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;
Contributing authors: ch330822@mail.ustc.edu.cn;
ZHANFENG001@ntu.edu.sg;

Proof of Proposition 1

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

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{\theta}_\infty) =&\alpha^2 \sum_{t\geqslant 0}(\mathbf{I}-\alpha\mathbf{H}_{\mathcal{S}})^t \mathrm{Cov}(\mathbf{u}_t)(\mathbf{I}-\alpha\mathbf{H}_{\mathcal{S}})^t \\
\leqslant&\alpha^2 \sum_{t\geqslant 0}\lambda_{\max}^{2t}(\mathbf{I}-\alpha\mathbf{H}_{\mathcal{S}})\mathbf{C} \\
=&\frac{\alpha^2}{1-\lambda_{\max}^2(\mathbf{I}-\alpha\mathbf{H}_{\mathcal{S}})}\mathbf{C} \\
=&\mathcal{O}(\alpha).
\end{aligned}$$

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

$$\begin{aligned}
\phi_{\boldsymbol{\theta}_\infty}(\mathbf{s}) =& \prod_{t\geqslant 0}\phi_{u_t}(\alpha(\mathbf{I}-\alpha\mathbf{H}_{\mathcal{S}})^t\mathbf{s}) \\
=&\prod_{t\geqslant 0}(1-\alpha^2\mathbf{s}^\top(\mathbf{I}-\alpha\mathbf{H}_{\mathcal{S}})^t\mathrm{Cov}(\mathbf{u}_t)(\mathbf{I}-\alpha\mathbf{H}_{\mathcal{S}})^t\mathbf{s}+o(\alpha^2\|\mathbf{s}\|_2^2)) \\
=&1-\mathbf{s}^\top\mathrm{Cov}(\boldsymbol{\theta}_\infty)\mathbf{s}+o(\|\mathbf{s}\|_2^2\alpha^2),
\end{aligned}$$

By the proof of (1), $\phi_{\boldsymbol{\theta}_\infty}(\mathbf{s}) \to 1-\mathbf{s}^\top \mathrm{Cov}(\boldsymbol{\theta}_\infty)\mathbf{s}$ as $\alpha \to 0$, thus $\alpha^{-1/2}(P(\alpha)-\hat{P}(\alpha)) \overset{\text{law}}{\to} \mathbf{0}$.
(3): Let event $A = \{\boldsymbol{\theta} \mid \ |\boldsymbol{\theta}[i] - \boldsymbol{\theta}^*_{\mathcal{S}}[i]| \leqslant K\sqrt{\boldsymbol{\Sigma}[i][i]}, i = 1, ..., p\}$.

$$\begin{aligned}
\mathcal{W}^{(1)}(P|_{\Theta_K}, \hat{P}|_{\Theta_K}) = & \inf_{F_{\boldsymbol{\theta}_1}=F_{P|_{\Theta_K}}, F_{\boldsymbol{\theta}_2}=F_{\hat{P}|_{\Theta_K}}} \mathbb{E}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_1 & (1) \\
\leqslant & \inf_{F_{\boldsymbol{\theta}_1}=F_P, F_{\boldsymbol{\theta}_2}=F_{\hat{P}}} \mathbb{E}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}[\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_1 \cdot \chi_A(\boldsymbol{\theta}_1) \cdot \chi_A(\boldsymbol{\theta}_2)] & (2) \\
\leqslant & \inf_{F_{\theta_1}=F_P, F_{\boldsymbol{\theta}_2}=F_{\hat{P}}} \sum_{i=1}^{p} \int_{\theta^*_{\mathcal{S}}[i]-K\sqrt{\boldsymbol{\Sigma}[i][i]}}^{\theta^*_{\mathcal{S}}[i]+K\sqrt{\boldsymbol{\Sigma}[i][i]}} |F_{P_i}(x) - F_{\hat{P}_i}(x)| \mathrm{d}x & (3) \\
\leqslant & 2K \sum_{i=1}^{p} \sqrt{\boldsymbol{\Sigma}[i][i]} \cdot C\mathbb{E}\left|\boldsymbol{\theta}[i]/\sqrt{\Sigma[i][i]}\right|^3 & (4) \\
\leqslant & 2\tilde{C}K \sum_{i=1}^{p} (\boldsymbol{\Sigma}[i][i])^{-1} \cdot \big(\sum_{t\geqslant 0} \alpha^3(1-\alpha\lambda_{\min}(\mathbf{H}_{\mathcal{S}}))^{3t}\Gamma\big)[i] & (5) \\
\leqslant & \frac{2\alpha^2\tilde{C}K\boldsymbol{\Gamma}}{3\lambda_{\min}(\mathbf{H}_{\mathcal{S}})} \mathrm{tr}(\boldsymbol{\Sigma}^{-1}), & (6)
\end{aligned}$$

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

Proof of Lemma 2

Proof. Let's start with a claim: Suppose the parameter space Θ 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 $\|L_S - L\|_{\mathrm{lip}} \leqslant C(\delta, \Theta)/\sqrt{n}$.

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

$$\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla_\theta l(f_\theta(x_i), y) - \nabla_\theta L(\theta)\right) \overset{d}{\to} \mathcal{N}(0, \mathrm{Cov}(\nabla_\theta l(f_\theta(x), y))).$$

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

$$\sup_{\theta\in\Theta} \|\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla_\theta l(f_\theta(x_i), y) - \nabla_\theta L(\theta)\|^2 \leqslant \sup_{\theta\in\Theta} \mathrm{tr}(\mathrm{Cov}(\nabla_\theta l(f_\theta(x), y)))/(\delta n),$$

where Θ is the compact parameter space. Then, the proof is completed by taking

$$C(\delta, \Theta) = 2\sqrt{\sup_{\theta\in\Theta} \mathrm{tr}(\mathrm{Cov}(\nabla_\theta l(f_\theta(x), y)))/\delta}.$$

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

$$|(L(P) - L_{\mathcal{S}}(P)) - (L(\hat{P}) - L_{\mathcal{S}}(\hat{P}))| \quad (7)$$

$$=|\mathbb{E}_{\theta\sim P}(L(\boldsymbol{\theta})-L_{\mathcal{S}}(\boldsymbol{\theta}))-\mathbb{E}_{\boldsymbol{\theta}\sim\hat{P}}(L(\boldsymbol{\theta})-L_{\mathcal{S}}(\boldsymbol{\theta}))| \tag{8}$$

$$\leqslant|\mathbb{E}_{\boldsymbol{\theta}\sim P|_{\boldsymbol{\Theta}_K}}L(\boldsymbol{\theta})-\mathbb{E}_{\boldsymbol{\theta}\sim\hat{P}|_{\boldsymbol{\Theta}_K}}L(\boldsymbol{\theta})|+\max\{P(A^c),\hat{P}(A^c)\}\cdot\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_K}|L(\theta)| \tag{9}$$

$$\leqslant\rho\mathcal{W}^{(1)}(P|_{\boldsymbol{\Theta}_K},\hat{P}|_{\boldsymbol{\Theta}_K})+\max\{P(A^c),\hat{P}(A^c)\}\cdot\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_K}|L(\boldsymbol{\theta})| \tag{10}$$

$$\leqslant\frac{2C(\delta)\alpha^2\tilde{C}K\Gamma}{3\lambda_{\min}(\mathbf{H}_{\mathcal{S}})\sqrt{n}}\mathrm{tr}(\Sigma^{-1})+\sup_{\theta\in\Theta}|L(\theta)|\cdot\frac{2p}{K\sqrt{2\pi}}e^{-K^2/2} \tag{11}$$

$$\triangleq C_1\alpha^2K+C_2\frac{p}{Ke^{K^2/2}}, \tag{12}$$

where $C_1\triangleq\frac{2C(\delta)\tilde{C}\Gamma}{3\lambda_{\min}(\mathbf{H}_{\mathcal{S}})\sqrt{n}}\mathrm{tr}(\Sigma^{-1})$, $C_2\triangleq\sup_{\theta\in\Theta}|L(\theta)|\cdot\sqrt{\frac{2}{\pi}}$. Let $K\triangleq\sqrt{2\log(\frac{C_2p}{C_1\alpha})}$, we have

$$|(L(P)-L_{\mathcal{S}}(P))-(L(\hat{P})-L_{\mathcal{S}}(\hat{P}))|\leqslant C_1\alpha^2(\sqrt{2\log(\frac{C_2p}{C_1\alpha})}+\sqrt{2\log(\frac{C_2p}{C_1\alpha})}^{-1}).$$

□

Proof of Lemma 3

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

$$D_{\mathrm{KL}}(\hat{P}\|\sigma(\mathcal{S})^{\perp}) \tag{13}$$

$$\leqslant D_{\mathrm{KL}}(\hat{P}\|\bar{P}) \tag{14}$$

$$=\frac{1}{2}\int_{\theta\in\Theta}-\log\frac{|\Sigma_{\mathcal{S}}|}{|\Sigma|}+(\theta-\theta^*_{\mathcal{S}})^\top(\Sigma^{-1}-\Sigma_{\mathcal{S}}^{-1})(\theta-\theta^*_{\mathcal{S}})+2(\theta-\theta^*_{\mathcal{S}})^\top\Sigma^{-1}(\theta^*_{\mathcal{S}}-\theta^*) \tag{15}$$

$$+(\theta^*_{\mathcal{S}}-\theta^*)^\top\Sigma^{-1}(\theta^*_{\mathcal{S}}-\theta^*)\mathrm{d}\theta \tag{16}$$

$$=-\frac{1}{2}\log|\Sigma^{-1}\Sigma_{\mathcal{S}}|+\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\Sigma_{\mathcal{S}}-I)+\frac{1}{2}(\theta^*_{\mathcal{S}}-\theta^*)^\top\Sigma^{-1}(\theta^*_{\mathcal{S}}-\theta^*). \tag{17}$$

Let $0<a_*\leqslant a_1\leqslant...\leqslant a_k\leqslant 1\leqslant a_{k+1}\leqslant...\leqslant a_p$ be the eigenvalues of $\mathbf{M}_{\mathcal{S}}\triangleq\Sigma^{-1}\Sigma_{\mathcal{S}}$, thus

$$D_{\mathrm{KL}}(\hat{P}\|\bar{P})=\frac{1}{2}\sum_{i=1}^{p}(-\log a_i+a_i-1)+\frac{1}{2}(\theta^*_{\mathcal{S}}-\theta^*)^\top\Sigma^{-1}(\theta^*_{\mathcal{S}}-\theta^*). \tag{18}$$

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

$$-\log(1-x^{1/2})+(1-x^{1/2})-1<\frac{-\log a_*+a_*-1}{(1-a_*)^2}x$$

$$-\log(1+x^{1/2})+(1+x^{1/2})-1<\frac{1}{2(1+\sqrt{x_0}}(x-x_0)+-\log(1+\sqrt{x_0})+(1+\sqrt{x_0})-1.$$

Where $x_0 = \frac{V_2}{p-k}$. Therefore,

$$\sum_{i=1}^{k} -\log a_i + a_i - 1 \leqslant \frac{-\log a_* + a_* - 1}{(1-a_*)^2} V_1, \tag{19}$$

$$\sum_{i=k+1}^{p} -\log a_i + a_i - 1 \leqslant -(p-k)\log(1+\sqrt{\frac{V_2}{p-k}}) + (p-k)\sqrt{\frac{V_2}{p-k}} \leqslant V_2, \tag{20}$$

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

$$\mathbb{E}_{\mathcal{S}} D_{\mathrm{KL}}(\hat{P}\|\sigma(\mathcal{S})^{\perp}) \leqslant \frac{1}{2}\max\{\frac{-\log a_* + a_* - 1}{1-a_*}, 1\} M + \frac{1}{2}(\theta^*_{\mathcal{S}} - \theta^*)^{\top}\Sigma^{-1}(\theta^*_{\mathcal{S}} - \theta^*). \tag{21}$$

The final result follows the Chebyshev's inequality. □

Proof of Proposition 3

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

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{\theta}_T) =& \sum_{t=0}^{T-1} \alpha^2 (\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^{T-t-1}\mathrm{Cov}(\mathbf{u}_t)(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^{T-t-1} \\ \leqslant& T\alpha^2\mathbf{C} \\ =& \mathcal{O}(T\alpha^2).\end{aligned}$$

(2): Let ϕ_x be the characteristic function of $\boldsymbol{x}$, thus

$$\begin{aligned}\phi_{\boldsymbol{\theta}_T - \mathbb{E}[\boldsymbol{\theta}_T]}(\mathbf{s}) =& \prod_{t=0}^{T-1} \phi_{\boldsymbol{u}_t}(\alpha(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t \mathbf{s}) \\ =& \prod_{t=0}^{T-1}(1 - \alpha^2 \mathbf{s}^{\top}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t \mathrm{Cov}(\mathbf{u}_t)(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t\mathbf{s} + o(\alpha^2\|\mathbf{s}\|_2^2)) \\ =& 1 - \mathbf{s}^{\top}\mathrm{Cov}(\boldsymbol{\theta}_T)\mathbf{s} + o(\|\mathbf{s}\|_2^2\alpha^2),\end{aligned}$$

By the proof of (1), $\phi_{\boldsymbol{\theta}_\infty}(\mathbf{s}) \to 1 - \mathbf{s}^{\top}\mathrm{Cov}(\boldsymbol{\theta}_\infty)\mathbf{s}$ as $\max \alpha_t \to 0$, thus $(\sum_{t=0}^{T-1}\alpha_t^2)^{-1/2}(P(\alpha) - \hat{P}(\alpha)) \overset{\text{law}}{\to} \mathbf{0}$.
(3): Without loss of generality, assume that the eigenvector direction of $\mathbf{H}_{\mathcal{S}}$ is consistent with the coordinate axis. Let event $A = \{\boldsymbol{\theta} \mid |\boldsymbol{\theta}[i] - \boldsymbol{\theta}^*_{\mathcal{S}}[i]| \leqslant K\sqrt{\boldsymbol{\Sigma}[i][i]}, i = 1, ..., p\}$.

$$\mathcal{W}^{(1)}(P|_{\Theta_K}, \hat{P}|_{\Theta_K}) = \inf_{F_{\boldsymbol{\theta}_1} = F_{P|_{\Theta_K}}, F_{\boldsymbol{\theta}_2} = F_{\hat{P}|_{\Theta_K}}} \mathbb{E}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_1 \tag{22}$$

$$\leqslant \inf_{F_{\boldsymbol{\theta}_1}=F_P, F_{\boldsymbol{\theta}_2}=F_{\hat{P}}} \mathbb{E}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}[\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_1 \cdot \chi_A(\boldsymbol{\theta}_1) \cdot \chi_A(\boldsymbol{\theta}_2)] \tag{23}$$

$$\leqslant \inf_{F_{\theta_1}=F_P, F_{\theta_2}=F_{\hat{P}}} \sum_{i=1}^{p} \int_{\theta^*_{\mathcal{S}}[i]-K\sqrt{\boldsymbol{\Sigma}[i][i]}}^{\theta^*_{\mathcal{S}}[i]+K\sqrt{\boldsymbol{\Sigma}[i][i]}} |F_{P_i}(x) - F_{\hat{P}_i}(x)| \mathrm{d}x \tag{24}$$

$$\leqslant 2K \sum_{i=1}^{p} \sqrt{\boldsymbol{\Sigma}[i][i]} \cdot \tilde{C}\mathbb{E}\big|\boldsymbol{\theta}[i]/\sqrt{\Sigma[i][i]}\big|^3 \tag{25}$$

$$\leqslant 2\tilde{C}K\bigg(\sum_{i=1}^{q} (\boldsymbol{\Sigma}[i][i])^{-1} \cdot \big(\sum_{t=0}^{T-1} \alpha^3 (1-\alpha\tilde{\lambda}_{\min}(\mathbf{H}_{\mathcal{S}}))^{3t}\Gamma\big)[i] \tag{26}$$

$$+ \sum_{i=q+1}^{p} (\boldsymbol{\Sigma}[i][i])^{-1} \cdot \big(\sum_{t=0}^{T-1} \alpha^3 \mathbb{E}|\boldsymbol{u}_t[i]|^3\big)\bigg) \tag{27}$$

$$\leqslant \tilde{C}'K\bigg(\frac{\alpha\boldsymbol{\Gamma}}{3\tilde{\lambda}_{min}} + \frac{\sum_{i=1}^{T}\alpha_t^3}{\sum_{i=1}^{T}\alpha_t^2} \bigg), \tag{28}$$

where (26) is obtained by Berry-Essen inequality. □

Proof of Lemma 4

Proof.

$$|(L(P) - L_{\mathcal{S}}(P)) - (L(\hat{P}) - L_{\mathcal{S}}(\hat{P}))| \tag{29}$$

$$= |\mathbb{E}_{\theta\sim P} L(\boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{\theta}\sim\hat{P}} L(\boldsymbol{\theta})| \tag{30}$$

$$\leqslant |\mathbb{E}_{\boldsymbol{\theta}\sim P|_{\boldsymbol{\Theta}_K}} L(\boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{\theta}\sim\hat{P}|_{\boldsymbol{\Theta}_K}} L(\boldsymbol{\theta})| + \max\{P(A^c), \hat{P}(A^c)\} \cdot \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_K} |L(\theta)| \tag{31}$$

$$\leqslant \frac{C(\delta)}{\sqrt{n}} \mathcal{W}^{(1)}(P|_{\boldsymbol{\Theta}_K}, \hat{P}|_{\boldsymbol{\Theta}_K}) + \max\{P(A^c), \hat{P}(A^c)\} \cdot \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_K} |L(\boldsymbol{\theta})| \tag{32}$$

$$\leqslant \frac{2C(\delta)\tilde{C}K\bigg(\frac{\alpha\boldsymbol{\Gamma}}{3\tilde{\lambda}_{min}} + \frac{\sum_{i=1}^{T}\alpha_t^3}{\sum_{i=1}^{T}\alpha_t^2}\bigg)}{\sqrt{n}} + \sup_{\theta\in\Theta}|L(\theta)| \cdot \frac{2p}{K\sqrt{2\pi}} e^{-K^2/2} \tag{33}$$

$$\triangleq C_1 K + C_2 \frac{p}{Ke^{K^2/2}}, \tag{34}$$

where $C_1 \triangleq \frac{2C(\delta)\tilde{C}K\left(\frac{\alpha\boldsymbol{\Gamma}}{3\tilde{\lambda}_{min}} + \frac{\sum_{i=1}^{T}\alpha_t^3}{\sum_{i=1}^{T}\alpha_t^2}\right)}{\sqrt{n}}$, $C_2 \triangleq \sup_{\theta\in\Theta}|L(\theta)| \cdot \sqrt{\frac{2}{\pi}}$. Let $K \triangleq \sqrt{2\log(\frac{C_2 p}{C_1})}$, we have

$$|(L(P) - L_{\mathcal{S}}(P)) - (L(\hat{P}) - L_{\mathcal{S}}(\hat{P}))| \leqslant C_1\Big(\sqrt{2\log(\frac{C_2 p}{C_1})} + \sqrt{2\log(\frac{C_2 p}{C_1})}^{-1}\Big).$$

□

Proof of Proposition 5

(1) Additive Noise Insertion: By substituting $\mathbf{u}_t$ in the proof of Proposition 1 with $\mathbf{u} - \boldsymbol{\eta}_t$, then our conclusion directly follows $\text{Cov}(\mathbf{u} - \boldsymbol{\eta}_t) = \text{Cov}(\mathbf{u}) + \text{Var}(\boldsymbol{\eta}_0[1])\mathbf{I}$.

(2) Multiplicative Noise Insertion: The dynamic of SGD with multiplicative noise is:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha\boldsymbol{\gamma}^{(t)} \odot g_{B_t} = (I - \alpha\mathbf{H}_{\mathcal{S}} \odot \boldsymbol{\gamma}^{(t)})\boldsymbol{\theta}_t - \alpha\boldsymbol{\gamma}^{(t)} \odot \mathbf{u}_t$$

$$\Longrightarrow$$

$$\boldsymbol{\theta}_T = \sum_{t=0}^{T-1} \prod_{i=t+1}^{T-1} (\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}} \odot \boldsymbol{\gamma}^{(i)}) \cdot \alpha\boldsymbol{\gamma}^{(i)} \odot \mathbf{u}_t + \prod_{t=1}^{T}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}} \odot \boldsymbol{\gamma}^{(t)})\theta_0.$$

By taking the covariance of $\boldsymbol{\theta}_T$, we have

$$\begin{aligned}
\text{Cov}(\boldsymbol{\theta}_T) =& \mathbb{E}_{\boldsymbol{\gamma}_T,\mathbf{u}}[(\ \mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}} \odot \boldsymbol{\gamma}^{(T)})\text{Cov}(\boldsymbol{\theta}_{T-1})(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}} \odot \boldsymbol{\gamma}^{(T)})] + \text{Cov}(\alpha\boldsymbol{\gamma}_T \odot \mathbf{u}_t) \\
=& (\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})\text{Cov}(\boldsymbol{\theta}_{T-1})(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}}) + \text{Cov}(\alpha\boldsymbol{\gamma}_T \odot \mathbf{u}_t) + \mathcal{O}(\alpha^2\text{Cov}(\boldsymbol{\theta}_{T-1}))
\end{aligned}$$

$$\Longrightarrow$$

$$\begin{aligned}
\lim_{\alpha\to 0} \alpha^{-1}\text{Cov}(\boldsymbol{\theta}_T) &= \lim_{\alpha\to 0} \alpha^{-1}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})\text{Cov}(\boldsymbol{\theta}_{T-1})(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}}) + \alpha^{-1}\text{Cov}(\alpha\boldsymbol{\gamma}_T \odot \mathbf{u}_t) \\
&= \lim_{\alpha\to 0} \alpha \sum_{t=0}^{T}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t\mathbf{C}^{'}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t,
\end{aligned}$$

where $\mathbf{C}^{'} = (\mathbf{C} + (\mathbb{E}\boldsymbol{\gamma}_0[1]^2 - 1)\text{diag}(\mathbf{C}))$. By taking $T = \infty$, we have

$$\lim_{\alpha\to 0} \alpha^{-1}\text{Cov}(\boldsymbol{\theta}_{\infty}^{'}) = \alpha \sum_{t\geqslant 0}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t\mathbf{C}^{'}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t.$$

Appendix B. Detailed Experiments

Experimental Settings

Our experiments on nerual 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 × 200 FC)-ReLU-(200 × 200 FC)-ReLU-(200 × 10 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 × 6 5 × 5C)-ReLU-MP2-(6 × 16 5 × 5C)-ReLU-MP2-(400 × 120 FC)-ReLU-(120 × 84 FC)-ReLU-(84 × 10 FC), where (5 × 5C) 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 $\{\boldsymbol{\theta}^{(i)}_{\text{MNIST}}\}_{i=1}^{3000}, \{\boldsymbol{\theta}^{(i)}_{\text{CIFAR10}}\}_{i=1}^{3000}$.
4. Take MNIST for example, for each marginal $j = 1, ..., p_{\text{MNIST}}$ with $p_{\text{MNIST}} = 198800$, we perform the Person test on $\{\boldsymbol{\theta}^{(i)}_{\text{MNIST}}[j]\}_{i=1}^{3000}$ 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 δ. The same procedures are conducted on CIFAR-10.
5. Take MNIST for example, we calculate the precentage of the marginals with p-values smaller than a given threshold, which takes values in $\{0.001, 0.002, ..., 0.999, 1\}$. 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 $\{\boldsymbol{\theta}_0[1], \boldsymbol{\theta}_0[2]\} \overset{\text{i.i.d}}{\sim} \text{U}(0,1)$ 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

[1] LeCun, Y., Cortes, C.: MNIST handwritten digit database (2010)

[2] Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto (2009)

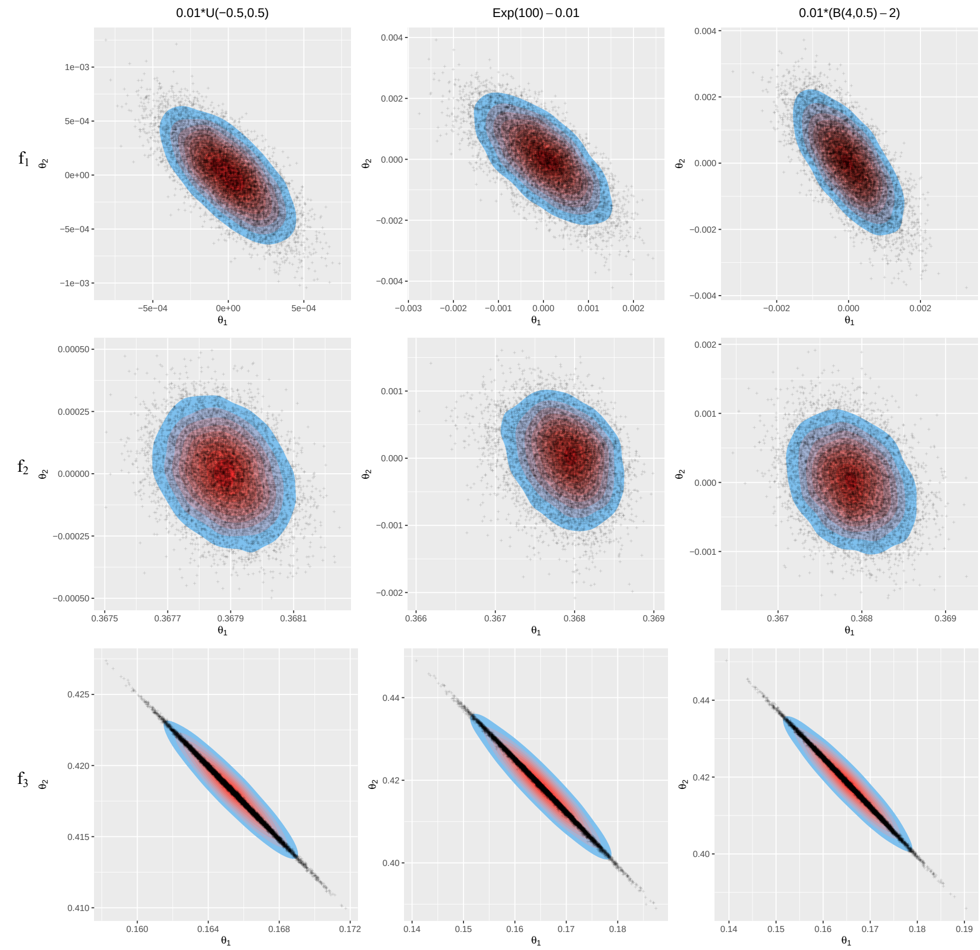


Fig. 1 For each gradient noise implements (including adding uniform, exponential and binomial gradient noise) and each loss functions f_1, f_2 and f_3, experiments are ran with $\alpha = 0.01$ and $\boldsymbol{\theta}_0 = (1, 1)^\top$. We visualize the empirical limiting distribution by a 2D-kernel density plot.

[3] He, K., Zhang, X., Ren, S., Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification (2015)

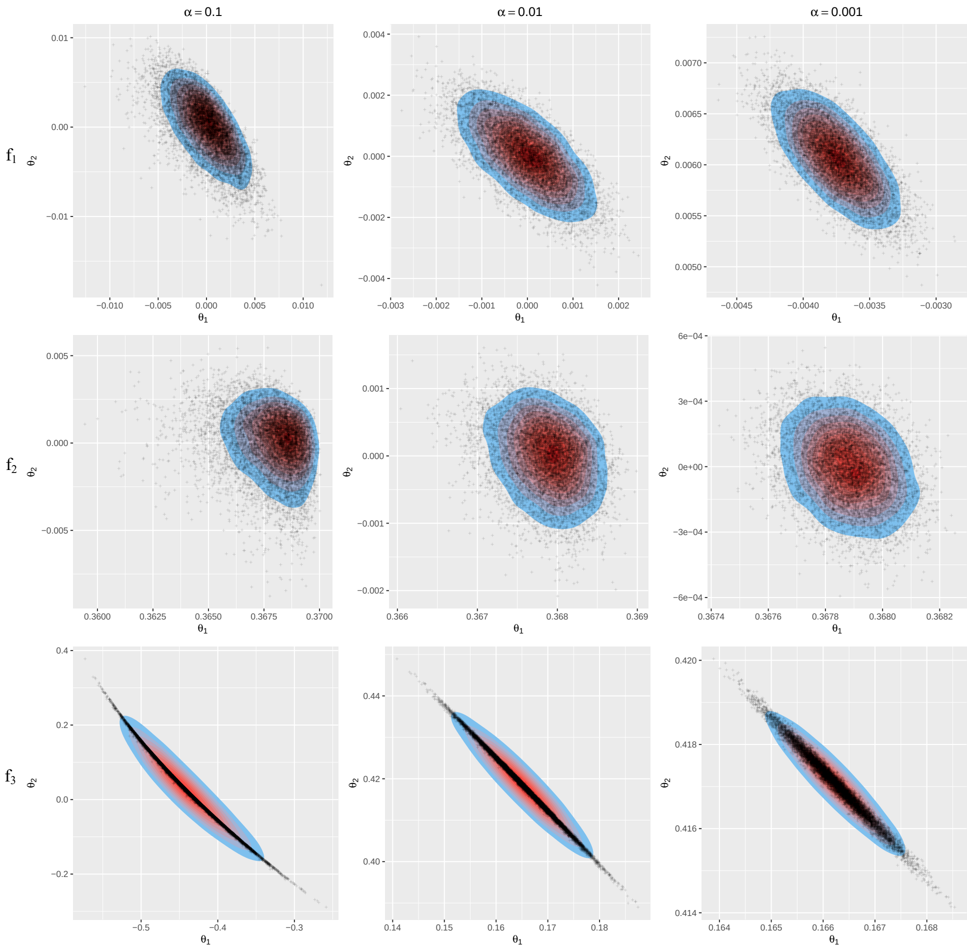


Fig. 2 The gradient noise is fixed to be exponential. For each loss functions f_1, f_2, f_2, the experiments are ran with $\alpha \in \{0.1, 0.01, 0.001\}$ and a fixed initialization $\boldsymbol{\theta}_0 = (1, 1)^\top$. We visualize the empirical limiting distribution by a 2D-kernel density plot.

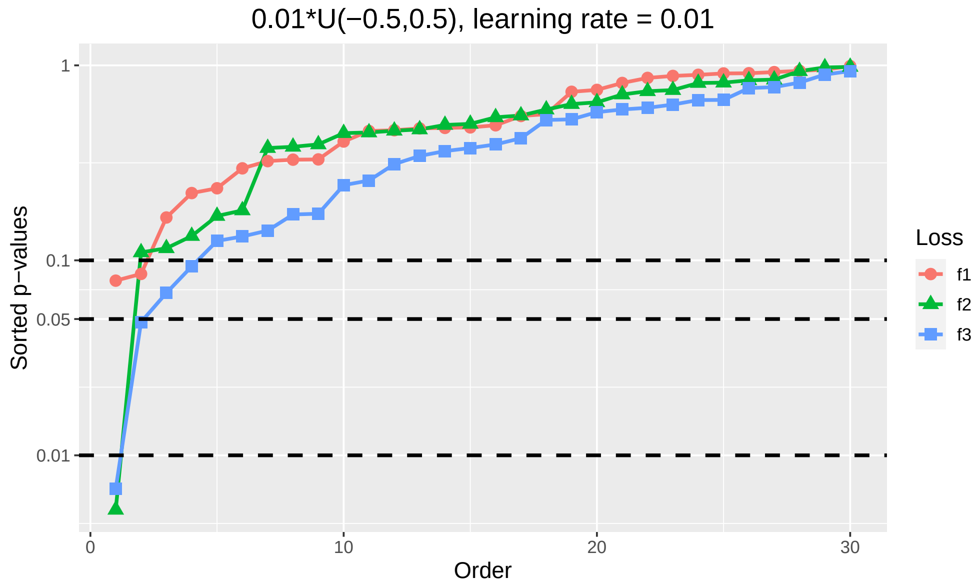


Fig. 3 For loss functions f_1, f_2, f_3, we perform SGDs with uniformly distributed gradient noise and $\alpha = 0.01$. At the confidential level of 0.99, about 29/30 of the 30 repetitions fail to provide statistically significant evidence against the two-dimensional Gaussianity of the limiting parameter distributions.

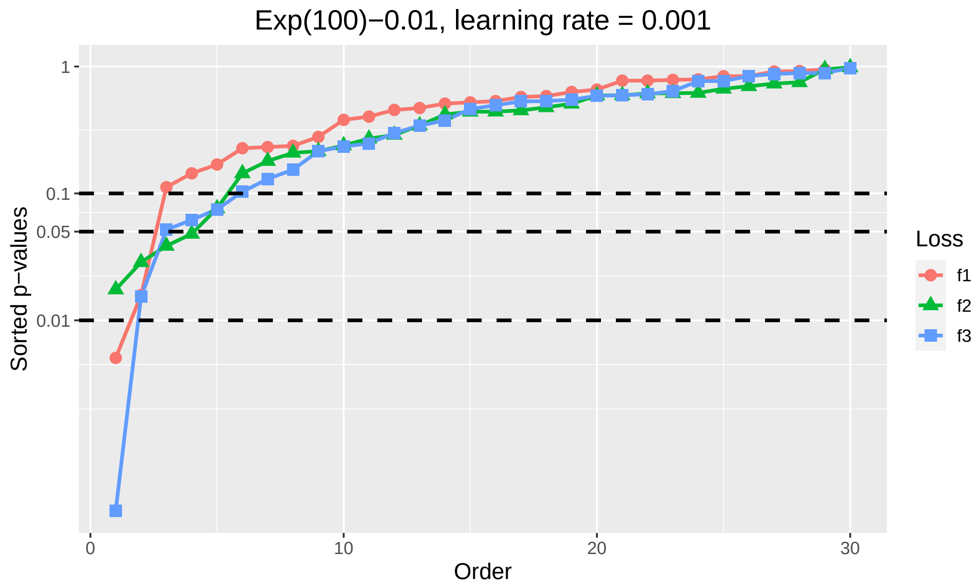


Fig. 4 For loss functions f_1, f_2, f_3, we perform SGDs with exponentially distributed gradient noise and $\alpha = 0.001$. At the confidential level of 0.99, about 29/30 of the 30 repetitions fail to provide statistically significant evidence against the two-dimensional Gaussianity of the limiting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

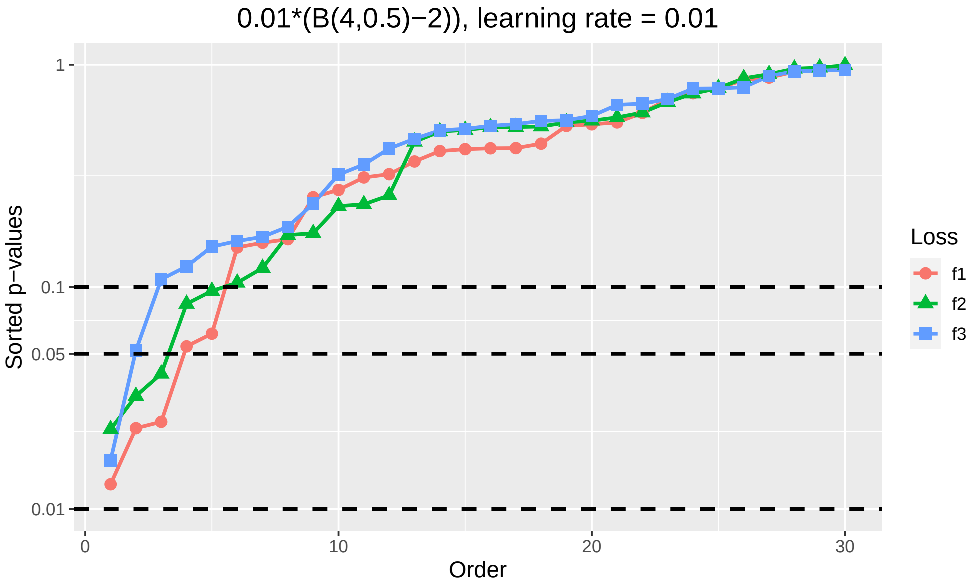
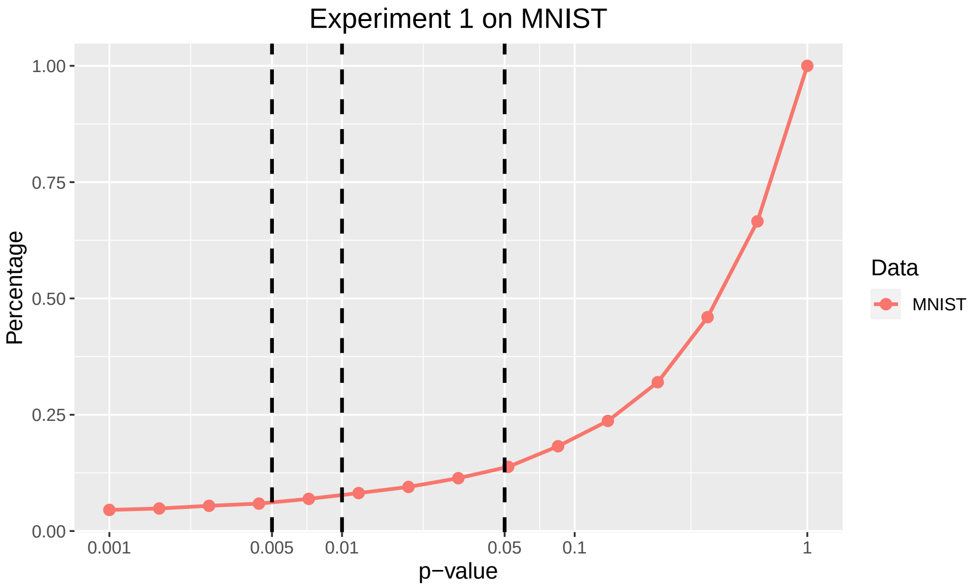
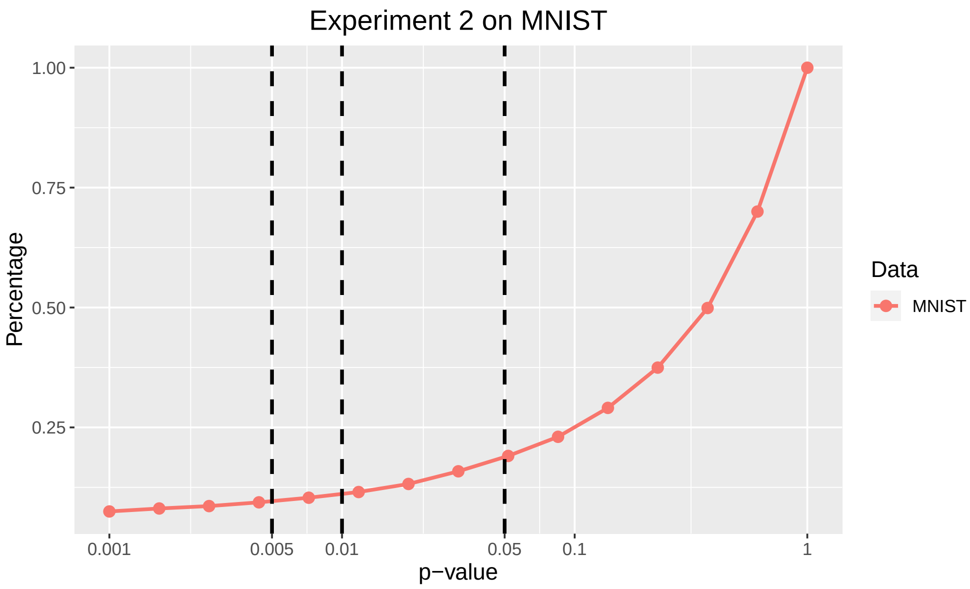


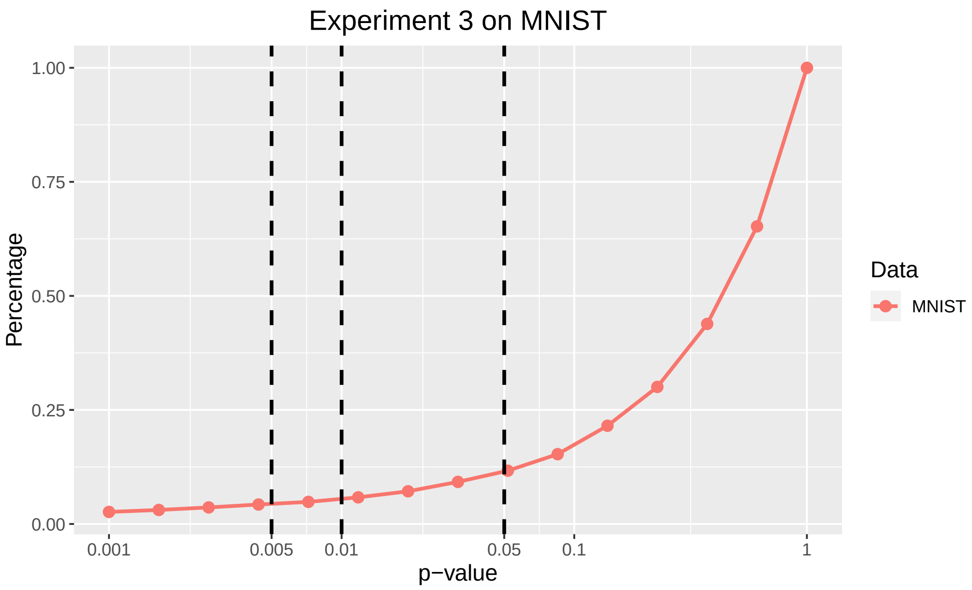
Fig. 5 For loss functions f_1, f_2, f_3, we perform SGDs with binomially distributed gradient noise and $\alpha = 0.01$. At the confidential level of 0.99, all of the 30 repetitions fail to provide statistically significant evidence against the two-dimensional Gaussianity of the limiting parameter distributions.

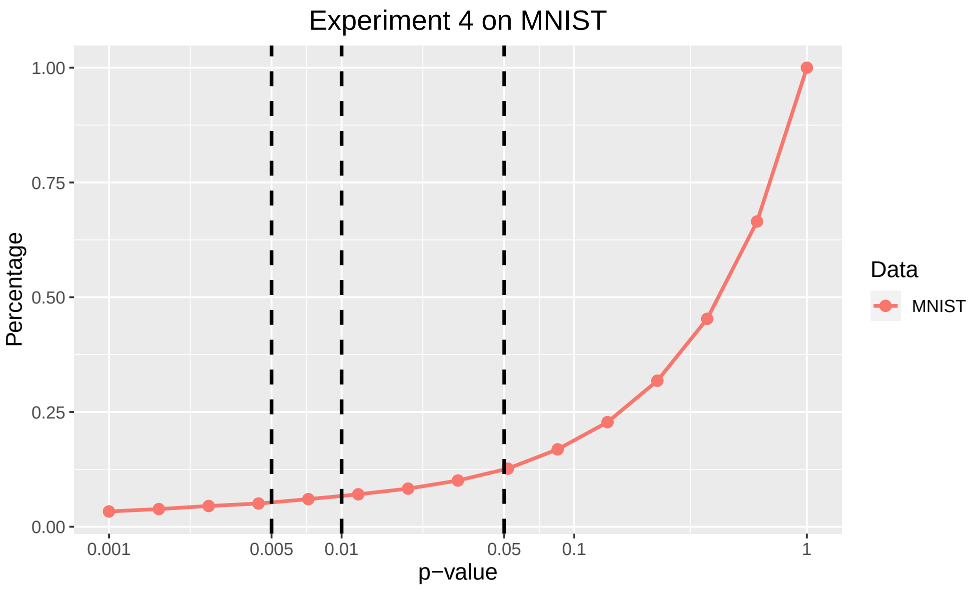
Table 1 For each experiment, we calculate the percentages of dimensions with marginal p-values smaller than $0.1, 0.05$ and 0.01, respectively. For a marginal with a p-value smaller than $\delta \in (0, 1)$, we can reject the null hypothesis that this marginal follows a Gaussian distribution at a confidential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

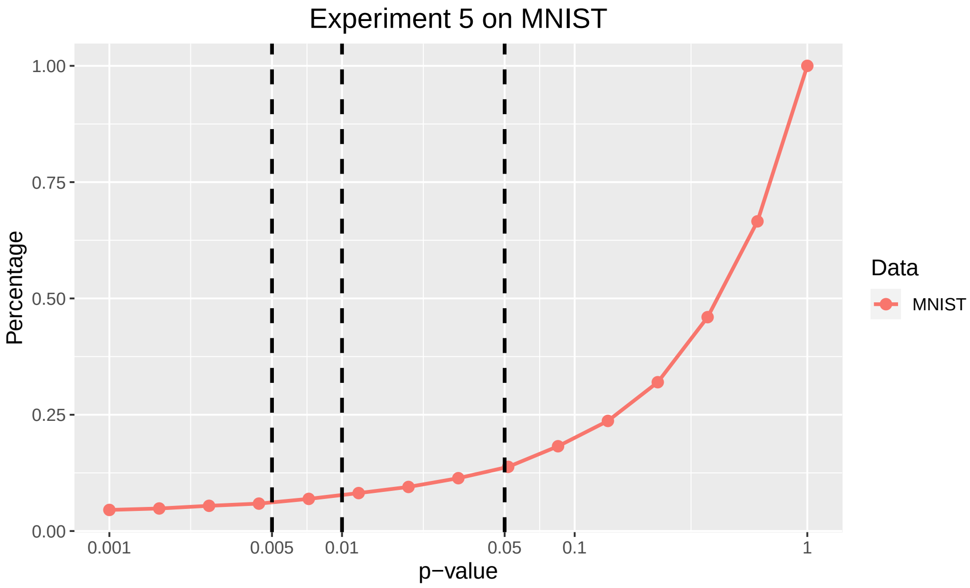
Percentage	$\leqslant 0.1$	$\leqslant 0.05$	$\leqslant 0.01$
MNIST Exp. 1	10.8%	5.7%	1.5%
MNIST Exp. 2	11.8%	6.9%	2.6%
MNIST Exp. 3	12.3%	7.2%	2.7%
MNIST Exp. 4	12.6%	7.5%	3.2%
MNIST Exp. 5	12.2%	7.0%	2.6%
CIFAR-10 Exp. 1	10.8%	5.7%	1.5%
CIFAR-10 Exp. 2	11.8%	6.9%	2.6%
CIFAR-10 Exp. 3	12.4%	7.2%	2.7%
CIFAR-10 Exp. 4	12.6%	7.5%	3.2%
CIFAR-10 Exp. 5	12.2%	7.0%	2.6%

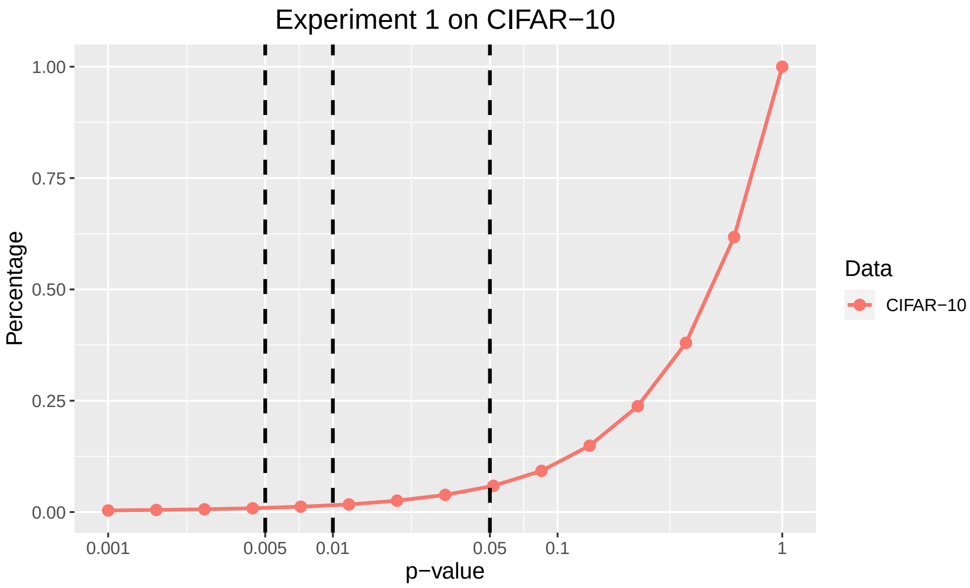

Experiment 1 on MNIST
Percentage
1.00
0.75
0.50
0.25
0.00
0.001
0.005
0.01
0.05
0.1
1
p−value
Data
MNIST

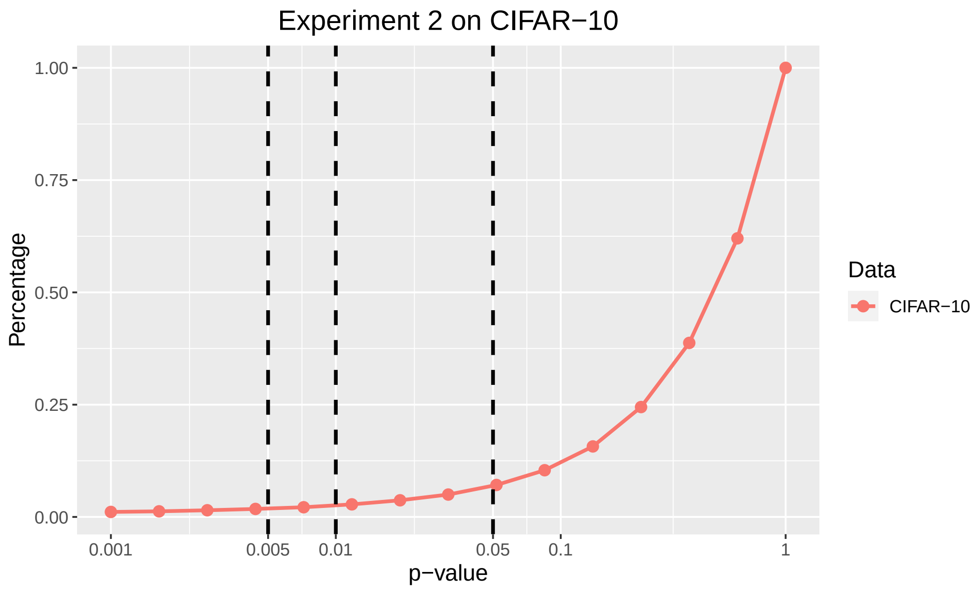

Experiment 2 on MNIST
Percentage
1.00
0.75
0.50
0.25
0.001
0.005
0.01
0.05
0.1
1
p−value
Data
MNIST

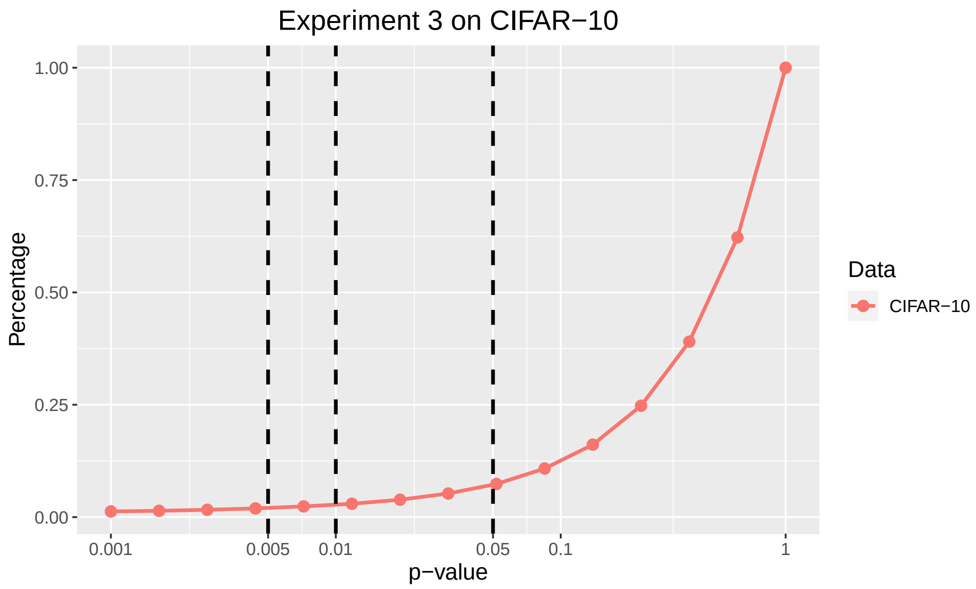

Experiment 3 on MNIST
Percentage
1.00
0.75
0.50
0.25
0.00
0.001
0.005
0.01
0.05
0.1
1
p−value
Data
MNIST

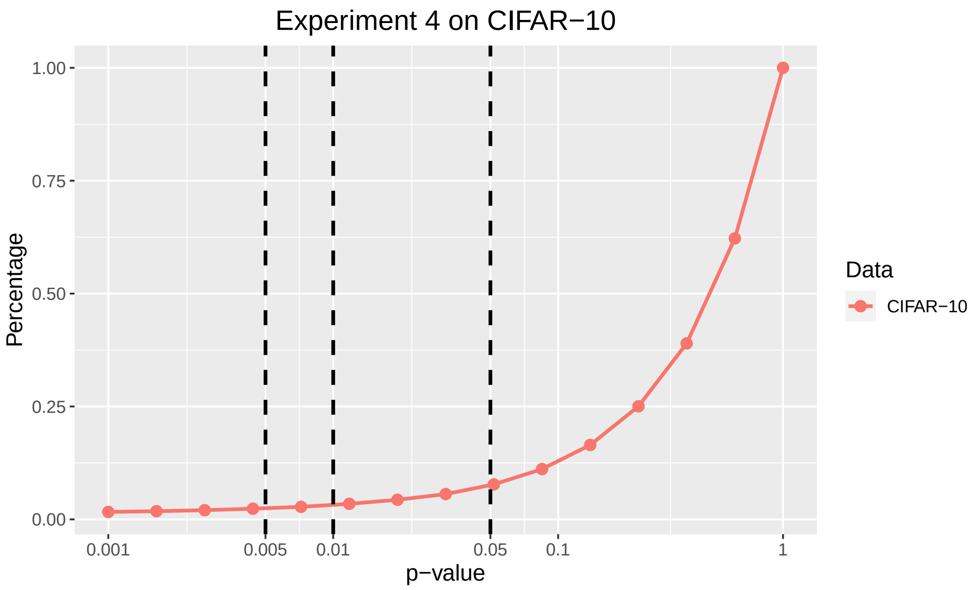

Experiment 4 on MNIST
Percentage
1.00
0.75
0.50
0.25
0.00
0.001
0.005
0.01
0.05
0.1
1
p−value
Data
MNIST

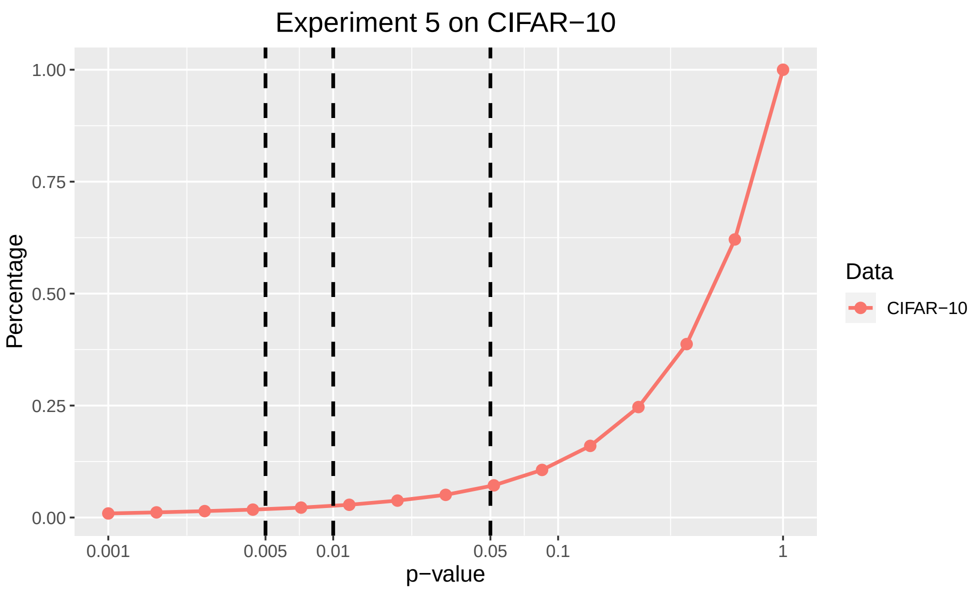

Experiment 5 on MNIST
1.00
0.75
0.50
0.25
0.00
Percentage
0.001
0.005
0.01
0.05
0.1
1
p−value
Data
MNIST


Experiment 1 on CIFAR−10
1.00
0.75
0.50
0.25
0.00
Percentage
0.001
0.005
0.01
0.05
0.1
1
p−value
Data
CIFAR−10


Experiment 2 on CIFAR−10
Percentage
1.00
0.75
0.50
0.25
0.00
0.001
0.005
0.01
0.05
0.1
1
p−value
Data
CIFAR−10


Experiment 3 on CIFAR−10
Percentage
1.00
0.75
0.50
0.25
0.00
0.001
0.005
0.01
0.05
0.1
1
p−value
Data
CIFAR−10


Experiment 4 on CIFAR−10
Percentage
1.00
0.75
0.50
0.25
0.00
0.001
0.005
0.01
0.05
0.1
1
p−value
Data
CIFAR−10


Experiment 5 on CIFAR−10
Percentage
1.00
0.75
0.50
0.25
0.00
0.001
0.005
0.01
0.05
0.1
1
p−value
Data
CIFAR−10