Generalization Error Bound for an SGD Family via a Gaussian Approximation Method

Hao Chen  
University of Science and Technology of China

Zhanfeng Mo  
Nanyang Technological University

Zhouwang Yang (yangzw@ustc.edu.cn)  
University of Science and Technology of China

Research Article

Keywords:

Posted Date: June 7th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3020104/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Generalization Error Bound for an SGD Family via a Gaussian Approximation Method

Zhouwang Yang$^1*$, Hao Chen$^1$ and Zhanfeng Mo$^2$

$^1$University of Science and Technology of China, Hefei, 230026, China.
$^2$Nanyang Technological University, 639798, Singapore.

*Corresponding author(s). E-mail(s): yangzw@ustc.edu.cn; ch330822@mail.ustc.edu.cn; ZHANFENG001@ntu.edu.sg;

Proof of Proposition 1

Proof. (1): Since $u_t$ is uniformly bounded, $\exists C \in \mathbb{R}^{p \times p}, C \succ 0$ such that $\text{Cov}(u_t) \prec C$ holds for any $t$. Then we have

$$\text{Cov}(\theta_{\infty}) = \alpha^2 \sum_{t \geq 0} (I - \alpha H_S)^t \text{Cov}(u_t) (I - \alpha H_S)^t$$

$$\leq \alpha^2 \sum_{t \geq 0} \lambda_{\max}^2 (I - \alpha H_S) C$$

$$= \frac{\alpha^2}{1 - \lambda_{\max}^2 (I - \alpha H_S)} C$$

$$= O(\alpha).$$

(2): Let $\phi_{\theta_t}$ be the characteristic function of $\theta_t$, thus

$$\phi_{\theta_{\infty}}(s) = \prod_{t \geq 0} \phi_{u_t}(\alpha(I - \alpha H_S)^t s)$$

$$= \prod_{t \geq 0} (1 - \alpha^2 s^\top (I - \alpha H_S)^t \text{Cov}(u_t) (I - \alpha H_S)^t s + o(\alpha^2 \|s\|^2_2))$$

$$= 1 - s^\top \text{Cov}(\theta_{\infty}) s + o(\|s\|^2_2 \alpha^2).$$

1
By the proof of (1), \(\phi_{\theta_0}(s) \to 1 - s^\top \text{Cov}(\theta_0) s\) as \(\alpha \to 0\), thus \(\alpha^{-1/2}(P(\alpha) - \hat{P}(\alpha)) \xrightarrow{\text{law}} 0\).

(3): Let event \(A = \{\theta \mid \|\theta[i] - \theta_0[i]\| \leq K \sqrt{\Sigma[i][i]}, i = 1, \ldots, p\}\).

\[
\mathcal{W}(P|_{\Theta}, \hat{P}|_{\Theta}) = \inf_{F_{\theta_1}=F_{\theta_2}=F_{\hat{\theta}_0}, F_{\hat{\theta}_0}=F_{\hat{\theta}_1}} \mathbb{E}_{\theta_1, \theta_2} \|\theta_1 - \theta_2\|_1 \\
\leq \inf_{F_{\theta_1}=F_{\theta_2}=F_{\hat{\theta}_0}} \mathbb{E}_{\theta_1, \theta_2} \|\theta_1 - \theta_2\|_1 \cdot \chi_A(\theta_1) \cdot \chi_A(\theta_2) \\
\leq \inf_{F_{\theta_1}=F_{\theta_2}=F_{\hat{\theta}_0}} \frac{p}{\int_{\theta_0[i] - K \sqrt{\Sigma[i][i]} \leq x < \theta_0[i] + K \sqrt{\Sigma[i][i]}} |F_{\hat{\theta}_0}(x) - F_{\theta_1}(x)| \, dx \\
\leq 2K \sum_{i=1}^{p} \sqrt{\Sigma[i][i]} \cdot \mathbb{E}_{\theta_1}[\|\theta_1[i]\|/\sqrt{\Sigma[i][i]}]^3 \\
\leq 2\hat{C}K \sum_{i=1}^{p} (\Sigma[i][i])^{-1} \cdot \left( \sum_{i=0}^{\alpha^3(1 - \alpha \lambda_{\text{min}}(H_S))^{3/2}} \right) [i] \\
\leq \frac{2\alpha^2\hat{C}K\Gamma \text{tr}(\Sigma^{-1})}{3\lambda_{\text{min}}(H_S)} 
\]

where \(F_{\theta_1}\) is the cumulative function of \(\theta[i]\), (4) is obtained by Berry-Essen inequality. \(\square\)

**Proof of Lemma 2**

Proof. Let’s start with a claim: Suppose the parameter space \(\Theta\) is compact, for \(\forall \delta \in (0, 1)\), with probability of at least \(1 - \delta\) over the choice of \(S\), there exists a constant \(C(\delta, \Theta)\) such that \(\|L_S - L\|_{\text{lip}} \leq C(\delta, \Theta)/\sqrt{n}\).

Proof of claim: By CLT, as \(n \to \infty\),

\[
\left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla \ell(f_\theta(x_i), y) - \nabla \ell(L(\theta)) \right) \overset{d}{\to} N(0, \text{Cov}(\nabla \ell(f_\theta(x), y))).
\]

Hence, by standard Chebyshev inequality, for \(\forall \delta \in (0, 1)\), with probability of at least \(1 - \delta\) over the choice of \(S\), we have

\[
\sup_{\theta \in \Theta} \|\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla \ell(f_\theta(x_i), y) - \nabla \ell(L(\theta))\|^2 \leq \sup_{\theta \in \Theta} \text{tr}(\text{Cov}(\nabla \ell(f_\theta(x), y)))/\delta n,
\]

where \(\Theta\) is the compact parameter space. Then, the proof is completed by taking

\[
C(\delta, \Theta) = 2 \sqrt{\sup_{\theta \in \Theta} \text{tr}(\text{Cov}(\nabla \ell(f_\theta(x), y)))/\delta}.
\]

Now let’s move on to the proof of Lemma 2:

\[
|(L(P) - L_S(P)) - (L(\hat{P}) - L_S(\hat{P}))|
\]
\[ \begin{align*}
&= \|\mathbb{E}_{\theta \sim P}(L(\theta) - L_S(\theta)) - \mathbb{E}_{\theta \sim P}(L(\theta) - L_S(\theta))\| \\
&\leq \mathbb{E}\{P(A^c) \cdot \hat{P}(A^c) \} \cdot \sup_{\theta \in \Theta} |L(\theta)| \tag{8} \\
&\leq \rho \mathbb{W}(P|\Theta, \hat{P}|\Theta) + \max\{P(A^c), \hat{P}(A^c)\} \cdot \sup_{\theta \in \Theta} |L(\theta)| \tag{9} \\
&\leq \frac{2C_1^2 \alpha^2 \hat{C} \mathbb{K} \mathbb{L}}{\sqrt{\log(1 + \frac{1}{\hat{C} \mathbb{K} \mathbb{L}})}} + \frac{2p}{K^2} e^{-K^2/2} \tag{10} \\
&\leq \frac{2C_1^2 \alpha^2 \hat{C} \mathbb{K} \mathbb{L}}{\sqrt{\log(1 + \frac{1}{\hat{C} \mathbb{K} \mathbb{L}})}} + C_2^2 \frac{p}{K} e^{-K^2/2}, \tag{11} \\
&\leq 2C_1^2 \alpha^2 \frac{\hat{C} \mathbb{K} \mathbb{L}}{\sqrt{\alpha}} + C_2^2 \frac{p}{K} e^{-K^2/2}, \tag{12}
\end{align*} \]

where \( C_1 \triangleq \frac{2C_1^2 \alpha^2 \hat{C} \mathbb{K} \mathbb{L}}{\sqrt{\log(1 + \frac{1}{\hat{C} \mathbb{K} \mathbb{L}})}} \), \( C_2 \triangleq \sup_{\theta \in \Theta} |L(\theta)| \cdot \sqrt{\frac{2}{\alpha}} \). Let \( K \triangleq \sqrt{2 \log(\frac{\hat{C} \mathbb{K} \mathbb{L}}{\alpha})} \), we have

\[ \begin{align*}
&\frac{2C_1^2 \alpha^2 \hat{C} \mathbb{K} \mathbb{L}}{\sqrt{\alpha}} + C_2^2 \frac{p}{K} e^{-K^2/2}, \\
&= (L(P) - L_S(P)) - (L(\hat{P}) - L_S(\hat{P})) \leq C_1^2 \alpha^2 (2 \log(\frac{C_2 p}{C_1^2 \alpha}) + 2 \log(\frac{C_2 p}{C_1^2 \alpha})).
\end{align*} \]

\( \square \)

**Proof of Lemma 3**

*Proof.* Let \( \hat{P} = N(\theta^*, \Sigma) \), by definition,

\[ D_{KL}(\hat{P}||\sigma(S)^{-1}) \]

\[ \leq D_{KL}(\hat{P}||\hat{P}) \]

\[ \leq \frac{1}{2} \int_{\theta \in \Theta} \frac{|\Sigma_S|}{|\Sigma|} (\theta - \theta^*)^\top (\Sigma^{-1} - \Sigma_S^{-1})(\theta - \theta^*) + 2(\theta - \theta^*)^\top \Sigma^{-1}(\theta^* - \theta^*) \]

\[ + (\theta^* - \theta^*)^\top \Sigma^{-1}(\theta^* - \theta^*) d\theta \]

\[ = - \frac{1}{2} \log |\Sigma^{-1} \Sigma_S| + \frac{1}{2} \text{tr}(\Sigma^{-1} \Sigma_S - I) + \frac{1}{2} (\theta^* - \theta^*)^\top \Sigma^{-1}(\theta^* - \theta^*). \]

Let \( 0 < a_1 \leq a_2 \leq \ldots \leq a_k \leq 1 \leq a_{k+1} \leq \ldots \leq a_p \) be the eigenvalues of \( M_S \triangleq \Sigma^{-1} \Sigma_S \), thus

\[ D_{KL}(\hat{P}||\hat{P}) = \frac{1}{2} \sum_{i=1}^{p} (- \log a_i + a_i - 1) + \frac{1}{2}(\theta^* - \theta^*)^\top \Sigma^{-1}(\theta^* - \theta^*). \]

Since \( - \log(1 - x^{1/2}) + (1 - x^{1/2}) - 1 \) is convex for \( x \in (0, (1 - a_2)^2) \) and \( - \log(1 + x^{1/2}) + (1 + x^{1/2}) - 1 \) is concave for \( x > 0 \),

\[ - \log(1 - x^{1/2}) + (1 - x^{1/2}) - 1 < \frac{- \log a_k + a_k - 1}{(1 - a_k)^2} x \]

\[ - \log(1 + x^{1/2}) + (1 + x^{1/2}) - 1 < \frac{1}{2(1 + \sqrt{x_0})(x - x_0)} + \log(1 + \sqrt{x_0}) + (1 + \sqrt{x_0}) - 1. \]
Where \( x_0 = \frac{V_2}{p-k} \). Therefore,

\[
\sum_{i=1}^k - \log a_i + a_i - 1 \leq \frac{- \log a_* + a_* - 1}{(1-a_*)^2} V_1, \\
\sum_{i=k+1}^p - \log a_i + a_i - 1 \leq -(p-k) \log(1 + \sqrt{\frac{V_2}{p-k}}) + (p-k) \sqrt{\frac{V_2}{p-k}} \leq V_2,
\]

where \( V_1 = \sum_{i=1}^k (a_i - 1)^2 \), \( V_2 = \sum_{i=k+1}^p (a_i - 1)^2 \). Combine 19 and 20 we have

\[
E SD KL(\hat{P} || \sigma(S)^\perp) \leq \frac{1}{2} \max \{- \log a_* + a_* - 1, 1\} M + \frac{1}{2} (\theta_* - \theta^*)^\top \Sigma^{-1} (\theta_* - \theta^*).
\]

The final result follows the Chebyshev’s inequality.

Proof of Proposition 3

Proof. (1): Since \( u_t \) is uniformly bounded, \( \exists C \in \mathbb{R}^{p \times p}, C \succ 0 \) such that \( \text{Cov}(u_t) \prec C \) holds for any \( t \). Then we have

\[
\text{Cov}(\theta_T) = \sum_{t=0}^{T-1} \alpha^2 (I - \alpha H_S)^{T-t-1} \text{Cov}(u_t) (I - \alpha H_S)^{T-t-1} \\
\leq T \alpha^2 C \\
= \mathcal{O}(T \alpha^2).
\]

(2): Let \( \phi_x \) be the characteristic function of \( x \), thus

\[
\phi_{\theta_T - \theta}(s) = \prod_{t=0}^{T-1} \phi_{u_t} (\alpha (I - \alpha H_S)^t s) \\
= \prod_{t=0}^{T-1} (1 - \alpha^2 s^\top (I - \alpha H_S)^t \text{Cov}(u_t) (I - \alpha H_S)^t s + o(\alpha^2 \|s\|^2_2)) \\
= 1 - s^\top \text{Cov}(\theta_T) s + o(\|s\|^2_2 \alpha^2),
\]

By the proof of (1), \( \phi_{\theta_*} (s) \to 1 - s^\top \text{Cov}(\theta_* \infty) s \) as \( \max \alpha_t \to 0 \), thus \( \frac{\sum_{t=0}^{T-1} \alpha^2}{(\sum_{t=0}^{T-1} \alpha^2)^{-1/2}(\hat{P}(\alpha) - \hat{P}(\alpha))} \to 0 \).

(3): Without loss of generality, assume that the eigenvector direction of \( H_S \) is consistent with the coordinate axis. Let event \( A = \{ \theta | \|\theta[i] - \theta_* S[i]\| \leq K \sqrt{\Sigma[i]} \|i\|, i = 1, \ldots, p \} \).

\[
\mathcal{W}(P|\theta_K, \hat{P}|\theta_K) = \inf_{F_{\theta_1}=F_{\theta_2}=F_{\hat{P}|\theta_K}} \mathbb{E}_{\theta_1, \theta_2} \|\theta_1 - \theta_2\|_1
\]
where (26) is obtained by Berry–Essen inequality.

Proof of Lemma 4

Proof.

\[ |(L(P) - L_S(P)) - (L(\hat{P}) - L_S(\hat{P}))| \]  
\[ \leq \inf_{F_{\theta_1}, F_{\theta_2}} \mathbb{E}_{\theta_1, \theta_2} [||\theta_1 - \theta_2||_1 \cdot \chi_A(\theta_1) \cdot \chi_A(\theta_2)] \]  
\[ \leq \inf_{F_{\theta_1}, F_{\theta_2}} \mathbb{E}_{\theta_1, \theta_2} \sum_{i=1}^{p} \int_{S_i}^{\delta_2} \sqrt{\sum_{i}^{\delta_2}} |F_{\theta_2}(x) - F_{\theta_1}(x)| \text{d}x \]  
\[ \leq 2K \sum_{i=1}^{p} \sqrt{\sum_{i}^{\delta_2}} \cdot \tilde{C} \sqrt{\theta_i} \sqrt{\sum_{i}^{\delta_2}} \]  
\[ \leq 2\tilde{C} K \left( \sum_{i=1}^{q} (\sum_{i}^{\delta_2})^{-1} \cdot \left( \sum_{t=0}^{T-1} \alpha_{3} (1 - \alpha_{3} \lambda_{\text{min}}(H_{S}))^{3} \Gamma_{i} \right) \right) \]  
\[ + \sum_{t=q+1}^{p} (\sum_{i}^{\delta_2})^{-1} \cdot \left( \sum_{t=0}^{T-1} \alpha_{3} \mathbb{E} \left| u_{i} \right|^{3} \right) \]  
\[ \leq \tilde{C} K \left( \frac{\alpha_{3}}{3 \lambda_{\text{min}}} + \sum_{i=1}^{\delta_2} \alpha_{i}^{3} \right) \]  
(28)

where (26) is obtained by Berry–Essen inequality.

\[ \quad \]

Proof.

\[ (L(P) - L_S(P)) - (L(\hat{P}) - L_S(\hat{P})) \]  
\[ = |\mathbb{E}_{\theta \sim P} L(\theta) - \mathbb{E}_{\theta \sim \hat{P}} L(\hat{\theta})| \]  
\[ \leq |\mathbb{E}_{\theta \sim P} [L(\theta) - L(\hat{\theta})]| + \max_{\theta \in \Theta} \sup_{\theta \in \Theta} |L(\theta)| \]  
(31)

\[ \leq \frac{C(\delta)}{\sqrt{n}} \mathcal{W}(1) (P|\Theta, \hat{P}|\Theta_{\hat{\kappa}}) + \max \{P(A^{c}), \hat{P}(A^{c})\} \sup_{\theta \in \Theta} |L(\theta)| \]  
(32)

\[ \leq 2C(\delta) K \left( \frac{\alpha_{3}}{3 \lambda_{\text{min}}} + \sum_{i=1}^{\delta_2} \alpha_{i}^{3} \right) \]  
\[ + \sup_{\theta \in \Theta} |L(\theta)| \cdot \frac{2p}{K^{2}} e^{-K^{2}/2} \]  
(33)

\[ \triangleq C_1 K + C_2 \frac{p}{K^{2}} e^{-K^{2}/2} \]  
(34)

where \( C_1 \triangleq \frac{2C(\delta) K \left( \frac{\alpha_{3}}{3 \lambda_{\text{min}}} + \sum_{i=1}^{\delta_2} \alpha_{i}^{3} \right)}{\sqrt{n}} \), \( C_2 \triangleq \sup_{\theta \in \Theta} |L(\theta)| \cdot \sqrt{\frac{2}{\pi}} \). Let \( K \triangleq \sqrt{2 \log \left( \frac{C_{1} p}{C_{2}} \right)} \),

we have

\[ |(L(P) - L_S(P)) - (L(\hat{P}) - L_S(\hat{P}))| \leq C_1 \left( 2 \log \left( \frac{C_{2} p}{C_{1}} \right) + \sqrt{2 \log \left( \frac{C_{2} p}{C_{1}} \right)} \right)^{-1} \]  

\[ \square \]
Proof of Proposition 5
(1) Additive Noise Insertion: By substituting \( u \) in the proof of Proposition 1 with \( u - \eta \), then our conclusion directly follows 
\[
\text{Cov}(u - \eta) = \text{Cov}(u) + \text{Var}(\eta[1])I.
\]
(2) Multiplicative Noise Insertion: The dynamic of SGD with multiplicative noise is:
\[
\theta_{t+1} = \theta_t - \alpha \gamma(t) \odot g_B = (I - \alpha H_S \odot \gamma(t)) \theta_t - \alpha \gamma(t) \odot u_t
\]
\[
\implies \theta_T = \sum_{t=0}^{T-1} \prod_{i=t+1}^{T-1} (I - \alpha H_S \odot \gamma(i)) \cdot \alpha \gamma(i) \odot u_t + \prod_{t=1}^{T} (I - \alpha H_S \odot \gamma(t)) \theta_0.
\]
By taking the covariance of \( \theta_T \), we have
\[
\text{Cov}(\theta_T) = \text{E}_{\gamma_T, u}[(I - \alpha H_S \odot \gamma(T)) \text{Cov}(\theta_{T-1})(I - \alpha H_S \odot \gamma(T))] + \text{Cov}(\alpha \gamma \odot u_t)
\]
\[
= (I - \alpha H_S) \text{Cov}(\theta_{T-1})(I - \alpha H_S) + \text{Cov}(\alpha \gamma \odot u_t) + O(\alpha^2 \text{Cov}(\theta_{T-1}))
\]
\[
\implies \lim_{\alpha \to 0} \alpha^{-1} \text{Cov}(\theta_T) = \lim_{\alpha \to 0} \alpha^{-1} (I - \alpha H_S) \text{Cov}(\theta_{T-1})(I - \alpha H_S) + \alpha^{-1} \text{Cov}(\alpha \gamma \odot u_t)
\]
\[
= \lim_{\alpha \to 0} \alpha \sum_{t=0}^{T} (I - \alpha H_S)^t C'(I - \alpha H_S)^t,
\]
where \( C' = (C + (E\gamma_0[1]^2 - 1) \text{diag}(C)) \). By taking \( T = \infty \), we have
\[
\lim_{\alpha \to 0} \alpha^{-1} \text{Cov}(\theta_{\infty}) = \alpha \sum_{t \geq 0} (I - \alpha H_S)^t C'(I - \alpha H_S)^t.
\]

Appendix B. Detailed Experiments

Experimental Settings

Our experiments on neural networks are conducted on different models and different datasets, namely MNIST [1] and CIFAR-10 [2]. On MNIST dataset, we train a three-layer network (model 1) with \((784 \times 200 \text{ FC})\text{-ReLU-}(200 \times 200 \text{ FC})\text{-ReLU-}(200 \times 10 \text{ FC})\), where FC denotes a fully connected layer. We use the optimizer of SGD with batch size=200 and learning rate=0.01 for the network. For CIFAR-10 dataset, we use a convolution network (model 2) with \((3 \times 6 \times 5 \times 5)\text{-ReLU-MP2-(6 \times 16 \times 5 \times 5)-ReLU-MP2-(400 \times 120 \text{ FC})-ReLU-(84 \times 120 \times 84 \text{ FC})-ReLU-(84 \times 10 \text{ FC})\), where \((5 \times 5)\) denotes a 5 × 5 convolution layer and MP2 denotes a 2 × 2 max pooling layer. The optimizer of SGD is used again but the settings changes to batch size=4 and learning rate=0.001. Experiments are executed as follows:

1. Initialize the model at a fixed point in the vicinity of the optima. In each experiments, we get this fixed point by training 5 epochs on model 1 and 10 epochs in model 2 with a Xavier and Kaiming initialization [3].
2. Train the models until the training loss and accuracy are stable. We train 30 epochs on model 1 and 50 epochs on model 2.

3. Repeat the second step for 3000 times and collect the parameters of the final epochs. We obtain \( \{\theta^{(i)}_{\text{MNIST}}\}_{i=1}^{3000}, \{\theta^{(i)}_{\text{CIFAR10}}\}_{i=1}^{3000} \).

4. Take MNIST for example, for each marginal \( j = 1, \ldots, p_{\text{MNIST}} \) with \( p_{\text{MNIST}} = 198800 \), we perform the Person test on \( \{\theta^{(i)}_{\text{MNIST}}[j]\}_{i=1}^{3000} \) to check where marginal-Gaussianity holds for the \( j \)th dimension. This results to 198800 marginal p-values. At a confidential level of \( 1 - \delta \), we reject the null hypothesis that the \( j \)th marginal is Gaussian if the corresponding p-value is smaller than \( \delta \). The same procedures are conducted on CIFAR-10.

5. Take MNIST for example, we calculate the percentage of the marginals with p-values smaller than a given threshold, which takes values in \( \{0.001, 0.002, \ldots, 0.999, 1\} \). Then we obtain the percentage v.s. p-values thresholds plot, which show us how the marginal Gaussianity is violated at different confidential level.

All these procedures are repeated 5 times.

Experimental Results

Experiments on Two-Dimensional Loss functions

The scatter plots (see Figure 1 and Figure 2) of the limiting parameter distributions again coincide with our understanding: the limiting parameter distribution of SGD with non-Gaussian gradient noise tends to be Gaussian-like. To further examine the two-dimensional Gaussianity of the limiting distribution, the aforementioned procedures with a random initialization \( \{\theta_0[1], \theta_0[2]\} \sim \text{U}(0,1) \) are repeated 30 times. For each initialization, we perform the Henze-Zirkler multivariate normality test on the limiting distributions. We then collect the p-values of each repetition. As we can see in Figure 3, Figure 4 and Figure 5, there is no statistically significant evidence against the null hypothesis that the limiting distribution is Gaussian.

Experiments on Neural Networks

For a given threshold (horizontal-axis), we calculate the percentage (vertical-axis) of marginals with p-values smaller than the threshold. The horizontal-axis of the lower figure is log-scaled. Table 1 shows that the marginal-Gaussianity holds for most of the dimensions and strongly suggests that the limited distributions of parameters are Gaussian-like.

References


For each gradient noise implements (including adding uniform, exponential and binomial gradient noise) and each loss functions $f_1, f_2$ and $f_3$, experiments are ran with $\alpha = 0.01$ and $\theta_0 = (1, 1)^T$. We visualize the empirical limiting distribution by a 2D-kernel density plot.

The gradient noise is fixed to be exponential. For each loss functions $f_1$, $f_2$, $f_3$, the experiments are ran with $\alpha \in \{0.1, 0.01, 0.001\}$ and a fixed initialization $\theta_0 = (1, 1)^T$. We visualize the empirical limiting distribution by a 2D-kernel density plot.
Fig. 3 For loss functions $f_1, f_2, f_3$, we perform SGDs with uniformly distributed gradient noise and $\alpha = 0.01$. At the confidential level of 0.99, about 29/30 of the 30 repetitions fail to provide statistically significant evidence against the two-dimensional Gaussianity of the limiting parameter distributions.

Fig. 4 For loss functions $f_1, f_2, f_3$, we perform SGDs with exponentially distributed gradient noise and $\alpha = 0.001$. At the confidential level of 0.99, about 29/30 of the 30 repetitions fail to provide statistically significant evidence against the two-dimensional Gaussianity of the limiting parameter distributions.
Fig. 5 For loss functions $f_1, f_2, f_3$, we perform SGDs with binomially distributed gradient noise and $\alpha = 0.01$. At the confidential level of 0.99, all of the 30 repetitions fail to provide statistically significant evidence against the two-dimensional Gaussianity of the limiting parameter distributions.

Table 1 For each experiment, we calculate the percentages of dimensions with marginal p-values smaller than 0.1, 0.05 and 0.01, respectively. For a marginal with a p-value smaller than $\delta \in (0, 1)$, we can reject the null hypothesis that this marginal follows a Gaussian distribution at a confidential level of $1 - \delta$. As we can see, at the confidential level of 0.99, marginal Gaussianity holds for most of the marginals.

<table>
<thead>
<tr>
<th>Percentage</th>
<th>$\leq 0.1$</th>
<th>$\leq 0.05$</th>
<th>$\leq 0.01$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST Exp. 1</td>
<td>10.8%</td>
<td>5.7%</td>
<td>1.5%</td>
</tr>
<tr>
<td>MNIST Exp. 2</td>
<td>11.8%</td>
<td>6.9%</td>
<td>2.6%</td>
</tr>
<tr>
<td>MNIST Exp. 3</td>
<td>12.3%</td>
<td>7.2%</td>
<td>2.7%</td>
</tr>
<tr>
<td>MNIST Exp. 4</td>
<td>12.6%</td>
<td>7.5%</td>
<td>3.2%</td>
</tr>
<tr>
<td>MNIST Exp. 5</td>
<td>12.2%</td>
<td>7.0%</td>
<td>2.6%</td>
</tr>
<tr>
<td>CIFAR-10 Exp. 1</td>
<td>10.8%</td>
<td>5.7%</td>
<td>1.5%</td>
</tr>
<tr>
<td>CIFAR-10 Exp. 2</td>
<td>11.8%</td>
<td>6.9%</td>
<td>2.6%</td>
</tr>
<tr>
<td>CIFAR-10 Exp. 3</td>
<td>12.4%</td>
<td>7.2%</td>
<td>2.7%</td>
</tr>
<tr>
<td>CIFAR-10 Exp. 4</td>
<td>12.6%</td>
<td>7.5%</td>
<td>3.2%</td>
</tr>
<tr>
<td>CIFAR-10 Exp. 5</td>
<td>12.2%</td>
<td>7.0%</td>
<td>2.6%</td>
</tr>
</tbody>
</table>
Experiment 1 on MNIST

Experiment 2 on MNIST

Data

MNIST

p-value
Percentage

0.001 0.005 0.01 0.05 0.1 1

0.00 0.25 0.50 0.75 1.00

0.00 0.05 0.1 0.5 1

12
Experiment 5 on MNIST

Experiment 1 on CIFAR−10

Data

MNIST

CIFAR−10
Experiment 2 on CIFAR-10

Experiment 3 on CIFAR-10

Data

CIFAR-10
Experiment 4 on CIFAR-10

Experiment 5 on CIFAR-10
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.pdf