirical risk $\nabla\widehat{L}(\theta)$ around θ^*.

2. By the law of large numbers, we have that $\widehat{L}(\theta) \xrightarrow{p} L(\theta)$, $\nabla\widehat{L}(\theta) \xrightarrow{p} \nabla L(\theta)$ and $\nabla^2\widehat{L}(\theta) \xrightarrow{p} \nabla^2 L(\theta)$ as $n \to \infty$.

3. Central limit theorem (CLT).

First, we state the CLT for i.i.d. means and a lemma that we will use in the proof.

Theorem 2.2 (Central Limit Theorem)**.** *Let $X_1, \cdots, X_n$, be i.i.d. random variables, where $\widehat{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and the covariance matrix Σ is finite. Then, as $n \to \infty$ we have*

1. *$\widehat{X} \xrightarrow{p} \mathbb{E}[X]$, and*

2. *$\sqrt{n}(\widehat{X}-\mathbb{E}[X]) \xrightarrow{d} \mathcal{N}(0,\Sigma)$. In particular, $\sqrt{n}(\widehat{X}-\mathbb{E}[X]) = O_P(1)$.*

Lemma 2.3.

1. *If $Z \sim N(0,\Sigma)$ and A is a deterministic matrix, then $AZ \sim N(0, A\Sigma A^\top)$.*

2. *If $Z \sim N(0,\Sigma^{-1})$ and $Z \in \mathbb{R}^p$, then $Z^\top \Sigma Z \sim \chi^2(p)$, where $\sim \chi^2(p)$ is the chi-squared distribution with p degrees of freedom.*

2.1.2 Main proof

Let us start with heuristic arguments for Parts 1 and 2. First, note that by definition, the gradient of the empirical risk at the empirical risk minimizer, $\nabla\widehat{L}(\hat{\theta})$, is equal to 0. From the Taylor expansion of $\nabla\widehat{L}$ around θ^*, we have that

$$0 = \nabla\widehat{L}(\hat{\theta}) = \nabla\widehat{L}(\theta^*) + \nabla^2\widehat{L}(\theta^*)(\hat{\theta}-\theta^*) + O(\|\hat{\theta}-\theta^*\|_2^2). \tag{2.2}$$

Rearranging, we have

$$\hat{\theta}-\theta^* = -(\nabla^2\widehat{L}(\theta^*))^{-1}\nabla\widehat{L}(\theta^*) + O(||\hat{\theta}-\theta^*||_2^2). \tag{2.3}$$

Multiplying by $\sqrt{n}$ on both sides,

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\theta^*) &= -(\nabla^2\widehat{L}(\theta^*))^{-1}\sqrt{n}(\nabla\widehat{L}(\theta^*)) + O(\sqrt{n}||\hat{\theta}-\theta^*||_2^2) &(2.4)\\ &\approx -(\nabla^2\widehat{L}(\theta^*))^{-1}\sqrt{n}(\nabla\widehat{L}(\theta^*)). &(2.5)\end{aligned}$$

Applying the Central Limit Theorem (Theorem 2.2) using $X_i = \nabla \ell((x^{(i)}, y^{(i)}), \theta^*)$ and $\widehat{X} = \nabla \widehat{L}(\theta^*)$, and noticing that $\mathbb{E}[\nabla \widehat{L}(\theta^*)] = \nabla L(\theta^*)$, we have

$$\sqrt{n}(\nabla \widehat{L}(\theta^*) - \nabla L(\theta^*)) \xrightarrow{d} \mathcal{N}(0, \operatorname{Cov}(\nabla \ell((x, y), \theta^*))). \tag{2.6}$$

Note that $\nabla L(\theta^*) = 0$ because θ^* is the minimizer of L, so $\sqrt{n}(\nabla \widehat{L}(\theta^*)) \xrightarrow{d} \mathcal{N}(0, \operatorname{Cov}(\nabla \ell((x, y), \theta^*)))$. By the law of large numbers, $\nabla^2 \widehat{L}(\theta^*) \xrightarrow{p} \nabla^2 L(\theta^*)$. Applying these results to (2.5) (together with an application of Slutsky's theorem),

$$\begin{align}
\sqrt{n}(\hat{\theta} - \theta^*) &\xrightarrow{d} \nabla^2 L(\theta^*)^{-1} \mathcal{N}(0, \operatorname{Cov}(\nabla \ell((x, y), \theta^*))) \tag{2.7}\\
&\overset{d}{=} \mathcal{N}\left(0, \nabla^2 L(\theta^*)^{-1} \operatorname{Cov}(\nabla \ell((x, y), \theta^*)) \nabla^2 L(\theta^*)^{-1}\right), \tag{2.8}
\end{align}$$

where the second step is due to Lemma 2.3. This proves Part 2 of Theorem 2.1.

Part 1 follows directly from Part 2 by the following fact: If $X_n \xrightarrow{d} P$ for some probability distribution P, then $X_n = O_P(1)$.

We now turn to proving Parts 3 and 4. Using a Taylor expansion of L with respect to θ at θ^*, we find

$$L(\hat{\theta}) = L(\theta^*) + \langle \nabla L(\theta^*), \hat{\theta} - \theta^* \rangle + \frac{1}{2} \langle \hat{\theta} - \theta^*, \nabla^2 L(\theta^*)(\hat{\theta} - \theta^*) \rangle + o(\|\hat{\theta} - \theta^*\|_2^2). \tag{2.9}$$

Since θ^* is the minimizer of the population risk L, we know that $\nabla L(\theta^*) = 0$ and the linear term is equal to 0. Rearranging and multiplying by n, we can write

$$\begin{align}
n(L(\hat{\theta}) - L(\theta^*)) &= \frac{n}{2} \langle \hat{\theta} - \theta^*, \nabla^2 L(\theta^*)(\hat{\theta} - \theta^*) \rangle + o(\|\hat{\theta} - \theta^*\|_2^2) \tag{2.10}\\
&\approx \frac{1}{2} \langle \sqrt{n}(\hat{\theta} - \theta^*), \nabla^2 L(\theta^*) \sqrt{n}(\hat{\theta} - \theta^*) \rangle \tag{2.11}\\
&= \frac{1}{2} \left\| \nabla^2 L(\theta^*)^{1/2} \sqrt{n}(\hat{\theta} - \theta^*) \right\|_2^2, \tag{2.12}
\end{align}$$

where the last equality follows from the fact that for any vector v and positive semi-definite matrix A of appropriate dimensions, the inner product $\langle v, Av \rangle = v^\top A v = \|A^{1/2} v\|_2^2$. Let $S = \nabla^2 L(\theta^*)^{1/2} \sqrt{n}(\hat{\theta} - \theta^*)$, i.e. the random vector inside the norm. By Part 2, we know the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta^*)$ is Gaussian. Thus as $n \to \infty$, $n(L(\hat{\theta}) - L(\theta^*)) \xrightarrow{d} \frac{1}{2}\|S\|_2^2$ where

$$\begin{align}
S &\sim \nabla^2 L(\theta^*)^{1/2} \cdot \mathcal{N}\left(0, \nabla^2 L(\theta^*)^{-1} \operatorname{Cov}(\nabla \ell((x, y), \theta^*)) \nabla^2 L(\theta^*)^{-1}\right) \tag{2.13}\\
&\overset{d}{=} \mathcal{N}\left(0, \nabla^2 L(\theta^*)^{-1/2} \operatorname{Cov}(\nabla \ell((x, y), \theta^*)) \nabla^2 L(\theta^*)^{-1/2}\right). \tag{2.14}
\end{align}$$

This proves Part 4, and Part 3 follows directly from the definition of the O_P notation.

Finally, for Part 5, using the fact that the trace operator is invariant under cyclic permutations, the fact that $\mathbb{E}[S] = 0$, and some regularity conditions,

$$\begin{align}
\lim_{n \to \infty} \mathbb{E}\left[n(L(\hat{\theta}) - L(\theta^*))\right] &= \frac{1}{2} \mathbb{E}\left[\|S\|_2^2\right] = \frac{1}{2} \mathbb{E}\left[\operatorname{tr}(S^\top S)\right] \tag{2.15}\\
&= \frac{1}{2} \mathbb{E}\left[\operatorname{tr}(S S^\top)\right] = \frac{1}{2} \operatorname{tr}\left(\mathbb{E}[S S^\top]\right) \tag{2.16}\\
&= \frac{1}{2} \operatorname{tr}\left(\operatorname{Cov}(S)\right) \tag{2.17}\\
&= \frac{1}{2} \operatorname{tr}\left(\nabla^2 L(\theta^*)^{-1} \operatorname{Cov}(\nabla \ell((x, y), \theta^*))\right). \tag{2.18}
\end{align}$$

2.1.3 Well-specified case

Theorem 2.1 is powerful because it is general, avoiding any assumptions of a probabilistic model of our data. However in many applications, we assume a model of our data and we define the log-likelihood with respect to this model. Formally, suppose that we have a family of probability distributions P_θ, parameterized by $\theta \in \Theta$, such that P_{θ_*} is the true data-generating distribution. This is known as the well-specified case. To make the results of Theorem 2.1 more applicable, we derive analogous results for this well-specified case in Theorem 2.4.

Theorem 2.4. *In addition to the assumptions of Theorem 2.1, suppose there exists a parametric model $P(y \mid x; \theta)$, $\theta \in \Theta$, such that $\{y^{(i)} \mid x^{(i)}\}_{i=1}^n \sim P(y^{(i)} \mid x^{(i)}; \theta_*)$ for some $\theta_* \in \Theta$. Assume that we performing maximum likelihood estimation (MLE), i.e. our loss function is the negative log-likelihood $\ell((x^{(i)}, y^{(i)}), \theta) = -\log P(y^{(i)} \mid x^{(i)}; \theta)$. As before, let $\hat{\theta}$ and θ^* denote the minimizers of empirical risk and population risk, respectively. Then*

$$\theta^* = \theta_*, \tag{2.19}$$

$$\mathbb{E}\left[\nabla \ell((x, y), \theta^*)\right] = 0, \tag{2.20}$$

$$\operatorname{Cov}\left(\nabla \ell((x, y), \theta^*)\right) = \nabla^2 L(\theta^*), \text{ and} \tag{2.21}$$

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \nabla^2 L(\theta^*)^{-1}). \tag{2.22}$$

Remark 1: You may also have seen (2.22) in the following form: under the maximum likelihood estimation (MLE) paradigm, the MLE is asymptotically efficient as it achieves the Cramer-Rao lower bound. That is, the parameter error of the MLE estimate converges in distribution to $\mathcal{N}(0, \mathcal{I}(\theta)^{-1})$, where $\mathcal{I}(\theta)$ is the Fisher information matrix (in this case, equivalent to the risk Hessian $\nabla^2 L(\theta^*)$) [Rice, 2006].

Remark 2: (2.21) is also known as Bartlett's identity [Liang, 2016].

Although the proofs were not presented in live lecture, we include them here.

Proof. From the definition of the population loss,

$$\begin{aligned}
L(\theta) &= \mathbb{E}\left[\ell((x^{(i)}, y^{(i)}), \theta)\right] & (2.23)\\
&= \mathbb{E}\left[-\log P(y \mid x; \theta)\right] & (2.24)\\
&= \mathbb{E}\left[-\log P(y \mid x; \theta) + \log P(y \mid x; \theta_*)\right] + \mathbb{E}\left[-\log P(y \mid x; \theta_*)\right] & (2.25)\\
&= \mathbb{E}\left[\log \frac{P(y \mid x; \theta_*)}{P(y \mid x; \theta)}\right] + \mathbb{E}\left[-\log P(y \mid x; \theta_*)\right]. & (2.26)
\end{aligned}$$

Notice that the second term is a constant which we will express as $\mathcal{H}(y \mid x; \theta_*)$. We expand the first term using the tower rule (or law of total expectation):

$$L(\theta) = \mathbb{E}\left[\mathbb{E}\left[\log \frac{P(y \mid x; \theta_*)}{P(y \mid x; \theta)} \middle| x\right]\right] + \mathcal{H}(y \mid x; \theta_*). \tag{2.27}$$

The term in the expectation is just the KL divergence between the two probabilities, so

$$\begin{aligned}
L(\theta) &= \mathbb{E}\left[\mathrm{KL}\left(y \mid x; \theta_* \| y \mid x; \theta\right)\right] + \mathcal{H}(y \mid x; \theta_*) & (2.28)\\
&\geq \mathcal{H}(y \mid x; \theta_*), & (2.29)
\end{aligned}$$

since KL divergence is always non-negative. Since θ_* makes the KL divergence term 0, it minimizes $L(\theta)$ and so $\theta_* \in \operatorname{argmin}_\theta L(\theta)$. However, the minimizer of $L(\theta)$ is unique because of consistency, so we must have $\operatorname{argmin}_\theta L(\theta) = \theta^*$ which proves (2.19).

For (2.20), recall $\nabla L(\theta^*) = 0$, so we have

$$0 = \nabla L(\theta^*) = \nabla \mathbb{E}\left[\ell((x^{(i)}, y^{(i)}), \theta^*)\right] = \mathbb{E}\left[\nabla \ell((x^{(i)}, y^{(i)}), \theta^*)\right], \tag{2.30}$$

where we can switch the gradient and expectation under some regularity conditions.

To prove (2.21), we first expand the RHS using the definition of covariance and express the marginal distributions as integrals:

$$\begin{aligned}\operatorname{Cov}\left(\nabla \ell((x,y),\theta^*)\right) &= \mathbb{E}\left[\nabla \ell((x,y),\theta^*)\nabla \ell((x,y),\theta^*)^\top\right] &(2.31)\\ &= \int P(x)\left(\int P(y \mid x;\theta^*)\nabla \log P(y^{(i)} \mid x^{(i)};\theta^*)\nabla \log P(y^{(i)} \mid x^{(i)};\theta^*)^\top dy\right)dx &(2.32)\\ &= \int P(x)\left(\int \frac{\nabla P(y \mid x;\theta^*)\nabla P(y \mid x;\theta^*)^\top}{P(y \mid x;\theta^*)}dy\right)dx. &(2.33)\end{aligned}$$

Now we expand the LHS using the definition of the population loss and differentiate repeatedly:

$$\begin{aligned}\nabla^2 L(\theta^*) &= \mathbb{E}\left[-\nabla^2 \log P(y \mid x;\theta^*)\right] &(2.34)\\ &= \int P(x)\left(\int -\nabla^2 P(y \mid x;\theta^*) + \frac{\nabla P(y \mid x;\theta^*)\nabla P(y \mid x;\theta^*)^\top}{P(y \mid x;\theta^*)}dy\right)dx. &(2.35)\end{aligned}$$

Note that we can express

$$\int \nabla^2 P(y \mid x;\theta^*)dy = \nabla^2 \int P(y \mid x;\theta^*)dy = \nabla 1 = 0 \tag{2.36}$$

so we find

$$\nabla^2 L(\theta^*) = \int P(x)\left(\int \frac{\nabla P(y \mid x;\theta^*)\nabla P(y \mid x;\theta^*)^\top}{P(y \mid x;\theta^*)}dy\right)dx = \operatorname{Cov}\left(\nabla \ell((x,y),\theta^*)\right). \tag{2.37}$$

Finally, (2.22) follows directly from Part 2 of Theorem 2.1 and (2.21). □

Using similar logic to our proof of Part 4 and 5 of Theorem 2.1, we can see that $n(L(\hat{\theta}) - L(\theta^*)) \xrightarrow{d} \frac{1}{2}\|S\|_2^2$ where $S \sim N(0, I)$. Since a chi-squared distribution with p degrees of freedom is defined as a sum of the squares of p independent standard normals, it quickly follows that $2n(L(\hat{\theta}) - L(\theta^*)) \sim \chi^2(p)$, where $\theta \in \mathbb{R}^p$ and $n \to \infty$. We can thus characterize the excess risk in this case using the properties of a chi-squared distribution:

$$\lim_{n\to\infty} \mathbb{E}\left[L(\hat{\theta}) - L(\theta^*)\right] = \frac{p}{2n}. \tag{2.38}$$

2.2 Limitations of asymptotic analysis

One limitation of asymptotic analysis is that our bounds often obscure dependencies on higher order terms. As an example, suppose we have a bound of the form

$$\frac{p}{2n} + o\left(\frac{1}{n}\right). \tag{2.39}$$

(Here $o(\cdot)$ treats the parameter p as a constant as n goes to infinity.) We have no idea how large n needs to be for asymptotic bounds to be "reasonable." Compare two possible versions of (2.39):

$$\frac{p}{2n} + \frac{1}{n^2} \quad \text{vs.} \quad \frac{p}{2n} + \frac{p^{100}}{n^2}. \tag{2.40}$$

Asymptotic analysis treats both of these bounds as the same, hiding the polynomial dependence on p in the second bound. Clearly, the second bound is significantly more data-intensive than the first: we would need $n > p^{50}$ for $\frac{p^{100}}{n^2}$ to be less than one. Since p represents the dimensionality of the data, this may be an unreasonable assumption.

This is where non-asymptotic analysis can be helpful. Whereas asymptotic analysis uses large-sample theorems such as the central limit theorem and the law of large numbers to provide convergence guarantees, non-asymptotic analysis relies on concentration inequalities to develop alternative techniques for reasoning about the performance of learning algorithms.