CS229M/STATS214: Machine Learning Theory

Lecturer: Tengyu Ma Lecture # 2
Scribe: Alexander Ke and Trenton Chang Jan 13th, 2021

2.1 Review and overview

In this lecture, we finish the proof on the asymptotic behavior of the excess risk, showing that it is asymptotically bounded in probability by $1/n$ (Section 2.2). We then extend these results to the maximum likelihood estimation case and discuss the limitations of asymptotics. Motivated by these limitations, we introduce non-asymptotic analysis and apply it to uniform convergence.

Over the next few weeks, we will focus on how to bound the excess risk $L(\hat{\theta}) - L(\theta^*)$. On this we start with the asymptotic case and pick up from Lecture 1 by proving Parts 3 and 4 from Theorem 2.1.

2.2 Asymptotic analysis of the excess risk

In Lecture 1, we were introduced to Theorem 2.1 (this version has an additional Part 5, which is a consequence of Part 4).

Theorem 2.1. *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. 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. *$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, (\nabla^2 L(\theta^*))^{-1} Cov(\ell((x,y),\theta^*))(\nabla^2 L(\theta^*))^{-1}\right)$.*
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 $S \sim \mathcal{N}\left(0, (\nabla^2 L(\theta^*))^{-1/2} Cov(\ell((x,y),\theta^*))(\nabla^2 L(\theta^*))^{-1/2}\right)$.*
5. *$\lim_{n\to\infty} \mathbb{E}\left[n(L(\hat{\theta}) - L(\theta^*))\right] = \frac{1}{2}\operatorname{tr}\left(\nabla^2 L(\theta^*)^{-1} \operatorname{Cov}(\nabla \ell((x,y),\theta^*)\right)$.*

Proof. Recall that Parts 1 and 2 were proven in Lecture 1, so we proceed with the proof of the remaining parts. Using a Taylor expansion of L with respect to θ at θ^*, we find

$$L(\hat{\theta}) = L(\theta^*) + \langle \nabla L(\theta^*), \hat{\theta} - \theta^* \rangle + \frac{1}{2}\langle \hat{\theta} - \theta^*, \nabla^2 L(\theta^*)(\hat{\theta} - \theta^*) \rangle + o(\|\hat{\theta} - \theta^*\|_2^2). \tag{2.1}$$

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

$$\begin{aligned} n(L(\hat{\theta}) - L(\theta^*)) &= \frac{n}{2}\langle \hat{\theta} - \theta^*, \nabla^2 L(\theta^*)(\hat{\theta} - \theta^*) \rangle + o(\|\hat{\theta} - \theta^*\|_2^2) && (2.2) \\ &\approx \frac{1}{2}\langle \sqrt{n}(\hat{\theta} - \theta^*), \nabla^2 L(\theta^*)\sqrt{n}(\hat{\theta} - \theta^*) \rangle && (2.3) \\ &= \frac{1}{2}\left\| \nabla^2 L(\theta^*)^{1/2}\sqrt{n}(\hat{\theta} - \theta^*) \right\|_2^2, && (2.4) \end{aligned}$$

where the last equality follows by the fact that for any vector v and square matrix A of appropriate dimensions, the inner product $\langle v, Av\rangle = v^T Av = \|A^{1/2}v\|_2^2$. 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

$$S \sim \nabla^2 L(\theta^*)^{1/2} \cdot \mathcal{N}\left(0, \nabla^2 L(\theta^*)^{-1} \operatorname{Cov}(\nabla \ell((x, y), \theta^*) \nabla^2 L(\theta^*)^{-1}\right) \tag{2.5}$$

$$\overset{d}{=} \mathcal{N}\left(0, \nabla^2 L(\theta^*)^{-1/2} \operatorname{Cov}(\nabla \ell((x, y), \theta^*) \nabla^2 L(\theta^*)^{-1/2}\right). \tag{2.6}$$

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

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[n(L(\hat{\theta}) - L(\theta^*))\right] &= \frac{1}{2}\mathbb{E}\left[\|S\|_2^2\right] && (2.7)\\
&= \frac{1}{2}\mathbb{E}\left[\operatorname{tr}(S^\top S)\right] && (2.8)\\
&= \frac{1}{2}\mathbb{E}\left[\operatorname{tr}(SS^\top)\right] && (2.9)\\
&= \frac{1}{2}\operatorname{tr}\left(\mathbb{E}[SS^\top]\right) && (2.10)\\
&= \frac{1}{2}\operatorname{tr}\left(\operatorname{Cov}(S)\right) && (2.11)\\
&= \frac{1}{2}\operatorname{tr}\left(\nabla^2 L(\theta^*)^{-1}\operatorname{Cov}(\nabla \ell((x, y), \theta^*)\right). && (2.12)
\end{aligned}$$

□

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 $1/n$.

2.2.1 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 2.2.

Theorem 2.2. *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*

$$\theta^* = \theta_*, \tag{2.13}$$

$$\mathbb{E}\left[\nabla \ell((x, y), \theta^*)\right] = 0, \tag{2.14}$$

$$\mathrm{Cov}\left(\nabla \ell((x, y), \theta^*)\right) = \nabla^2 L(\theta^*), \;\; \textit{and} \tag{2.15}$$

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \nabla^2 L(\theta^*)^{-1}). \tag{2.16}$$

Remark 1: You may also have seen (2.16) in the following form: under the maximum likelihood estimation (MLE) paradigm, the MLE is asymptotically efficient in 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 $\nabla^2 L(\theta^*)$) [Ric06].

Remark 2: (2.15) is also known as Bartlett's identity [Lia16].

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

Proof. From the definition of the population loss,

$$\begin{aligned}
L(\theta) &= \mathbb{E}\left[\ell((x^{(i)}, y^{(i)}), \theta)\right] && (2.17)\\
&= \mathbb{E}\left[-\log P(y \mid x; \theta)\right] && (2.18)\\
&= \mathbb{E}\left[-\log P(y \mid x; \theta) + \log P(y \mid x; \theta_*)\right] + \mathbb{E}\left[-\log P(y \mid x; \theta_*)\right] && (2.19)\\
&= \mathbb{E}\left[\log \frac{P(y \mid x; \theta_*)}{P(y \mid x; \theta)}\right] + \mathbb{E}\left[-\log P(y \mid x; \theta_*)\right]. && (2.20)
\end{aligned}$$

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):

$$L(\theta) = \mathbb{E}\left[\mathbb{E}\left[\log \frac{P(y \mid x; \theta_*)}{P(y \mid x; \theta)} \middle| x\right]\right] + \mathcal{H}(y \mid x; \theta_*). \tag{2.21}$$

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

$$\begin{aligned}
L(\theta) &= \mathbb{E}\left[\mathrm{KL}\left(y \mid x; \theta_* \| y \mid x; \theta\right)\right] + \mathcal{H}(y \mid x; \theta_*) && (2.22)\\
&\geq \mathcal{H}(y \mid x; \theta_*), && (2.23)
\end{aligned}$$

since KL divergence is always non-negative. Since θ_* makes the KL divergence term 0, it minimizes $L(\theta)$ and so $\theta_* \in \mathrm{argmin}_\theta L(\theta)$. However, the minimizer of $L(\theta)$ is unique because of consistency, so we must have $\mathrm{argmin}_\theta L(\theta) = \theta^*$ which proves (2.13).

For (2.14), recall $\nabla L(\theta^*) = 0$, so we have

$$0 = \nabla L(\theta^*) = \nabla \mathbb{E}\left[\ell((x^{(i)}, y^{(i)}), \theta^*)\right] = \mathbb{E}\left[\nabla \ell((x^{(i)}, y^{(i)}), \theta^*)\right], \tag{2.24}$$

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

To prove (2.15), we first expand the RHS using the definition of covariance and express the marginal distributions as integrals:

$$\begin{aligned}
\mathrm{Cov}\left(\nabla\ell((x,y),\theta^*)\right) &= \mathbb{E}\left[\nabla\ell((x,y),\theta^*)\nabla\ell((x,y),\theta^*)^\top\right] && (2.25)\\
&= \int P(x)\left(\int P(y \mid x;\theta^*)\nabla\log P(y^{(i)} \mid x^{(i)};\theta^*)\nabla\log P(y^{(i)} \mid x^{(i)};\theta^*)^\top dy\right)dx && (2.26)\\
&= \int P(x)\left(\int \frac{\nabla P(y \mid x;\theta^*)\nabla P(y \mid x;\theta^*)^\top}{P(y \mid x;\theta^*)}dy\right)dx. && (2.27)
\end{aligned}$$

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

$$\begin{aligned}
\nabla^2 L(\theta^*) &= \mathbb{E}\left[\nabla^2 \log P(y \mid x;\theta^*)\right] && (2.28)\\
&= \int P(x)\left(\int -\nabla^2 P(y \mid x;\theta^*) + \frac{\nabla P(y \mid x;\theta^*)\nabla P(y \mid x;\theta^*)^\top}{P(y \mid x;\theta^*)}dy\right)dx. && (2.29)
\end{aligned}$$

Note that we can express

$$\int \nabla^2 P(y \mid x;\theta^*)dy = \nabla^2\int P(y \mid x;\theta^*)dy = \nabla 1 = 0 \tag{2.30}$$

so we find

$$\nabla^2 L(\theta^*) = \int P(x)\left(\int \frac{\nabla P(y \mid x;\theta^*)\nabla P(y \mid x;\theta^*)^\top}{P(y \mid x;\theta^*)}dy\right)dx = \mathrm{Cov}\left(\nabla\ell((x,y),\theta^*)\right). \tag{2.31}$$

Finally, (2.16) follows directly from Part 2 of Theorem 2.1 and (2.15). □

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:

$$\lim_{n\to\infty}\mathbb{E}\left[L(\hat{\theta}) - L(\theta^*)\right] = \frac{p}{2n}. \tag{2.32}$$

2.3 Introduction to non-asymptotic analysis

This section introduces the concept of *non-asymptotic analysis*. In the first half of this lecture and the previous lecture, we covered asymptotic analysis which tells you how learning algorithms behave as $n \to \infty$. This section will:

- Motivate non-asymptotic analysis by highlighting a core limitation of asymptotic analysis.
- Introduce forms of Hoeffding's inequality, a concentration inequality used to derive bounds everywhere in non-asymptotic analysis.

2.3.1 Motivation

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

$$\frac{p}{2n} + o\left(\frac{1}{n}\right). \tag{2.33}$$

(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 (2.33):

$$\frac{p}{2n} + \frac{1}{n^2} \quad \text{vs.} \quad \frac{p}{2n} + \frac{p^{100}}{n^2}. \tag{2.34}$$

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 $n > p^{50}$ before $\frac{p^{100}}{n^2}$ is 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.

2.3.2 Clarifications on notation

For the rest of this course, we will use "big-O" notation in the following sense: every occurrence of $O(x)$ is a placeholder for some function $f(x)$ such that for every x, $|f(x)| \leq Cx$ for some absolute/universal constant C. (The difference with traditional "big-O" notation is that we do not need to send $n \to \infty$ in order to define "big-O".)

Also, for any $a, b \geq 0$, we will let $a \lesssim b$ mean that there is some absolute constant $c > 0$ such that $a \leq cb$.

In this lecture, we are not going to care about the constants much, so the absolute constants in the bounds may not be tight.

2.3.3 Preliminaries: Hoeffding's inequality

We provide a brief overview of Hoeffding's inequality, a foundational concentration inequality:

Theorem 2.3 (Hoeffding's inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. real-valued random variables drawn from some distribution, such that $a_i \leq X_i \leq b_i$ almost surely. Define $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, and let $\mu = \mathbb{E}[\bar{X}]$. Then for any $\varepsilon > 0$,*

$$\Pr\left[|\bar{X} - \mu| \leq \varepsilon\right] \geq 1 - 2\exp\left(\frac{-2n^2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{2.35}$$

Note that the demoninator within the exponential term, $\sum_{i=1}^n (b_i - a_i)^2$, can be thought of as an upper bound or proxy for the variance $\mathrm{Var}(X_i)$. In fact, under the i.i.d. assumption, we can show

$$\mathrm{Var}\left(\bar{X}\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(x_i) \leq \frac{1}{n^2}\sum_{i=1}^n (b_i - a_i)^2. \tag{2.36}$$

Let $\sigma^2 = \frac{1}{n^2}\sum_{i=1}^n (b_i - a_i)^2$. If we take $\varepsilon = O(\sigma\sqrt{\log n}) = \sigma\sqrt{c\log n}$; i.e. ε bounded above by some large (i.e., $c \geq 10$) multiple of the standard deviation of the x_i's times $\sqrt{\log n}$, we can substitute this value of ε into (2.35) to reach the following conclusion:

$$\begin{aligned}\Pr\left[|\bar{X} - \mu| \leq \varepsilon\right] &\geq 1 - 2\exp\left(\frac{-2\varepsilon^2}{\sigma^2}\right) & (2.37)\\ &= 1 - 2\exp(-2c\log n) & (2.38)\\ &= 1 - 2n^{-2c} & (2.39)\end{aligned}$$

We can see that as n grows, the right-most term tends to zero such that $\Pr[|\bar{X} - \mu| \leq \varepsilon]$ very quickly approaches 1. Intuitively, this result tells us that, with high probability, the sample mean $\bar{X}$ will not be "much farther" from the population mean μ by some sublogarithmic ($\sqrt{c\log n}$) factor of the standard deviation.[1] Thus, we can restate the above claim we reached as follows:

Remark 2.4. For sufficiently large n, $|\bar{X} - \mu| \leq O(\sigma\sqrt{\log n})$ with high probability.

Remark 2.5. If, in addition, we have $a_i = -O(1)$ and $b_i = O(1)$, then $\sigma^2 = O\left(\frac{1}{n}\right)$, and $|\bar{X} - \mu| \leq O\left(\sqrt{\frac{\log n}{n}}\right) = \widetilde{O}\left(\frac{1}{\sqrt{n}}\right)$.[2]

Remark 2.5 provides a compact form of the Hoeffding bound that we can use when the X_i are bounded almost surely. Now, we will apply this mathematical machinery to learning theory.

2.4 Introduction to uniform convergence

One problem area where non-asymptotic analysis is useful is *uniform convergence.* A central goal of learning theory is to bound the *excess risk* $L(\hat{\theta}) - L(\theta^*)$. This is important as we don't want the expected risk of our ERM to be much larger than the expected risk of the best possible model. As we will see in the remainder of this section, uniform convergence is a technique that helps us achieve such bounds.

Uniform convergence is a property of a parameter set Θ, which gives us bounds of the form

$$\Pr\left[|\hat{L}(\theta) - L(\theta)| \geq \varepsilon\right] \leq \delta;\ \forall\theta \in \Theta. \tag{2.40}$$

In other words, uniform convergence tells us that for any choice of θ, our empirical risk is always close to our population risk with high probability. Let's look at a motivating example for why this type of bound is useful.

2.4.1 Motivation: Uniform convergence implies generalization

Consider the standard supervised learning setup where we have some i.i.d. $(x^{(i)}, y^{(i)})$. Furthermore, assume that we have a bounded loss function; specifically, suppose that $0 \leq \ell((x, y); \theta) \leq 1$, as in the case of the zero-one loss function. We show that uniform convergence implies generalization.

[1] This is with the caveat, of course, that σ is not exactly standard deviation but a loose upper bound on standard deviation.

[2] $\widetilde{O}$ is analogous to Big-O notation, in that $\widetilde{O}$ hides logarithmic factors. That is; if $f(n) = O(\log n)$, then $f(n) = \widetilde{O}(1)$.

First, via telescoping sums, we can decompose the excess risk into three terms:

$$L(\hat{\theta}) - L(\theta^*) = \underbrace{L(\hat{\theta}) - \hat{L}(\hat{\theta})}_{\textcircled{1}} + \underbrace{\hat{L}(\hat{\theta}) - \hat{L}(\theta^*)}_{\textcircled{2}} + \underbrace{\hat{L}(\theta^*) - L(\theta^*)}_{\textcircled{3}}. \tag{2.41}$$

We know that $\hat{L}(\hat{\theta}) - \hat{L}(\theta^*) \leq 0$ since $\hat{\theta}$ is a minimizer of $\hat{L}$. This allows us to write

$$\begin{aligned} L(\hat{\theta}) - L(\theta^*) &\leq |L(\hat{\theta}) - \hat{L}(\hat{\theta})| + \hat{L}(\hat{\theta}) - \hat{L}(\theta^*) + |\hat{L}(\theta^*) - L(\theta^*)| && (2.42)\\ &\leq |L(\hat{\theta}) - \hat{L}(\hat{\theta})| + 0 + |\hat{L}(\theta^*) - L(\theta^*)| && (2.43)\\ &\leq 2 \sup_{\theta \in \Theta} |L(\theta) - \hat{L}(\theta)|. && (2.44) \end{aligned}$$

This result tells us that if $\sup_{\theta \in \Theta} |L(\theta) - \hat{L}(\theta)|$ is small (say, less than $\varepsilon/2$), then excess risk $L(\hat{\theta}) - L(\theta^*)$ is less than ε. But this is exactly in the form of the bound in (2.40). Hence, if we can show that a parameter family exhibits uniform convergence, we can get a bound on excess risk as well.

For future references, Equation (2.44) can be strengthened straightforwardly into the following with slightly more careful treatment of the signs of each term:

$$L(\hat{\theta}) - L(\theta^*) \leq |\hat{L}(\theta^*) - L(\theta^*)| + L(\hat{\theta}) - \hat{L}(\hat{\theta}) \leq |\hat{L}(\theta^*) - L(\theta^*)| + \sup_{\theta \in \Theta} \left(L(\hat{\theta}) - \hat{L}(\hat{\theta}) \right) \tag{2.45}$$

This will make some of our future derivations technically slightly more convenient, but the nuanced difference between Equations (2.44) and (2.45) does not change the fundamental idea and the discussions in this lecture.

Let us try to apply our knowledge of concentration inequalities to this problem. Earlier we assumed that $\ell((x, y); \theta)$ is bounded, so we can bound $\textcircled{3}$ by $\widetilde{O}\left(\frac{1}{\sqrt{n}}\right)$ via Hoeffding's inequality (Remark 2.5). However, we cannot apply the same concentration inequality to $\textcircled{1}$: since $\hat{\theta}$ is data-dependent by definition, the i.i.d. assumption no longer holds. (To see this, note that $\hat{\theta}$ depends on the training dataset $(x^{(i)}, y^{(i)})$, so the terms in $\widehat{L}(\theta)$, $\ell((x^{(i)}, y^{(i)}); \hat{\theta})$, all depend on the training dataset too.) This is concerning: it is certainly possible that $L(\hat{\theta}) - \hat{L}(\hat{\theta})$ is large. You've probably encountered this yourself when a model exhibits low training loss, but high validation/testing loss.

2.4.2 Deriving uniform convergence bounds

Uniform convergence is one way we can fix this issue. The high-level idea is as follows:

- Suppose we have a bound of the form $\Pr[|\hat{L}(\theta) - L(\theta)| \geq \varepsilon'] \leq \delta'$ for some single, fixed choice of θ.
- If we know *all possible values of* θ in advance, we can use the above bound to create a more general bound over all values of θ.

In particular, we can use the union-bound inequality to create the general bound described in the second bullet point, using the bound in the first bullet point:

$$\Pr\left[\forall \theta \in \Theta; |\hat{L}(\theta) - L(\theta)| \geq \varepsilon'\right] \leq \sum_{\theta \in \Theta} \Pr\left[|\hat{L}(\theta) - L(\theta)| \geq \varepsilon'\right]. \tag{2.46}$$

We can then use Hoeffding's inequality to deal with the summands as θ there is no longer data-dependent. We'll talk more in future lectures about proving statements of this form.

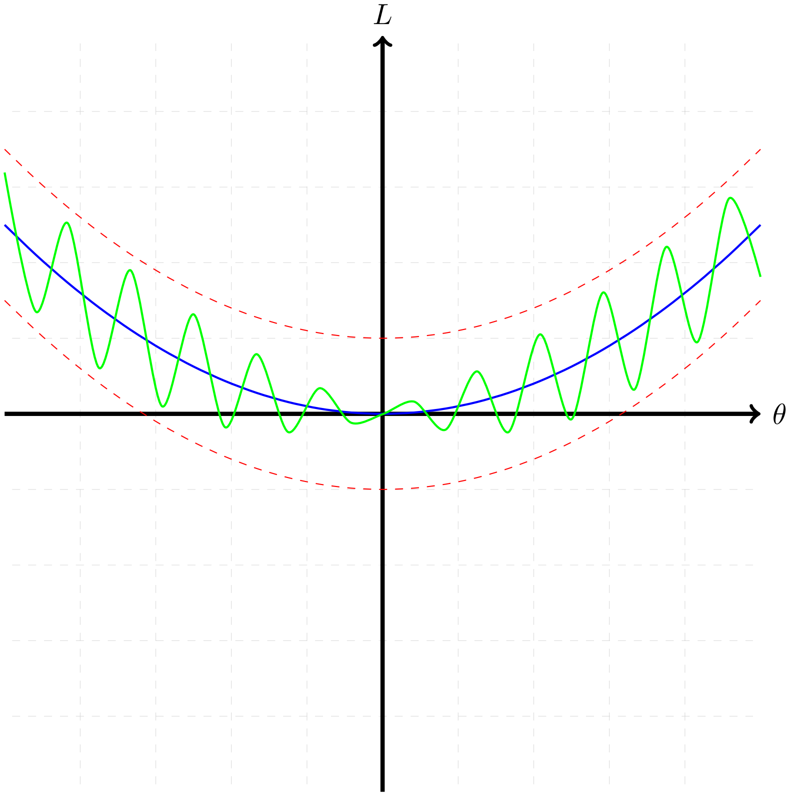

Figure 2.1: Empirical risk landscape under uniform convergence: Green: empirical risk, blue: population risk, red, dashed: ε additive error bounds for excess risk.

2.4.3 Intuitive Interpretation of Uniform Convergence.

Since uniform convergence implies generalization, if we know that population risk and empirical risk are always "close," then excess risk is "small" as well (Figure 2.1). In fact, it is possible to show that not only is $L(\theta)$ "close" to $\hat{L}(\theta)$ for sufficiently large data, but that the "shape" of $\hat{L}$ is "close" to the shape of L as well (Figure 2.2). This holds for the convex case; furthermore, there are conditions under which this holds in the non-convex case, for which a rigorous treatment can be found in [MBM17].

Figure design and some wording in this section was inspired by [Lia16, LT18].

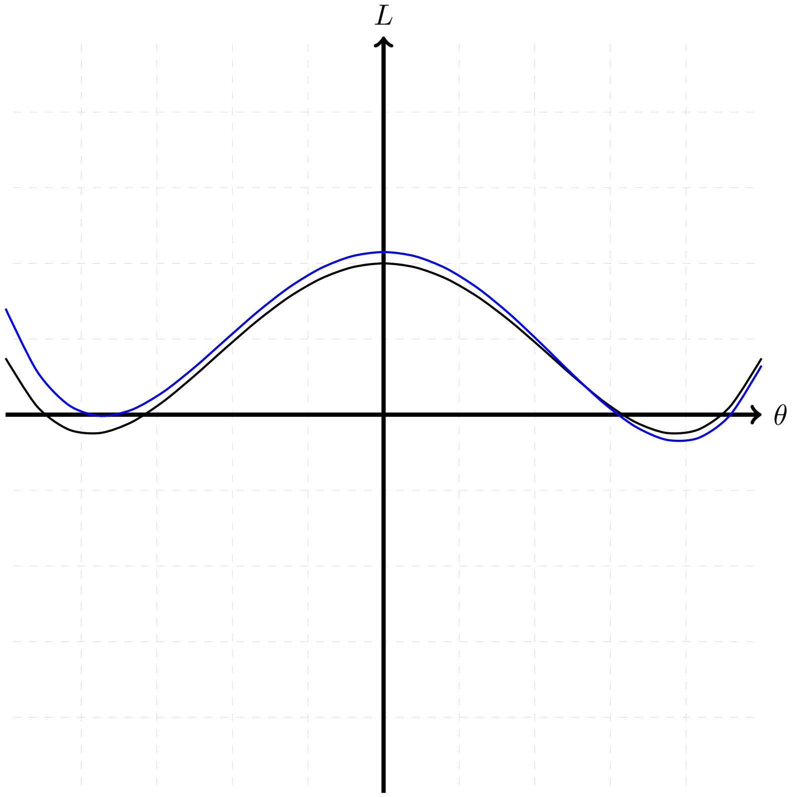

Figure 2.2: Empirical risk landscape under uniform convergence: Blue: empirical risk, black: population risk.

Bibliography

[Lia16] Percy Liang, *Cs229t/stat231: Statistical learning theory (winter 2016)*, April 2016.

[LT18] Pengda Liu and Garrett Thomas, *Cs229t/stat231: Statistical learning theory (fall 2018)*, October 2018.

[MBM17] Song Mei, Yu Bai, and Andrea Montanari, *The landscape of empirical risk for non-convex losses*, 2017.

[Ric06] John A. Rice, *Mathematical statistics and data analysis.*, third ed., Belmont, CA: Duxbury Press., 2006.