Supplementary Material: How Does Loss Function Affect Generalization Performance of Deep Learning? Application to Human Age Estimation

Ali Akbari [1] **Muhammad Awais** [1] **Manijeh Bashar** [2] **Josef Kittler** [1]

Abstract

This document provides supplementary material, including proof of theorem and statements in the main text.

1. Preliminaries

For simplicity, we review some preliminaries described in the paper.

Definition 1 (Generalization Error)**.** *Given a training set $\mathcal{S}$ and staring with a set of random indices $\mathcal{R}$ of samples in $\mathcal{S}$, the generalization error of the output model $f^{\theta}_{\mathcal{S},\mathcal{R}}$, trained by SGD, is defined as the difference between the empirical risk and true risk, i.e. $E(\mathcal{S},\mathcal{R}) = R_{\text{true}}(f^{\theta}_{\mathcal{S},\mathcal{R}}) - R_{\text{emp}}(f^{\theta}_{\mathcal{S},\mathcal{R}})$. It should be noted that due to the randomness of $\mathcal{S}$ and $\mathcal{R}$, $f^{\theta}_{\mathcal{S},\mathcal{R}}$ and consequently $E(\mathcal{S},\mathcal{R})$ are random variable.*

Definition 2 (Lipschitzness)**.** *A loss function $\ell(\hat{\mathbf{y}},\mathbf{y})$ is γ-Lipschitz with regard to the estimated output vector $\hat{\mathbf{y}}$, if for $\gamma \geq 0$ and $\forall \mathbf{u},\mathbf{v} \in \mathbb{R}^K$ we have*

$$|\ell(\mathbf{u},\mathbf{y}) - \ell(\mathbf{v},\mathbf{y})| \leq \gamma \|\mathbf{u}-\mathbf{v}\|, \tag{1}$$

where $\|\cdot\|$ denotes the ℓ_2-norm of vectors. Intuitively, a Lipschitz loss function is upper-bounded in terms of its rate of change.

Definition 3 (Smoothness)**.** *A loss function $\ell(\hat{\mathbf{y}},\mathbf{y})$ is η-smooth with regard to the estimated output vector $\hat{\mathbf{y}}$, if its gradient $\nabla\ell(\hat{\mathbf{y}},\mathbf{y})$ is η-Lipschitz, that is for $\eta \geq 0$ and $\forall \mathbf{u},\mathbf{v} \in \mathbb{R}^K$ we have*

$$\|\nabla\ell(\mathbf{u},\mathbf{y}) - \nabla\ell(\mathbf{v},\mathbf{y})\| \leq \eta \|\mathbf{u}-\mathbf{v}\|. \tag{2}$$

Intuitively, the curvature of the loss function is upper-bounded by the η-smoothness property.

Definition 4 (Bounded Difference Inequality (BDI))**.** *Let $\mathcal{Z}$ be some set and $G : \mathcal{Z}^n \rightarrow \mathbb{R}$ be any measurable function. Consider two sets $\mathcal{Q}, \mathcal{Q}' \in \mathcal{Z}^n$, such that $\mathcal{Q}$ and $\mathcal{Q}'$ differ in at most one element. If there exists constant ρ such that the following condition, namely bounded difference condition (BDC),*

$$\sup_{\mathcal{Q},\mathcal{Q}' \in \mathcal{Z}^n} |G(\mathcal{Q}) - G(\mathcal{Q}')| \leq \rho, \tag{3}$$

holds, then $\forall \epsilon > 0$

$$\mathbb{P}_{\mathcal{Q}}\big[G(\mathcal{Q}) - \mathbb{E}_{\mathcal{Q}}[G(\mathcal{Q})] \geq \epsilon\big] \leq \exp(-2\epsilon^2/n\rho^2). \tag{4}$$

Definition 5 (Uniform Stability)**.** *Let $\mathcal{S}'$ and $\mathcal{S}$ denote two training sets of equal size, following an unknown distribution $\mathbb{P}$, such that $\mathcal{S}$ and $\mathcal{S}'$ vary in one entity. Let $f_{\mathcal{S},\mathcal{R}}$ and $f_{\mathcal{S}',\mathcal{R}}$ be the optimal models obtained by SGD, with the set of random indices $\mathcal{R}$ of the training samples in $\mathcal{S}$ and $\mathcal{S}'$, respectively. SGD is then β-uniformly stable with regard to a certain loss function ℓ, if the following inequality holds:*

$$\forall\, \mathcal{S},\, \mathcal{S}' \quad \sup_{\mathbf{z}} \mathbb{E}_{\mathcal{R}}\big[\,|\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{z}) - \ell(f_{\mathcal{S}',\mathcal{R}};\mathbf{z})|\,\big] \leq \beta, \tag{5}$$

where the expectation is taken over the randomness of SGD which is a function of the random choice of data $\mathcal{S}$ for training.

2. Proofs

Theorem 1. *Let function $h : (0,\infty) \rightarrow \mathbb{R}$ be convex, such that $h(1) = 0$. Let's define the following function:*

$$I(\hat{\mathbf{y}},\mathbf{y}) = \sum_{k=1}^{K} y_k h\left(\frac{\hat{y}_k}{y_k}\right). \tag{6}$$

If $h(\cdot)$ is γ-Lipschitz, i.e.

$$|h(x) - h(z)| \leq \gamma |x - z| \quad \forall x, z, \tag{7}$$

then $I(\hat{\mathbf{y}},\mathbf{y})$ is also γ-Lipschitz. Furthermore, since $h(\cdot)$ is convex, $I(\hat{\mathbf{y}},\mathbf{y})$ is also convex with regard to its first argument.

Proof. Let $x = \frac{u_k}{y_k}$ and $z = \frac{v_k}{y_k}$. Then, from (7), we have

$$\left| h\left(\frac{u_k}{y_k}\right) - h\left(\frac{v_k}{y_k}\right) \right| \leq \gamma \left| \frac{u_k}{y_k} - \frac{v_k}{y_k} \right| \quad \forall k \in \{1, \cdots L\}. \tag{8}$$

[1] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK. [2] Institute for Communication Systems (ICS), University of Surrey, Guildford, UK. Correspondence to: Ali Akbari <ali.akbari@surrey.ac.uk>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Multiplying both sides of (8) by y_k, and then employing summation on all the obtained inequalities for all $k \in \{1, \cdots L\}$, we obtain

$$\sum_{k=1}^{K} \left| y_k h\left(\frac{u_k}{y_k}\right) - y_k h\left(\frac{v_k}{y_k}\right) \right| \leq \gamma \sum_{k=1}^{K} |u_k - v_k| . \quad (9)$$

Using the generalised triangle inequality, we get

$$\left| \sum_{k=1}^{K} y_k h\left(\frac{u_k}{y_k}\right) - \sum_{k=1}^{K} y_k h\left(\frac{v_k}{y_k}\right) \right| \leq \sum_{k=1}^{K} \left| y_k h\left(\frac{u_k}{y_k}\right) - y_k h\left(\frac{v_k}{y_k}\right) \right| \leq \gamma \sum_{k=1}^{K} |u_k - v_k| . \quad (10)$$

Finally, we obtain

$$|I(\mathbf{u}, \mathbf{y}) - I(\mathbf{v}, \mathbf{y})| \leq \gamma \|\mathbf{u} - \mathbf{v}\|. \quad (11)$$

and the Lipschitz property of $I(\hat{\mathbf{y}}, \mathbf{y})$ is proved.

We now prove that the convexity of $h(\cdot)$ implies the convexity of $I(\hat{\mathbf{y}}, \mathbf{y})$ with respect to its first argument. Let $\mathbf{u}, \mathbf{v} \in \mathcal{Y}$ be two probability distributions with all values nonzero and $t \in [0.1]$. Then we have

$$I(t\mathbf{u} + (1-t)\mathbf{u}, \mathbf{y}) = \sum_{k=1}^{K} y_k h\left(\frac{tu_k + (1-t)v_k}{y_k}\right). \quad (12)$$

Due to the convexity of h, we have $\forall k \in \{1, \cdots, K\}$

$$h\left(\frac{tu_k + (1-t)v_k}{y_k}\right) \leq th\left(\frac{u_k}{y_k}\right) + (1-t)h\left(\frac{v_k}{y_k}\right). \quad (13)$$

Summing over k from 1 to K and utilizing (12) we get

$$I(t\mathbf{u} + (1-t)\mathbf{u}, \mathbf{y}) \leq tI(\mathbf{u}, \mathbf{y}) + (1-t)I(\mathbf{v}, \mathbf{y}) \quad (14)$$

proving the desired result. ■

Lemma 1. *A function $h : (0, \infty) \to \mathbb{R}$ is γ-Lipschitz, if γ satisfies the following condition:*

$$\gamma = \sup_x |h'(x)|. \quad (15)$$

This implies the value of γ must be equal to the maximum value of $|h'(x)|$.

Proof. This lemma can be easily derived from the definition of Lipschitz property. ■

Corollary 1. *Let the GJM and KL loss functions are γ_{GJM}-Lipschitz and γ_{KL}-Lipschitz, respectively. Then, the following inequality holds:*

$$\gamma_{GJM} \leq \gamma_{KL}. \quad (16)$$

Proof. As $h_{KL}(x) = -\log(x), x > 0$, then obviously $\ell_{KL}(\hat{\mathbf{y}}, \mathbf{y}) = I(\hat{\mathbf{y}}, \mathbf{y})$. It is also obvious that if $h_{GJM}(x) = |1 - x^{\alpha}|^{\frac{1}{\alpha}}, x > 0$, then $\ell_{GJM}(\hat{\mathbf{y}}, \mathbf{y}) = I(\hat{\mathbf{y}}, \mathbf{y})$. Then, we have

$$h'_{KL}(x) = -\frac{1}{x} \quad (17)$$

and

$$h'_{GJM}(x) = \text{sign}(x^{\alpha} - 1)x^{\alpha - 1}|x^{\alpha} - 1|^{\frac{1-\alpha}{\alpha}}. \quad (18)$$

Fig. 1 shows the absolute value of the derivative for the KL and GJM loss functions as a function of x. As can be seen $|h'_{GJM}(x)|$ is smaller than $|h'_{KL}(x)|$. From Lemma 1, this implies the inequality in (16) holds. We also theoretically prove that $|h'_{GJM}(x)| \leq |h'_{KL}(x)|$ for $\alpha = 0.5$, *i.e.*

$$\left| 1 - \frac{1}{\sqrt{x}} \right| \leq \left| \frac{1}{x} \right|. \quad (19)$$

(19) is equivalent to $|x - \sqrt{x}| \leq 1$, which results in $x \leq 2.6$ after some mathematical simplification. We experimentally found that the variable x always satisfies this condition when the model starts to converge. Note that $|h'_{GJM}(x)|$ and $|h'_{KL}(x)|$ meet each other at some point. For instance, for $\alpha = 0.2, 0.4, 0.5, 0.8$, the intersection point is $x_p = 22.06, 3.75, 2.61, 1.42$ respectively. After this point, $|h'_{GJM}(x)|$ starts to be slightly larger than $|h'_{KL}(x)|$, but, the difference in this area is very small and negligible compared with the points smaller than the intersection point. In addition, since the large values of x usually occur at the beginning of training (first few epoch), the stability of both GJM and KL are close to each other. This can be seen in Fig. 2 (min paper), where the curves of KL and GJM are close to each other. However, when training continues GJM has better stability, and so better generalization, than KL. ■

Theorem 2. *For two distribution $\mathbf{y}, \hat{\mathbf{y}} \in \mathbb{R}^K$, the GJM loss function with $\alpha = 0.5$ is upper-bounded by the KL divergence, i.e. we have the following inequality:*

$$\ell_{GJM}(\hat{\mathbf{y}}, \mathbf{y}) \leq \ell_{KL}(\hat{\mathbf{y}}, \mathbf{y}). \quad (20)$$

Proof. The use of binomial theorem and Jensen's inequality

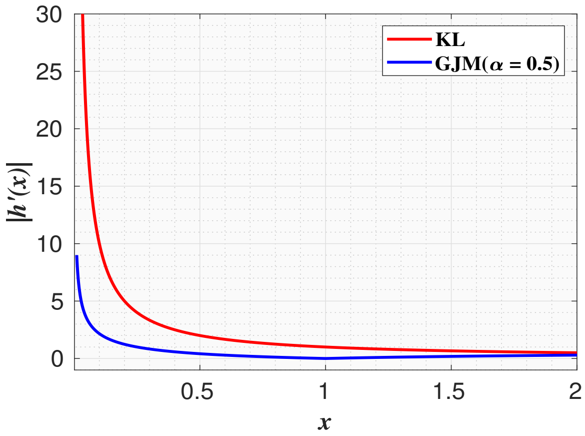

Figure 1. Absolute value of derivative of loss functions at different points x.

gives

$$\begin{aligned}
\ell_{GJM}(\hat{\mathbf{y}},\mathbf{y}) &= \sum_{k=1}^{K} y_k \left|1-\left(\frac{\hat{y}_k}{y_k}\right)^{\alpha}\right|^{\frac{1}{\alpha}} \\
&\overset{(a)}{=} \sum_{k=1}^{K} y_k \sum_{i=0}^{1/\alpha} (-1)^i \binom{1/\alpha}{i} \left(\frac{\hat{y}_k}{y_k}\right)^{i\alpha} \\
&= \sum_{k=1}^{K} y_k \sum_{i=0}^{1/\alpha} (-1)^i \binom{1/\alpha}{i} \exp\left(i\alpha \log\left(\frac{\hat{y}_k}{y_k}\right)\right) \\
&\overset{(b)}{\leq} \sum_{i=0}^{1/\alpha} (-1)^i \binom{1/\alpha}{i} \exp\left(i\alpha \sum_{k=1}^{K} y_k \log\left(\frac{\hat{y}_k}{y_k}\right)\right) \\
&= \sum_{i=0}^{1/\alpha} (-1)^i \binom{1/\alpha}{i} \exp(-i\alpha \ell_{KL}(\hat{\mathbf{y}},\mathbf{y})) \\
&= (1-\exp(-\alpha \ell_{KL}(\hat{\mathbf{y}},\mathbf{y})))^{\frac{1}{\alpha}},
\end{aligned} \tag{21}$$

where inequalities a and b are due to the Binomial theorem and the Jensen's inequality, respectively. Note that $\ell_{KL}(\hat{\mathbf{y}},\mathbf{y}) \in [0,\infty)$ (Gibbs & Su, 2002). The use of Bernoulli's inequality $\exp(x) \geq 1+x$ gives

$$\begin{aligned}
\ell_{GJM}(\hat{\mathbf{y}},\mathbf{y}) &\leq (1-\exp(-\alpha \ell_{KL}(\hat{\mathbf{y}},\mathbf{y})))^{\frac{1}{\alpha}} \\
&\leq 1-\exp(-\alpha \ell_{KL}(\hat{\mathbf{y}},\mathbf{y})) \leq \ell_{KL}(\hat{\mathbf{y}},\mathbf{y}),
\end{aligned} \tag{22}$$

and the inequality is proved. ■

Theorem 3. *Consider a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, L]$. Let $f_{\mathcal{S},\mathcal{R}}$ denote the optimal model obtained by SGD with the set of random indices $\mathcal{R}$ of the training samples in $\mathcal{S}$. Let SGD be β-uniformly stable with regard to the employed loss function ℓ. Furthermore, assume there is a constant ρ for which $\ell(f_{\mathcal{S},\mathcal{R}}, \mathbf{z})$ satisfies the bounded difference condition* (3) *with respect to $\mathcal{R}$. Then, with probability at least $1-\delta$, the following bounds hold $\forall \mathcal{S}, \mathcal{R}$:*

$$\begin{aligned}
&R_{\text{true}}(f_{\mathcal{S},\mathcal{R}}) - R_{\text{emp}}(f_{\mathcal{S},\mathcal{R}}) \leq \\
&\rho\sqrt{T\log(2/\delta)} + \beta\left(1+\sqrt{2N\log(2/\delta)}\right) + L\sqrt{\frac{\log(2/\delta)}{2N}}.
\end{aligned} \tag{23}$$

Proof. Let $E(\mathcal{S},\mathcal{R}) = R_{\text{true}}(f_{\mathcal{S},\mathcal{R}}) - R_{\text{emp}}(f_{\mathcal{S},\mathcal{R}})$, be a random variable depending on $\mathcal{S}$ and $\mathcal{R}$. Let $f_{\mathcal{S},\mathcal{R}}$ and $f_{\mathcal{S},\mathcal{R}'}$ be two output models using the learning algorithm $\mathcal{A}$ applied on the training set $\mathcal{S}$ with the two sets of random parameters $\mathcal{R}$ and $\mathcal{R}'$, respectively. We apply the BDI (4) by considering function G and set Q as $\ell(f_{\mathcal{S},\mathcal{R}}; \mathbf{z})$ and $\mathcal{R}$, respectively. Assume $\mathcal{R}$ and $\mathcal{R}'$ differ only in two elements[1]. Note the BDI cannot be applied directly in this case. So we partition each $\mathcal{R}$ and $\mathcal{R}'$ in two subsets $\mathcal{R}_1, \mathcal{R}_2$ and $\mathcal{R}'_1, \mathcal{R}'_2$ such that the corresponding subsets, *i.e.* $\mathcal{R}_1, \mathcal{R}'_1$ and $\mathcal{R}_2, \mathcal{R}'_2$ differ only in one element. Using the bounded difference conditions (3), there would be a constant $\rho = \max(\rho_1, \rho_2)$ such that for every $\mathbf{z}$ and $\mathcal{S}$, we have the following bounded difference conditions with respect to ℓ:

$$\sup_{\mathcal{R}_1,\mathcal{R}'_1} \left|\ell(f_{\mathcal{S},\mathcal{R}_1}; \mathbf{z}) - \ell(f_{\mathcal{S},\mathcal{R}'_1}; \mathbf{z})\right| \leq \rho \tag{24}$$

and

$$\sup_{\mathcal{R}_2,\mathcal{R}'_2} \left|\ell(f_{\mathcal{S},\mathcal{R}_2}; \mathbf{z}) - \ell(f_{\mathcal{S},\mathcal{R}'_2}; \mathbf{z})\right| \leq \rho, \tag{25}$$

[1] Recall that $\mathcal{R}$ is a set of random indices in $\mathcal{S}$. So, if $\mathcal{R}$ and $\mathcal{R}'$ differ in one element, unavoidably $\mathcal{R}$ and $\mathcal{R}'$ would differ in another element as well.

Then, for every $S, \mathcal{R}_1, \mathcal{R}_1'$, we have

$$\begin{aligned}
&|E(\mathcal{S}, \mathcal{R}_1) - E(\mathcal{S}, \mathcal{R}_1')| \\
&\quad = \Big|\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\big[\ell(f_{\mathcal{S},\mathcal{R}_1};\mathbf{z})\big] - \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\big[\ell(f_{\mathcal{S},\mathcal{R}_1'};\mathbf{z})\big] \\
&\qquad\qquad -\frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathcal{S},\mathcal{R}_1};\mathbf{z}_i) + \frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathcal{S},\mathcal{R}_1'};\mathbf{z}_i)\Big| \\
&\quad \le \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\Big[\,\big|\ell(f_{\mathcal{S},\mathcal{R}_1};\mathbf{z}) - \ell(f_{\mathcal{S},\mathcal{R}_1'};\mathbf{z})\big|\,\Big] \\
&\qquad\qquad + \frac{1}{N}\sum_{i=1}^{N}\big|\ell(f_{\mathcal{S},\mathcal{R}_1};\mathbf{z}_i) - \ell(f_{\mathcal{S},\mathcal{R}_1'};\mathbf{z}_i)\big| \\
&\quad \le 2\rho.
\end{aligned} \tag{26}$$

Applying the BDI (4) results in the following inequality

$$\mathbb{P}_{\mathcal{R}_1}\big[E(\mathcal{S},\mathcal{R}_1) - \mathbb{E}_{\mathcal{R}_1}\left[E(\mathcal{S},\mathcal{R}_1)\right] \ge \epsilon\big] \le \exp(-\epsilon^2/2T\rho^2). \tag{27}$$

Following the same lines for $\mathcal{R}_2, \mathcal{R}_2'$, the following inequality also holds

$$\mathbb{P}_{\mathcal{R}_2}\big[E(\mathcal{S},\mathcal{R}_2) - \mathbb{E}_{\mathcal{R}_2}\left[E(\mathcal{S},\mathcal{R}_2)\right] \ge \epsilon\big] \le \exp(-\epsilon^2/2T\rho^2). \tag{28}$$

By combining the above two inequalities, we obtain:

$$\mathbb{P}_{\mathcal{R}}\big[E(\mathcal{S},\mathcal{R}) - \mathbb{E}_{\mathcal{R}}\left[E(\mathcal{S},\mathcal{R})\right] \ge \epsilon\big] \le \exp(-\epsilon^2/T\rho^2). \tag{29}$$

By setting the r.h.s. equal to ν, the following inequality holds with probability at least $1-\nu$:

$$E(\mathcal{S},\mathcal{R}) \le \rho\sqrt{T\log(1/\nu)} + \mathbb{E}_{\mathcal{R}}\left[E(\mathcal{S},\mathcal{R})\right]. \tag{30}$$

To bound the random variable $E(\mathcal{S},\mathcal{R})$, we now provide the upper bound for $\mathbb{E}_{\mathcal{R}}\left[E(\mathcal{S},\mathcal{R})\right]$. To this end, we again apply the BDI (4) for function G and set Q being $\mathbb{E}_{\mathcal{R}}\left[E(\mathcal{S},\mathcal{R})\right]$ and set of training samples $\mathcal{S}$, respectively. Note that, in this case, the bounded difference condition (3) equals the uniform stability (5), so $\rho = \beta$. Let $f_{\mathcal{S},\mathcal{R}}$ and $f_{\mathcal{S}',\mathcal{R}}$ be the two output models using the learning algorithm $\mathcal{A}$ with the set of random parameters $\mathcal{R}$ applied on two training sets $\mathcal{S},\mathcal{S}'$, respectively. Assume $\mathcal{S},\mathcal{S}'$ differ only in j-th element. For every $S, S', \mathcal{R}$, we have

$$\begin{aligned}
&\Big|\mathbb{E}_{\mathcal{R}}\left[E(\mathcal{S},\mathcal{R})\right] - \mathbb{E}_{\mathcal{R}}\left[E(\mathcal{S}',\mathcal{R})\right]\Big| \\
&\quad = \Bigg|\mathbb{E}_{\mathcal{R}}\Bigg[\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\big[\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{z})\big] - \frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{z}_i)\Bigg] \\
&\quad -\mathbb{E}_{\mathcal{R}}\Bigg[\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\big[\ell(f_{\mathcal{S}',\mathcal{R}};\mathbf{z})\big] - \frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathcal{S}',\mathcal{R}};\mathbf{z}_i)\Bigg]\Bigg| \\
&\quad = \Bigg|\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\Big[\mathbb{E}_{\mathcal{R}}\big[\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{z}) - \ell(f_{\mathcal{S}',\mathcal{R}};\mathbf{z})\big]\Big] \\
&\quad -\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathcal{R}}\big[\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{z}_i) - \ell(f_{\mathcal{S}',\mathcal{R}};\mathbf{z}_i)\big]\Bigg| \\
&\quad \le \underbrace{\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\Big[\mathbb{E}_{\mathcal{R}}\big[\,|\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{z}) - \ell(f_{\mathcal{S}',\mathcal{R}};\mathbf{z})|\,\big]\Big]}_{a} \\
&\quad + \underbrace{\frac{1}{N}\sum_{i=1,i\neq j}^{N}\mathbb{E}_{\mathcal{R}}\big[\,|\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{z}_i) - \ell(f_{\mathcal{S}',\mathcal{R}};\mathbf{z}_i)|\,\big]}_{b} \\
&\quad + \frac{1}{N}\mathbb{E}_{\mathcal{R}}\big[\,|\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{z}_i) - \ell(f_{\mathcal{S}',\mathcal{R}};\mathbf{z}_i)|\,\big] \\
&\quad \le 2\beta + \frac{L}{N},
\end{aligned} \tag{31}$$

where the terms (a) and (b) are upper bounded by β using the definition of uniform stability. Therefore, we have

$$\sup_{\mathcal{S},\mathcal{S}'\in\mathbb{R}^N}\Big|\mathbb{E}_{\mathcal{R}}\left[E(\mathcal{S},\mathcal{R})\right] - \mathbb{E}_{\mathcal{R}}\left[E(\mathcal{S}',\mathcal{R})\right]\Big| \le 2\beta + \frac{L}{N}. \tag{32}$$

Applying the BDI (4) results in the following inequality

$$\begin{aligned}
&\mathbb{P}_{\mathcal{S}}\big[\mathbb{E}_{\mathcal{R}}\left[E(\mathcal{S},\mathcal{R})\right] - \mathbb{E}_{\mathcal{S},\mathcal{R}}\left[E(\mathcal{S},\mathcal{R})\right] \ge \epsilon\big] \\
&\qquad\qquad \le \exp(-2N\epsilon^2/(2N\beta + L)^2).
\end{aligned} \tag{33}$$

By setting the r.h.s. equal to ν, the following inequality holds with probability at least $1-\nu$:

$$\mathbb{E}_{\mathcal{R}}\left[E(\mathcal{S},\mathcal{R})\right] \le \frac{(2N\beta + L)\sqrt{\log(1/\nu)}}{\sqrt{2N}} + \mathbb{E}_{\mathcal{S},\mathcal{R}}\left[E(\mathcal{S},\mathcal{R})\right]. \tag{34}$$

Now, we provide the upper bound for $\mathbb{E}_{\mathcal{S},\mathcal{R}}\left[E(\mathcal{S},\mathcal{R})\right]$. Denote by $\mathcal{T} = \{\mathbf{t}_i, i = 1, 2, \cdots, N\}$, a set of N training samples that are independent from $\mathcal{S}$ and are drawn from an unknown distribution $\mathcal{D}$. Denote $\mathcal{S}'$ the set obtained by replacing the i-th sample in the set $\mathcal{S}$ with i-th sample from the set $\mathcal{T}$.

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S},\mathcal{R}}\left[E(\mathcal{S},\mathcal{R})\right] \\
&= \mathbb{E}_{\mathcal{S},\mathcal{R}}\left[\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\big[\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{z})\big] - \frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{z}_i)\right] \\
&= \mathbb{E}_{\mathcal{S},\mathcal{R}}\left[\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\big[\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{z})\big]\right] \\
&\qquad - \mathbb{E}_{\mathcal{S},\mathcal{T},\mathcal{R}}\left[\frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{t}_i)\right] \\
&+ \mathbb{E}_{\mathcal{S},\mathcal{T},\mathcal{R}}\left[\frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{t}_i)\right] \\
&\qquad - \mathbb{E}_{\mathcal{S},\mathcal{R}}\left[\frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{z}_i)\right] \\
&= \mathbb{E}_{\mathcal{S},\mathcal{R}}\left[\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\big[\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{z})\big]\right] \\
&\qquad - \mathbb{E}_{\mathcal{S},\mathcal{R}}\left[\mathbb{E}_{\mathbf{t}\sim\mathcal{D}}\big[\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{t})\big]\right] \\
&\qquad + \mathbb{E}_{\mathcal{S},\mathcal{T},\mathcal{R}}\left[\frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{t}_i)\right] \\
&\qquad - \mathbb{E}_{\mathcal{S},\mathcal{T},\mathcal{R}}\left[\frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathcal{S}',\mathcal{R}};\mathbf{t}_i)\right] \\
&= \mathbb{E}_{\mathcal{S},\mathcal{T},\mathcal{R}}\left[\frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{t}_i)\right] \\
&\qquad - \mathbb{E}_{\mathcal{S},\mathcal{T},\mathcal{R}}\left[\frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathcal{S}',\mathcal{R}};\mathbf{t}_i)\right] \\
&= \mathbb{E}_{\mathcal{S},\mathcal{T},\mathcal{R}}\left[\frac{1}{N}\sum_{i=1}^{N}\big(\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{t}_i) - \ell(f_{\mathcal{S}',\mathcal{R}};\mathbf{t}_i)\big)\right] \\
&\leq \sup_{\mathcal{S},\mathcal{S}',\mathbf{z}} \mathbb{E}_{\mathcal{R}}\big[\,|\ell(f_{\mathcal{S},\mathcal{R}};\mathbf{z}) - \ell(f_{\mathcal{S}',\mathcal{R}};\mathbf{z})|\,\big] \leq \beta.
\end{aligned}
\tag{35}
$$

The last line is derived from the uniform stability definition (5) and amounts to changing $\mathbf{t}$ to $\mathbf{z}$. By combining (30), (34) and (35) , the following inequality holds with probability at least $1 - 2\nu$.

$$
\begin{aligned}
&E(\mathcal{S},\mathcal{R}) \\
&\leq \rho\sqrt{T\log(1/\nu)} + \beta\left(1 + \sqrt{2N\log(1/\nu)}\right) + L\sqrt{\frac{\log(1/\nu)}{2N}}.
\end{aligned}
\tag{36}
$$

The results follows by setting $\delta = 2\nu$. ■

Theorem 4. *Assume that SGD is run for T iterations with an annealing learning rate λ_t to find the optimal solution. Let $\ell(f^{\theta}(\mathbf{x}),\mathbf{y})$ be convex, γ-Lipschitz and η-smooth with regard to its first argument for each $\mathbf{z} = (\mathbf{x},\mathbf{y})$. Then SGD is β-uniformly stable and holds the ρ-BDC* (3) *with regard to $\ell(f_{\mathcal{S},\mathcal{R}},\mathbf{z})$ and $\mathcal{R}$. Consequently, we have*

$$
\beta \leq \frac{2\gamma^2}{N}\sum_{t=1}^{T}\lambda_t \qquad \textit{and} \qquad \rho \leq \frac{4\gamma^2}{T}\sum_{t=1}^{T}\lambda_t. \tag{37}
$$

Proof. We first prove the first inequality which is similar to (Hardt et al., 2016)[Theorem 3.7]. We include the proof here for the sake of completeness. Then we show how to prove the second inequality. For the sake of simplicity of notation, we represent the output model $f^{\theta}_{\mathcal{S},\mathcal{R}}$ as $\theta_{\mathcal{S},\mathcal{R}}$ in this proof. We will omit $\mathcal{S},\mathcal{R}$ when it is clear from the context. Given a learning rate $\lambda \geq 0$ and a training set $\mathcal{S}$, SGD performs the gradient descent update rule, defined as $G(\theta) = \theta - \lambda\nabla_{\theta}\ell(\theta;\mathbf{z})$, T steps over all samples in $\mathcal{S}$. Here, sample $\mathbf{z}$ is randomly picked from $\mathcal{S}$. Assume the gradient update G is τ-expansive, *i.e.* $\sup_{u,v\in\mathcal{H}} \|\frac{G(u)-G(v)\|}{\|u-v\|} \leq \tau$, and σ-bounded, *i.e.* $\sup_{\theta\in\mathcal{H}}\|\theta - G(\theta)\| \leq \sigma$. Since $\ell(f^{\theta}(\mathbf{x}),\mathbf{y}) = \ell(<\theta,\mathbf{x}>,\mathbf{y})$ is convex, γ-Lipschitz and η-smooth with respect to its first argument for every $\mathbf{z} = (\mathbf{x},\mathbf{y})$, we have $\|\theta - G(\theta)\| \leq \lambda\|\nabla_{\theta}\ell(\theta;\mathbf{z})\| = \lambda\|\nabla_{\theta}\ell(<\theta,\mathbf{x}>,\mathbf{y})\| \leq \lambda\gamma$. Therefore the update rule is $\lambda\gamma$-bounded.

Let $\theta^1_{\mathcal{S}},\cdots\theta^T_{\mathcal{S}}$ and $\theta^1_{\mathcal{S}'},\cdots\theta^T_{\mathcal{S}'}$ be two sequences of output models resulting respectively from performing two sequences of the gradient updates $G(\theta^1_{\mathcal{S}}),\cdots G(\theta^T_{\mathcal{S}})$ and $G(\theta^1_{\mathcal{S}'}),\cdots G(\theta^T_{\mathcal{S}'})$ applied to two training sets $\mathcal{S},\mathcal{S}'$. Assume sets $\mathcal{S},\mathcal{S}'$ differ only in one element and the initialisation weights $\theta^0_{\mathcal{S}} = \theta^0_{\mathcal{S}'}$. Let $\Delta^t = \|\theta^t_{\mathcal{S}} - \theta^t_{\mathcal{S}'}\|$. The proof is based on the growth recursion concept (Hardt et al., 2016)[Lemma 2.4] which investigates how two distinct sequences of update rules applied to a deep neural model diverge when they start from the same initialisation point and the training set is perturbed at each step. For simplicity, we recall here the growth recursion result.

Growth recursion rule. ***(Hardt et al., 2016)[Lemma 2.4]*** *There exists the following relation recurrence between Δ^{t+1} and Δ^t:*

- *If $G(\theta^t_{\mathcal{S}})$ and $G(\theta^t_{\mathcal{S}'})$ are equal and τ-expansive, then $\Delta^{t+1} \leq \tau\Delta^t$*
- *$G(\theta^t_{\mathcal{S}})$ and $G(\theta^t_{\mathcal{S}'})$ are σ-bounded and τ-expansive, then $\Delta^{t+1} \leq \min(\tau,1)\Delta^t + 2\sigma_t$*

At each step t, there are two cases when two samples $\mathbf{z}$ and $\mathbf{z}'$ are picked by SGD from $\mathcal{S}$ and $\mathcal{S}'$ respectively: 1) $\mathbf{z}$ and $\mathbf{z}'$ are the same with probability $1 - 1/N$, which implies $G^t_{\mathcal{S}} = G^t_{\mathcal{S}'}$, 2) $\mathbf{z}$ and $\mathbf{z}'$ are different with probability $1/N$. Further, if the loss function is smooth and convex and the learning rate λ_t is small enough, it is proved that the gradient update rule $G^t_{\mathcal{S}}$ is 1-expansive (Hardt et al., 2016). Beside that, as mentioned before, $G^t_{\mathcal{S}}$ is $\lambda\gamma$-bounded. Applying the growth recursion rule results:

$$\Delta^{t+1} \leq \left(1 - \frac{1}{N}\right)\Delta^t + \frac{1}{N}\Delta^t + \frac{2\lambda\gamma}{N}. \tag{38}$$

Considering this inequality recursively through all steps, we obtain

$$\Delta^T \leq \frac{2\gamma}{N}\sum_{t=1}^{T}\lambda_t. \tag{39}$$

Using (38) and the fact that the loss function $\ell(f^\theta(\mathbf{x}), \mathbf{y}) = \ell(<\theta, \mathbf{x}>, \mathbf{y})$ is γ-Lipschitz with respect to its first argument, the following inequality is obtained for any $\mathbf{z}$, $\mathcal{S}$, $\mathcal{S}'$:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{R}}\left[\left|\ell(\theta^T_{\mathcal{S},\mathcal{R}}; \mathbf{z}) - \ell(\theta^T_{\mathcal{S}',\mathcal{R}}; \mathbf{z})\right|\right] \\
&\quad \leq \mathbb{E}_{\mathcal{R}}\left[\left|\ell(<\theta^T_{\mathcal{S},\mathcal{R}}, \mathbf{x}>, \mathbf{y})) - \ell(<\theta^T_{\mathcal{S}',\mathcal{R}}, \mathbf{x}>, \mathbf{y}))\right|\right] \\
&\quad \leq \gamma\mathbb{E}_{\mathcal{R}}\left[\left\|<\theta^T_{\mathcal{S},\mathcal{R}}, \mathbf{x}> - <\theta^T_{\mathcal{S}',\mathcal{R}}, \mathbf{x}>\right\|\right] \\
&\quad \leq \gamma\mathbb{E}_{\mathcal{R}}\left[\left\|\theta^T_{\mathcal{S},\mathcal{R}} - \theta^T_{\mathcal{S}',\mathcal{R}}\right\|\right] = \gamma\mathbb{E}_{\mathcal{R}}\left[\Delta^T\right] \leq \frac{2\gamma^2}{N}\sum_{t=1}^{T}\lambda_t,
\end{aligned} \tag{40}$$

where the expectation is taken over randomness of the SGD algorithm, which appears in the random set $\mathcal{R}$. Without loss of generality, we assume $\|\mathbf{x}\| \leq 1$ in the last inequality. The inequality (40) implies the uniform stability (5), which renders the desired inequality.

Now, we derive the second inequality. The proof follows the same reasoning as that used for deriving the first inequality, except that the sequences of the update rule relate to $\mathcal{R}, \mathcal{R}'$. Let $\theta^1_{\mathcal{R}}, \cdots \theta^T_{\mathcal{R}}$ and $\theta^1_{\mathcal{R}'}, \cdots \theta^T_{\mathcal{R}'}$ be two sequences of output models resulting respectively from performing two sequences of the gradient updates $G(\theta^1_{\mathcal{R}}), \cdots G(\theta^T_{\mathcal{R}})$ and $G(\theta^1_{\mathcal{R}'}), \cdots G(\theta^T_{\mathcal{R}'})$ applied to the training sets $\mathcal{S}$ with two different random sets $\mathcal{R}, \mathcal{R}'$. Assume sets $\mathcal{R}, \mathcal{R}'$ differ only at two element and the initialisation weights $\theta^0_{\mathcal{S}} = \theta^0_{\mathcal{S}'}$. Let $\Delta^t = \|\theta^t_R - \theta^t_{R'}\|$. At each step t, there are two cases when two samples $\mathbf{z}$ and $\mathbf{z}'$ are picked by SGD by the permutation order in $\mathcal{R}$ and $\mathcal{R}'$ respectively: 1) $\mathbf{z}$ and $\mathbf{z}'$ are the same with probability $1 - 2/N$, which implies $G^t_R = G^t_{R'}$, 2) $\mathbf{z}$ and $\mathbf{z}'$ are different with probability $2/N$. Following the same chain of equations (38)-(40) results in:

$$\begin{aligned}
&\left|\ell(\theta^T_{\mathcal{S},\mathcal{R}}; \mathbf{z}) - \ell(\theta^T_{\mathcal{S},\mathcal{R}'}; \mathbf{z})\right| \\
&\quad \leq \left|\ell(<\theta^T_{\mathcal{S},\mathcal{R}}, \mathbf{x}>, \mathbf{y})) - \ell(<\theta^T_{\mathcal{S},\mathcal{R}'}, \mathbf{x}>, \mathbf{y}))\right| \\
&\quad \leq \gamma\left\|<\theta^T_{\mathcal{S},\mathcal{R}}, \mathbf{x}> - <\theta^T_{\mathcal{S},\mathcal{R}'}, \mathbf{x}>\right\| \\
&\quad \leq \gamma\left\|\theta^T_{\mathcal{S},\mathcal{R}} - \theta^T_{\mathcal{S},\mathcal{R}'}\right\| = \gamma\Delta^T \leq \frac{4\gamma^2}{T}\sum_{t=1}^{T}\lambda_t,
\end{aligned} \tag{41}$$

The inequality (41) implies the desired inequality. ■

Theorem 5. *Consider a loss function ℓ such that $0 \leq \ell(f(\cdot; \mathbf{z}) \leq L$ for any point $\mathbf{z}$. Suppose that the SGD update rule is executed for T iterations with an annealing learning rate λ_t. Then, we have the following generalization error bound with probability at least $1 - \delta$:*

$$\begin{aligned}
&R_{\text{true}}(f_{\mathcal{S},\mathcal{R}}) - R_{\text{emp}}(f_{\mathcal{S},\mathcal{R}}) \leq L\sqrt{\frac{\log(2/\delta)}{2N}} + \\
&2\gamma^2\sum_{t=1}^{T}\lambda_t\left(2\sqrt{\frac{\log(2/\delta)}{T}} + \sqrt{\frac{2\log(2/\delta)}{N}} + \frac{1}{N}\right).
\end{aligned} \tag{42}$$

Proof. These inequality immediately follows from combining Theorem 3 and Theorem 4. ■

Corollary 2. *Consider two models $f^{GJM}_{\mathcal{S},\mathcal{R}}$ and $f^{KL}_{\mathcal{S},\mathcal{R}}$ trained under the same settings using the GJM and KL loss functions, respectively, using the training set, $\mathcal{S}$. We have the following inequality:*

$$E(f^{GJM}_{\mathcal{S},\mathcal{R}}) \leq E_{CE}(f^{KL}_{\mathcal{S},\mathcal{R}}), \tag{43}$$

where $E(f_{\mathcal{S},\mathcal{R}}) = R_{\text{true}}(f_{\mathcal{S},\mathcal{R}}) - R_{\text{emp}}(f_{\mathcal{S},\mathcal{R}})$.

Proof. These inequalities immediately follow from Corollary 1 and Theorem 2. From Corollary 1, while T and N are fixed, the second term is always smaller for GJM. From Theorem 2, $\ell_{GJN} \leq \ell_{KL}$. This implies that $L_{GJN} \leq \ell_{GJM}$. ■

References

Gibbs, A. L. and Su, F. E. On choosing and bounding probability metrics. *International Statistical Review*, 70 (3):419–435, 2002.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1225–1234, New York, New York, USA, Jun 2016.