1 PAC Learning Framework

Definition 1.1 (Generalization error)**.** Given a hypothesis $h \in \mathcal{H}$, a target concept $c \in \mathcal{C}$, and an underlying distribution $\mathcal{D}$, the generalization error or risk of h is defined by

$$R(h) = \Pr_{x \sim \mathcal{D}}[h(x) \neq c(x)] = \mathbb{E}_{x \sim \mathcal{D}}\Big[\mathbf{1}_{h(x) \neq c(x)}\Big],$$

where $\mathbf{1}_\omega$ is the indicator function of the event ω.

Definition 1.2 (Empirical error)**.** Given a hypothesis $h \in \mathcal{H}$, a target concept $c \in \mathcal{C}$, and a sample $S = (x_1, \ldots, x_m)$, the empirical error or empirical risk of h is defined by

$$\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{h(x_i) \neq c(x_i)}.$$

Definition 1.3 (PAC-learning)**.** A concept class $\mathcal{C}$ is said to be PAC-learnable *if there exists an algorithm A and a polynomial function* $\mathrm{poly}(\cdot, \cdot, \cdot, \cdot)$ *such that for any $\epsilon > 0$ and $\delta > 0$, for all distributions $\mathcal{D}$ on $\mathcal{X}$ and for any target concept $c \in \mathcal{C}$, the following holds for any sample size $m \geq \mathrm{poly}(1/\epsilon, 1/\delta, n, \mathrm{size}(c))$:*

$$\Pr_{S \sim \mathcal{D}^m}[R(h_S) \leq \epsilon] \geq 1 - \delta.$$

If A further runs in $\mathrm{poly}(1/\epsilon, 1/\delta, n, \mathrm{size}(c))$, then $\mathcal{C}$ is said to be *efficiently PAC-learnable.* When such an algorithm A exists, it is called a PAC-learning algorithm for $\mathcal{C}$.

Theorem 1.4 (Learning bound — finite $\mathcal{H}$, consistent case)**.** *Let $\mathcal{H}$ be a finite set of functions mapping from X to Y. Let A be an algorithm that for any target concept $c \in \mathcal{H}$ and i.i.d. sample S returns a consistent hypothesis h_S: $\hat{R}_S(h_S) = 0$. Then, for any $\epsilon, \delta > 0$, the inequality*

$$\Pr_{S \sim \mathcal{D}^m}\Big[R(h_S) \leq \epsilon\Big] \geq 1 - \delta$$

holds if

$$m \geq \frac{1}{\epsilon}\left(\log |\mathcal{H}| + \log \frac{1}{\delta}\right). \tag{2.8}$$

This sample complexity result admits the following equivalent statement as a generalization bound: for any $\epsilon, \delta > 0$, with probability at least $1 - \delta$,

$$R(h_S) \leq \frac{1}{m}\left(\log |\mathcal{H}| + \log \frac{1}{\delta}\right). \tag{2.9}$$

Corollary 1.5. *Fix $\epsilon > 0$. Then, for any hypothesis $h : X \to \{0, 1\}$, the following inequalities hold:*

$$\Pr_{S \sim \mathcal{D}^m}\Big[\hat{R}_S(h) - R(h) \geq \epsilon\Big] \leq \exp(-2m\epsilon^2). \tag{2.14}$$

$$\Pr_{S \sim \mathcal{D}^m}\Big[\hat{R}_S(h) - R(h) \leq -\epsilon\Big] \leq \exp(-2m\epsilon^2). \tag{2.15}$$

By the union bound, this implies the following two-sided inequality:

$$\Pr_{S \sim \mathcal{D}^m}\Big[\,|\hat{R}_S(h) - R(h)| \geq \epsilon\,\Big] \leq 2\exp(-2m\epsilon^2). \tag{2.16}$$

Corollary 1.6 (Generalization bound — single hypothesis)**.** *Fix a hypothesis $h : X \to \{0, 1\}$. Then, for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$:*

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{2.17}$$

Theorem 1.7 (Learning bound — finite $\mathcal{H}$, inconsistent case)**.** *Let $\mathcal{H}$ be a finite hypothesis set. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds:*

$$\forall h \in \mathcal{H}, \quad R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{2}{\delta}}{2m}}. \tag{2.20}$$

Definition 1.8 (Agnostic PAC-learning)**.** Let $\mathcal{H}$ be a hypothesis set. An algorithm A is an agnostic PAC-learning algorithm if there exists a polynomial function $\text{poly}(\cdot, \cdot, \cdot, \cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, the following holds for any sample size $m \geq \text{poly}(1/\epsilon, 1/\delta, n, \text{size}(c))$:

$$\Pr_{S \sim \mathcal{D}^m}\left[R(h_S) - \min_{h \in \mathcal{H}} R(h) \leq \epsilon\right] \geq 1 - \delta. \tag{2.21}$$

If A further runs in $\text{poly}(1/\epsilon, 1/\delta, n)$, then it is said to be an efficient agnostic PAC-learning algorithm.

Definition 1.9 (Bayes error)**.** Given a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, the Bayes error R^* is defined as the infimum of the errors achieved by measurable functions $h : \mathcal{X} \to \mathcal{Y}$:

$$R^* = \inf_{h \text{ measurable}} R(h). \tag{2.22}$$

A hypothesis h with $R(h) = R^*$ is called a *Bayes hypothesis* or *Bayes classifier*.

By definition, in the deterministic case, we have $R^* = 0$, but, in the stochastic case, $R^* \neq 0$. Clearly, the Bayes classifier h_{Bayes} can be defined in terms of the conditional probabilities as:

$$\forall x \in \mathcal{X}, \quad h_{\text{Bayes}}(x) = \arg\max_{y \in \{0,1\}} \Pr[y \mid x]. \tag{2.23}$$

The average error made by the bayes hypothesis on $x \in \mathcal{X}$ is thus $\min\{\Pr[0|x], \Pr[1|x]\}$, and this is the minimum possible error.

Definition 1.10 (Noise)**.** Given a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, the noise at point $x \in \mathcal{X}$ is defined by

$$\text{noise}(x) = \min\{\Pr[1 \mid x], \Pr[0 \mid x]\}. \tag{2.24}$$

The average noise or the noise associated to $\mathcal{D}$ is $\mathbb{E}[\text{noise}(x)]$.

Thus, the average noise is precisely the Bayes error:

$$\text{noise} = \mathbb{E}[\text{noise}(x)] = R^*.$$

The noise is a characteristic of the learning task indicative of its level of difficulty. A point $x \in \mathcal{X}$, for which $\text{noise}(x)$ is close to $1/2$, is sometimes referred to as *noisy*, and is of course a challenge for accurate prediction.

2 Rademacher Complexity and VC-Dimension

Definition 2.1 (Empirical Rademacher complexity)**.** Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[a, b]$ and $S = (z_1, \ldots, z_m)$ a fixed sample of size m with elements in $\mathcal{Z}$. Then, the empirical Rademacher complexity of $\mathcal{G}$ with respect to the sample S is defined as:

$$\hat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_\sigma\left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i)\right], \tag{3.1}$$

where $\sigma = (\sigma_1, \ldots, \sigma_m)^\top$, with σ_i independent uniform random variables taking values in $\{-1, +1\}$. The random variables σ_i are called *Rademacher variables*.

Let $\mathbf{g}_S$ denote the vector of values taken by function g over the sample S: $\mathbf{g}_S = (g(z_1), \ldots, g(z_m))^\top$. Then, the empirical Rademacher complexity can be rewritten as

$$\hat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{\sigma \cdot \mathbf{g}_S}{m} \right].$$

The inner product $\sigma \cdot \mathbf{g}_S$ measures the correlation of $\mathbf{g}_S$ with the vector of random noise σ. The supremum $\sup_{g \in \mathcal{G}} \frac{\sigma \cdot \mathbf{g}_S}{m}$ is a measure of how well the function class $\mathcal{G}$ correlates with σ over the sample S. Thus, the empirical Rademacher complexity measures on average how well the function class $\mathcal{G}$ correlates with random noise on S. This describes the richness of the family $\mathcal{G}$: richer or more complex families $\mathcal{G}$ can generate more vectors $\mathbf{g}_S$ and thus better correlate with random noise, on average.

Definition 2.2 (Rademacher complexity)**.** Let $\mathcal{D}$ denote the distribution according to which samples are drawn. For any integer $m \geq 1$, the Rademacher complexity of $\mathcal{G}$ is the expectation of the empirical Rademacher complexity over all samples of size m drawn according to $\mathcal{D}$:

$$\mathfrak{R}_m(\mathcal{G}) = \mathbb{E}_{S \sim \mathcal{D}^m} \left[\hat{\mathfrak{R}}_S(\mathcal{G}) \right]. \tag{3.2}$$

Theorem 2.3. *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[0, 1]$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m, each of the following holds for all $g \in \mathcal{G}$:*

$$\mathbb{E}[g(z)] \;\leq\; \frac{1}{m} \sum_{i=1}^{m} g(z_i) + 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log(1/\delta)}{2m}}, \tag{3.3}$$

and

$$\mathbb{E}[g(z)] \;\leq\; \frac{1}{m} \sum_{i=1}^{m} g(z_i) + 2\hat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log(2/\delta)}{2m}}. \tag{3.4}$$

Theorem 2.4 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_m$ be independent random variables with X_i taking values in $[a_i, b_i]$ for all $i \in [m]$. Then, for any $\epsilon > 0$, the following inequalities hold for*

$$S_m = \sum_{i=1}^{m} X_i :$$

$$\Pr\Big[S_m - \mathbb{E}[S_m] \geq \epsilon \Big] \;\leq\; \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{m} (b_i - a_i)^2} \right), \tag{D.4}$$

$$\Pr\Big[S_m - \mathbb{E}[S_m] \leq -\epsilon \Big] \;\leq\; \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{m} (b_i - a_i)^2} \right). \tag{D.5}$$

Theorem 2.5 (McDiarmid's inequality)**.** *Let $X_1, \ldots, X_m \in \mathcal{X}^m$ be a set of $m \geq 1$ independent random variables and assume that there exist $c_1, \ldots, c_m > 0$ such that $f : \mathcal{X}^m \to \mathbb{R}$ satisfies the following conditions:*

$$\Big| f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m) \Big| \leq c_i, \quad \forall i \in [m], \;\; \forall x_1, \ldots, x_m, x_i' \in \mathcal{X}. \tag{D.15}$$

Let $f(S)$ denote $f(X_1, \ldots, X_m)$. Then, for all $\epsilon > 0$, the following inequalities hold:

$$\Pr\Big[f(S) - \mathbb{E}[f(S)] \geq \epsilon\Big] \;\leq\; \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right), \tag{D.16}$$

$$\Pr\Big[f(S) - \mathbb{E}[f(S)] \leq -\epsilon\Big] \;\leq\; \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right). \tag{D.17}$$

Lemma 2.6. *Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ and let $\mathcal{G}$ be the family of loss functions associated to $\mathcal{H}$ for the zero-one loss:*

$$\mathcal{G} = \{(x, y) \mapsto \mathbf{1}_{h(x) \neq y} : h \in \mathcal{H}\}.$$

For any sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ of elements in $\mathcal{X} \times \{-1, +1\}$, let S_X denote its projection over $\mathcal{X}$: $S_X = (x_1, \ldots, x_m)$. Then, the following relation holds between the empirical Rademacher complexities of $\mathcal{G}$ and $\mathcal{H}$:

$$\hat{\mathfrak{R}}_S(\mathcal{G}) = \tfrac{1}{2}\,\hat{\mathfrak{R}}_{S_X}(\mathcal{H}). \tag{3.16}$$

Theorem 2.7 (Rademacher complexity bounds – binary classification)**.** *Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ and let $\mathcal{D}$ be the distribution over the input space $\mathcal{X}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over a sample S of size m drawn according to $\mathcal{D}$, each of the following holds for any $h \in \mathcal{H}$:*

$$R(h) \;\leq\; \hat{R}_S(h) + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}, \tag{3.17}$$

and

$$R(h) \;\leq\; \hat{R}_S(h) + \hat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{3.18}$$

Definition 2.8 (Growth function)**.** The growth function $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ for a hypothesis set $\mathcal{H}$ is defined by:

$$\forall m \in \mathbb{N}, \quad \Pi_{\mathcal{H}}(m) = \max_{\{x_1, \ldots, x_m\} \subseteq \mathcal{X}} \left|\{(h(x_1), \ldots, h(x_m)) : h \in \mathcal{H}\}\right|. \tag{3.19}$$

In other words, $\Pi_{\mathcal{H}}(m)$ is the maximum number of distinct ways in which m points can be classified using hypotheses in $\mathcal{H}$. Each one of these distinct classifications is called a *dichotomy* and, thus, the growth function counts the number of dichotomies that are realized by the hypothesis. This provides another measure of the richness of the hypothesis set $\mathcal{H}$. However, unlike the Rademacher complexity, this measure does not depend on the distribution; it is purely combinatorial.

Theorem 2.9 (Massart's lemma)**.** *Let $\mathcal{A} \subseteq \mathbb{R}^m$ be a finite set, with*

$$r = \max_{\mathbf{x} \in \mathcal{A}} \|\mathbf{x}\|_2,$$

then the following holds:

$$\mathbb{E}_\sigma\left[\frac{1}{m} \sup_{\mathbf{x} \in \mathcal{A}} \sum_{i=1}^m \sigma_i x_i\right] \leq \frac{r\sqrt{2 \log |\mathcal{A}|}}{m}, \tag{3.20}$$

where σ_is are independent uniform random variables taking values in $\{-1, +1\}$ and $x_1, \ldots, x_m$ are the components of vector $\mathbf{x}$.

Corollary 2.10. *Let $\mathcal{G}$ be a family of functions taking values in $\{-1, +1\}$. Then the following holds:*

$$\mathfrak{R}_m(\mathcal{G}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{G}}(m)}{m}}. \tag{3.21}$$

Corollary 2.11 (Growth function generalization bound)**.** *Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in \mathcal{H}$,*

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(m)}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \tag{3.22}$$

Definition 2.12 (shattering)**.** A set S of at least one point is said to be shattered by a hypothesis set $\mathcal{H}$ when $\mathcal{H}$ realizes all possible dichotomies of S, i.e. $\Pi_{\mathcal{H}}(m) = 2^m$.

Definition 2.13 (VC-dimension)**.** The VC-dimension of a hypothesis set $\mathcal{H}$ is the size of the largest set that can be shattered by $\mathcal{H}$:

$$\mathrm{VCdim}(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\}. \tag{3.24}$$

Note that, by definition, if $\mathrm{VCdim}(\mathcal{H}) = d$, there exists a set of size d that can be shattered. However, this does not imply that all sets of size d or less are shattered and, in fact, this is typically not the case.

Theorem 2.14 (Radon's theorem)**.** *Any set X of $d + 2$ points in $\mathbb{R}^d$ can be partitioned into two subsets X_1 and X_2 such that the convex hulls of X_1 and X_2 intersect.*

Theorem 2.15 (Sauer's lemma)**.** *Let $\mathcal{H}$ be a hypothesis set with $\mathrm{VCdim}(\mathcal{H}) = d$. Then, for all $m \in \mathbb{N}$, the following inequality holds:*

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}.$$

Corollary 2.16. *Let $\mathcal{H}$ be a hypothesis set with $\mathrm{VCdim}(\mathcal{H}) = d$. Then for all $m \geq d$,*

$$\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d = O(m^d).$$

Corollary 2.17 (VC-dimension generalization bounds)**.** *Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ with VC-dimension d. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in \mathcal{H}$:*

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \tag{3.29}$$

Thus, the form of this generalization bound is

$$R(h) \leq \hat{R}_S(h) + O\left(\sqrt{\frac{\log(m/d)}{(m/d)}}\right). \tag{3.30}$$

Theorem 2.18 (Lower bound, realizable case)**.** *Let $\mathcal{H}$ be a hypothesis set with VC-dimension $d > 1$. Then, for any $m \geq 1$ and any learning algorithm A, there exist a distribution $\mathcal{D}$ over $\mathcal{X}$ and a target function $f \in \mathcal{H}$ such that*

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left[R_{\mathcal{D}}(h_S, f) > \frac{d-1}{32m}\right] \geq \frac{1}{100}. \tag{3.31}$$

Lemma 2.19 (Lemma 3.21). *Let α be a uniformly distributed random variable taking values in $\{\alpha_-, \alpha_+\}$, where $\alpha_- = \frac{1}{2} - \frac{\epsilon}{2}$ and $\alpha_+ = \frac{1}{2} + \frac{\epsilon}{2}$, and let S be a sample of $m \geq 1$ random variables $X_1, \ldots, X_m$ taking values in $\{0, 1\}$ and drawn i.i.d. according to the distribution $\mathcal{D}_\alpha$ defined by $\mathbb{P}_{\mathcal{D}_\alpha}[X = 1] = \alpha$. Let h be a function from X^m to $\{\alpha_-, \alpha_+\}$. Then, the following holds:*

$$\mathbb{E}_\alpha \left[\mathbb{P}_{S \sim \mathcal{D}_\alpha^m} \Big[h(S) \neq \alpha \Big] \right] \geq \Phi(2\lceil m/2 \rceil, \epsilon),$$

where

$$\Phi(m, \epsilon) = \tfrac{1}{4} \left(1 - \sqrt{1 - \exp\left(- \tfrac{m\epsilon^2}{1-\epsilon^2} \right)} \right), \quad \textit{for all } m \textit{ and } \epsilon.$$

Lemma 2.20 (Lemma 3.22). *Let Z be a random variable taking values in $[0, 1]$. Then, for any $\gamma \in [0, 1)$,*

$$\mathbb{P}[Z > \gamma] \;\geq\; \frac{\mathbb{E}[Z] - \gamma}{1 - \gamma} \;>\; \mathbb{E}[Z] - \gamma.$$

Theorem 2.21 (Theorem 3.23, Lower bound, non-realizable case). *Let $\mathcal{H}$ be a hypothesis set with VC-dimension $d > 1$. Then, for any $m \geq 1$ and any learning algorithm A, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$ such that:*

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left[R_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} R_{\mathcal{D}}(h) > \sqrt{\tfrac{d}{320m}} \right] \;\geq\; \tfrac{1}{64}.$$

Equivalently, for any learning algorithm, the sample complexity verifies

$$m \;\geq\; \frac{d}{320\epsilon^2}.$$

3 Model Selection

3.1 Convex surrogate losses

Problem: Solving ERM optimization problem is NP-hard since the 0-1 loss function is not convex. Solution: Use alternative convex surrogate losses that upper bounds the 0-1 loss. For $h : \mathcal{X} \to \mathbb{R}$, define following binary classifier:

$$f_h(x) = \begin{cases} +1 & \text{if } h(x) \geq 0, \\ -1 & \text{if } h(x) < 0. \end{cases}$$

We have

$$1_{f_h(x) \neq y} = 1_{yh(x) < 0} + 1_{h(x) = 0 \wedge y = -1} \leq 1_{yh(x) \leq 0}.$$

For any $x \in \mathcal{X}$, let $\eta(x) = \mathbb{P}[y = +1 \mid x]$ and let $\mathcal{D}_\mathcal{X}$ denote the marginal distribution over $\mathcal{X}$. Then, for any h, we can write

$$\begin{aligned} R(h) &= \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}} \Big[\eta(x) 1_{h(x) < 0} + (1 - \eta(x)) 1_{h(x) > 0} + (1 - \eta(x)) 1_{h(x) = 0} \Big] \\ &= \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}} \Big[\eta(x) 1_{h(x) < 0} + (1 - \eta(x)) 1_{h(x) \geq 0} \Big]. \end{aligned}$$

In view of that, the Bayes classifier can be defined as assigning label $+1$ to x when $\eta(x) \geq \frac{1}{2}$, and -1 otherwise. It can therefore be induced by the function h^* defined by

$$h^*(x) = \eta(x) - \tfrac{1}{2}. \tag{4.9}$$

We will refer to $h^* : \mathcal{X} \to \mathbb{R}$ as the *Bayes scoring function* and will denote by R^* the Bayes error:

$$R^* = R(h^*).$$

Lemma 3.1 (Lemma 4.5)**.** *The excess error of any hypothesis $h : \mathcal{X} \to \mathbb{R}$ can be expressed as follows in terms of η and the Bayes scoring function h^*:*

$$R(h) - R^* \;=\; 2\,\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\Big[\,|h^*(x)|\,1_{\,h(x)h^*(x)\leq 0}\,\Big].$$

$|h^*(x)|$接近0时说明贝叶斯分类器也不确定x的分类，因此这时候如果h分类与h^*不一致，对excess error的贡献小，反之大。因为在二分类中，错误概率和正确概率是互补的，差值计算时会出现一个 2 倍的加权项。