ML notes 10.21

Zixiao Wang, Yidan Yuan, Yixuan Zou, Tianyu Xu, Yifan Li,
Qibin Yang, Jinkai Fan, Yuchen Lin, Modi Wu, Shi Sheng, Yufei Ding, Dachao Hao

October 2025

1 Step I. Double Sampling Trick

1.1 Goal

Recall:
As our goal, we want to estimate the upperbound of the inference error:

$$\mathbb{P}\left(P_D\left(Y \neq \hat{f}(X)\right) - \frac{1}{n}\sum_{i=1}^{n} I\left[y_i \neq \hat{f}(x_i)\right] \geq \epsilon\right). \tag{1}$$

where $|\mathcal{F}| = \infty$ and $f \in \mathcal{F}$

$P_D\left(Y \neq \hat{f}(X)\right)$ is an expectation, which we desire to turn into an statistic estimation.Thus, we need to recall the double sampling trick:

1.2 Trick

$X_1, \ldots, X_{2n}$ are i.i.d. Bernoulli random variables. Def:

$$\nu_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \nu_2 = \frac{1}{n}\sum_{i=n+1}^{2n} X_i$$

With $n \geq \frac{\ln 2}{\varepsilon^2}$, the following Inequality exists:

$$\frac{1}{2}\mathbb{P}(|\nu_1 - \mathbb{E}[X]| \geq 2\varepsilon) \leq \mathbb{P}(|\nu_1 - \nu_2| \geq \varepsilon) \leq 2\mathbb{P}\left(|\nu_1 - \mathbb{E}[X]| \geq \frac{\varepsilon}{2}\right). \tag{2}$$

Here, the key point is that, the statistical average and the theoretical expectation is somehow convertable between each other. To be more precise, they can use each other as both the upper and lower bound.
Applying Inequation (2) to (1), we can obtain:

$$\Pr[\exists f \in \mathcal{F}, \Pr_D(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} I[Y_i \neq f(X_i)] \geq \epsilon] \tag{3}$$

$$\leq 2\Pr[\exists f \in \mathcal{F}, \frac{1}{n}\sum_{i=1}^{n} I[Y_i \neq f(X_i)] - \frac{1}{n}\sum_{i=n+1}^{2n} I[Y_i \neq f(X_i)] \geq \frac{\epsilon}{2}] \tag{4}$$

Therefore, we successfully eliminate the "Expectation" part.
Something a little tricky is that: directly using the llama, we can only obtain the above Inequation with a given f, instead of the $\exists f$.
The reason why it still works is that,

$$\mathbb{P}[\exists f \in \mathcal{F}, \Pr_D(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} I[Y_i \neq f(X_i)] \geq \epsilon] \tag{5}$$

$$= \mathbb{P}\left[\sup_{f\in\mathcal{F}}\left(\Pr_D(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} I[Y_i \neq f(X_i)]\right) \geq \epsilon\right] \tag{6}$$

The sup case has already been proved in last lecture's notes.

Step II. Symmetrization

We assume we draw $2n$ independent and identically distributed (i.i.d.) samples from the distribution $\mathcal{D}_{XY}$:

$$(x_i, y_i)_{i=1}^{2n} \overset{i.i.d.}{\sim} \mathcal{D}_{XY}.$$

Let $Z_i = (x_i, y_i)$ be the i-th sample.

The loss function is $\Phi_f(Z_i) = [y_i \neq f(x_i)]$, where $[\cdot]$ is the indicator function. The Symmetrization technique is used to bound the probability that the empirical risk deviates significantly from the true risk.

2.1 The Two-Sample Deviation

The proof begins by bounding the deviation between the empirical risks of two disjoint halves of the sample set:

$$\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\Phi_f(Z_i) - \frac{1}{n}\sum_{i=n+1}^{2n}\Phi_f(Z_i)\right| \geq \varepsilon'\right)$$

2.2 Random Permutation and Expectation

Let $\mathbf{Z} = \{Z_1, \ldots, Z_{2n}\}$ be the fixed sample set. We introduce a random permutation $\sigma \in S_{2n}$ (the symmetric group):

$$\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \Phi_f(Z_i) - \frac{1}{n} \sum_{i=n+1}^{2n} \Phi_f(Z_i) \right| \geq \varepsilon' \right)$$
$$\leq_{\mathbf{Z}} \left[\sigma \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \Phi_f(Z_{\sigma(i)}) - \frac{1}{n} \sum_{i=n+1}^{2n} \Phi_f(Z_{\sigma(i)}) \right| \geq \varepsilon' \right) \right]$$

The expectation $_{\mathbf{Z}}$ is over the samples, and $_{\sigma}$ is the probability over the random permutation.

2.3 Concentration Inequality

For a fixed function f, let $W_i = \Phi_f(Z_i) \in \{0, 1\}$. The term inside the supremum involves concentration for sampling without replacement.

$$\sigma \left(\left| \frac{1}{n} \sum_{i=1}^{n} W_{\sigma(i)} - \frac{1}{2n} \sum_{i=1}^{2n} W_{\sigma(i)} \right| \geq \varepsilon \right) \leq e^{-O(n\varepsilon^2)}.$$

2.4 Introducing the Growth Function

We use the Union Bound and the exponential decay from the concentration inequality.

Definition: Growth Function on a Sample Set

$$N^{\mathcal{F}}(Z_1, \ldots, Z_{2n}) = |\{(\Phi_f(Z_1), \ldots, \Phi_f(Z_{2n})) : f \in \mathcal{F}\}|$$

Applying the Union Bound over the $N^{\mathcal{F}}$ distinct labelings:

$$\mathbf{Z} \left[\sigma \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \Phi_f(Z_{\sigma(i)}) - \frac{1}{n} \sum_{i=n+1}^{2n} \Phi_f(Z_{\sigma(i)}) \right| \geq \varepsilon' \right) \right]$$
$$\leq_{\mathbf{Z}} \left[N^{\mathcal{F}}(Z_1, \ldots, Z_{2n}) \cdot \max_{f \in} \left\{ \sigma \left(\left| \frac{1}{n} \sum_{i=1}^{n} \Phi_f(Z_{\sigma(i)}) - \frac{1}{n} \sum_{i=n+1}^{2n} \Phi_f(Z_{\sigma(i)}) \right| \geq \varepsilon' \right) \right\} \right]$$
$$\leq_{\mathbf{Z}} \left[N^{\mathcal{F}}(Z_1, \ldots, Z_{2n}) \cdot e^{-O(n(\varepsilon')^2)} \right]$$
$$\leq \left(\max_{Z_1, \ldots, Z_{2n}} N^{\mathcal{F}}(Z_1, \ldots, Z_{2n}) \right) \cdot e^{-O(n(\varepsilon')^2)}$$

Definition: Maximum Growth Function

$$N^{\mathcal{F}}(m) = \max_{Z_1, \ldots, Z_m} N^{\mathcal{F}}(Z_1, \ldots, Z_m)$$

The bound is therefore:

$$(\ldots) \leq N^{\mathcal{F}}(2n) \cdot e^{-O(n(\varepsilon')^2)}$$

The full Symmetrization Lemma often yields a bound of $2 \cdot N^{\mathcal{F}}(2n) e^{-O(n\varepsilon^2)}$.

Step III: Bounding by VC dimension (general case)

Setup. Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ be a binary classifier class. For a sample $S = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $f \in \mathcal{F}$, denote the empirical error $\widehat{L}_S(f) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{f(x_i) \neq y_i\}$. We aim to control

$$\mathbb{P}\Big\{ \exists f \in \mathcal{F} :\ L(f) - \widehat{L}_S(f) > \varepsilon \Big\},$$

where $L(f) = \mathbb{P}\{f(X) \neq Y\}$ is the true risk under the (unknown) data distribution. By Step I (sampling trick) and Step II (symmetrization), it suffices to bound, for i.i.d. Rademacher signs σ_i,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i\, \mathbf{1}\{f(x_i) \neq y_i\} > \tfrac{\varepsilon}{2} \;\middle|\; x_{1:n}, y_{1:n} \right).$$

Intuition for VC dimension

VC dimension measures the capacity of a model class. Intuitively, if d is small relative to n,the class cannot fit arbitrary labels, which helps generalization.

Growth function and shattering. For $S = \{x_1, \ldots, x_n\}$, let $\mathcal{F}|_S = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}$ be the set of dichotomies induced by $\mathcal{F}$ on S. Define the growth function

$$\Pi_{\mathcal{F}}(n) \;=\; \max_{|S|=n} \big|\mathcal{F}|_S\big|.$$

The VC dimension $d = (\mathcal{F})$ is the largest n such that $\Pi_{\mathcal{F}}(n) = 2^n$ (i.e. some S of size n is shattered).

Uniform convergence interpretation

Equation (eq:vc-uniform-simplified) provides a probabilistic guarantee that the empirical error approximates the true error for all functions in F simultaneously.

Sauer–Shelah lemma. If $(\mathcal{F}) = d < \infty$, then for all $n \geq d$,

$$\Pi_{\mathcal{F}}(n) \;\leq\; \sum_{k=0}^{d} \binom{n}{k} \;\leq\; \left(\tfrac{en}{d}\right)^d. \tag{7}$$

From finite classes to VC classes. Condition on S. Replace the supremum over $\mathcal{F}$ by a maximum over the finite projection $\mathcal{F}|_S$ (at most $\Pi_{\mathcal{F}}(n)$ functions). Applying a Hoeffding bound and a union bound over $\mathcal{F}|_S$ yields

$$\mathbb{P}\left\{ \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i\, \mathbf{1}\{f(x_i) \neq y_i\} > \tfrac{\varepsilon}{2} \right\} \;\leq\; \Pi_{\mathcal{F}}(n)\, \exp\big(-\tfrac{n\varepsilon^2}{2}\big).$$

Taking expectation over S and combining with Steps I–II gives the uniform deviation bound

$$\mathbb{P}\Big\{ \exists f \in \mathcal{F} :\ L(f) - \widehat{L}_S(f) > \varepsilon \Big\} \ \le\ 2\,\Pi_{\mathcal{F}}(n)\, \exp\big(-\tfrac{n\varepsilon^2}{2}\big). \tag{8}$$

Using (7),

$$\mathbb{P}\left\{ \sup_{f\in\mathcal{F}} \big(L(f) - \widehat{L}_S(f)\big) > \varepsilon \right\} \ \le\ 2\Big(\tfrac{en}{d}\Big)^d \exp\big(-\tfrac{n\varepsilon^2}{2}\big). \tag{9}$$

High-probability generalization (one-sided). Equivalently, with probability at least $1-\delta$ (over the draw of S), simultaneously for all $f \in \mathcal{F}$,

$$L(f) \ \le\ \widehat{L}_S(f) \ +\ \sqrt{\frac{2}{n}\Big(d\log\frac{en}{d} + \log\frac{2}{\delta}\Big)}. \tag{10}$$

A symmetric bound holds for $\widehat{L}_S(f) - L(f)$, which yields a two-sided deviation. Implications for model selection

- Choose model classes with VC dimension not too large to avoid overfitting.
- For fixed n, increasing d increases the uniform deviation bound.
- For fixed d, increasing n decreases the bound (more samples help).

Sample complexity (realizable case). If there exists $f^\star \in \mathcal{F}$ with $L(f^\star) = 0$, then choosing

$$n \ \gtrsim\ \frac{1}{\varepsilon}\Big(d\log\tfrac{1}{\varepsilon} + \log\tfrac{1}{\delta}\Big)$$

ensures that Empirical Risk Minimization achieves $L(\hat{f}) = O(\varepsilon)$ with probability at least $1-\delta$. In the agnostic case, the dependence becomes

$$n \ \gtrsim\ \frac{1}{\varepsilon^2}\Big(d\log\tfrac{1}{\varepsilon} + \log\tfrac{1}{\delta}\Big).$$

Sample complexity summary

Realizable case: fewer samples needed for uniform convergence.

Agnostic case: more samples required, scales as $1/\epsilon^2$.

Takeaway. Step III replaces the (possibly infinite) class $\mathcal{F}$ by the finite projection $\mathcal{F}|_S$ and controls its cardinality via the VC dimension through Sauer–Shelah. This yields uniform convergence rates that scale with $d = (\mathcal{F})$.

VC dimension helps determine how much data is needed relative to model complexity.

Combining VC-based bounds with regularization or data augmentation further improves generalization.

5.2 Proof of Inequality (5.8)

We begin by examining a particular scenario—Based on our initial assumptions, when considering $d+1$ components, there exists a configuration

$$(\phi(x_1), \phi(x_2), \ldots, \phi(x_n))$$

that cannot be realized.
In the special case, we assume that:

$$\begin{aligned} &\forall i_1, i_2, ..., i_{d+1} \in \{1, 2, ..., n\}, i_j \neq i_k (j \neq k), \\ &\quad (0, 0, ..., 0) \neq (\phi(x_{i_1}), \phi(x_{i_2}), ..., \phi(x_{i_{d+1}})) \end{aligned}$$

which means configurations containing $d+1$ zeros are unattainable. This implies that the maximum number of zeros permissible in the equation is d. The total count of possible value assignments with no more than d zeros is given by

$$\sum_{k=0}^{d} \binom{n}{k}$$

So in the special case,

$$|\{(\phi(x_1), \phi(x_2), ..., \phi(x_n)), \phi \in \Phi\}| \leq \sum_{k=0}^{d} \binom{n}{k} = O(n^d),\ n > d$$

Since specific cases are constrained, we explore the transformation of general cases into these special cases. The comprehensive proof follows.

Proof: We commence by enumerating all unrealizable configurations and analyzing the behavior at the first component.

Three distinct scenarios emerge:

Zero as the first component:

$$\begin{cases} 0, *, 1, \cdots & \text{(n bits)} \\ 0, 1, *, \cdots & \cdots \\ 0, 0, *, \cdots & \cdots \end{cases}$$

One as the first component:

$$\begin{cases} 1, *, 1, \cdots & \cdots \\ 1, 1, *, \cdots & \cdots \\ 1, 0, *, \cdots & \cdots \end{cases}$$

No restriction on the first component:

$$\begin{cases} *, *, 1, \cdots & \cdots \\ *, 1, *, \cdots & \cdots \\ *, 0, *, \cdots & \cdots \end{cases}$$

When we convert a 1 in the first component to 0, we observe a reduction in all possible configurations.

This pattern persists when applying the same transformation to any component—converting 1 to 0 consistently diminishes the space of possibilities.

Consequently, if we systematically convert all 1's to 0's across every component, all configurations converge to the special case where precisely $d+1$ zeros cannot be achieved. Therefore, the number of attainable configurations in the general case must exceed that of the restricted special case.

In this specialized scenario, the count of zeros can range from 0 to d. Thus, we establish $N^*(n) \leq \sum_{k=0}^{d} \binom{n}{k}$

$$\sum_{k=0}^{d} \binom{n}{k} \leq \left(\frac{en}{d}\right)^d$$

Applying the Chernoff bound under the assumption $d < \frac{n}{2}$.