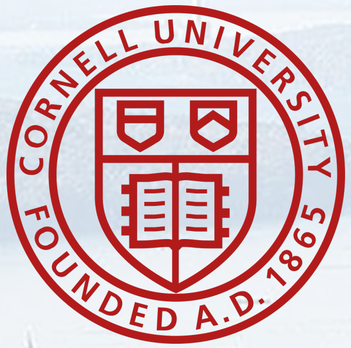



40th Annual Conference On Neural Information Processing Systems

# Provably Reliable Classifier Guidance via Cross-Entropy Control

Sharan Sahu[†], Arisina Banerjee[†], Yuchen Wu[‡]

[†] Department of Statistics and Data Science

[‡] School of Operations Research and Information Engineering

Cornell University

# Score-based diffusion models

**Forward process**
Ornstein–Uhlenbeck (OU) process

$$d\overrightarrow{X}_t = -\overrightarrow{X}_t\,dt + \sqrt{2}\,dB_t, \qquad \overrightarrow{X}_0 \sim P_{\text{data}}$$

**Explicit solution**
Gaussian transition kernel

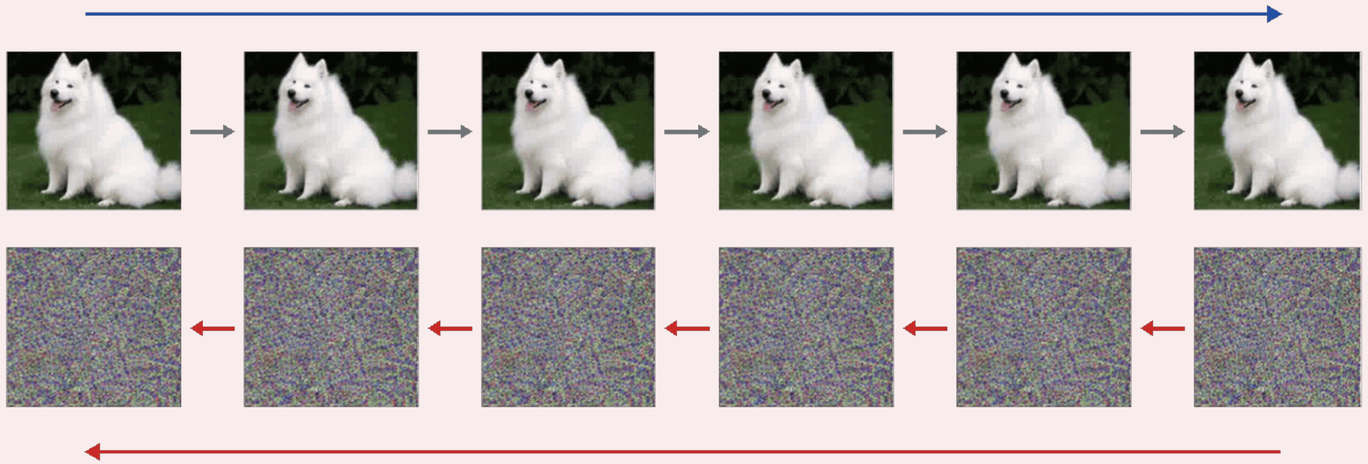
$$\overrightarrow{X}_t = \lambda_t \overrightarrow{X}_0 + \sigma_t Z, \qquad \lambda_t = e^{-t}, \quad \sigma_t^2 = 1 - e^{-2t}$$

**Reverse process**
Time-reversal of the forward process

$$d\overleftarrow{X}_t = \left\{\overleftarrow{X}_t + 2\,\nabla_x \log p_{T-t}(\overleftarrow{X}_t)\right\} dt + \sqrt{2}\,dB'_t, \qquad \overleftarrow{X}_0 \sim p_T$$

**Score matching**
Early stopping cutoff $\delta$

$$\mathcal{L}_{\text{score}}(s_\theta) = \int_\delta^T \mathbb{E}_{X_t \sim p_t} \left\| s_\theta(X_t, t) - \nabla_x \log p_t(X_t) \right\|_2^2 dt$$



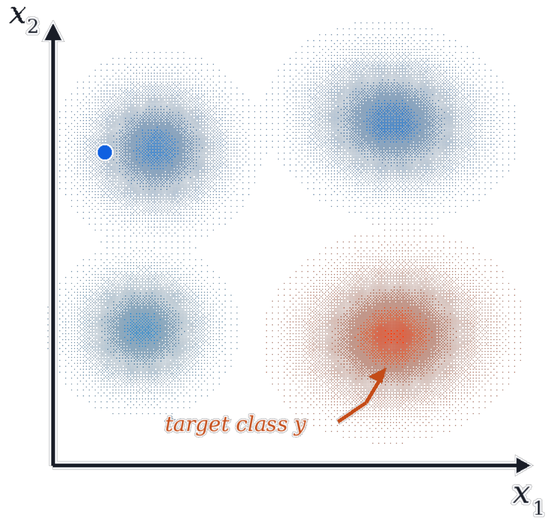
**forward process:** $p_T \approx N(0, I_d)$ for sufficiently large $T$ **reverse process:** sequential denoising with $s_\theta(x,t) \approx \nabla_\theta p_t(x)$ yields samples from $P_{\text{data}}$

# Conditional sampling through classifier guidance

Goal: steer an unconditional diffusion model toward the conditional distribution $X_0 \mid Y = y$

$$\nabla_x \log p_t(x \mid y) = \underbrace{\nabla_x \log p_t(x)}_{\text{unconditional score}} + \underbrace{\nabla_x \log p_t(y \mid x)}_{\text{guidance term}}$$

A trained probabilistic classifier $\hat{p}_t(y \mid x)$ defines an approximate guidance vector field where $\gamma$ controls the guidance strength




$$\mathrm{d}\overleftarrow{X}_t^{(y)} = \left\{ \overleftarrow{X}_t^{(y)} + 2\nabla_x \log p_{T-t}(\overleftarrow{X}_t^{(y)}) + 2\,\nabla_x \log p_{T-t}(y \mid \overleftarrow{X}_t^{(y)}) \right\} \mathrm{d}t + \sqrt{2}\,\mathrm{d}B'_t, \qquad \overleftarrow{X}_0^{(y)} \sim p_T(\cdot \mid y)$$

$$\mathrm{d}\hat{X}_t^{(y)} = \left\{ \hat{X}_t^{(y)} + 2s_\theta(\hat{X}_t^{(y)}, T-t) + 2\gamma\nabla_x \log \hat{p}_{T-t}(y \mid \hat{X}_t^{(y)}) \right\} \mathrm{d}t + \sqrt{2}\,\mathrm{d}B'_t, \qquad \hat{X}_0^{(y)} \sim \pi_d$$

$\gamma = 1$ approximates $P_{\text{data}}(\cdot \mid y)$ whereas $\gamma > 1$ enhances sample fidelity at the expense of sample diversity [Dhariwal and Nichol. (2021), Ho and Salimans. (2021)].

# Open research question

## Classifier training

Conditional label KL, minimized via the empirical cross-entropy loss

$$\mathcal{L}_{\text{cls}}(\hat{p}_t; t) = \mathbb{E}_{X_t \sim p_t}\Big[D_{\text{KL}}\big(p_t(\cdot \mid X_t) \,\|\, \hat{p}_t(\cdot \mid X_t)\big)\Big]$$

*When does small conditional KL guarantee effective guidance?*

## Sampling dynamics

Depend on the accuracy of the vector field $\nabla_x \log p_t(y \mid x)$

$$\mathcal{E}_{\text{guid}}(t, y) = \mathbb{E}_{X_t \sim p_t(\cdot|y)}\Big[\big\|\nabla_x \log p_t(y \mid X_t) - \nabla_x \log \hat{p}_t(y \mid X_t)\big\|_2^2\Big]$$

**Low classification error alone does not guarantee effective guidance.** Classifier guidance can be brittle: even when the classifier predicts labels correctly, its gradients can point in the wrong direction, leading to oversaturated or semantically inconsistent samples [Dinh et al. (2023), Ganz et al. (2023), Wallace et al. (2023)].

# Our contributions

1. Without sufficient smoothness, accurate classifiers do not provide sufficient guidance.

2. With sufficient smoothness, accurate classifiers with conditional KL of order $\mathcal{O}(\varepsilon^2)$ provide sufficient guidance of order $\tilde{\mathcal{O}}(d\varepsilon)$ with this rate being optimal.

3. Using our bound on the guidance error, we bound the DDPM sampling error $D_{\mathrm{KL}}(q_{\gamma,T-\delta} \,||\, \hat{q}_{\gamma,T-\delta})$ for guidance strength $\gamma > 1$ and early stopping time $\delta$, where $q_{\gamma,t}$ and $\hat{q}_{\gamma,t}$ are the target and generated distributions, respectively.

# Without smoothness, accurate classifiers do not provide good guidance

**Theorem 3.1**

For any density $p$ such that $p(\{x : 0 < p(y \mid x) < 1\} \mid y) > 0$ and the map $x \mapsto p(y \mid x)$ is differentiable, there exists a sequence of classifiers $\{\hat{p}_n\}$ such that, almost surely over $x \sim p$,

$$D_{\mathrm{KL}}\big(p(\cdot|x) \,\|\, \hat{p}_n(\cdot|x)\big) \to 0 \quad \text{while} \quad \liminf_{n\to\infty} \mathbb{E}_{x\sim p(\cdot|y)} \big\| \nabla_x \log p(y|x) - \nabla_x \log \hat{p}_n(y|x) \big\|_2^2 > 0$$

In fact, we can choose $\{\hat{p}_n\}$ such that either

1. $\sup_x D_{\mathrm{KL}}\big(p(\cdot|x) \,\|\, \hat{p}_n(\cdot|x)\big) = O(1/n)$ and $\mathbb{E}_{x\sim p(\cdot|y)} \big\| \nabla_x \log p(y|x) - \nabla_x \log \hat{p}_n(y|x) \big\|_2^2 = \Omega(1)$
2. $\sup_x D_{\mathrm{KL}}\big(p(\cdot|x) \,\|\, \hat{p}_n(\cdot|x)\big) = O(1/\sqrt{n})$ and $\mathbb{E}_{x\sim p(\cdot|y)} \big\| \nabla_x \log p(y|x) - \nabla_x \log \hat{p}_n(y|x) \big\|_2^2 = \Omega(n)$

**Remark 3.2.** The conditions only require differentiability and that we cannot identify the target label $y$ with probability one. Any positive result must rely on additional regularity assumptions on the classifier.

# With smoothness, accurate classifiers can provide good guidance

Reverse-time grid $0 = t_0 < t_1 < \cdots < t_N = T - \delta$; $s_k = T - t_k$ is the forward time at which the classifier is used.

**(1)** $P_{\text{data}}$ is supported on a bounded set $K \subseteq \mathbb{R}^d$, with $R = \sup_{x \in K} \|x\|_2 < \infty$

**(2)** The predicted probability $x \mapsto \hat{p}_{s_k}(y \mid x)$ belongs to $C^2(\mathbb{R}^d)$, and

$$\|\nabla_x \log \hat{p}_{s_k}(y|x)\|_2 \le 2\lambda_{s_k}\sigma_{s_k}^{-2}R, \qquad \left|\mathrm{Tr}\left(\nabla_x^2 \log \hat{p}_{s_k}(y|x)\right)\right| \le 2\lambda_{s_k}^2\sigma_{s_k}^{-4}R^2$$

**(3)** High-probability bounds under $p_{s_k}(\cdot \mid y)$, for sufficiently small $\epsilon$:

$$\mathbb{P}_{x\sim p_{s_k}(\cdot|y)}\left(\|\nabla_x \log \hat{p}_{s_k}(y|x)\|_2 \le \frac{C\sqrt{d+\log(1/\epsilon)}}{\sigma_{s_k}}\right) \ge 1 - \frac{\epsilon}{P_{\text{data}}(y)}$$

$$\mathbb{P}_{x\sim p_{s_k}(\cdot|y)}\left(\left|\mathrm{Tr}\left(\nabla_x^2 \log \hat{p}_{s_k}(y|x)\right)\right| \le \frac{C(d+\log(1/\epsilon))}{\sigma_{s_k}^2}\right) \ge 1 - \frac{\epsilon}{P_{\text{data}}(y)}$$

**(4)** The target label satisfies $P_{\text{data}}(y) > 0$

**(5)** Classification error, measured by the conditional KL, at each step:

$$\mathbb{E}_{x\sim p_{s_k}}\left[D_{\text{KL}}\left(p_{s_k}(\cdot|x) \,\|\, \hat{p}_{s_k}(\cdot|x)\right)\right] \le \varepsilon_{s_k}^2, \qquad \varepsilon_{s_k} \in \left[0,\, P_{\text{data}}(y)/\sqrt{2}\right]$$

**Remark 3.5**

Assumptions (2)–(3) hold if $\hat{p}_{s_k}$ is replaced with the true probability $p_{s_k}$ (Lemma B.1, Corollary B.7): the OU forward process smooths the data distribution at all positive times $t > 0$.

# With smoothness, accurate classifiers can provide good guidance

**Theorem 3.3** Under Assumptions (1)–(5), for all $k$ there exists a numerical constant $C_0 > 0$ such that

$$\mathbb{E}_{x\sim p_{s_k}(\cdot|y)}\left\|\nabla_x \log p_{s_k}(y|x) - \nabla_x \log \hat{p}_{s_k}(y|x)\right\|_2^2 \;\le\; \frac{C_0\,\varepsilon_{s_k}}{\sigma_{s_k}^2\,P_{\mathrm{data}}(y)}\left[d + \log\frac{R^2+d}{\sigma_{s_k}^2} + \log\frac{1}{\varepsilon_{s_k}} + \log\frac{1}{P_{\mathrm{data}}(y)}\right]$$

**Theorem 3.7** There exists a data distribution $P_{\mathrm{data}}$ such that for any sufficiently small $\varepsilon$, there exist probabilistic classifiers satisfying Assumptions (1)–(5) with

$$\sup_t \mathbb{E}_{X_t\sim p_t}\left[D_{\mathrm{KL}}(p_t(\cdot|X_t)\,\|\,\hat{p}_t(\cdot|X_t))\right] \le C_1\,\varepsilon^2 \qquad \inf_t \mathbb{E}_{X_t\sim p_t(\cdot|y)}\left\|\nabla_x \log p_t(y|X_t) - \nabla_x \log \hat{p}_t(y|X_t)\right\|_2^2 \ge C_2\,\varepsilon$$

1. **Scales linearly with the dimension $d$**
   This dependence cannot be improved (Remark 3.10).

2. **Complete characterization**
   With Theorem 3.1 and Theorem 3.3, we provide sufficient conditions for when accurate classifiers induce effective classifier guidance.

# Sampling error of classifier-guided diffusion models

DDPM-type sampler with an exponential integrator on $0 = t_0 < t_1 < \cdots < t_N = T - \delta$, with $\tau_k = t_{k+1} - t_k$ and $s_k = T - t_k$, starting from $\hat{X}_0 \sim \mathcal{N}(0, I_d)$; at each step it adds the score estimate $s_\theta(\cdot, s_k)$ plus $\gamma$ times the classifier guidance $\nabla \log \hat{p}_{s_k}(y \mid \cdot)$, plus Gaussian noise. Output distribution: $\hat{q}_{\gamma, T-\delta}$. Compared against the ideal guided marginal $q_{\gamma, T-\delta}$ of the ideal classifier-guided reverse process started from $p_{T,\gamma}(x) = p_T(x)\, p_T(y \mid x)^\gamma / Z_{T,\gamma}$.

**Theorem 3.9** Suppose $T \geq 1$ and $\tau_k \leq \kappa \min\{1, s_{k+1}\}$. Under Assumption 3.8 (score estimation error $\varepsilon^2_{\text{score}}$) and the assumptions of Theorem 3.3, for $\gamma > 1$,

$$D_{\mathrm{KL}}\left(q_{\gamma,T-\delta} \,\|\, \hat{q}_{\gamma,T-\delta}\right) \leq C_{\mathrm{s}} \left( \frac{\underbrace{\varepsilon^2_{\text{score}}}_{\text{score estimation}} + \underbrace{\gamma^2 \varepsilon^2_{\text{guide}}}_{\text{guidance}} + \underbrace{(2\gamma^2 - 2\gamma + 1)(d + R^2)(\kappa T + \kappa^2 N)}_{\text{discretization}}}{P_{\text{data}}(y)^\gamma} + \underbrace{\frac{\gamma^2 (d + R^2) e^{-2T}}{P_{\text{data}}(y)^{2\gamma}}}_{\text{initialization}} \right)$$

## Guidance error

$\varepsilon^2_{\text{guide}} = \sum \tau_k \, \mathbb{E}_{x \sim p_{s_k}(\cdot|y)} \|\nabla_x \log p_{s_k}(y|x) - \nabla_x \log \hat{p}_{s_k}(y|x)\|_2^2$:

the average guidance term MSE, controlled by the conditional KL at each step through Theorem 3.3.

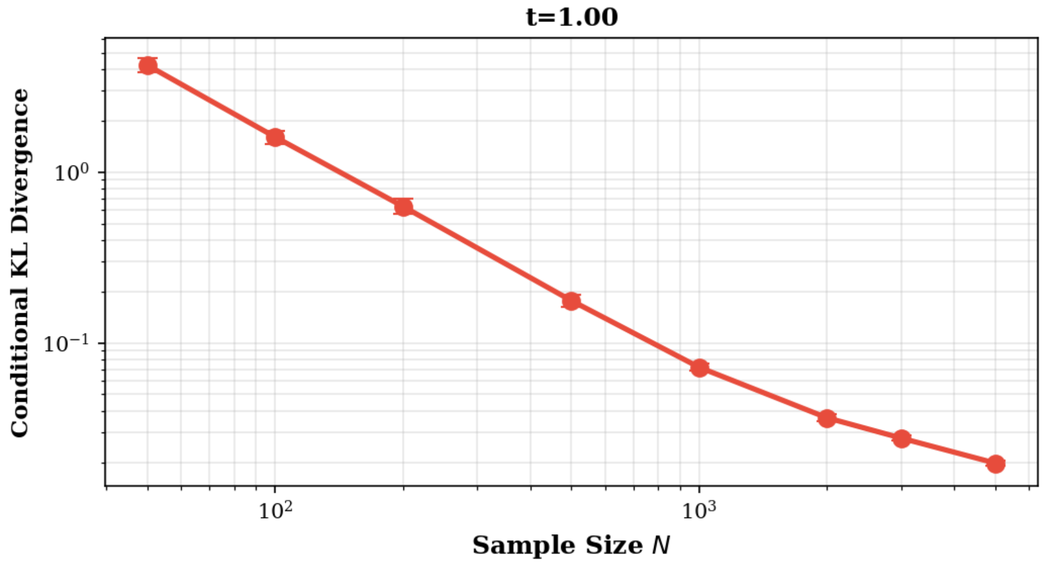
## Discretization and initialization

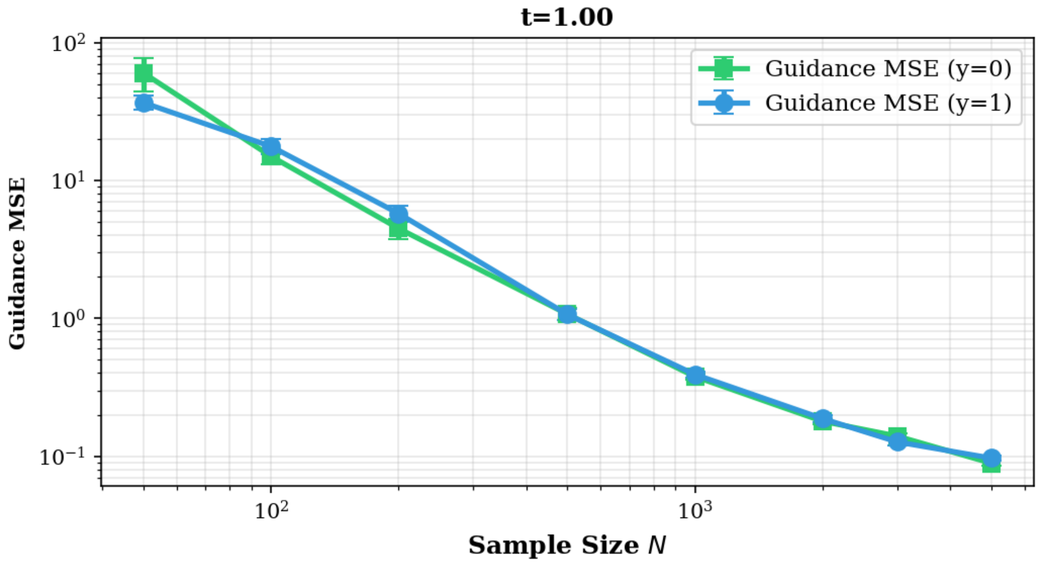
Controlled following [BBDD24] (Lemmas B.10, B.11); the initialization error decays as $e^{-2T}$.

# Guidance error control with classifier smoothness: GMM

$$p(x \mid y = j) = \mathcal{N}(\mu_j,\, 0.5\, I_{50}), \quad \mu_0 = -\mu_1 = (-2, 0, \ldots, 0)^\top, \quad p(y = 0) = p(y = 1) = \tfrac{1}{2}$$

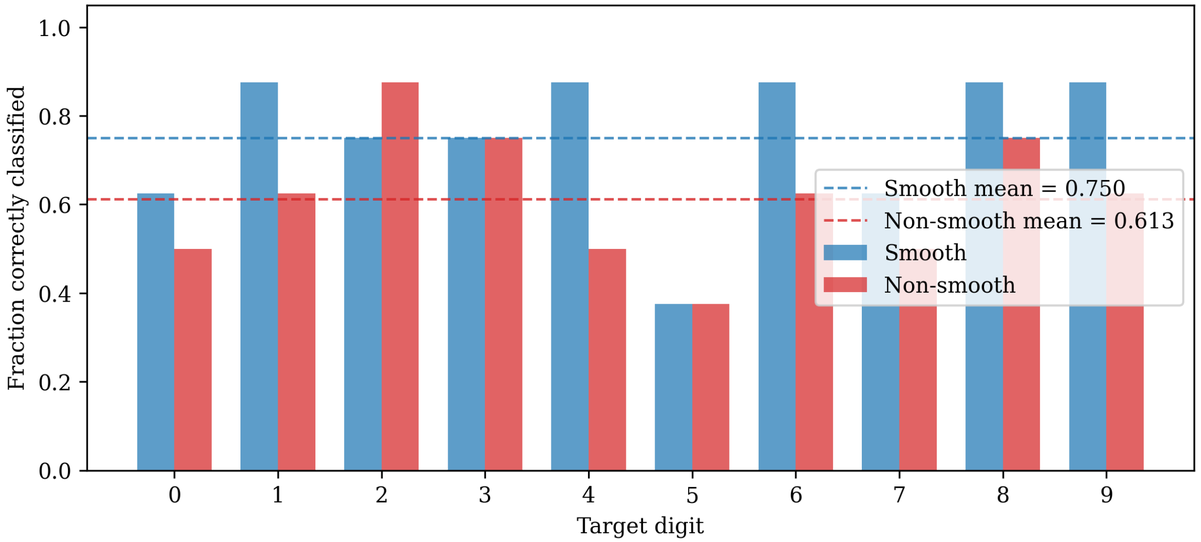
Forward noising process; $p_t(y \mid x)$ computed analytically via Bayes' rule: $p_t(x \mid y = j) = \mathcal{N}\big(\lambda_t \mu_j,\ (0.5\lambda_t^2 + \sigma_t^2)\, I_{50}\big)$

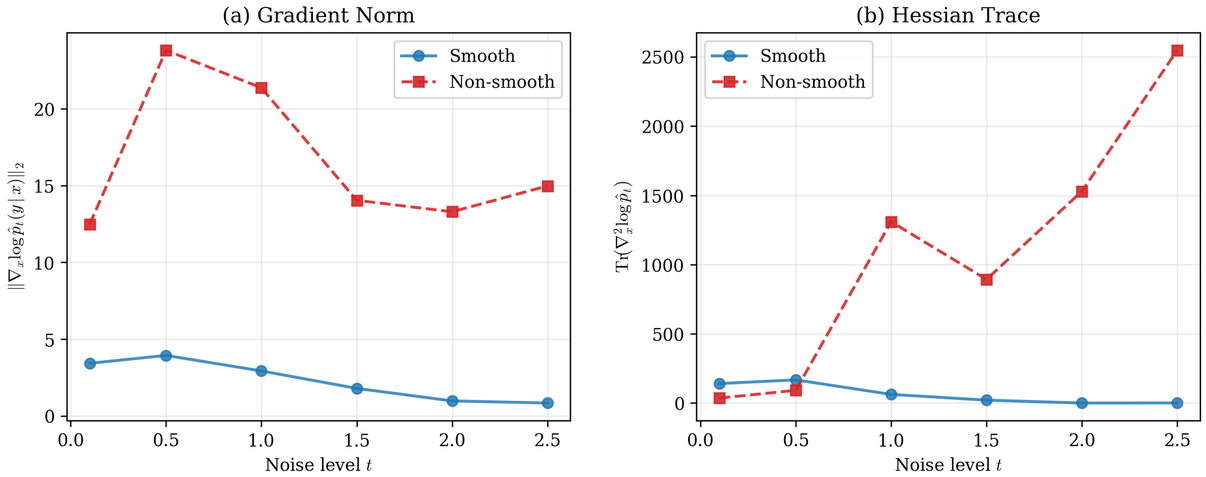


# MNIST: guidance effectiveness

We train two neural classifiers with identical architectures, differing only in their activation functions: a smooth classifier using GELU and tanh, and a non-smooth classifier using ReLU and Hardtanh. Both achieve comparable cross-entropy loss and classification accuracy.




A held-out MNIST classifier identifies images generated using smooth classifier guidance at a substantially higher rate than those generated using non-smooth guidance, demonstrating the benefits of classifier smoothness for effective guidance.

# Thank you for listening!

For further details, please visit:

- Our paper: https://openreview.net/forum?id=5i317VTcld
- Our poster: https://neurips.cc/virtual/2026/loc/sydney/poster/155431

Hall 1-4, Wed, Dec 9, 2026 • 5:00 PM – 8:00 PM AEDT