Understanding Flatness in Generative Models: Its Role and Benefits

1. Mathematical claims and proofs

For the main claims, we follow $\mathcal{L}(\mathbf{x};\boldsymbol{\theta},t,p_t) := ||s_{\boldsymbol{\theta}}(\mathbf{x},t) - \nabla_{\mathbf{x}} \log p_t(\mathbf{x})||_2^2$, while dropping the timestep t without loss of generality. Our mathematical claims are valid for all timesteps.

Definition 1. *(Δ-flat minima) Let us consider a SGM with loss function $\mathcal{L}(\mathbf{x};\boldsymbol{\theta},p)$. A minimum $\boldsymbol{\theta}^*$ is Δ-flat minima when the following constraints are hold:*

$$\forall\, \delta \in \mathbb{R}^{d\times m} \text{ s.t. } \|\delta\|_2 \le \Delta,\ \ \mathcal{L}(\mathbf{x};\boldsymbol{\theta}^*+\delta,p) = l^*$$
$$\exists\, \delta \in \mathbb{R}^{d\times m} \text{ s.t. } \|\delta\|_2 > \Delta,\ \ \mathcal{L}(\mathbf{x};\boldsymbol{\theta}^*+\delta,p) > l^*,$$

where $l^* := \mathcal{L}(\mathbf{x};\boldsymbol{\theta}^*,p)$ and $\Delta \in \mathbb{R}^+$.[1]

Definition 2. *($\mathcal{E}$-distribution gap robustness) A minimum $\boldsymbol{\theta}^*$ is $\mathcal{E}$-distribution gap robust when the following constraints are hold:*

$$\forall\, \hat{p}(\mathbf{x}) \text{ s.t. } D(p||\hat{p}) \le \mathcal{E},\ \ \mathcal{L}(\mathbf{x};\boldsymbol{\theta}^*,\hat{p}) = l^*$$
$$\exists\, \hat{p}(\mathbf{x}) \text{ s.t. } D(p||\hat{p}) > \mathcal{E},\ \ \mathcal{L}(\mathbf{x};\boldsymbol{\theta}^*,\hat{p}) > l^*,$$

where $D(\cdot||\cdot)$ is the divergence between two probability density functions, $\hat{p}$ is the perturbed prior distribution of $\mathbf{x}$, and $\mathcal{E}$ is a positive real number.

Theorem 1. *(A perturbed distribution) For a given prior distribution of $p(\mathbf{x})$ and the δ-perturbed minimum, i.e., $\boldsymbol{\theta}+\delta$, the following $\hat{p}(\mathbf{x})$ satisfies the $\mathcal{L}(\mathbf{x};\boldsymbol{\theta}+\delta,p) = \mathcal{L}(\mathbf{x};\boldsymbol{\theta},\hat{p})$:*

$$\hat{p}(\mathbf{x}) = e^{-I(\mathbf{x},\delta)} p(\mathbf{x}), \tag{1}$$

where $I(\mathbf{x},\delta) := \frac{1}{2}\mathbf{x}^\top(\delta\mathbf{W}^\top)\mathbf{x} + \mathbf{x}^\top\delta(\mathbf{U}^\top\mathbf{e}) + C$, and $C \in \mathbb{R}$ is set to satisfy $\int_{\mathbb{R}^d} e^{-I(\mathbf{x},\delta)}p(\mathbf{x})d\mathbf{x} = 1$.

Proof. By following [**?**], we formulate the score model $s_{\boldsymbol{\theta}}(\cdot,\cdot)$ as a random feature model:

$$s_{\boldsymbol{\theta}}(\mathbf{x},t) := \frac{1}{m}\boldsymbol{\theta}\sigma(\mathbf{W}^\top\mathbf{x} + \mathbf{U}^\top\mathbf{e}_t) \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^{d\times 1}$, $\boldsymbol{\theta} \in \mathbb{R}^{d\times m}$, $\mathbf{W} \in \mathbb{R}^{d\times m}$, $\mathbf{U} \in \mathbb{R}^{d_e\times m}$, $\mathbf{e}_t \in \mathbb{R}^{d_e\times 1}$, and d, m, d_e are positive integers.

The score matching loss objective is defined as

$$\mathcal{L}(\mathbf{x};\boldsymbol{\theta},p) := ||s_{\boldsymbol{\theta}}(\mathbf{x}) - \nabla_{\mathbf{x}}\log p(\mathbf{x})||_2^2, \tag{3}$$

For the perturbation $\delta \in \mathbb{R}^{d\times m}$ in the diffusion model parameters $\boldsymbol{\theta}$, the perturbed loss value becomes:

$$\mathcal{L}(\mathbf{x};\boldsymbol{\theta}+\delta,p) := ||s_{\boldsymbol{\theta}+\delta}(\mathbf{x}) - \nabla_{\mathbf{x}}\log p(\mathbf{x})||_2^2. \tag{4}$$

$$s_{\boldsymbol{\theta}+\delta}(\mathbf{x}) - \nabla_{\mathbf{x}}\log p(\mathbf{x}) \tag{5}$$
$$= \frac{1}{m}(\boldsymbol{\theta}+\delta)\sigma(\mathbf{W}^\top\mathbf{x} + \mathbf{U}^\top\mathbf{e}) - \nabla_{\mathbf{x}}\log p(\mathbf{x}) \tag{6}$$
$$= \frac{1}{m}\boldsymbol{\theta}\sigma(\mathbf{W}^\top\mathbf{x} + \mathbf{U}^\top\mathbf{e}) + \frac{1}{m}\delta\sigma(\mathbf{W}^\top\mathbf{x} + \mathbf{U}^\top\mathbf{e}) - \nabla_{\mathbf{x}}\log p(\mathbf{x}) \tag{7}$$

[1] $\forall$ means 'for all,' $\exists$ means 'there exists,' and $\mathbb{R}^+$ indicates the set of positive real numbers

Let us focus on the second and third terms with the assumptions of the positive outputs for the activation function:

$$\frac{1}{m}\delta(\mathbf{W}^\top \mathbf{x} + \mathbf{U}^\top \mathbf{e}) - \nabla_{\mathbf{x}} \log p(\mathbf{x}) \tag{8}$$

Here, let us define $I(\mathbf{x})$ as a function of $\mathbf{x}$, whose derivative is the first term of the previous equation:

$$\frac{\partial I(\mathbf{x})}{\partial \mathbf{x}} = \frac{1}{m}\delta(\mathbf{W}^\top \mathbf{x} + \mathbf{U}^\top \mathbf{e}) \tag{9}$$

Based on it, $I(\mathbf{x}) \in \mathbb{R}$ is

$$I(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top(\delta\mathbf{W}^\top)\mathbf{x} + \mathbf{x}^\top\delta(\mathbf{U}^\top \mathbf{e}) + C, \tag{10}$$

with the assumption $\delta\mathbf{W}^\top$ is symmetric and where C is a constant real number.

$$\begin{aligned}
&\frac{1}{m}\delta(\mathbf{W}^\top \mathbf{x} + \mathbf{U}^\top \mathbf{e}) - \nabla_{\mathbf{x}} \log p(\mathbf{x}) && (11)\\
&= \nabla_{\mathbf{x}} I(\mathbf{x}) - \nabla_{\mathbf{x}} \log p(\mathbf{x}) && (12)\\
&= -\nabla \log \left(e^{-I(\mathbf{x})} p(\mathbf{x}) \right) && (13)
\end{aligned}$$

When C is the real number that satisfies the following condition for the function I with C:

$$\int_{\mathbb{R}^d} e^{-I(\mathbf{x})} p(\mathbf{x}) = 1, \tag{14}$$

then we can define $\hat{p}(\mathbf{x})$ to be a perturbed PDF of inputs:

$$\begin{aligned}
&\hat{p}(\mathbf{x}) := e^{-I(\mathbf{x})} p(\mathbf{x}) && (15)\\
&= \exp\Big\{ -\frac{1}{2m}\mathbf{x}^\top(\delta\mathbf{W}^\top)\mathbf{x} - \frac{1}{m}\mathbf{x}^\top\delta(\mathbf{U}^\top \mathbf{e}) - C) \Big\} p(\mathbf{x}) && (16)
\end{aligned}$$

□

Corollary 1. *(Diffusion version of **Theorem 1**) For a given prior Gaussian distribution of noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and the δ-perturbed minimum, i.e., $\boldsymbol{\theta} + \delta$, the following $\hat{\epsilon}$ satisfies the $\mathcal{L}(\mathbf{x}; \boldsymbol{\theta} + \delta, p) = \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}, \hat{p})$:*

$$\hat{\epsilon} = e^{-I(\mathbf{x}, \delta)} \epsilon = \mathcal{N}(\boldsymbol{\mu}_\delta, \Sigma_\delta), \tag{17}$$

where $\Sigma_\delta := \left(\mathbf{I} + \frac{\delta_w}{m} \right)^{-1}$, $\boldsymbol{\mu}_\delta := \frac{1}{m}\Sigma_\delta \delta_u$.

Proof. We provide the theoretical link that the model satisfying the $\mathcal{E}$-flat in Theorem **1** is also robust to distribution shift caused by the exposure bias.

Before that, we introduce the notations:

- $\epsilon(\mathbf{x})$: the true Gaussian distribution that is known in the training process.
- $\hat{\epsilon}(\mathbf{x})$: the perturbed distribution that caused by the δ model perturbation in Eq. (15).

When we train the diffusion model, we add the noise $\boldsymbol{\epsilon}$ in the forward process and want the diffusion model to predict the $\boldsymbol{\epsilon}$ in the reverse process where $\boldsymbol{\epsilon}$ follows the normal Gaussian distribution, i.e., $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Therefore, the distribution that

the model trains is the normal Gaussian, and we can define the perturbed Gaussian distribution as follows:

$$\hat{\epsilon}(\mathbf{x}) := e^{-I(\mathbf{x},\delta)}\epsilon \tag{18}$$

$$= \exp\left(-\frac{1}{2m}\mathbf{x}^\top\delta_w\mathbf{x} - \frac{1}{m}\mathbf{x}^\top\delta_u - C\right)\cdot\frac{1}{\sqrt{(2\pi)^d}}\exp\left(-\frac{1}{2}\mathbf{x}^\top\mathbf{x}\right), \tag{19}$$

$$\text{where } \delta_w := \delta\mathbf{W}^\top, \delta_u := \delta\mathbf{U}^\top\mathbf{e}, \text{ and } \epsilon \sim \mathcal{N}(0,\mathbf{I})$$

$$= \frac{1}{\sqrt{(2\pi)^d}}\exp\left(-\frac{1}{2}\mathbf{x}^\top(\mathbf{I}+\frac{1}{m}\delta_w)\mathbf{x} - \frac{1}{m}\mathbf{x}^\top\delta_u - C\right) \tag{20}$$

$$= \frac{1}{\sqrt{(2\pi)^d}}\exp\left(-\frac{1}{2}\mathbf{x}^\top(\mathbf{I}+\frac{1}{m}\delta_w)\mathbf{x} - \frac{1}{m}\mathbf{x}^\top\delta_u - \frac{1}{2m^2}\delta_u^\top\Sigma_\delta\delta_u - \frac{1}{2}\log|\Sigma_\delta|\right), \tag{21}$$

$$\text{where } C = \frac{1}{2m^2}\delta_u\Sigma_\delta\delta_u + \frac{1}{2}\log|\Sigma_\delta| \tag{22}$$

$$= \frac{1}{\sqrt{(2\pi)^d|\Sigma_\delta|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\delta)^\top{\Sigma_\delta}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\delta)\right), \text{ where } \Sigma_\delta := \left(\mathbf{I}+\frac{\delta_w}{m}\right)^{-1}, \text{ and } \boldsymbol{\mu}_\delta := -\frac{1}{m}\Sigma_\delta\delta_u \tag{23}$$

Because the $\hat{\epsilon}(\mathbf{x})$ is also the Gaussian distribution, we present the KL Divergence between $\hat{\epsilon}(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\delta, \Sigma_\delta)$ and $\epsilon(\mathbf{x}) = \mathcal{N}(0,\mathbf{I})$ as follows: □

Definition 3. *(A set of perturbed distribution) For a given distribution of $p(\mathbf{x})$, a set of distributions $\hat{\mathcal{P}}(\mathbf{x};p,\Delta)$ is defined as the set of perturbed distributions $\hat{p}(\mathbf{x})$:*

$$\hat{\mathcal{P}}(\mathbf{x};p,\Delta) := \{e^{-I(\mathbf{x},\delta)}p(\mathbf{x}) \mid \|\delta\|_2 \le \Delta\}. \tag{24}$$

Proposition 1. *(A link from Δ-flatness to $\hat{\mathcal{P}}$) A Δ-flat minimum $\boldsymbol{\theta}^*$ achieves the flat loss values for all distributions sampled from the set of perturbed distribution:*

$$\forall\, p \sim \hat{\mathcal{P}}(\mathbf{x};p,\Delta),\ \ \mathcal{L}(\mathbf{x};\boldsymbol{\theta}^*,p) = l^* \tag{25}$$

$$\exists\, p \nsim \hat{\mathcal{P}}(\mathbf{x};p,\Delta),\ \ \mathcal{L}(\mathbf{x};\boldsymbol{\theta}^*,p) > l^*. \tag{26}$$

Theorem 2. *(A link from Δ-flatness to $\mathcal{E}$-gap robustness) A Δ-flat minimum achieves $\mathcal{E}$-distribution gap robustness, such that $\mathcal{E}$ is upper-bounded as follows:*

$$\mathcal{E} \le \max_{\hat{p}\sim\hat{\mathcal{P}}(\mathbf{x};p,\Delta)} D(p||\hat{p}). \tag{27}$$

Corollary 2. *(Diffusion version of **Theorem 2**) For a diffusion model, a Δ-flat minimum achieves $\mathcal{E}$-distribution gap robustness, such that $\mathcal{E}$ is upper-bounded as follows:*

$$\mathcal{E} \le \max_{\|\delta\|_2\le\Delta}\frac{1}{2}\left[\log|\Sigma_\delta| - d + tr(\Sigma_\delta^{-1}) + {\boldsymbol{\mu}_\delta}^\top\Sigma_\delta^{-1}\boldsymbol{\mu}_\delta\right] \tag{28}$$

$$\le \frac{1}{2}\left[\sum_i^d(\sigma_i - \log\sigma_i) - d + \frac{\sigma_d}{m^2}\|\mathbf{U}^\top\mathbf{e}\|_2^2\Delta^2\right] \tag{29}$$

where σ_i is an eigenvalue of Σ_δ^{-1} with the increasing order of $\sigma_1 \le \sigma_2 \le \ldots \le \sigma_d$.

Proof. From the definition **2**, a minimum θ^* hold following:

$$\forall\, \hat{p}(\mathbf{x}) \text{ s.t. } D(p||\hat{p}) \le \mathcal{E},\ \ \mathcal{L}(\mathbf{x};\boldsymbol{\theta}^*,\hat{p}) = l^*.$$

Let σ_i is an eigenvalue of Σ_δ^{-1} with the increasing order $\sigma_1 \leq \sigma_2 \leq \ldots \leq \sigma_d$. Then, the Diffusion version of Theorem **2** is represented as follows:

$$\mathcal{E} \leq \max_{\|\delta\|_2 \leq \Delta} D_{KL}(\epsilon \| \hat{\epsilon}) \tag{30}$$

$$= \max_{\|\delta\|_2 \leq \Delta} \frac{1}{2} \left[\log |\Sigma_\delta| - d + tr(\Sigma_\delta^{-1}) + {\boldsymbol{\mu}_\delta}^\top {\Sigma_\delta}^{-1} \boldsymbol{\mu}_\delta \right] \tag{31}$$

$$= \max_{\|\delta\|_2 \leq \Delta} \frac{1}{2} \left[\sum_i^d \log \frac{1}{\sigma_i} - d + \sum_i^d \sigma_i + {\boldsymbol{\mu}_\delta}^\top {\Sigma_\delta}^{-1} \boldsymbol{\mu}_\delta \right] \tag{32}$$

$$\leq \frac{1}{2} \left[\sum_i^d (\sigma_i - \log \sigma_i) - d + \frac{\sigma_d}{m^2} \|\mathbf{U}^\top \mathbf{e}\|_2^2 \Delta^2 \right], \tag{33}$$

where inequality Eq. (33) holds when $\boldsymbol{\mu}_\delta$ is the eigenvector satisfying $\Sigma_\delta^{-1} \boldsymbol{\mu}_\delta = \sigma_d \boldsymbol{\mu}_\delta$. □

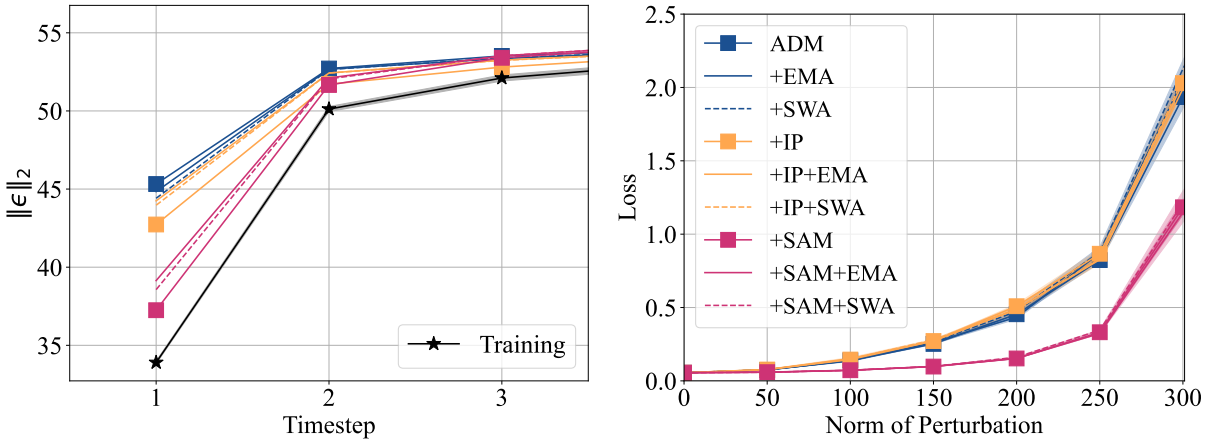

Figure 1. Additional results for CIFAR-10. We measure the L2 norm of predicted noise and loss plots under perturbation for all algorithms including `+IP+EMA`, `+IP+SWA`, `+SAM+EMA`, `+SAM+SWA`.

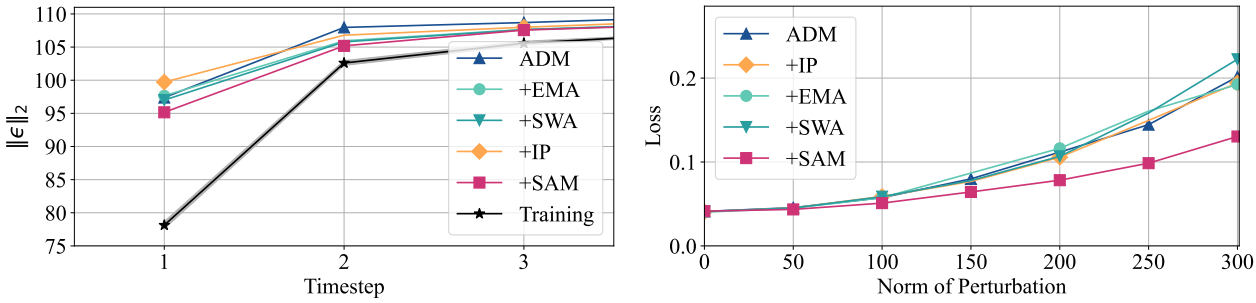

Figure 2. (Left)L2 norm of the predicted noise for LSUN Tower dataset, (Right) Loss plots under perturbation for LSUN Tower.

LPF ↓	w/o	+EMA	+SWA
ADM	0.091	0.090	0.092
+IP	0.089	0.092	0.097
+SAM	**0.072**	**0.070**	**0.071**

↓: a lower value is preferred.

Table 1. Flatness measure on LSUN Tower. We calculate the loss with the perturbed model with Gaussian noise. Lower values indicate a flatter loss landscape.

2. Additional experimental results

2.1. Further results for mixture of baselines

In Fig. 1, we report additional results for `+IP+EMA`, `+IP+SWA`, `+SAM+EMA`, `+SAM+SWA` for CIFAR-10. We observe that ADM already possesses a certain level of flatness supporting `+SWA` and `+EMA` fail to induce additional flatness. We also report L2 norm of predicted noise loss plots under perturbation in Fig. 2 and LPF flatness in Table. 1 for LSUN Tower dataset. It coincides with the result of CIFAR-10 that `+SAM` induces the lower exposure bias and flatter minima.

2.2. Sampling results using fixed random seed

We did not fix the random seed in Fig. 4 of the main paper. We report FID under varying quantization levels with a fixed seed in Fig. 3. As shown, the results remain consistent with our main findings, and all randomly selected samples follow the same trends discussed in the paper.

Figure 3. FID under quantization with fixed random seed.

Dataset	**CIFAR-10 (32x32)**	
T'	20 steps	100 steps
ADM+pEMA	6.58	7.39
SAM+pEMA	6.50	5.46

Table 2. FID of post-hoc EMA (+pEMA) with ADM, **+SAM**.

2.3. Post-hoc EMA

EMA provides significant performance improvements with a simple approach, but it requires cumbersome accumulation of checkpoints during training, and it is also hard to obtain a new combination of EMA after training is completed. Post-hoc EMA (pEMA) defines averaged coefficients as a power function at time t:

$$\hat{\theta}_\gamma(t) = \frac{\gamma+1}{t^{\gamma+1}} \int_0^t \tau^\gamma \theta(\tau) d\tau, \tag{34}$$

where constant γ controls the sharpness of merged checkpoints and τ^γ determines the time weighting. $\hat{\theta}_\gamma(t)$ is updated as follows:

$$\hat{\theta}_\gamma(t) = \beta_\gamma(t)\hat{\theta}_\gamma(t-1) + (1-\beta_\gamma(t))\theta(t), \quad \beta_\gamma(t) = (1-\frac{1}{t})^{\gamma+1}, \tag{35}$$

which is quite similar to conventional EMA. Tab. 2 and Fig. 3 (c), (e) show that FID performance and sample visualization of post-hoc EMA(pEMA). Compared with Table 2 in the main paper, pEMA achieves the highest FID improvements; it still suffers from LPF value (0.103) even though it searches a wider range of combinations than EMA. Also, pEMA exhibits a lack of robustness, evidenced by a sharp FID degradation of $6.58 \rightarrow 170.09$ under 4-bit quantization. Building upon the analysis of EMA and SWA in the main paper, we argue that well-crafted weight averaging also suffers from poor flatness and robustness.