DIFFUSION REMOVES LANGEVIN'S CONDITIONING DEPENDENCE: A SHARP GAUSSIAN ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite their empirical success, why diffusion models overcome the bottlenecks of classical score-based samplers remains unclear. In this work, we leverage Gaussian distributions to isolate this phenomenon. We establish 2-Wasserstein convergence bounds for optimized hyperparameters, showing that diffusion processes achieve a sampling error of $O(\sqrt{d\lambda_{\max}}\log N/N)$, where d is the dimension, N the number of sampling steps, and $\lambda_{\max}$ the largest eigenvalue of the target covariance matrix. Unadjusted and underdamped Langevin dynamics suffer from an additional $\sqrt{\kappa}$ factor, where κ is the condition number. These rates follow from spectral bounds which are sharp: we confirm them via matching first-order asymptotics as $N \to \infty$. Our analysis provides a rigorous characterization, in the Gaussian setting, of how time-dependent score trajectories remove condition-number dependence during sampling. By contrast, in the learning phase, we show that estimating the unnoised score by gradient descent leads to essentially the same estimator as estimating a noisy score, which suggests that the benefits of noising do not come from the learning phase.

1 INTRODUCTION

Since their introduction, diffusion models (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020; Song et al., 2021) have achieved state-of-the-art performance across a wide range of domains (Dhariwal and Nichol, 2021; Kong et al., 2021; Hoogeboom et al., 2022; Watson et al., 2023). Training diffusion models implies estimating the gradient of the log-density (the so-called *score*) of a noisy version of the target. Then samples are produced by iterative denoising through a formula involving this score. Taking a step back, iterative score-based sampling strategies predate by far modern diffusion models, as epitomized by the celebrated Langevin algorithm (Pavliotis, 2014; Roberts and Tweedie, 1996) and its many variants such as underdamped Langevin (Leimkuhler and Matthews, 2015). The key factors explaining their success compared to those more classical algorithms are still unclear: what exact role does the noising process in diffusion models play, and how does it interact with the learning and sampling phases?

Our primary goal in this work is to prove a separation in sampling efficiency between diffusion models and Langevin dynamics in terms of their dependence on the condition number. This sheds light on the practical effectiveness of diffusion models in ill-conditioned distributions such as natural images (Guth et al., 2022; Field, 1987). To do so, we consider the class of (centered) Gaussian targets. Naturally, sampling from a Gaussian distribution is not our primary concern here. Rather, we use this analytically tractable setting to derive sharp convergence rates accompanied by matching asymptotics, in order to highlight how diffusion models overcome limitations of other score-based sampling methods. As discussed below, deriving tight condition-number dependence for more general target classes such as log-concave distributions is a long-standing open problem (Cheng et al., 2018; Dalalyan and Riou-Durand, 2020), which is out of the scope of the present paper. Additionally, we focus for most of the paper on the sampling phase, comparing the sampling convergence rates of diffusion, unadjusted Langevin algorithm (ULA) and underdamped Langevin dynamics (ULD). Assuming access to the true score yields a clean analysis and the separation between diffusion and Langevin, hence for simplicity we focus on this case. The connection with the estimation of the score in the learning phase is discussed in the latter part of our work.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Convergence rates and ε-complexities for the Gaussian target. Rates are the orders of $W_2(\mathbb{P}_N, \mathbb{P})$ as $N \to \infty$ for fixed Σ (Theorems 1 and 3). The ε-complexity is the iteration budget N needed to achieve $W_2(\mathbb{P}_N, \mathbb{P}) \leq \varepsilon$, where $\tilde{O}$ hides logarithmic factors. Diffusion models remove both the conditioning factor $\sqrt{\kappa}$ of Langevin samplers and the burn-in of order κ.

Algorithm	**Asymptotic equivalent**	**ε-complexity**
ULA	$\Theta\left(\lambda_{\max}\sqrt{\mathrm{Tr}(\Sigma^{-1})}\,\frac{\log N}{N}\right)$	$\tilde{O}\left(\kappa + \frac{\sqrt{d\kappa\lambda_{\max}}}{\varepsilon}\right)$
ULD	$\Theta\left(\lambda_{\max}\sqrt{\mathrm{Tr}(\Sigma^{-1})}\,\frac{\log N}{N}\right)$	$\tilde{O}\left(\kappa + \frac{\sqrt{d\kappa\lambda_{\max}}}{\varepsilon}\right)$
Diffusion	$\Theta\left(\sqrt{\mathrm{Tr}\,\Sigma}\,\frac{\log N}{N}\right)$	$\tilde{O}\left(\frac{\sqrt{d\lambda_{\max}}}{\varepsilon}\right)$

Contributions. We establish upper bounds on the 2-Wasserstein error between the target distribution and the distribution sampled from the diffusion, ULA and ULD. Convergence rates are derived in Section 2 by optimizing these bounds with respect to hyperparameters (see Table 1 for summary). The bounds for diffusion and for Langevin differ by a factor proportional to the square root of the condition number. Thanks to explicit computations of the Wasserstein errors, we confirm these rates via matching first-order asymptotics. By contrast, in the learning phase, we show in Section 3 that estimating the unnoised score by gradient descent leads to essentially the same estimator as estimating a noisy score, the two differing only in how they regularize the directions unseen in the data, which suggests that the benefits of noising do not come from the learning phase.

Related work. For ULA, a similar analysis of the two error terms was already performed, for example, by Durmus et al. (2019); Wibisono (2018) and specifically in the Gaussian case by Pedregosa (2023). Building upon these works, we perform hyperparameter optimization and compute the asymptotic equivalent (rather than upper bounds). Many works have studied ULD in the log-concave case, see, *e.g.*, Cheng et al. (2018); Dalalyan and Riou-Durand (2020); Leimkuhler et al. (2024); Kim et al. (2026) for studies of the stochastic Euler scheme and Euler–Maruyama (EM) integrators. This line of work yields untight dependence on the condition number due to the bias term induced by non-vanishing step-sizes (see, *e.g.*, Eq. (11) in Dalalyan and Riou-Durand (2020) with a $\kappa^{3/2}$ in the complexity bound compared to our $\sqrt{\kappa}$). A line of work has proved improved condition dependence for Gaussian targets (Shen and Lee, 2019; Apers et al., 2024; Gouraud et al., 2025), however with more complex integrators (splitting, random step-sizes, or Metropolis corrections). To our knowledge, the tight study of the bias of ULD for a Gaussian target and the EM integrator is a novel contribution.

For diffusion models, Pierret and Galerne (2025) derive and numerically analyze error terms for Gaussian targets, however do not analytically derive optimal hyperparameters, discuss the dependence on the condition number, nor derive asymptotic equivalents. Closer are the works by Hurault et al. (2025) and Hurault et al. (2026), which derive Wasserstein bounds for Gaussian data and discuss optimal hyperparameter choices depending on the covariance spectrum. A key difference with our work is that their analysis holds asymptotically for small step-size, whereas we make a fully non-asymptotic analysis (albeit in less generality), which enables us to capture the interaction between the step-size and other parameters, and to derive optimal step complexity with respect to our bounds. Wasserstein bounds beyond the Gaussian case are given for instance by Strasman et al. (2025); Beyler and Bach (2025).

2 Sampling: diffusion removes the condition-number dependence

2.1 Problem setup

We denote by $\mathbb{P}$ the target d-dimensional centered Gaussian distribution $\mathcal{N}(0, \Sigma)$. Let $(\lambda_i)_{1\leq i\leq d}$ denote the eigenvalues of the positive-definite covariance matrix Σ, and $\lambda_{\min}$ and $\lambda_{\max}$ be respectively the smallest and largest eigenvalues. The ratio

$$\kappa := \frac{\lambda_{\max}}{\lambda_{\min}}$$

is the condition number of Σ, which measures the degree of anisotropy of $\mathbb{P}$. Ill-conditioning is a classical source of difficulty for Langevin-based sampling methods, as their convergence rates typically deteriorate with κ (Cheng et al., 2018; Dalalyan and Riou-Durand, 2020).

Let p denote the Lebesgue density of $\mathbb{P}$. We define the score s of the target distribution as

$$s : x \mapsto \nabla \log p(x) = -\Sigma^{-1} x.$$

We assume that the score is known exactly in this section; the learning problem is studied separately in Section 3.

We measure sampling error in the 2-Wasserstein distance, denoted by W_2. For two centered Gaussians whose covariance matrices Σ_1 and Σ_2 commute, the distance reduces to the Fréchet distance $\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\|_F$, with $\|\cdot\|_F$ the Frobenius norm (Gelbrich, 1990). For each sampling scheme, we denote by $\mathbb{P}_k$ the law of its k-th iterate.

Time-homogeneous Langevin samplers. We first consider the Langevin process associated with $\mathbb{P}$,

$$\mathrm{d}X_t = s(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t,\ t \in \mathbb{R}_+,\ X_0 = 0, \tag{1}$$

where (W_t) is a standard Brownian motion.

The unadjusted Langevin algorithm (ULA) (Dalalyan, 2017) is simply the Euler–Maruyama discretization of the SDE in (1). Given a step-size η, an iteration budget N, and a sequence (ξ_k) of independent standard Gaussian vectors in $\mathbb{R}^d$, it is defined by

$$\begin{cases} X_0 = 0, \\ \forall k \in \{0, \dots, N-1\},\ X_{k+1} = X_k + \eta s(X_k) + \sqrt{2\eta}\xi_k. \end{cases} \tag{2}$$

We next consider underdamped Langevin dynamics (ULD) (Cheng et al., 2018), which introduces an auxiliary velocity variable into the Langevin dynamics with the aim of accelerating convergence:

$$\begin{cases} \mathrm{d}X_t = V_t\,\mathrm{d}t, \\ \mathrm{d}V_t = -\gamma V_t\,\mathrm{d}t + s(X_t)\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}W_t, \end{cases} \quad t \in \mathbb{R}_+, \tag{3}$$

with initialization $X_0 = V_0 = 0$, where $\gamma > 0$ is the friction coefficient. We discretize this dynamics using the Euler–Maruyama scheme, yielding with the same notation as above

$$\begin{cases} X_{k+1} = X_k + \eta V_k, \\ V_{k+1} = (1 - \eta\gamma)V_k + \eta s(X_k) + \sqrt{2\eta\gamma}\xi_k. \end{cases} \tag{4}$$

Time-inhomogeneous diffusion sampler. Diffusion-based sampling follows a different principle (Sohl-Dickstein et al., 2015; Song et al., 2021): rather than sampling directly with the score of the target distribution, it constructs a family of perturbed distributions and uses a time-dependent score along this family. Many conventions exist; we study the following one because it leads to telescoping cancellations in the computations, greatly simplifying our analysis (see Appendix C.2.1). Concretely, for a general (centered) target $\mathbb{Q}$, consider the forward noising process

$$\begin{cases} \mathrm{d}X_t = \mathrm{d}W_t,\ t \in [0, T], \\ X_0 = X \sim \mathbb{Q}, \end{cases}$$

whose marginal $\mathbb{Q}_t$ at time t has Lebesgue density q_t and score $s_t := \nabla \log q_t$. The forward diffusion gradually smooths the target distribution, and T is chosen large enough so that $\mathbb{Q}_T \approx \mathcal{N}(0, T\,\mathrm{I})$, a distribution that is straightforward to sample from. Sampling proceeds by reversing this transformation, *i.e.*, by simulating the reverse-time SDE (Anderson, 1982)

$$\begin{cases} \mathrm{d}Y_t = s_{T-t}(Y_t)\,\mathrm{d}t + \mathrm{d}W_t,\ t \in [0, T-\delta], \\ Y_0 \sim \mathcal{N}(0, T\,\mathrm{I}), \end{cases} \tag{5}$$

where $\mathcal{N}(0, T\,\mathrm{I})$ replaces the intractable true marginal $\mathbb{Q}_T$ and $\delta > 0$ is an early-stopping threshold, standard in the analysis of diffusion models, where it prevents a possible blow-up of the score as

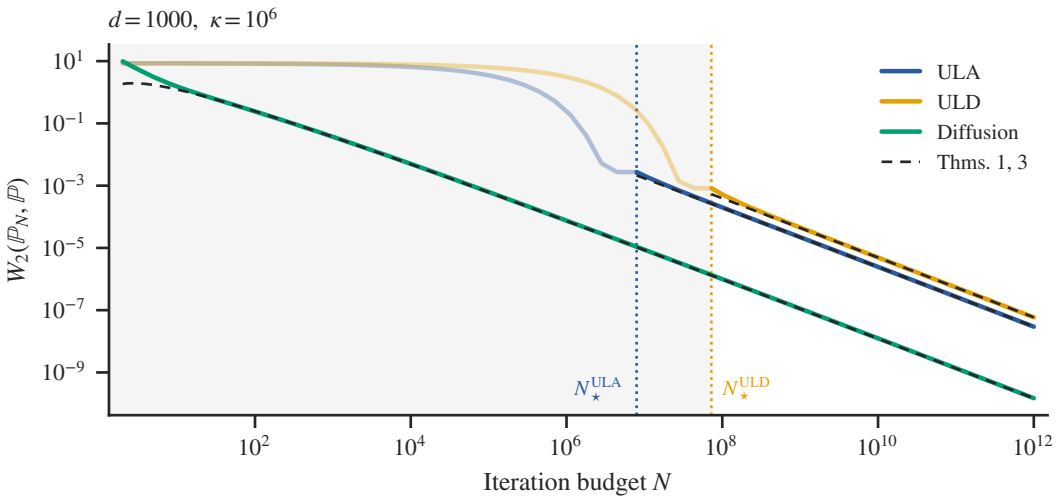


Figure 1: Sampling error $W_2(\mathbb{P}_N, \mathbb{P})$ as a function of the iteration budget N (log–log scale) on an ill-conditioned Gaussian target: $d = 1000$, log-uniform spectrum $\lambda_i = \kappa^{-(i-1)/(d-1)}$ with $\kappa = 10^6$. For this spectrum, $\lambda_{\max}\sqrt{\mathrm{Tr}(\Sigma^{-1})/\,\mathrm{Tr}\,\Sigma} = \sqrt{\kappa}$. Each point is a separate run: for every N, the sampler is tuned as in Theorems 1 and 3, run for N steps, and the error is evaluated in closed form. Dashed lines: asymptotic equivalents of the two theorems, drawn from $N_\star$ onwards for the Langevin samplers. Segments $N < N_\star$: the step-size of Theorem 1 is inadmissible (Remark 2) and the largest admissible one is used instead, $\eta = \lambda_{\min}$ for ULA and $\eta = 1/\gamma$ for ULD; see the discussion after Remark 2.

$t \to 0$ (De Bortoli, 2022; Benton et al., 2024). In our Gaussian setting $\mathbb{Q} = \mathbb{P}$, all these quantities are explicit: the marginal at time t is $\mathbb{Q}_t = \mathcal{N}(0, C_t)$ with $C_t = \Sigma + t\,\mathrm{I}$, and

$$s_t : x \mapsto -C_t^{-1}x$$

whose Lipschitz constant $(\lambda_{\min} + t)^{-1}$ is bounded by $(\lambda_{\min} + \delta)^{-1}$ on $[\delta, T]$.

Since information about the target is concentrated near $t = T - \delta$, we discretize (5) using a decreasing sequence of step-sizes (η_k) instead of a constant η. We fix a decreasing grid $T = \tau_0 > \tau_1 > \cdots > \tau_N = \delta$ defining $\tau_k = \tau_{k-1} - \eta_k$ for $k \in \{1, \dots, N\}$. The diffusion algorithm we study is then

$$\begin{cases} Y_0 \sim \mathcal{N}(0, T\,\mathrm{I}), \\ \forall k = 1, \dots, N,\ Y_k = Y_{k-1} + \eta_k s_{\tau_{k-1}}(Y_{k-1}) + \sqrt{\eta_k}\xi_k. \end{cases} \tag{6}$$

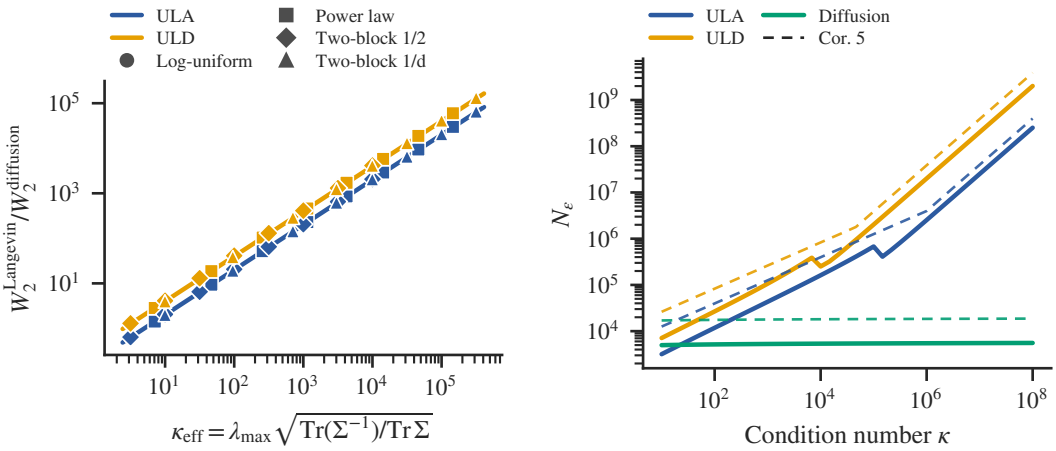
The main difference between diffusion and Langevin lies in the use of a time-dependent score s_t in the former. Langevin methods only use the target score s, and therefore operate under a fixed geometry governed by κ. This leads, as we will show in Theorem 1, to a penalizing $\sqrt{\kappa}$ factor in the convergence rate. In contrast, diffusion progressively modifies the geometry of the score along the sampling trajectory. Indeed, note that the condition number associated with s_t is

$$\kappa_t = \frac{t + \lambda_{\max}}{t + \lambda_{\min}} \underset{t\to\infty}{\longrightarrow} 1,$$

decreasing from $\kappa_0 = \kappa$ to 1. Sampling therefore begins in a near-isotropic regime, and only gradually recovers the original conditioning of the target as the trajectory approaches $t = 0$. This intuition is formalized in Theorem 3, showing that the sampling complexity of diffusion is indeed independent of κ.

2.2 Langevin samplers: the condition-number dependence

Table 1 summarizes our guarantees, with hyperparameters optimized for each method, and Figures 1 and 2 illustrate them. Proofs are given in the appendices. We first consider ULA and ULD. The ULA bound builds on the Gaussian analysis of Pedregosa (2023). The convergence rates for ULA and ULD differ only by numerical constants.



(a) Error ratio versus κ_{eff}, $N = 10^{12}$.

(b) ε-complexity versus κ, $\varepsilon = 10^{-3}\sqrt{\text{Tr}\,\Sigma}$.

Figure 2: Tightness of the theory in dimension $d = 1000$. **(a)** Ratio $W_2^{\text{Langevin}}/W_2^{\text{diffusion}}$ at budget $N = 10^{12}$, as a function of $\kappa_{\text{eff}} := \lambda_{\max}\sqrt{\text{Tr}(\Sigma^{-1})/\,\text{Tr}\,\Sigma}$, for four spectral families with $\lambda_{\max} = 1$ and $\kappa \in \{10, 10^2, \dots, 10^8\}$: "log-uniform" ($\lambda_i = \kappa^{-(i-1)/(d-1)}$); "power law" ($\lambda_i = i^{-a}$ with $a = \log\kappa/\log d$); "two-block $1/2$" (half of the eigenvalues equal to 1 and half to $1/\kappa$); "two-block $1/d$" (one eigenvalue equal to 1 and $d-1$ equal to $1/\kappa$). Langevin samplers are tuned as in Theorem 1; the diffusion uses the spectrum-agnostic schedule of Proposition 4 ($T = \sqrt{N}$, $\delta = 1/N^2$). Lines are the quotients of the asymptotic equivalents, $\frac{1}{5}\kappa_{\text{eff}}$ (ULA) and $\frac{2}{5}\kappa_{\text{eff}}$ (ULD), with no fitted constant: all four families collapse onto them within 3.4%, consistent with a finite-N correction of order $1/\log N \approx 3.6\%$. **(b)** Budget N_ε after which the error stays below $\varepsilon = 10^{-3}\sqrt{\text{Tr}\,\Sigma}$, where $\sqrt{\text{Tr}\,\Sigma} = W_2(\delta_0, \mathbb{P})$ is the scale of the target, for the log-uniform spectrum. Dashed: the sufficient budgets of Corollary 5 ($r = 10$ in Remark 6 for the diffusion), within a factor 3 of the observed ones. The diffusion budget does not depend on κ. The Langevin budgets follow the $\sqrt{\kappa}$ rate of Table 1 as long as $\varepsilon \lesssim \lambda_{\min}\sqrt{\text{Tr}(\Sigma^{-1})} = \sqrt{\text{Tr}\,\Sigma/\kappa}$; beyond, the burn-in $N_\star = \Theta(\kappa\log\kappa)$ of Remark 2 takes over and the budget grows linearly in κ. Kinks on the ULA curve mark the switch between the two branches of its step-size.

Theorem 1 (Langevin samplers)**.** *Let $N \geq 2$ be the iteration budget, and set the step-size η (and the friction γ for ULD) as*

- *ULA: $\eta = \frac{1}{2}\lambda_{\max}\frac{\log N}{N}$, provided $N \geq \frac{1}{2}\kappa\log N$;*
- *ULD: $\eta = \sqrt{2\lambda_{\max}\kappa}\,\frac{\log N}{N}$ and $\gamma = \sqrt{8/\lambda_{\min}}$, provided $N \geq 4\kappa\log N$.*

Then

$$W_2(\mathbb{P}_N, \mathbb{P}) \leq \frac{\lambda_{\max}\sqrt{\text{Tr}(\Sigma^{-1})}}{N}\left(C_1\log N + C_2\right),$$

with $C_1 = \frac{1}{4}$, $C_2 = \sqrt{2}$ for ULA and $C_1 = \frac{3}{7}$, $C_2 = 8$ for ULD. Furthermore, for fixed Σ, these bounds are tight up to a constant factor:

$$W_2(\mathbb{P}_N, \mathbb{P}) \underset{N\to\infty}{\sim} C_3\,\lambda_{\max}\sqrt{\text{Tr}(\Sigma^{-1})}\,\frac{\log N}{N},$$

with $C_3 = \frac{1}{8}$ for ULA and $C_3 = \frac{1}{4}$ for ULD.

Both Langevin samplers deteriorate with the condition number through a $\sqrt{\kappa}$ factor, which can be made explicit by noting that $\lambda_{\max}\sqrt{\text{Tr}(\Sigma^{-1})} = \sqrt{\lambda_{\max}}\sqrt{\sum_{i=1}^d \frac{\lambda_{\max}}{\lambda_i}} \in [\sqrt{\lambda_{\max}\kappa}, \sqrt{\lambda_{\max} d\kappa}]$. The proof of this theorem, given in Appendix B, exploits the Gaussian structure of the target, reducing the analysis to the eigenvalues of the covariance matrix, for which a closed-form expression can be obtained. This yields a Wasserstein error bound consisting of two terms: a discretization error,

arising from the discretization of the continuous-time Langevin dynamics, and an early-stopping error, resulting from terminating the algorithm at time $N\eta$ rather than letting it run indefinitely. The resulting hyperparameter choices and convergence rates therefore reflect a trade-off between these two errors.

The ratio $\phi = \eta/\gamma$ of ULD equals the step-size η of ULA. This is expected: for large friction, the position approximately follows the overdamped Langevin dynamics slowed down by a factor γ (Pavliotis, 2014), so that ϕ acts as an effective step-size. The small-friction regime yields the same rate (Corollary B.13): with the EM scheme, neither brings the acceleration expected from ULD.

Remark 2 (Admissibility and burn-in)**.** *The conditions on N in Theorem 1 are exactly the requirements under which our analysis operates, namely $\eta \le \lambda_{\min}$ for ULA and $\eta\gamma \le 1$ for ULD (Appendix B). They amount to a burn-in phase: writing*

$$N_\star^{\mathrm{ULA}} := \min\{n \ge 3 : n \ge \tfrac{1}{2}\kappa \log n\} \quad \textit{and} \quad N_\star^{\mathrm{ULD}} := \min\{n \ge 3 : n \ge 4\kappa \log n\},$$

provided $N \ge 3$, the condition on N for ULA (resp. ULD) holds if and only if $N \ge N_\star^{\mathrm{ULA}}$ (resp. $N \ge N_\star^{\mathrm{ULD}}$), and both thresholds grow as $\Theta(\kappa \log \kappa)$. The guarantee thus applies only after $\Theta(\kappa \log \kappa)$ iterations. Ill-conditioning penalizes Langevin samplers twice: through the $\sqrt{\kappa}$ factor in the rate, and through the budget required before the guarantee applies. Neither penalty has a counterpart for diffusion: Theorem 3 holds for every $N \ge 1$.

Figure 1 illustrates this burn-in. Below $N_\star$, the step-size of Theorem 1 violates the admissibility conditions of Remark 2. Since our bounds are convex (with respect to the variable η for ULA, $\phi := \eta/\gamma$ for ULD), their constrained minimizer is the largest admissible value, $\eta = \lambda_{\min}$ and $\eta\gamma = 1$ respectively (Appendix B). This step is dictated by the stiffest direction of the target, so the flattest ones barely move: the error stays at its initial value $W_2(\mathbb{P}_0, \mathbb{P}) = \sqrt{\mathrm{Tr}\,\Sigma}$ for $N \lesssim \kappa$, then decays like $e^{-cN/\kappa}$ ($c = 2$ for ULA, $c = 1/4$ for ULD) down to the discretization bias of that step, reached at $N \approx N_\star = \Theta(\kappa \log \kappa)$. Only then does the $\log N/N$ regime of Theorem 1 begin, whereas the diffusion sampler follows its asymptote from $N \approx 10^2$ onwards.

2.3 Diffusion models: circumventing the condition-number dependence

In contrast, diffusion models follow the scores s_t, whose conditioning κ_t tends to 1 at large times, and achieve rates that depend on neither κ nor $\lambda_{\min}$.

Theorem 3 (Diffusion sampler)**.** *Consider an iteration budget $N \ge 1$. Set horizon $T = (\mathrm{Tr}(\Sigma^5)/\,\mathrm{Tr}\,\Sigma)^{1/4}\sqrt{N}$, early-stopping threshold $\delta = \mathrm{Tr}\,\Sigma/(dN^2)$, and geometric time discretization grid $\tau_k = T(\delta/T)^{k/N}$ for $k \in \{0, \dots, N\}$. Then the diffusion algorithm* (6) *yields*

$$W_2(\mathbb{P}_N, \mathbb{P}) \le \frac{\sqrt{\mathrm{Tr}\,\Sigma}}{N}\Big(\tfrac{5}{8}\, a_N\, \rho^{1/N} \log(\rho^{2/5} N) + 1 + \sqrt{8}\Big), \quad \rho := \frac{d\lambda_{\max}}{\mathrm{Tr}\,\Sigma} \in [1, d],$$

where $a_N := (1 + 4N^{-2})^{1/2} N^{5/(2N)}$ satisfies $a_N \le 2^{7/4}$ for all $N \ge 1$ and $a_N \to 1$. Moreover, this bound is asymptotically exact: under the same hyperparameters,

$$W_2(\mathbb{P}_N, \mathbb{P}) \underset{N\to\infty}{\sim} \frac{5}{8}\sqrt{\mathrm{Tr}\,\Sigma}\,\frac{\log N}{N}, \tag{7}$$

so that, for fixed Σ, the ratio of the bound to the actual error tends to 1.

The previous theorem gives a spectrum-adaptive bound. The simplified $O(\sqrt{\mathrm{Tr}\,\Sigma}\log N/N)$ rate is obtained when d grows at most polynomially with N. The following proposition shows that this growth assumption can be removed at the price of a coarser dependence on the spectrum.

Proposition 4. *With the alternative choice $T = \lambda_{\max}\sqrt{N}$, early-stopping threshold $\delta = \lambda_{\max}/N^2$, and geometric time discretization grid $\tau_k = T(\delta/T)^{k/N}$ for $k \in \{0, \dots, N\}$, the diffusion algorithm* (6) *satisfies, for every $N \ge 1$,*

$$W_2(\mathbb{P}_N, \mathbb{P}) \le \frac{\sqrt{d\lambda_{\max}}}{N}\left(\frac{5}{2^{5/4}}\log N + \left(\frac{5}{4}\right)^{3/2} + 1\right).$$

On the one hand, Theorem 3 provides our sharpest, spectrum-adaptive guarantee, with a prefactor governed by the effective scale $\sqrt{\mathrm{Tr}\,\Sigma}$. However, achieving this sharp rate requires choosing the

horizon T and early-stopping threshold δ using spectral information about Σ, and its simplified $O\big(\sqrt{\operatorname{Tr}\Sigma}\frac{\log N}{N}\big)$ form is stated under the growth condition $d \le P(N)$ for some polynomial P. On the other hand, Proposition 4 holds unconditionally in d, at the price of a coarser prefactor. Since $\operatorname{Tr}\Sigma \le d\lambda_{\max}$ (with equality if and only if the target is isotropic), the coarse bound exceeds the spectral bound by a factor $\sqrt{\rho} \in [1, \sqrt{d}]$. Importantly, this gap is merely a finite-N phenomenon reflecting specific hyperparameter choices rather than a different asymptotic sampling regime. We show in Appendix C.4 that any geometric schedule with $T \propto \sqrt{N}$ and $\delta \propto N^{-2}$ yields the same asymptotic equivalent (7): the rate is robust to schedule tuning. This covers the parameter choices of both Theorem 3 and Proposition 4.

As in Theorem 1, the proof of Theorem 3 and Proposition 4, given in Appendix C, exploits the Gaussian structure of the target distribution. Since the resulting distribution $\mathbb{P}_N$ is also Gaussian, the analysis reduces to the eigenvalues of its covariance matrix, which can be characterized explicitly. This leads to a Wasserstein error bound with three distinct contributions: a discretization error, an early-stopping error controlled by the threshold δ, and an initialization error arising from starting the process at $\mathbb{P}_0 = \mathcal{N}(0, T\,\mathrm{I})$ rather than at the true marginal $\mathcal{N}(0, \Sigma + T\,\mathrm{I})$, which is not directly tractable. The resulting hyperparameter choices and convergence rates reflect a trade-off between these three sources of error. Consistently with the intuition of Section 2.1, the geometric grid $\tau_k = T(\delta/T)^{k/N}$ takes coarse steps while s_t remains close to isotropic, and only refines them as $t \to \delta$, when the anisotropy of the true score re-emerges. Such geometrically decreasing step-sizes near the data are standard in the analysis of diffusion models: the schedule of Benton et al. (2024), for instance, combines constant steps far from the data with steps decaying geometrically as $t \to 0$. We adopt the geometric grid over the whole horizon $[\delta, T]$, which makes the computations of Appendix C easier.

The rates above describe the sampling error for a fixed iteration budget N. Sampling guarantees are, however, more commonly stated in the reverse direction in the literature: given a target accuracy, how many iterations are required? We follow this convention below.

Corollary 5 (ε-complexity)**.** *Let $0 < \varepsilon \le \sqrt{\operatorname{Tr}\Sigma}$. In order for the sampling error to satisfy $W_2(\mathbb{P}_N, \mathbb{P}) \le \varepsilon$, it suffices*

- *for ULA with $\eta = \min\big(\lambda_{\min}, \varepsilon/\sqrt{\operatorname{Tr}(\Sigma^{-1})}\big)$, to take*

$$N \ge \frac{\kappa}{2}\max\left(1, \frac{\lambda_{\min}\sqrt{\operatorname{Tr}(\Sigma^{-1})}}{\varepsilon}\right)\log\frac{\sqrt{8\operatorname{Tr}\Sigma}}{\varepsilon},$$

- *for ULD with $\eta = \min\big(\sqrt{\lambda_{\min}/8}, \frac{7\sqrt{2}}{6}\varepsilon/\sqrt{\lambda_{\min}\operatorname{Tr}(\Sigma^{-1})}\big)$ and $\gamma = \sqrt{8/\lambda_{\min}}$, to take*

$$N \ge 4\kappa\max\left(1, \frac{3}{14}\frac{\lambda_{\min}\sqrt{\operatorname{Tr}(\Sigma^{-1})}}{\varepsilon}\right)\log\frac{16\sqrt{\operatorname{Tr}\Sigma}}{\varepsilon},$$

- *for the diffusion algorithm with $\delta = \varepsilon^2/(9d)$, $T = \sqrt{3/\varepsilon}\operatorname{Tr}\big((\Sigma + \delta\,\mathrm{I})^3\Sigma^2\big)^{1/4}$, and the geometric grid $\tau_k = T(\delta/T)^{k/N}$, to take*

$$N \ge \frac{3e}{4}\frac{\sqrt{\operatorname{Tr}\Sigma + 4\varepsilon^2/9}}{\varepsilon}\log\frac{T}{\delta}.$$

Using $\operatorname{Tr}\Sigma \le d\lambda_{\max}$ and $\operatorname{Tr}(\Sigma^{-1}) \le d/\lambda_{\min}$, we retrieve the ε-complexities of Table 1, *i.e.*, $\tilde{O}(\kappa + \sqrt{d\kappa\lambda_{\max}}/\varepsilon)$ for Langevin dynamics versus the condition-number-free $\tilde{O}(\sqrt{d\lambda_{\max}}/\varepsilon)$ for diffusion. The spectral form is more informative. For the Langevin samplers, the $1/\varepsilon$ rate only sets in once $\varepsilon \le \lambda_{\min}\sqrt{\operatorname{Tr}(\Sigma^{-1})}$ for ULA and $\varepsilon \le \frac{3}{14}\lambda_{\min}\sqrt{\operatorname{Tr}(\Sigma^{-1})}$ for ULD; above these thresholds, the budget is a burn-in of order $\kappa\log(\sqrt{\operatorname{Tr}\Sigma}/\varepsilon)$, in line with Remark 2. The diffusion sampler, on the contrary, has no burn-in: its budget is proportional to $1/\varepsilon$ for every $\varepsilon \le \sqrt{\operatorname{Tr}\Sigma}$, up to the factor $\log(T/\delta) = O(\log(\sqrt{\operatorname{Tr}\Sigma}/\varepsilon))$, which does not involve κ. Finally, the constants are better for ULA than for ULD; this reflects our proof technique rather than an intrinsic advantage, as we did not seek to optimize them.

Remark 6. *The restriction $\varepsilon \le \sqrt{\operatorname{Tr}\Sigma}$ is free: $\sqrt{\operatorname{Tr}\Sigma} = W_2(\delta_0, \mathbb{P})$ is the error of the trivial output 0, so larger error tolerances require no iterations at all. Moreover, the constant $3e/4$ of the diffusion budget can be traded for a mild burn-in. We show in Appendix D that, with the same δ, T and grid, it suffices to take, for any $r \ge 1$,*

$$N \ge r \max\left(1, \frac{3e^{1/r}}{4r} \frac{\sqrt{\operatorname{Tr}\Sigma + 4\varepsilon^2/9}}{\varepsilon}\right) \log \frac{T}{\delta}.$$

For $r = 1$ the maximum is always attained by its second argument, which gives the corollary. As r grows, the constant $3e^{1/r}/4$ decreases towards $3/4$, while the first argument, a κ-free burn-in of $r \log(T/\delta)$ iterations, prevails for $\varepsilon > c_r \sqrt{\operatorname{Tr}\Sigma}$, with $c_r = \frac{3}{2}\left(4r^2 e^{-2/r} - 1\right)^{-1/2}$.

3 Learning: the impact of noise on the implicit bias

The previous results assume the score of the target distribution—or, for diffusion, its noisy versions s_t—to be known exactly. In practice, these scores must first be learned from (noisy) samples. The goal of this section is to study whether the noise introduced in diffusion models leads to a different score estimator. We are given a dataset $\mathcal{D} = \{x_i\}_{1 \le i \le m}$ of m independent realizations of $\mathbb{P}$ and denote by $\hat{\Sigma} = \frac{1}{m}\sum_{i=1}^m x_i x_i^\top$ the empirical covariance matrix. Since the score of a Gaussian distribution is linear, we consider the score model

$$s_A : x \mapsto -Ax,$$

where $A \in \mathbb{R}^{d \times d}$, which recovers the true score s when $A = \Sigma^{-1}$. When $\hat{\Sigma}$ is invertible, which is almost surely the case if $m \ge d$, the score-matching loss (Hyvärinen, 2005)

$$L : A \mapsto \frac{1}{m}\sum_{i=1}^m \left(\nabla \cdot s_A(x_i) + \frac{1}{2}\|s_A(x_i)\|^2\right) \tag{8}$$

where $\|\cdot\|$ denotes the Euclidean norm, admits the unique minimizer $\hat{\Sigma}^{-1}$. In the overparameterized regime $m < d$, however, it has no minimizer as it diverges towards $-\infty$ along $\operatorname{Ker}\hat{\Sigma}$. In this regime, the direction of divergence followed by a given procedure is not determined by L alone but also depends on the optimization algorithm and its initialization, which is usually referred to as an *implicit bias* of the optimization algorithm (Gunasekar et al., 2018). Our goal is to compare the implicit biases of two algorithms. First, we study gradient descent initialized at $A_0 = 0$ and write A_K for the parameter obtained after $K \ge 1$ gradient steps on (8). We then compare with the iterate A_K^σ obtained by running K steps of gradient descent, from the same initialization, on a diffusion-style noised objective. Fix a noise level $\sigma > 0$ and let $\omega \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$. The noised objective is

$$L_\sigma : A \mapsto \frac{1}{m}\sum_{i=1}^m \mathbb{E}_\omega\left[\nabla \cdot s_A(x_i + \omega) + \frac{1}{2}\|s_A(x_i + \omega)\|^2\right].$$

As shown by Vincent (2011), this coincides, up to an A-independent constant, with the standard denoising score-matching regression objective. Additionally, a direct computation shows that

$$L_\sigma(A) = L(A) + \frac{\sigma^2}{2}\|A\|_F^2.$$

Thus, it is equivalent to adding a ridge-type regularization term.

We show that, in this Gaussian setting, noise injection does not fundamentally change the learned score: both procedures recover the empirical covariance $\hat{\Sigma}$ on its range, up to a shift of order σ^2, and differ only in how they regularize its kernel. The key takeaway is that the benefit of the noising process identified in the sampling phase has no counterpart in the learning phase, at least in the Gaussian setting considered here.

Since A plays the role of the precision matrix Σ^{-1} in the score model $s_A(x) = -Ax$, Theorem 7 below characterizes, up to an exponentially decaying error, the covariance matrices estimated by each procedure—namely A_K^{-1} and $(A_K^\sigma)^{-1}$. Let us denote by $\hat{\lambda}_{\max}$ the largest eigenvalue of $\hat{\Sigma}$, $\hat{\lambda}_{\min}$ its smallest eigenvalue, $\hat{\lambda}_{\min}^+$ its smallest positive eigenvalue and $\Pi_{\operatorname{Ker}\hat{\Sigma}}$ the orthogonal projector onto its kernel, of dimension $\max(0, d - m)$ almost surely.

Theorem 7 (Score learning)**.** *Consider gradient descent initialized at $A_0 = 0$, with step-size $\alpha > 0$, on both L and L_σ. Let $K \geq 1$.*

- *If $\alpha < \frac{2}{\hat{\lambda}_{\max}}$, the iterate A_K is invertible and*

$$A_K^{-1} = \hat{\Sigma} + \frac{1}{K\alpha}\Pi_{\operatorname{Ker}\hat{\Sigma}} + O(\rho_1^K),$$

  with $\rho_1 = \max\left(|1 - \alpha\hat{\lambda}^+_{\min}|, |1 - \alpha\hat{\lambda}_{\max}|\right) < 1$.

- *If $\alpha < \frac{2}{\hat{\lambda}_{\max}+\sigma^2}$, the iterate A_K^σ is invertible and*

$$(A_K^\sigma)^{-1} = \hat{\Sigma} + \sigma^2\,\mathrm{I} + O(\rho_2^K),$$

  with $\rho_2 = \max\left(|1 - \alpha(\hat{\lambda}_{\min} + \sigma^2)|, |1 - \alpha(\hat{\lambda}_{\max} + \sigma^2)|\right) < 1$.

The distinction between the two procedures is clearest when $\hat{\Sigma}$ is singular, that is, almost surely when $m < d$. For the standard score-matching loss L, the resulting non-identifiability is resolved implicitly by gradient descent: up to an exponentially small optimization error, A_K^{-1} coincides with $\hat{\Sigma}$ on its range, while a correction of order $1/(K\alpha)$ appears only in the kernel of $\hat{\Sigma}$. In contrast, the noised objective L_σ induces an explicit ridge regularization, and this translates into a uniform covariance shift $\sigma^2\,\mathrm{I}$: all directions are regularized by the same amount. Thus the two procedures recover a regularized empirical covariance matrix, but differ slightly in the regularization: the implicit bias of L acts only in directions that are not identified by the data, whereas noise injection regularizes all directions indiscriminately. Yet, in order to recover the true covariance directions, one should take $\sigma^2 \ll \hat{\lambda}^+_{\min}$ and, respectively, $1/(K\alpha) \ll \hat{\lambda}^+_{\min}$. In this regime, both regularizations are essentially identical, meaning that the noising process does not help regularize the estimated score compared to learning the original score (with early stopping). We finally note that this constitutes an instance where the gradient descent path follows the explicit ridge regularization path, which is a structure appearing in many contexts (Ali et al., 2020).

4 Conclusion

This work demonstrates an advantage of modeling the score at various levels of noise in the Gaussian setting: more efficient sampling by gradually adding anisotropy. This benefit is specific to sampling: in the same setting, the score learned with noise coincides, up to the regularization of the empirical nullspace, with the one learned without. Our sampling results take the form of scaling laws, both for the sampling error and for the hyperparameters that achieve it. We emphasize that, despite its simplicity, the Gaussian framework enables a rigorous proof by characterizing the exact dependence on the condition number. Weaker assumptions, *e.g.*, log-concavity, lead to loose dependence on the condition number, which is insufficient for our purpose.

Natural directions for future work include incorporating the score estimation error into the sampling analysis, extending the comparison beyond the Gaussian setting, and investigating whether variable step-sizes could benefit Langevin samplers as well.

Reproducibility Statement

The figures presented in Section 2 are fully reproducible without stochastic computation. The Wasserstein distances are computed in closed form using the exact expressions of the eigenvalues of the covariance matrices, which are presented in Appendices B.2.1, B.3.1 and C.2.1. Consequently, no random seed or stochastic procedure is involved in obtaining the reported results. The code used to generate the figures will be made publicly available in the future.

References

Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 233–244, 2020.

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Simon Apers, Sander Gribling, and Dániel Szilágyi. Hamiltonian Monte Carlo for efficient Gaussian sampling: long and random steps. *Journal of Machine Learning Research*, 25(348):1–30, 2024.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on Learning Representations (ICLR)*, 2024.

Eliot Beyler and Francis Bach. Convergence of deterministic and stochastic diffusion-model samplers: A simple analysis in Wasserstein distance. *arXiv preprint arXiv:2508.03210*, 2025.

Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: a non-asymptotic analysis. In *Conference on Learning Theory (COLT)*, pages 300–323, 2018.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):651–676, 2017.

Arnak S. Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 8780–8794, 2021.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987.

Matthias Gelbrich. On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Nicolaï Gouraud, Pierre Le Bris, Adrien Majka, and Pierre Monmarché. HMC and underdamped Langevin united in the unadjusted convex smooth case. *SIAM/ASA Journal on Uncertainty Quantification*, 13(1):278–303, 2025.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning (ICML)*, 2018.

Florentin Guth, Simon Coste, Valentin De Bortoli, and Stéphane Mallat. Wavelet score-based generative modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 478–491, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *International Conference on Machine Learning (ICML)*, pages 8867–8887, 2022.

Samuel Hurault, Matthieu Terris, Thomas Moreau, and Gabriel Peyré. From score matching to diffusion: A fine-grained error analysis in the Gaussian setting. *arXiv preprint arXiv:2503.11615*, 2025.

Samuel Hurault, Thomas Moreau, and Gabriel Peyré. Geometry-aware discretization error of diffusion models. *arXiv preprint arXiv:2605.08392*, 2026.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Eliahu I. Jury. A simplified stability criterion for linear discrete systems. *Proceedings of the IRE*, 50(6):1493–1500, 1962.

Kyurae Kim, Samuel Gruffaz, Ji Won Park, and Alain Oliviero Durmus. Analysis of kinetic Langevin Monte Carlo under the stochastic exponential Euler discretization from underdamped all the way to overdamped. *Electronic Journal of Statistics*, 20(2):3106–3142, 2026.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

Ben Leimkuhler and Charles Matthews. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*, volume 39 of *Interdisciplinary Applied Mathematics*. Springer, 2015.

Benedict Leimkuhler, Daniel Paulin, and Peter A. Whalley. Contraction and convergence rates for discretized kinetic Langevin dynamics. *SIAM Journal on Numerical Analysis*, 62(3):1226–1258, 2024.

Grigorios A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations*, volume 60 of *Texts in Applied Mathematics*. Springer, 2014.

Fabian Pedregosa. On the convergence of the unadjusted Langevin algorithm. `http://fa.bianp.net/blog/2023/ulaq/`, 2023.

Emile Pierret and Bruno Galerne. Diffusion models for Gaussian distributions: Exact solutions and Wasserstein errors. In *International Conference on Machine Learning (ICML)*, pages 49355–49381, 2025.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pages 2256–2265, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Stanislas Strasman, Antonio Ocello, Claire Boyer, Sylvain Le Corff, and Vincent Lemaire. An analysis of the noise schedule for score-based generative models. *Transactions on Machine Learning Research*, 2025.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory (COLT)*, 2018.

TABLE OF CONTENTS

A Preliminaries **12**

B Proof of Theorem 1 (Langevin samplers) **14**

B.1 Overview 14

B.2 Unadjusted Langevin algorithm 15

B.3 Underdamped Langevin dynamics 19

C Proof of Theorem 3 and Proposition 4 (diffusion sampler) **26**

C.1 Overview 26

C.2 Parametric bound for the diffusion sampler 27

C.3 Choice of hyperparameters 31

C.4 Asymptotic equivalent 33

D Proof of Corollary 5 (ε-complexity) **34**

E Proof of Theorem 7 (score learning) **37**

E.1 Losses and gradients 37

E.2 Closed form of the iterates 38

E.3 Inverses of the iterates 39

LLM usage **41**

A PRELIMINARIES

We start with a lemma that serves for all three algorithms. Throughout, a deterministic random vector equal to 0 is regarded as a degenerate centered Gaussian vector with a vanishing covariance matrix; this convention lets us handle the zero initializations of ULA and ULD without special treatment.

Lemma A.1. *Let P be an orthogonal matrix diagonalizing Σ, i.e., $\Sigma = P\Lambda P^\top$ with $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$, and let $\mathcal{A}(\Sigma)$ denote the set of matrices that are diagonal in the eigenbasis of Σ:*

$$\mathcal{A}(\Sigma) = \{PDP^\top,\ D \text{ diagonal}\}.$$

For $M \in \mathcal{A}(\Sigma)$, we write $[M]_i$ for its i-th diagonal entry in this basis. Let $(A_k^{xx}, A_k^{xv}, A_k^{vx}, A_k^{vv}, B_k^x, B_k^v)_{k\geq 0}$ be elements of $\mathcal{A}(\Sigma)$, (ξ_k) a sequence of independent standard Gaussian vectors in $\mathbb{R}^d$, and (X_0, V_0) a centered, possibly degenerate, Gaussian pair, independent of (ξ_k), whose covariance blocks $\mathbb{E}[X_0X_0^\top]$, $\mathbb{E}[V_0V_0^\top]$ and $\mathbb{E}[X_0V_0^\top]$ all belong to $\mathcal{A}(\Sigma)$. Consider the process (X_k, V_k) defined by

$$\forall k \geq 0, \begin{cases} X_{k+1} = A_k^{xx}X_k + A_k^{xv}V_k + B_k^x\xi_k, \\ V_{k+1} = A_k^{vx}X_k + A_k^{vv}V_k + B_k^v\xi_k, \end{cases} \tag{9}$$

and set $\Sigma_k := \mathbb{E}[X_kX_k^\top]$, $S_k := \mathbb{E}[V_kV_k^\top]$ and $\Gamma_k := \mathbb{E}[X_kV_k^\top]$. Finally, for $1 \leq i \leq d$, denote by $x_k^{(i)}$, $v_k^{(i)}$ and $\zeta_k^{(i)}$ the i-th coordinates of $P^\top X_k$, $P^\top V_k$ and $P^\top\xi_k$ respectively. Then:

(i) (X_k, V_k) is a centered Gaussian vector, and Σ_k, S_k and Γ_k belong to $\mathcal{A}(\Sigma)$.

(ii) The d pairs $(x_k^{(i)}, v_k^{(i)})$, $1 \le i \le d$, are mutually independent, and each of them follows the scalar recursion

$$\forall k \ge 0, \begin{cases} x_{k+1}^{(i)} = [A_k^{xx}]_i x_k^{(i)} + [A_k^{xv}]_i v_k^{(i)} + [B_k^x]_i \zeta_k^{(i)}, \\ v_{k+1}^{(i)} = [A_k^{vx}]_i x_k^{(i)} + [A_k^{vv}]_i v_k^{(i)} + [B_k^v]_i \zeta_k^{(i)}, \end{cases}$$

where the $(\zeta_k^{(i)})_{k,i}$ are mutually independent standard one-dimensional Gaussian variables, independent of (X_0, V_0).

(iii) The eigenvalues $\lambda_i^{(k)}$ of Σ_k are the variances of the variables $x_k^{(i)}$:

$$\lambda_i^{(k)} = \mathrm{Var}(x_k^{(i)}).$$

Remark A.2. *ULA* (2)*, ULD* (4) *and the diffusion sampler* (6) *are all instances of the process (X_k, V_k) above. For ULD the identification is immediate: $A_k^{xx} = \mathrm{I}$, $A_k^{xv} = \eta\,\mathrm{I}$, $B_k^x = 0$, $A_k^{vx} = -\eta\Sigma^{-1}$, $A_k^{vv} = (1-\eta\gamma)\,\mathrm{I}$ and $B_k^v = \sqrt{2\eta\gamma}\,\mathrm{I}$. ULA and the diffusion have no velocity component, but they fit the same framework upon taking $V_0 = 0$ and $A_k^{xv} = A_k^{vx} = A_k^{vv} = B_k^v = 0$. ULA then corresponds to $A_k^{xx} = \mathrm{I} - \eta\Sigma^{-1}$ and $B_k^x = \sqrt{2\eta}\,\mathrm{I}$, and the diffusion to $A_k^{xx} = \mathrm{I} - \eta_{k+1} C_{\tau_k}^{-1}$ and $B_k^x = \sqrt{\eta_{k+1}}\,\mathrm{I}$, the covariance $T\,\mathrm{I}$ of its initialization $\mathcal{N}(0, T\,\mathrm{I})$ being indeed in $\mathcal{A}(\Sigma)$.*

Proof of Lemma A.1. We first prove (ii). For every $k \ge 0$, set $\tilde{X}_k := P^\top X_k$, $\tilde{V}_k := P^\top V_k$ and $\tilde{\xi}_k := P^\top \xi_k$. Since P is orthogonal, $(\tilde{\xi}_k)_{k\ge 0}$ is again a sequence of independent standard Gaussian vectors, independent of (X_0, V_0); hence the family $(\zeta_k^{(i)})_{k\ge 0,\, 1\le i\le d}$ consists of mutually independent $\mathcal{N}(0,1)$ variables, independent of (X_0, V_0), which is the last claim of (ii). Multiplying (9) by $P^\top$ and inserting $PP^\top = \mathrm{I}$ in front of X_k, V_k and ξ_k gives

$$\begin{cases} \tilde{X}_{k+1} = P^\top A_k^{xx} P \tilde{X}_k + P^\top A_k^{xv} P \tilde{V}_k + P^\top B_k^x P \tilde{\xi}_k, \\ \tilde{V}_{k+1} = P^\top A_k^{vx} P \tilde{X}_k + P^\top A_k^{vv} P \tilde{V}_k + P^\top B_k^v P \tilde{\xi}_k. \end{cases}$$

By definition of $\mathcal{A}(\Sigma)$, every matrix in this system is diagonal, so reading it coordinate by coordinate yields exactly the announced recursion

$$\begin{cases} x_{k+1}^{(i)} = [A_k^{xx}]_i x_k^{(i)} + [A_k^{xv}]_i v_k^{(i)} + [B_k^x]_i \zeta_k^{(i)}, \\ v_{k+1}^{(i)} = [A_k^{vx}]_i x_k^{(i)} + [A_k^{vv}]_i v_k^{(i)} + [B_k^v]_i \zeta_k^{(i)}, \end{cases} \qquad 1 \le i \le d.$$

It remains to show that the d pairs $(x_k^{(i)}, v_k^{(i)})$ are independent, which we do by induction on k.

Base case. $(\tilde{X}_0, \tilde{V}_0)$ is a centered Gaussian vector, as (X_0, V_0) is, and its covariance matrix

$$\begin{bmatrix} P^\top \Sigma_0 P & P^\top \Gamma_0 P \\ P^\top \Gamma_0^\top P & P^\top S_0 P \end{bmatrix}$$

has four diagonal blocks, since Σ_0, S_0 and Γ_0 belong to $\mathcal{A}(\Sigma)$. The pairs $(x_0^{(i)}, v_0^{(i)})_{1\le i\le d}$ are therefore uncorrelated and, being jointly Gaussian, independent.

Induction step. Assume the pairs $(x_k^{(i)}, v_k^{(i)})_{1\le i\le d}$ are independent for some $k \ge 0$. The vector ξ_k is independent of X_0, V_0 and $(\xi_j)_{j<k}$, of which (X_k, V_k) is a deterministic function; hence ξ_k is independent of (X_k, V_k), and $\tilde{\xi}_k$ of $(\tilde{X}_k, \tilde{V}_k)$. As the coordinates $\zeta_k^{(1)}, \dots, \zeta_k^{(d)}$ of $\tilde{\xi}_k$ are themselves independent, the d triplets $(x_k^{(i)}, v_k^{(i)}, \zeta_k^{(i)})$ are mutually independent. Each pair $(x_{k+1}^{(i)}, v_{k+1}^{(i)})$ is a deterministic function of the i-th triplet, so the pairs at step $k+1$ are independent as well. This completes the induction, and the proof of (ii).

We now turn to (i). The vector (X_{k+1}, V_{k+1}) is a linear function of (X_k, V_k, ξ_k). If (X_k, V_k) is a centered Gaussian vector, so is (X_k, V_k, ξ_k), because ξ_k is centered Gaussian and independent of (X_k, V_k) (see the proof of (ii)); hence (X_{k+1}, V_{k+1}) is centered Gaussian too, and the claim follows by induction from the assumption on (X_0, V_0). As for the covariance matrices, Σ_k belongs to $\mathcal{A}(\Sigma)$ if and only if $P^\top \Sigma_k P$ is diagonal. But $P^\top \Sigma_k P$ is the covariance matrix of $\tilde{X}_k$, whose (i,j) entry is $\mathrm{Cov}(x_k^{(i)}, x_k^{(j)})$, which vanishes for $i \ne j$ by (ii). Thus $\Sigma_k \in \mathcal{A}(\Sigma)$, and the same argument applies to S_k and Γ_k.

Finally, (iii) is a direct consequence of the above: the eigenvalues of Σ_k are the diagonal entries of $P^\top \Sigma_k P$, namely the variances $\mathrm{Var}(x_k^{(i)})$. □

We also rely on the following elementary fact, which expresses the Wasserstein distance between two centered Gaussians whose covariances are diagonal in the same orthonormal basis in terms of their eigenvalues.

Lemma A.3. *Let $\Sigma_1, \Sigma_2 \in \mathcal{A}(\Sigma)$ be positive semi-definite, with diagonal entries (hence eigenvalues) $(a_i)_{1\le i\le d}$ and $(b_i)_{1\le i\le d}$ in the basis P. Then*

$$W_2\big(\mathcal{N}(0,\Sigma_1),\mathcal{N}(0,\Sigma_2)\big) = \left(\sum_{i=1}^{d}\left(\sqrt{a_i}-\sqrt{b_i}\right)^2\right)^{1/2}.$$

Proof. Since Σ_1 and Σ_2 belong to $\mathcal{A}(\Sigma)$, they commute, and the formula recalled in Section 2.1 applies (Gelbrich, 1990):

$$W_2(\mathcal{N}(0,\Sigma_1),\mathcal{N}(0,\Sigma_2)) = \|\Sigma_1^{1/2}-\Sigma_2^{1/2}\|_F.$$

Now $\Sigma_1^{1/2} = P\,\mathrm{diag}(\sqrt{a_1},\dots,\sqrt{a_d})\,P^\top$ and $\Sigma_2^{1/2} = P\,\mathrm{diag}(\sqrt{b_1},\dots,\sqrt{b_d})\,P^\top$, and the Frobenius norm is invariant under orthogonal conjugation, whence the result. □

Finally, the following elementary lemma is what we use to optimize the bounds of Propositions B.1 and B.7.

Lemma A.4. *Let A, B, C and D be positive constants, and let*

$$\begin{array}{rccc} f:(0,D] & \longrightarrow & \mathbb{R}, \\ x & \longmapsto & Ae^{-Bx}+Cx. \end{array}$$

Then f is strictly convex and, provided $0<\log\frac{AB}{C}\le BD$, it admits a unique minimizer,

$$x_* = \frac{1}{B}\log\frac{AB}{C}, \quad \textit{with} \quad f(x_*) = \frac{C}{B}\left(1+\log\frac{AB}{C}\right).$$

Proof. Strict convexity follows from $f''>0$ on $(0,D]$. Solving $f'(x)=0$ gives x_*, the condition $0<\log\frac{AB}{C}\le BD$ is exactly what ensures $0<x_*\le D$, and the value of $f(x_*)$ follows by substitution. □

B Proof of Theorem 1 (Langevin samplers)

B.1 Overview

This section is devoted to the proof of the following theorem, stated in Section 2.2.

Theorem 1 (Langevin samplers)**.** *Let $N\ge 2$ be the iteration budget, and set the step-size η (and the friction γ for ULD) as*

- *ULA: $\eta = \frac{1}{2}\lambda_{\max}\frac{\log N}{N}$, provided $N\ge\frac{1}{2}\kappa\log N$;*
- *ULD: $\eta=\sqrt{2\lambda_{\max}\kappa}\,\frac{\log N}{N}$ and $\gamma=\sqrt{8/\lambda_{\min}}$, provided $N\ge 4\kappa\log N$.*

Then

$$W_2(\mathbb{P}_N,\mathbb{P}) \le \frac{\lambda_{\max}\sqrt{\mathrm{Tr}(\Sigma^{-1})}}{N}\left(C_1\log N + C_2\right),$$

with $C_1=\frac{1}{4}$, $C_2=\sqrt{2}$ for ULA and $C_1=\frac{3}{7}$, $C_2=8$ for ULD. Furthermore, for fixed Σ, these bounds are tight up to a constant factor:

$$W_2(\mathbb{P}_N,\mathbb{P}) \underset{N\to\infty}{\sim} C_3\,\lambda_{\max}\sqrt{\mathrm{Tr}(\Sigma^{-1})}\,\frac{\log N}{N},$$

with $C_3=\frac{1}{8}$ for ULA and $C_3=\frac{1}{4}$ for ULD.

Error decomposition. ULA and ULD define time-homogeneous Markov chains. Under stability conditions on the hyperparameters, established in step (i) below, these chains converge, not to the target $\mathbb{P}$, but to an invariant law $\mathbb{P}_\infty$ that depends on the step-size η (and on the friction γ for ULD). The triangle inequality

$$W_2(\mathbb{P}_N, \mathbb{P}) \le W_2(\mathbb{P}_N, \mathbb{P}_\infty) + W_2(\mathbb{P}_\infty, \mathbb{P}) \tag{10}$$

separates two contributions. The first term is the early-stopping error: after N iterations, the chain has not yet reached its invariant law. The second is the discretization error: the bias of $\mathbb{P}_\infty$ with respect to $\mathbb{P}$, introduced by the Euler–Maruyama scheme.

Reduction to eigenvalues. Lemma A.1 applies to both algorithms (Remark A.2): each $\mathbb{P}_N$ is a centered Gaussian whose covariance Σ_N lies in $\mathcal{A}(\Sigma)$, and so does the covariance Σ_∞ of $\mathbb{P}_\infty$, the limit of (Σ_N). By Lemma A.3, both terms in (10) then depend only on the eigenvalues $\lambda_i^{(N)}$ of Σ_N, λ_i^∞ of Σ_∞ and λ_i of Σ:

$$W_2^2(\mathbb{P}_N, \mathbb{P}_\infty) = \sum_{i=1}^d \left(\sqrt{\lambda_i^{(N)}} - \sqrt{\lambda_i^\infty}\right)^2, \quad W_2^2(\mathbb{P}_\infty, \mathbb{P}) = \sum_{i=1}^d \left(\sqrt{\lambda_i^\infty} - \sqrt{\lambda_i}\right)^2. \tag{11}$$

This is the payoff of the Gaussian setting: the analysis reduces to real sequences for which closed forms are available.

Throughout, a choice of hyperparameters is called *admissible* if it lies in the stability range of the chain, shrunk by numerical constants that give the estimates some slack, keeping the denominators involved bounded away from zero. Remarks B.4 and B.11 make the gap with the exact stability conditions explicit.

Outline. Sections B.2 (ULA) and B.3 (ULD) follow the same four steps.

(i) *Closed form.* Compute the eigenvalues $\lambda_i^{(N)}$ and λ_i^∞ from the scalar recursions of Lemma A.1, and identify the stability conditions under which (Σ_N) converges.

(ii) *Parametric bound.* Bound each of the two terms in (10) and sum: this gives a bound on $W_2(\mathbb{P}_N, \mathbb{P})$ valid over the whole admissible range identified in (i) (Proposition B.1 for ULA, Proposition B.7 for ULD).

(iii) *Choice of hyperparameters.* Minimize, up to a numerical constant, the bound obtained in (ii), and check that the selected values satisfy the stability conditions as soon as $N \ge N_\star$.

(iv) *Asymptotic equivalent.* For the hyperparameters of step (iii), derive an exact equivalent of $W_2(\mathbb{P}_N, \mathbb{P})$ as $N \to \infty$ from the closed forms of step (i). This equivalent has the same order $\log N/N$ and the same spectral prefactor as the bound of step (iii), up to a numerical constant: the bound is of the right order of magnitude.

B.2 Unadjusted Langevin algorithm

The goal of this section is to prove the following proposition, which bounds $W_2(\mathbb{P}_N, \mathbb{P})$ for every small enough step-size η. To lighten notation, we drop the index i whenever we argue at a fixed eigendirection i.

Proposition B.1 (Parametric bound for ULA)**.** *For any step-size $0 < \eta \le \lambda_{\min}$ and any $N \ge 0$, the error of ULA satisfies*

$$W_2(\mathbb{P}_N, \mathbb{P}) \le \sqrt{2 \operatorname{Tr} \Sigma}\, e^{-2\eta N/\lambda_{\max}} + \frac{1}{2}\sqrt{\operatorname{Tr}(\Sigma^{-1})}\, \eta.$$

The two terms of this bound come from bounding the early-stopping error and the discretization error respectively. The first decays exponentially in $N\eta$, the time at which the process is stopped, while the second grows linearly with the step-size η. The step-size selected in Theorem 1 is the best compromise between these two competing effects.

B.2.1 CLOSED FORM OF THE EIGENVALUES

We first establish a closed form for the eigenvalues $\lambda_i^{(N)}$ and their limit λ_i^∞.

Lemma B.2. *Assume* $0 < \eta < 2\lambda_{\min}$ *and set*

$$\lambda_i^\infty := \frac{\lambda_i}{1 - \frac{\eta}{2}\lambda_i^{-1}}.$$

Then, for every $N \geq 0$*,*

$$\lambda_i^{(N)} = \lambda_i^\infty \left(1 - \left(1 - \frac{\eta}{\lambda_i}\right)^{2N}\right),$$

and in particular $\lambda_i^{(N)} \to \lambda_i^\infty$ *as* $N \to \infty$*.*

Proof. Fix a direction i and consider the corresponding sequence $(\lambda^{(k)})$ of eigenvalues of Σ_k (the index i is dropped, as announced). By Lemma A.1(iii) and the identification of Remark A.2, $\lambda^{(k)}$ is the variance of x_k, where $x_0 = 0$ and

$$\forall k \geq 0, \quad x_{k+1} = \left(1 - \frac{\eta}{\lambda}\right) x_k + \sqrt{2\eta}\,\zeta_k,$$

with (ζ_k) independent standard one-dimensional Gaussians. Since x_k is a function of $(\zeta_j)_{j<k}$, it is independent of ζ_k, and taking variances gives $\lambda^{(0)} = 0$ and

$$\forall k \geq 0, \quad \lambda^{(k+1)} = \left(1 - \frac{\eta}{\lambda}\right)^2 \lambda^{(k)} + 2\eta,$$

a first-order affine recursion. Its ratio satisfies

$$\left(1 - \frac{\eta}{\lambda}\right)^2 < 1, \tag{12}$$

because $0 < \eta < 2\lambda_{\min}$ gives $0 < \eta/\lambda < 2$, *i.e.*, $-1 < 1 - \eta/\lambda < 1$. In particular the ratio is not 1, so the recursion has a unique fixed point

$$\ell = \frac{2\eta}{1 - (1 - \eta\lambda^{-1})^2} = \frac{2\eta}{2\eta\lambda^{-1} - \eta^2\lambda^{-2}} = \frac{\lambda}{1 - \frac{\eta}{2}\lambda^{-1}} = \lambda^\infty,$$

and

$$\forall k \geq 0, \quad \lambda^{(k)} = \left(1 - \frac{\eta}{\lambda}\right)^{2k} (\lambda^{(0)} - \ell) + \ell = \lambda^\infty \left(1 - \left(1 - \frac{\eta}{\lambda}\right)^{2k}\right).$$

Taking $k = N$ gives the closed form, and (12) gives the convergence $\lambda^{(N)} \to \lambda^\infty$. Observe that $\lambda^\infty \to \lambda$ as $\eta \to 0$: the finer the Euler–Maruyama discretization, the smaller the bias. □

B.2.2 EARLY-STOPPING ERROR

We now bound the early-stopping error of ULA.

Lemma B.3. *Assume* $0 < \eta \leq \lambda_{\min}$*. Then, for every* $N \geq 0$*,*

$$W_2(\mathbb{P}_N, \mathbb{P}_\infty) \leq \sqrt{2\,\mathrm{Tr}\,\Sigma}\, e^{-2\eta N/\lambda_{\max}}.$$

Proof. Starting from the first identity in (11) and writing $\sqrt{a} - \sqrt{b} = (a-b)/(\sqrt{a} + \sqrt{b})$,

$$W_2^2(\mathbb{P}_N, \mathbb{P}_\infty) = \sum_{i=1}^d \left(\frac{\lambda_i^\infty - \lambda_i^{(N)}}{\sqrt{\lambda_i^\infty} + \sqrt{\lambda_i^{(N)}}}\right)^2 \leq \sum_{i=1}^d \frac{\left(\lambda_i^\infty - \lambda_i^{(N)}\right)^2}{\lambda_i^\infty}. \tag{13}$$

Since $\lambda^{(N)} \to \lambda^\infty$, this step only costs a factor 2 asymptotically (on W_2 itself, once the square root is taken). By Lemma B.2, $\lambda^\infty - \lambda^{(N)} = \lambda^\infty \left(1 - \frac{\eta}{\lambda}\right)^{2N}$, hence

$$\frac{(\lambda^\infty - \lambda^{(N)})^2}{\lambda^\infty} = \lambda^\infty \left(1 - \frac{\eta}{\lambda}\right)^{4N}.$$

The assumption $\eta \leq \lambda_{\min}$ gives both $0 \leq 1 - \frac{\eta}{\lambda} \leq 1 - \frac{\eta}{\lambda_{\max}}$ and $1 - \frac{\eta}{2}\lambda^{-1} \geq \frac{1}{2}$, *i.e.*, $0 < \lambda^\infty \leq 2\lambda$, so that

$$\frac{(\lambda^\infty - \lambda^{(N)})^2}{\lambda^\infty} \leq 2\lambda \left(1 - \frac{\eta}{\lambda_{\max}}\right)^{4N}.$$

Summing over i and plugging into (13) yields

$$W_2(\mathbb{P}_N, \mathbb{P}_\infty) \leq \sqrt{2 \operatorname{Tr} \Sigma} \left(1 - \frac{\eta}{\lambda_{\max}}\right)^{2N} \leq \sqrt{2 \operatorname{Tr} \Sigma}\, e^{-2\eta N/\lambda_{\max}},$$

where the last inequality uses $1 - u \leq e^{-u}$ together with the positivity of $1 - \eta/\lambda_{\max}$. □

Remark B.4. *Lemma B.2 only requires $\eta < 2\lambda_{\min}$, which is what the convergence of $\lambda^{(N)}$ to λ^∞ needs, whereas Lemma B.3 assumes $\eta \leq \lambda_{\min}$. The latter is a convenience assumption: it keeps $1 - \eta\lambda^{-1}$ nonnegative and, more importantly, keeps the denominator $1 - \frac{\eta}{2}\lambda^{-1}$ bounded away from zero. Since η is meant to go to 0 with N to reduce the discretization error, this is a mild restriction.*

B.2.3 Discretization error

We turn to the discretization error.

Lemma B.5. *Assume $0 < \eta \leq \lambda_{\min}$. Then*

$$W_2(\mathbb{P}_\infty, \mathbb{P}) \leq \frac{1}{2}\sqrt{\operatorname{Tr}(\Sigma^{-1})}\, \eta.$$

Proof. This time we start from the second identity in (11):

$$W_2^2(\mathbb{P}_\infty, \mathbb{P}) = \sum_{i=1}^d \left(\frac{\lambda_i^\infty - \lambda_i}{\sqrt{\lambda_i^\infty} + \sqrt{\lambda_i}}\right)^2 \leq \frac{1}{4}\sum_{i=1}^d \frac{(\lambda_i^\infty - \lambda_i)^2}{\lambda_i}, \tag{14}$$

where the inequality uses

$$\lambda^\infty > \lambda, \tag{15}$$

a consequence of $1 - \frac{\eta}{2}\lambda^{-1} < 1$. By Lemma B.2,

$$\lambda^\infty - \lambda = \lambda\left(\frac{1}{1 - \frac{\eta}{2}\lambda^{-1}} - 1\right) = \frac{\eta}{2 - \eta\lambda^{-1}} \leq \eta,$$

where we used $\eta \leq \lambda_{\min}$ once more. Since $\lambda^\infty - \lambda > 0$ by (15), this bound can be squared, and (14) gives the result. □

Proof of Proposition B.1. Combine Lemmas B.3 and B.5 through the triangle inequality (10). □

B.2.4 Choice of the step-size

We now prove the first part of Theorem 1, namely the non-asymptotic bound for ULA.

Proof of Theorem 1, non-asymptotic bound for ULA. Let $\eta_N := \frac{1}{2}\lambda_{\max}\frac{\log N}{N}$ be the ULA step-size of Theorem 1, and let $N \geq \frac{1}{2}\kappa \log N$. This condition is exactly $\eta_N \leq \lambda_{\min}$, and since $N > 1$ we have $0 < \eta_N \leq \lambda_{\min}$. Proposition B.1 therefore applies with $\eta = \eta_N$, and since $e^{-2\eta_N N/\lambda_{\max}} = 1/N$,

$$W_2(\mathbb{P}_N, \mathbb{P}) \leq \frac{\sqrt{2 \operatorname{Tr} \Sigma}}{N} + \frac{1}{4}\lambda_{\max}\sqrt{\operatorname{Tr}(\Sigma^{-1})}\frac{\log N}{N}.$$

Finally,

$$\sqrt{\operatorname{Tr} \Sigma} \leq \lambda_{\max}\sqrt{\operatorname{Tr}(\Sigma^{-1})}, \tag{16}$$

since the left-hand side is at most $\sqrt{d\lambda_{\max}}$ while the right-hand side is at least $\lambda_{\max}\sqrt{d/\lambda_{\max}} = \sqrt{d\lambda_{\max}}$. This gives the bound of Theorem 1,

$$W_2(\mathbb{P}_N, \mathbb{P}) \leq \frac{\lambda_{\max}\sqrt{\operatorname{Tr}(\Sigma^{-1})}}{N}\left(C_1 \log N + C_2\right),$$

with $C_1 = \frac{1}{4}$ and $C_2 = \sqrt{2}$. □

Remark B.6 (On the choice of η)**.** *The step-size η_N is not the exact minimizer of the bound of Proposition B.1, but an asymptotically equivalent simplification of it. Indeed, applying Lemma A.4 with*

$$A = \sqrt{2\,\mathrm{Tr}\,\Sigma}, \quad B = \frac{2N}{\lambda_{\max}}, \quad C = \frac{1}{2}\sqrt{\mathrm{Tr}(\Sigma^{-1})}, \quad D = \lambda_{\min},$$

the infimum of that bound is attained at

$$\eta_* = \frac{\lambda_{\max}}{2N}\log\frac{2NA}{\lambda_{\max}C} = \frac{\lambda_{\max}}{2N}\left(\log N + \log\left(\frac{4\sqrt{2}}{\lambda_{\max}}\sqrt{\frac{\mathrm{Tr}\,\Sigma}{\mathrm{Tr}(\Sigma^{-1})}}\right)\right),$$

which is admissible as soon as $2NA > \lambda_{\max}C$ and $2N \geq \kappa \log\frac{2NA}{\lambda_{\max}C}$, and the minimum equals

$$f(\eta_*) = \frac{\lambda_{\max}C}{2N}\Big(1 + \log\frac{2NA}{\lambda_{\max}C}\Big).$$

The choice η_N amounts to keeping only the $\log N$ term in the parentheses. As a result, $f(\eta_N)$ and $f(\eta_*)$ share the leading term $\frac{\lambda_{\max}C}{2}\frac{\log N}{N}$ and differ by an $O(1/N)$; in particular $f(\eta_N)/f(\eta_*) \to 1$. The only loss is on the constant C_2, in exchange for a step-size that requires no spectral information beyond $\lambda_{\max}$, and an admissibility condition that depends on κ only.

B.2.5 Asymptotic equivalent

We finally prove the second part of Theorem 1 for ULA: with the step-size $\eta = \eta_N = \frac{1}{2}\lambda_{\max}\frac{\log N}{N}$,

$$W_2(\mathbb{P}_N, \mathbb{P}) \underset{N\to\infty}{\sim} \frac{1}{8}\lambda_{\max}\sqrt{\mathrm{Tr}(\Sigma^{-1})}\frac{\log N}{N}.$$

Proof of Theorem 1, asymptotic equivalent for ULA. We take N large enough for η_N to be admissible in Proposition B.1. The starting point is the closed form of Lemma B.2,

$$\lambda^{(N)} = \frac{\lambda}{1 - \frac{\eta_N}{2}\lambda^{-1}} - \frac{\lambda}{1 - \frac{\eta_N}{2}\lambda^{-1}}\left(1 - \frac{\eta_N}{\lambda}\right)^{2N}. \tag{17}$$

We first show that $\lambda^{(N)} \to \lambda$. On the one hand, since $\eta_N \to 0$,

$$\frac{\lambda}{1 - \frac{\eta_N}{2}\lambda^{-1}} \underset{N\to\infty}{\longrightarrow} \lambda. \tag{18}$$

On the other hand, using $\log(1-u) \leq -u$ for $u < 1$ and then $\lambda \leq \lambda_{\max}$,

$$\left(1 - \frac{\eta_N}{\lambda}\right)^{2N} = \exp\left(2N\log\left(1 - \frac{\eta_N}{\lambda}\right)\right) \leq \exp\left(-\frac{2N\eta_N}{\lambda}\right) = N^{-\lambda_{\max}/\lambda} \leq \frac{1}{N}, \tag{19}$$

so the second term in (17) tends to 0, and with (18) we get $\lambda^{(N)} \to \lambda$. Applying Lemma A.3 directly to $W_2(\mathbb{P}_N, \mathbb{P})$ then gives

$$W_2^2(\mathbb{P}_N, \mathbb{P}) = \sum_{i=1}^{d}\left(\sqrt{\lambda_i^{(N)}} - \sqrt{\lambda_i}\right)^2 = \sum_{i=1}^{d}\left(\frac{\lambda_i^{(N)} - \lambda_i}{\sqrt{\lambda_i^{(N)}} + \sqrt{\lambda_i}}\right)^2 \sim \frac{1}{4}\sum_{i=1}^{d}\frac{(\lambda_i^{(N)} - \lambda_i)^2}{\lambda_i}. \tag{20}$$

It remains to estimate $\lambda^{(N)} - \lambda$. By (17) and (19),

$$\begin{aligned}\lambda^{(N)} - \lambda &= \frac{\lambda}{1 - \frac{\eta_N}{2}\lambda^{-1}} - \lambda + O\left(\frac{1}{N}\right) = \frac{\eta_N}{2 - \eta_N\lambda^{-1}} + O\left(\frac{1}{N}\right) \\ &= \frac{1}{2 - \eta_N\lambda^{-1}} \times \frac{1}{2}\lambda_{\max}\frac{\log N}{N} + O\left(\frac{1}{N}\right) \sim \frac{1}{4}\lambda_{\max}\frac{\log N}{N},\end{aligned}$$

an equivalent which does not depend on λ. Plugging it into (20),

$$W_2(\mathbb{P}_N, \mathbb{P}) \underset{N\to\infty}{\sim} \frac{1}{8}\lambda_{\max}\sqrt{\mathrm{Tr}(\Sigma^{-1})}\frac{\log N}{N},$$

which is the announced equivalent, consistent with the non-asymptotic bound. □

B.3 Underdamped Langevin dynamics

The proof of Theorem 1 for ULD follows the structure of the previous section. We first prove the proposition below, which bounds $W_2(\mathbb{P}_N, \mathbb{P})$ for all (η, γ) satisfying one of two sets of assumptions: a *large-friction* regime, which is the one adopted in Theorem 1, and a *small-friction* regime. Both lead to similar bounds, up to numerical constants; Remark B.8 explains why the first one is retained in Theorem 1.

Proposition B.7 (Parametric bound for ULD)**.** *Let $N \geq 0$. In the large-friction regime*

$$\gamma \geq \sqrt{8/\lambda_{\min}} \quad \textit{and} \quad \eta \leq \frac{1}{\gamma}, \tag{LF}$$

the error of ULD satisfies

$$W_2(\mathbb{P}_N, \mathbb{P}) \leq 8\sqrt{\operatorname{Tr}\Sigma}\, e^{-2\eta N/(\lambda_{\max}\gamma)} + \frac{6}{7}\sqrt{\operatorname{Tr}(\Sigma^{-1})}\,\frac{\eta}{\gamma},$$

and in the small-friction regime

$$\gamma \leq \sqrt{2/\lambda_{\max}} \quad \textit{and} \quad \eta \leq \frac{1}{2}\lambda_{\min}\gamma, \tag{SF}$$

it satisfies

$$W_2(\mathbb{P}_N, \mathbb{P}) \leq 8\sqrt{\operatorname{Tr}\Sigma}\, e^{-\eta\gamma N/2} + \frac{3}{2}\sqrt{\operatorname{Tr}(\Sigma^{-1})}\,\frac{\eta}{\gamma}.$$

Here too, each bound is the sum of an early-stopping term and a discretization term. Up to a numerical constant, the discretization term has the same expression in both regimes. Only the early-stopping term changes, with the friction γ appearing in the denominator or in the numerator of the exponent according to whether it is large or small.

Remark B.8. *The small-friction bound of Proposition B.7 is strictly decreasing in γ. There is no trade-off: one simply saturates the constraint $\gamma \leq \sqrt{2/\lambda_{\max}}$, which also makes the admissible range $\eta \leq \frac{1}{2}\lambda_{\min}\gamma$ as wide as possible. The resulting rates are the same in both regimes. We retain the large-friction regime* (LF) *in Theorem 1 because the roots $r_\pm$ introduced below are then real and positive, which keeps the analysis simplest, and because the bound has exactly the form of the ULA bound with $\phi := \eta/\gamma$ in place of η; with the hyperparameters of the theorem, $\phi = \frac{1}{2}\lambda_{\max}\frac{\log N}{N}$ is precisely the ULA step-size.*

B.3.1 Closed form of the eigenvalues

As before, we start with a closed form for the eigenvalues of Σ_N and Σ_∞.

Lemma B.9. *Fix an iteration budget $N \geq 0$. We consider two sets of assumptions:*

- *either*

$$\gamma > \frac{2}{\sqrt{\lambda_{\min}}} \quad \textit{and} \quad 0 < \eta < \frac{2}{\gamma}, \tag{LF$_0$}$$

  in which case we set

$$r_\pm = 1 - \frac{1}{2}\eta\big(\gamma \mp \sqrt{\gamma^2 - 4\lambda^{-1}}\big); \tag{21}$$

- *or*

$$0 < \gamma < \frac{2}{\sqrt{\lambda_{\max}}} \quad \textit{and} \quad 0 < \eta < \lambda_{\min}\gamma, \tag{SF$_0$}$$

  in which case we set

$$r_\pm = 1 - \frac{1}{2}\eta\big(\gamma \mp i\sqrt{4\lambda^{-1} - \gamma^2}\big),$$

  where i denotes the imaginary unit.

In both cases, the eigenvalues $\lambda^{(N)}$ are exactly

$$\lambda^{(N)} = \frac{2\gamma(2-\eta\gamma+\eta^2\lambda^{-1})}{(\gamma-\eta\lambda^{-1})(4-2\eta\gamma+\eta^2\lambda^{-1})}\lambda - \frac{2\eta\gamma}{\gamma^2-4\lambda^{-1}}\left(\frac{r_+^{2N}}{1-r_+^2} - \frac{2(r_+r_-)^N}{1-r_+r_-} + \frac{r_-^{2N}}{1-r_-^2}\right), \tag{22}$$

and they converge as $N \to \infty$ to

$$\lambda^{\infty} = \frac{2\gamma(2-\eta\gamma+\eta^2\lambda^{-1})}{(\gamma-\eta\lambda^{-1})(4-2\eta\gamma+\eta^2\lambda^{-1})}\lambda. \tag{23}$$

Proof. Fix an index i and consider the corresponding sequence of eigenvalues $(\lambda^{(k)})$. By Lemma A.1(iii) and Remark A.2,

$$\forall k \geq 0, \quad \lambda^{(k)} = \mathrm{Var}(x_k), \tag{24}$$

where (x_k, v_k) is defined by $x_0 = v_0 = 0$ and

$$\begin{cases} x_{k+1} = x_k + \eta v_k, \\ v_{k+1} = (1-\eta\gamma)v_k - \eta\lambda^{-1}x_k + \sqrt{2\eta\gamma}\,\zeta_k, \end{cases}$$

with (ζ_k) independent standard one-dimensional Gaussians. In matrix form, $z_{k+1} = Mz_k + u\zeta_k$ with

$$z_k = \begin{bmatrix} x_k \\ v_k \end{bmatrix}, \quad M = \begin{bmatrix} 1 & \eta \\ -\eta\lambda^{-1} & 1-\eta\gamma \end{bmatrix}, \quad u = \begin{bmatrix} 0 \\ \sqrt{2\eta\gamma} \end{bmatrix}, \tag{25}$$

so that, since $z_0 = 0$,

$$z_k = \sum_{j=0}^{k-1} M^j u\, \zeta_{k-1-j}. \tag{26}$$

Let β_j denote the $(1,2)$ entry of M^j. Reading the first coordinate of (26) at $k = N$ gives $x_N = \sqrt{2\eta\gamma}\sum_{j=0}^{N-1}\beta_j\zeta_{N-1-j}$, hence by (24)

$$\lambda^{(N)} = 2\eta\gamma\sum_{j=0}^{N-1}\beta_j^2. \tag{27}$$

It remains to compute the β_j. By the Cayley-Hamilton theorem, $M^2 - \mathrm{Tr}(M)M + \det(M)\,\mathrm{I} = 0$, hence $M^{j+2} = \mathrm{Tr}(M)M^{j+1} - \det(M)M^j$ for all j, and taking $(1,2)$ entries,

$$\forall j \geq 0, \quad \beta_{j+2} = \mathrm{Tr}(M)\beta_{j+1} - \det(M)\beta_j. \tag{28}$$

From (25), $\mathrm{Tr}(M) = 2-\eta\gamma$ and $\det(M) = 1-\eta\gamma+\eta^2\lambda^{-1}$, so the behavior of (β_j) is governed by the roots of

$$P(X) = X^2 - (2-\eta\gamma)X + 1 - \eta\gamma + \eta^2\lambda^{-1},$$

whose discriminant is

$$\Delta = \eta^2(\gamma^2 - 4\lambda^{-1}). \tag{29}$$

This formula shows that the two assumptions on γ in Lemma B.9 are exactly the cases where Δ is positive for all eigenvalues (large friction, real roots), or negative for all eigenvalues (small friction, complex roots): these are the two homogeneous regimes we study.

Stability. By (27), for $\lambda^{(N)}$ to converge the sequence (β_j) must tend to 0, *i.e.*, both roots of P must have modulus strictly less than 1. By the Jury stability criterion (Jury, 1962), this amounts to

$$|\det(M)| < 1, \quad P(1) > 0, \quad P(-1) > 0,$$

that is,

$$|1-\eta\gamma+\eta^2\lambda^{-1}| < 1, \quad \eta^2\lambda^{-1} > 0, \quad 4-2\eta\gamma+\eta^2\lambda^{-1} > 0.$$

The second condition holds since $\eta > 0$. For the third, as $\eta^2\lambda^{-1} > 0$, it suffices that $\eta\gamma \leq 2$: this is immediate under (LF_0), whose assumption on η is precisely the strict version of this inequality, and under (SF_0) it follows from

$$P(-1) = \frac{(\eta-\gamma\lambda)^2}{\lambda} + 4 - \gamma^2\lambda \geq 4 - \gamma^2\lambda_{\max} > 0.$$

As for the first condition, $\det(M) < 1$ is equivalent to $\eta/\gamma < \lambda$, which holds since $\eta/\gamma < \lambda_{\min} \le \lambda$ under both (LF_0) (where $\eta/\gamma < 2/\gamma^2 < \lambda_{\min}/2$) and ($\mathrm{SF}_0$); and $\det(M) > -1$ follows from $P(1) + P(-1) = 2(1 + \det(M))$, which is positive by the two conditions just established. Thus, in both cases,

$$|\det(M)| < 1, \quad P(1) > 0, \quad P(-1) > 0, \tag{30}$$

and the assumptions of Lemma B.9 guarantee stability.

Computation of $\lambda^{(N)}$. By (29), P has two distinct roots, which are the $r_\pm$ of the statement: $r_\pm = 1 - \frac{1}{2}\eta\big(\gamma \mp \sqrt{\gamma^2 - 4\lambda^{-1}}\big)$ under (LF_0) and $r_\pm = 1 - \frac{1}{2}\eta\big(\gamma \mp i\sqrt{4\lambda^{-1} - \gamma^2}\big)$ under (SF_0). They satisfy

$$r_+ + r_- = \mathrm{Tr}(M) = 2 - \eta\gamma \tag{31}$$

and

$$r_+ r_- = \det(M) = 1 - \eta\gamma + \eta^2\lambda^{-1}, \tag{32}$$

two relations we will use repeatedly. By (28), $\beta_j = A r_+^j + B r_-^j$ for some constants A, B determined by $\beta_0 = 0$ and $\beta_1 = \eta$, which gives

$$\forall j \ge 0, \quad \beta_j = \frac{\eta}{r_+ - r_-}(r_+^j - r_-^j),$$

and with (27),

$$\lambda^{(N)} = \frac{2\eta^3\gamma}{(r_+ - r_-)^2} \sum_{j=0}^{N-1} (r_+^{2j} - 2(r_+ r_-)^j + r_-^{2j}).$$

The Jury criterion, satisfied under (LF_0) and (SF_0), ensures

$$|r_\pm| < 1, \tag{33}$$

so r_+^2, r_-^2 and $r_+ r_-$ all differ from 1, and summing the three geometric series yields

$$\lambda^{(N)} = \lambda^\infty - \frac{2\eta^3\gamma}{(r_+ - r_-)^2}\left(\frac{r_+^{2N}}{1 - r_+^2} - \frac{2(r_+ r_-)^N}{1 - r_+ r_-} + \frac{r_-^{2N}}{1 - r_-^2}\right),$$

with

$$\lambda^\infty = \frac{2\eta^3\gamma}{(r_+ - r_-)^2}\left(\frac{1}{1 - r_+^2} - \frac{2}{1 - r_+ r_-} + \frac{1}{1 - r_-^2}\right). \tag{34}$$

By (33), $\lambda^{(N)}$ converges to λ^∞. It remains to show that λ^∞ has the announced form. Putting the three fractions in (34) over the common denominator $(1 - r_+^2)(1 - r_-^2)(1 - r_+ r_-)$, the numerator C factorizes:

$$\begin{aligned} C &= (1 - r_+^2)(1 - r_+ r_-) - 2(1 - r_+^2)(1 - r_-^2) + (1 - r_+ r_-)(1 - r_-^2) \\ &= r_+^2 - 2r_+ r_- + r_-^2 + r_+ r_-(r_+^2 - 2r_+ r_- + r_-^2) \\ &= (r_+ - r_-)^2(1 + r_+ r_-), \end{aligned}$$

so that

$$\lambda^\infty = 2\eta^3\gamma\,\frac{1 + r_+ r_-}{(1 - r_+^2)(1 - r_-^2)(1 - r_+ r_-)}. \tag{35}$$

To conclude independently of the regime, we express everything through the sum and product of the roots. On the one hand, by (31) and (32),

$$\begin{aligned} (1 - r_+^2)(1 - r_-^2) &= 1 + (r_+ r_-)^2 - (r_+ + r_-)^2 + 2r_+ r_- \\ &= 1 + (1 - \eta\gamma + \eta^2\lambda^{-1})^2 - (2 - \eta\gamma)^2 + 2(1 - \eta\gamma + \eta^2\lambda^{-1}) \\ &= \eta^4\lambda^{-2} + 4\eta^2\lambda^{-1} - 2\eta^3\gamma\lambda^{-1} = \eta^2\lambda^{-1}(4 - 2\eta\gamma + \eta^2\lambda^{-1}), \end{aligned}$$

the third line being a direct expansion of the second. On the other hand, (32) gives $1 + r_+ r_- = 2 - \eta\gamma + \eta^2\lambda^{-1}$ and $1 - r_+ r_- = \eta(\gamma - \eta\lambda^{-1})$. Plugging these into (35),

$$\lambda^\infty = \frac{2\gamma(2 - \eta\gamma + \eta^2\lambda^{-1})}{(\gamma - \eta\lambda^{-1})(4 - 2\eta\gamma + \eta^2\lambda^{-1})}\lambda,$$

which is (23). As expected, this expression does not depend on the regime, and once again $\lambda^\infty \to \lambda$ as $\eta \to 0$. Finally, (22) follows from $(r_+ - r_-)^2 = (r_+ + r_-)^2 - 4r_+ r_- = \eta^2(\gamma^2 - 4\lambda^{-1})$, by (31) and (32) once more. □

B.3.2 EARLY-STOPPING ERROR

We now bound the early-stopping error of ULD.

Lemma B.10. *Let $N \geq 0$. In the large-friction regime* (LF)*, the early-stopping error satisfies*

$$W_2(\mathbb{P}_N, \mathbb{P}_\infty) \leq 8\sqrt{\operatorname{Tr}\Sigma}\, e^{-2\eta N/(\lambda_{\max}\gamma)},$$

and in the small-friction regime (SF)*,*

$$W_2(\mathbb{P}_N, \mathbb{P}_\infty) \leq 8\sqrt{\operatorname{Tr}\Sigma}\, e^{-\eta\gamma N/2}.$$

Proof. Once again, (11) gives

$$W_2^2(\mathbb{P}_N, \mathbb{P}_\infty) = \sum_{i=1}^{d} \left(\frac{\lambda_i^\infty - \lambda_i^{(N)}}{\sqrt{\lambda_i^\infty} + \sqrt{\lambda_i^{(N)}}} \right)^2 \leq \sum_{i=1}^{d} \frac{\left(\lambda_i^\infty - \lambda_i^{(N)}\right)^2}{\lambda_i^\infty}. \tag{36}$$

We first check that

$$\lambda^\infty > \lambda. \tag{37}$$

Starting from (23),

$$\begin{aligned}
\lambda^\infty - \lambda &= \frac{\gamma(4 - 2\eta\gamma + 2\eta^2\lambda^{-1}) - (\gamma - \eta\lambda^{-1})(4 - 2\eta\gamma + \eta^2\lambda^{-1})}{(\gamma - \eta\lambda^{-1})(4 - 2\eta\gamma + \eta^2\lambda^{-1})}\lambda \\
&= \frac{\gamma\eta^2\lambda^{-1} + \eta\lambda^{-1}(4 - 2\eta\gamma + \eta^2\lambda^{-1})}{(\gamma - \eta\lambda^{-1})(4 - 2\eta\gamma + \eta^2\lambda^{-1})}\lambda \\
&= \frac{(4 - \eta\gamma + \eta^2\lambda^{-1})\,\eta}{(\gamma - \eta\lambda^{-1})(4 - 2\eta\gamma + \eta^2\lambda^{-1})}.
\end{aligned} \tag{38}$$

Recognizing $P(-1) = 4 - 2\eta\gamma + \eta^2\lambda^{-1}$, which we showed to be positive, this reads

$$\lambda^\infty - \lambda = \frac{\eta}{\gamma - \eta\lambda^{-1}} \times \frac{P(-1) + \eta\gamma}{P(-1)}. \tag{39}$$

The difference therefore has the sign of $\gamma - \eta\lambda^{-1}$, which is positive by the first Jury condition in (30). This proves (37), and (36) becomes

$$W_2^2(\mathbb{P}_N, \mathbb{P}_\infty) \leq \sum_{i=1}^{d} \frac{\left(\lambda_i^\infty - \lambda_i^{(N)}\right)^2}{\lambda_i}. \tag{40}$$

We are thus led to bound $|\lambda_i^\infty - \lambda_i^{(N)}|$, for which Lemma B.9 provides

$$|\lambda^\infty - \lambda^{(N)}| = \frac{2\eta\gamma}{|\gamma^2 - 4\lambda^{-1}|} \left| \frac{r_+^{2N}}{1 - r_+^2} - \frac{2(r_+ r_-)^N}{1 - r_+ r_-} + \frac{r_-^{2N}}{1 - r_-^2} \right|. \tag{41}$$

Large-friction regime (LF)*.* By (27), λ^∞ is the limit of $2\eta\gamma \sum_{j=0}^{N-1} \beta_j^2$ as $N \to \infty$, so the difference $\lambda^\infty - \lambda^{(N)}$ is the tail

$$\lambda^\infty - \lambda^{(N)} = 2\eta\gamma \sum_{j=N}^{\infty} \beta_j^2 \geq 0.$$

Since moreover $\gamma^2 - 4\lambda^{-1} > 0$ under (LF), (41) holds without absolute values:

$$|\lambda^\infty - \lambda^{(N)}| = \frac{2\eta\gamma}{\gamma^2 - 4\lambda^{-1}} \left(\frac{r_+^{2N}}{1 - r_+^2} - \frac{2(r_+ r_-)^N}{1 - r_+ r_-} + \frac{r_-^{2N}}{1 - r_-^2} \right). \tag{42}$$

Now $1 - r_+ r_- > 0$ by (33), and (LF) makes the product of the roots positive:

$$r_+ r_- = 1 - \eta\gamma + \eta^2\lambda^{-1} > 1 - \eta\gamma \geq 0. \tag{43}$$

The middle term in (42) is therefore negative, and

$$|\lambda^\infty - \lambda^{(N)}| < \frac{2\eta\gamma}{\gamma^2 - 4\lambda^{-1}} \left(\frac{r_+^{2N}}{1 - r_+^2} + \frac{r_-^{2N}}{1 - r_-^2} \right), \tag{44}$$

where every numerator and denominator is positive: (LF) for $\gamma^2 - 4\lambda^{-1}$, and (33) for $1 - r_\pm^2$. We next compare $|r_+|$ and $|r_-|$. By (43) the two roots have the same sign, namely that of their sum $2 - \eta\gamma > 0$ by (31) and (LF), so both are positive, and their expression (21) gives

$$r_+ = |r_+| > |r_-| = r_-. \tag{45}$$

Hence (44) yields

$$|\lambda^\infty - \lambda^{(N)}| < \frac{2\eta\gamma}{\gamma^2 - 4\lambda^{-1}} \left(\frac{1}{1 - r_+^2} + \frac{1}{1 - r_-^2} \right) r_+^{2N}.$$

One could bound the two terms in the parenthesis separately and get a slightly sharper estimate, but this would only improve a numerical constant, so we simply use $1 - r_-^2 > 1 - r_+^2$, a consequence of (45), to obtain

$$|\lambda^\infty - \lambda^{(N)}| < \frac{4\eta\gamma}{\gamma^2 - 4\lambda^{-1}} \times \frac{r_+^{2N}}{1 - r_+^2}.$$

By (LF), $\gamma^2 - 4\lambda^{-1} \geq \gamma^2 - 4\lambda_{\min}^{-1} \geq \frac{1}{2}\gamma^2$, hence

$$|\lambda^\infty - \lambda^{(N)}| < \frac{8\eta}{\gamma} \times \frac{r_+^{2N}}{1 - r_+^2}. \tag{46}$$

Again without optimizing constants, since $r_+ > 0$,

$$\frac{1}{1 - r_+^2} = \frac{1}{(1 + r_+)(1 - r_+)} \leq \frac{1}{1 - r_+}, \tag{47}$$

and by (21),

$$\frac{1}{1 - r_+} = \frac{2}{\eta\big(\gamma - \sqrt{\gamma^2 - 4\lambda^{-1}}\big)} = \frac{2\big(\gamma + \sqrt{\gamma^2 - 4\lambda^{-1}}\big)}{\eta\big(\gamma^2 - (\gamma^2 - 4\lambda^{-1})\big)} < \frac{\gamma\lambda}{\eta}.$$

Combining with (46) and (47),

$$|\lambda^\infty - \lambda^{(N)}| < 8\lambda\, r_+^{2N}. \tag{48}$$

To make this bound usable we finally bound r_+. The concavity inequality $\sqrt{1 - u} \leq 1 - \frac{1}{2}u$, valid for $u \in [0, 1]$, gives

$$r_+ = 1 - \frac{1}{2}\eta\gamma \left(1 - \sqrt{1 - 4\lambda^{-1}\gamma^{-2}}\right) \leq 1 - \frac{1}{2}\eta\gamma \times \frac{2}{\lambda\gamma^2} = 1 - \frac{\eta}{\gamma\lambda} \leq 1 - \frac{\eta}{\gamma\lambda_{\max}}.$$

Since $1 - \eta/(\gamma\lambda_{\max})$ is positive under (LF), raising to the power $2N$ and using, as in Section B.2, the convexity inequality

$$\forall u \in \mathbb{R}, \quad 1 - u \leq e^{-u}, \tag{49}$$

we deduce from (48)

$$|\lambda^\infty - \lambda^{(N)}| < 8\lambda \left(1 - \frac{\eta}{\gamma\lambda_{\max}}\right)^{2N} \leq 8\lambda\, e^{-2\eta N/(\gamma\lambda_{\max})}. \tag{50}$$

Small-friction regime (SF). We return to (41). The roots $r_\pm$ are now complex conjugates, so

$$|r_+|^2 = |r_-|^2 = r_+ r_- = 1 - \eta\gamma + \eta^2\lambda^{-1} =: R^2,$$

the common value being the constant coefficient of P. Moreover $\frac{r_+^{2N}}{1 - r_+^2}$ and $\frac{r_-^{2N}}{1 - r_-^2}$ are conjugate as well, hence

$$\begin{aligned}
\left| \frac{r_+^{2N}}{1 - r_+^2} - \frac{2(r_+ r_-)^N}{1 - r_+ r_-} + \frac{r_-^{2N}}{1 - r_-^2} \right| &= 2 \left| \mathrm{Re} \left(\frac{r_+^{2N}}{1 - r_+^2} \right) - \frac{R^{2N}}{1 - R^2} \right| \\
&\leq 2 \left| \frac{r_+^{2N}}{1 - r_+^2} \right| + \frac{2R^{2N}}{1 - R^2} = 2R^{2N} \left(\frac{1}{|1 - r_+^2|} + \frac{1}{1 - R^2} \right) \\
&\leq \frac{4R^{2N}}{1 - R^2},
\end{aligned}$$

where the last step uses $|1 - r_+^2| \geq 1 - |r_+|^2 = 1 - R^2$. Now $R^2 = 1 - \eta(\gamma - \eta\lambda^{-1})$ and, by (SF),

$$\gamma - \eta\lambda^{-1} \geq \gamma - \frac{1}{2}\lambda_{\min}\gamma\,\lambda_{\min}^{-1} = \frac{1}{2}\gamma, \tag{51}$$

so $R^2 \leq 1 - \frac{1}{2}\eta\gamma$ and

$$\left| \frac{r_+^{2N}}{1 - r_+^2} - \frac{2(r_+ r_-)^N}{1 - r_+ r_-} + \frac{r_-^{2N}}{1 - r_-^2} \right| \leq \frac{8\left(1 - \frac{1}{2}\eta\gamma\right)^N}{\eta\gamma}.$$

Plugging this into (41) and using $|\gamma^2 - 4\lambda^{-1}| = 4\lambda^{-1} - \gamma^2 \geq 4\lambda^{-1} - 2\lambda_{\max}^{-1} \geq 2\lambda^{-1}$,

$$|\lambda^\infty - \lambda^{(N)}| \leq \frac{16}{|\gamma^2 - 4\lambda^{-1}|}\left(1 - \frac{1}{2}\eta\gamma\right)^N \leq 8\lambda\left(1 - \frac{1}{2}\eta\gamma\right)^N.$$

Since $1 - \frac{1}{2}\eta\gamma \geq 0$ under (SF), the convexity inequality (49) finally gives

$$|\lambda^\infty - \lambda^{(N)}| \leq 8\lambda\, e^{-\eta\gamma N/2}. \tag{52}$$

Conclusion. Plugging (50), respectively (52), into (40) gives

$$W_2(\mathbb{P}_N, \mathbb{P}_\infty) \leq 8\sqrt{\operatorname{Tr}\Sigma}\, e^{-2\eta N/(\lambda_{\max}\gamma)} \qquad \text{under (LF)},$$
$$W_2(\mathbb{P}_N, \mathbb{P}_\infty) \leq 8\sqrt{\operatorname{Tr}\Sigma}\, e^{-\eta\gamma N/2} \qquad \text{under (SF)}.$$

□

Remark B.11. *The large- and small-friction assumptions* (LF) *and* (SF) *are stricter than conditions* (LF_0) *and* (SF_0) *of Lemma B.9, but only through numerical factors: their nature is the same. This stricter framework, which defines the admissible hyperparameters for ULD, keeps the denominator* $\gamma^2 - 4\lambda^{-1}$ *bounded away from zero, and affects the early-stopping and discretization bounds by a constant factor only.*

B.3.3 Discretization error

We now bound the discretization error.

Lemma B.12. *Under* (LF) *or* (SF)*, the discretization error satisfies*

$$W_2(\mathbb{P}_\infty, \mathbb{P}) \leq C\sqrt{\operatorname{Tr}(\Sigma^{-1})}\,\frac{\eta}{\gamma},$$

with $C = \frac{6}{7}$ *in the large-friction regime and* $C = \frac{3}{2}$ *in the small-friction regime.*

Proof. By (11) and the inequality (37) established in the previous section,

$$W_2^2(\mathbb{P}_\infty, \mathbb{P}) = \sum_{i=1}^d \left(\frac{\lambda_i^\infty - \lambda_i}{\sqrt{\lambda_i^\infty} + \sqrt{\lambda_i}}\right)^2 \leq \frac{1}{4}\sum_{i=1}^d \frac{(\lambda_i^\infty - \lambda_i)^2}{\lambda_i}. \tag{53}$$

The lemma will again follow from a bound on $|\lambda^\infty - \lambda| = \lambda^\infty - \lambda$, for which we return to (39):

$$\lambda^\infty - \lambda = \frac{\eta}{\gamma - \eta\lambda^{-1}} \times \frac{P(-1) + \eta\gamma}{P(-1)} = \frac{\eta}{\gamma - \eta\lambda^{-1}}\left(1 + \frac{\eta\gamma}{P(-1)}\right), \tag{54}$$

where $P(-1)$ is the quantity introduced with the Jury criterion. The assumptions of Proposition B.7, stronger than those of Lemma B.9 where this criterion appears, keep $P(-1)$ bounded away from zero. Indeed, in both regimes $\eta\gamma \leq 1$: this is immediate under (LF), and under (SF),

$$\eta\gamma \leq \frac{1}{2}\lambda_{\min}\gamma^2 \leq \frac{1}{\kappa} \leq 1.$$

Consequently $P(-1) = 4 - 2\eta\gamma + \eta^2\lambda^{-1} > 4 - 2\eta\gamma \geq 2$ in both regimes, and (54) gives

$$\lambda^\infty - \lambda < \frac{3}{2} \times \frac{\eta}{\gamma - \eta\lambda^{-1}}.$$

It remains to bound $\gamma - \eta\lambda^{-1}$ from below. Under (LF),

$$\gamma - \eta\lambda^{-1} = \gamma\left(1 - \frac{\eta}{\gamma\lambda}\right) \geq \gamma\left(1 - \frac{1}{\lambda\gamma^2}\right) \geq \gamma\left(1 - \frac{\lambda_{\min}}{8\lambda}\right) \geq \frac{7}{8}\gamma,$$

hence

$$\lambda^\infty - \lambda < \frac{12\eta}{7\gamma}. \tag{55}$$

Under (SF), (51) gives $\gamma - \eta\lambda^{-1} \geq \frac{1}{2}\gamma$, hence

$$\lambda^\infty - \lambda < \frac{3\eta}{\gamma}. \tag{56}$$

Plugging (55) and (56) into (53) completes the proof. □

Proof of Proposition B.7. Combine Lemmas B.10 and B.12 through the triangle inequality (10). □

B.3.4 Choice of hyperparameters

We can now prove the first part of Theorem 1 for ULD.

Proof of Theorem 1, non-asymptotic bound for ULD. Let $N \geq 2$ satisfy $N \geq 4\kappa \log N$, and recall the ULD hyperparameters of Theorem 1: $\eta_N = \sqrt{2\lambda_{\max}\kappa}\,\frac{\log N}{N}$ and $\gamma = \sqrt{8/\lambda_{\min}}$. They satisfy the large-friction assumption (LF): the only point to check is $\eta_N\gamma \leq 1$, and indeed

$$\eta_N\gamma = 4\kappa\frac{\log N}{N} \leq 1$$

by assumption. The first bound of Proposition B.7 therefore applies. Since $\frac{\eta_N}{\gamma} = \frac{1}{2}\lambda_{\max}\frac{\log N}{N}$, the exponential equals $e^{-\log N} = 1/N$, and

$$W_2(\mathbb{P}_N, \mathbb{P}) \leq \frac{8\sqrt{\operatorname{Tr}\Sigma}}{N} + \frac{3}{7}\lambda_{\max}\sqrt{\operatorname{Tr}(\Sigma^{-1})}\frac{\log N}{N}.$$

Using (16) once more gives the announced bound,

$$W_2(\mathbb{P}_N, \mathbb{P}) \leq \frac{\lambda_{\max}\sqrt{\operatorname{Tr}(\Sigma^{-1})}}{N}\Big(\frac{3}{7}\log N + 8\Big).$$

□

The same argument in the small-friction regime (SF) gives a bound of the same order.

Corollary B.13. *Let $N \geq 2\kappa \log N$. With $\eta = \sqrt{2\lambda_{\max}}\,\frac{\log N}{N}$ and $\gamma = \sqrt{2/\lambda_{\max}}$,*

$$W_2(\mathbb{P}_N, \mathbb{P}) \leq \frac{\lambda_{\max}\sqrt{\operatorname{Tr}(\Sigma^{-1})}}{N}\Big(\frac{3}{2}\log N + 8\Big).$$

Proof. With these hyperparameters, $\eta\gamma = 2\frac{\log N}{N}$ and $\frac{\eta}{\gamma} = \lambda_{\max}\frac{\log N}{N}$, and the condition $N \geq 2\kappa \log N$ is exactly $\eta \leq \frac{1}{2}\lambda_{\min}\gamma$, so (SF) holds. The second bound of Proposition B.7 gives $W_2(\mathbb{P}_N, \mathbb{P}) \leq 8\sqrt{\operatorname{Tr}\Sigma}/N + \frac{3}{2}\lambda_{\max}\sqrt{\operatorname{Tr}(\Sigma^{-1})}\frac{\log N}{N}$, and (16) concludes. □

Remark B.14. *To determine the optimal parameters η_*, γ_*, one applies Lemma A.4 to the variable $\phi := \eta/\gamma$. As for ULA, the exact minimizers are slightly different, with spectral coefficients inside the logarithms, but lead to the same bound up to numerical constants. The hyperparameters we retain are simpler: they require no knowledge of the spectrum of Σ beyond its extreme eigenvalues. The same principles guided the hyperparameters of Corollary B.13, where Lemma A.4 applies thanks to Remark B.8, which removes the variable γ. In what follows we nevertheless focus on the regime of Theorem 1. As for ULA, the asymptotic equivalent of the next section shows that, for the chosen hyperparameters, the actual error is also of order $\lambda_{\max}\sqrt{\operatorname{Tr}(\Sigma^{-1})}\log N/N$.*

B.3.5 ASYMPTOTIC EQUIVALENT

To complete the proof of Theorem 1, it remains to establish the asymptotic equivalent

$$W_2(\mathbb{P}_N, \mathbb{P}) \underset{N\to\infty}{\sim} \frac{1}{4}\lambda_{\max}\sqrt{\mathrm{Tr}(\Sigma^{-1})}\frac{\log N}{N}$$

for $\eta = \eta_N = \sqrt{2\lambda_{\max}\kappa}\,\frac{\log N}{N}$ and $\gamma = \sqrt{8/\lambda_{\min}}$.

Proof of Theorem 1, asymptotic equivalent for ULD. Take N large enough for the hyperparameters to satisfy (LF), so that Lemma B.9 and the bound (50) apply. Recall from (23) that

$$\lambda^\infty = \frac{2\gamma(2-\eta_N\gamma+\eta_N^2\lambda^{-1})}{(\gamma-\eta_N\lambda^{-1})(4-2\eta_N\gamma+\eta_N^2\lambda^{-1})}\lambda.$$

We first show that $\lambda^{(N)} \to \lambda$. On the one hand, $\eta_N \to 0$, so

$$\lambda^\infty \longrightarrow \frac{4\gamma}{4\gamma}\lambda = \lambda. \tag{57}$$

On the other hand, (50) with the values of the hyperparameters gives

$$|\lambda^\infty - \lambda^{(N)}| \le 8\lambda\, e^{-2\eta_N N/(\gamma\lambda_{\max})} = \frac{8\lambda}{N},$$

and with (57), $\lambda^{(N)} \to \lambda$ indeed. Applying Lemma A.3 to $\mathbb{P}_N$ and $\mathbb{P}$ then gives

$$W_2^2(\mathbb{P}_N, \mathbb{P}) = \sum_{i=1}^d \left(\sqrt{\lambda_i^{(N)}} - \sqrt{\lambda_i}\right)^2 = \sum_{i=1}^d \left(\frac{\lambda_i^{(N)} - \lambda_i}{\sqrt{\lambda_i^{(N)}} + \sqrt{\lambda_i}}\right)^2 \sim \frac{1}{4}\sum_{i=1}^d \frac{(\lambda_i^{(N)} - \lambda_i)^2}{\lambda_i}. \tag{58}$$

It remains to estimate $\lambda^{(N)} - \lambda$. By the bound above,

$$\lambda^{(N)} - \lambda = (\lambda^{(N)} - \lambda^\infty) + (\lambda^\infty - \lambda) = \lambda^\infty - \lambda + O\left(\frac{1}{N}\right), \tag{59}$$

and by (38),

$$\lambda^\infty - \lambda = \frac{(4-\eta_N\gamma+\eta_N^2\lambda^{-1})\,\eta_N}{(\gamma-\eta_N\lambda^{-1})(4-2\eta_N\gamma+\eta_N^2\lambda^{-1})} \underset{N\to\infty}{\sim} \frac{4\eta_N}{4\gamma} = \frac{\eta_N}{\gamma} = \frac{1}{2}\lambda_{\max}\frac{\log N}{N}.$$

Together with (59),

$$\lambda^{(N)} - \lambda = \frac{1}{2}\lambda_{\max}\frac{\log N}{N} + O\left(\frac{1}{N}\right),$$

an equivalent which, as for ULA, does not depend on λ. The equivalent (58) then yields

$$W_2(\mathbb{P}_N, \mathbb{P}) \underset{N\to\infty}{\sim} \frac{1}{4}\lambda_{\max}\sqrt{\mathrm{Tr}(\Sigma^{-1})}\frac{\log N}{N}.$$

□

C PROOF OF THEOREM 3 AND PROPOSITION 4 (DIFFUSION SAMPLER)

C.1 OVERVIEW

We now turn to the proof of the following theorem, stated in Section 2.3.

Theorem 3 (Diffusion sampler)**.** *Consider an iteration budget $N \ge 1$. Set horizon $T = (\mathrm{Tr}(\Sigma^5)/\,\mathrm{Tr}\,\Sigma)^{1/4}\sqrt{N}$, early-stopping threshold $\delta = \mathrm{Tr}\,\Sigma/(dN^2)$, and geometric time discretization grid $\tau_k = T(\delta/T)^{k/N}$ for $k \in \{0,\dots,N\}$. Then the diffusion algorithm* (6) *yields*

$$W_2(\mathbb{P}_N, \mathbb{P}) \le \frac{\sqrt{\mathrm{Tr}\,\Sigma}}{N}\Big(\tfrac{5}{8}\,a_N\,\rho^{1/N}\log(\rho^{2/5}N) + 1 + \sqrt{8}\Big), \quad \rho := \frac{d\lambda_{\max}}{\mathrm{Tr}\,\Sigma} \in [1, d],$$

where $a_N := (1+4N^{-2})^{1/2}N^{5/(2N)}$ satisfies $a_N \le 2^{7/4}$ for all $N \ge 1$ and $a_N \to 1$. Moreover, this bound is asymptotically exact: under the same hyperparameters,

$$W_2(\mathbb{P}_N, \mathbb{P}) \underset{N\to\infty}{\sim} \frac{5}{8}\sqrt{\mathrm{Tr}\,\Sigma}\,\frac{\log N}{N}, \tag{7}$$

so that, for fixed Σ, the ratio of the bound to the actual error tends to 1.

Proposition 4, which differs from Theorem 3 only through the choice of hyperparameters, is restated and proved in Section C.3.

Error decomposition. The error of the diffusion algorithm of Section 2.1 combines three sources. The first is the initialization error: the process is started from $Y_0 \sim \mathcal{N}(0, T\,\mathrm{I})$ as a proxy for the true marginal $\mathbb{Q}_T = \mathcal{N}(0, \Sigma + T\,\mathrm{I})$, which is unknown *a priori*. It is measured by the Wasserstein distance $W_2(\mathbb{P}_N, \tilde{\mathbb{P}}_N)$ between the law $\mathbb{P}_N$ actually sampled and the law $\tilde{\mathbb{P}}_N$ that would have been sampled from the exact initialization $Y_0 \sim \mathcal{N}(0, \Sigma + T\,\mathrm{I})$. The second is the discretization error incurred when simulating the reverse SDE (5): it is $W_2(\tilde{\mathbb{P}}_N, \mathbb{Q}_\delta)$, the distance, for an exact initialization, between the law produced by the discrete scheme and the law $\mathbb{Q}_\delta = \mathcal{N}(0, \Sigma + \delta\,\mathrm{I})$ reached by the continuous process. The third is the early-stopping error $W_2(\mathbb{Q}_\delta, \mathbb{P})$, due to stopping the continuous process at the threshold $\delta > 0$. The triangle inequality formalizes this decomposition:

$$W_2(\mathbb{P}_N, \mathbb{P}) \le W_2(\mathbb{P}_N, \tilde{\mathbb{P}}_N) + W_2(\tilde{\mathbb{P}}_N, \mathbb{Q}_\delta) + W_2(\mathbb{Q}_\delta, \mathbb{P}). \tag{60}$$

Reduction to eigenvalues. All the laws in (60) are centered Gaussians with covariance in $\mathcal{A}(\Sigma)$: this is immediate for $\mathbb{P}$ and $\mathbb{Q}_\delta$, and follows from Lemma A.1 for $\mathbb{P}_N$ and $\tilde{\mathbb{P}}_N$. Lemma A.3 therefore applies and, denoting by $\lambda_i^{(N)}$ and $\tilde{\lambda}_i^{(N)}$ the eigenvalues of the covariances of $\mathbb{P}_N$ and $\tilde{\mathbb{P}}_N$, we record for later use

$$W_2^2(\mathbb{P}_N, \tilde{\mathbb{P}}_N) = \sum_{i=1}^{d} \left(\sqrt{\lambda_i^{(N)}} - \sqrt{\tilde{\lambda}_i^{(N)}}\right)^2, \tag{61}$$

$$W_2^2(\tilde{\mathbb{P}}_N, \mathbb{Q}_\delta) = \sum_{i=1}^{d} \left(\sqrt{\tilde{\lambda}_i^{(N)}} - \sqrt{\lambda_i + \delta}\right)^2, \tag{62}$$

$$W_2^2(\mathbb{Q}_\delta, \mathbb{P}) = \sum_{i=1}^{d} \left(\sqrt{\lambda_i + \delta} - \sqrt{\lambda_i}\right)^2. \tag{63}$$

Outline. The proof of Theorem 3 follows the same steps as that of Theorem 1.

(i)–(ii) *Parametric bound* (Section C.2). Closed form of the eigenvalues $\lambda_i^{(N)}$ and $\tilde{\lambda}_i^{(N)}$ from the scalar recursions of Lemma A.1, then separate bounds on the three terms of (60), and summation. This gives a bound on $W_2(\mathbb{P}_N, \mathbb{P})$ in which T and δ remain free (Proposition C.1).

(iii) *Choice of hyperparameters* (Section C.3). Plug the hyperparameters of Theorem 3 and Proposition 4 into the bound of Proposition C.1.

(iv) *Asymptotic equivalent* (Section C.4). Exact equivalent of $W_2(\mathbb{P}_N, \mathbb{P})$ as $N \to \infty$ for any geometric grid with $T \propto \sqrt{N}$ and $\delta \propto N^{-2}$, from the closed forms of step (i). This equivalent does not depend on the proportionality constants, which shows that the spectral rate is a robust feature of the scheme.

C.2 Parametric bound for the diffusion sampler

The goal of this section is to prove the following proposition, which bounds $W_2(\mathbb{P}_N, \mathbb{P})$ for all values of the hyperparameters. As before, we freely drop the eigenvalue index i in the proofs.

Proposition C.1 (Parametric bound for the diffusion sampler)**.** *For every $N \ge 1$ and every $T > \delta > 0$, with the geometric discretization $\tau_k = T(\delta/T)^{k/N}$, the error of the diffusion algorithm satisfies*

$$W_2(\mathbb{P}_N, \mathbb{P}) \le \frac{\sqrt{\mathrm{Tr}((\Sigma + \delta\,\mathrm{I})^3\Sigma^2)}}{T^2} + \frac{\sqrt{\mathrm{Tr}\,\Sigma + 4d\delta}}{4N}\left(\frac{T}{\delta}\right)^{1/N} \log\frac{T}{\delta} + \sqrt{d\delta}.$$

The three terms bound the initialization, discretization and early-stopping errors respectively.

C.2.1 Closed form of the eigenvalues

We first establish a closed form for the eigenvalues $\lambda^{(N)}$ and $\tilde{\lambda}^{(N)}$, valid for any discretization grid.

Lemma C.2. *Fix $N \ge 1$ and consider the following generalized diffusion algorithm:*

$$\begin{cases} Y_0 \sim \mathcal{N}(0, M), \\ \forall k = 1, \dots, N\,,\ Y_k = Y_{k-1} + \eta_k s_{\tau_{k-1}}(Y_{k-1}) + \sqrt{\eta_k}\xi_k, \end{cases}$$

for some $M \in \mathcal{A}(\Sigma)$, with eigenvalues $\mu_i := [M]_i$. Then the iterates (Y_k) are centered Gaussian vectors whose covariance matrices (M_k) belong to $\mathcal{A}(\Sigma)$, and the eigenvalues $\mu_i^{(N)} := [M_N]_i$ are given by

$$\forall i, \quad \mu_i^{(N)} = \left(\frac{\lambda_i + \delta}{\lambda_i + T}\right)^2 \mu_i + (\lambda_i + \delta)^2 \sum_{j=1}^{N} \frac{\eta_j}{(\lambda_i + \tau_j)^2}.$$

Proof. This generalization of (6) falls within the scope of Lemma A.1: the identification is that of Remark A.2, the only change being the initial covariance M, which lies in $\mathcal{A}(\Sigma)$ by assumption. Point (ii) of the lemma gives the recursion

$$\forall k \geq 0, \quad \mu^{(k+1)} = \left(1 - \frac{\eta_{k+1}}{\lambda + \tau_k}\right)^2 \mu^{(k)} + \eta_{k+1}.$$

Since $\eta_{k+1} = \tau_k - \tau_{k+1}$, the contraction factor telescopes, leading to the simplification announced in Section 2.1:

$$\mu^{(k+1)} = \left(\frac{\lambda + \tau_k - \eta_{k+1}}{\lambda + \tau_k}\right)^2 \mu^{(k)} + \eta_{k+1} = \left(\frac{\lambda + \tau_{k+1}}{\lambda + \tau_k}\right)^2 \mu^{(k)} + \eta_{k+1}.$$

Dividing by $(\lambda + \tau_{k+1})^2$ and setting

$$\forall j \geq 0, \quad \nu_j := \frac{\mu^{(j)}}{(\lambda + \tau_j)^2},$$

we are left with the much simpler recursion

$$\forall j \geq 0, \quad \nu_{j+1} = \nu_j + \frac{\eta_{j+1}}{(\lambda + \tau_{j+1})^2}, \quad \nu_0 = \frac{\mu}{(\lambda + T)^2},$$

whence

$$\nu_k = \frac{\mu}{(\lambda + T)^2} + \sum_{j=1}^{k} \frac{\eta_j}{(\lambda + \tau_j)^2} \quad \text{and} \quad \mu^{(k)} = \left(\frac{\lambda + \tau_k}{\lambda + T}\right)^2 \mu + (\lambda + \tau_k)^2 \sum_{j=1}^{k} \frac{\eta_j}{(\lambda + \tau_j)^2}.$$

Taking $k = N$, with $\tau_N = \delta$, gives the announced formula. □

Remark C.3. *Applying this lemma with $M = T\,\mathrm{I}$ and with $M = \Sigma + T\,\mathrm{I}$, both in $\mathcal{A}(\Sigma)$, gives for every $N \geq 1$*

$$\lambda^{(N)} = \left(\frac{\lambda + \delta}{\lambda + T}\right)^2 T + (\lambda + \delta)^2 \sum_{j=1}^{N} \frac{\eta_j}{(\lambda + \tau_j)^2},$$

$$\tilde{\lambda}^{(N)} = \frac{(\lambda + \delta)^2}{\lambda + T} + (\lambda + \delta)^2 \sum_{j=1}^{N} \frac{\eta_j}{(\lambda + \tau_j)^2}.$$

C.2.2 Initialization error

Lemma C.4. *For every $N \geq 1$ and every $T > \delta > 0$, the initialization error is bounded uniformly in N:*

$$W_2(\mathbb{P}_N, \tilde{\mathbb{P}}_N) \leq \frac{\sqrt{\mathrm{Tr}((\Sigma + \delta\,\mathrm{I})^3 \Sigma^2)}}{T^2}.$$

Proof. Fix $N \geq 1$ and $T > \delta > 0$ throughout. By (61),

$$W_2^2(\mathbb{P}_N, \tilde{\mathbb{P}}_N) = \sum_{i=1}^{d} \left(\sqrt{\lambda_i^{(N)}} - \sqrt{\tilde{\lambda}_i^{(N)}}\right)^2. \tag{64}$$

Starting from the expression of $\tilde{\lambda}^{(N)}$ in Remark C.3, and using $(\lambda+\delta)^2 \cdot \frac{1}{\lambda+\delta} = \lambda+\delta$,

$$\tilde{\lambda}^{(N)} = (\lambda+\delta)^2\left(\sum_{j=1}^{N}\frac{\eta_j}{(\lambda+\tau_j)^2} + \frac{1}{\lambda+T}\right)$$
$$= \lambda+\delta+(\lambda+\delta)^2\left(\sum_{j=1}^{N}\frac{\eta_j}{(\lambda+\tau_j)^2} - \left(\frac{1}{\lambda+\delta}-\frac{1}{\lambda+T}\right)\right) = \lambda+\delta+(\lambda+\delta)^2\Delta, \quad (65)$$

where

$$\Delta := \sum_{j=1}^{N}\frac{\eta_j}{(\lambda+\tau_j)^2} - \left(\frac{1}{\lambda+\delta}-\frac{1}{\lambda+T}\right) = \sum_{j=1}^{N}\frac{\eta_j}{(\lambda+\tau_j)^2} - \int_{\delta}^{T}\frac{\mathrm{d}\tau}{(\lambda+\tau)^2}.$$

Δ is the quadrature error made when approximating $\int_{\delta}^{T}(\lambda+\tau)^{-2}\,\mathrm{d}\tau$ by the Riemann sum $\sum_{j=1}^{N}\eta_j(\lambda+\tau_j)^{-2}$, in which the decreasing integrand is evaluated at the left endpoint of each interval $[\tau_j, \tau_{j-1}]$; it is therefore positive. Explicitly, since $\tau_0 = T$ and $\tau_N = \delta$,

$$\begin{aligned}\Delta &= \sum_{j=1}^{N}\frac{\eta_j}{(\lambda+\tau_j)^2} - \sum_{j=1}^{N}\left(\frac{1}{\lambda+\tau_j}-\frac{1}{\lambda+\tau_{j-1}}\right)\\ &= \sum_{j=1}^{N}\left(\frac{\eta_j}{(\lambda+\tau_j)^2} - \frac{\eta_j}{(\lambda+\tau_j)(\lambda+\tau_{j-1})}\right)\\ &= \sum_{j=1}^{N}\frac{\eta_j^2}{(\lambda+\tau_j)^2(\lambda+\tau_{j-1})} > 0,\end{aligned} \quad (66)$$

where we used $\tau_{j-1}-\tau_j = \eta_j$ twice. With (65), this gives

$$\tilde{\lambda}^{(N)} > \lambda+\delta. \quad (67)$$

Remark C.3 also gives the difference of the two eigenvalues,

$$\tilde{\lambda}^{(N)} - \lambda^{(N)} = \left(\frac{\lambda+\delta}{\lambda+T}\right)^2\lambda, \quad (68)$$

so that, by (67),

$$\sqrt{\tilde{\lambda}^{(N)}} - \sqrt{\lambda^{(N)}} = \frac{\tilde{\lambda}^{(N)}-\lambda^{(N)}}{\sqrt{\tilde{\lambda}^{(N)}}+\sqrt{\lambda^{(N)}}} \le \frac{\tilde{\lambda}^{(N)}-\lambda^{(N)}}{\sqrt{\lambda+\delta}} = \frac{(\lambda+\delta)^{3/2}\lambda}{(\lambda+T)^2} \le \frac{(\lambda+\delta)^{3/2}\lambda}{T^2}.$$

Squaring, summing over i and using (64) gives the bound of the lemma. □

C.2.3 Discretization error

Lemma C.5. *For every $N \ge 1$ and every $T > \delta > 0$, with the geometric discretization $\tau_k = T(\delta/T)^{k/N}$, the discretization error satisfies*

$$W_2(\tilde{\mathbb{P}}_N, \mathbb{Q}_\delta) \le \frac{\sqrt{\operatorname{Tr}\Sigma + 4d\delta}}{4N}\left(\frac{T}{\delta}\right)^{1/N}\log\frac{T}{\delta}.$$

Proof. Fix $N \ge 1$ and $T > \delta > 0$. By (62) and (67),

$$W_2^2(\tilde{\mathbb{P}}_N, \mathbb{Q}_\delta) = \sum_{i=1}^{d}\left(\sqrt{\tilde{\lambda}_i^{(N)}} - \sqrt{\lambda_i+\delta}\right)^2 = \sum_{i=1}^{d}\left(\frac{\tilde{\lambda}_i^{(N)} - (\lambda_i+\delta)}{\sqrt{\tilde{\lambda}_i^{(N)}}+\sqrt{\lambda_i+\delta}}\right)^2$$
$$\le \frac{1}{4}\sum_{i=1}^{d}\frac{\left(\tilde{\lambda}_i^{(N)} - (\lambda_i+\delta)\right)^2}{\lambda_i+\delta},$$

and with (65),

$$W_2^2(\tilde{\mathbb{P}}_N, \mathbb{Q}_\delta) \le \frac{1}{4}\sum_{i=1}^{d}(\lambda_i+\delta)^3\Delta_i^2. \tag{69}$$

It thus remains to bound Δ for the geometric grid $\tau_k = T(\delta/T)^{k/N}$. Let $q := (\delta/T)^{1/N}$ denote its ratio, so that $\tau_k = q\tau_{k-1}$ and

$$\forall k \ge 1, \quad \eta_k = (q^{-1}-1)\tau_k.$$

Returning to (66),

$$\Delta = (q^{-1}-1)\sum_{j=1}^{N}\frac{\tau_j}{\lambda+\tau_j}\left(\frac{1}{\lambda+\tau_j} - \frac{1}{\lambda+\tau_{j-1}}\right) \tag{70}$$

$$= (q^{-1}-1)\sum_{j=1}^{N} f(\tau_j)\big(G(\tau_j) - G(\tau_{j-1})\big), \tag{71}$$

where

$$f : \tau \longmapsto \frac{\tau}{\lambda+\tau}, \quad G : \tau \longmapsto \frac{1}{\lambda+\tau}.$$

Since $\tau_j < \tau_{j-1}$ and f is increasing,

$$f(\tau_j)\big(G(\tau_j) - G(\tau_{j-1})\big) = f(\tau_j)\int_{\tau_j}^{\tau_{j-1}}\frac{\mathrm{d}\tau}{(\lambda+\tau)^2} \le \int_{\tau_j}^{\tau_{j-1}}\frac{f(\tau)\,\mathrm{d}\tau}{(\lambda+\tau)^2} = \int_{\tau_j}^{\tau_{j-1}}\frac{\tau\,\mathrm{d}\tau}{(\lambda+\tau)^3},$$

and summing over j in (71),

$$\Delta \le (q^{-1}-1)\int_{\delta}^{T}\frac{\tau\,\mathrm{d}\tau}{(\lambda+\tau)^3}. \tag{72}$$

One checks that $\tau \mapsto \frac{\lambda}{2(\lambda+\tau)^2} - \frac{1}{\lambda+\tau}$ is an antiderivative of the integrand, hence

$$\Delta \le (q^{-1}-1)\left[\frac{\lambda}{2(\lambda+\tau)^2} - \frac{1}{\lambda+\tau}\right]_{\delta}^{T} = (q^{-1}-1)\left(\frac{\lambda+2\delta}{2(\lambda+\delta)^2} - \frac{\lambda+2T}{2(\lambda+T)^2}\right)$$
$$\le (q^{-1}-1)\frac{\lambda+2\delta}{2(\lambda+\delta)^2},$$

and therefore

$$(\lambda+\delta)^3\Delta^2 \le \frac{1}{4}(q^{-1}-1)^2\frac{(\lambda+2\delta)^2}{\lambda+\delta} \le \frac{1}{4}(q^{-1}-1)^2(\lambda+4\delta),$$

where the last inequality is $(\lambda+2\delta)^2 \le (\lambda+2\delta)^2 + \lambda\delta = (\lambda+\delta)(\lambda+4\delta)$. Plugging this into (69),

$$W_2(\tilde{\mathbb{P}}_N, \mathbb{Q}_\delta) \le \frac{1}{4}(q^{-1}-1)\sqrt{\operatorname{Tr}\Sigma + 4d\delta}. \tag{73}$$

Finally, the inequality $e^u - 1 \le ue^u$, valid for all $u \in \mathbb{R}$, applied to $u = \frac{1}{N}\log\frac{T}{\delta}$ gives

$$q^{-1} - 1 = \left(\frac{T}{\delta}\right)^{1/N} - 1 \le \frac{1}{N}\left(\frac{T}{\delta}\right)^{1/N}\log\frac{T}{\delta},$$

which together with (73) concludes the proof. □

C.2.4 EARLY-STOPPING ERROR

Lemma C.6. *For every $\delta > 0$, the early-stopping error satisfies*

$$W_2(\mathbb{Q}_\delta, \mathbb{P}) \le \sqrt{d\delta}.$$

Proof. By (63), it suffices to note that for every $\lambda > 0$,

$$(\sqrt{\lambda+\delta} - \sqrt{\lambda})^2 = 2\lambda + \delta - 2\sqrt{\lambda}\sqrt{\lambda+\delta} < 2\lambda + \delta - 2\lambda = \delta.$$

□

Proof of Proposition C.1. Combine Lemmas C.4, C.5 and C.6 through the triangle inequality (60). □

C.3 Choice of hyperparameters

Thanks to Proposition C.1, we can now prove the non-asymptotic bounds of Theorem 3 (including the properties of (a_N) stated there) and of Proposition 4.

Proof of Theorem 3, non-asymptotic bound. Let $N \geq 1$ and recall the hyperparameters of Theorem 3:

$$T = \left(\frac{\mathrm{Tr}(\Sigma^5)}{\mathrm{Tr}\,\Sigma}\right)^{1/4}\sqrt{N}, \quad \delta = \frac{\mathrm{Tr}\,\Sigma}{dN^2}.$$

We first check that $T > \delta$, as required by Proposition C.1. By Hölder's inequality, $\mathrm{Tr}(\Sigma^5) \geq (\mathrm{Tr}\,\Sigma)^5/d^4$, hence

$$T \geq \left(\frac{(\mathrm{Tr}\,\Sigma)^4}{d^4}\right)^{1/4}\sqrt{N} = \frac{\mathrm{Tr}\,\Sigma}{d}\sqrt{N},$$

which exceeds $\delta = \frac{\mathrm{Tr}\,\Sigma}{d}N^{-2}$ as soon as $N \geq 2$. For $N = 1$, the inequality $T > \delta$ still holds unless Hölder's inequality is an equality, *i.e.*, unless Σ is a scalar matrix. In that isotropic case one has $T = \delta$, the algorithm performs a single step of size $\eta_1 = 0$, and $\mathbb{P}_1 = \mathcal{N}(0, T\,\mathrm{I}) = \mathbb{P}$, so that $W_2(\mathbb{P}_1, \mathbb{P}) = 0$ and the bound holds trivially. This degenerate case is of no interest, but covering it allows Theorem 3 to be stated for all $N \geq 1$. We assume $T > \delta$ from now on.

Proposition C.1 then gives

$$W_2(\mathbb{P}_N, \mathbb{P}) \leq \frac{\sqrt{\mathrm{Tr}((\Sigma+\delta\,\mathrm{I})^3\Sigma^2)}}{T^2} + \frac{\sqrt{\mathrm{Tr}\,\Sigma + 4d\delta}}{4N}\left(\frac{T}{\delta}\right)^{1/N}\log\frac{T}{\delta} + \sqrt{d\delta}, \tag{74}$$

and we bound the three terms in turn.

Initialization. With the value of T,

$$\frac{\sqrt{\mathrm{Tr}((\Sigma+\delta\,\mathrm{I})^3\Sigma^2)}}{T^2} = \frac{\sqrt{\mathrm{Tr}\,\Sigma}}{N} \times \sqrt{\frac{\mathrm{Tr}((\Sigma+\delta\,\mathrm{I})^3\Sigma^2)}{\mathrm{Tr}(\Sigma^5)}}. \tag{75}$$

Expanding the binomial and inserting the value of δ,

$$\mathrm{Tr}((\Sigma+\delta\,\mathrm{I})^3\Sigma^2) = \sum_{j=0}^{3}\binom{3}{j}\delta^j\,\mathrm{Tr}(\Sigma^{5-j}) = \sum_{j=0}^{3}\binom{3}{j}\left(\frac{\mathrm{Tr}\,\Sigma}{dN^2}\right)^j\mathrm{Tr}(\Sigma^{5-j}).$$

By Hölder's inequality again, $\mathrm{Tr}(\Sigma^k) \leq d^{1-k/5}\,\mathrm{Tr}(\Sigma^5)^{k/5}$ for $1 \leq k \leq 5$; applied to $k = 1$ and $k = 5 - j$, this gives $(\mathrm{Tr}\,\Sigma/d)^j \leq d^{-j/5}\,\mathrm{Tr}(\Sigma^5)^{j/5}$ and $\mathrm{Tr}(\Sigma^{5-j}) \leq d^{j/5}\,\mathrm{Tr}(\Sigma^5)^{1-j/5}$, so the powers of d cancel and

$$\mathrm{Tr}((\Sigma+\delta\,\mathrm{I})^3\Sigma^2) \leq \sum_{j=0}^{3}\binom{3}{j}\frac{1}{N^{2j}}\,\mathrm{Tr}(\Sigma^5) = \left(1+\frac{1}{N^2}\right)^3\mathrm{Tr}(\Sigma^5).$$

Returning to (75),

$$\frac{\sqrt{\mathrm{Tr}((\Sigma+\delta\,\mathrm{I})^3\Sigma^2)}}{T^2} \leq \left(1+\frac{1}{N^2}\right)^{3/2}\frac{\sqrt{\mathrm{Tr}\,\Sigma}}{N} \leq \sqrt{8}\,\frac{\sqrt{\mathrm{Tr}\,\Sigma}}{N}.$$

Discretization. On the one hand, $\mathrm{Tr}\,\Sigma + 4d\delta = (1 + 4N^{-2})\,\mathrm{Tr}\,\Sigma$. On the other hand,

$$\frac{T}{\delta} = \left(\frac{\mathrm{Tr}(\Sigma^5)}{\mathrm{Tr}\,\Sigma}\right)^{1/4}\frac{dN^{5/2}}{\mathrm{Tr}\,\Sigma} \leq \frac{d\lambda_{\max}}{\mathrm{Tr}\,\Sigma}N^{5/2} = \rho N^{5/2},$$

where we used $\mathrm{Tr}(\Sigma^5) \leq \lambda_{\max}^4\,\mathrm{Tr}\,\Sigma$, and where $\rho := d\lambda_{\max}/\,\mathrm{Tr}\,\Sigma \in [1, d]$ is the ratio between the largest eigenvalue of Σ and the arithmetic mean of its spectrum: the more anisotropic Σ, the larger ρ. Since $x \mapsto x^{1/N}\log x$ is increasing on $[1, \infty)$ and $T/\delta > 1$, the discretization term is at most

$$\frac{(1+4N^{-2})^{1/2}\sqrt{\mathrm{Tr}\,\Sigma}}{4N}\,(\rho N^{5/2})^{1/N}\log(\rho N^{5/2}) = \frac{5}{8}\,a_N\,\frac{\sqrt{\mathrm{Tr}\,\Sigma}}{N}\,\rho^{1/N}\log(\rho^{2/5}N),$$

where we wrote $\log(\rho N^{5/2}) = \frac{5}{2}\log(\rho^{2/5}N)$ and $a_N := (1+4N^{-2})^{1/2}N^{5/(2N)}$ as in the statement.

Early stopping. Immediately, $\sqrt{d\delta} = \sqrt{\operatorname{Tr}\Sigma}/N$.

Summing the three bounds in (74) gives the bound of Theorem 3,

$$W_2(\mathbb{P}_N, \mathbb{P}) \le \frac{\sqrt{\operatorname{Tr}\Sigma}}{N}\Big(\frac{5}{8}\, a_N\, \rho^{1/N}\log(\rho^{2/5}N) + 1 + \sqrt{8}\Big).$$

Properties of (a_N). Both factors of a_N tend to 1, so $a_N \to 1$. Moreover, $(1+4N^{-2})^{1/2}$ is decreasing in N, and $N^{1/N}$ is decreasing for $N \ge 3$ (as the study of $x \mapsto x^{1/x}$ shows), so $(a_N)_{N\ge 3}$ is decreasing and

$$\sup_{N\ge 1} a_N = \max\big(a_1, a_2, a_3\big) = \max\Big(\sqrt{5},\ \sqrt{2}\times 2^{5/4},\ \tfrac{1}{3}\sqrt{13}\times 3^{5/6}\Big) = 2^{7/4}. \tag{76}$$

In particular $\frac{5}{8}a_N \le 5/2^{5/4}$, which yields a version of the bound with fixed numerical constants. □

We now turn to Proposition 4, which we restate for convenience.

Proposition 4. *With the alternative choice $T = \lambda_{\max}\sqrt{N}$, early-stopping threshold $\delta = \lambda_{\max}/N^2$, and geometric time discretization grid $\tau_k = T(\delta/T)^{k/N}$ for $k \in \{0,\dots,N\}$, the diffusion algorithm* (6) *satisfies, for every $N \ge 1$,*

$$W_2(\mathbb{P}_N, \mathbb{P}) \le \frac{\sqrt{d\lambda_{\max}}}{N}\left(\frac{5}{2^{5/4}}\log N + \Big(\frac{5}{4}\Big)^{3/2} + 1\right).$$

Proof. Let $N \ge 1$ and set $T = \lambda_{\max}\sqrt{N}$ and $\delta = \lambda_{\max}/N^2$ as in the statement. The condition $T > \delta$ of Proposition C.1 holds as soon as $N \ge 2$. For $N = 1$, we have $T = \delta$; as in the proof of Theorem 3, the algorithm performs a single step of size $\eta_1 = 0$, and $\mathbb{P}_1 = \mathcal{N}(0, \lambda_{\max}\,\mathrm{I})$. Hence, $W_2(\mathbb{P}_1, \mathbb{P}) \le \sqrt{d\lambda_{\max}}$ and the bound of Proposition 4 holds. As before, this degenerate case only serves to state the proposition for all $N \ge 1$.

From now on, we assume $N \ge 2$, so that $T > \delta$ and Proposition C.1 applies:

$$W_2(\mathbb{P}_N, \mathbb{P}) \le \frac{\sqrt{\operatorname{Tr}((\Sigma+\delta\,\mathrm{I})^3\Sigma^2)}}{T^2} + \frac{\sqrt{\operatorname{Tr}\Sigma + 4d\delta}}{4N}\Big(\frac{T}{\delta}\Big)^{1/N}\log\frac{T}{\delta} + \sqrt{d\delta}. \tag{77}$$

We bound the three terms in turn.

Initialization. Bounding each eigenvalue λ_i by $\lambda_{\max}$, we have

$$\frac{\sqrt{\operatorname{Tr}((\Sigma+\delta\,\mathrm{I})^3\Sigma^2)}}{T^2} \le \frac{\sqrt{(1+N^{-2})^3 d\lambda_{\max}^5}}{\lambda_{\max}^2 N} = (1+N^{-2})^{3/2}\frac{\sqrt{d\lambda_{\max}}}{N}.$$

Since $N \ge 2$, we get

$$\frac{\sqrt{\operatorname{Tr}((\Sigma+\delta\,\mathrm{I})^3\Sigma^2)}}{T^2} \le \Big(\frac{5}{4}\Big)^{3/2}\frac{\sqrt{d\lambda_{\max}}}{N}.$$

Discretization. Since $T/\delta = N^{5/2}$, using $\operatorname{Tr}\Sigma \le d\lambda_{\max}$ gives

$$\begin{aligned}
\frac{\sqrt{\operatorname{Tr}\Sigma + 4d\delta}}{4N}\Big(\frac{T}{\delta}\Big)^{1/N}\log\frac{T}{\delta} &\le \frac{\sqrt{d\lambda_{\max}(1+4N^{-2})}}{4N}N^{5/(2N)}\log(N^{5/2}) \\
&= \frac{5}{8}(1+4N^{-2})^{1/2}N^{5/(2N)}\sqrt{d\lambda_{\max}}\frac{\log N}{N} \\
&= \frac{5}{8}a_N\sqrt{d\lambda_{\max}}\frac{\log N}{N} \\
&\le \frac{5}{2^{5/4}}\sqrt{d\lambda_{\max}}\frac{\log N}{N},
\end{aligned}$$

where the last inequality is (76).

Early stopping. Immediately, $\sqrt{d\delta} = \sqrt{d\lambda_{\max}}/N$.

Summing the three bounds in (77) then yields the announced bound. □

The equivalent (7), on which the asymptotic exactness claimed in Theorem 3 rests, is the object of the next section.

C.4 Asymptotic equivalent

We prove the following proposition, of which the asymptotic equivalent of Theorem 3 is a special case. It also covers the hyperparameters of Proposition 4, namely $A = B = \lambda_{\max}$, so that the asymptotic equivalent (7) holds in that case as well.

Proposition C.7. *Let $A, B > 0$ be constants independent of N. With hyperparameters of the form*

$$T = A\sqrt{N}, \quad \delta = \frac{B}{N^2},$$

and the geometric discretization $\tau_k = T(\delta/T)^{k/N}$, the error of the diffusion algorithm satisfies

$$W_2(\mathbb{P}_N, \mathbb{P}) \underset{N\to\infty}{\sim} \frac{5}{8}\sqrt{\operatorname{Tr}\Sigma}\,\frac{\log N}{N}.$$

Proof. Let $T = A\sqrt{N}$, $\delta = BN^{-2}$ and $\tau_k = T(\delta/T)^{k/N}$ as in the statement, and take N large enough that $T > \delta$. By Lemma A.3,

$$W_2^2(\mathbb{P}_N, \mathbb{P}) = \sum_{i=1}^{d}\left(\sqrt{\lambda_i^{(N)}} - \sqrt{\lambda_i}\right)^2 = \sum_{i=1}^{d}\left(\frac{\lambda_i^{(N)} - \lambda_i}{\sqrt{\lambda_i^{(N)}} + \sqrt{\lambda_i}}\right)^2, \tag{78}$$

so we need an equivalent of $\lambda^{(N)} - \lambda$, which we split as

$$\lambda^{(N)} - \lambda = (\lambda^{(N)} - \tilde{\lambda}^{(N)}) + (\tilde{\lambda}^{(N)} - (\lambda + \delta)) + \delta.$$

By (68), the first term is $-\left(\frac{\lambda+\delta}{\lambda+T}\right)^2\lambda = O(T^{-2}) = O(N^{-1})$, and the third is $\delta = O(N^{-2})$. Hence

$$\lambda^{(N)} - \lambda = \tilde{\lambda}^{(N)} - (\lambda + \delta) + O\left(\frac{1}{N}\right), \tag{79}$$

and by (65) the remaining term is

$$\tilde{\lambda}^{(N)} - (\lambda + \delta) = (\lambda + \delta)^2\Delta. \tag{80}$$

Everything thus reduces to an equivalent of the quadrature error Δ, which we obtain by squeezing it between two quantities with the same asymptotics.

Upper bound. By (72), since the integrand is positive,

$$\begin{aligned}\Delta &\le (q^{-1} - 1)\int_\delta^T \frac{\tau\,\mathrm{d}\tau}{(\lambda+\tau)^3} \le (q^{-1} - 1)\int_0^\infty \frac{\tau\,\mathrm{d}\tau}{(\lambda+\tau)^3} \\ &= (q^{-1} - 1)\left[\frac{\lambda}{2(\lambda+\tau)^2} - \frac{1}{\lambda+\tau}\right]_0^\infty = \frac{q^{-1} - 1}{2\lambda},\end{aligned}$$

and, since $\frac{1}{N}\log\frac{AN^{5/2}}{B} \to 0$,

$$q^{-1} - 1 = \left(\frac{AN^{5/2}}{B}\right)^{1/N} - 1 = \exp\left(\frac{1}{N}\log\frac{AN^{5/2}}{B}\right) - 1 \sim \frac{1}{N}\log\frac{AN^{5/2}}{B} \sim \frac{5}{2}\frac{\log N}{N}. \tag{81}$$

Lower bound. Since $q \le 1$ and $\tau_j = q\tau_{j-1}$, for every $j \ge 1$,

$$f(\tau_j) = \frac{q\tau_{j-1}}{\lambda + q\tau_{j-1}} \ge \frac{q\tau_{j-1}}{\lambda + \tau_{j-1}} = qf(\tau_{j-1}),$$

where $f : \tau \mapsto \frac{\tau}{\lambda+\tau}$ is the increasing function of Section C.2. Returning to (70) and using $(q^{-1} - 1)q = 1 - q$,

$$\begin{aligned}\Delta &\ge (1-q)\sum_{j=1}^{N} f(\tau_{j-1})\left(\frac{1}{\lambda+\tau_j} - \frac{1}{\lambda+\tau_{j-1}}\right) = (1-q)\sum_{j=1}^{N} f(\tau_{j-1})\int_{\tau_j}^{\tau_{j-1}} \frac{\mathrm{d}\tau}{(\lambda+\tau)^2} \\ &\ge (1-q)\sum_{j=1}^{N}\int_{\tau_j}^{\tau_{j-1}} \frac{f(\tau)\,\mathrm{d}\tau}{(\lambda+\tau)^2} = (1-q)\int_\delta^T \frac{\tau\,\mathrm{d}\tau}{(\lambda+\tau)^3},\end{aligned}$$

the second inequality using $f(\tau) \le f(\tau_{j-1})$ on $[\tau_j, \tau_{j-1}]$. As $\delta \to 0$ and $T \to \infty$, the integrability of $\tau \mapsto \tau(\lambda+\tau)^{-3}$ gives

$$\int_\delta^T \frac{\tau \,\mathrm{d}\tau}{(\lambda+\tau)^3} \underset{N\to\infty}{\longrightarrow} \int_0^\infty \frac{\tau \,\mathrm{d}\tau}{(\lambda+\tau)^3} = \frac{1}{2\lambda},$$

while $q \to 1$ and (81) give $1 - q = q(q^{-1}-1) \sim \frac{5}{2}\frac{\log N}{N}$.

Conclusion. We have shown

$$\underbrace{(1-q)\int_\delta^T \frac{\tau \,\mathrm{d}\tau}{(\lambda+\tau)^3}}_{\sim \frac{5}{4\lambda}\frac{\log N}{N}} \le \Delta \le \underbrace{\frac{q^{-1}-1}{2\lambda}}_{\sim \frac{5}{4\lambda}\frac{\log N}{N}}, \quad \text{hence} \quad \Delta \underset{N\to\infty}{\sim} \frac{5}{4\lambda}\frac{\log N}{N}.$$

Plugging this into (80) and then (79), and using $\delta \to 0$,

$$\lambda^{(N)} - \lambda \sim \tilde\lambda^{(N)} - (\lambda+\delta) \sim \frac{5\lambda}{4}\frac{\log N}{N}.$$

In particular $\lambda^{(N)} \to \lambda$, so the denominators in (78) tend to $2\sqrt{\lambda}$ and

$$W_2^2(\mathbb{P}_N, \mathbb{P}) \sim \sum_{i=1}^d \frac{(\lambda_i^{(N)} - \lambda_i)^2}{4\lambda_i} \sim \left(\frac{5}{8}\frac{\log N}{N}\right)^2 \sum_{i=1}^d \lambda_i = \left(\frac{5}{8}\frac{\log N}{N}\right)^2 \operatorname{Tr}\Sigma,$$

which is the announced equivalent. □

Proof of Theorem 3, asymptotic equivalent and exactness. The hyperparameters of Theorem 3 are of the form of Proposition C.7 with $A = (\operatorname{Tr}(\Sigma^5)/\operatorname{Tr}\Sigma)^{1/4}$ and $B = \operatorname{Tr}\Sigma/d$, which gives the equivalent (7). Dividing the non-asymptotic bound by $\frac{5}{8}\sqrt{\operatorname{Tr}\Sigma}\frac{\log N}{N}$, the ratio of the bound to the error is asymptotically

$$a_N\, \rho^{1/N}\, \frac{\log(\rho^{2/5}N)}{\log N} + \frac{8(1+\sqrt{8})}{5\log N} = a_N\, \rho^{1/N}\left(1 + \frac{2\log\rho}{5\log N}\right) + \frac{8(1+\sqrt{8})}{5\log N} \underset{N\to\infty}{\longrightarrow} 1$$

for fixed Σ, since $a_N \to 1$, $\rho^{1/N} \to 1$ and ρ does not depend on N. The bound is therefore asymptotically exact, which completes the proof of Theorem 3. □

D Proof of Corollary 5 (ε-complexity)

This section proves the following corollary, stated in Section 2.3.

Corollary 5 (ε-complexity)**.** *Let $0 < \varepsilon \le \sqrt{\operatorname{Tr}\Sigma}$. In order for the sampling error to satisfy $W_2(\mathbb{P}_N, \mathbb{P}) \le \varepsilon$, it suffices*

- *for ULA with $\eta = \min\left(\lambda_{\min}, \varepsilon/\sqrt{\operatorname{Tr}(\Sigma^{-1})}\right)$, to take*
$$N \ge \frac{\kappa}{2}\max\left(1, \frac{\lambda_{\min}\sqrt{\operatorname{Tr}(\Sigma^{-1})}}{\varepsilon}\right)\log\frac{\sqrt{8\operatorname{Tr}\Sigma}}{\varepsilon},$$
- *for ULD with $\eta = \min\left(\sqrt{\lambda_{\min}/8}, \frac{7\sqrt{2}}{6}\varepsilon/\sqrt{\lambda_{\min}\operatorname{Tr}(\Sigma^{-1})}\right)$ and $\gamma = \sqrt{8/\lambda_{\min}}$, to take*
$$N \ge 4\kappa\max\left(1, \frac{3}{14}\frac{\lambda_{\min}\sqrt{\operatorname{Tr}(\Sigma^{-1})}}{\varepsilon}\right)\log\frac{16\sqrt{\operatorname{Tr}\Sigma}}{\varepsilon},$$
- *for the diffusion algorithm with $\delta = \varepsilon^2/(9d)$, $T = \sqrt{3/\varepsilon}\operatorname{Tr}\left((\Sigma+\delta\,\mathrm{I})^3\Sigma^2\right)^{1/4}$, and the geometric grid $\tau_k = T(\delta/T)^{k/N}$, to take*
$$N \ge \frac{3e}{4}\frac{\sqrt{\operatorname{Tr}\Sigma + 4\varepsilon^2/9}}{\varepsilon}\log\frac{T}{\delta}.$$

The three ε-complexities follow from the parametric bounds of Propositions B.1, B.7 and C.1 respectively.

Proof. Let $\varepsilon \le \sqrt{\operatorname{Tr}\Sigma}$. In each case we choose the hyperparameters and the budget N so that the error terms are all at most a fixed fraction of ε.

ULA. By Proposition B.1, for every $0 < \eta \le \lambda_{\min}$ and every $N \ge 0$,

$$W_2(\mathbb{P}_N, \mathbb{P}) \le \sqrt{2 \operatorname{Tr} \Sigma}\, e^{-2\eta N/\lambda_{\max}} + \frac{1}{2}\sqrt{\operatorname{Tr}(\Sigma^{-1})}\, \eta, \tag{82}$$

and we make each term at most $\varepsilon/2$. For the discretization term,

$$\frac{1}{2}\sqrt{\operatorname{Tr}(\Sigma^{-1})}\, \eta \le \frac{\varepsilon}{2} \iff \eta \le \frac{\varepsilon}{\sqrt{\operatorname{Tr}(\Sigma^{-1})}}. \tag{83}$$

We take

$$\eta = \min\left(\lambda_{\min}, \frac{\varepsilon}{\sqrt{\operatorname{Tr}(\Sigma^{-1})}}\right), \tag{84}$$

which is admissible for Proposition B.1 and saturates (83) whenever possible: it is the largest admissible step-size, hence the one that makes the exponential factor decay fastest. For the early-stopping term,

$$\sqrt{2 \operatorname{Tr} \Sigma}\, e^{-2\eta N/\lambda_{\max}} \le \frac{\varepsilon}{2} \iff N \ge \frac{\lambda_{\max}}{2\eta} \log \frac{\sqrt{8 \operatorname{Tr} \Sigma}}{\varepsilon},$$

and with (84),

$$\frac{\lambda_{\max}}{2\eta} = \frac{\lambda_{\max}}{2} \max\left(\frac{1}{\lambda_{\min}}, \frac{\sqrt{\operatorname{Tr}(\Sigma^{-1})}}{\varepsilon}\right) = \frac{\kappa}{2} \max\left(1, \frac{\lambda_{\min}\sqrt{\operatorname{Tr}(\Sigma^{-1})}}{\varepsilon}\right).$$

This is the announced condition

$$N \ge \frac{\kappa}{2} \max\left(1, \frac{\lambda_{\min}\sqrt{\operatorname{Tr}(\Sigma^{-1})}}{\varepsilon}\right) \log \frac{\sqrt{8 \operatorname{Tr} \Sigma}}{\varepsilon}.$$

ULD. By the first bound of Proposition B.7, for every $\gamma \ge \sqrt{8/\lambda_{\min}}$, every $\eta \le 1/\gamma$ and every $N \ge 0$,

$$W_2(\mathbb{P}_N, \mathbb{P}) \le 8\sqrt{\operatorname{Tr} \Sigma}\, e^{-2\eta N/(\lambda_{\max}\gamma)} + \frac{6}{7}\sqrt{\operatorname{Tr}(\Sigma^{-1})}\, \frac{\eta}{\gamma}.$$

This bound depends on (η, γ) only through $\phi := \eta/\gamma$, and the constraints translate into $\phi \le 1/\gamma^2 \le \lambda_{\min}/8$; conversely, every $\phi \in (0, \lambda_{\min}/8]$ is reached by an admissible pair, for instance $\gamma = \sqrt{8/\lambda_{\min}}$ and $\eta = \phi\gamma$. We therefore work with the bound

$$W_2(\mathbb{P}_N, \mathbb{P}) \le 8\sqrt{\operatorname{Tr} \Sigma}\, e^{-2\phi N/\lambda_{\max}} + \frac{6}{7}\sqrt{\operatorname{Tr}(\Sigma^{-1})}\, \phi, \quad \phi \le \frac{\lambda_{\min}}{8}.$$

Up to some numerical constants, this bound has the same form as in (82). We proceed as before and obtain the condition of Corollary 5, namely

$$N \ge 4\kappa \max\left(1, \frac{3}{14}\frac{\lambda_{\min}\sqrt{\operatorname{Tr}(\Sigma^{-1})}}{\varepsilon}\right) \log \frac{16\sqrt{\operatorname{Tr} \Sigma}}{\varepsilon},$$

with

$$\eta = \min\left(\sqrt{\frac{\lambda_{\min}}{8}}, \frac{7\sqrt{2}}{6}\frac{\varepsilon}{\sqrt{\lambda_{\min} \operatorname{Tr}(\Sigma^{-1})}}\right), \quad \gamma = \sqrt{\frac{8}{\lambda_{\min}}}.$$

Diffusion. We show the more general statement of Remark 6, namely that, with $\delta = \varepsilon^2/(9d)$, $T = \sqrt{3/\varepsilon}\operatorname{Tr}\left((\Sigma + \delta \mathrm{I})^3\Sigma^2\right)^{1/4}$ and the geometric grid $\tau_k = T(\delta/T)^{k/N}$, an error ε is achieved by taking, for any $r \ge 1$,

$$N \ge r \max\left(1, \frac{3e^{1/r}}{4r}\frac{\sqrt{\operatorname{Tr} \Sigma + 4\varepsilon^2/9}}{\varepsilon}\right) \log \frac{T}{\delta}.$$

By Proposition C.1, for every $N \ge 1$ and every $T > \delta > 0$,

$$W_2(\mathbb{P}_N, \mathbb{P}) \le \frac{\sqrt{\operatorname{Tr}\left((\Sigma + \delta \mathrm{I})^3\Sigma^2\right)}}{T^2} + \frac{\sqrt{\operatorname{Tr} \Sigma + 4d\delta}}{4N}\left(\frac{T}{\delta}\right)^{1/N} \log \frac{T}{\delta} + \sqrt{d\delta},$$

and we make each of the three terms at most $\varepsilon/3$.

Early stopping. We have

$$\sqrt{d\delta} \le \frac{\varepsilon}{3} \iff \delta \le \frac{\varepsilon^2}{9d}.$$

We saturate this constraint by taking

$$\delta = \frac{\varepsilon^2}{9d}. \tag{85}$$

Initialization. We have

$$\frac{\operatorname{Tr}\left((\Sigma+\delta\,\mathrm{I})^3\Sigma^2\right)^{1/2}}{T^2} \le \frac{\varepsilon}{3} \iff T \ge \sqrt{\frac{3}{\varepsilon}}\operatorname{Tr}\left((\Sigma+\delta\,\mathrm{I})^3\Sigma^2\right)^{1/4},$$

so we take

$$T = \sqrt{\frac{3}{\varepsilon}}\operatorname{Tr}\left((\Sigma+\delta\,\mathrm{I})^3\Sigma^2\right)^{1/4}. \tag{86}$$

This choice is compatible with the constraint $T > \delta$. Indeed, with (85) and (86),

$$\frac{T}{\delta} = \left(\frac{3}{\varepsilon}\right)^{5/2} d\operatorname{Tr}\left((\Sigma+\delta\,\mathrm{I})^3\Sigma^2\right)^{1/4} > \left(\frac{3}{\varepsilon}\right)^{5/2} d\operatorname{Tr}(\Sigma^5)^{1/4}. \tag{87}$$

But, by Hölder's inequality, $\operatorname{Tr}(\Sigma^5) \ge (\operatorname{Tr}\Sigma)^5/d^4$, so (87) yields

$$\frac{T}{\delta} > 3^{5/2}\left(\frac{\operatorname{Tr}\Sigma}{\varepsilon^2}\right)^{5/4} > 1, \tag{88}$$

since we assumed $\varepsilon \le \sqrt{\operatorname{Tr}\Sigma}$.

Discretization. It remains to ensure

$$E := \frac{\sqrt{\operatorname{Tr}\Sigma + 4d\delta}}{4N}\left(\frac{T}{\delta}\right)^{1/N}\log\frac{T}{\delta} \le \frac{\varepsilon}{3}.$$

The factor $(T/\delta)^{1/N}$ is what prevents solving $E \le \varepsilon/3$ in closed form; but it tends to 1 as N grows, and it is under control as soon as N exceeds a few multiples of $\log(T/\delta)$. Precisely, for any $r \ge 1$,

$$\left(\frac{T}{\delta}\right)^{1/N} \le e^{1/r} \iff N \ge r\log\frac{T}{\delta}. \tag{89}$$

Under (89) for a given $r \ge 1$,

$$E \le \frac{e^{1/r}}{4}\frac{\sqrt{\operatorname{Tr}\Sigma + 4d\delta}}{N}\log\frac{T}{\delta},$$

and

$$\frac{e^{1/r}}{4}\frac{\sqrt{\operatorname{Tr}\Sigma + 4d\delta}}{N}\log\frac{T}{\delta} \le \frac{\varepsilon}{3} \iff N \ge \frac{3e^{1/r}}{4}\frac{\sqrt{\operatorname{Tr}\Sigma + 4d\delta}}{\varepsilon}\log\frac{T}{\delta}.$$

Both requirements are met as soon as

$$N \ge \max\left(r\log\frac{T}{\delta}, \frac{3e^{1/r}}{4}\frac{\sqrt{\operatorname{Tr}\Sigma + 4d\delta}}{\varepsilon}\log\frac{T}{\delta}\right) = r\max\left(1, \frac{3e^{1/r}}{4r}\frac{\sqrt{\operatorname{Tr}\Sigma + 4\varepsilon^2/9}}{\varepsilon}\right)\log\frac{T}{\delta}, \tag{90}$$

which is the budget of Remark 6. The parameter r tunes a trade-off: a larger r brings the constant $3e^{1/r}/4$ down towards $3/4$, at the price of a higher floor $r\log(T/\delta)$.

Taking $r = 1$, since $\sqrt{\operatorname{Tr}\Sigma + 4\varepsilon^2/9} \ge \frac{2}{3}\varepsilon$, the second argument of the maximum in (90) is at least $\frac{3e}{4} \times \frac{2}{3} = \frac{e}{2} > 1$. The maximum is therefore always attained by the second argument, and (90) reduces to

$$N \ge \frac{3e}{4}\frac{\sqrt{\operatorname{Tr}\Sigma + 4\varepsilon^2/9}}{\varepsilon}\log\frac{T}{\delta}, \tag{91}$$

the budget of Corollary 5. □

Remark D.1 (Order of magnitude of $\log(T/\delta)$)**.** *We can make the logarithmic factor in* (90) *explicit. By* (88)*, we have the lower bound*

$$\log \frac{T}{\delta} \geq \frac{5}{2} \log \frac{3\sqrt{\operatorname{Tr} \Sigma}}{\varepsilon}.$$

For the upper bound, with (85) *we have*

$$\delta \leq \frac{\operatorname{Tr} \Sigma}{9d} \leq \frac{1}{9} \lambda_{\max},$$

whence

$$\operatorname{Tr}((\Sigma + \delta \, \mathrm{I})^3 \Sigma^2) \leq (\lambda_{\max} + \delta)^3 \lambda_{\max} \operatorname{Tr} \Sigma \leq \Big(\frac{10}{9}\Big)^3 \lambda_{\max}^4 \operatorname{Tr} \Sigma.$$

With $\rho = d\lambda_{\max} / \operatorname{Tr} \Sigma \in [1, d]$ *as in Theorem 3,* (87) *yields*

$$\frac{T}{\delta} \leq \Big(\frac{10}{9}\Big)^{3/4} \Big(\frac{3}{\varepsilon}\Big)^{5/2} d\lambda_{\max} (\operatorname{Tr} \Sigma)^{1/4} = 3^{5/2} \Big(\frac{10}{9}\Big)^{3/4} \rho \Big(\frac{\sqrt{\operatorname{Tr} \Sigma}}{\varepsilon}\Big)^{5/2}.$$

Since $\log \big(3^{5/2} (10/9)^{3/4}\big) \approx 2.83 < 3$*, we finally obtain*

$$\frac{5}{2} \log \frac{3\sqrt{\operatorname{Tr} \Sigma}}{\varepsilon} \;\leq\; \log \frac{T}{\delta} \;\leq\; \frac{5}{2} \log \frac{\sqrt{\operatorname{Tr} \Sigma}}{\varepsilon} + \log \rho + 3. \tag{92}$$

In particular, $\log(T/\delta) = \Theta(\log(\sqrt{\operatorname{Tr} \Sigma}/\varepsilon))$ *for fixed* Σ*. Combined with* $\sqrt{\operatorname{Tr} \Sigma + 4\varepsilon^2/9} \leq \frac{\sqrt{13}}{3} \sqrt{\operatorname{Tr} \Sigma}$*, equation* (91) *and the upper bound in* (92) *lead to the fully explicit budget condition*

$$N \geq \frac{\sqrt{13} e}{4} \frac{\sqrt{\operatorname{Tr} \Sigma}}{\varepsilon} \Big(\frac{5}{2} \log \frac{\sqrt{\operatorname{Tr} \Sigma}}{\varepsilon} + \log \rho + 3 \Big),$$

of order $\tilde{O}(\sqrt{\operatorname{Tr} \Sigma}/\varepsilon)$ *with no dependence on* κ*.*

E Proof of Theorem 7 (score learning)

This section proves the following theorem, stated in Section 3.

Theorem 7 (Score learning)**.** *Consider gradient descent initialized at* $A_0 = 0$*, with step-size* $\alpha > 0$*, on both* L *and* L_σ*. Let* $K \geq 1$*.*

- *If* $\alpha < \frac{2}{\hat{\lambda}_{\max}}$*, the iterate* A_K *is invertible and*

$$A_K^{-1} = \hat{\Sigma} + \frac{1}{K\alpha} \Pi_{\operatorname{Ker} \hat{\Sigma}} + O(\rho_1^K),$$

  with $\rho_1 = \max\big(|1 - \alpha \hat{\lambda}_{\min}^+|, |1 - \alpha \hat{\lambda}_{\max}|\big) < 1.$

- *If* $\alpha < \frac{2}{\hat{\lambda}_{\max} + \sigma^2}$*, the iterate* A_K^σ *is invertible and*

$$(A_K^\sigma)^{-1} = \hat{\Sigma} + \sigma^2 \, \mathrm{I} + O(\rho_2^K),$$

  with $\rho_2 = \max\big(|1 - \alpha(\hat{\lambda}_{\min} + \sigma^2)|, |1 - \alpha(\hat{\lambda}_{\max} + \sigma^2)|\big) < 1.$

The proof is a direct computation: we first write both losses in closed form and derive their gradients, then compute the iterates A_K and A_K^σ explicitly, and finally invert these expressions.

E.1 Losses and gradients

We treat the two losses simultaneously by introducing, for $t \geq 0$,

$$L^{(t)}(A) = \frac{1}{m} \sum_{i=1}^m \mathbb{E}_\omega [\nabla \cdot s_A(x_i + \omega) + \frac{1}{2} \|s_A(x_i + \omega)\|^2],$$

with $\omega \sim \mathcal{N}(0, t \, \mathrm{I})$. Hence, $L = L^{(0)}$ and $L_\sigma = L^{(\sigma^2)}$. We first compute the losses and their gradients:

Lemma E.1. *For all $t \geq 0$ and $A \in \mathbb{R}^{d\times d}$, we have*

$$L^{(t)}(A) = \frac{1}{2}\operatorname{Tr}(A\hat{\Sigma}A^\top) - \operatorname{Tr} A + \frac{t}{2}\|A\|_F^2,$$

with gradient

$$\nabla L^{(t)}(A) = A(\hat{\Sigma} + t\,\mathrm{I}) - \mathrm{I}\,.$$

Proof. Fix $t \geq 0$, $A \in \mathbb{R}^{d\times d}$. For any $x \in \mathbb{R}^d$, we have $s_A(x) = -Ax$, hence

$$\begin{aligned}
L^{(t)}(A) &= \frac{1}{m}\sum_{i=1}^m \mathbb{E}_\omega[-\operatorname{Tr} A + \frac{1}{2}\|A(x_i+\omega)\|^2] \\
&= -\operatorname{Tr} A + \frac{1}{2m}\sum_{i=1}^m \mathbb{E}_\omega[\|Ax_i\|^2 + 2x_i^\top A^\top A\omega + \|A\omega\|^2] \\
&= -\operatorname{Tr} A + \frac{1}{2}\mathbb{E}_\omega[\omega^\top A^\top A\omega] + \frac{1}{2m}\sum_{i=1}^m \operatorname{Tr}(Ax_ix_i^\top A^\top) + \frac{1}{m}\left(\sum_{i=1}^m x_i^\top A^\top A\right)\mathbb{E}\omega \\
&= -\operatorname{Tr} A + \frac{1}{2}\operatorname{Tr}(t\,\mathrm{I}\,AA^\top) + \frac{1}{2}\operatorname{Tr}(A\hat{\Sigma}A^\top) \\
&= \frac{1}{2}\operatorname{Tr}(A\hat{\Sigma}A^\top) - \operatorname{Tr} A + \frac{t}{2}\|A\|_F^2.
\end{aligned}$$

Denoting

$$f : A \longmapsto \frac{1}{2}\operatorname{Tr}(A\hat{\Sigma}A^\top), \quad g : A \longmapsto -\operatorname{Tr} A, \quad h : A \longmapsto \frac{1}{2}\|A\|_F^2,$$

we compute, using the symmetry of $\hat{\Sigma}$,

$$\nabla f(A) = A\hat{\Sigma}, \quad \nabla g(A) = -\mathrm{I}, \quad \nabla h(A) = A.$$

The gradient of $L^{(t)}$ then follows:

$$\nabla L^{(t)}(A) = A(\hat{\Sigma} + t\,\mathrm{I}) - \mathrm{I}\,.$$

□

Note that in particular, we retrieve the fact that

$$\forall A, \quad L_\sigma(A) = L(A) + \frac{\sigma^2}{2}\|A\|_F^2,$$

so the loss L_σ is indeed a ridge-type regularization of L, as claimed in Section 3.

E.2 Closed form of the iterates

Let $\hat{\Sigma}_t = \hat{\Sigma} + t\,\mathrm{I}$ and denote by $(\hat{\lambda}_i)_{1\leq i\leq d}$ the eigenvalues of $\hat{\Sigma}$, with an orthonormal eigenbasis (u_i); $\hat{\Sigma}_t$ has eigenvalues $\hat{\lambda}_i + t$ in the same basis. Moreover, denote by $\hat{\Sigma}^+$ the Moore-Penrose inverse of $\hat{\Sigma}$, *i.e.*, the matrix defined by

$$\hat{\Sigma}^+ = \sum_{\substack{i=1 \\ \hat{\lambda}_i \neq 0}}^d \tfrac{1}{\hat{\lambda}_i} u_i u_i^\top.$$

Proposition E.2. *Let $\alpha > 0$ be the gradient step-size. For every $K \geq 0$, the iterate of gradient descent on $L^{(t)}$ is*

$$A_K^{(t)} = \alpha\sum_{j=0}^{K-1}(\mathrm{I} - \alpha\hat{\Sigma}_t)^j = \sum_{i=1}^d \psi_K(\hat{\lambda}_i + t)u_iu_i^\top,$$

with

$$\psi_K : x \longmapsto \begin{cases} \frac{1-(1-\alpha x)^K}{x} & \text{if } x \neq 0, \\ K\alpha & \text{otherwise.} \end{cases}$$

Hence, for $t > 0$ we have

$$A_K^{(t)} = \hat{\Sigma}_t^{-1}\big(\mathrm{I} - (\mathrm{I} - \alpha\hat{\Sigma}_t)^K\big), \tag{93}$$

while

$$A_K^{(0)} = \hat{\Sigma}^{+}\big(\mathrm{I} - (\mathrm{I} - \alpha\hat{\Sigma})^K\big) + K\alpha\Pi_{\operatorname{Ker}\hat{\Sigma}}, \tag{94}$$

with $\hat{\Sigma}^{+} = \hat{\Sigma}^{-1}$ and $\Pi_{\operatorname{Ker}\hat{\Sigma}} = 0$ when $\hat{\Sigma}$ is invertible.

Proof. Let $K \geq 0$. The process $(A_j^{(t)})_{0\leq j\leq K}$ is defined by the gradient descent on the loss $L^{(t)}$ with step-size α, initialized at $A_0^{(t)} = 0$:

$$\begin{cases} \forall j, \quad A_{j+1}^{(t)} = A_j^{(t)} - \alpha\nabla L^{(t)}(A_j^{(t)}), \\ A_0^{(t)} = 0. \end{cases}$$

Therefore, by Lemma E.1, we have

$$\forall j, \quad A_{j+1}^{(t)} = A_j^{(t)}(\mathrm{I} - \alpha\hat{\Sigma}_t) + \alpha\,\mathrm{I},$$

thus

$$\begin{aligned} A_K^{(t)} &= \alpha\sum_{j=0}^{K-1}(\mathrm{I} - \alpha\hat{\Sigma}_t)^j \\ &= \sum_{i=1}^{d}\left(\alpha\sum_{j=0}^{K-1}\big(1 - \alpha(\hat{\lambda}_i + t)\big)^j u_i u_i^\top\right). \end{aligned}$$

The equality

$$A_K^{(t)} = \sum_{i=1}^{d}\psi_K(\hat{\lambda}_i + t)u_i u_i^\top$$

is then a simple consequence of the identity

$$\sum_{j=0}^{K-1} x^j = \begin{cases} \frac{1-x^K}{1-x} & \text{if } x \neq 1, \\ K & \text{otherwise,} \end{cases}$$

and the relations (93) and (94) follow. □

E.3 Inverses of the iterates

We can now prove Theorem 7.

Proof of Theorem 7. Fix $K \geq 1$. By Proposition E.2, we have for any $t \geq 0$, $\alpha > 0$:

$$A_K^{(t)} = \sum_{i=1}^{d}\psi_K(\hat{\lambda}_i + t)u_i u_i^\top.$$

In particular, for every $\alpha < \frac{2}{\hat{\lambda}_{\max}+t}$, the eigenvalues $\psi_K(\hat{\lambda}_i + t)$ of $A_K^{(t)}$ are all positive, thus non-zero, and $A_K^{(t)}$ is invertible with

$$(A_K^{(t)})^{-1} = \sum_{i=1}^{d}\frac{u_i u_i^\top}{\psi_K(\hat{\lambda}_i + t)}. \tag{95}$$

Loss L ($t = 0$). We have

$$\psi_K(\hat\lambda_i) = \begin{cases} \frac{1-(1-\alpha\hat\lambda_i)^K}{\hat\lambda_i} & \text{if } \hat\lambda_i \neq 0, \\ K\alpha & \text{otherwise,} \end{cases}$$

equation (95) therefore yields for $t = 0$:

$$\begin{aligned} A_K^{-1} &= \sum_{\substack{i=1 \\ \hat\lambda_i \neq 0}}^{d} \frac{u_i u_i^\top}{\psi_K(\hat\lambda_i)} + \sum_{\substack{i=1 \\ \hat\lambda_i = 0}}^{d} \frac{u_i u_i^\top}{\psi_K(\hat\lambda_i)} \\ &= \sum_{\substack{i=1 \\ \hat\lambda_i \neq 0}}^{d} \frac{\hat\lambda_i}{1-(1-\alpha\hat\lambda_i)^K} u_i u_i^\top + \sum_{\substack{i=1 \\ \hat\lambda_i = 0}}^{d} \frac{1}{K\alpha} u_i u_i^\top \\ &= \sum_{\substack{i=1 \\ \hat\lambda_i \neq 0}}^{d} \hat\lambda_i u_i u_i^\top + \sum_{\substack{i=1 \\ \hat\lambda_i \neq 0}}^{d} \frac{\hat\lambda_i (1-\alpha\hat\lambda_i)^K}{1-(1-\alpha\hat\lambda_i)^K} u_i u_i^\top + \frac{1}{K\alpha} \Pi_{\operatorname{Ker}\hat\Sigma} \\ &= \hat\Sigma + \frac{1}{K\alpha}\Pi_{\operatorname{Ker}\hat\Sigma} + \sum_{\substack{i=1 \\ \hat\lambda_i \neq 0}}^{d} \frac{\hat\lambda_i (1-\alpha\hat\lambda_i)^K}{1-(1-\alpha\hat\lambda_i)^K} u_i u_i^\top. \end{aligned} \tag{96}$$

Denote by R_K the last sum. It is a symmetric matrix, diagonal in the basis (u_i), whose eigenvalues are the numbers

$$r_i := \frac{\hat\lambda_i (1-\alpha\hat\lambda_i)^K}{1-(1-\alpha\hat\lambda_i)^K}, \quad \hat\lambda_i \neq 0.$$

Since $x \mapsto |1-\alpha x|$ is convex, its maximum over the interval $[\hat\lambda_{\min}^+, \hat\lambda_{\max}]$ is attained at one of the endpoints: $|1-\alpha\hat\lambda_i| \leq \max\big(|1-\alpha\hat\lambda_{\min}^+|, |1-\alpha\hat\lambda_{\max}|\big) = \rho_1$ for every i with $\hat\lambda_i \neq 0$. The assumption $\alpha < 2/\hat\lambda_{\max}$ gives $\rho_1 < 1$; therefore

$$|r_i| \leq \frac{\hat\lambda_i \rho_1^K}{1-\rho_1^K} \leq \frac{\hat\lambda_i}{1-\rho_1}\rho_1^K.$$

Hence, denoting $\|\cdot\|_2$ the spectral norm, we have

$$\|R_K\|_2 \leq \frac{\hat\lambda_{\max}}{1-\rho_1}\rho_1^K = O(\rho_1^K),$$

which, along with (96), leads to the first claim of Theorem 7.

Loss L_σ ($t = \sigma^2$). The argument is the same, without any kernel to single out; the inverse formula (95) gives

$$\begin{aligned} (A_K^\sigma)^{-1} &= \sum_{i=1}^{d} \frac{\hat\lambda_i + \sigma^2}{1-(1-\alpha(\hat\lambda_i+\sigma^2))^K} u_i u_i^\top \\ &= \sum_{i=1}^{d} \left(\hat\lambda_i + \sigma^2 + \frac{(\hat\lambda_i+\sigma^2)(1-\alpha(\hat\lambda_i+\sigma^2))^K}{1-(1-\alpha(\hat\lambda_i+\sigma^2))^K} \right) u_i u_i^\top \\ &= \hat\Sigma + \sigma^2 \operatorname{I} + R_K^\sigma, \end{aligned} \tag{97}$$

where

$$R_K^\sigma = \sum_{i=1}^{d} \frac{(\hat\lambda_i+\sigma^2)(1-\alpha(\hat\lambda_i+\sigma^2))^K}{1-(1-\alpha(\hat\lambda_i+\sigma^2))^K} u_i u_i^\top.$$

Under the assumption $\alpha < 2/(\hat\lambda_{\max}+\sigma^2)$,

$$\|R_K^\sigma\|_2 \leq \frac{\hat\lambda_{\max}+\sigma^2}{1-\rho_2}\rho_2^K = O(\rho_2^K),$$

with $\rho_2 = \max\big(|1-\alpha(\hat\lambda_{\min}+\sigma^2)|, |1-\alpha(\hat\lambda_{\max}+\sigma^2)|\big)$. With (97), this gives the second claim of Theorem 7. □

LLM USAGE

LLMs were used to translate the appendix into English, which was initially written in another language by the authors. They were also used to assist in checking the proofs. All LLM-assisted content was reviewed and verified by the authors.