On Forgetting and Stability of Score-based Generative models

Based on joint works with **Stanislas strasman**, Gabriel V. Cardoso, V. Lemaire and A. Ocello

---

Sylvain Le Corff

Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université

Table of contents

Propos liminaire important

<footer>Forgetting of SBGMs 3/23</footer>

Score-based generative models (G. Cardoso)

SGMs as Markovian models

Forward diffusion (noising)

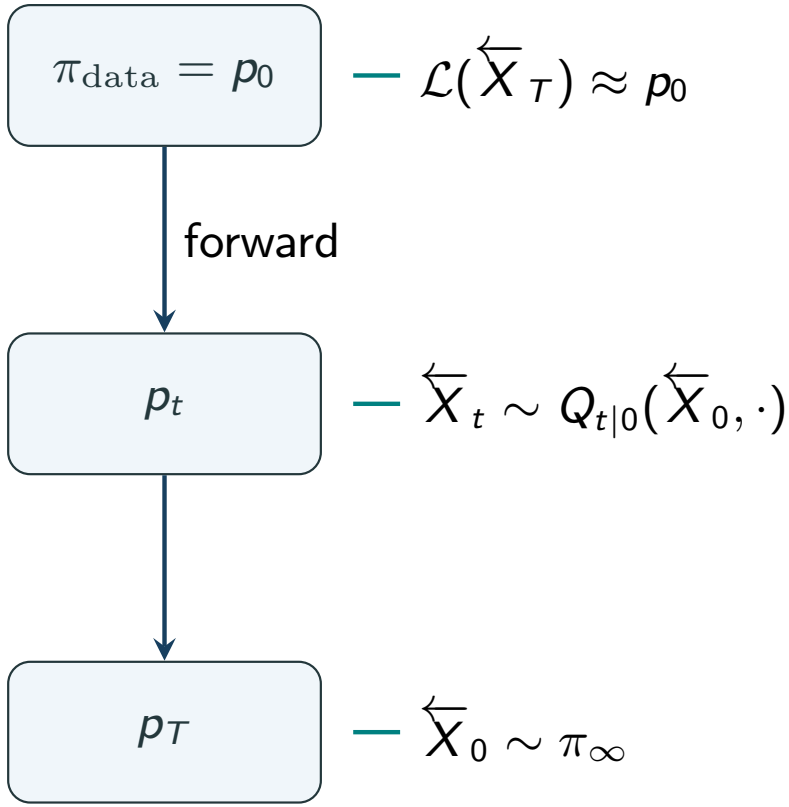
$$d\overrightarrow{X}_t = -\alpha\beta_t \overrightarrow{X}_t\, dt + \sqrt{2\beta_t}\, dB_t, \quad \overrightarrow{X}_0 \sim \pi_{\text{data}}.$$

Reverse diffusion (sampling)

$$d\overleftarrow{X}_t = \left(\alpha\bar{\beta}_t \overleftarrow{X}_t + 2\bar{\beta}_t S_{T-t}(\overleftarrow{X}_t)\right) dt + \sqrt{2\bar{\beta}_t}\, dB_t,$$

with score $S_t(x) = \nabla \log p_t(x)$ and Markov semigroup

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



Viewpoint. The generative model is a time-inhomogeneous Markov sampler that transports a simple reference law toward π_{data}.

Where do sampling errors come from?

Sampling is exact only if we can start from p_T, know the **true score**, and **simulate the reverse SDE exactly**.

Three approximations used in practice

1. **Initialization / mixing error:** replace p_T by a tractable reference law π_∞.
2. **Score approximation:** replace S_t by a learned network $s_\theta(t, x)$.
3. **Discretization error:** replace the reverse diffusion by Euler–Maruyama.

This yields the discrete Markov chain

$$\bar{X}^\theta_{t_{k+1}} = \bar{X}^\theta_{t_k} + \Delta_k \left(\alpha \bar{X}^\theta_{t_k} + 2 s_\theta(\bar{X}^\theta_{t_k}, T - t_k)\right) + \sqrt{2\Delta_k}\, \xi_k,$$

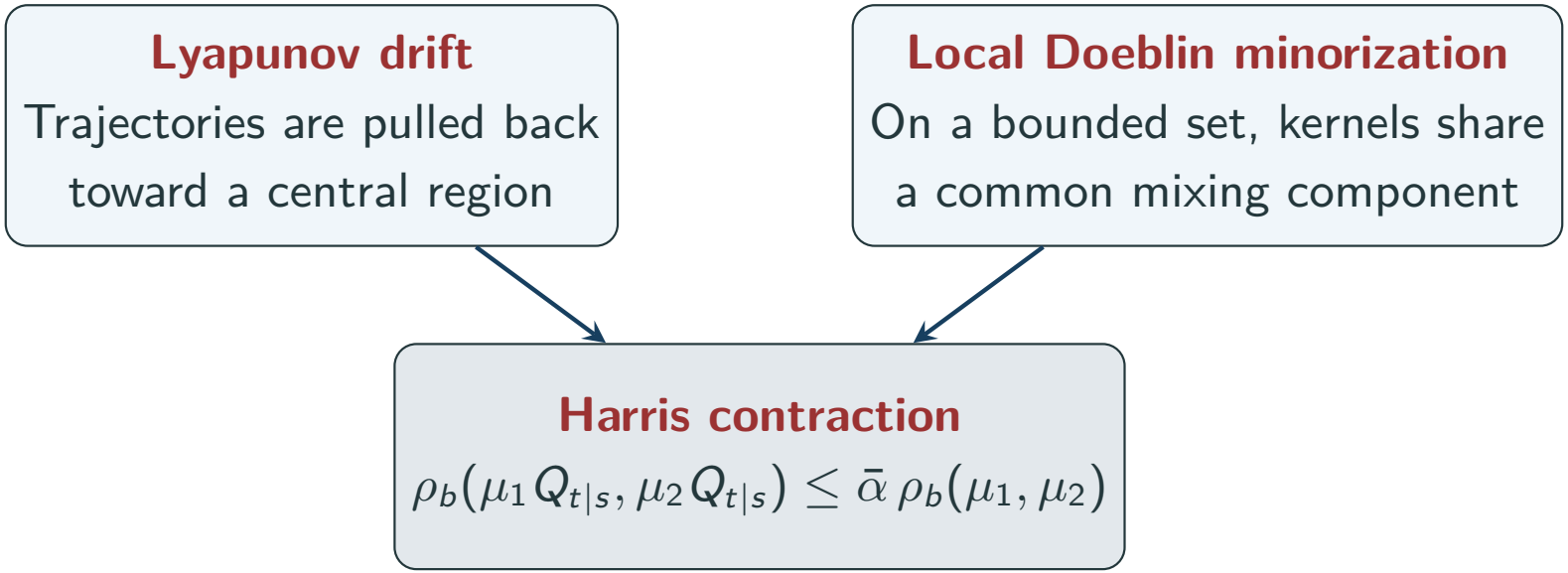
with one-step kernel $Q^\theta_{t_{k+1}|t_k}$.

Question: how do these **local errors** propagate along the reverse trajectory?

Harris roadmap: why Lyapunov + local Doeblin?

The objective is to analyze the reverse kernels using a metric ρ_b, $b > 0$ such that for all measures μ_1 and μ_2,

$$\|\mu_1 - \mu_2\|_{\mathrm{TV}} \leq c_b \rho_b(\mu_1, \mu_2) \quad \text{and} \quad \mathcal{W}_2^2(\mu_1, \mu_2) \leq \tilde{c}_b \rho_b(\mu_1, \mu_2).$$



Interpretation: one condition controls the tails, the other creates overlap on a "small set"; together they imply forgetting of the initial state.

Exponential convergence of Markov chains (G. Lang)

Exponential convergence of Markov chains

Let P be a Markov transition kernel on a measurable space X.

The goal is to obtain **unique ergodicity** and **exponential convergence** toward the invariant measure on an unbounded state space.

For a given norm $\|\cdot\|$, this means that:

- P admits a **unique invariant probability measure** $\mu_\star$.
- There exist constants $C > 0$ and $\gamma_0 \in (0, 1)$ such that for every measurable φ with finite weighted norm,

$$\|P^n\varphi - \mu_\star(\varphi)\| \leq C\gamma_0^n \|\varphi - \mu_\star(\varphi)\|.$$

Hence convergence to equilibrium is exponential in $\|\cdot\|$.

Assumption 1: Geometric Drift

There exist a measurable function $V : \mathsf{X} \to [0, \infty)$ and constants $K \geq 0$, $\gamma \in (0, 1)$ such that

$$(PV)(x) \leq \gamma V(x) + K, \qquad x \in \mathsf{X}.$$

Interpretation

- Since $\gamma < 1$, the chain is pushed back toward the center (region with small values of V).
- Excursions to large values of V are controlled.

Role to obtain ergodicity

- Provides a tight control on return toward a central region.
- Geometric decay outside the small set.

Lyapunov in practice

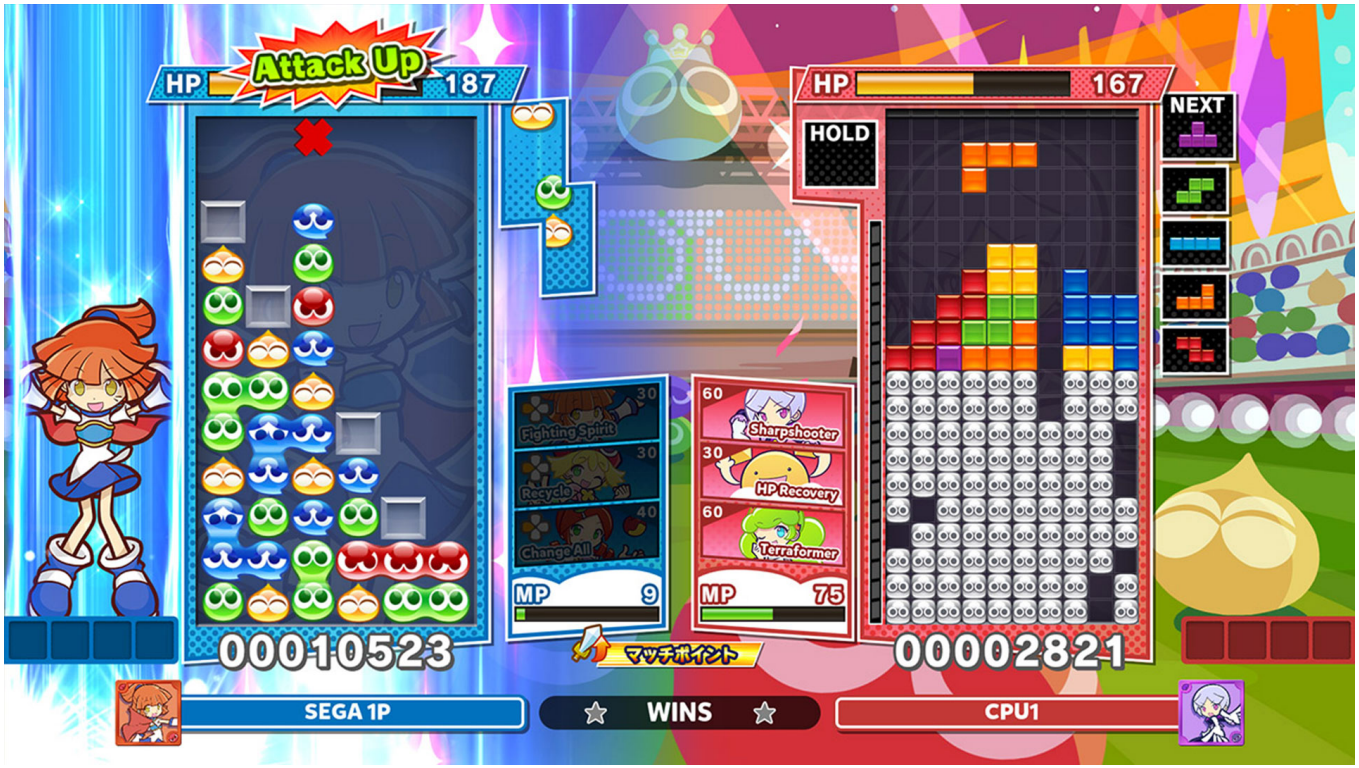


Figure 1: "Fonction de Lyapunov : hauteur de l'empilement. La hauteur révèle tout : stabilité d'un côté (joueur dit "fort"), dérive de l'autre (joueur dit "faible"). D'un côté, retour à l'équilibre en un mouvement, de l'autre, un simple empilement de dêchets, visiblement à l'infini. Le joueur "faible" est parfois dit enterré sous les "garbages".

Illustration d'un joueur faible

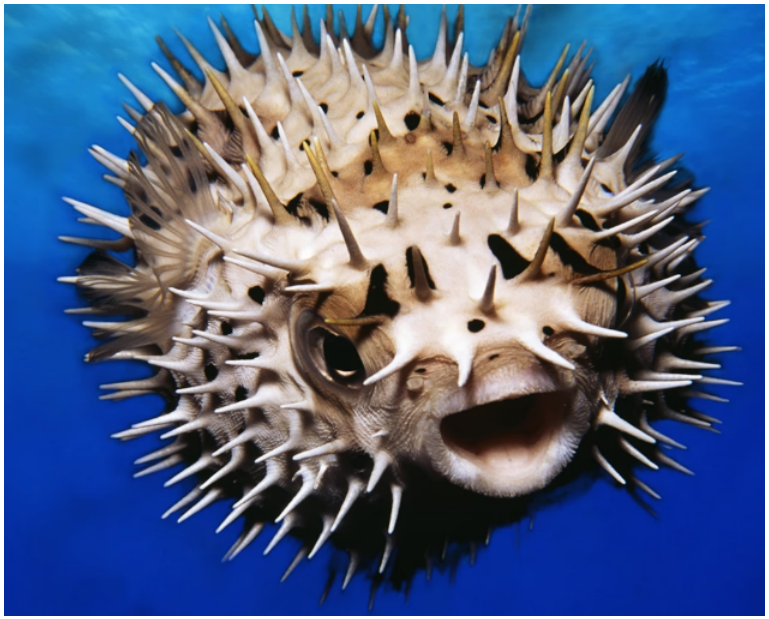



Figure 2: Les poissons-globes (Tetraodontidae, en français Tétraodontidés ; du grec ancien tetra = quatre et odous = dent) ou poissons-ballons.
Gonfle **soudainement**. Devient **incontrôlable**. Réactions **exagérées**.
Vous pensez à quelqu'un ?

Illustration d'un joueur fort

Figure 3: Nautilus vanuatuensis, organisation propre et anticipée. Les abeilles réalisent naturellement un pavage optimal de l'espace.
Chaque pièce s'ajoute selon une structure déjà prévue.

Assumption 2: Minorization on a Small Set

Choose

$$C = \{x \in \mathsf{X} : V(x) \leq R\}, \qquad R > \frac{2K}{1-\gamma}.$$

Assume there exist $\alpha \in (0,1)$ and a probability measure ν such that

$$\inf_{x \in C} P(x, \cdot) \geq \alpha \, \nu(\cdot).$$

Interpretation

- This is a localized version of Doeblin's condition.
- It creates a coupling / mixing mechanism on the set where the chain regularly returns.

Why the threshold on R?

- It ensures that the drift and the minorization estimates can be combined quantitatively.

Doeblin in practice

Figure 4: Condition de Doeblin respectée à gauche “peu importe d’où on part, il y a toujours une probabilité uniforme de retomber sur une zone commune bien mélangée.” Tout chemin a une chance de revenir dans une zone d’équilibre. A droite... est-ce utile de commenter ?

Example: Mean-reverting random walk

Consider the Markov chain on $\mathbb{R}$

$$X_{n+1} = aX_n + \xi_{n+1}, \qquad |a| < 1,$$

where (ξ_n) are i.i.d. $\mathcal{N}(0, \sigma^2)$.

Lyapunov function

$$V(x) = 1 + x^2$$

Drift condition

$$\mathbb{E}[V(X_{n+1}) \mid X_n] = 1 + a^2 X_n^2 + \sigma^2 \leq a^2 V(X_n) + (1 + \sigma^2).$$

Hence Hypothesis 1 holds with

$$\lambda = a^2 < 1, \qquad b = 1 + \sigma^2.$$

Contraction toward the origin, large values are pulled back.

Weighted Norms by Hairer et al.

Define, for $\beta > 0$,

$$\|\varphi\|_\beta = \sup_x \frac{|\varphi(x)|}{1 + \beta V(x)}$$

and the associated weighted total variation distance

$$\rho_\beta(\mu_1, \mu_2) = \sup_{\|\varphi\|_\beta \le 1} \int_X \varphi(x)\,(\mu_1 - \mu_2)(dx).$$

Main contraction result

Under Assumptions 1–2, there exist $\beta > 0$ and $\bar{\alpha} \in (0, 1)$ such that

$$\rho_\beta(P\mu_1, P\mu_2) \le \bar{\alpha}\,\rho_\beta(\mu_1, \mu_2)$$

for all probability measures μ_1, μ_2.

P becomes a **strict contraction in a weighted metric**.

What Does This Give?

Theorem 1.2: Geometric ergodicity

- P admits a unique invariant probability measure $\mu_\star$.
- There exist constants $C > 0$ and $\gamma_0 \in (0, 1)$ such that for every measurable φ with finite weighted norm,

$$\|P^n\varphi - \mu_\star(\varphi)\| \leq C\gamma_0^n \|\varphi - \mu_\star(\varphi)\|.$$

Cnvergence to equilibrium is exponential in a weighted supremum norm.

Theorem 3.2: Existence

- The contraction argument also gives an invariant measure μ_∞ with

$$\int_{\mathsf{X}} V(x)\, \mu_\infty(dx) < \infty.$$

Convergence of SBGMs (S. Strasman)

Lyapunov drift from dissipativity

Assumption on the data score $S_0 = \nabla \log p_0$

$$\langle S_0(x), x\rangle \le -\gamma_0 \|x\|^2 + \kappa_0, \qquad \|\nabla S_0(x)\|_F \le C_0(1 + \|x\|^p).$$

- **Dissipativity** condition: outside a compact set, the score points inward.
- **Polynomial growth** of the Jacobian: no global Lipschitz assumption.
- These assumptions cover nonconvex and multimodal targets.

The dissipativity propagates along the diffusion flow:

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

Hence, for $V_\ell(x) = \|x\|^\ell$:

Lyapunov

$$Q_{t|s} V_\ell(x) \le \lambda_{t|s}^{(\ell)} V_\ell(x) + K_{t|s}^{(\ell)}, \qquad \lambda_{t|s}^{(\ell)} = \exp\Big(-\int_s^t \tilde{\gamma}_{v,\ell}\, dv\Big).$$

Local Doeblin condition on a small set

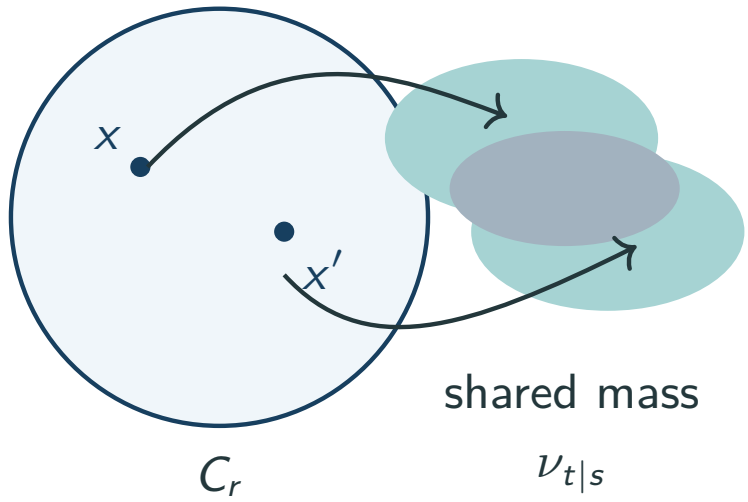
Fix a radius $r > 0$ and $C_r = B_r(0) = \{x : \|x\|^2 \le r^2\}$.

Local Doeblin

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

for some probability measure $\nu_{t|s}$ and $\varepsilon_{t|s}^{(r)} \in (0, 1)$.

- This is a **localized** Doeblin condition.
- A global minorization on all of $\mathbb{R}^d$ would be too strong.
- On C_r, different starting points share a common part of their transition law.



From local conditions to contraction

Decompose the global error into

$$\rho_b(\pi_{\text{data}}, \widehat{\pi}_N^\theta) \leq \rho_b(p_T Q_T, \pi_\infty Q_T) + \rho_b(\pi_\infty Q_T, \pi_\infty Q_{0:N}^\theta) \, .$$

By applying **Doeblin + Lyapunov** iteratively across the N transitions of the discretization grid, there exist $b > 0$ and $\bar{\alpha} \in (0, 1)$, such that

$$\rho_b(p_T Q_T, \pi_\infty Q_T) \leq \bar{\alpha}^N \rho_b(p_T, \pi_\infty) \, .$$

Telescoping argument along the grid to make **appear one step kernel error**. Define the intermediate measures:

$$\eta_k = \widehat{\pi}_{k-1}^\theta \, Q_{k|k-1} Q_{N|k} \, , \quad \tilde{\eta}_k = \widehat{\pi}_{k-1}^\theta \, Q_{k|k-1}^\theta Q_{N|k} \, .$$

We have,

$$\rho_b(\pi_\infty Q_T, \pi_\infty Q_{0:N}^\theta) \leq \sum_{k=1}^N \rho_b(\eta_k, \tilde{\eta}_k) \, .$$

Final bound

Final bound

$$\rho_b(\pi_{\mathrm{data}}, \hat{\pi}_N^\theta) \le \underbrace{\bar{\alpha}^N \Lambda(T) C_{\mathrm{mix}}}_{\text{initialization}} + \sum_{k=1}^{N} \bar{\alpha}^{N-k} \Big(\underbrace{\Delta_k C_{\mathrm{discr}}^{k-1}}_{\text{discretization}} + \underbrace{\sqrt{\Delta_k}\, C_{\mathrm{net}}^{k-1} \, \|E_{k-1}\|_{L^2(\hat{\pi}_{k-1}^\theta)}}_{\text{score approximation}} \Big).$$

The reverse diffusion **forgets** early perturbations.

Take-away for the talk

1. Interpret the generator as a Markov kernel.
2. Prove **Lyapunov drift** + **local Doeblin minorization**.
3. Invoke Harris theory to obtain contraction.

GMM (25 components)

All experiments use dimension $d = 50$, with 20 independent runs.

Run the reverse-time sampler over the full horizon $[0, T]$ and **introduce a *single* controlled score perturbation at one discretization step**.

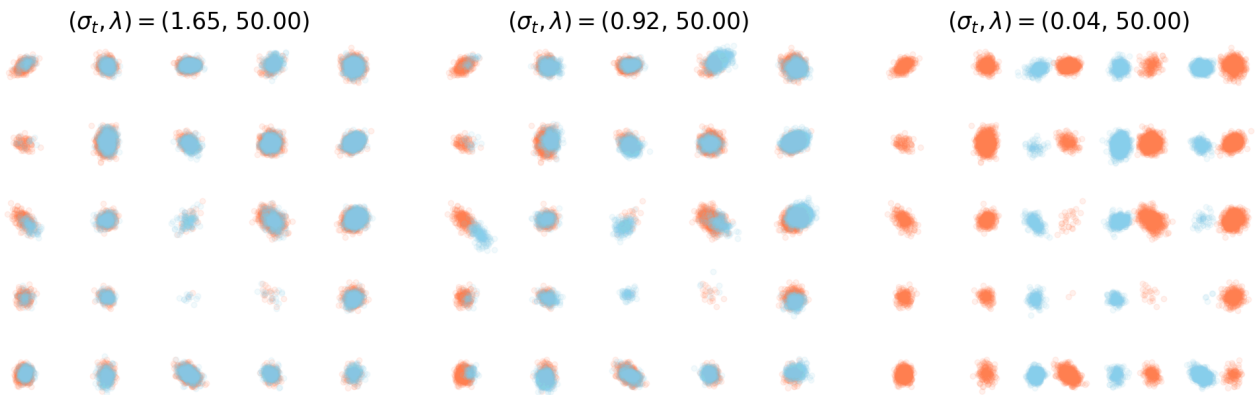


Figure 5: Local perturbation of the score in the GMM case for several noise levels σ_t and with $\lambda = 50$. Red points are samples from the true distribution and blue points the output of the perturbed score experiment.

GMM (25 components)

All experiments use dimension $d = 50$, with 20 independent runs.

Run the reverse-time sampler over the full horizon $[0, T]$ and **introduce a *single* controlled score perturbation at one discretization step**.

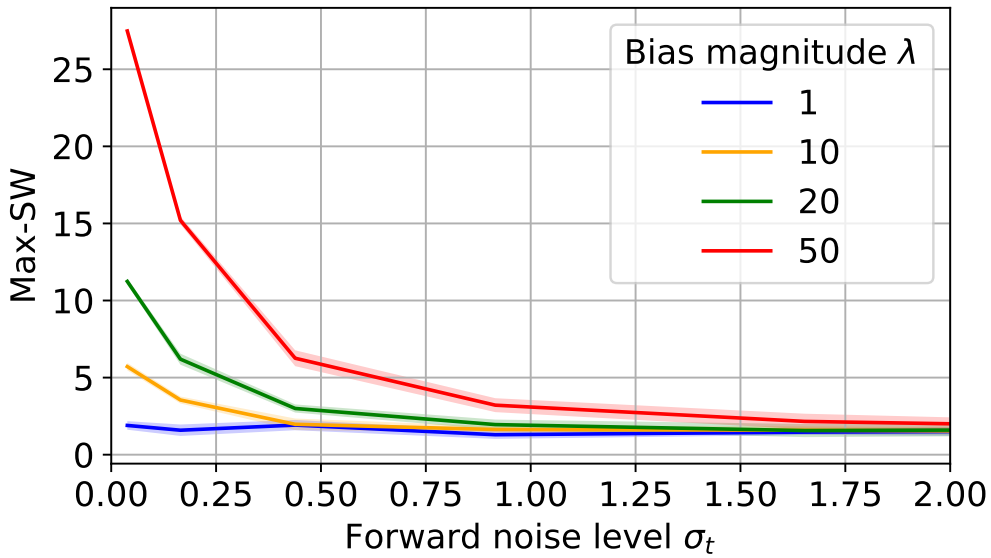


Figure 6: Max $\mathcal{W}_2$ as a function of the noise level and perturbation magnitude λ for the score-perturbation experiment. We use the forward-time convention ($t = 0$ is the data distribution).

En conclusion

Alors, pas mal cette petite danse non ?