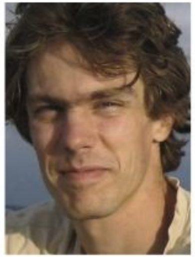
Overshoot and Shrinkage in **Classifier-Free Guidance**: From Theory to Practice

Kruno Lehman
ENS & Meta

Jakob Verbeek
Meta

Giulio Biroli
ENS

Marc Mezard
Bocconi University

Dynamical regimes of diffusion models

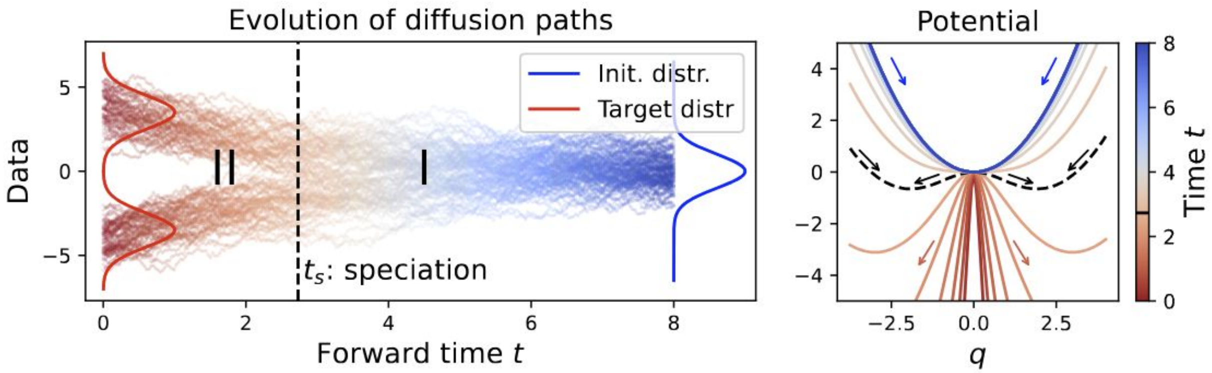


Figure 2. Illustration of Speciation phenomenon on a 1D example of a mixture of two Gaussians.

There exist two regimes of diffusion models (separated by speciation t_s):

Regime I: the class belonging has not been decided yet
Regime II: the class fate is set and class-specific features are generated

$$\Lambda e^{-2t_S} = 1, \qquad \textbf{Speciation: } \mathrm{I} \to \mathrm{II}$$

Classifier-Free Guidance (CFG)

Additional push proportional to the following score difference:

$$S_t^{\mathrm{CFG}}(\vec{x}, c) = S_t(\vec{x}, c) + \omega[S_t(\vec{x}, c) - S_t(\vec{x})]$$

This causes the following:
Theoretically, in low-dim., the generated distribution does not equal the target one.
In practice, the generated images have decreased diversity/can collapse to the mode.

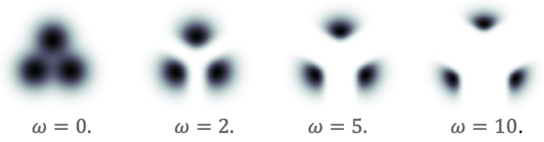


Figure 3a. Effect of CFG on a mixture of 3 Gaussians on varying ω.

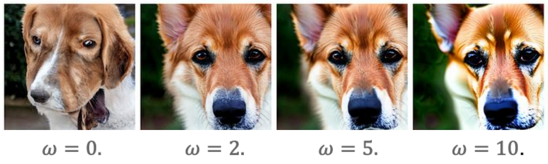


Figure 3b. Effect of CFG on real-world data with varying ω.

Research goal: Can CFG generate the correct distribution in large- or infinite- dimensional settings? Can we construct better guidance strategies?

Gaussian example taken from https://arxiv.org/abs/2207.12598.The image generated from Stable Diffusion v1.5 under different guidance scale. Image source: https://zhuanlan.zhihu.com/p/660518657

ENS ÉCOLE NORMALE SUPÉRIEURE Università Bocconi MILANO Meta

Classifier-Free Guidance (CFG)

Additional push proportional to the following score difference:

$$S_t^{\mathrm{CFG}}(\vec{x}, c) = S_t(\vec{x}, c) + \omega[S_t(\vec{x}, c) - S_t(\vec{x})]$$

Figure 3c. Comparison of unguided conditional diffusion and standard CFG

Key observation: effect only in low-dimensions

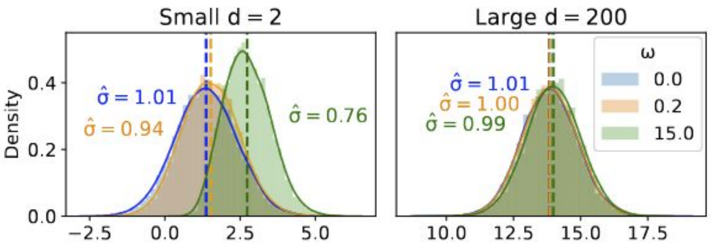
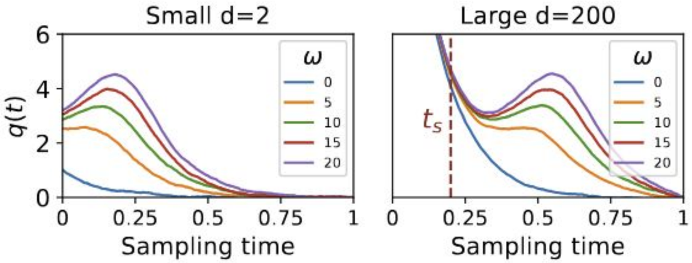


Figure 4a. Mean overshoot and variance shrinkage disappears in high-dim.



$$q(t=0) = \vec{x} \cdot \vec{m} / |\vec{m}|$$

Figure 4b. In high-dim., the paths realign with the unconditional one.

Effect of Classifier-Free guidance (CFG):

In one- and low-dimensions, CFG:

- Overshoots the mean of the target distribution
- Shrinks the variance of the target distribution

Analyzing CFG in infinite-dimensional:

- For suff. large and infinite dim., CFG generates the correct distribution
- CFG only affects the first part of the process
- This happens before the class has been decided

$$S_t^{\mathrm{CFG}}(\vec{x}, c) = S_t(\vec{x}, c) + \omega[S_t(\vec{x}, c) - S_t(\vec{x})]$$

CFG score formula.

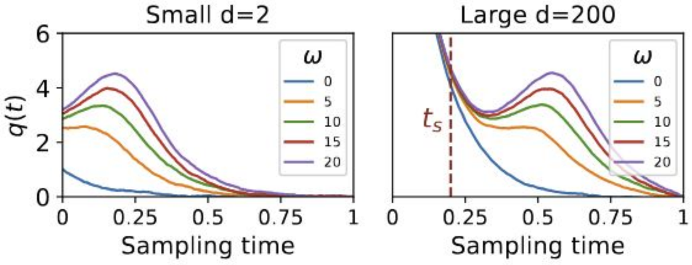
Main results

Intuitively, in Regime II:

$$\vec{S}_t(\vec{x}) = \vec{\nabla} \log P_t(\vec{x}) = \frac{\sum_{c=1}^{K} P_t(\vec{x}|c)p(c)\ \vec{S}_t(\vec{x},c)}{\sum_{c=1}^{K} P_t(\vec{x}|c)p(c)} \longrightarrow \sum_{c'=1}^{K} P_t(\vec{x}|c')p(c') \simeq P_t(\vec{x}|c)p(c) \longrightarrow \vec{S}_t(\vec{x},c) \simeq \vec{S}_t(\vec{x})$$

Our main results are as follows:

Result I. Before speciation time t_s, CFG is effective and speeds up convergence towards the target class.
Result II. Just before speciation time t_s, CFG-guided paths align with the unguided path.
Result III. After speciation time t_s, CFG has no effect on the generation process.



$q(t=0) = \vec{x}\cdot\vec{m}/|\vec{m}|$

Figure 4b. In high-dim., the paths realign with the unconditional one.

Generalized CFG

We can generalize CFG and propose non-linear guidances:

$$S_t^{\text{CFG-NL}}(\vec{x}, c) = S_t(\vec{x}, c) + \left[S_t(\vec{x}, c) - S_t(\vec{x})\right] \phi_t \left(\left|\vec{S}_t(\vec{x}, c) - \vec{S}_t(\vec{x})\right|\right)$$

As long as the $\lim_{s \to 0} \left[s\phi_t(s)\right]$ equals zero, our three results carry over. We can propose a simple **Power-Law guidance** scheme as follows:

$$\vec{S}_t^{\text{PL}}(\vec{x}, c) = S_t(\vec{x}, c) + \omega \left[S_t(\vec{x}, c) - S_t(\vec{x})\right] \left|\vec{S}_t(\vec{x}, c) - \vec{S}_t(\vec{x})\right|^{\alpha}$$

Theoretically, we can construct guidances with decreased overshoot and variance shrinkage. In practice, we can observe the same benefits.

Figure 6b. Comparison of non-guided, standard CFG and Power-Law CFG

Practical implications

$$\vec{S}_t^{\mathrm{PL}}(\vec{x}, c) = S_t(\vec{x}, c) + \omega \left[S_t(\vec{x}, c) - S_t(\vec{x}) \right] \left| \vec{S}_t(\vec{x}, c) - \vec{S}_t(\vec{x}) \right|^{\alpha}$$

Power-law CFG formula.

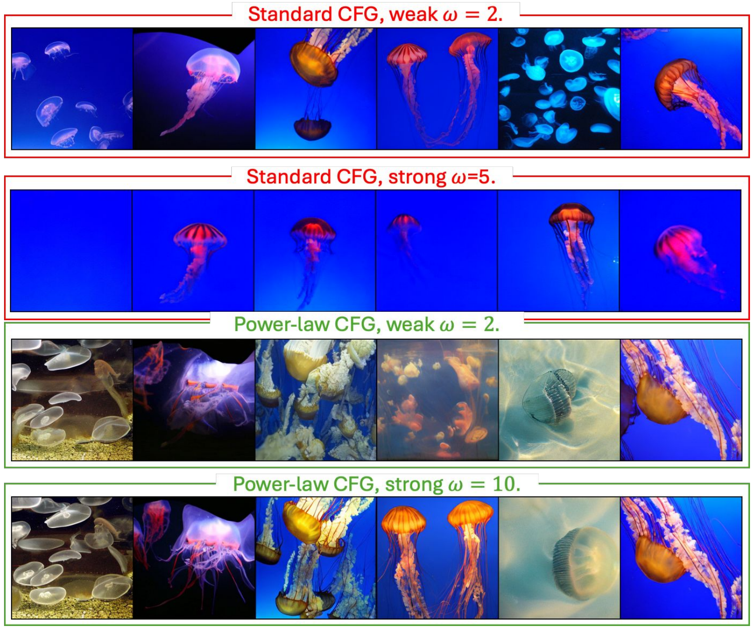


Figure 7a. Qualitative comparison of standard and Power-Law CFG for varying ω

Figure 7b. Qualitative comparison of Power-Law CFG for varying α

Sensitivity analysis and benchmarks

$$\vec{S}_t^{\mathrm{PL}}(\vec{x},c) = S_t(\vec{x},c) + \omega\left[S_t(\vec{x},c) - S_t(\vec{x})\right]\left|\vec{S}_t(\vec{x},c) - \vec{S}_t(\vec{x})\right|^{\alpha}$$

Power-law CFG formula.

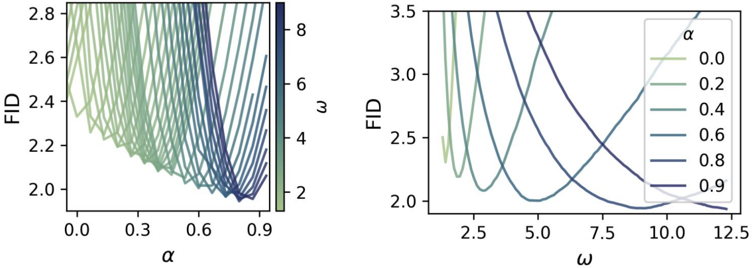


Figure 8. Sensitivity analysis of Power-Law CFG

	EDM2-S (CC, IM-1K 512)			DiT/XL-2 (CC, IM-1K 256)		
Model	FID	Precision	Recall	FID	Precision	Recall
Standard (Ho and Salimans, 2022)	2.29	0.751	0.582	2.27	0.829	0.584
Scheduler (Wang et al., 2024)	2.03	0.762	0.591	2.14	0.840	0.614
Limited (Kynkäänniemi et al., 2024)	1.87	0.760	0.598	1.97	0.801	0.632
Cosine (Gao et al., 2023)	2.15	0.770	0.619	2.30	**0.861**	0.520
CADS (Sadat et al., 2023)	<u>1.60</u>	**0.792**	0.619	<u>1.70</u>	0.772	0.627
APG (Sadat et al., 2024)	2.13	0.756	**0.640**	2.11	0.815	0.628
REG (Xia et al., 2024)	1.99	0.761	0.608	1.76	0.799	0.601
CFG++ (Chung et al., 2024)	N/A	N/A	N/A	N/A	N/A	N/A
Power-law CFG (Ours)	1.93 (↓)	<u>0.780</u> (↑)	<u>0.631</u> (↑)	2.05 (↓)	0.831 (↑)	0.595 (↑)
Power-law CFG + Limited (Ours)	1.73 (↓)	0.752 (↓)	0.600 (↑)	1.87 (↓)	<u>0.849</u> (↑)	**0.642** (↑)
Power-law CFG + CADS (Ours)	**1.52** (↓)	0.770 (↓)	0.622 (↑)	**1.63** (↓)	0.754 (↓)	<u>0.639</u> (↑)

Table 1. Adding power-law to other guidances improves their performance

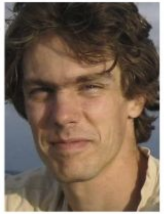
Benefits of non-linear CFG:

- Consistent improvement in FID for larger α
- Easier hyperparameter tuning for ω

ENS ÉCOLE NORMALE SUPÉRIEURE · Università Bocconi MILANO · Meta

Thank you!

Kruno Lehman
ENS & Meta



Jakob Verbeek
Meta

Giulio Biroli
ENS

Marc Mezard
Bocconi University

Paper →

References

[1] Biroli, G., Bonnaire, T., De Bortoli, V. and Mézard, M., 2024. Dynamical regimes of diffusion models.
[2] Peebles, W.S. and Xie, S., 2022. Scalable diffusion models with transformers. 2023
[3] Karras, T., Aittala, M., Lehtinen, J., Hellsten, J., Aila, T. and Laine, S., 2024. Analyzing and improving the training dynamics of diffusion models.