Correlation between Generalization Error and the Four Factors

1 Preliminaries

1.1 Description of Symbols

Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ a set of classes. We assume that the training and test samples are drawn *i.i.d* according to some distributions $\mathcal{D}^{tr}$ and $\mathcal{D}^{te}$ over $\mathcal{X} \times \mathcal{Y}$. The training set is denoted as $T = \{\boldsymbol{x}, y\} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ that contains n training samples, where $\boldsymbol{x}_i$ denotes the i-th sample's feature, and y_i is the associated label. Let d_i and $w\left(d_i\right)$ be the learning difficulty and the difficulty-based weight of $\boldsymbol{x}_i$.

The predictor is denoted by $f\left(\boldsymbol{\theta}, \boldsymbol{x}\right)$ and $\mathcal{F} = \{f\left(\boldsymbol{\theta}, \cdot\right) | \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^{\mathrm{d}}\}$. For the sake of notation, we focus on the binary setting $y_i \in \{-1, 1\}$ with $f\left(\boldsymbol{\theta}, \boldsymbol{x}\right) \in \mathbb{R}$. The sign of the model's output $f\left(\boldsymbol{\theta}, \boldsymbol{x}_i\right)$ is the predicted label. However, as to be clarified later, our results can be easily extended to the multi-class setting where $y_i \in \{1, 2, \cdots, C\}$. The softmax function is used, and the logits are given by $\{f_{y_j}\left(\boldsymbol{\theta}, \boldsymbol{x}\right)\}_{j=1}^C$. Given a non-negative loss ℓ and a classifier $f\left(\boldsymbol{\theta}, \cdot\right)$, the empirical risk can be expressed as $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n w\left(d_i\right) \cdot \ell\left(y_i f\left(\boldsymbol{\theta}, \boldsymbol{x}_i\right)\right)$. We focus particularly on the exponential loss $\ell\left(u\right) = \exp\left(-u\right)$ and logistic loss $\ell\left(u\right) = \log\left(1 + \exp\left(-u\right)\right)$. Let $\nabla l(u)$ be the loss gradient and $f\left(\boldsymbol{x}|T\right)$ is the trained model on T. The margin is denoted as $\gamma_i(T) = y_i f\left(\boldsymbol{\theta}, \boldsymbol{x}_i|T\right)$ for the binary setting, where it is denoted as $\gamma_i(T) = f_{y_i}\left(\boldsymbol{\theta}, \boldsymbol{x}_i|T\right) - \max_{i \neq j} f_{y_j}\left(\boldsymbol{\theta}, \boldsymbol{x}_i|T\right)$ for the multi-class setting.

1.2 Definition of the Generalization Error

Bias-variance tradeoff is a basic theory for the qualitative analysis of the generalization error in machine learning [Heskes, 1998]. This tradeoff is initially constructed via regression and mean square error, which is given by

$$\begin{aligned} Err &= E_{\boldsymbol{x},y}E_T[||y - f(\boldsymbol{x}|T)||_2^2] \\ &\approx \underbrace{E_{\boldsymbol{x},y}[||y - \overline{f}(\boldsymbol{x})||_2^2]}_{Bias} + \underbrace{E_{\boldsymbol{x},y}E_T[||f(\boldsymbol{x}|T) - \overline{f}(x)||_2^2]}_{Variance}, \end{aligned} \tag{1}$$

where $\overline{f}\left(\boldsymbol{x}\right) = E_T\left[f\left(\boldsymbol{x}|T\right)\right]$. Similarly, we define the generalization error of a single sample $\boldsymbol{x}_i$ as

$$\mathrm{err}_i = E_T\left[\ell\left(f\left(\boldsymbol{x}_i|T\right), y_i\right)\right] \approx B\left(\boldsymbol{x}_i\right) + V\left(\boldsymbol{x}_i\right), \tag{2}$$

where $B\left(\boldsymbol{x}_i\right)$ and $V(\boldsymbol{x}_i)$ are the bias and variance of $\boldsymbol{x}_i$.

1.3 Pre-Defined Definitions and Assumptions

A minimal set of assumptions for providing our theoretical results for a-homogeneous neural networks on the classification tasks are identified. Two assumptions are considered.

- The loss ℓ has an exponential tail. Following Lyu and Li [2019], there is a general definition of the exponential loss, where $\ell\left(u\right) = \exp\left(-f\left(u\right)\right)$,
  - f is smooth and $f'\left(u\right) \geq 0, \forall u$,
  - there exists $c > 0$ such that $f'\left(u\right)u$ is non-decreasing for $u > c$ and $f'\left(u\right)u \to \infty$ as $u \to \infty$.

  It is easy to verify that losses including the exponential loss, log loss, and cross-entropy loss satisfy the definition. Thus, $\lim_{u\to\infty} \ell(-u) = \lim_{u\to\infty} \nabla\ell(-u) = 0$.
- For $\forall \boldsymbol{x} \in \mathcal{X}$, $f(\boldsymbol{\theta}, \boldsymbol{x})$ is β-smooth and l-Lipschitz on $\mathbb{R}^{\mathrm{d}}$.

The second assumption provides certain regularities from $f(\boldsymbol{\theta}, \boldsymbol{x})$ to ensure the existence of critical points and the convergence of gradient descent. The gradient descent process is denoted as $\boldsymbol{\theta}_{t+1}\left(\boldsymbol{w}\right) = \boldsymbol{\theta}_t\left(\boldsymbol{w}\right) - \eta_t \nabla\mathcal{L}\left(\boldsymbol{\theta}_t\left[\boldsymbol{w}(\boldsymbol{d}\left[t\right])\right]\right)$,

where η_t is the learning rate which can be a constant or step-independent, $\nabla\mathcal{L}\left(\boldsymbol{\theta}_t\left[\boldsymbol{w}(\boldsymbol{d}\left[t\right])\right]\right)$ is the gradient of $\mathcal{L}$, and $\boldsymbol{w}(\boldsymbol{d}\left[t\right])$ is the difficulty-based weight of difficulty $\boldsymbol{d}$ at time t. The weight may change with time t if difficulty measures, such as loss and predicted probability, are used. The generalization error of the test set $\hat{\mathcal{L}}\left(f\right)$ is defined as

$$\hat{\mathcal{L}}\left(f\right) = \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}^{te}}[\gamma(f\left(\boldsymbol{x}, y\right)) \leq 0]. \tag{3}$$

2 Correlation between Generalization Error and the Four Factors

The generalization error is widely used in the theoretical analysis of machine learning. We analyze the connection between the generalization error and the four factors influencing the samples' difficulty. The exponential loss is adopted, and the positive samples are taken as the examples in the subsequent discussion. Without increasing the ambiguity, the generalization error of the samples is termed as error for brevity.

Take the exponential loss $\ell = \exp(-y_i f(\boldsymbol{x}_i))$ as an example in the subsequent analyses. Let T be a random training set from some distributions over $\mathcal{X} \times \mathcal{Y}$ and let $f\left(\cdot|T\right)$ be the trained model on T. The generalization error of sample $\boldsymbol{x}_i$ is

$$\begin{aligned} \mathrm{err}_i &= E_T\left[\ell\left(f\left(\boldsymbol{x}_i|T\right), y_i\right)\right] \\ &= \int_{T\in\mathcal{X}\times\mathcal{Y}} \exp(-y_i f\left(\boldsymbol{x}_i|T\right)) dP\left(T\right). \end{aligned} \tag{4}$$

For the sake of notation, we focus on the binary setting $y \in \{-1, 1\}$. The positive samples are taken as examples in the succeeding discussion.

2.1 Noise Factor

A number of studies consider noisy samples as hard ones [Castells *et al.*, 2020; Shin *et al.*, 2020]. The two kinds of noise are feature noise and label noise.

Feature Noise

For feature noise, we offer the following proposition:

Proposition 1. *Let $\Delta\boldsymbol{x}_i$ be the perturbation of sample $(\boldsymbol{x}_i, y_i)$, which is extremely small in that $o(\Delta\boldsymbol{x}_i)$ can be omitted. Let $\angle\varphi$ be the angle between the direction of $\Delta\boldsymbol{x}_i$ and the direction of $E_T[f'\left(\boldsymbol{x}_i|T\right)]$. If $E_T[f'\left(\boldsymbol{x}_i|T\right) \cdot \Delta\boldsymbol{x}_i] < 0$ (i.e., $\angle\varphi > 90°$), then the error of the noisy sample is increased relative to the clean one. Alternatively, the direction of the perturbation $\Delta\boldsymbol{x}_i$ and that of $E_T[f'\left(\boldsymbol{x}_i|T\right)]$ are contradictory. Otherwise, if $E_T[f'\left(\boldsymbol{x}_i|T\right) \cdot \Delta\boldsymbol{x}_i] > 0$, then $\angle\varphi < 90°$, and the error of the noisy sample is decreased.*

Proof. According to the definition of the generalization error, the error of a clean positive sample is

$$\mathrm{err}_i = E_T\left[\ell\left(f\left(\boldsymbol{x}_i|T\right), y_i\right)\right] = \int_{T\in\mathcal{X}\times\mathcal{Y}} \exp(-f\left(\boldsymbol{x}_i|T\right))dP\left(T\right). \tag{5}$$

Denote the perturbation of $\boldsymbol{x}_i$ as $\Delta\boldsymbol{x}_i$, which is usually sufficient small. After adding feature noise, the sample becomes $(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i, y_i = 1)$. Its error is

$$\begin{aligned}\mathrm{err}_{i'} &= E_T\left[\ell\left(f\left(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T\right), y_i\right)\right] \\ &= \int_{T\in\mathcal{X}\times\mathcal{Y}} \exp(-f\left(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T\right))dP\left(T\right).\end{aligned} \tag{6}$$

Based on the definition of the exponential-tail loss, the Taylor expansion of f can be adopted. Let $f'(\boldsymbol{x}_i|T)$ denote the first-order derivative. As we mainly concern about the direction of $f'(\boldsymbol{x}_i|T)$ and the perturbation $\Delta\boldsymbol{x}_i$ is small, the first-order Taylor expansion can be adopted. Thus, we have

$$f\left(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T\right) = f\left(\boldsymbol{x}_i|T\right) + f'\left(\boldsymbol{x}_i|T\right)\cdot\Delta\boldsymbol{x}_i + o\left(\Delta\boldsymbol{x}_i\right), \tag{7}$$

Here, $f(\boldsymbol{x}_i)$ is the output of the sigmoid layer, i.e., $f(\boldsymbol{x}_i) \in (0,1)$. Applying the first-order Taylor expansion on $\exp(-x)$, we yield $\exp(-x) = 1 - x + R(x)$. Then, the generalization error turns to

$$\begin{aligned}\mathrm{err}_i &= E_T\left[\exp(-f\left(\boldsymbol{x}_i|T\right))\right] \\ &= E_T\left[1 - f\left(\boldsymbol{x}_i|T\right) + R(f\left(\boldsymbol{x}_i|T\right))\right].\end{aligned} \tag{8}$$

After adding feature noise, its generalization error becomes

$$\begin{aligned}\mathrm{err}_{i'} &= E_T\left[\exp(-f\left(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T\right))\right] \\ &= E_T\left[1 - f\left(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T\right) + R(f\left(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T\right))\right].\end{aligned} \tag{9}$$

To compare the generalization error of the clean sample and the feature-noised sample, we separately study the first two terms and the residual term of the Taylor expansion shown in Formula (9).

$$\begin{aligned}&\mathrm{err}_i - \mathrm{err}_{i'} \\ =&E_T[1 - f(\boldsymbol{x}_i|T) + R(f(\boldsymbol{x}_i|T))] \\ &- E_T[1 - f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T) + R(f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T))] \\ =&E_T[1 - f(\boldsymbol{x}_i|T) - 1 + f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T) \\ &+ R(f(\boldsymbol{x}_i|T)) - R(f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T))] \\ =&E_T[1 - f(\boldsymbol{x}_i|T) - 1 + f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T)] \\ &+ E_T[R(f(\boldsymbol{x}_i|T)) - R(f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T))].\end{aligned} \tag{10}$$

For the first two terms,

$$\begin{aligned}&E_T[1 - f\left(\boldsymbol{x}_i|T\right) - 1 + f\left(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T\right)] \\ &= E_T[1 - f\left(\boldsymbol{x}_i|T\right) - 1 + f\left(\boldsymbol{x}_i|T\right) + f'\left(\boldsymbol{x}_i|T\right)\cdot\Delta\boldsymbol{x}_i + o(\Delta\boldsymbol{x}_i)] \\ &= E_T[f'\left(\boldsymbol{x}_i|T\right)\cdot\Delta\boldsymbol{x}_i + o(\Delta\boldsymbol{x}_i)].\end{aligned} \tag{11}$$

Comparing the residual term $R(x) = \exp(-x) + x - 1$,

$$\begin{aligned}&E_T[R(f(\boldsymbol{x}_i|T)) - R(f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T))] \\ &= E_T[\exp(-f\left(\boldsymbol{x}_i|T\right)) + f\left(\boldsymbol{x}_i|T\right)] \\ &- E_T[\exp(-f\left(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T\right)) + f\left(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T\right)] \\ &= E_T[-f'\left(\boldsymbol{x}_i|T\right)\Delta\boldsymbol{x}_i - o(\Delta\boldsymbol{x}_i)] \\ &+ E_T[\exp(-f\left(\boldsymbol{x}_i|T\right)) - \exp(-f\left(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T\right))].\end{aligned} \tag{12}$$

When $x \in (0,1)$, considering the relationship between the two functions $y = \exp(-x)$ and $y = -x + 1 + \frac{1}{e}$, $\exp(-f(\boldsymbol{x}_i|T)) - \exp(-f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T))$ can be bounded.

For the upper bound, when $f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T) \geq f(\boldsymbol{x}_i|T)$,

$$\begin{aligned}\exp(-f(\boldsymbol{x}_i|T)) &- \exp(-f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T)) \\ &\leq f'\left(\boldsymbol{x}_i|T\right)\cdot\Delta\boldsymbol{x}_i + o(\Delta\boldsymbol{x}_i);\end{aligned} \tag{13}$$

otherwise, when $f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T) < f(\boldsymbol{x}_i|T)$,

$$\exp(-f(\boldsymbol{x}_i|T)) - \exp(-f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T)) < 0 + o(\Delta\boldsymbol{x}_i). \tag{14}$$

Thus, we yield

$$\mathrm{err}_i - \mathrm{err}_{i'} \leq \mathcal{C}_1 E_T[f'(\boldsymbol{x}_i|T)\cdot\Delta\boldsymbol{x}_i + o(\Delta\boldsymbol{x}_i)], \tag{15}$$

where $\mathcal{C}_1 \in [0,1]$.

For the lower bound, when $f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T) \geq f(\boldsymbol{x}_i|T)$,

$$\exp(-f(\boldsymbol{x}_i|T)) - \exp(-f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T)) \geq 0 + o(\Delta\boldsymbol{x}_i); \tag{16}$$

when $f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T) < f(\boldsymbol{x}_i|T)$,

$$\begin{aligned}\exp(-f(\boldsymbol{x}_i|T)) &- \exp(-f(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i|T)) \\ &> f'\left(\boldsymbol{x}_i|T\right)\cdot\Delta\boldsymbol{x}_i + o(\Delta\boldsymbol{x}_i).\end{aligned} \tag{17}$$

Thus, we yield

$$\mathrm{err}_i - \mathrm{err}_{i'} \geq \mathcal{C}_2 E_T[f'(\boldsymbol{x}_i|T)\cdot\Delta\boldsymbol{x}_i + o(\Delta\boldsymbol{x}_i)], \tag{18}$$

where $\mathcal{C}_2 \in [0,1]$.

Obviously, $\mathcal{C}_2 \leq \mathcal{C}_1$. We consider the most cases, where $\mathcal{C}_1 \neq 0$ and $\mathcal{C}_2 \neq 0$. The difference between the two errors satisfies the following formula:

$$\begin{aligned}\mathcal{C}_2 E_T[f'(\boldsymbol{x}_i|T)\cdot\Delta\boldsymbol{x}_i + o(\Delta\boldsymbol{x}_i)] &\leq \mathrm{err}_i - \mathrm{err}_{i'} \\ &\leq \mathcal{C}_1 E_T[f'(\boldsymbol{x}_i|T)\cdot\Delta\boldsymbol{x}_i + o(\Delta\boldsymbol{x}_i)].\end{aligned} \tag{19}$$

Ignore the higher-order term $o(\Delta\boldsymbol{x}_i)$, we have

$$\begin{aligned}\mathcal{C}_2 E_T[f'(\boldsymbol{x}_i|T)\cdot\Delta\boldsymbol{x}_i] &\leq \mathrm{err}_i - \mathrm{err}_{i'} \\ &\leq \mathcal{C}_1 E_T[f'(\boldsymbol{x}_i|T)\cdot\Delta\boldsymbol{x}_i].\end{aligned} \tag{20}$$

From the formula above, there are three cases. Let $\angle\varphi$ be the angle between the direction of $\Delta\boldsymbol{x}_i$ and the direction of $E_T[f'\left(\boldsymbol{x}_i|T\right)]$. The cases are summarized as below.

1. If $E_T[f'\left(\boldsymbol{x}_i|T\right)\Delta\boldsymbol{x}_i] > 0$, then $\angle\varphi < 90°$. In this case, the direction of the perturbation $\Delta\boldsymbol{x}_i$ and the direction of $E_T[f'\left(\boldsymbol{x}_i|T\right)$ are consistent. Thus, the generalization error of the noisy sample is smaller than that of the clean one.
2. If $E_T[f'\left(\boldsymbol{x}_i|T\right)\Delta\boldsymbol{x}_i] < 0$, then $\angle\varphi > 90°$. In this case, the direction of the perturbation $\Delta\boldsymbol{x}_i$ and the direction of $E_T[f'\left(\boldsymbol{x}_i|T\right)$ are contradictory. Thus, the generalization error of the noisy sample is larger than that of the clean one.
3. If $E_T[f'\left(\boldsymbol{x}_i|T\right)\Delta\boldsymbol{x}_i] = 0$, then $\angle\varphi = 90°$ or $\Delta\boldsymbol{x}_i = \boldsymbol{0}$. The generalization error does not change in this case.

□

Therefore, the change of the generalization error with feature noise is dependent on the angle between the direction of the perturbation $\Delta \boldsymbol{x}_i$ and the direction of $E_T[f'(\boldsymbol{x}_i|T)]$.

Noise that increases the error is the adversarial type; otherwise, it is a promoted type. Therefore, the variation of the error under feature noise is determined by the noise type. For example, as all feature noises are adversarial in adversarial learning [Lowd and Meek, 2005], all of the samples' errors are increased with feature noise.

Label Noise

For label noise, we offer the following proposition:

Proposition 2. *Let π be the label corruption rate (i.e., the probability of each label flipping to another one). Denote the probability of correct classification for the original samples as p. If $p > 0.5$, then the errors of the noisy samples are larger than those of the clean ones.*

Proof. Let π be the label corruption rate, that is, the probability of each label flipping to another one. When label noise is added, the generalization error of the sample $(\boldsymbol{x}_i, y_i = 1)$ after adding label noise (i.e., $(\boldsymbol{x}_i, y'_i = -1)$) becomes

$$\begin{aligned}\mathrm{err}_{i''} &= E_T\left[\ell\left(f\left(\boldsymbol{x}_i|T\right), y'_i\right)\right] \\ &= \int_{T\in\mathcal{X}\times\mathcal{Y}} \left(1-\pi\right)e^{-f(\boldsymbol{x}_i|T)} + \pi e^{f(\boldsymbol{x}_i|T)} dP\left(T\right).\end{aligned} \tag{21}$$

The sign of the output $f(\boldsymbol{x}_i)$ indicates the predicted label. For samples which are classified correctly, $y_i f(\boldsymbol{x}_i) > 0$, otherwise, $y_i f(\boldsymbol{x}_i) < 0$. To clearly distinguish between correctly predicted and wrongly predicted samples, the absolute value of $f(\boldsymbol{x}_i)$ is utilized. Therefore, if a sample is correctly predicted, if loss is $e^{-|f(\boldsymbol{x}_i)|}$; otherwise, the loss is $e^{|f(\boldsymbol{x}_i)|}$. As the probability of a sample being correctly classified is p and $p > 0.5$, the generalization error of the original sample $(\boldsymbol{x}_i, y_i = 1)$ is

$$\begin{aligned}\mathrm{err}_i &= E_T\left[\ell\left(f\left(\boldsymbol{x}_i|T\right), y_i\right)\right] \\ &= \int_{T\in\mathcal{X}\times\mathcal{Y}} pe^{-|f(\boldsymbol{x}_i|T)|} + (1-p)e^{|f(\boldsymbol{x}_i|T)|} dP\left(T\right).\end{aligned} \tag{22}$$

After flipping the label of the sample, its generalization error becomes

$$\begin{aligned}\mathrm{err}_{i''} &= E_T\left[\ell\left(f\left(\boldsymbol{x}_i|T\right), y'_i\right)\right] \\ &= \int_{T\in\mathcal{X}\times\mathcal{Y}} (1-\pi)\, pe^{-|f(\boldsymbol{x}_i|T)|} + (1-\pi)(1-p)e^{|f(\boldsymbol{x}_i|T)|} dP(T) \\ &+ \int_{T\in\mathcal{X}\times\mathcal{Y}} \pi(1-p)e^{-|f(\boldsymbol{x}_i|T)|} + \pi pe^{|f(\boldsymbol{x}_i|T)|} dP\left(T\right).\end{aligned} \tag{23}$$

Comparing the two generalization errors above, we yield

$$\begin{aligned}\mathrm{err}_{i''} - \mathrm{err}_i &= \pi(2p-1)\int_{T\in\mathcal{X}\times\mathcal{Y}} e^{|f(\boldsymbol{x}_i|T)|} - e^{-|f(\boldsymbol{x}_i|T)|} dP\left(T\right) \\ &> 0.\end{aligned} \tag{24}$$

Therefore, for a sample that is more likely to be predicted correctly (i.e., the probability of being correctly classified p is greater than 0.5), its generalization error after adding label noise is larger than that of the original one. Therefore, the generalization errors of the samples with label noises are larger than those of the clean ones on the average. □

This proposition implies that the errors of the samples with label noises are larger than those of the clean ones on the average. Let $\mathcal{L}^*$ be the global optimum of the generalization error of the clean dataset. When the noise in Proposition 2 is added, the empirical error $\mathcal{L}'$ is

$$\mathcal{L}' = \left(1-\pi\right)\mathcal{L}^* + \pi\mathcal{L}\left(f\left(\boldsymbol{x}\right), -y\right), \tag{25}$$

where we have taken expectations over the noise.

When $\pi \to 0$, the noise disappears, and the optimal generalization is attained.

The noisy samples have larger errors than the clean ones on the average.

2.2 Imbalance Factor

Besides noise, imbalance is another common deviation of real world datasets. Samples in tail categories are considered to be hard ones [Lin *et al.*, 2017; Cui *et al.*, 2019]. The category distribution of the samples in the training set is non-uniform, which means that $p\left(y=1\right) \neq p\left(y=-1\right)$. Alternatively, if $\mathbb{c}_1$ and $\mathbb{c}_2$ samples exist in the two categories, then $\mathbb{c}_1 \neq \mathbb{c}_2$. The imbalance ratio is denoted by $c_r = \max\{\mathbb{c}_1, \mathbb{c}_2\} : \min\{\mathbb{c}_1, \mathbb{c}_2\}$. Then, we offer the following proposition.

Proposition 3. *If a predictor on an imbalanced dataset ($c_r > e : 1$) is an approximate Bayesian optimal classifier (as the exponential loss is an approximation for the zero-one loss), which is to minimize the total risk, then the average probability of the ground truth of the samples in the large category is greater than that of the samples in the small category.*

Proof. Take the output of the sigmoid layer as the model's output $f(\boldsymbol{x})$ and $f(\boldsymbol{x}) \in (0, 1)$. The disproved method is adopted to prove this proposition. We prove that if the average probability of ground truth of the large category, which contains the majority of samples, is smaller than that of the small category, then the classifier must not be the approximate Bayesian optimal classifier. There are two categories which are $y = 1$ and $y = -1$. The numbers of samples in the two categories are $\mathbb{c}_1$ and $\mathbb{c}_2$. The condition is that $\mathbb{c}_1 > e\mathbb{c}_2$ ($c_\tau > e : 1$), which means that the number of samples in the large category is e times the number of samples in the small category. Denote the average probabilities of ground truth for the large and small categories as f_1 and f_2, and $f_1 < f_2$. The total error is

$$\begin{aligned}\mathcal{L} &= \frac{1}{n}\left(\sum_{i=1}^{\mathbb{c}_1}\ell\left(f\left(\boldsymbol{x}_i\right), y_i\right) + \sum_{j=1}^{\mathbb{c}_2}\ell\left(f\left(\boldsymbol{x}_j\right), y_j\right)\right) \\ &= \frac{1}{n}\left[\mathbb{c}_1 e^{-f_1} + \mathbb{c}_2 e^{1-f_2}\right].\, (i)\end{aligned} \tag{26}$$

Denote $\frac{1}{n}\left[\mathbb{c}_1 e^{-f_2}+\mathbb{c}_2 e^{1-f_1}\right]$ as (ii). $(i)-(ii)$ gives

$$\begin{aligned}
&\frac{1}{n}\left[\mathbb{c}_1 e^{-f_1}+\mathbb{c}_2 e^{1-f_2}-\mathbb{c}_1 e^{-f_2}-\mathbb{c}_2 e^{1-f_1}\right]\\
&=\frac{1}{n}\left[\mathbb{c}_1\left(e^{-f_1}-e^{-f_2}\right)+\mathbb{c}_2\left(e^{1-f_2}-e^{1-f_1}\right)\right]\\
&=\frac{1}{n}\left[\mathbb{c}_1\frac{e^{f_2}-e^{f_1}}{e^{f_1}e^{f_2}}+\mathbb{c}_2\frac{e\left(e^{f_1}-e^{f_2}\right)}{e^{f_1}e^{f_2}}\right] \qquad (27)\\
&=\frac{1}{n}\frac{e^{f_2}-e^{f_1}}{e^{f_1}e^{f_2}}\left[\mathbb{c}_1-\mathbb{c}_2 e\right]\\
&>0
\end{aligned}$$

Obviously, if the average probability of the ground truth for the large category is smaller than that of the small category, then the predictor is not an approximate Bayesian optimal classifier. Thus, if a predictor is an approximate Bayesian optimal classifier, the average probability of the ground truth for the large category is greater than that of the small category, that is $f_1 > f_2$. Proposition 3 holds. □

According to Proposition 3, it is natural to get the following proposition.

Proposition 4. *The average generalization error $\overline{\mathrm{err}}$ of samples in the large category $\overline{\mathrm{err}}_1$ is larger than that of samples in the small category $\overline{\mathrm{err}}_2$.*

The proof is shown below.

Proof. If the positive category is the large category, the average generalization error of samples in this category is

$$\overline{\mathrm{err}}_1=\int_{T\in\mathcal{X}\times\mathcal{Y}} e^{-f_1}dP(T), \qquad (28)$$

If the positive category is the small category, the average generalization error of samples in this category is

$$\overline{\mathrm{err}}_2=\int_{T\in\mathcal{X}\times\mathcal{Y}} e^{-f_2}dP(T), \qquad (29)$$

where $\overline{\mathrm{err}}_1$ denotes the average generalization error of samples in the large category, and $\overline{\mathrm{err}}_2$ denotes the average generalization error of samples in the small category. All sampled datasets are imbalanced, in which $c_r > e : 1$. The difference between the two average generalization errors equals to

$$\overline{\mathrm{err}}_1-\overline{\mathrm{err}}_2=\int_{T\in\mathcal{X}\times\mathcal{Y}} e^{-f_1}-e^{-f_2}dP(T). \qquad (30)$$

From Proposition 3, we know that $f_1 > f_2$. Thus, $\overline{\mathrm{err}}_1 < \overline{\mathrm{err}}_2$, which means that the average generalization error of samples in the large category is smaller than that of samples in the small category. □

Accordingly, the average error of samples in the small category is larger than that of the samples in the large category (Proposition 4), indicating there are more hard samples in the small category. The tail categories contain more samples with larger errors. Thus, samples with larger errors are assigned with higher weights on the imbalanced training data in most existing studies [Cui *et al.*, 2019].

2.3 Margin Factor

The samples' margins measure the distances of the samples from the decision boundary [Elsayed *et al.*, 2018; Zhang *et al.*, 2021]. Intuitively, a small margin indicates a large difficulty, and it corresponds to a low confidence of the prediction. We offer the following proposition.

Proposition 5. *Let μ_i be the true margin of $\boldsymbol{x}_i$ corresponding to the oracle decision boundary. Assume the functional margins of a sample trained on random datasets obey a Gaussian distribution. In other words, for sample $\boldsymbol{x}_i$, its functional margin $\gamma_i \sim \mathcal{N}(\mu_i,\sigma_i^2)$. For sample $\boldsymbol{x}_j$, its margin $\gamma_j \sim \mathcal{N}(\mu_j,\sigma_j^2)$. If $\sigma_i^2 = \sigma_j^2$, when $\mu_i \le \mu_j$, then $\mathrm{err}_i \ge \mathrm{err}_j$. If $\mu_i = \mu_j$, when $\sigma_i^2 \ge \sigma_j^2$, then $\mathrm{err}_i \ge \mathrm{err}_j$.*

Proof. According to the moment-generating function, there is

$$E\left[e^{tx}\right]=e^{t\mu+\frac{1}{2}\sigma^2t^2}, \quad x\sim\mathcal{N}\left(\mu,\sigma^2\right). \qquad (31)$$

When $t=-1$, it can be drawn that

$$E\left[e^{-x}\right]=e^{-\mu+\frac{1}{2}\sigma^2}, \quad x\sim\mathcal{N}\left(\mu,\sigma^2\right). \qquad (32)$$

For sample $\boldsymbol{x}_i$, its generalization error and margin are denoted as err_i and γ_i, respectively. As the condition that the functional margins γ_i of sample $\boldsymbol{x}_i$ trained on random datasets obey the Gaussian distribution $\mathcal{N}$, and the mean of the distribution μ_i is the true margin corresponding to the oracle decision boundary. Thus, for samples $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, there are $\gamma_1 \sim \mathcal{N}(\mu_1,\sigma_1^2)$ and $\gamma_2 \sim \mathcal{N}(\mu_2,\sigma_2^2)$. Based on the moment-generating function, we yield

$$E_T\left[e^{-\gamma_1(T)}\right]=e^{-\mu_1+\frac{1}{2}\sigma_1^2}, \qquad (33)$$

and

$$E_T\left[e^{-\gamma_2(T)}\right]=e^{-\mu_2+\frac{1}{2}\sigma_2^2}. \qquad (34)$$

Therefore, when $\sigma_1=\sigma_2$, if $\mu_1\le\mu_2$, then we have

$$E_T[e^{-\gamma_1(T)}]\ge E_T[e^{-\gamma_2(T)}], \qquad (35)$$

which indicates that $\mathrm{err}_1 > \mathrm{err}_2$.

For the second case, when $\mu_1=\mu_2$, if $\sigma_1^2\ge\sigma_2^2$, we obtain

$$E_T[e^{-\gamma_1(T)}]\ge E_T[e^{-\gamma_2(T)}], \qquad (36)$$

which also indicates $\mathrm{err}_1 > \mathrm{err}_2$.

Thus, the true margin (the mean of the functional margin distribution) of a sample and its generalization error are negatively correlated when the margin variances of samples are equal, while the margin variance and the generalization error are positively correlated when the true margins are equal. □

This proposition indicates that the conclusion of [Xu *et al.*, 2020] that samples close to the oracle decision boundary are hard ones does not always hold. Even samples that have large true margins may have large errors.

The above proposition indicates that the true margin (i.e., the mean of the functional margin distribution) of a sample and error are negatively correlated when the margin variances of the samples are equal. By contrast, the margin variance and error are positively correlated when the true margins are equal. Thus, the conclusion in which samples close to the oracle decision boundary are hard ones [Xu *et al.*, 2020] is not completely correct. Indeed, the relation between the margin and error of sample $\boldsymbol{x}_i$ conforms with the following formula:

$$\mathrm{err}_i = E_T[e^{-\gamma_i(T)}] = e^{-\mu_i+\frac{1}{2}\sigma_i^2} \tag{37}$$

For the two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, if $\mu_i < \mu_j$ and $\sigma_i^2 < \sigma_j^2$, then we cannot judge whether err_i is greater than err_j. The average margin and error are negatively correlated for most samples, but it is not absolute, which accords with the above analyses.

2.4 Uncertainty Factor

Uncertainties [Kendall and Gal, 2017] in deep learning are classified into two types. The first type is aleatoric uncertainty (data uncertainty), which is caused by the noise in the observation data. Its correlation with the error has been discussed in Section 2.1. The second type is called epistemic uncertainty (model uncertainty), and it is used to indicate the consistency of multiple predictions. Let T be a training set, and let $P(\boldsymbol{\theta}|T)$ be the distribution of the training models based on T. The predictive variance $Var(f(\boldsymbol{x}_i|\boldsymbol{\theta}_1), \cdots, f(\boldsymbol{x}_i|\boldsymbol{\theta}_K))$ plus a precision constant is a typical manner of estimating epistemic uncertainty [Gal and Ghahramani, 2016; Abdar *et al.*, 2021]. Take the mean square loss as an example[1]. The epistemic uncertainty is

$$\begin{aligned}\widehat{\mathrm{Var}}\,[\boldsymbol{x}_i] :=& \tau^{-1} + \frac{1}{|K|}\sum\nolimits_k f(\boldsymbol{x}_i|\boldsymbol{\theta}_k)^\intercal f(\boldsymbol{x}_i|\boldsymbol{\theta}_k) \\ & - E[f(\boldsymbol{x}_i|\boldsymbol{\theta}_k)]^\intercal E[f(\boldsymbol{x}_i|\boldsymbol{\theta}_k)],\end{aligned} \tag{38}$$

where τ is a constant. The second term is the second raw moment of the predictive distribution and the third term is the square of the first moment on the right side of Eq. (38). When $K \to \infty$ and the constant term is ignored, Eq. (38) becomes

$$\widehat{\mathrm{Var}}\,[\boldsymbol{x}_i] := \int_{\boldsymbol{\theta}} ||f(\boldsymbol{x}_i|\boldsymbol{\theta}) - \overline{f}(\boldsymbol{x}_i)||_2^2 dP(\boldsymbol{\theta}|T). \tag{39}$$

If $P(\boldsymbol{\theta}|T)$ is approximated by the distribution of learned models on random training sets which conform to the Gaussian distribution $\mathcal{N}(T, \delta I)$, Eq. (39) is exactly the variance term of the error defined in Eq. (2) when the mean square loss is utilized.

As the bias term in the error can capture the aleatoric uncertainty and the variance term captures the epistemic uncertainty, the overall relationship between uncertainty and error is positively correlated. Nevertheless, the relationship between epistemic uncertainty and error is not simply positively or negatively correlated. For some samples with heavy noises, their epistemic uncertainties will be small as their predictions remain erroneous. However, their errors are large due to their large bias. Epistemic uncertainty and error are positively correlated for some samples, and the two variables are negatively correlated for other samples.

2.5 Relationship between Existing Difficulty Measures and Generalization Error

The commonly used difficulty measures, such as average loss and gradient norm, are mainly related to the bias term. Shin et al. [2020] emphasized that only using loss as the measurement cannot distinguish clean and noisy samples, especially for uniform label noise. Indeed, the variance term should not be underestimated when measuring the samples' learning difficulty[2]. Our analyses support that the error can capture four main factors influencing the samples' learning difficulty. Thus, the error can be leveraged as a more universal measure if efficient error calculation algorithms are to be proposed.

References

[Abdar *et al.*, 2021] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 2021.

[Agarwal *et al.*, 2020] Chirag Agarwal, Daniel D'souza, and Sara Hooker. Estimating example difficulty using variance of gradients. *arXiv preprint arXiv:2008.11600*, 2020.

[Castells *et al.*, 2020] Thibault Castells, Philippe Weinzaepfel, and Jerome Revaud. Superloss: A generic loss for robust curriculum learning. In *NeurIPS*, pages 1–12, 2020.

[Cui *et al.*, 2019] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pages 9260–9269, 2019.

[Elsayed *et al.*, 2018] Gamaleldin F. Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In *NeurIPS*, pages 850–860, Montréal, Canada, December 2018.

[Gal and Ghahramani, 2016] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059. PMLR, 2016.

[Heskes, 1998] Tom Heskes. Bias/variance decompositions for likelihood-based estimators. *Neural Computation*, 10(6):1425–1433, August 1998.

[Kendall and Gal, 2017] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *ICCV*, pages 2999–3007, 2017.

[Lowd and Meek, 2005] Daniel Lowd and Christopher Meek. Adversarial learning. In *SIGKDD*, pages 641–647, 2005.

[Lyu and Li, 2019] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2019.

[Shin *et al.*, 2020] Wonyoung Shin, Jung-Woo Ha, Shengzhe Li, Yongwoo Cho, Hoyean Song, and Sunyoung Kwon.

[1] For other losses, other methods can be used to calculate the predictive variance [Yang *et al.*, 2020].

[2] Limited existing studies use variance. Agarwal et al. [2020] applied the variance of gradient norms as the difficulty measure.

Which strategies matter for noisy label classification? insight into loss and uncertainty. *arXiv preprint arXiv:2008.06218*, 2020.

[Xu *et al.*, 2020] Da Xu, Yuting Ye, and Chuanwei Ruan. Understanding the role of importance weighting for deep learning. In *ICLR*, 2020.

[Yang *et al.*, 2020] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *ICML*, pages 10767–10777. PMLR, 2020.

[Zhang *et al.*, 2021] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *ICLR*, pages 1–29, Online, 2021.