Forgetting and stability of Score-based Generative Models.

Joint work with Gabriel Victorino Cardoso, Sylvain Le Corff,
Vincent Lemaire and Antonio Ocello

arXiv:2601.21868 ▪ arXiv:2607.04780

Stanislas Strasman

Sorbonne University, Laboratoire de Probabilités, Statistique et Modélisation (LPSM)

Generative AI

Data

We observe a **finite** dataset

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \pi_{\text{data}}\,, \qquad \pi_{\text{data}} \in \mathcal{P}(\mathbb{R}^d) \quad \text{unknown}\,.$$

Unsupervised Machine Learning

Learn a sampling mechanism whose output law $\widehat{\pi}$ satisfies

$$\widehat{\pi} \approx \pi_{\text{data}}\,.$$

⚠ both a learning and a sampling problem.

What does $\approx$ mean?

The notion of approximation depends on the application: visual quality, downstream performance, or proper mathematical distance on the space of probability measures.

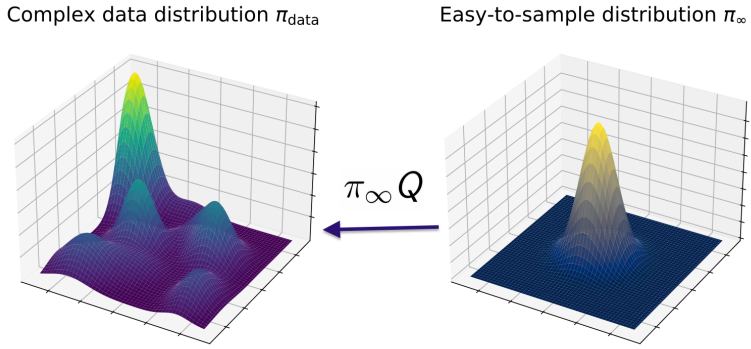
Generative Modeling design and stability

Simulable generation model

Find an easy-to-sample prior π_∞ and a simulable transition kernel Q such that

$$\pi_\infty Q \approx \pi_{\text{data}} .$$

What can be said about the properties of Q ?



Principles of SGMs and stability results

Reverse diffusion as a Markov sampler

We present the Brownian case: $\mathrm{d}\overrightarrow{X}_t = \sqrt{2}\,\mathrm{d}B_t, \qquad \overrightarrow{X}_0 \sim \pi_{\mathrm{data}}.$

Let p_t denote the marginal density of $\overrightarrow{X}_t$.

Time reversal (Haussmann and Pardoux, 1986)

For fixed $T > 0$,

$$(\overleftarrow{X}_t)_{t\in[0,T]} \overset{\mathcal{L}}{=} (\overrightarrow{X}_{T-t})_{t\in[0,T]},$$

and the reverse process satisfies

$$\mathrm{d}\overleftarrow{X}_t = 2\nabla \log p_{T-t}(\overleftarrow{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad \overleftarrow{X}_0 \sim p_T.$$

Reverse transition kernels

For $0 \le s < t \le T$, f bounded measurable

$$Q_{t|s} f(x) := \mathbb{E}\left[f(\overleftarrow{X}_t) \,\middle|\, \overleftarrow{X}_s = x \right].$$

The ideal sampler satisfies

$$\boxed{p_T Q_{T|0} = \pi_{\mathrm{data}}.}$$

SGMs in practice: three approximations

$\boxed{p_T Q_{T|0} = \pi_{\text{data}}}$ **is almost generative, but . . .**

1. Initialization

The exact terminal law still depends on the data:

$$\overrightarrow{X}_T \overset{\mathcal{L}}{=} \overrightarrow{X}_0 + \sqrt{2T}\, Z\,.$$

Equivalently,

$$p_T = p_0 * \mathcal{N}(0, 2T\mathrm{I}_d)\,.$$

For large T,

$$p_T \approx \pi_\infty = N(0, 2T\mathrm{I}_d)\,.$$

Thus, $p_T Q_{T|0} \approx \pi_\infty Q_{T|0}$.

2. Score approximation

The reverse drift uses the unknown score

$$\nabla \log p_{T-t}(x)\,.$$

In practice, learn

$$\mathsf{s}_\theta(x, T-t) \approx \nabla \log p_{T-t}(x)\,.$$

Hence

$$\pi_{\text{data}} \approx \pi_\infty Q^\theta_{T|0}\,.$$

3. Discretization

Choose a grid

$$t_k = kh, \qquad h = \frac{T}{N}\,.$$

to discretize

$$\mathrm{d}X^\theta_t = 2\mathsf{s}_\theta(X^\theta_t, T-t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$$

as

$$\widehat{\pi}^\theta_N = \pi_\infty \bar{Q}^\theta_0 \cdots \bar{Q}^\theta_{N-1}\,.$$

$$\boxed{\bar{X}^{\theta,N}_{k+1} = \bar{X}^{\theta,N}_k + 2h\,\mathsf{s}_\theta(\bar{X}^{\theta,N}_k, T-t_k) + \sqrt{2h}\,Z_{k+1}\,, \quad \bar{X}^{\theta,N}_0 \sim \pi_\infty\,, \quad Z_{k+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathrm{I}_d)\,.}$$

A classical route: path-space control by Girsanov

Let $\mathbb{P}_{[0,T]}$ denote the path law of the exact reverse process and $\mathbb{P}^{\theta}_{[0,T]}$ that of the practical continuous interpolation:

$$\begin{aligned} \mathrm{d}\overleftarrow{X}_t &= 2\nabla \log p_{T-t}(\overleftarrow{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad \overleftarrow{X}_0 \sim p_T, \\ \mathrm{d}X^{\theta}_t &= 2\mathsf{s}_{\theta}(X^{\theta}_{\underline{t}}, T-\underline{t})\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad X^{\theta}_0 \sim \pi_{\infty}, \end{aligned}$$

where $\underline{t} = t_k$ for $t \in [t_k, t_{k+1})$. Define the local score error

$$E_k(x) := \nabla \log p_{T-t_k}(x) - \mathsf{s}_{\theta}(x, T-t_k).$$

$$\begin{aligned} \mathsf{KL}\left(\pi_{\mathrm{data}} \,\middle\|\, \widehat{\pi}^{\theta}_N\right) &\le \mathsf{KL}(\mathbb{P}_{[0,T]} \| \mathbb{P}^{\theta}_{[0,T]}) \qquad \text{(data processing ineq.)} \\ &\le \mathsf{KL}(p_T \| \pi_{\infty}) + \int_0^T \mathbb{E}_{\mathbb{P}}\left[\left\|\nabla \log p_{T-t}(\overleftarrow{X}_t) - \mathsf{s}_{\theta}(\overleftarrow{X}_{\underline{t}}, T-\underline{t})\right\|^2\right] \mathrm{d}t \qquad \text{(Girsanov)} \\ &\lesssim \mathsf{KL}(p_T \| \pi_{\infty}) + \sum_{k=0}^{N-1} \left(\int_{t_k}^{t_{k+1}} \mathbb{E}\left[\left\|\nabla \log p_{T-t}(\overleftarrow{X}_t) - \nabla \log p_{T-t_k}(\overleftarrow{X}_{t_k})\right\|^2\right] \mathrm{d}t + h\|E_k\|^2_{L^2(p_{T-t_k})} \right) \end{aligned}$$

Schematically

$$\mathsf{KL}(\pi_{\mathrm{data}} \| \widehat{\pi}^{\theta}_N) \lesssim \mathcal{E}_{\mathrm{init}} + h \sum_k \left(\mathcal{E}^{\mathrm{disc}}_k + \|E_k\|^2_{L^2(p_{T-t_k})} \right).$$

The path-space argument controls the *magnitude* of accumulated local errors. It does not display how much of each local error at time t_k is damped by the stochastic sampler.

Convergence theory develops along several complementary axes

1. Quantitative complexity

Improve dependence on

$$d, \qquad T, \qquad N,$$

and on the target sampling accuracy.

Examples include: sharper discretization complexity; nearly linear dependence on d in KL, minimax statistical rates.

Lee–Lu–Tan, 2022 ; Chen–Lee–Lu, 2023 ; Benton et al., 2024 ; Stéphanovitch–Aamari–Levrard, 2025

2. Broader target classes

Relax assumptions such as global log-concavity, bounded support, or uniform score regularity.

Examples include: finite moments or Fisher information; sufficiently decaying tails; weak log-concavity / semiconvexity; weaker score regularity, early stopping.

Lee–Lu–Tan, 2023 ; Conforti–Durmus–Gentiloni Silveri, 2025 ; Gentiloni Silveri–Ocello, 2025 ; Strasman et al., 2025.

3. Probability metrics

Different metrics suggest different proof techniques:

- KL : path-space / Girsanov
- TV : Pinsker / χ^2
- $\mathcal{W}$: curvature / coupling methods.

Chen–Lee–Lu, 2023 ; Conforti–Durmus–Gentiloni Silveri, 2025 ; Gao–Nguyen–Zhu, 2025

Our question

How does the effect of a local error evolve along the remaining reverse flow?

Dynamical stability bound with forgetting

Forgetting is a property of the exact reverse kernels

For $0 \leq s < t \leq T$,

$$Q_{t|s} f(x) = \mathbb{E}\left[f(\overleftarrow{X}_t) \,\middle|\, \overleftarrow{X}_s = x \right].$$

Forgetting of initial conditions

We say that $Q_{t|s}$ forgets in a probability distance ρ if, for all admissible probability measures μ, ν,

$$\boxed{\rho\big(\mu Q_{t|s}, \nu Q_{t|s}\big) \leq \alpha_{s,t} \rho(\mu, \nu), \qquad \alpha_{s,t} < 1.}$$

Thus $\alpha_{s,t}$ quantifies how much of a discrepancy present at time s can "survive" until time t.

Why this matters

- **Stability and error propagation:** an error introduced at t_k is discounted by $\alpha_{t_k,T}$, rather than accumulated without attenuation.
- **Dynamics:** which features, modes, or classes are erased—and which remain distinguishable?

Generated distribution and target error

Fix a discretization grid: $\boxed{0 = t_0 < t_1 < \cdots < t_N = T}$

Exact reverse kernels vs. practical SGM kernels

For a bounded test function f, define

$$Q_k f(x) := \mathbb{E}\Big[f(\overleftarrow{X}_{t_{k+1}}) \,\Big|\, \overleftarrow{X}_{t_k} = x \Big], \qquad k = 0, \dots, N-1 .$$

and

$$\bar{Q}_k^\theta f(x) := \mathbb{E}\Big[f\Big(\bar{X}_{k+1}^{\theta,N}\Big) \Big| \bar{X}_k^{\theta,N} = x \Big], \qquad k = 0, \dots, N-1 .$$

Thus the generated law is

$$\widehat{\pi}_{\infty,N}^\theta = \pi_\infty \bar{Q}_0^\theta \cdots \bar{Q}_{N-1}^\theta .$$

Stability objective

For a discrepancy ρ between probability measures, the SGM error is

$$\rho\big(\pi_{\mathrm{data}}, \widehat{\pi}_{\infty,N}^\theta\big) = \rho\big(p_T Q_0 \cdots Q_{N-1}, \pi_\infty \bar{Q}_0^\theta \cdots \bar{Q}_{N-1}^\theta\big) .$$

Forgetting and stability analysis

Suppose the reverse kernel contracts: $\rho(\mu Q_k, \nu Q_k) \leq \alpha_k \rho(\mu, \nu), \qquad \alpha_k < 1.$

Define at step t_k (resp. k), the marginal of the reverse process

$$\mu_k = p_T Q_0 \cdots Q_{k-1}\,, \qquad \mu_0 = p_T$$

and the practical approximation of μ_k (i.e. the SGM sampler)

$$\widehat{\mu}_k = \pi_\infty \bar{Q}_0^\theta \cdots \bar{Q}_{k-1}^\theta\,, \qquad \widehat{\mu}_0 = \pi_\infty$$

The *accumulated error at time k*, writes as

$$\begin{aligned}\rho(\mu_{k+1}, \widehat{\mu}_{k+1}) = \rho(\mu_k Q_k, \widehat{\mu}_k \bar{Q}_k^\theta) &\leq \rho(\mu_k Q_k, \widehat{\mu}_k Q_k) + \rho(\widehat{\mu}_k Q_k, \widehat{\mu}_k \bar{Q}_k^\theta) \\ &\leq \alpha_k \underbrace{\rho(\mu_k, \widehat{\mu}_k)}_{\text{Old error}} + \underbrace{\rho\big(\widehat{\mu}_k Q_k, \widehat{\mu}_k \bar{Q}_k^\theta\big)}_{\text{Local error} := \delta_k}\,.\end{aligned}$$

Forgetting mechanism

$$\rho\big(\pi_{\mathrm{data}}, \widehat{\pi}_{\infty,N}^\theta\big) = \rho(\mu_N, \widehat{\mu}_N) \leq \left(\prod_{\ell=0}^{N-1} \alpha_\ell\right) \rho(\mu_0, \widehat{\mu}_0) + \sum_{k=0}^{N-1} \left(\prod_{\ell=k+1}^{N-1} \alpha_\ell\right) \delta_k\,.$$

Forgetting implies a discount of every local error

Suppose we have proven forgetting on $Q_{s,t}$ so that

$$\rho(\mu Q_k, \nu Q_k) \le \alpha_k \rho(\mu, \nu), \qquad \alpha_k \le \bar{\alpha} < 1$$

and we get a local error control, under the regularity and moment assumptions,

$$\delta_k \le \underbrace{h C_k^{\mathrm{disc}}}_{\text{discretization}} + \underbrace{\sqrt{h}\, C_k^{\mathrm{net}} \|E_k\|_{L^2(\widehat{\mu}_k)}}_{\text{score approximation}}.$$

then we get a bound of the form,

$$\rho\left(\pi_{\mathrm{data}}, \widehat{\pi}_N^\theta\right) \le \bar{\alpha}^N \rho(p_T, \pi_\infty) + \sum_{k=0}^{N-1} \bar{\alpha}^{N-1-k} \left(h\, C_k^{\mathrm{disc}} + \sqrt{h}\, C_k^{\mathrm{net}} \|E_k\|_{L^2(\widehat{\mu}_k)} \right).$$

$$\rho\left(\pi_{\mathrm{data}}, \widehat{\pi}_N^\theta\right) \lesssim \rho(p_T, \pi_\infty) + \sum_k \delta_k$$

no reverse-flow discount

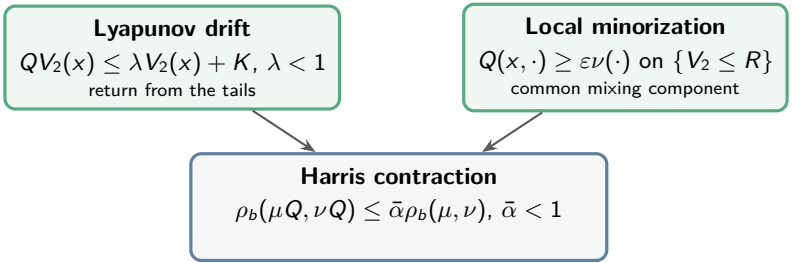
$$\rho\left(\pi_{\mathrm{data}}, \widehat{\pi}_N^\theta\right) \lesssim \bar{\alpha}^N \rho(p_T, \pi_\infty) + \sum_k \bar{\alpha}^{N-1-k} \delta_k$$

older errors are geometrically damped

Proving forgetting in SGMs

Harris theory: Hairer–Mattingly (2011)

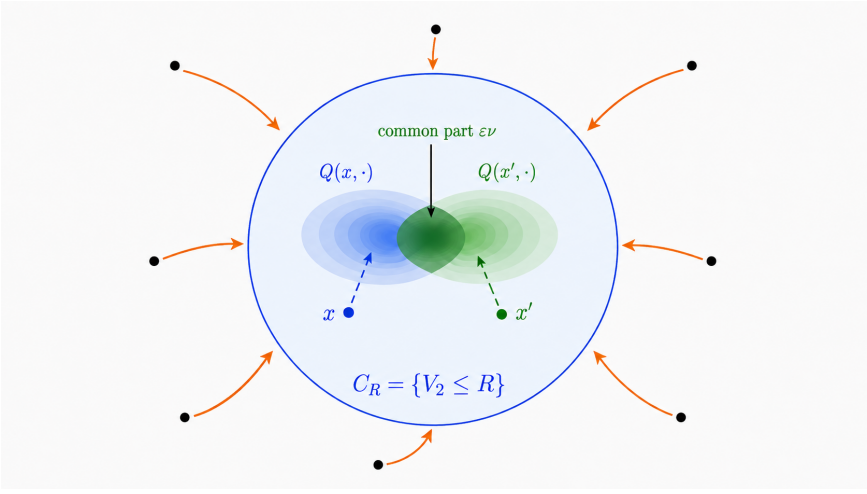
For $V_2(x) = \|x\|^2$, $b > 0$ consider $\rho_b(\mu, \nu) := \int_{\mathbb{R}^d} \big(1 + bV_2(x)\big)\, |\mu - \nu|(\mathrm{d}x)$.



Why local rather than global Doeblin?

On $\mathbb{R}^d$, having every region in the support of the common measure ν to have a uniformly non-vanishing chance of being visited, independently of where the chain starts x is usually infeasible.

Geometric intuition: return to the center, then mix



Relationship with TV and Wasserstein

Let $V_2(x) = \|x\|^2$ and $\rho_b(\mu,\nu) := \int_{\mathbb{R}^d} (1 + b\|x\|^2)|\mu - \nu|(\mathrm{d}x)$.

Control of TV and $\mathcal{W}_2$

For probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\boxed{\|\mu - \nu\|_{\mathrm{TV}} \leq \frac{1}{2}\rho_b(\mu,\nu)\,, \qquad \mathcal{W}_2^2(\mu,\nu) \leq \frac{2}{b}\rho_b(\mu,\nu)\,.}$$

Proof

Since $1 + b\|x\|^2 \geq 1$,

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2}|\mu - \nu|(\mathbb{R}^d) \leq \frac{1}{2}\rho_b(\mu,\nu)\,.$$

Moreover, using (Theorem 6.15 in Villani, 2009),

$$\mathcal{W}_2^2(\mu,\nu) \leq 2\int_{\mathbb{R}^d} \|x\|^2 |\mu - \nu|(\mathrm{d}x) \leq \frac{2}{b}\rho_b(\mu,\nu)\,.$$

Assumptions on the data score

Let $p_0 \in C^2(\mathbb{R}^d)$ and we assume that, for some $\gamma_0 > 0$, $\kappa_0 \geq 0$, $C_0 > 0$, and $m \geq 1$,

$$\boxed{\langle \nabla \log p_0(x), x \rangle \leq -\gamma_0 \|x\|^2 + \kappa_0} \quad \text{and} \quad \boxed{\|\nabla^2 \log p_0(x)\|_{\mathrm{F}} \leq C_0\big(1 + \|x\|^m\big).}$$

Dissipativity

Far from the origin, the score has an inward radial component. This condition yields sub-Gaussian tail control

Polynomial regularity

The Jacobian of the score may grow polynomially: we do *not* require a globally Lipschitz score or global log-concavity.

Scope

The framework accommodates nonconvex and multimodal targets, although it still imposes a confining tail condition.

Proof sketch: verifying the Harris conditions

1. Lyapunov drift

Dissipativity propagates through Gaussian smoothing:

$$\langle \nabla \log p_t(x), x \rangle \le -\gamma_t \|x\|^2 + \kappa_t.$$

For $V_2(x) = \|x\|^2$, the reverse generator satisfies

$$\mathcal{A}_u V_2(x) = 4 \langle \nabla \log p_{T-u}(x), x \rangle + 2d \le -\widetilde{\gamma}_u V_2(x) + \widetilde{\kappa}_u.$$

Dynkin's formula and Grönwall give

$$\boxed{Q_{t|s} V_2(x) \le \lambda_{t|s} V_2(x) + K_{t|s}, \qquad \lambda_{t|s} < 1.}$$

2. Local minorization

Let $\varphi_a(z)$ be the density of $\mathcal{N}(0, a\mathrm{I}_d)$. By Bayes' formula,

$$Q_{t|s}(x, \mathrm{d}y) = \frac{\varphi_{2(t-s)}(x-y) p_{T-t}(y)}{p_{T-s}(x)} \, \mathrm{d}y.$$

Since $x \in C_r = \{x : \|x\| \le r\}$,

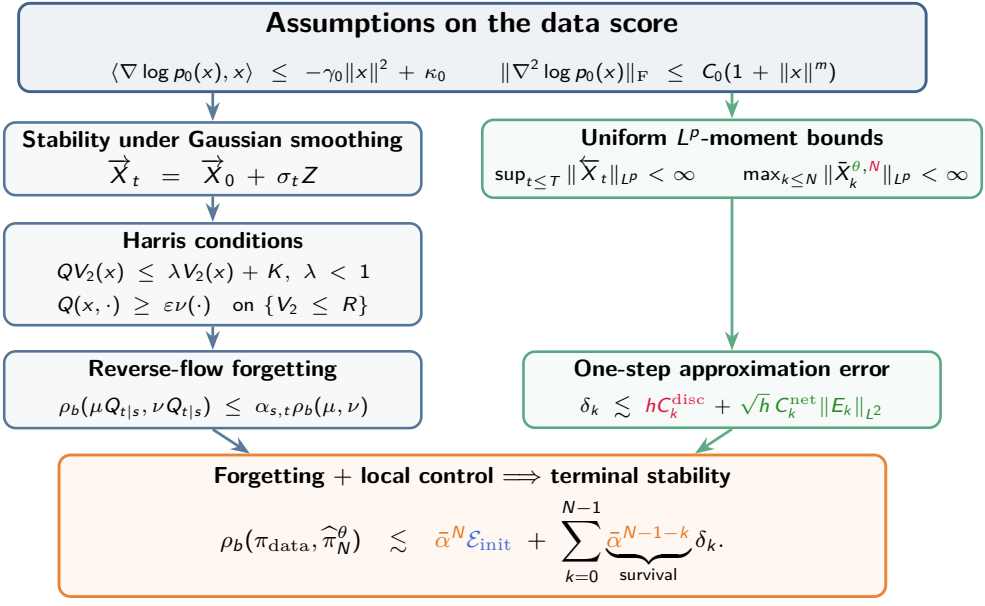
$$\varphi_{2(t-s)}(x-y) \ge c_r \, \varphi_{t-s}(y), \qquad x \in C_r,$$

for some $c_r > 0$. and $p_{T-s}(x)$ is uniformly bounded above. Hence

$$\boxed{Q_{t|s}(x, A) \ge \varepsilon_{t|s}^{(r)} \nu_{t|s}(A), \qquad x \in C_r.}$$

Conclusion

Each exact reverse transition satisfies quantitative Harris contraction. On a fixed finite discretization grid, the constants can be uniformized: there exist a common $b > 0$ and $\bar{\alpha} < 1$ usable throughout the grid.


Assumptions on the data score
$\langle \nabla \log p_0(x), x\rangle \le -\gamma_0\|x\|^2 + \kappa_0 \qquad \|\nabla^2 \log p_0(x)\|_{\mathrm{F}} \le C_0(1 + \|x\|^m)$
Stability under Gaussian smoothing
$\overrightarrow{X}_t = \overrightarrow{X}_0 + \sigma_t Z$
Harris conditions
$QV_2(x) \le \lambda V_2(x) + K,\ \lambda < 1$
$Q(x,\cdot) \ge \varepsilon\nu(\cdot)$ on $\{V_2 \le R\}$
Reverse-flow forgetting
$\rho_b(\mu Q_{t|s}, \nu Q_{t|s}) \le \alpha_{s,t}\rho_b(\mu,\nu)$
Uniform L^p-moment bounds
$\sup_{t\le T}\|\overleftarrow{X}_t\|_{L^p} < \infty \qquad \max_{k\le N}\|\bar{X}_k^{\theta,N}\|_{L^p} < \infty$
One-step approximation error
$\delta_k \lesssim hC_k^{\mathrm{disc}} + \sqrt{h}\, C_k^{\mathrm{net}}\|E_k\|_{L^2}$
Forgetting + local control $\Longrightarrow$ terminal stability
$\rho_b(\pi_{\mathrm{data}}, \widehat{\pi}_N^\theta) \lesssim \bar{\alpha}^N\mathcal{E}_{\mathrm{init}} + \sum_{k=0}^{N-1}\underbrace{\bar{\alpha}^{N-1-k}}_{\text{survival}}\delta_k.$

Numerical illustration

Numerical illustration: what is tested?

Goal

The experiments are designed to isolate the forgetting mechanism:

$$\text{perturb early in the reverse trajectory} \implies \text{small final effect,}$$

whereas

$$\text{perturb late in the reverse trajectory} \implies \text{large final effect.}$$

Two perturbation protocols

1. **Initialization perturbation:** perturb the cloud at an intermediate noise level, then run the reverse sampler to the data time.
2. **Local score perturbation:** keep the initialization fixed, but perturb the score at one single reverse step.

⚠ When the score is known and there is no discretization error our theory gives a precise estimate of such phenomenon for the *weigthed TV distance*.

Two controlled perturbation experiments

Let t_{pert} denote the forward noise level where the perturbation is introduced.

1. Initialization perturbation

1. Sample particles from the forward marginal $p_{t_{\mathrm{pert}}}$.
2. Shift the cloud in a fixed direction:
$$\overrightarrow{X}^{\mathrm{pert}}_{t_{\mathrm{pert}}} = \overrightarrow{X}_{t_{\mathrm{pert}}} + \lambda u\,.$$
3. Run the reverse sampler back to time 0.
4. Compare the final law with π_{data}.

2. Local score perturbation

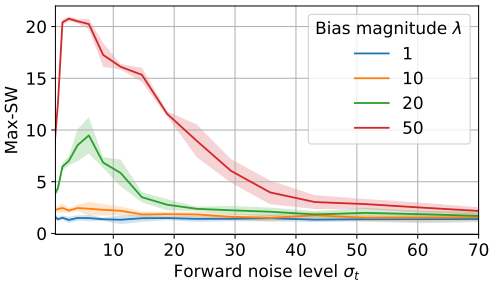
1. Start from the same terminal initialization.
2. Run the exact-score reverse sampler.
3. At one step k_{pert}, replace
$$\nabla \log p_{T-t_k} \quad \text{by} \quad \nabla \log p_{T-t_k} + \lambda u\,.$$
4. Complete the reverse sampler and measure the final error.

Expected signature of forgetting

The final error should be smaller when t_{pert} is far from the data time, because more reverse dynamics remains after the perturbation.

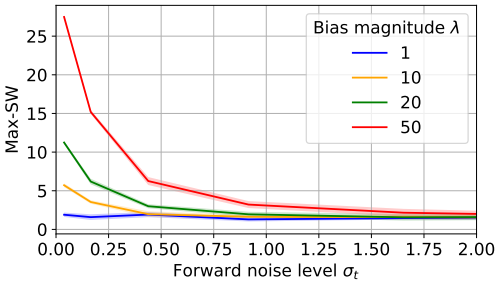
GMM: perturbation sensitivity

Initialization perturbation



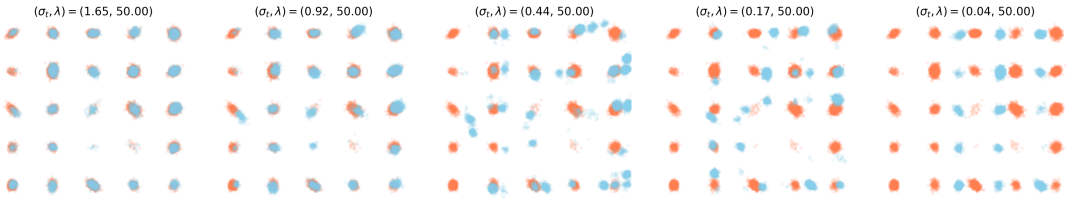
Perturb the cloud at a chosen noise level, then reverse.

One-step score perturbation



Perturb the score at one reverse step, then continue.

GMM: visualizing one-step score perturbations



Visual message

As the perturbation is made closer to the data time, the final samples deviate more visibly from the target mixture. Late score errors have little time left to be forgotten.

CIFAR-10: one-step score perturbation

Real-data perturbation experiment

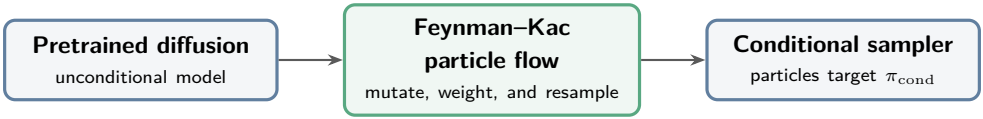
Using a pretrained EDM VP model on CIFAR-10, perturb the denoiser once.

Step	0	25	50	70	75	80	85	90	95
FID	13.3	13.0	13.1	13.6	14.4	16.4	28.3	153	304
Max-SW	0.011	0.014	0.016	0.033	0.044	0.060	0.094	0.266	0.779

Beyond unconditional generation: conditional sampling with SMC

A conditioning likelihood changes the target while the pretrained diffusion prior is reused:

$$\pi_{\mathrm{cond}}(\mathrm{d}x) \propto p(y \mid x)\, \pi_{\mathrm{data}}(\mathrm{d}x).$$



Forward-smoothing forgetting discounts both sources of error

If $\rho_{\mathrm{FK}} \in (0,1)$, an error introduced at step ℓ is discounted by $\rho_{\mathrm{FK}}^{N-\ell}$:

$$\mathcal{E}_N^{\mathrm{total}} \lesssim \underbrace{\sum_{\ell=0}^{N-1} \rho_{\mathrm{FK}}^{N-\ell}\, \mathcal{E}_\ell^{\mathrm{MC}}}_{\text{finite-particle error}} + \underbrace{\sum_{\ell=0}^{N-1} \rho_{\mathrm{FK}}^{N-\ell}\, \mathcal{E}_\ell^{\mathrm{bias}}}_{\text{initialization, discretization, score approximation}}.$$

Strasman et al. (2026), *Non-Asymptotic Error Bounds for SMC with Biased Proposals*, arXiv:2607.04780.

References

U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4):1188–1205, 1986.

M. Hairer and J. C. Mattingly. Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, Progress in Probability 63, pages 109–117. Birkhäuser/Springer, 2011.

C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften 338. Springer, 2009.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021.

S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *International Conference on Learning Representations*, 2023.

H. Lee, J. Lu, and Y. Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

H. Lee, J. Lu, and Y. Tan. Convergence of score-based generative modeling for general data distributions. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory*, PMLR 201:946–985, 2023.

H. Chen, H. Lee, and J. Lu. Improved analysis of score-based generative modeling: user-friendly bounds under minimal smoothness assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202:4735–4763, 2023.

References ii

J. Benton, V. De Bortoli, A. Doucet, and G. Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. *International Conference on Learning Representations*, 2024.

G. Conforti, A. Durmus, and M. Gentiloni Silveri. KL convergence guarantees for score diffusion models under minimal data assumptions. *SIAM Journal on Mathematics of Data Science*, 7(1):86–109, 2025.

M. Gentiloni Silveri and A. Ocello. Beyond log-concavity and score regularity: improved convergence bounds for score-based generative models in $\mathcal{W}_2$-distance. In *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267:55631–55656, 2025.

X. Gao, H. M. Nguyen, and L. Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of Machine Learning Research*, 26(43):1–54, 2025.

A. Stéphanovitch, E. Aamari, and C. Levrard. Generalization bounds for score-based generative models: a synthetic proof. arXiv:2507.04794, 2025.

S. Strasman, S. Surendran, C. Boyer, S. Le Corff, V. Lemaire, and A. Ocello. Wasserstein convergence of critically damped Langevin diffusions. *Advances in Neural Information Processing Systems 38*, 2025.

T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

S. Strasman, G. Victorino Cardoso, S. Le Corff, V. Lemaire, and A. Ocello. On forgetting and stability of score-based generative models. arXiv:2601.21868, 2026.

S. Strasman, G. Victorino Cardoso, S. Le Corff, V. Lemaire, and A. Ocello. Non-asymptotic error bounds for SMC with biased proposals: application to conditional diffusion sampling. arXiv:2607.04780, 2026.

How this relates to other stability analyses

Viewpoint	Main mechanism	Typical output	What it emphasizes
Path-space KL / Girsanov	Compare exact and learned path laws	Terminal KL bounded by an integral of local drift errors	Powerful local control; the standard global form does not display a downstream discount
Curvature / synchronous coupling	Monotonicity or log-concavity of the score	Wasserstein contraction on contractive time intervals	Geometric contraction, often under curvature-type assumptions
Harris viewpoint here	Tail return + local overlap	Weighted-TV contraction and a discounted telescoping bound	Forgetting without global convexity; explicit time-of-error dependence

Important nuance

The approaches are complementary. In fact, the one-step approximation estimate in our proof uses a *local* Girsanov argument; Harris contraction is what discounts that local error through the remaining flow.

Proof sketch I: dissipativity yields a Lyapunov drift

Step 1: propagate the data assumption

Assume that the data score is dissipative: $\langle \nabla \log p_0(x), x \rangle \le -\gamma_0 \|x\|^2 + \kappa_0$. Gaussian smoothing preserves this structure:

$$\boxed{\langle \nabla \log p_t(x), x \rangle \le -\gamma_t \|x\|^2 + \kappa_t, \qquad t \in [0, T],}$$

where $\gamma_t > 0$ and $\kappa_t \ge 0$ are explicit functions of γ_0, κ_0, d, and t.

Step 2: apply the reverse generator to $V_2 = \|\cdot\|^2$

At time u, the generator is $\mathcal{A}_u f(x) = 2\langle \nabla \log p_{T-u}(x), \nabla f(x) \rangle + \Delta f(x)$.

Since $\nabla V_2(x) = 2x$ and $\Delta V_2(x) = 2d$, we obtain

$$\mathcal{A}_u V_2(x) = 4\langle \nabla \log p_{T-u}(x), x \rangle + 2d \le -4\gamma_{T-u} V_2(x) + 4\kappa_{T-u} + 2d = -\widetilde{\gamma}_u V_2(x) + \widetilde{\kappa}_u.$$

Dynkin's formula and Grönwall's lemma yield

$$\boxed{Q_{t|s} V_2(x) \le \lambda_{t|s} V_2(x) + K_{t|s}, \qquad \lambda_{t|s} < 1,}$$

where

$$\lambda_{t|s} = \exp\left(-\int_s^t \widetilde{\gamma}_u \, \mathrm{d}u \right), \qquad K_{t|s} = \int_s^t \exp\left(-\int_u^t \widetilde{\gamma}_v \, \mathrm{d}v \right) \widetilde{\kappa}_u \, \mathrm{d}u.$$

Proof sketch II: local minorization

Explicit density of the reverse kernel

Bayes' formula gives

$$\boxed{Q_{t|s}(x, \mathrm{d}y) = \frac{\varphi_{2(t-s)}(x-y)\, p_{T-t}(y)}{p_{T-s}(x)}\, \mathrm{d}y,}$$

where φ_a denotes the density of $\mathcal{N}(0, a\mathrm{I}_d)$, $a > 0$.

A common lower envelope on C_R

For every $x \in C_R$ and $y \in \mathbb{R}^d$,

$$\varphi_{2h}(x-y) \geq \underbrace{2^{-d/2} \exp\left(-\frac{R}{2(t-s)}\right)}_{c_R > 0} \varphi_h(y).$$

Moreover, $M_{T-s}^{(R)} := \sup_{x \in C_R} p_{T-s}(x) < \infty$.

Hence, for $x \in C_R$,

$$Q_{t|s}(x, \mathrm{d}y) \geq \frac{c_R}{M_{T-s}^{(R)}} \varphi_h(y) p_{T-t}(y)\, \mathrm{d}y.$$

The dependence in x is removed...

A strong route to forgetting: global Doeblin I

Global Doeblin condition

Let Q be a Markov kernel on E. Assume that there exist $\varepsilon \in (0,1]$ and a probability measure $\nu_\star$ such that

$$Q(x,\cdot) \geq \varepsilon \nu_\star(\cdot)\,, \qquad \forall x \in E\,.$$

Mixture decomposition

If $\varepsilon < 1$, define

$$R(x,\cdot) := \frac{Q(x,\cdot) - \varepsilon \nu_\star(\cdot)}{1-\varepsilon}\,.$$

By minorization, the denominator is a positive measure and $R(x,E) = 1$. Then R is a Markov kernel and

$$Q(x,\cdot) = \varepsilon \nu_\star(\cdot) + (1-\varepsilon) R(x,\cdot)\,.$$

Interpretation

With probability at least ε, the chain is refreshed from the same law $\nu_\star$, independently of its starting point.

A strong route to forgetting: global Doeblin II

For any two probability measures $\mu, \tilde{\mu}$, the decomposition gives

$$\mu Q = \varepsilon \nu_\star + (1-\varepsilon)\mu R, \qquad \tilde{\mu} Q = \varepsilon \nu_\star + (1-\varepsilon)\tilde{\mu} R\,.$$

Hence the common part cancels:

$$\mu Q - \tilde{\mu} Q = (1-\varepsilon)(\mu R - \tilde{\mu} R)\,.$$

Therefore

$$\|\mu Q - \tilde{\mu} Q\|_{\mathrm{TV}} = (1-\varepsilon)\|\mu R - \tilde{\mu} R\|_{\mathrm{TV}} \le (1-\varepsilon)\|\mu - \tilde{\mu}\|_{\mathrm{TV}}\,.$$

Iterating the contraction, gives geometric discounting

$$\boxed{\|\mu Q^n - \tilde{\mu} Q^n\|_{\mathrm{TV}} \le (1-\varepsilon)^n \|\mu - \tilde{\mu}\|_{\mathrm{TV}}\,.}$$

Harris assumptions for one Markov kernel

Global Doeblin too strong on $\mathbb{R}^d$

A global Doeblin condition asks for $Q(x,\cdot) \geq \varepsilon\nu_\star(\cdot)$, $\forall x \in \mathbb{R}^d$.

This means that every $Q(x,\cdot)$, even when x is far away, must contain the same fixed amount of mass $\varepsilon\nu_\star$.

Gaussian data example

If $\pi_{\mathrm{data}} = \mathcal{N}(0, \sigma_0^2)$, then

$$Q_{t|s}(x,\cdot) = \mathcal{L}(\overleftarrow{X}_t \mid \overleftarrow{X}_s = x) = \mathcal{N}(m_{s,t}x, \Gamma_{s,t}) \,,$$

with

$$m_{s,t} = \frac{\sigma_0^2 + 2(T-t)}{\sigma_0^2 + 2(T-s)} > 0 \,.$$

Thus, for every fixed $R > 0$,

$$Q_{t|s}(x, [-R, R]) \to 0 \qquad \text{as } |x| \to \infty \,.$$

⚠ No fixed positive common mass can be shared by all $Q_{t|s}(x,\cdot)$.

How to read the Lyapunov drift condition

The condition is an expectation inequality

For a Markov kernel Q, $QV(x) = \mathbb{E}[V(X_{n+1}) \mid X_n = x]$.

Thus the Lyapunov condition $QV(x) \le \lambda V(x) + K, \qquad \lambda < 1$ becomes

$$\mathbb{E}\left[\|X_{n+1}\|^2 \,\middle|\, X_n = x\right] \le \lambda \|x\|^2 + K.$$

Subtracting the current value, $QV_2(x) - V_2(x) \le -(1-\lambda)\|x\|^2 + K$.

Hence, whenever

$$\boxed{\|x\|^2 > \frac{K}{1-\lambda},}$$

we have

$$\boxed{\mathbb{E}\left[\|X_{n+1}\|^2 \,\middle|\, X_n = x\right] < \|x\|^2.}$$

Interpretation

Far enough in the tails, the chain has an *average restoring tendency*: its expected Lyapunov value decreases after one step.

Harris theorem: contraction in weighted total variation

Harris contraction (Hairer and Mattingly, 2011)

Under Lyapunov drift and a small-set minorization conditions there exist $\beta > 0$ and $\bar{\alpha} \in (0,1)$ such that

$$\boxed{\rho_\beta(\mu Q, \nu Q) \le \bar{\alpha}\, \rho_\beta(\mu, \nu)},$$

for all probability measures μ, ν with finite V-moment.

Weighted total variation

Let $V : \mathbb{R}^d \to [0, \infty)$ be a Lyapunov function and let $b > 0$. For probability measures μ, ν with finite V-moment, define

$$\rho_b(\mu, \nu) := \int_{\mathbb{R}^d} \big(1 + bV(x)\big)\, |\mu - \nu|(\mathrm{d}x)\,.$$

Interpretation

If $V(x) \to \infty$ as $\|x\| \to \infty$, then

$$\begin{aligned} \text{error near the center} \quad &: \quad (1 + bV_2(x)) \approx 1 \quad \Longrightarrow \quad \text{small cost}\,, \\ \text{error in the tails} \quad &: \quad (1 + bV_2(x)) \gg 1 \quad \Longrightarrow \quad \text{large cost}\,. \end{aligned}$$

Why ordinary TV does not capture the geometry of the state space

Consider the toy deterministic $1d$ kernel that brings points back toward the origin:

$$Q(x,\cdot) = \delta_{x/2}(\cdot)\,,$$

Total variation distance

For $x \neq 0$,

$$\|\delta_x - \delta_0\|_{\mathrm{TV}} = 1\,, \qquad \|\delta_x Q - \delta_0 Q\|_{\mathrm{TV}} = \|\delta_{x/2} - \delta_0\|_{\mathrm{TV}} = 1\,.$$

So ordinary TV does not see that x moved closer to the center.

Weighted total variation

With $V(x) = \|x\|^2$, for $x \neq 0$, using that $|\delta_x - \delta_0| = \delta_x + \delta_0$,

$$\rho_b(\delta_x, \delta_0) = \int (1 + b\|z\|^2)\,|\delta_x - \delta_0|(\mathrm{d}z) = (1 + b\|x\|^2) + (1 + b\|0\|^2) = 2 + b\|x\|^2$$

and

$$\rho_b(\delta_x Q, \delta_0 Q) = \rho_b(\delta_{x/2}, \delta_0) = (1 + b\|x/2\|^2) + (1 + b\|0\|^2) = 2 + \frac{b}{4}\|x\|^2\,.$$

Assumptions on the data distribution: Brownian case

Data assumptions

Let $\pi_{\mathrm{data}}(\mathrm{d}x) = p_0(x)\mathrm{d}x$, $p_0 \in C^2(\mathbb{R}^d)$ and assume there exist constants

$$\gamma_0 > 0\,, \qquad \kappa_0 \geq 0\,, \qquad C_0 > 0\,, \qquad m \geq 1\,,$$

such that, for all $x \in \mathbb{R}^d$,

$$\langle \nabla \log p_0(x), x\rangle \leq -\gamma_0 \|x\|^2 + \kappa_0\,, \qquad \text{(dissipativity)}$$

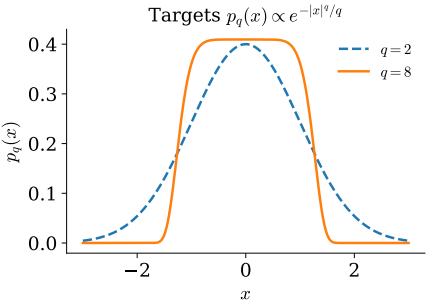
and

$$\|\nabla^2 \log p_0(x)\|_{\mathrm{F}} \leq C_0(1 + \|x\|^m)\,. \qquad \text{(polynomial regularity)}$$

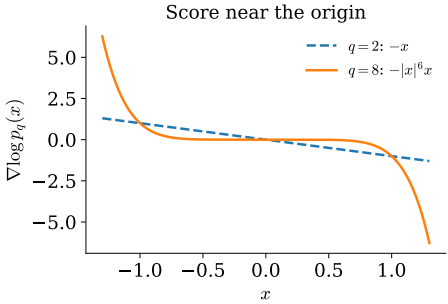
Interpretation: polynomial Jacobian growth

- The condition Jacobian condition is not a Lipschitz assumption, which would require instead $\sup_{x\in\mathbb{R}^d} \|\nabla^2 \log p_0(x)\| < \infty$.
- Here the Jacobian of the score may grow polynomially. This includes strongly confining distribution.

Example: Gaussian vs. strongly confining target



Densities $p_q(x) \propto \exp(-|x|^q/q)$, for $q = 2$ and $q = 8$.



Corresponding scores $\nabla \log p_q(x) = -|x|^{q-2}x$

Compactly supported data

Many data distributions are naturally supported on a bounded set, as density on $\mathbb{R}^d$, this means $p_0(x) = 0$ outside the support, so $\log p_0$ and $\nabla \log p_0$ not well-defined. One may consider

$$p_0^\varepsilon(x) \propto \int \exp\left(-\frac{\|x-y\|^q}{\varepsilon}\right) \pi_{\text{data}}(\mathrm{d}y)\,, \qquad q \geq 2\,.$$

Dissipativity: inward score and return to the mixing set

Dissipativity means inward radial drift

Outside the ball

$$\|x\|^2 \geq \frac{2\kappa_0}{\gamma_0},$$

the dissipativity condition gives

$$\langle \nabla \log p_0(x), x \rangle \leq -\frac{\gamma_0}{2}\|x\|^2.$$

Equivalently,

$$\left\langle \nabla \log p_0(x), \frac{x}{\|x\|} \right\rangle \leq -\frac{\gamma_0}{2}\|x\|.$$

Thus, far from the origin, the score points back to the center.

Why this matters for Harris

The reverse dynamics is driven by the score. The dissipativity condition prevents the process from escaping to infinity and pushes it back toward a region where minorization can produce mixing.

Dissipativity implies Gaussian-type tail control

Fix a direction $u \in \mathbb{S}^{d-1}$ and write $x = ru$. Then

$$\frac{\mathrm{d}}{\mathrm{d}r} \log p_0(ru) = \frac{\langle \nabla \log p_0(ru), ru \rangle}{r} .$$

By dissipativity,

$$\frac{\mathrm{d}}{\mathrm{d}r} \log p_0(ru) \leq -\gamma_0 r + \frac{\kappa_0}{r} .$$

For $r \geq \sqrt{2\kappa_0/\gamma_0}$, this gives

$$\frac{\mathrm{d}}{\mathrm{d}r} \log p_0(ru) \leq -\frac{\gamma_0}{2} r .$$

Integrating the above

$$\log p_0(ru) \leq C - \frac{\gamma_0}{4} r^2 ,$$

and therefore

$$p_0(x) \leq C \exp\left(-\frac{\gamma_0}{4} \|x\|^2\right) \qquad \text{for large } \|x\| .$$

Message

Dissipativity means enforce (sub-)Gaussian-type tail decay.

Stability under Gaussian perturbation

Dissipativity propagation through Gaussian smoothing

$$\langle \nabla \log p_0(x), x \rangle \leq -\gamma_0 \|x\|^2 + \kappa_0 \quad \Longrightarrow \quad \langle \nabla \log p_t(x), x \rangle \leq -\gamma_t \|x\|^2 + \kappa_t \,, \text{for all } t \geq 0 \,.$$

The smoothed score remains dissipative under Gaussian perturbation and, where, in the Brownian convention $p_t = p_0 * \mathcal{N}(0, 2t\mathrm{I}_d)$, one can take

$$\gamma_t = \frac{\gamma_0}{1 + 4\gamma_0 t} \,, \qquad \kappa_t = \frac{\kappa_0 + d}{1 + 4\gamma_0 t} \,.$$

⚠ the estimates are quantitative !

Remark: polynomial regularity is also stable

The polynomial growth control on the score Jacobian is also propagated by the Gaussian perturbation.

⚠ This transfers assumptions made on π_{data} to the smoothed marginals of the forward and therefore to the score function driving the reverse dynamics.

Ingredient 1: From dissipativity to a Lyapunov drift bound

The reverse process has generator, at reverse time u,

$$\mathcal{A}_u f(x) = 2\,\langle \nabla \log p_{T-u}(x)\,, \nabla f(x)\rangle + \Delta f(x)\,.$$

Apply the generator to $V_2(x) = \|x\|^2$

Since $\nabla V_2(x) = 2x$ and $\Delta V_2(x) = 2d$, we get

$$\mathcal{A}_u V_2(x) = 4\langle \nabla \log p_{T-u}(x), x\rangle + 2d\,.$$

Using the propagated dissipativity, $\mathcal{A}_u V_2(x) \le -4\gamma_{T-u} V_2(x) + (4\kappa_{T-u} + 2d) \le -\widetilde{\gamma}_u V_2 + \widetilde{\kappa}_u$.

Semigroup Lyapunov drift

By Dynkin's formula and Grönwall,

$$Q_{t|s} V_2(x) \le \lambda_{t|s} V_2(x) + K_{t|s}\,,$$

with

$$\lambda_{t|s} = \exp\left(-\int_s^t \widetilde{\gamma}_u \,\mathrm{d}u\right), \quad \text{and} \quad K_{t|s} = \int_s^t \exp\left(-\int_u^t \widetilde{\gamma}_v \,\mathrm{d}v\right) \widetilde{\kappa}_u \,\mathrm{d}u\,.$$

Ingredient 2: Local minorization

The reverse kernel admits the Bayes representation

$$Q_{t|s}(x, \mathrm{d}y) = \frac{\varphi_{2(t-s)}(x-y)\, p_{T-t}(y)}{p_{T-s}(x)}\, \mathrm{d}y, \qquad 0 \le s < t \le T\,,$$

where φ_a denotes the density of $\mathcal{N}(0, a\mathrm{I}_d)$.

Restriction to a compact set

Fix

$$C_r = \{x \in \mathbb{R}^d : \|x\| \le r\}\,.$$

For $x \in C_r$, the Gaussian factor admits a common lower envelope:

$$\varphi_{2(t-s)}(x-y) \ge c_{r,t-s}\, \varphi_{t-s}(y)\,, \qquad c_{r,t-s} > 0\,.$$

Moreover, since p_{T-s} is Gaussian-smoothed, $p_{T-s}(x) \le M_{T-s} < \infty$.

Localized Doeblin condition

Therefore, there exist a probability measure $\nu_{t|s}$ and $\varepsilon^{(r)}_{t|s} > 0$ such that, for all $x \in C_r$,

$$Q_{t|s}(x, A) \ge \varepsilon^{(r)}_{t|s}\, \nu_{t|s}(A)\,, \qquad A \in \mathcal{B}(\mathbb{R}^d)\,.$$

Conclusion: Harris forgetting of the reverse kernel

Two ingredients

For $V_2(x) = \|x\|^2$, for $0 \le s < t \le T$, the previous estimates give:

$$Q_{t|s} V_2(x) \le \lambda^{(2)}_{t|s} V_2(x) + K^{(2)}_{t|s}\,, \qquad \lambda^{(2)}_{t|s} < 1\,,$$

and, on $C_r = \{x : \|x\| \le r\}$,

$$Q_{t|s}(x,\cdot) \ge \varepsilon^{(r)}_{t|s} \nu_{t|s}(\cdot)\,, \qquad x \in C_r\,.$$

Harris contraction

For any, $r^2 > \frac{2K^{(2)}_{t|s}}{1-\lambda^{(2)}_{t|s}}$, $a_0 \in (0, \varepsilon^{(r)}_{t|s})$, $\eta_0 \in \left(\lambda^{(2)}_{t|s} + \frac{2K^{(2)}_{t|s}}{r^2}, 1\right)$, $b^{(r)}_{s,t} := \frac{a_0}{K^{(2)}_{t|s}}$. Then,

$$\rho_{b^{(r)}_{s,t}}\left(\mu Q_{t|s}, \nu Q_{t|s}\right) \le \bar{\alpha}_{t|s} \rho_{b^{(r)}_{s,t}}(\mu, \nu)\,,$$

with

$$\bar{\alpha}_{t|s} = \left[1 - \left(\varepsilon^{(r)}_{t|s} - a_0\right)\right] \vee \frac{2 + r^2 b^{(r)}_{s,t} \eta_0}{2 + r^2 b^{(r)}_{s,t}} < 1\,.$$