Lecture Notes (in Progress)
STATS214 / CS229M: Machine Learning Theory (Winter 2021)
@ Stanford University

Tianyu Du

January 14, 2021

Note: CS229M is different from CS229: Machine Learning

1 Preliminary

Lecture 1. Jan. 11, 2021

1.1 Formulation and Asymptotics

For components of standard supervised learning problems, we use the following notations.

- Input space: $\mathcal{X}$.
- Output space: $\mathcal{Y}$.
- Joint probability distribution P over $\mathcal{X} \times \mathcal{Y}$.
- Training data $(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)}) \overset{i.i.d.}{\sim} P$.
- Predictors/model/hypothesis $h : \mathcal{X} \to \mathcal{Y}$.
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, typically we assume $\ell(\hat{y}, y) \geq 0$ for all $\hat{y}, y \in \mathcal{Y}$.
- The expected/population risk/loss $L(h) \overset{\triangle}{=} \mathbb{E}_{(x,y)\sim P}\ell(h(x), y))]$, *the goal of supervised learning problems is to minimize the population risk.*
- Hypothesis class/family $\mathcal{H}$ is the set of all functions from $\mathcal{X}$ to $\mathcal{Y}$.
- Excess risk (w.r.t. $\mathcal{H}$) of a particular $h \in \mathcal{H}$ is defined as $L(h) - \inf_{g \in \mathcal{H}} L(g)$, the excess risk is always non-negative.

Example 1.1. For regression problems, $\mathcal{Y} = \mathbb{R}$ and typically $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$. For k-class classification problems, $\mathcal{Y} = \{1, 2, \dots, k\}$ and $\ell_{0\text{-}1}(\hat{y}, y) = \mathbb{1}\{\hat{y} \neq y\}$.

1.2 Empirical Risk Minimization (ERM)

The training loss / empirical loss / empirical risk associated a particular dataset $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ is defined as

$$\hat{L} \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(h(x^{(i)}), y^{(i)}) \tag{1}$$

The ERM estimator is

$$\hat{h} \triangleq \underset{h\in\mathcal{H}}{\operatorname{argmin}}\, \hat{L}(h) \tag{2}$$

Because $(x^{(i)}, y^{(i)}) \sim P$, for every $h \in \mathcal{H}$,

$$\mathbb{E}_{\{(x^{(i)},y^{(i)})\}_{i=1}^n \overset{i.i.d.}{\sim} P}[\hat{L}(h)] = L(h) \tag{3}$$

1.3 Parameterization

Consider the family of hypothesis parameterized by θ: $\mathcal{H} = \{h_\theta \mid \theta \in \Theta\}$. For instance, with $\Theta = \mathbb{R}^d$ and $h_\theta(x) = \theta^T x$, $\mathcal{H}$ becomes the family of linear models. The ERM for parameterized family is

$$\hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(h_\theta(x^{(i)}), y^{(i)}) \tag{4}$$

$$\hat{\theta} = \hat{\theta}_{ERM} = \underset{\theta\in\Theta}{\operatorname{argmin}}\, \hat{L}(\theta) \tag{5}$$

Sometimes we use the alternative notion $\ell((x^{(i)}, y^{(i)}), \theta)$ for $\ell(h_\theta(x^{(i)}), y^{(i)})$.

Goal: bound the excess risk of $\hat{\theta}$.

1.4 Asymptotic Analysis

Let $n \to \infty$, we wish to obtain a bound with form

$$L(\hat{\theta}) - \operatorname{argmin} L(\theta) \leq \frac{c}{n} + o(\frac{1}{n}) \tag{6}$$

where c depends on the problem.

From now on,

$$\Theta = \mathbb{R}^p \tag{7}$$

$$\hat{\theta} = \underset{\theta\in\mathbb{R}^p}{\operatorname{argmin}}\, \hat{L}(\theta) \tag{8}$$

$$\theta^* = \underset{\theta\in\mathbb{R}^p}{\operatorname{argmin}}\, L(\theta) \tag{9}$$

$$\text{excess risk} = L(\hat{\theta}) - L(\theta^*) \tag{10}$$

Theorem 1.1. Assume the consistency of $\hat{\theta}$,

$$\hat{\theta} \xrightarrow{p} \theta^* \text{ as } n \to \infty \tag{11}$$

Further, suppose $\nabla^2 L(\theta^*)$ has full-rank, and mild regularity conditions, there exists absolute constants $c_0, c_1 \in \mathbb{R}_+$ such that

1. $\sqrt{n}||\hat{\theta} - \theta^*|| \xrightarrow{p} c_0$,
2. $n[L(\hat{\theta}) - L(\theta^*)] \xrightarrow{p} c_1$,
3. $\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \nabla^2 L(\theta^*)^{-1}\text{cov}(\nabla \ell((x,y),\theta^*))\nabla^2 L(\theta^*)^{-1})$,
4. Let $S \sim \mathcal{N}(0, \underbrace{\nabla^2 L(\theta^*)^{-1/2}\text{cov}(\nabla \ell((x,y),\theta))\nabla^2 L(\theta^*)^{-1/2}}_{W})$, then

$$n(L(\hat{\theta}) - L(\theta^*)) \xrightarrow{d} \frac{1}{2}||S||_2^2$$

and

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

Proof. Together with the optimality of $\hat{\theta}$ with respect to $\hat{L}$, the Taylor expansion of $\hat{L}$ around θ^* indicates

$$0 = \nabla \hat{L}(\hat{\theta}) = \nabla \hat{L}(\theta^*) + \nabla^2 \hat{L}(\theta^*)(\hat{\theta} - \theta^*) + \mathcal{O}(||\hat{\theta} - \theta^*||_2^2) \tag{12}$$

$$\implies \hat{\theta} - \theta^* = -\nabla^2 \hat{L}(\theta^*)^{-1}\nabla \hat{L}(\theta^*) + \mathcal{O}(||\hat{\theta} - \theta^*||_2^2) \tag{13}$$

Let $\ell_i(\theta) = \ell((x^{(i)}, y^{(i)}), \theta)$ denote the individual loss, then the following holds

- $\nabla \hat{L}(\theta^*) = \frac{1}{n}\sum_{i=1}^n \nabla \ell_i(\theta^*)$.
- $\nabla^2 \hat{L}(\theta^*) = \frac{1}{n}\sum_{i=1}^n \nabla^2 \ell_i(\theta^*)$.

Moreover, by law of large numbers (LLN),

- $\nabla \hat{L}(\theta^*) \xrightarrow{p} \nabla L(\theta^*) = 0$ and $\mathbb{E}\left[\nabla \hat{L}(\theta^*)\right] = \nabla L(\theta^*)$.
- $\nabla^2 \hat{L}(\theta^*) \xrightarrow{p} \nabla^2 L(\theta^*) \neq 0$ and $\mathbb{E}\left[\nabla^2 \hat{L}(\theta^*)\right] = \nabla^2 L(\theta^*)$

Theorem 1.2 (Central Limit Theorem)**.** Let $X_1, \ldots, X_n$ be n i.i.d. random variables, let $\Sigma = \text{cov}(X_i)$. As $n \to \infty$, define $\hat{X} = \frac{1}{n}\sum_{i=1}^n X_i$,

- $\hat{X} \xrightarrow{p} \mathbb{E}[\hat{X}]$,
- $\sqrt{n}(\hat{X} - \mathbb{E}[\hat{X}]) \xrightarrow{d} \mathcal{N}(0, \Sigma)$.

Since $\nabla \hat{L}(\theta^*)$ is the mean of n i.i.d. random variables $\ell_i(\theta^*)$, by the central limit theorem (CLT),

$$\sqrt{n}(\nabla \hat{L}(\theta^*) - \nabla L(\theta^*)) \to \mathcal{N}(0, \mathrm{cov}(\nabla \ell_i)) \tag{14}$$
$$\sqrt{n}\nabla \hat{L}(\theta^*) \to \mathcal{N}(0, \mathrm{cov}(\nabla \ell_i)) \tag{15}$$

where $\Sigma = \mathrm{cov}(\ell_i)$.

$$\begin{aligned}
\hat{\theta} - \theta^* &= -\nabla^2 \hat{L}(\theta^*)^{-1} \frac{1}{n} \sum_{i=1}^{n} \nabla \ell_i(\theta^*) + \mathcal{O}(||\hat{\theta} - \theta^*||_2^2) && (16)\\
&= -\left(\nabla^2 L(\theta^*) + \mathcal{O}(\frac{1}{\sqrt{n}})\right)^{-1} \mathcal{O}(\frac{1}{\sqrt{n}}) + \mathcal{O}(||\hat{\theta} - \theta^*||_2^2) && (17)\\
&= \nabla^2 L(\theta^*) \mathcal{O}(\frac{1}{\sqrt{n}}) \approx \frac{1}{\sqrt{n}} && (18)
\end{aligned}$$

More precisely,

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta^*) &= -\underbrace{\nabla^2 \hat{L}(\theta^*)^{-1}}_{\approx \nabla^2 L(\theta^*)^{-1}} \underbrace{\sqrt{n}[\nabla \hat{L}(\theta^*) - \nabla L(\theta^*)]}_{\mathcal{N}(0,\Sigma)} + \mathcal{O}(||\hat{\theta} - \theta^*||_2^2) && (19)\\
&= \nabla^2 L(\theta^*)^{-1} Z \text{ where } Z \sim \mathcal{N}(0, \mathrm{cov}(\nabla \ell_i)) && (20)\\
&\overset{d}{=} \mathcal{N}(0, \nabla^2 L(\theta^*)^{-1} \mathrm{cov}(\nabla \ell_i) \nabla^2 L(\theta^*)^{-1}) && (21)
\end{aligned}$$

Lecture 2. Jan. 13, 2021

The Taylor's expansion of L around θ^* implies

$$L(\hat{\theta}) - L(\theta^*) = \langle \nabla L(\theta^*), \hat{\theta} - \theta^* \rangle + \frac{1}{2} \langle \hat{\theta} - \theta^*, \nabla^2 L(\theta^*)(\hat{\theta} - \theta^*) \rangle + \mathcal{O}(||\hat{\theta} - \theta^*||_2^2) \tag{22}$$

Since $\theta^* \equiv \mathrm{argmin}_{\theta \in \Theta} L(\theta)$, $\nabla L(\theta^*) = 0$. Multiply both sides by n,

$$n[L(\hat{\theta}) - L(\theta^*)] = \frac{1}{2} \langle \sqrt{n}(\hat{\theta} - \theta^*), \nabla^2 L(\theta^*) \sqrt{n}(\hat{\theta} - \theta^*) \rangle + \text{higher order terms} \tag{23}$$

Note that $\langle v, Av \rangle = ||A^{1/2} v||_2^2$,

$$(23) = \frac{1}{2} ||\nabla^2 L(\theta^*)^{1/2} \sqrt{n}(\hat{\theta} - \theta^*)||_2^2 + \text{higher order terms} \tag{24}$$

By result (3) and property of Gaussian distribution,

$$\begin{aligned}
\nabla^2 L(\theta^*)^{1/2} \sqrt{n}(\hat{\theta} - \theta^*) &\sim \mathcal{N}(0, \nabla^2 L(\theta^*)^{1/2} \nabla^2 L(\theta^*)^{-1} \mathrm{cov}(\nabla \ell((x,y),\theta)) \nabla^2 L(\theta^*)^{-1} \nabla^2 L(\theta^*)^{1/2}) && (25)\\
&= \mathcal{N}(0, \nabla^2 L(\theta^*)^{-1/2} \mathrm{cov}(\nabla \ell((x,y),\theta)) \nabla^2 L(\theta^*)^{-1/2}) \overset{d}{=} S && (26)
\end{aligned}$$

Consequently,

$$(24) \stackrel{d}{=} \frac{1}{2}||S||_2^2 + \text{higher order terms} \tag{27}$$

The first moment of $n[L(\hat{\theta}) - L(\theta^*)]$ converges as well, and because $\mathbb{E}\left[||v||_2^2\right] = \mathbb{E}\left[\operatorname{tr}(vv^T)\right] = \operatorname{tr}(\mathbb{E}\left[vv^T\right])$,

$$\mathbb{E}\left[n[L(\hat{\theta}) - L(\theta^*)]\right] \xrightarrow{p} \frac{1}{2}\mathbb{E}\left[||S||_2^2\right] \tag{28}$$

$$= \frac{1}{2}\operatorname{tr}(\nabla^2 L(\theta^*)^{-1/2}\operatorname{cov}(\nabla\ell)\nabla^2 L(\theta^*)^{-1/2}) \tag{29}$$

$$= \frac{1}{2}\operatorname{tr}(\nabla^2 L(\theta^*)^{-1}\operatorname{cov}(\nabla\ell)) \tag{30}$$

■

1.5 Well-Specified Case

Theorem 1.3 (Well-Specification)**.** In addition to assumptions in Theorem 1.1, suppose there exists some probabilistic model $P(y|x;\theta)$ parameterized by θ, that is,

$$\exists\theta_* \text{ s.t. } y^{(i)}|x^{(i)} \sim P(y|x;\theta_*)\ \forall i \in [n] \tag{31}$$

take the loss function to be the negative log likelihood

$$\ell((x^{(i)}, y^{(i)}); \theta) = -\log P(y^{(i)}|x^{(i)}; \theta) \tag{32}$$

then,

(1) The excess risk minimizer equals the ground truth: $\theta^* \equiv \operatorname{argmin}_\theta L(\theta) = \theta_*$.

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

(3) $\operatorname{cov}(\nabla\ell((x,y),\theta^*)) = \nabla^2 L(\theta^*)$.

(4) $\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \nabla^2 L(\theta^*)^{-1})$, suppose $S \sim \mathcal{N}(0,1)$,

$$n(L(\hat{\theta}) - L(\theta^*)) \xrightarrow{d} \frac{1}{2}||S||_2^2 \sim \chi^2(p) \tag{33}$$

So that

$$\mathbb{E}\left[L(\hat{\theta}) - L(\theta^*)\right] \approx \frac{p}{2n} \tag{34}$$

1.6 Limitation of Asymptotic Analysis

Asymptotic analysis hides dependencies on p, for instance, both $\frac{p}{2n} + \frac{1}{n^2}$ and $\frac{p}{2n} + \frac{p^{100}}{n^2}$ are classified into $\frac{p}{2n} + o(1/n)$ by asymptotic analysis.

In contrast, non-asymptotic analysis only hides absolute constants and we can bound model performance with form $L(\hat{\theta}) - L(\theta^*) \leq \mathcal{O}(f(p,n))\ \forall p, n \geq 1$.

In the following non-asymptotic analysis, every occurrence of $\mathcal{O}(x)$ is a placeholder for some function $f \in \mathcal{O}(x)$.

For all $a, b \geq 0$, $a \precsim b \iff \exists$ absolute constant $c \geq 0$ s.t. $a \leq cb$.

1.7 Uniform Convergence

Key Idea For every $\theta \in \Theta$, $\hat{L}(\theta)$ is an empirical estimate of $L(\theta)$, so $\hat{L}(\theta) \approx L(\theta)$ (we still need to prove this). If we can bound

$$\left|\hat{L}(\theta^*) - L(\theta^*)\right| \leq \alpha \tag{35}$$

$$L(\hat{\theta}) - \hat{L}(\hat{\theta}) \leq \alpha \tag{36}$$

Recall that we wanted to bound the excess risk of $\hat{\theta}$, which is

$$L(\hat{\theta}) - L(\theta^*) = [L(\hat{\theta}) - \hat{L}(\hat{\theta})] + [\hat{L}(\hat{\theta}) - \hat{L}(\theta^*)] + [\hat{L}(\theta^*) - L(\theta^*)] \tag{37}$$

$$\leq \alpha + 0 + \alpha = 2\alpha \tag{38}$$

1.8 Contraction Inequality (to show $L(\theta) \approx \hat{L}(\theta)$)

Theorem 1.4 (Hoeffding's Inequality)**.** Let $X_1, \ldots, X_n$ be i.i.d. real-valued random variables, assume $a_i \leq x_i \leq b_i$ for all i almost surely. Let $\mu = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right]$, then

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \leq \varepsilon\right] \geq 1 - 2\exp\left(\frac{-2n^2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \tag{39}$$

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \geq \varepsilon\right] \leq 2\exp\left(\frac{-2n^2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \tag{40}$$

To use this theorem, consider

$$\sigma^2 = \frac{1}{n^2}\sum_{i=1}^n (b_i - a_i)^2 \tag{41}$$

as a proxy for the variance of $\frac{1}{n}\sum_{i=1}^n X_i$:

$$Var(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) \leq \frac{1}{n^2} \leq \frac{1}{n^2}\sum_{i=1}^n (b_i - a_i)^2 \tag{42}$$

Take $\varepsilon = \mathcal{O}(\sqrt{\sigma^2 \log(n)}) = \mathcal{O}(\sqrt{c\sigma^2 \log(n)})$, where c is a large constant.

$$\Pr\left[\left|\frac{1}{n}\sum X_i - \mu\right| \leq \sqrt{c\sigma^2 \log n}\right] \geq 1 - 2\exp\left(\frac{-2n^2 c\sigma^2 \log n}{n^2\sigma^2}\right) \tag{43}$$
$$= 1 - 2\exp\left(-2c\log n\right) \tag{44}$$
$$= 1 - 2\exp\left(\log n^{-2c}\right) \tag{45}$$
$$= 1 - 2n^{-2c} \approx 1 \tag{46}$$

Moreover, if $a_i = -\mathcal{O}(1)$ and $b_i = \mathcal{O}(1)$, then $\sigma^2 = \frac{1}{n}$. With high probability,

$$\left|\frac{1}{n}\sum X_i - \mu\right| \leq \mathcal{O}(\sqrt{\sigma^2 \log n}) = \mathcal{O}(\sqrt{\frac{\log n}{n}}) = \tilde{\mathcal{O}}(\frac{1}{\sqrt{n}}) \tag{47}$$

1.9 Back to Learning Theory

Take $X_i = \ell((x^{(i)}, y^{(i)}); \theta)$, assume $\ell((x,y);\theta) \in [0,1]$ (such as 0-1 loss).

Lemma 1.1. For any θ , with high probability,

$$\left|\hat{L}(\theta) - L(\theta)\right| \leq \tilde{\mathcal{O}}(\frac{1}{\sqrt{n}}) \tag{48}$$

In particular, $\left|\hat{L}(\theta^*) - L(\theta^*)\right| \leq \tilde{\mathcal{O}}(\frac{1}{\sqrt{n}})$.

Still need to check.

1.10 Uniform Convergence

$\hat{L} \to L$ uniformly on Θ if

$$\Pr\left[\forall\theta \in \Theta, \left|\hat{L}(\theta) - L(\theta)\right| \leq \varepsilon'\right] \geq 1 - \delta' \tag{49}$$
$$\Pr\left[\exists\theta \in \Theta, \left|\hat{L}(\theta) - L(\theta)\right| \geq \varepsilon'\right] \leq \sum_{\theta\in\Theta} \Pr\left[\left|\hat{L}(\theta) - L(\theta)\right| \geq \varepsilon'\right] \tag{50}$$

$$\Pr\left[\forall\theta \in \Theta, \left|\hat{L}(\theta) - L(\theta)\right| \geq \varepsilon'\right] = 1 - \Pr\left[\forall\theta \in \Theta, \left|\hat{L}(\theta) - L(\theta)\right| \leq \varepsilon'\right] \tag{51}$$