Appendices

A Experiment Details

A.1 Architectures and Hyperparameters

In Section 4, MLP has one hidden layer with 512 hidden units, and AlexNet has five convolution layers (conv. 3×3 (64 filters) $\rightarrow$ max-pool $3 \times 3 \rightarrow$ conv. 5×5 (192 filters) $\rightarrow$ max-pool $3 \times 3 \rightarrow$ conv. 3×3 (384 filters) $\rightarrow$ conv. 3×3 (256 filters) $\rightarrow$ conv. 3×3 (256 filters) $\rightarrow$ max-pool 3×3) followed by two fully connected layers both with 4096 units and a 10-way linear layer as the output layer. All of the convolution layers and the fully connected layers use standard rectified linear activation functions (ReLU).

The fixed learning rates used for MLP and AlexNet are 0.01 and 0.001, respectively. The batch size is set to 60. For the corrupted label experiment, we train the models until the models achieve 100% training accuracy. For other cases, we train the neural networks until the training loss converges (e.g., < 0.0001). Other settings are either described in Section 4 or apparent in the figures. Standard techniques such as weight decay and batch normalization are not used.

In Section 5, we compare GMP with other advanced regularization methods. The results of other methods are reported directly from [83], and we now give their hyperparameter settings here for completeness. For Dropout, 10% of neurons are randomly selected to be deactivated in each layer. For label smoothing, the coefficient is 0.2. For flooding, the level is set to 0.02. For MixUp, we lineally combine random pairs of training data where the coefficient is drawn from $\mathrm{Beta}(1, 1)$. For adversarial training, the perturbation size is 1 for each pixel and we take one step to generate adversarial examples. For AMP, the number of inner iteration is 1, and the L_2 norm ball radius values are 0.5 for PreActResNet18 and 0.1 for VGG16, respectively.

A.2 Algorithm of Dynamic Gradient Clipping

The dynamic gradient clipping algorithm is described in Algorithm 1. For both MLP and AlexNet, we let $\alpha = 0.1$. The start step for clipping, T_c, is also an important hyperparameter. However, it can be removed by detecting the evolution of the average gradient norm for each epoch. Specifically, whenever the average gradient norm of epoch j is larger than the average gradient norm of epoch $j - 1$, the clipping operation begins.

Algorithm 1 Dynamic Gradient Clipping

Require: Training set S, Batch size b, Loss function ℓ, Initial model parameter $\boldsymbol{w}_0$, Learning rate λ, Initial minimum gradient norm $\mathcal{G}$, Number of iterations T, Clipping parameter α, Clipping step T_c

for $t \leftarrow 1$ to T **do**
  Sample $\mathcal{B} = \{\boldsymbol{z}_i\}_{i=1}^{b}$ from training set S
  Compute gradient:
  $g_{\mathcal{B}} \leftarrow \sum_{i=1}^{b} \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}_{t-1}, \boldsymbol{z}_i)/b$
  if $t > T_c$ **then**
    if $||g_{\mathcal{B}}||_2 > \mathcal{G}$ **then**
      $g_{\mathcal{B}} \leftarrow \alpha \cdot \mathcal{G} \cdot g_{\mathcal{B}}/||g_{\mathcal{B}}||_2$
    else
      $\mathcal{G} \leftarrow ||g_{\mathcal{B}}||_2$
    end if
  end if
  Update parameter: $\boldsymbol{w}_t \leftarrow \boldsymbol{w}_{t-1} - \lambda \cdot g_{\mathcal{B}}$
end for

From Figure 4 we can see that dynamic gradient clipping effectively alleviates overfitting by conspicuously slowing down the transition of training to the memorization regime, without changing the convergence speed of testing accuracy. Unfortunately, the current design of the dynamic gradient clipping algorithm does not provide a significant improvement for models trained on a true dataset.

PreActResNet18	Top-1 Acc. (%)
GMP^3	97.43±0.037
GMP^5	97.40±0.043
GMP^{10}	97.34±0.058
VGG16	Top-1 Acc. (%)
GMP^3	97.18±0.057
GMP^5	97.17±0.072
GMP^{10}	97.09±0.068

(a) SVHN

PreActResNet18	Top-1 Acc. (%)
GMP^3	95.64±0.053
GMP^5	95.63±0.073
GMP^{10}	95.71±0.073
VGG16	Top-1 Acc. (%)
GMP^3	94.33±0.094
GMP^5	94.49±0.118
GMP^{10}	94.45±0.158

(b) CIFAR-10

PreActResNet18	Top-1 Acc. (%)
GMP^3	78.05±0.208
GMP^5	77.93±0.188
GMP^{10}	78.07±0.170
VGG16	Top-1 Acc. (%)
GMP^3	74.45±0.256
GMP^5	74.91±0.389
GMP^{10}	75.09±0.285

(c) CIFAR-100

Table 2: Top-1 classification accuracy on (a) SVHN, (b) CIFAR-10 and (c) CIFAR-100. Superscript denotes the number of sampled Gaussian noises during training.

Designing better regularization algorithms may require understanding the epoch-wise double descent curve of gradient dispersion where the model is trained on a true dataset.

A.3 Discussion on Gradient Dispersion of Models trained on True Dataset

In the case of no noise injected, Figure 3a shows that the model with good generalization property has a exponentially-decaying gradient dispersion. This is consistent with our discussion of Lemma 5 in Section 3, that is, small $I(G_t + N_t; Z_i|\widetilde{W}_{t-1})$ indicates good generalization. Notably, gradient dispersion of AlexNet trained on the ture CIFAR10 data still has a epoch-wise double descent curve. The difference between Figure 3e with Figure 3f-3h is that the testing accuracy does not decrease in the second phase/memorization regime for AlexNet trained on the true CIFAR10 data. Loosely speaking, we conjugate that memorizing random labels will hurt the performance on unseen true data but memorizing true labels will not. This explains why dynamic gradient clipping or preventing the training entering the memorization regime cannot improve the performance on a true dataset.

A.4 Algorithm of Gaussian Model Perturbation

Algorithm 2 Gaussian Model Perturbation Training

Require: Training set S, Batch size b, Loss function ℓ, Initial model parameter $\boldsymbol{w}_0$, Learning rate λ, Number of noise k, Standard deviation of Gaussian distribution σ, Lagrange multiplier ρ

while $\boldsymbol{w}_t$ not converged **do**

Update iteration: $t \leftarrow t + 1$

Sample $\mathcal{B} = \{\boldsymbol{z}_i\}_{i=1}^b$ from training set S

Sample $\Delta_j \sim \mathcal{N}(0, \sigma^2)$ for $j \in [k]$

Compute gradient:

$$g_{\mathcal{B}} \leftarrow \sum_{i=1}^{b} \left(\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}_t, z_i) + \rho \sum_{j=1}^{k} \left(\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}_t + \Delta_j, z_i) - \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}_t, z_i) \right) / k \right) / b$$

Update parameter: $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \lambda \cdot g_{\mathcal{B}}$

end while

The GMP algorithm is given in Algorithm 2. Table 1 shows that our method is competitive to the state-of-the-art regularization techniques. Specifically, our method achieves the best performance on SVHN for both models and on CIFAR-100 where VGG16 is employed. Particularly, testing accuracy is improved by nealy 2% on the CIFAR-100 dataset with VGG16. For other tasks, GMP is always able to achieve the top-3 performance. From Table 2, we find that increasing the number of sampled noises does not guarantee the improvement of testing accuracy and may even degrade the performance on some datasets (e.g., SVHN). This hints that we can use small number of noises to reduce the running time without losing performance. Moreover, we observe that **GMP with** $k = 3$ **usually takes around** $1.76\times$ **that of ERM training time**, which is affordable.

One potential extension of GMP is letting the variance of the noise distribution be a function of the iteration step t. In other words, using the time-dependent σ_t instead of a constant σ.

A.5 License of the Assets

MNIST is made available under the terms of the Creative Commons Attribution-Share Alike 3.0 license. CIFAR10/CIFAR100 is licensed under the MIT License. SVHN is licensed under the GNU General Public License v3.0. Two open source packages used in this paper, BackPACK and PyHessian, are licensed under the MIT License.

B Proofs for Section 3

B.1 Proof of Lemma 4

Proof. The proof given here is a simple extension of the proof of [60, Lemma 3.4.2], which is a special instance of the (weak) HWI inequality.

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(P_{X+\sqrt{t}N}||P_{Y+\sqrt{t}N}) \leq & \mathrm{D}_{\mathrm{KL}}(P_{X,Y,X+\sqrt{t}N}||P_{X,Y,Y+\sqrt{t}N}) && (1)\\
= & \mathop{\mathbb{E}}_{X,Y}\left[\mathrm{D}_{\mathrm{KL}}(P_{X+\sqrt{t}N|X,Y}||P_{Y+\sqrt{t}N|X,Y})\right] && (2)\\
= & \mathop{\mathbb{E}}_{X,Y}\left[\mathrm{D}_{\mathrm{KL}}(\mathcal{N}(X,t)||\mathcal{N}(Y,t))|X,Y\right] && (3)\\
= & \frac{1}{2t}\mathop{\mathbb{E}}_{X,Y}\left[||X-Y||^2\right], && (4)
\end{aligned}$$

where Eq.1 is by the chain rule of the KL divergence, Eq.3 holds since N is independent of (X,Y) and Eq.4 is a special case of the equality

$$\mathcal{N}(\mu_1,\Sigma_1)||\mathcal{N}(\mu_2,\Sigma_2)) = \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - d + \mathrm{Tr}\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)\right].$$

Eq.4 holds for any joint distribution of (X,Y), so by the definition of Wasserstein distance, we have

$$\mathrm{D}_{\mathrm{KL}}(P_{X+\sqrt{t}N}||P_{Y+\sqrt{t}N}) \leq \frac{1}{2t}\mathbb{W}_2^2(P_X,P_Y). \tag{5}$$

This completes the proof. □

B.2 Proof of Theorem 1

Proof. The main parts of the proof follow directly from [49]. We first bound the mutual information $I(\widetilde{W}_T;S)$,

$$\begin{aligned}
I(\widetilde{W}_T;S) = & \int_s\int_w dP_{\widetilde{W}_T|S=s}(w)\log\frac{dP_{\widetilde{W}_T|S=s}(w)}{dP_{\widetilde{W}_T}(w)}d\nu(s) && (6)\\
= & \int_s\int_w dP_{\widetilde{W}_T|S=s}(w)\log\frac{dP_{\widetilde{W}_T|S=s}(w)}{\int_s' dP_{\widetilde{W}_T|S'=s'}(w)d\nu(s')}d\nu(s) && (7)\\
\leq & \int_{s,s'}\int_w dP_{\widetilde{W}_T|S=s}(w)\log\frac{dP_{\widetilde{W}_T|S=s}(w)}{dP_{\widetilde{W}_T|S'=s'}(w)}d\nu(s)d\nu(s') && (8)\\
= & \mathop{\mathbb{E}}_{S,S'}\left[\mathrm{D}_{\mathrm{KL}}\left(P_{\widetilde{W}_T|S=s}||P_{\widetilde{W}_T|S'=s'}\right)\middle|S=s,S'=s'\right] && (9)\\
\leq & \mathop{\mathbb{E}}_{S,S'}\left[\mathrm{D}_{\mathrm{KL}}\left(P_{\widetilde{W}_1,\widetilde{W}_2,\ldots,\widetilde{W}_T|S=s}||P_{\widetilde{W}_1,\widetilde{W}_2,\ldots,\widetilde{W}_T|S'=s'}\right)\middle|S=s,S'=s'\right] && (10)\\
= & \sum_{t=1}^{T}\mathop{\mathbb{E}}_{S,S',\widetilde{W}_{t-1}}\left[\mathrm{D}_{\mathrm{KL}}\left(P_{\widetilde{W}_t|\widetilde{W}_{t-1}=\widetilde{w}_{t-1},S=s}||P_{\widetilde{W}_t|\widetilde{W}_{t-1}=\widetilde{w}_{t-1},S'=s'}\right)\middle|\widetilde{W}_{t-1}=\widetilde{w}_{t-1},S=s,S'=s'\right], && (11)
\end{aligned}$$

where Eq.8 is by Jensen's inequality, Eq.10 is by the data-processing inequality of the KL divergence and Eq.11 is by the chain rule of the KL divergence. The main difference between the above process and [49] is that we don't need to use an independent copy $\widetilde{W}_T'$ of $\widetilde{W}_T$ in Eq.6.

Now apply Lemma 4, we have

$$\begin{aligned}
&\mathrm{D_{KL}}\left(P_{\widetilde{W}_t|\widetilde{W}_{t-1}=\widetilde{w}_{t-1},S=s}||P_{\widetilde{W}_t|\widetilde{W}_{t-1}=\widetilde{w}_{t-1},S'=s'}\right)\\
=&\mathrm{D_{KL}}\left(P_{g(\widetilde{w}_{t-1}-\Delta_{t-1},B_t)-\frac{N_t}{\lambda_t}|\widetilde{W}_{t-1}=\widetilde{w}_{t-1},S=s}||P_{g(\widetilde{w}_{t-1}-\Delta_{t-1},B'_t)-\frac{N_t}{\lambda_t}|\widetilde{W}_{t-1}=\widetilde{w}_{t-1},S'=s'}\right) \quad (12)\\
\leq&\frac{\lambda_t^2}{2\sigma_t^2}\mathbb{E}\left[||g(\widetilde{w}_{t-1}-\Delta_{t-1},B_t)-g(\widetilde{w}_{t-1}-\Delta_{t-1},B'_t)||_2^2\right]. \quad (13)
\end{aligned}$$

Since each instance is independently sampled during training, we have

$$\begin{aligned}
\mathbb{E}\left[||g(w_{t-1},B_t)-g(w_{t-1},B'_t)||_2^2\right] =&2\mathbb{E}\left[||g(w_{t-1},B_t)-\mathbb{E}\left[g(w_{t-1},B_t)\right]||_2^2\right] \quad (14)\\
=&\frac{2}{b}\mathbb{E}\left[||\nabla_w\ell(w_{t-1},Z)-\mathbb{E}\left[\nabla_w\ell(w_{t-1},Z)\right]||_2^2\right]. \quad (15)
\end{aligned}$$

Plugging Eq.13 and Eq.15 into Eq.11,

$$\begin{aligned}
I(\widetilde{W}_T;S)\leq&\sum_{t=1}^{T}\mathop{\mathbb{E}}_{S,S',W_{t-1}}\left[\frac{\lambda_t^2}{2\sigma_t^2}||g(W_{t-1},B_t)-g(W_{t-1},B'_t)||_2^2\right] \quad (16)\\
=&\sum_{t=1}^{T}\mathop{\mathbb{E}}_{Z,W_{t-1}}\left[\frac{\lambda_t^2}{b\sigma_t^2}||\nabla_w\ell(W_{t-1},Z)-\mathop{\mathbb{E}}_{Z}\left[\nabla_w\ell(W_{t-1},Z)\right]||_2^2\right] \quad (17)\\
=&\sum_{t=1}^{T}\frac{\lambda_t^2}{b\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}\left[\mathbb{V}(W_{t-1})\right]. \quad (18)
\end{aligned}$$

Thus,

$$|\mathrm{gen}(\mu,P_{W_T|S})|=\left|\mathrm{gen}(\mu,P_{\widetilde{W}_T|S})+\mathop{\mathbb{E}}_{W_T,\Delta_T}\left[L_\mu(W_T)-L_\mu(\widetilde{W}_T)\right]+\mathop{\mathbb{E}}_{W_T,\Delta_T,S}\left[L_S(\widetilde{W}_T)-L_S(W_T)\right]\right| \quad (19)$$

$$\leq\sqrt{\frac{2R^2I(\widetilde{W}_T;S)}{n}}+\left|\mathop{\mathbb{E}}_{W_T,S,S'}\left[\gamma(W_T,S)-\gamma(W_T,S')\right]\right| \quad (20)$$

$$\leq\sqrt{\frac{2R^2}{nb}\sum_{t=1}^{T}\frac{\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}\left[\mathbb{V}(W_{t-1})\right]}+\left|\mathop{\mathbb{E}}_{W_T,S,S'}\left[\gamma(W_T,S)-\gamma(W_T,S')\right]\right|, \quad (21)$$

where Eq.20 is by Lemma 1 and the triangle inequality. This completes the proof. □

B.3 Proof of Lemma 5

Proof. Let Z_i be used in t^{th} step, the LHS of the inequality can be rewritten as

$$\begin{aligned}
&I\left(-g(W_{t-1},B_t)+\frac{1}{\lambda_t}N_t;Z_i|\widetilde{W}_{t-1}\right)\\
=&\mathop{\mathbb{E}}_{\widetilde{W}_{t-1},Z_i}\left[\mathrm{D_{KL}}\left(P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|Z_i=z,\widetilde{W}_{t-1}=\widetilde{w}_{t-1}}||P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|\widetilde{W}_{t-1}=\widetilde{w}_{t-1}}\right)\right]. \quad (22)
\end{aligned}$$

By Lemma 4, we have

$$\begin{aligned}
&\mathrm{D_{KL}}\left(P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|Z_i=z,\widetilde{W}_{t-1}=\widetilde{w}_{t-1}}||P_{-g(W_{t-1},B_t)+\frac{N_t}{\lambda_t}|\widetilde{W}_{t-1}=\widetilde{w}_{t-1}}\right)\\
\leq&\frac{\lambda_t^2}{2\sigma_t^2}\mathbb{W}_2^2\left(P_{-g(W_{t-1},B_t)|Z_i=z,\widetilde{W}_{t-1}=\widetilde{w}_{t-1}},P_{-g(W_{t-1},B_t)|\widetilde{W}_{t-1}=\widetilde{w}_{t-1}}\right). \quad (23)
\end{aligned}$$

Let Z_i' be an independent copy of Z_i such that $(W_{t-1}, Z_i', Z_i) \sim P_{W_{t-1}} \otimes P_{Z_i|W_{t-1}} \otimes P_{Z_i'|W_{t-1}}$. We now consider a special coupling: define two random vectors $\widetilde{G}_t^z$ and $\widetilde{G}_t$ as follows,

$$\widetilde{G}_t^z \triangleq -\frac{1}{b}\left(\sum_{Z_j \in B_t \setminus \{Z_i\}} \nabla_w \ell(\widetilde{w}_{t-1} - \Delta_{t-1}, Z_j) + \nabla_w \ell(\widetilde{w}_{t-1} - \Delta_{t-1}, z)\right), \tag{24}$$

$$\widetilde{G}_t \triangleq -\frac{1}{b}\left(\sum_{Z_j \in B_t} \nabla_w \ell(\widetilde{w}_{t-1} - \Delta_{t-1}, Z_j))\right). \tag{25}$$

Thus, $\widetilde{G}_t^z$ and $\widetilde{G}_t$ have marginals $P_{-g(W_{t-1},B_t)|Z_i=z,\widetilde{W}_{t-1}=\widetilde{w}_{t-1}}$ and $P_{-g(W_{t-1},B_t)|\widetilde{W}_{t-1}=\widetilde{w}_{t-1}}$, respectively. Combining Eq.22, Eq.23 and the definition of Wasserstein distance, we have

$$I\left(-g(W_{t-1}, B_t) + \frac{N_t}{\lambda_t}; Z_i|\widetilde{W}_{t-1}\right) \leq \frac{\lambda_t^2}{2\sigma_t^2} \mathop{\mathbb{E}}_{\widetilde{W}_{t-1}, Z_i, \Delta_{t-1}, Z_i'} \left[||\widetilde{G}_t^z - \widetilde{G}_t||_2^2\right] \tag{26}$$

$$= \frac{\lambda_t^2}{2\sigma_t^2} \mathop{\mathbb{E}}_{W_{t-1}, Z_i, Z_i'} \left[\frac{1}{b^2}||\nabla_w \ell(W_{t-1}, Z_i) - \nabla_w \ell(W_{t-1}, Z_i')||_2^2\right] \tag{27}$$

$$= \frac{\lambda_t^2}{\sigma_t^2 b^2} \mathop{\mathbb{E}}_{W_{t-1}} [\mathbb{V}(W_{t-1})], \tag{28}$$

where Eq.28 is by

$$\mathop{\mathbb{E}}_{Z_i, Z_i'} \left[||\nabla_w \ell(w_{t-1}, Z_i) - \nabla_w \ell(w_{t-1}, Z_i')||_2^2\right] = 2\mathbb{V}(w_{t-1}).$$

This completes the proof. □

B.4 Proof of Theorem 2

Since $\eta_{i,\tau} \leq 1$ for every i and τ in Theorem 3, Theorem 2 follows directly from Theorem 3.

Or alternatively, using the following proof.

Proof. Given a specific sample path, let the step t_K be the last iteration that Z_i is used (e.g., see Figure 5). By using the data processing inequality, we can easily have

$$I(\widetilde{W}_T; Z_i) \leq I(\widetilde{W}_{t_K}; Z_i). \tag{29}$$

Then notice that

$$I(\widetilde{W}_{t_K}; Z_i) = I\left(\widetilde{W}_{t_K - 1} + G_{t_K} + N_{t_K}; Z_i\right) \tag{30}$$

$$\leq I\left(\widetilde{W}_{t_K - 1}, G_{t_K} + N_{t_K}; Z_i\right) \tag{31}$$

$$= I(\widetilde{W}_{t_K - 1}; Z_i) + I\left(-g(W_{t_K - 1}, B_{t_K}) + \frac{1}{\lambda_{t_K}} N_{t_K}; Z_i|\widetilde{W}_{t_K - 1}\right), \tag{32}$$

where $G_{t_K} = -\lambda_{t_K} g(W_{t_K - 1}, B_{t_K})$, Eq.31 is by $I(f(X); Y) \leq I(X; Y)$ and Eq.32 is by the chain rule of mutual information. The second term in Eq.32 can be upper bounded by using Lemma 5. The first term can be upper bounded by following the similar procedure recursively, namely, Eq.29-32. Let t_1 be the first time that Z_i is used in training. Given the fact that Z_i is independent of $\widetilde{W}_t$ when $t < t_1$, it's easy to see

$$I(\widetilde{W}_T; Z_i) \leq \sum_{t \in \mathcal{T}_i} \frac{\lambda_t^2}{\sigma_t^2 b^2} \mathop{\mathbb{E}}_{W_{t-1}} [\mathbb{V}(W_{t-1})], \tag{33}$$

where $\mathcal{T}_i$ is the set of all indices of iterations that contain Z_i.

Thus, the following bound holds,

$$|\mathrm{gen}(\mu, P_{W_T|S})| = \left|\mathrm{gen}(\mu, P_{\widetilde{W}_T|S}) + \mathop{\mathbb{E}}_{W_T,\Delta_T}\left[L_\mu(W_T) - L_\mu(\widetilde{W}_T)\right] + \mathop{\mathbb{E}}_{W_T,\Delta_T,S}\left[L_S(\widetilde{W}_T) - L_S(W_T)\right]\right|$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2R^2 I(W;Z_i)} + \left|\mathop{\mathbb{E}}_{W_T,S,S'}\left[\gamma(W_T,S) - \gamma(W_T,S')\right]\right| \tag{34}$$

$$\leq \frac{R}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\frac{2\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}\left[\mathbb{V}(W_{t-1})\right]} + \left|\mathop{\mathbb{E}}_{W_T,S,S'}\left[\gamma(W_T,S) - \gamma(W_T,S')\right]\right|, \tag{35}$$

where Eq.34 is by Lemma 2 and the triangle inequality. This concludes the proof. □

B.5 Proof of Corollary 1

Proof. To handle the mismatch between the outputs of perturbed SGD and SGD, we apply Taylor expansion around $\Delta_T = \vec{0}$,

$$\mathop{\mathbb{E}}_{W_T,S,\Delta_T}\left[L_S(W_T+\Delta_T) - L_S(W_T)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{W_T,Z_i,\Delta_T}\left[\ell(W_T+\Delta_T,Z_i) - \ell(W_T,Z_i)\right] \tag{36}$$

$$\approx \mathop{\mathbb{E}}_{W_T,Z,\Delta_T}\left[\langle\nabla_w\ell(W_T,Z),\Delta_T\rangle + \frac{1}{2}\Delta_T^T\mathrm{H}_{W_T}(Z)\Delta_T\right] \tag{37}$$

$$= \mathop{\mathbb{E}}_{W_T,Z,\Delta_T}\left[\frac{1}{2}\Delta_T^T\mathrm{H}_{W_T}(Z)\Delta_T\right] \tag{38}$$

$$= \frac{1}{2}\langle\mathop{\mathbb{E}}_{W_T,Z}\left[\mathrm{H}_{W_T}(Z)\right], \mathop{\mathbb{E}}_{\Delta_T}\left[\Delta_T\Delta_T^T\right]\rangle \tag{39}$$

$$= \frac{1}{2}\langle\mathop{\mathbb{E}}_{W_T,Z}\left[\mathrm{H}_{W_T}(Z)\right], \mathrm{diag}(\sum_{t=1}^{T}\sigma_t^2)\rangle \tag{40}$$

$$= \frac{\sum_{t=1}^{T}\sigma_t^2}{2}\mathrm{Tr}(\mathop{\mathbb{E}}_{W_T,Z}\left[\mathrm{H}_{W_T}(Z)\right]), \tag{41}$$

where Eq.38 is by the zero mean of the perturbation, Eq.40 is by the independence of the coordinates of Δ_T, $\langle\cdot,\cdot\rangle$ denotes the inner product of two matrices, $\mathrm{diag}(A)$ is the diagonal matrix with element A and $\mathrm{Tr}(\cdot)$ is the trace of a matrix.

Under the condition $\mathop{\mathbb{E}}_{W_T,S'}\left[\gamma(W_T,S')\right] \geq 0$, we now bound $\mathrm{gen}(\mu, \mathrm{P}_{\widetilde{W}_T|S})$ instead of its absolute value, $|\mathrm{gen}(\mu, \mathrm{P}_{\widetilde{W}_T|S})|$. The following is straightforward,

$$\mathrm{gen}(\mu, \mathrm{P}_{\widetilde{W}_T|S}) \leq \frac{R}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\frac{2\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}\left[\mathbb{V}(W_{t-1})\right]} + \mathop{\mathbb{E}}_{W_T,S,S'}\left[\gamma(W_T,S) - \gamma(W_T,S')\right] \tag{42}$$

$$\leq \frac{R}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\frac{2\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}\left[\mathbb{V}(W_{t-1})\right]} + \mathop{\mathbb{E}}_{W_T,S}\left[\gamma(W_T,S)\right] \tag{43}$$

$$\lesssim \frac{R}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\frac{2\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}\left[\mathbb{V}(W_{t-1})\right]} + \frac{\sum_{t=1}^{T}\sigma_t^2}{2}\mathrm{Tr}(\mathop{\mathbb{E}}_{W_T,Z}\left[\mathrm{H}_{W_T}(Z)\right]). \tag{44}$$

This completes the proof. □

B.6 Proof of Lemma 6

Proof. Notice that if each instance is trained only once for every epoch, it's easy to see that $\sum_{i=1}^{n}\sum_{t\in\mathcal{T}_i} = b\sum_{t=1}^{T}$. By $\sqrt{\sum_i x_i} \le \sum_i \sqrt{x_i}$, we have

$$\frac{R}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\frac{2\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})]} \le \frac{R}{n}\sum_{t=1}^{T}\sqrt{\frac{2\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})]}. \tag{45}$$

Or alternatively, square root is a concave function. By Jensen's inequality, we have

$$\frac{R}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\frac{2\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})]} \le \sqrt{\frac{2R^2}{nb}\sum_{t=1}^{T}\frac{\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})]}. \tag{46}$$

This completes the proof. □

B.7 Proof of Corollary 2

Proof. Recall the smoothness implies $f(\mathbf{v}) \le f(\mathbf{w}) + \langle\nabla f(\mathbf{w}), \mathbf{v}-\mathbf{w}\rangle + \frac{\beta}{2}||\mathbf{v}-\mathbf{w}||^2$ for all $\mathbf{v}$ and $\mathbf{w}$. By the triangle inequality, we have

$$|\mathbb{E}\left[L_\mu(W_T) - L_\mu(W_T+\Delta_T)\right]| \le |\mathbb{E}\left[\langle\nabla_w\ell(W_T,Z),\Delta_T\rangle\right]| + \frac{\beta}{2}\mathbb{E}\left[||\Delta_T||^2\right] = \frac{\beta d\sum_{t=1}^{T}\sigma_t^2}{2} \tag{47}$$

Thus, we can see that $|\mathbb{E}\left[L_\mu(W_T) - L_\mu(W_T+\Delta_T)\right]| + |\mathbb{E}\left[L_S(W_T+\Delta_T) - L_S(W_T)\right]| \le \beta d\sum_{t=1}^{T}\sigma_t^2$. □

C Background on Strong Data Processing Inequality

For a Markov chain $U \to X \to Y$, the well-known data-processing inequality [18] states that $I(U;Y) \le I(U;X)$. It turns out that this inequality can often be tightened in the form of

$$I(U;Y) \le \eta I(U;X) \tag{48}$$

for some $\eta < 1$. Such an inequality is referred to as a strong data-processing inequality (SDPI). Note the coefficient η fundamentally depends on the contraction property of the stochastic kernel $P_{Y|X}$, characterizing the extent by which the kernel brings two different distributions P_X and $P_{X'}$ closer after the stochastic mapping. As the reader may refer to [57–59] for a full exposition of this subject, here for self-containedness, we here make an modest revision of the settings in [57–59] and develop slightly finer results concerning the contraction properties of $P_{Y|X}$. Specifically, the contraction coefficient η in our setting will not only depend on the kernel $P_{Y|X}$, it will also depend on the "effective input space" which the kernel acts upon.

To that end, we will denote by $\mathcal{U}$, $\mathcal{X}$, and $\mathcal{Y}$ the spaces in which U, X, Y take values, respectively. For any distribution P on $\mathcal{X}$, we will use $P_{Y|X}\circ P$ to denote the distribution on $\mathcal{Y}$ induced by the push-forward of the distribution P by $P_{Y|X}$, namely, for any $y\in\mathcal{Y}$,

$$\left(P_{Y|X}\circ P\right)(y) \triangleq \int P_{Y|X}(y|x)P(x)dx$$

For the Markov chain $U \to X \to Y$, we will denote by $\Omega(U)$ the support of the distribution P_U. That is, $\mathcal{S}(U)$ is the subset of $\mathcal{U}$ on which P_U is strictly positive. Let $\mathcal{H}(U, P_{X|U})$ be the convex hull of $\{P_{X|U=u} : u\in\mathcal{S}(U)\}$, namely, $\mathcal{H}(U,P_{X|U})$ contains all distributions on $\mathcal{X}$ which can be expressed as $P_{X|U}\circ P$ for some distribution P on $\mathcal{U}$ whose support is a subset (not necessarily proper) of $\mathcal{S}(U)$. It is apparent that $P_X \in \mathcal{H}(U,P_{X|U})$.

Given the Markov chain, $U\to X\to Y$, we now define the contraction coefficient η as

$$\eta(U\to X\to Y) \triangleq \sup_{P,Q\in\mathcal{H}(U,P_{X|U})}\frac{\mathrm{D_{KL}}(P_{Y|X}\circ P||P_{Y|X}\circ Q)}{\mathrm{D_{KL}}(P||Q)} \tag{49}$$

where $\mathrm{D_{KL}}$ denotes the KL divergence. We note that this definition of the contraction coefficient differs from the standard definition [57–59] in that the supremization in the latter is over all P, Q which are distributions on $\mathcal{X}$, making the contraction coefficient only depends on $P_{Y|X}$. In our definition (49), the coefficient η also depends on the "effective input space" of the kernel, namely, $\mathcal{H}(U, P_{X|U})$.

Some standard results concerning the contraction coefficient can be easily extended to this revised definition of η in (49), which we state below.

Lemma 8. *For any Markov chain $U \to X \to Y$,*

$$I(U;Y) \le \eta(U \to X \to Y) I(U;X). \tag{50}$$

Additionally, if $Y = X + N$ for a Gaussian noise $N \sim \mathcal{N}(0, \delta^2 \mathbf{I})$ independent of (U, X) then

$$\eta(U \to X \to Y) \le 1 - 2\mathsf{Q}\left(\frac{\mathrm{Diam}(\mathcal{H}(U, P_{X|U}))}{2\delta}\right) \tag{51}$$

where $\mathrm{Diam}(\mathcal{H}(U, P_{X|U}))$ refers to the diameter of $\mathcal{H}(U, P_{X|U})$, measured under L2-distance, and Q is standard Q-function, or the complementary of Gaussian CDF.

We note that the SDPI (50) follows easily from the definition of mutual information and that of η. The inequality (51) is an adaptation, to this context, of the result that the contraction coefficient is upper bounded by the *Dobrushin's coefficient* of the kernel $P_{Y|X}$ [21]. More details about the proof of Lemma 8 can be found in [57].

Concerning the contraction coefficients $\eta_{i,t}$, we note that from a purely theoretical perspective, it is rather difficult to estimate them or upper-bound them by a quantity smaller than 1. This is because of the Gaussian noise N_t injected at each step t having unbounded support. As a consequence, $\mathcal{H}(Z_i, P_{\widetilde{W}_t|V_t})$ is unbounded, making the restriction of the range of supremization to $\mathcal{H}(Z_i, P_{\widetilde{W}_t|V_t})$ useless.

D Proofs for Section 6

In this section, we will prove Theorem 3 that is a stronger version of Theorem 2. To prove Theorem 2, we only need to let the contraction coefficient η be 1, which means using a weak version of the data-processing inequality.

We first apply the SDPI in Lemma 8 to auxiliary weight process and the SGD weight process.

For notational convenience, we denote

$$G_t \triangleq -\lambda_t g(W_{t-1}, B_t), \text{ and } V_t \triangleq \widetilde{W}_{t-1} + G_t. \tag{52}$$

Thus V_t and $\widetilde{W}_t$ differ only by the noise N_t, or $\widetilde{W}_t = V_t + N_t$.

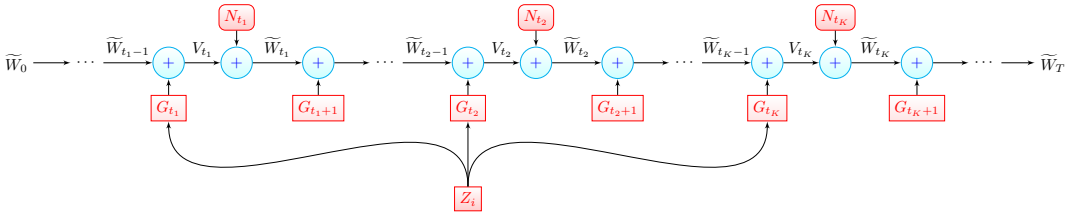

Figure 5: The relationship between Z_i, $\widetilde{W}_t$'s V_t's, and G_t's

Consider a fixed example Z_i in the training set S. Let $\mathcal{T}_i$ denote the of all batch indices t for which B_t contains Z_i. We will assume that $\mathcal{T}_i$ contains K indices and consider $\mathcal{T}_i$ as $\{t_1, t_2, \ldots, t_K\}$. The relationship between Z_i, $\widetilde{W}_t$'s, V_t's, and G_t's are shown in Figure 5. It is clear that $Z_i \to V_t \to \widetilde{W}_t$ form a Markov chain[3]. Denote

$$\eta_{i,t} \triangleq \eta(Z_i \to V_t \to \widetilde{W}_t).$$

[3]Note that even for $t < t_1$, in which Z_i and V_t are independent, the Markov chain still holds, although it degenerates.

We have the following result.

Lemma 9. *Suppose that $I(G_t + N_t; Z_i|\widetilde{W}_{t-1}) \leq C_t$ for some positive value C_t. Then*

$$I(\widetilde{W}_T; Z_i) \leq \sum_{k=1}^{K} C_{t_k} \prod_{\tau \in \Gamma_i^k} \eta_{i,t},$$

where $\Gamma_i^k = \{t_k + 1, t_k + 2, \ldots, T\} \setminus \mathcal{T}_i$.

Proof. For the step t_K, which is the last time Z_i appeared in the training process, we have

$$\begin{aligned} I(\widetilde{W}_T; Z_i) &\leq \eta_{i,T} I(V_T; Z_i) && (53)\\ &\leq \eta_{i,T} I(\widetilde{W}_{T-1}; Z_i) && (54)\\ &\leq \left(\prod_{\tau=t_K+1}^{T} \eta_{i,\tau} \right) I(\widetilde{W}_{t_K}; Z_i), && (55) \end{aligned}$$

where Eq.53 and Eq.54 is due to Eq.50 and the data processing inequality, respectively. Eq.55 is by applying these two steps recursively.

Notice that

$$\begin{aligned} I(\widetilde{W}_{t_K}; Z_i) = I(\widetilde{W}_{t_K-1} + G_{t_K} + N_{t_K}; Z_i) &\leq I(\widetilde{W}_{t_K-1}, G_{t_K} + N_{t_K}; Z_i) && (56)\\ &= I(\widetilde{W}_{t_K-1}; Z_i) + I(G_{t_K} + N_{t_K}; Z_i|\widetilde{W}_{t_K-1}) && (57)\\ &\leq I(\widetilde{W}_{t_K-1}; Z_i) + C_{t_K}, && (58) \end{aligned}$$

where Eq.56 is by $I(f(X); Y) \leq I(X; Y)$ and Eq.57 is by the chain rule of mutual information. Combine Eq.55 and Eq.58, we have

$$I(\widetilde{W}_T; Z_i) \leq \left(\prod_{\tau=t_K+1}^{T} \eta_{i,\tau} \right) \left(I(\widetilde{W}_{t_K-1}; Z_i) + C_{t_K} \right). \tag{59}$$

Then we can apply the similar procedure, namely Eq.53-58, to $I(\widetilde{W}_{t_K-1}; Z_i)$ and get

$$I(\widetilde{W}_{t_K-1}; Z_i) \leq \left(\prod_{\tau=t_{K-1}+1}^{t_K-1} \eta_{i,\tau} \right) \left(I(\widetilde{W}_{t_{K-1}-1}; Z_i) + C_{t_{K-1}} \right), \tag{60}$$

where t_{K-1} is the second-to-last time that Z_i is used in the training process. Plugging Eq.60 into Eq.59,

$$I(\widetilde{W}_T; Z_i) \leq C_{t_K} \cdot \prod_{\tau=t_K+1}^{T} \eta_{i,\tau} + \left(C_{t_{K-1}} + I(\widetilde{W}_{t_{K-1}-1}; Z_i) \right) \cdot \prod_{\substack{\tau=t_{K-1}+1 \\ \tau \neq t_K}}^{T} \eta_{i,\tau}. \tag{61}$$

Finally, we apply this procedure recursively and given the fact that $I(\widetilde{W}_t; Z_i) = 0$ for $t < t_1$, we have

$$I(\widetilde{W}_T; Z_i) \leq \sum_{k=1}^{K} C_{t_k} \prod_{\tau \in \Gamma_i^k} \eta_{i,\tau}, \tag{62}$$

where $\Gamma_i^k = \{t_k + 1, t_k + 2, \ldots, T\} \setminus \mathcal{T}_i$. □

D.1 Proof of Theorem 3

Proof. Let the generalization error of SGD be decomposed by

$$\mathrm{gen}(\mu, P_{W_T|S}) = \mathrm{gen}(\mu, P_{\widetilde{W}_T|S}) + \mathbb{E}\left[L_\mu(W_T) - L_\mu(\widetilde{W}_T) \right] + \mathbb{E}\left[L_S(\widetilde{W}_T) - L_S(W_T) \right]. \tag{63}$$

Then we use Lemma 2 to bound the first term,

$$\operatorname{gen}(\mu, P_{\widetilde{W}_T|S}) \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2R^2 I(\widetilde{W}_T; Z_i)} + \mathbb{E}\left[L_\mu(W_T) - L_\mu(\widetilde{W}_T)\right] + \mathbb{E}\left[L_S(\widetilde{W}_T) - L_S(W_T)\right]. \tag{64}$$

Given a specific sample path, let the step t be the last iteration that Z_i is used. Then the following Markov chain holds,

$$Z_i \to V_t \to \widetilde{W}_t \to V_{t+1} \to \widetilde{W}_{t+1} \to \cdots \to V_T \to \widetilde{W}_T.$$

For mutual information $I(\widetilde{W}_T; Z_i)$ in Eq.64, by using the strong data processing inequality and the data processing inequality, we have

$$I(\widetilde{W}_T; Z_i) \leq I(V_{t+1}; Z_i) \cdot \prod_{\tau=t+1}^{T} \eta_{i,\tau} \leq I(\widetilde{W}_t; Z_i) \cdot \prod_{\tau=t+1}^{T} \eta_{i,\tau}. \tag{65}$$

Further, we would like to bound the term $I(\widetilde{W}_t; Z_i)$. Notice that

$$\begin{aligned} I(\widetilde{W}_t; Z_i) &= I\left(\widetilde{W}_{t-1} + G_t + N_t; Z_i\right) && (66)\\ &\leq I\left(\widetilde{W}_{t-1}, G_t + N_t; Z_i\right) && (67)\\ &= I(\widetilde{W}_{t-1}; Z_i) + I\left(-g(W_{t-1}, B_t) + \frac{1}{\lambda_t}N_t; Z_i | \widetilde{W}_{t-1}\right). && (68) \end{aligned}$$

Thus, applying Lemma 9 and Lemma 5 (i.e., by letting $C_t = \frac{\lambda_t^2}{\sigma_t^2 b^2}\mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})]$), we have

$$I(\widetilde{W}_T; Z_i) \leq \sum_{t\in\mathcal{T}_i} \frac{\lambda_t^2}{\sigma_t^2 b^2}\mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})] \cdot \prod_{\tau\in\Gamma_i^t} \eta_{i,\tau}, \tag{69}$$

where $\mathcal{T}_i$ is the set of all indices of iterations that contains Z_i. Thus, the following bound holds,

$$|\operatorname{gen}(\mu, \mathrm{P}_{\widetilde{W}_T|S})| \leq \frac{R}{nb}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\frac{2\lambda_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{W_{t-1}}[\mathbb{V}(W_{t-1})] \cdot \prod_{\tau\in\Gamma_i^t}\eta_{i,\tau}} + \left|\mathop{\mathbb{E}}_{W_T,S,S'}[\gamma(W_T, S) - \gamma(W_T, S')]\right|, \tag{70}$$

which concludes the proof. □

E Potential Negative Societal Impacts

In this paper, we derived information-theoretic generalization bounds for SGD and proposed a regularization scheme. Our work is a foundational research and we aim to understand a fundamental problem of deep learning, that is, the generalization ability of SGD. Although our work does not have direct negative societal impacts, it's possible to design new algorithms based on the discussion in this paper, which could be tied to any particular application. In this case, our work may have some unexpected negative societal impacts. For example, if these new algorithms were abused by high-tech companies, people's privacy would be easily violated.