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

1.1 Formulation and Notations

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

$$\hat{\theta} \xrightarrow{p} \theta^* \text{ as } n \to \infty \tag{1}$$

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

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

$$\implies \hat{\theta} - \theta^* = -\nabla^2 \hat{L}(\theta^*)^{-1} \nabla \hat{L}(\theta^*) + \mathcal{O}(||\hat{\theta} - \theta^*||_2^2) \tag{3}$$

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 $\Sigma = \mathrm{cov}(X_i)$. 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),

$$\sqrt{n}(\nabla \hat{L}(\theta^*) - \nabla L(\theta^*)) \to \mathcal{N}(0, \mathrm{cov}(\nabla \ell_i)) \tag{4}$$
$$\sqrt{n}\nabla \hat{L}(\theta^*) \to \mathcal{N}(0, \mathrm{cov}(\nabla \ell_i)) \tag{5}$$

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

$$\begin{aligned}
\hat{\theta} - \theta^* &= -\nabla^2 \hat{L}(\theta^*)^{-1} \frac{1}{n}\sum_{i=1}^n \nabla \ell_i(\theta^*) + \mathcal{O}(||\hat{\theta} - \theta^*||_2^2) && (6)\\
&= -\left(\nabla^2 L(\theta^*) + \mathcal{O}(\frac{1}{\sqrt{n}})\right)^{-1} \mathcal{O}(\frac{1}{\sqrt{n}}) + \mathcal{O}(||\hat{\theta} - \theta^*||_2^2) && (7)\\
&= \nabla^2 L(\theta^*)\mathcal{O}(\frac{1}{\sqrt{n}}) \approx \frac{1}{\sqrt{n}} && (8)
\end{aligned}$$

More precisely,

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta^*) &= -\underbrace{\nabla^2 \hat{L}(\theta^*)^{-1}}_{\approx \nabla^2 L(\theta^*)^{-1}} \underbrace{\sqrt{n}[\nabla \hat{L}(\theta^*) - \nabla L(\theta^*)]}_{\mathcal{N}(0,\Sigma)} + \mathcal{O}(||\hat{\theta} - \theta^*||_2^2) && (9)\\
&= \nabla^2 L(\theta^*)^{-1} Z \text{ where } Z \sim \mathcal{N}(0, \mathrm{cov}(\nabla \ell_i)) && (10)\\
&\overset{d}{=} \mathcal{N}(0, \nabla^2 L(\theta^*)^{-1} \mathrm{cov}(\nabla \ell_i) \nabla^2 L(\theta^*)^{-1}) && (11)
\end{aligned}$$

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

$$L(\hat{\theta}) - L(\theta^*) = \langle \nabla L(\theta^*), \hat{\theta} - \theta^* \rangle + \frac{1}{2}\langle \hat{\theta} - \theta^*, \nabla^2 L(\theta^*)(\hat{\theta} - \theta^*) \rangle + \mathcal{O}(||\hat{\theta} - \theta^*||_2^2) \tag{12}$$

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

$$n[L(\hat{\theta}) - L(\theta^*)] = \frac{1}{2}\langle\sqrt{n}(\hat{\theta} - \theta^*), \nabla^2 L(\theta^*)\sqrt{n}(\hat{\theta} - \theta^*)\rangle + \text{higher order terms} \tag{13}$$

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

$$(13) = \frac{1}{2}||\nabla^2 L(\theta^*)^{1/2}\sqrt{n}(\hat{\theta} - \theta^*)||_2^2 + \text{higher order terms} \tag{14}$$

By result (3) and property of Gaussian distribution,

$$\nabla^2 L(\theta^*)^{1/2}\sqrt{n}(\hat{\theta} - \theta^*) \sim \mathcal{N}(0, \nabla^2 L(\theta^*)^{1/2}\nabla^2 L(\theta^*)^{-1}\text{cov}(\nabla\ell((x,y),\theta))\nabla^2 L(\theta^*)^{-1}\nabla^2 L(\theta^*)^{1/2}) \tag{15}$$

$$= \mathcal{N}(0, \nabla^2 L(\theta^*)^{-1/2}\text{cov}(\nabla\ell((x,y),\theta))\nabla^2 L(\theta^*)^{-1/2}) \overset{d}{=} S \tag{16}$$

Consequently,

$$(14) \overset{d}{=} \frac{1}{2}||S||_2^2 + \text{higher order terms} \tag{17}$$

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

$$\mathbb{E}\left[n[L(\hat{\theta}) - L(\theta^*)]\right] \xrightarrow{p} \frac{1}{2}\mathbb{E}\left[||S||_2^2\right] \tag{18}$$

$$= \frac{1}{2}\text{tr}(\nabla^2 L(\theta^*)^{-1/2}\text{cov}(\nabla\ell)\nabla^2 L(\theta^*)^{-1/2}) \tag{19}$$

$$= \frac{1}{2}\text{tr}(\nabla^2 L(\theta^*)^{-1}\text{cov}(\nabla\ell)) \tag{20}$$

■

1.2 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,

$$\exists\theta_* \text{ s.t. } y^{(i)}|x^{(i)} \sim P(y|x;\theta_*)\ \forall i \in [n] \tag{21}$$

take the loss function to be the negative log likelihood

$$\ell((x^{(i)}, y^{(i)}); \theta) = -\log P(y^{(i)}|x^{(i)};\theta) \tag{22}$$

then,

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

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

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

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

$$n(L(\hat{\theta}) - L(\theta^*)) \xrightarrow{d} \frac{1}{2}||S||_2^2 \sim \chi^2(p) \tag{23}$$

So that

$$\mathbb{E}\left[L(\hat{\theta}) - L(\theta^*)\right] \approx \frac{p}{2n} \tag{24}$$

Remark 1.1 (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 $L(\hat{\theta}) - L(\theta^*) \leq \mathcal{O}(f(p, n))\ \forall p, n \geq 1$.

Notation 1.1. 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 $a, b \geq 0$,

$$a \precsim b \iff \exists \text{ absolute constant } c \geq 0 \text{ s.t. } a \leq cb \tag{25}$$

2 Uniform Convergence

Key Idea For every $\theta \in \Theta$, $\hat{L}(\theta)$ is an empirical estimate of $L(\theta)$ and $\hat{L}(\theta) \approx L(\theta)$. If we can bound

$$\left|\hat{L}(\theta^*) - L(\theta^*)\right| \leq \alpha \tag{1}$$

$$L(\hat{\theta}) - \hat{L}(\hat{\theta}) \leq \alpha \tag{2}$$

then

$$\begin{aligned} L(\hat{\theta}) - L(\theta^*) &= [L(\hat{\theta}) - \hat{L}(\hat{\theta})] + [\hat{L}(\hat{\theta}) - \hat{L}(\theta^*)] + [\hat{L}(\theta^*) - L(\theta^*)] && (3) \\ &\leq \alpha + 0 + \alpha = 2\alpha && (4) \end{aligned}$$

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

Theorem 2.1 (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

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \leq \varepsilon\right] \geq 1 - 2\exp\left(\frac{-2n^2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \tag{5}$$

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \geq \varepsilon\right] \leq 2\exp\left(\frac{-2n^2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \tag{6}$$