Theoretical guidelines for annealed Langevin dynamics in compositional simulation-based inference

Camille Touron
Univ. Grenoble Alpes, Inria
CNRS, Grenoble INP, LJK, France
camille.touron@inria.fr

Gabriel V. Cardoso
Geostatistics team, Centre for geosciences and geoengineering
Mines Paris, PSL University, Fontaineableau, France
gabriel.victorino_cardoso@minesparis.psl.eu

Julyan Arbel
Univ. Grenoble Alpes, Inria
CNRS, Grenoble INP, LJK, France
julyan.arbel@inria.fr

Pedro L. C. Rodrigues
Univ. Grenoble Alpes, Inria
CNRS, Grenoble INP, LJK, France
pedro.rodrigues@inria.fr

Abstract

Compositional score-based approaches to simulation-based inference (SBI) approximate the posterior over a shared parameter given n independent observations by aggregating individually learned posterior scores: currently, there are two main propositions of such methods (Geffner et al. [13], Linhart et al. [16]). As the resulting composite score does not correspond to the score of any distribution along the forward diffusion path of the true multi-observation posterior, sampling from it via a reverse SDE leads to an irreducible bias. Annealed Langevin dynamics provides a principled alternative: it treats the composite score as the genuine score of a sequence of tractable bridging densities and samples from them in succession. When properly tuned, it could lead to a controllable bias. However, its hyperparameters, namely step sizes, the number of steps per level, and the number of annealing levels, have so far been chosen empirically. We derive Wasserstein bounds for annealed Langevin with approximate scores and translate them into explicit decision rules for these hyperparameters that guarantee a prescribed sampling accuracy, while highlighting different theoretical aspects of each composite score formulation. In the Gaussian setting, we obtain closed-form expressions for all relevant quantities and prove that the bridging densities of Linhart et al. [16] consistently admit larger step sizes and require fewer total Langevin steps than those of Geffner et al. [13]. Furthermore, we show empirically that the tuning obtained in the Gaussian setting generalizes to more complex problems, thus providing a well-understood and theoretically grounded starting point for practitioners using compositional score-based approaches.

1 Introduction

In simulation-based inference (SBI), the goal is to approximate the posterior distribution $p(\theta \mid x)$ of a stochastic model whose likelihood is intractable but which can be simulated [6, 8]. A recurrent challenge arises when n independent observations $x_1, \dots, x_n$ share a common unknown parameter $\theta^\star$:

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

as n grows, the multi-observation posterior $p(\theta \mid x_{1:n})$ concentrates, and one wants inference to sharpen accordingly. *Pooling strategies* train a single network on all n observations jointly [19, 20], but their simulation budget scales with n and the network must be retrained for each new choice of n. *Compositional strategies* [13, 16] instead combine individual posterior estimators trained once on single-observation data, amortizing over n at no additional simulation cost: a key practical advantage that motivates our focus on this paradigm.

Score-based diffusion models [24, 26, 22] provide a natural starting point: for a single observation, one learns the individual posterior score $\nabla_\theta \log p_t(\theta \mid x_i)$ along a forward diffusion path and uses it in a reverse SDE to draw samples from $p(\theta \mid x_i)$. In the multi-observation setting, the tempting shortcut is to aggregate these individual scores into a composite score and run the same reverse SDE to target $p(\theta \mid x_{1:n})$. This shortcut is, however, theoretically unjustified. The reverse SDE must be driven by the *exact* score of the joint forward diffusion of $p(\theta \mid x_{1:n})$, i.e. $\nabla_\theta \log p_t(\theta \mid x_{1:n})$, which satisfies the Fokker–Planck equation of that specific process. A composite of individual scores does *not* satisfy this equation: it is the score of a different distribution $\pi_t \neq p_t(\theta \mid x_{1:n})$ that does not lie on the forward diffusion path of the target. Using such a score in a reverse SDE therefore produces samples from the wrong distribution, as detailed in Appendix B.5, and no amount of additional reverse steps can correct this structural mismatch [9]. This also explains why some works [26, 14] attempt to correct this bias by incorporating additional *Langevin* refinement steps into the standard reverse diffusion process. However, this comes at the cost of sacrificing the efficiency of diffusion as a fast sampling method while keeping the challenge of tuning the Langevin hyperparameters.

Several works recognized this problem and proposed composite scores that, while still mismatched with the true forward diffusion, deliberately define their own tractable sequence of bridging densities [13, 1, 16, 14]. In this paper, we focus on the following two main composite score formulations considered in the literature. Geffner et al. [13] use

$$s_t^G(\theta \mid x_{1:n}) = (1-n)\,\nabla_\theta \log \lambda_t(\theta) + \textstyle\sum_{i=1}^n \nabla_\theta \log p_t(\theta \mid x_i), \tag{1}$$

whose associated bridging density π_t^G is well-defined and tractable. Linhart et al. [16] incorporate a second-order Gaussian approximation of the backward diffusion kernels to obtain s_t^L, whose bridging density π_t^L is closer to the true diffused posterior $p_t(\theta \mid x_{1:n})$, at the cost of estimating per-observation and prior covariance matrices $\Sigma_{t,i}$, $\Sigma_{t,\lambda}$. Their composite score reads

$$s_t^L(\theta \mid x_{1:n}) = \Lambda_t^{-1}\Big(\sum_{i=1}^n \Sigma_{t,i}^{-1}\nabla_\theta \log p_t(\theta \mid x_i) + (1-n)\Sigma_{t,\lambda}^{-1}\nabla_\theta \log \lambda_t(\theta)\Big), \tag{2}$$

where Λ_t is a linear combination of $\Sigma_{t,i}$, $\Sigma_{t,\lambda}$ and $\lambda_t(\theta)$ denotes the diffused prior. s_t^G and s_t^L coincide with the score of the target multi-observation posterior when $t = 0$ but deviate from its diffused versions for $t > 0$. In both cases, the key insight is to stop trying to target $p_t(\theta \mid x_{1:n})$ through a reverse SDE, and instead *define* π_t^G or π_t^L as the actual sampling target: the question becomes how to efficiently draw samples from these bridging densities given access to their scores.

Annealed Langevin dynamics is one theoretically sound alternative [24]. It runs T successive Unadjusted Langevin Algorithm (ULA) chains, each with step size h_{t_p} and k_{t_p} number of steps,

$$\theta^{(j+1)} = \theta^{(j)} + h_{t_p}\,\nabla_\theta \log \pi_{t_p}(\theta^{(j)}) + \sqrt{2h_{t_p}}\,z^{(j)}, \quad z^{(j)} \sim \mathcal{N}(0, I_d), \;\; j = 0, \dots, k_{t_p} - 1 \tag{3}$$

on bridging densities $\{\pi_{t_p}\}_{p=0}^T$ interpolating from $\pi_{t_T} = \mathcal{N}(0, I_d)$ to the target π_{t_0} (Figure 1). It requires only the score of each π_{t_p}, which the composite scores evaluated at time t_p provide by construction, and its gradual annealing improves mixing for concentrated posteriors at large n. Since Song et al. [26] established the SDE interpretation of diffusion models, the field has largely shifted to reverse-SDE and DDPM/DDIM samplers [15], which require far fewer score evaluations. Those samplers, however, are theoretically grounded only when the score is computed or trained to match the *exact* forward score $\nabla_\theta \log p_t(\theta \mid x_{1:n})$, a condition violated here by construction. Annealed Langevin avoids this mismatch entirely, at the cost of more score evaluations and the need to choose, at each level t_p, a step size h_{t_p}, a number of steps k_{t_p}, and a total number of levels T. Selecting these hyperparameters in a principled way, with guarantees on the resulting sampling quality, has remained an open problem; both Geffner et al. [13] and Linhart et al. [16] resort to empirical tuning. In the context of image generation, Song and Ermon [25] propose some techniques to choose these hyperparameters, most relying on empirical validation, but leaving theoretical guarantee on the sampling quality unaddressed. Our paper closes that gap.

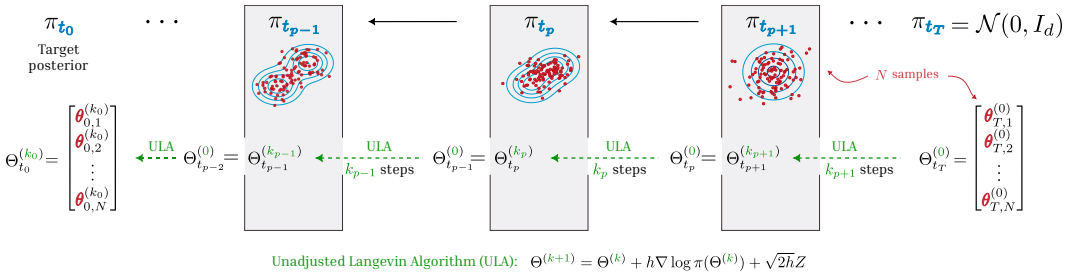

Figure 1: The Annealed Langevin algorithm consists in T ULA steps in succession. It allows to sample from bridging densities $\{\pi_{t_p}\}_{p=0}^{T-1}$, represented in blue: the first ULA yields samples (in red) from $\pi_{t_{T-1}}$, the second from $\pi_{t_{T-2}}$, etc... and the last one from the target density π_{t_0}. Green dashed arrows stand for the T ULA: each of them has to define its own constant step size h_{t_p} and number of steps k_{t_p} (simply denoted k_p for readability concerns). The first process starts from Gaussian samples; then the start point of each process corresponds to the last iterates of the previous process, which is symbolized by the equal signs on the scheme.

Contributions. We provide the first theoretically grounded guidelines for selecting the hyperparameters of annealed Langevin dynamics in compositional SBI.

(i) Wasserstein bounds (Section 2). Building on Dalalyan and Karagulyan [7], we derive a per-level bound for a single ULA chain accounting for both composite score error and the Wasserstein distance between consecutive bridging densities (Proposition 1); composing T such bounds yields a global sampling guarantee (Proposition 2). We show that log-concavity and smoothness of the target are inherited by both Geffner's and Linhart's bridging densities, in closed form for Gaussians (Lemma 2) and under mild conditions in general (Lemma 1).

(ii) Hyperparameter decision rule. Given a target accuracy $\gamma > 0$ and a bias–variance split $\omega \in (0, 1)$, we derive explicit formulas for the hyperparameters h_t and k_t ensuring the Wasserstein error stays below γ at every annealing level $t \in \{t_0, \dots, t_T\}$, with the choice of compositional score entering directly as a design parameter (Section 3). We illustrate the practical use of our decision rule in Gaussian and non-Gaussian settings (resp. in Sections 4 and 5).

(iii) Geffner vs. Linhart. In the Gaussian setting, we show that Linhart's algorithm always admits larger step sizes and its bridging densities move less between consecutive levels near $t = 1$ (Section 4), requiring fewer total steps. We extend the comparison to non-Gaussian settings empirically (Section 5). In summary, Linhart's composite score yields a more efficient sampler.

2 Wasserstein error bounds for annealed Langevin dynamics

To measure the sampling quality and the rate of convergence of ULA, different works derive upper-bounds of the Wasserstein distance [10] or the KL divergence [29] between the target distribution π_{t_0} and the distribution of the kth iterate $\theta^{(k)}$ generated by the algorithm: these bounds rely on assumptions on π_{t_0} (smoothness, log-concavity or log-Sobolev inequality) and are valid only with an exact score function. Theorem 4 of Dalalyan and Karagulyan [7] is one of the first results to establish a bound when the score is not known accurately. Based on these results, we are able to upper-bound the Wasserstein error of an annealed Langevin algorithm with T annealing steps, as we show below.

We denote by $\mathcal{L}(\theta^{(k)})$ the distribution of the kth iterate of a Langevin process. A density π is m-log concave and M-smooth when $mI_d \preceq -\nabla^2_\theta \log \pi(\theta) \preceq MI_d$ for all θ in the support of π. It is said to be Gaussian log-Lipschitz if it admits a density with respect to the Gaussian measure whose logarithm is Lipschitz continuous. The first proposition bounds the Wasserstein error at a *given* annealing level $t_p \in \{t_0, \dots, t_T\}$.

Proposition 1 (Intermediate Wasserstein error bound) *Suppose that π_{t_p} is m_p-log concave and M_p-smooth in $\mathbb{R}^d$. Let k_p be the number of steps, h_p the constant step size such that $h_p < \frac{2}{M_p + m_p}$, $B_p = 1.65\frac{M_p}{m_p}(h_p d)^{0.5} + \frac{\epsilon_{DSM}}{m_p}$ and ϵ_{DSM} upper bounds the L_2-score-matching error. Then,*

$$\mathcal{W}_2(\mathcal{L}(\theta^{(k_p)}_{t_p}), \pi_{t_p}) \leq (1 - m_p h_p)^{k_p} \left(\mathcal{W}_2(\mathcal{L}(\theta^{(k_{p+1})}_{t_{p+1}}), \pi_{t_{p+1}}) + \mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p}) \right) + B_p.$$

Proof is given in Appendix A.1. For readability concerns, we replace the index t_p simply by p for the constants M, m, h, k, B, but they remain level-dependent. With respect to Theorem 4 in Dalalyan and Karagulyan [7], we have included a new term based on the Wasserstein distance between two consecutive intermediate densities $\mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p})$ and also simplified the role of the bias of the score approximation to consider a biased but deterministic score estimate controlled by the score-matching error ϵ_{DSM}. Note that in the case of compositional scores, we only have access to the score-matching errors of individual score estimates $s_\phi(\theta, x_i, t) \approx \nabla_\theta \log p_t(\theta \mid x_i)$: it is, however, possible to obtain a bound of the compositional score error by composing these individual errors (see Section 5).

We also emphasize that B_p depends only on the choice of step size h_p, while the number of steps k_p influences only the Wasserstein distances: this observation will motivate the guidelines we propose in Section 3. The following proposition gives a theoretical Wasserstein guarantee for the annealed Langevin algorithm by composing the intermediate bound T times, while keeping the same notations as before. Proof is given in Appendix A.2.

Proposition 2 (Global Wasserstein error bound) *Suppose that π_{t_p} is m_p-log concave and M_p-smooth for all $p \in \{0, \dots, T-1\}$. Let h_p be the constant step size at the pth annealing level such that $h_p < \frac{2}{M_p + m_p}$ and k_p be the number of steps. Then,*

$$\mathcal{W}_2(\mathcal{L}(\theta_{t_0}^{(k_0)}), \pi_{t_0}) \leq \sum_{p=0}^{T-1} \left(\prod_{s=0}^{p} (1 - m_s h_s)^{k_s} \right) \mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p}) + B_p \prod_{s=0}^{p-1} (1 - m_s h_s)^{k_s},$$

where $B_p = 1.65 \frac{M_p}{m_p} (h_p d)^{0.5} + \frac{\epsilon_{DSM}}{m_p}$ for all $p \in \{0, \dots, T-1\}$.

Proposition 2 highlights that an efficient Langevin sampling is governed by the quantities $\{m_p, M_p, \mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p})\}_{p \in \{0, \cdots, T-1\}}$, which depend on the choice of the intermediate distributions. We provide in Lemma 4 (Appendix A.3) additional bounds for $\mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p})$ for both π_t^L and π_t^G showing that while $\mathcal{W}_2(\pi_{t_{p+1}}^G, \pi_{t_p}^G)$ is governed by the L^2 proximity of the scores, $\mathcal{W}_2(\pi_{t_{p+1}}^L, \pi_{t_p}^L)$ takes into account the covariance of the backward diffusion kernels.

Although Proposition 2 relies on strong assumptions, requiring all intermediate densities π_{t_p} to retain log-concavity and smoothness properties of the target density, especially when $t_p \to t_0$, Lemma 1 relaxes these assumptions in the following training setting used throughout the rest of the paper.

Diffusion setting. From now on, we consider training individual scores using the Variance-Preserving (VP) diffusion scheme [26]. The corresponding forward transition kernel is of the form $q(\theta_t \mid \theta_0) = \mathcal{N}(\theta_t; \sqrt{\alpha_t}\theta_0, v_t I_d)$ where $v_t = 1 - \alpha_t$ and α_t is a strictly decreasing function of time such that $\alpha_0 = 1$ and $\alpha_t \underset{t \to 1}{\to} 0$. In this setting, we consider T annealing levels spaced over the interval $[0, 1]$, with the convention that $t_0 = 0$ and $t_T = 1$. This choice is motivated by the fact that, under VP training, the intermediate densities π_t^L and π_t^G converge to a standard Gaussian as $t \to 1$.

Lemma 1 (Preservation of smoothness and log-concavity - General case) *Let $\sigma_{min}(A)$ (resp. $\sigma_{max}(A)$) denote the minimal (resp. maximal) eigenvalue of any matrix A. Let $\Sigma_{t,i}$, $\Sigma_{t,\lambda}$ denote the covariance matrices of the Gaussian backward kernels in Linhart's compositional score (Equation 2). Assume that the prior and each individual posterior are Gaussian log-Lipschitz, then they become smooth along the VP scheme, which is equivalent to:*

$$\begin{aligned} m_{t,\lambda} I_d &\preceq -\nabla_\theta^2 \log \lambda_t(\theta) \preceq M_{t,\lambda} I_d, \\ m_{t,i} I_d &\preceq -\nabla_\theta^2 \log p_t(\theta \mid x_i) \preceq M_{t,i} I_d \quad \forall\, t \in [0, 1], \end{aligned}$$

where $m_{t,\lambda}, m_{t,i} \in \mathbb{R}$ and $M_{t,\lambda}, M_{t,i} \in \mathbb{R}_+$. Moreover, π_t^G and π_t^L are also smooth, with smoothness constants having an explicit formula given in Appendix A.5. If $\frac{1}{n} \sum_{i=1}^n \frac{1}{\gamma_{t,i}} m_{t,i} \geq \frac{n-1}{n} \frac{1}{\gamma_{t,\lambda}} M_{t,\lambda}$ for one given $t \in [0, 1]$, then π_t^G is log-concave when $\gamma_{t,i} = \gamma_{t,\lambda} = 1$; under additional symmetry conditions, π_t^L is also log-concave, when $\gamma_{t,i} = \sigma_{min}(\Sigma_{t,i})$ if $m_{t,i} \leq 0$, $\gamma_{t,i} = \sigma_{max}(\Sigma_{t,i})$ if $m_{t,i} \geq 0$ and $\gamma_{t,\lambda} = \sigma_{min}(\Sigma_{t,\lambda})$.

Proof is derived in Appendix A.5. It highlights that mild first-order regularity conditions on initial distributions can yield a stronger smoothness property for the intermediate densities over time.

Furthermore, if the individual posteriors and the prior are only smooth (not log-concave) at some time t, their negative smoothness constants can be compensated by positive contributions for sufficiently large n, making π_t^G and π_t^L smooth and log-concave. This is also consistent with Bernstein-von Mises theorem, under which the multi-observation posterior is asymptotically Gaussian (thus smooth and log-concave) as n increases [28]. In addition, Silveri and Ocello [23] prove that log-concavity is strengthened along the forward diffusion path, driving smoothness constants $m_{t,i}$ and $m_{t,\lambda}$ towards less negative values provided that the initial distributions are weakly-log-concave: thus, the condition that ensures log-concavity for π_t^G and π_t^L becomes increasingly mild as $t \to 1$.

Therefore, Lemma 1 allows one to use the theoretical bounds in Propositions 1 and 2 under milder assumptions, when first order regularity conditions on individual posteriors and prior induce smoothness *and* log-concavity properties for the target multi-observation posterior and bridging densities π_t^G and π_t^L. This is notably the case for multimodal individual posteriors, which when considered on their own may be only smooth, without being log-concave. In the following section, we propose a method to fine-tune the hyperparameters of the annealed Langevin algorithm using the theoretical bounds previously derived and under the relaxed smoothness assumptions of Lemma 1.

3 Guidelines for fine-tuning annealed Langevin hyperparameters

Despite its apparent freedom to choose intermediate densities, annealed Langevin dynamics remains hyperparameter-heavy. Previous works using this sampling scheme typically fix the number of steps a priori and tune the constant step size at each annealing level using taming rules [16, 13] or some heuristics [25] to improve sampling quality [5]. In this paper, we propose to choose these hyperparameters so that the annealed Langevin algorithm achieves a theoretically guaranteed small Wasserstein error. Our decision rule is based on the intermediate theoretical bound derived in Proposition 1: assuming that the Wasserstein error of the $(p+1)^{\text{th}}$ Langevin process is controlled by a threshold $\gamma > 0$, i.e. $\mathcal{W}_2(\mathcal{L}(\theta_{t_{p+1}}^{(k_{p+1})}), \pi_{t_{p+1}}) \leq \gamma$, the idea is to choose h_{t_p} and k_{t_p} so that the Wasserstein error of the following process remains lower than γ, i.e. $\mathcal{W}_2(\mathcal{L}(\theta_{t_p}^{(k_p)}), \pi_{t_p}) \leq \gamma$. Therefore, the final Wasserstein error of the algorithm given in Proposition 2 will also be bounded by γ. Let $\omega \in (0,1)$ be a weight parameter, such that $\omega\gamma > \frac{\epsilon_{\text{DSM}}}{m_t}$ then choosing

$$0 \leq h_{t_p} \leq \min\left(\frac{(\omega\gamma - \frac{\epsilon_{\text{DSM}}}{m_{t_p}})^2}{d(1.65)^2}\frac{m_{t_p}^2}{M_{t_p}^2}, \frac{2}{m_{t_p} + M_{t_p}}\right),$$
$$k_{t_p} \geq \log\left(\frac{(1-\omega)\gamma}{\gamma + \mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p})}\right)\frac{1}{\log(1 - m_{t_p}h_{t_p})} \geq 0,$$

achieves this objective. With these choices, the bias term B_{t_p} in Proposition 1 becomes lower than $\omega\gamma$, while the rest of the upper bound is lower than $(1-\omega)\gamma$ (see Appendix A.4 for more details). Note that the condition on $\omega\gamma$ seems natural: the intermediate bound in Proposition 1 shows a linear dependence between the score error ϵ_{DSM} and the Wasserstein error; therefore, it seems difficult to achieve a high sampling quality (small γ) using a very biased score estimate (high ϵ_{DSM}).

We highlight that the choice of bridging densities π_{t_p} (Geffner-like or Linhart-like) significantly affects hyperparameter selection under this decision rule: indeed, the choice of step sizes mainly involves the ratio m_{t_p}/M_{t_p} that depends on each bridging density π_{t_p}, while the choice of the number of steps depends on the Wasserstein distance between two consecutive bridging densities $\mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p})$ and the log-concavity constant m_{t_p}. These theoretical quantities may be difficult to compute in practice. In the following section, we consider Gaussian distributions, yielding analytical expressions for these quantities. We compare the step sizes and numbers of steps chosen with this decision rule when the bridging densities are derived from Geffner or Linhart.

4 Theoretical comparison of guidelines in a Gaussian setting

In this section, we consider a d-dimensional standard Gaussian prior $\lambda(\theta) = \mathcal{N}(\theta; 0, I_d)$ and a Gaussian likelihood $p(x \mid \theta) = \mathcal{N}(x; \theta, \Sigma)$. Given n IID observations $x_1, \ldots, x_n$, the goal of Bayesian inference is to sample from the (Gaussian) multi-observation posterior $p(\theta \mid x_{1:n}) := \pi_0(\theta)$. In this Gaussian setting, the (diffused) individual posteriors $p_t(\theta \mid x_i)$ and the bridging densities (π_t^G

or π_t^L) are also Gaussian and analytically tractable for all $t \in [0,1]$. To simplify the comparison between both algorithms in terms of step sizes and number of steps, we consider a null score error, i.e. $\epsilon_{\text{DSM}} = 0$ by analytically computing all the scores involved.

4.1 Geffner's and Linhart's intermediate densities

We start with the densities π_t^L defined by Linhart et al. [16] for $t \in [0,1]$. As all the assumptions of their work are fulfilled in the Gaussian setting, the distributions underlying their composite scores (Equation 2) correspond to the diffused versions of the target multi-observation posterior along the VP forward process and have the following Gaussian form :

$$\pi_t^L(\theta \mid x_{1:n}) = p_t(\theta \mid x_{1:n}) = \mathcal{N}(\theta; \mu_t^L, \Sigma_t^L) \quad \text{with} \quad \begin{cases} \Sigma_t^L = \alpha_t(n\Sigma^{-1} + I_d)^{-1} + v_t I_d \\ \mu_t^L = \sqrt{\alpha_t}(nI_d + \Sigma)^{-1} \sum_{i=1}^n x_i. \end{cases}$$

Let $\Sigma_t = \alpha_t(\Sigma^{-1} + I_d)^{-1} + v_t I_d$ be the covariance matrix of any diffused individual posterior $p_t(\theta \mid x_i)$. Then, the bridging densities π_t^G defined by Geffner et al. [13] are also Gaussian, with the following first two moments:

$$\pi_t^G(\theta \mid x_{1:n}) = \mathcal{N}(\theta; \mu_t^G, \Sigma_t^G) \quad \text{with} \quad \begin{cases} \Sigma_t^G = \left((1-n)I_d + n\Sigma_t^{-1}\right)^{-1} \\ \mu_t^G = \sqrt{\alpha_t}\Sigma_t^G\Sigma_t^{-1}(I_d + \Sigma)^{-1} \sum_{i=1}^n x_i. \end{cases}$$

Note that the covariance of Geffner's densities is well defined for all time $t \in [0,1]$ as the covariance matrices of the prior and the likelihood commute (see Appendix A.6). In the Gaussian case, these intermediate densities naturally retain the log-concavity and smoothness properties of the target distribution π_0 over time. Lemma 2 provides an affine relation between the log-concavity constant m_0 (resp. smoothness constant M_0) of π_0 and Linhart's bridging densities constants m_t^L (resp. M_t^L) or Geffner's counterparts m_t^G (resp. M_t^G) for all $t \in [0,1]$. Proof is given in Appendix A.7.

Lemma 2 (Preservation of smoothness and log-concavity - Gaussian case) *Let the target posterior distribution π_0 be M_0-smooth and m_0-log-concave. Let $\sigma_{\max}$ (resp. $\sigma_{\min}$) denote the maximal (resp. minimal) eigenvalue of the likelihood covariance matrix Σ. Then M_t^G and M_t^L (resp. m_t^G and m_t^L) are time-dependent affine functions of M_0 (resp. m_0) as per*

$$M_t^L = M_0 \frac{\sigma_{\min}}{\sigma_{\min} + nv_t} \quad \textit{and} \quad m_t^L = m_0 \frac{\sigma_{\max}}{\sigma_{\max} + nv_t},$$
$$M_t^G = M_0 \frac{\sigma_{\min}}{\sigma_{\min} + v_t} + \frac{(1-n)v_t}{\sigma_{\min} + v_t} \quad \textit{and} \quad m_t^G = m_0 \frac{\sigma_{\max}}{\sigma_{\max} + v_t} + \frac{(1-n)v_t}{\sigma_{\max} + v_t}.$$

4.2 Comparison of step sizes

With the decision rule provided in Section 3, the choice of step sizes depends mainly on the ratio $\frac{m_t}{M_t}$. In the Gaussian setting, the log-concavity constant m_t and smoothness constant M_t respectively correspond to the minimal and maximal eigenvalue of the precision matrix of the intermediate density π_t^G or π_t^L. The following proposition shows that these constants are always higher for Geffner's densities than for Linhart's counterparts.

Proposition 3 (Link between M_t^G and m_t^G with M_t^L and m_t^L) *Let σ_{min} and σ_{max} denote the minimal and maximal eigenvalue of the likelihood covariance matrix Σ. Then, we have*

$$M_t^G = M_t^L + \frac{n(n-1)\alpha_t v_t}{D(\sigma_{min}, v_t)} \quad \textit{and} \quad m_t^G = m_t^L + \frac{n(n-1)\alpha_t v_t}{D(\sigma_{max}, v_t)},$$

where $D(\sigma, v_t) = (\sigma + v_t)(\sigma + nv_t)$ and $t \in [0,1]$.

Proof is provided in Appendix A.8. Proposition 3 gives us insights into the ratio of both constants, therefore into the choice of step sizes for each algorithm presented in Section 3.

Corollary 1 (Comparison of the ratios m_t^G/M_t^G and m_t^L/M_t^L) *Let M_t^L, m_t^L (resp. M_t^G, m_t^G) denote the smoothness and log-concavity constants of Linhart's (resp. Geffner's) intermediate densities. Then we have :*

$$\frac{m_t^G}{M_t^G} \leq \frac{m_t^L}{M_t^L} \quad \forall t \in [0,1].$$

Consequently, the step sizes h_t^G used with Geffner's composite score in annealed Langevin are smaller than their counterparts h_t^L used with Linhart's score for any time $t \in [0,1]$.

Proof is given in Appendix A.9. Choosing Linhart's composite score instead of Geffner's counterpart allows one to use larger step sizes at each annealing level without degrading the final sampling quality, measured in Wasserstein distance. The following section provides a comparison of the number of steps for both algorithms.

4.3 Comparison of the number of steps

For a given level $t_p \in [0,1]$, the choice of the number of steps k_{t_p} prescribed in Section 3 depends on the Wasserstein distance between intermediate densities $\mathcal{W}_2(\pi_{t_p}, \pi_{t_{p+1}})$ and the log-concavity constant m_{t_p}. When $t_p \to 0$ or $t_p \to 1$, the contribution of m_{t_p} in the fine-tuning objective is negligible, since this term is essentially the same for both algorithms. As a result, the weighting of the Wasserstein distance in the choice of k_{t_p} is the only factor that enables a meaningful comparison between the two approaches.

As intermediate distributions $\pi_{t_p}^L$ or $\pi_{t_p}^G$ are Gaussian for all $t_p \in [0,1]$, the Wasserstein distance has a closed form expression including a L_2-mean difference and a Bures distance between covariance matrices [2]. The following proposition compares the intermediate Wasserstein distances for both algorithms when $t_p \to 0$ and $t_p \to 1$.

Proposition 4 (Comparison of Wasserstein distances between intermediate densities) *Let $\pi_{t_p}^L$ (resp. $\pi_{t_p}^G$) denote the (Gaussian) intermediate density defined by Linhart (resp. by Geffner) at level t_p. Then, we have :*

$$\begin{aligned} &\text{As } t_p \to 0, \quad \mathcal{W}_2^2(\pi_{t_p}^G, \pi_{t_{p+1}}^G) \leq \mathcal{W}_2^2(\pi_{t_p}^L, \pi_{t_{p+1}}^L). \\ &\text{As } t_p \to 1, \quad \mathcal{W}_2^2(\pi_{t_p}^G, \pi_{t_{p+1}}^G) \geq \mathcal{W}_2^2(\pi_{t_p}^L, \pi_{t_{p+1}}^L). \end{aligned}$$

Therefore, the number of Langevin steps $k_{t_p}^G$ (chosen as in Section 3) used with Geffner's composite score in annealed Langevin is greater than its counterpart $k_{t_p}^L$ used with Linhart's score as $t_p \to 1$, and conversely as $t_p \to 0$.

Proof is given in Appendix A.10. We quantify computational complexity by the total number of score function evaluations[1], that is also equal to the total number of Langevin steps $\sum_{p=0}^{T} k_{t_p}$. Proposition 4 gives us insight into the complexity of each algorithm in two different extreme regimes. As $t \to 0$, the annealed Langevin algorithm tends to sample from intermediate distributions close to the target density π_0: in this case, using Geffner's composite score yields less number of steps for an equivalent sampling quality than Linhart's composite score. Conversely, when $t \to 1$, the annealed Langevin algorithm tends to sample from distributions close to a standard Gaussian: in this case, using Linhart's score seems beneficial, as the number of steps is lower and the step size higher.

A theoretical characterization of the evolution of the number of steps between these two regimes remains an open question. Empirical tests indicate that there are only two regimes between $t = 0$ and $t = 1$ (see Figure 2): there seems to exist a time level $\tilde{t}$ such that on $[0, \tilde{t}]$, k_t^L is always higher than k_t^G, and on $[\tilde{t}, 1]$ the situation is reverse. This switching time level varies depending on the number of observations n, the eigenvalues of the likelihood covariance matrix, and the dimension of the space d in a complex manner. Though, we observe that $\tilde{t}$ is often close to 0 regardless of dimension $d \geq 3$: exceptions appear when $d = 2$ where the switching point may be later (see Figure 6 in Appendix B.2). This unclear evolution may also stem from the contribution of the term $h_t m_t$ in the choice of k_t

[1]The computational complexity does not include the additional cost of estimating the covariance matrices required by Linhart's composite score s_t^L, as we treat this as a *shared* cost, as detailed in Appendix B.4

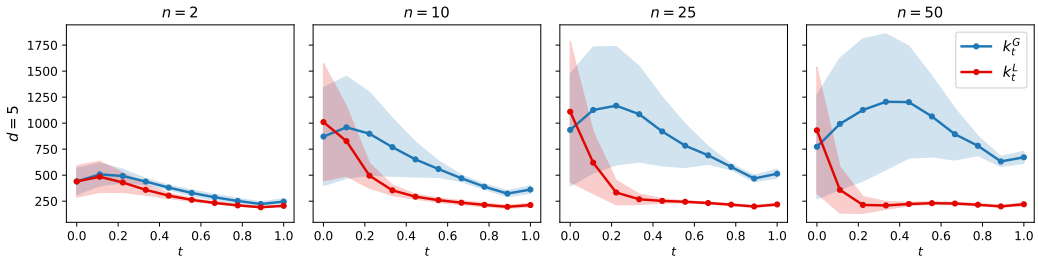

Figure 2: Evolution of the number of Langevin steps in the 5-dimensional Gaussian setting for different numbers n of conditional observations. Blue (resp. red) curves represent the mean number of steps k_t^G (resp. k_t^L) at each annealing level when using Geffner's (resp. Linhart's) compositional score. We choose $T = 10$ annealing levels uniformly spaced over $[0, 1]$. The mean (and std) is computed over 5 different Gaussian models, each with a different likelihood matrix. Empirical results for $d = 2, 10$ are provided in Appendix B.2.

which may outweigh that of the Wasserstein distance: in extreme regimes, this term is very similar for Geffner's and Linhart's densities, making k_t depend on the intermediate Wasserstein distances only.

Globally, the theoretical results accompanied by the empirical observations in the d-dimensional Gaussian setting indicate that achieving a Wasserstein error comparable to Geffner's method, requires Linhart's algorithm to use larger step sizes and fewer Langevin steps for most annealing levels, thus a lower total number of steps. Figure 4 (Appendix B.1) provides an empirical validation of the Wasserstein error given by the prescribed hyperparameters for both compositional score choices.

5 Hyperparameter selection in practice

In this section, we illustrate empirically that the hyperparameter guidelines developed in Section 3 are useful beyond the purely Gaussian case. The main difficulty to generalize these guidelines is to compute the smoothness (M_t^G, M_t^L) and log-concavity constants (m_t^G, m_t^L) that drive the choice of the step sizes h_t^G and h_t^L, as well as the Wasserstein distances between consecutive densities that affect the choice of the number of steps per annealing level k_t^G and k_t^L. We therefore consider a second-order approximation of each individual posterior and the prior, which allows us to get an analytical Gaussian approximation of the target multi-observation posterior by composing individual posteriors and prior as per $p(\theta \mid x_{1:n}) \propto \lambda(\theta)^{1-n} \prod_{i=1}^n p(\theta \mid x_i)$. Note that this approximation becomes increasingly reasonable according to the Bernstein-von Mises theorem [28], as the number of observations n increases (see Appendix B.6 for more details). Intermediate densities π_t^G or π_t^L also admit analytical expressions, similar to those in Section 4.1. This makes their smoothness and log-concavity constants, together with Wasserstein distances between intermediate densities, tractable and enables the selection of (h_t^G, k_t^G) or (h_t^L, k_t^L) according to our decision rule. Importantly, we do not introduce any additional Gaussianity assumptions here but instead rely on those used by Linhart et al. [16] when estimating the covariance matrices in Equation 2 with their `GAUSS` algorithm. We also note that we cannot provide a strong theoretical upper bound on the Wasserstein error under such approximations, but we expect them not to compromise the theoretical Wasserstein control, especially for large n.

General setting. For all the following experiments, we use T annealing levels uniformly spaced over the interval $[0, 1]$, so that $t_p = \frac{p}{T}$ for $p \in \{0, \dots, T\}$. We train a time-dependent amortized score estimate s_ϕ with score matching on samples of the joint distribution $(\theta, x) \sim \lambda(\theta)p(x \mid \theta)$ (see Appendix B.4 for details). This score is an estimate of the *exact* score of individual posteriors along the VP diffusion scheme, i.e. $s_\phi(\theta, x_i, t) \approx \nabla_\theta \log p_t(\theta \mid x_i)$ for all $t \in [0, 1]$ and all x_i. Prior scores can often be derived analytically (see Appendix A.12). As prescribed in Linhart et al. [16] for their `GAUSS` algorithm, we use the score s_ϕ to sample individual posteriors using a reverse diffusion scheme and compute empirical covariances, involved in Equation 2. Contrary to the previous section, we need to take into account the score-matching error ϵ_{DSM} of each compositional score in the choice of step sizes: as we compose individual scores according to Equation 1 or 2, we only have access to the score-matching errors of individual score estimates. However, it is possible to obtain a bound of

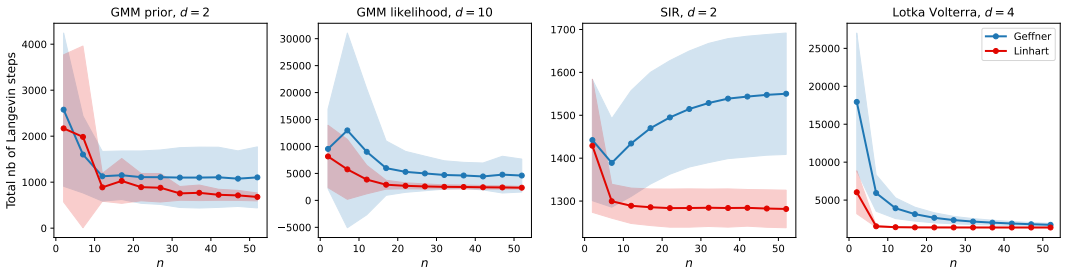

Figure 3: Evolution of the mean total number of Langevin steps $\sum_{p=0}^{T} k_{t_p}$ for different numbers of conditional observations n. These choices achieve a prescribed low final Wasserstein error for both compositional score formulations (see Appendix B.4). Blue curves stand for Geffner's complexity while red curves represent that of Linhart. Mean and std are computed over 5 different seeds.

this compositional score error for both algorithms: Touron et al. [27] provide a bound of the mean squared error of Linhart's compositional score s_t^L based on individual score errors. Adopting the authors' notations, we derive a bound for Geffner's score error in Lemma 3, where $\tilde{s}_t^G$ is the estimate of s_t^G (Equation 1) with inexact individual scores. Proof is given in Appendix A.11. Appendix B.7 provides additional technical details on score error estimation.

Lemma 3 *Let s_ϕ (resp. s_λ) be an amortized neural-based approximation of the individual posteriors $p(\theta \mid x_i)$ (resp. the prior $\lambda(\theta)$). Assume that $\mathrm{E}_\theta\left[\|\nabla_\theta \log \lambda_t(\theta) - s_\lambda(\theta, t)\|^2\right] \leq \epsilon^2_{DSM,\lambda}$ and $\mathrm{E}_\theta\left[\|\nabla_\theta \log p_t(\theta \mid x_i) - s_\phi(\theta, x_i, t)\|^2\right] \leq \epsilon^2_{DSM}$ for all $i = 1, \dots, n$. Then,*

$$\mathrm{E}_\theta\left[\|\nabla_\theta \log \pi_t^G(\theta \mid x_{1:n}) - \tilde{s}_t^G(\theta \mid x_{1:n})\|^2\right] \leq \Big((n-1)\epsilon_{DSM,\lambda} + n\epsilon_{DSM}\Big)^2.$$

Numerical applications. We consider the following four inference tasks, where d denotes the dimension of the parameter space:

GMM prior ($d = 2$): The prior is a mixture of two Gaussians with tunable means, and the likelihood is Gaussian. The resulting posterior is a mixture of Gaussians, whose modes become more separated as the prior means move far apart. Note that as n grows, the posterior becomes approximately Gaussian and unimodal. This task allows us to test the robustness of our decision rule for multimodal target posteriors, where non-Gaussianity keeps mastered (see Figure 7 in Appendix B.3).

GMM likelihood ($d = 10$): The prior is a standard Gaussian, while the likelihood is a mixture of two Gaussians, yielding a Gaussian mixture posterior. This task is inspired by Geffner et al. [13] and deviates slightly more from Gaussianity than the first one.

SIR ($d = 2$): This epidemiological model tracks the evolution of susceptible (S), infected (I) and recovered (R) populations. The prior is Log-Normal, the simulator is based on a set of differential equations that outputs samples from a Binomial distribution. The resulting posterior is unimodal.

Lotka-Volterra ($d = 4$): This predator-prey model describes the evolution of the populations of two interacting species. The prior is Log-Normal, the simulator is based on a set of differential equations that outputs samples from a Log-Normal distribution. The resulting posterior is unimodal.

For the first two tasks, analytical individual posterior scores can be derived (see Appendix A.12). The last two tasks, inspired by the SBI benchmark [17], allow us to test the robustness of our decision rule when individual posterior scores need to be learned and the compositional error estimated for different parameter dimensions d. The hyperparameters chosen according to our decision rule allow us to run annealed Langevin *without* intermediate clipping steps for all tasks. Figure 8 (in Appendix B.4) further shows that they yield a *similar* final Wasserstein error, chosen to be *small*, for both choices of intermediate densities: this empirical observation also holds for different choices of threshold γ and number of annealing levels T, as detailed in Appendices B.8 and B.9. Overall, Linhart's densities use larger step sizes and require fewer steps than Geffner's counterparts across most inference tasks (see Figure 3): Linhart's composite score thus yields a more efficient sampler in terms of score function evaluations.

6 Conclusion

We propose the first theoretically grounded guidelines to fine-tune the hyperparameters of annealed Langevin algorithm in compositional SBI, supported by Wasserstein error bounds under smoothness and log-concavity assumptions. These guidelines remain valid under weaker smoothness assumptions on individual posteriors, particularly for large n, thereby extending their applicability to broader settings. We highlight the central role of intermediate densities that underlie the compositional scores used in annealed Langevin, in the hyperparameter fine-tuning. We provide explicit hyperparameter prescriptions in the Gaussian case and show that they can be effectively transferred to non-Gaussian settings, thus avoiding the need for manual tuning while achieving a prescribed low Wasserstein error. In addition, we show that Linhart's compositional score yields a more efficient sampler than Geffner's one in both settings in terms of score function evaluations.

Our analysis could serve as a theoretically grounded starting point for practitioners using compositional score-based approaches for designing novel intermediate densities that improve the annealed Langevin sampling performance. Although our theoretical guarantees provide useful guidance for hyperparameter tuning, they still rely on smoothness and (approximate) log-concavity assumptions that may not hold in highly multimodal regimes; extending these guarantees to more general distributions, e.g. those satisfying a log-Sobolev inequality only, is a key direction for future work.

Acknowledgements

PLCR and JA were supported by national grants managed by the French National Research Agency (Agence Nationale de la Recherche) attributed to the SBI4C project (PLCR) and the THUNDER project (JA) of the MIAI Cluster, under the reference ANR-23-IACL-0006. JA was supported by the ANR project Bayes-Duality (grant ANR-21-JSTM-0001). GVC was partially funded by the Geolearning chair of Mines Paris – PSL, funded by Andra, BNP Paribas, CCR and the SCOR Foundation for Science.

References

[1] Arruda, J., Pandey, V., Sherry, C., Barroso, M., Intes, X., Hasenauer, J., and Radev, S. (2026). Compositional amortized inference for large-scale hierarchical bayesian models. In *International Conference on Learning Representations*.

[2] Bhatia, R., Jain, T., and Lim, Y. (2017). On the Bures-Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37.

[3] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[4] Bradbury, J., Frostig, R., Hawkins, P., et al. (2018). Jax: composable transformations of python+numpy programs.

[5] Brosse, N., Durmus, A., Moulines, É., and Sabanis, S. (2019). The tamed unadjusted Langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663.

[6] Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062.

[7] Dalalyan, A. S. and Karagulyan, A. (2019). User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311.

[8] Deistler, M., Boelts, J., Steinbach, P., Moss, G., Moreau, T., Gloeckler, M., Rodrigues, P. L., Linhart, J., Lappalainen, J. K., and Miller, B. K. (2025). Simulation-based inference: A practical guide. *arXiv preprint arXiv:2508.12939*.

[9] Du, Y., Durkan, C., Strudel, R., Tenenbaum, J. B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., and Grathwohl, W. (2023). Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and MCMC. In *International Conference on Machine Learning*.

[10] Durmus, A. and Moulines, E. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551 – 1587.

[11] Flamary, R., Courty, N., Gramfort, A., et al. (2021). Python optimal transport (pot): A python library for optimal transport analysis. *Journal of Machine Learning Research*, 22(78):1–8.

[12] Gao, Y., Huang, J., and Jiao, Y. (2024). Gaussian interpolation flows. *Journal of Machine Learning Research*, 25(253):1–52.

[13] Geffner, T., Papamakarios, G., and Mnih, A. (2023). Compositional score modeling for simulation-based inference. In *International Conference on Machine Learning*.

[14] Gloeckler, M., Toyota, S., Fukumizu, K., and Macke, J. H. (2025). Compositional simulation-based inference for time series. In *International Conference on Learning Representations*.

[15] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*.

[16] Linhart, J., Cardoso, G. V., Gramfort, A., Corff, S. L., and Rodrigues, P. L. (2026). Diffusion posterior sampling for simulation-based inference in tall data settings. *TMLR*.

[17] Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., and Macke, J. (2021). Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*.

[18] Phan, D., Pradhan, N., and Jankowiak, M. (2020). Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*.

[19] Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., and Köthe, U. (2020). Bayesflow: Learning complex stochastic models with invertible neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(4):1452–1466.

[20] Rodrigues, P., Moreau, T., Louppe, G., and Gramfort, A. (2021). HNPE: leveraging global parameters for neural posterior estimation. *Advances in Neural Information Processing Systems*.

[21] Saumard, A. and Wellner, J. A. (2014). Log-concavity and strong log-concavity: a review. *Statistics Surveys*, 8:45–114.

[22] Sharrock, L., Simons, J., Liu, S., and Beaumont, M. (2024). Sequential neural score estimation: Likelihood-free inference with conditional score based diffusion models. In *International Conference on Machine Learning*.

[23] Silveri, M. G. and Ocello, A. (2025). Beyond log-concavity and score regularity: Improved convergence bounds for score-based generative models in W2-distance. In *International Conference on Machine Learning*.

[24] Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In *International Conference on Neural Information Processing Systems*.

[25] Song, Y. and Ermon, S. (2020). Improved techniques for training score-based generative models. In *International Conference on Neural Information Processing Systems*.

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

[27] Touron, C., Victorino Cardoso, G., Arbel, J., and Coelho Rodrigues, P. L. (2025). Error analysis of a compositional score-based algorithm for simulation-based inference. In *Workshop on Principles of Generative Modeling at EurIPS 2025*, Copenhagen, Denmark.

[28] Vaart, A. W. v. d. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[29] Vempala, S. and Wibisono, A. (2019). Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*.

Contents

1 Introduction **1**

2 Wasserstein error bounds for annealed Langevin dynamics **3**

3 Guidelines for fine-tuning annealed Langevin hyperparameters **5**

4 Theoretical comparison of guidelines in a Gaussian setting **5**

4.1 Geffner's and Linhart's intermediate densities 6

4.2 Comparison of step sizes 6

4.3 Comparison of the number of steps 7

5 Hyperparameter selection in practice **8**

6 Conclusion **10**

A Theoretical results **14**

A.1 Proof of Proposition 1 14

A.2 Proof of Proposition 2 16

A.3 Bound of intermediate Wasserstein distances 17

A.4 Fine-tuning of hyperparameters 19

A.5 Proof of Lemma 1 - Smoothness and log-concavity preservation 19

A.5.1 Smoothness 20

A.5.2 Log-concavity 23

A.6 Analytical bridging densities in a Gaussian setting 24

A.7 Proof of Lemma 2 25

A.8 Proof of Proposition 3 25

A.9 Proof of Corollary 1 27

A.10 Proof of Proposition 4 27

A.10.1 Comparison of Bures distances 28

A.10.2 Comparison of mean differences 31

A.10.3 Final derivation 34

A.11 Proof of Lemma 3 36

A.12 Analytical derivation of prior and individual posterior scores 36

B Empirical results and technical details **38**

B.1 Empirical validation of the Wasserstein error - Gaussian setting 38

B.2 Comparison of the number of steps - Gaussian setting 39

B.3 Additional figures - GMM Prior 40

B.4 Technical details - Non Gaussian setting 41

B.5 Comparison with score-based diffusion sampling and alternative Langevin tuning . 43

B.6 Discussion on the Gaussian approximations 45

B.7 Empirical estimation of the score error ϵ_{DSM} 46
B.8 Additional results on the choice of γ 47
B.9 Additional results on the choice of T 49

A Theoretical results

A.1 Proof of Proposition 1

Suppose that we want to sample from a distribution π_{t_p} that is m_{t_p}-log concave and M_{t_p}-smooth, for a given $p \in \{0, \dots, T-1\}$. π_{t_p} stands for one bridging density in the annealed Langevin process (defined by Geffner or Linhart). For readability concerns, we replace the index t_p simply by p for constants m, M, B, h, k. Suppose that the constant step size is such that $h_p < \frac{2}{m_p + M_p}$, then according to Dalalyan and Karagulyan [7] (Theorem 4), the following holds:

$$
\begin{aligned}
\mathcal{W}_2(\mathcal{L}(\theta_{t_p}^{(k_p)}), \pi_{t_p}) &\leq (1 - m_p h_p)^{k_p} \mathcal{W}_2(\mathcal{L}(\theta_{t_p}^{(0)}), \pi_{t_p}) + B_p \\
&\leq (1 - m_p h_p)^{k_p} (\mathcal{W}_2(\mathcal{L}(\theta_{t_p}^{(0)}), \pi_{t_{p+1}}) + \mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p})) + B_p \\
&\qquad \text{by triangular inequality for Wasserstein distance} \\
&= (1 - m_p h_p)^{k_p} \mathcal{W}_2(\mathcal{L}(\theta_{t_p}^{(0)}), \pi_{t_{p+1}}) + (1 - m_p h_p)^{k_p} \mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p}) + B_p \\
&\boxed{= (1 - m_p h_p)^{k_p} \mathcal{W}_2(\mathcal{L}(\theta_{t_{p+1}}^{(k_{p+1})}), \pi_{t_{p+1}}) + (1 - m_p h_p)^{k_p} \mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p}) + B_p} \\
&\qquad \text{since } \theta_{t_{p+1}}^{(k_{p+1})} = \theta_{t_p}^{(0)},
\end{aligned}
$$

where $B_p = 1.65 \frac{M_p}{m_p} (h_p d)^{0.5} + \frac{\delta \sqrt{d}}{m_p} + \frac{\sigma^2 (h_p d)^{0.5}}{1.65 M_p + \sigma \sqrt{m_p}}$, δ bounds the bias of the score estimate, σ bounds its variance and d is the dimension of the space.

Let's now simplify the bias terms B_p when the score estimate is obtained by denoising score matching. In this case, we obtain a deterministic *but* biased version of the exact score. Recall that the score estimate in Dalalyan and Karagulyan [7] $s_k = \nabla_\theta \log \pi(\theta_k) + \zeta_k$ has a bounded bias at any iterate θ_k of the LMC algorithm as per

$$
\mathrm{E}[\|\mathrm{E}(\zeta_k \mid \theta_k)\|_2^2] \leq \delta^2 d.
$$

If ϵ_{DSM}^2 controls the mean square error (MSE) between the true score and its estimate in average over the LMC iterates, then we have :

$$
\begin{aligned}
\mathrm{E}_{\theta_k}[\|\mathrm{E}(\zeta_k \mid \theta_k)\|_2^2] &= \mathrm{E}_{\theta_k}[\|\mathrm{E}(s_k - \nabla_\theta \log \pi(\theta_k) \mid \theta_k)\|_2^2] \\
&\leq \mathrm{E}_{\theta_k}[\mathrm{E}(\|s_k - \nabla_\theta \log \pi(\theta_k)\|_2^2 \mid \theta_k)] \quad \text{by Jensen inequality} \\
&= \mathrm{E}_{\theta_k, s_k}[\|s_k - \nabla_\theta \log \pi(\theta_k)\|_2^2] \\
&= \mathrm{E}_{\theta_k}[\|s(\theta_k) - \nabla_\theta \log \pi(\theta_k)\|_2^2] \\
&\qquad \text{if } s_k = s(\theta_k) \text{ i.e. for each point the score estimate is deterministic} \\
&\leq \epsilon_{\text{DSM}}^2.
\end{aligned}
$$

This is valid for any bridging distribution π_t. If ϵ_{DSM}^2 upperbounds the maximum MSE over time (especially over levels $t_0, \dots, t_T$), then we can set $\delta = \frac{\epsilon_{\text{DSM}}}{\sqrt{d}}$. Moreover, training a score estimate with denoising score-matching provides a deterministic score estimate: we can thus assume a null variance of the score estimate ($\sigma = 0$) in the theoretical bound. In Dalalyan and Karagulyan [7], the variance is bounded as follows:

$$
\mathrm{E}[\|\zeta_k - \mathrm{E}(\zeta_k \mid \theta_k)\|_2^2] \leq \frac{\sigma^2}{d}.
$$

In our setting, we obtain

$$\begin{aligned}
\mathrm{E}_{\theta_k}[\|\zeta_k - \mathrm{E}(\zeta_k \mid \theta_k)\|_2^2] &= \mathrm{E}_{\theta_k}[\|s_k - \nabla_\theta \log \pi(\theta_k) - \mathrm{E}(s_k - \nabla_\theta \log \pi(\theta_k) \mid \theta_k)\|_2^2] \\
&= \mathrm{E}_{\theta_k}[\|s(\theta_k) - \nabla_\theta \log \pi(\theta_k) - \mathrm{E}(s(\theta_k) - \nabla_\theta \log \pi(\theta_k) \mid \theta_k)\|_2^2] \\
&\quad \text{when we assume that } s_k = s(\theta_k) \text{ is deterministic} \\
&= \mathrm{E}_{\theta_k}[\|s(\theta_k) - \nabla_\theta \log \pi(\theta_k) - s(\theta_k) - \nabla_\theta \log \pi(\theta_k)\|_2^2] \\
&= 0.
\end{aligned}$$

So the bias terms simplify to $B_p = 1.65 \frac{M_p}{m_p} (h_p d)^{0.5} + \frac{\epsilon_{\text{DSM}}}{m_p}$ in the case where the score estimate is deterministic but biased, as in denoising score matching. Note that to estimate the bias of the score and compute ϵ_{DSM} in practice, we need to bound the score matching error in average over the annealed Langevin iterates at each annealing level.

A.2 Proof of Proposition 2

To get the global error bound in Wasserstein distance of an annealed Langevin algorithm with T annealing steps, it is sufficient to compose the previous intermediate bound T times, assuming that each intermediate distribution π_t is M_t-smooth and m_t-concave for all $t = t_0, \dots, t_{T-1}$. Here again, for readability concerns, we replace the time level t_p simply by p for the constants m, M, k, h, B. As we start from Gaussian samples $\theta_{t_T}^{(0)} \sim \mathcal{N}(0, I_d)$, we use the convention that $k_T = 0$, therefore, $\theta_{t_{T-1}}^{(0)} = \theta_{t_T}^{(k_T)} = \theta_{t_T}^{(0)}$. Suppose that $h_p < \frac{2}{m_p + M_p}$ for all $p = 0, \dots, T-1$, then we get:

$$
\begin{aligned}
&\mathcal{W}_2(\mathcal{L}(\theta_0^{(k_0)}), \pi_{t_0}) \\
&\quad \leq (1 - m_0 h_0)^{k_0} \mathcal{W}_2(\mathcal{L}(\theta_{t_1}^{(k_1)}), \pi_{t_1}) + (1 - m_0 h_0)^{k_0} \mathcal{W}_2(\pi_{t_1}, \pi_{t_0}) + B_0 \\
&\quad \leq (1 - m_0 h_0)^{k_0} \left((1 - m_1 h_1)^{k_1} \mathcal{W}_2(\mathcal{L}(\theta_{t_2}^{(k_2)}), \pi_{t_2}) + (1 - m_1 h_1)^{k_1} \mathcal{W}_2(\pi_{t_2}, \pi_{t_1}) + B_1 \right) \\
&\qquad + (1 - m_0 h_0)^{k_0} \mathcal{W}_2(\pi_{t_1}, \pi_{t_0}) + B_0 \\
&\quad = (1 - m_0 h_0)^{k_0} (1 - m_1 h_1)^{k_1} \mathcal{W}_2(\mathcal{L}(\theta_{t_2}^{(k_2)}), \pi_{t_2}) + (1 - m_0 h_0)^{k_0} (1 - m_1 h_1)^{k_1} \mathcal{W}_2(\pi_{t_2}, \pi_{t_1}) \\
&\qquad + (1 - m_0 h_0)^{k_0} B_1 + (1 - m_0 h_0)^{k_0} \mathcal{W}_2(\pi_{t_1}, \pi_{t_0}) + B_0 \\
&\quad = \dots \quad \text{after several annealing steps} \\
&\quad \leq \prod_{s=0}^{T-2} (1 - m_s h_s)^{k_s} \mathcal{W}_2(\mathcal{L}(\theta_{t_{T-1}}^{(k_{T-1})}), \pi_{t_{T-1}}) + \sum_{p=0}^{T-2} \left(\prod_{s=0}^{p} (1 - m_s h_s)^{k_s} \mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p}) + \underbrace{\prod_{s=0}^{p-1} (1 - m_s h_s)^{k_s}}_{=1 \text{ if } p=0} B_p \right) \\
&\quad \leq \prod_{s=0}^{T-2} (1 - m_s h_s)^{k_s} \left((1 - m_{T-1} h_{T-1})^{k_{T-1}} \underbrace{\mathcal{W}_2(\mathcal{L}(\theta_{t_T}^{(k_T)}), \pi_{t_T})}_{=0} + (1 - m_{T-1} h_{T-1})^{k_{T-1}} \mathcal{W}_2(\pi_{T-1}, \pi_T) + B_{T-1} \right) \\
&\qquad + \sum_{p=0}^{T-2} \left(\prod_{s=0}^{p} (1 - m_s h_s)^{k_s} \mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p}) + \prod_{s=0}^{p-1} (1 - m_s h_s)^{k_s} B_p \right) \\
&\qquad \text{using the bound of Proposition 1 and since } \theta_{t_T}^{(k_T)} \sim \pi_{t_T} \\
&\quad = \boxed{\sum_{p=0}^{T-1} \left(\prod_{s=0}^{p} (1 - m_s h_s)^{k_s} \right) \mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p}) + \left(\prod_{s=0}^{p-1} (1 - m_s h_s)^{k_s} \right) B_p},
\end{aligned}
$$

where $B_p = 1.65 \frac{M_p}{m_p} (h_p d)^{0.5} + \frac{\delta \sqrt{d}}{m_p} + \frac{\sigma^2 (h_p d)^{0.5}}{1.65 M_p + \sigma \sqrt{m_p}}$ for all $p = 0, \dots, T-1$. If we consider taking the same step size and number of Langevin steps per Langevin process, i.e. $h_p = h$ and $k_p = k$ for all p such that $h < \min_p \frac{2}{m_p + M_p}$ and we set $m = \min_p m_p$, then we get a simplified bound :

$$
\mathcal{W}_2(\mathcal{L}(\theta_0^{(k)}), \pi_{t_0}) \leq \sum_{p=0}^{T-1} (1 - mh)^{k(p+1)} \mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p}) + (1 - mh)^{kp} B_p.
$$

A.3 Bound of intermediate Wasserstein distances

Lemma 4 *Let $\|x\|_{A^{-1}} := \sqrt{x^T A^{-1} x}$ be the Mahalanobis norm induced by a SPD matrix A. Then,*

$$\begin{aligned}
\mathcal{W}_2(\pi^G_{t+1}, \pi^G_t) \leq \frac{1}{m^G_t}\Big((n-1)\sqrt{\mathrm{E}_\theta\big\|\nabla_\theta \log \lambda_t(\theta) - \nabla_\theta \log \lambda_{t+1}(\theta)\|^2} \\
+ \sum_{i=1}^n \sqrt{\mathrm{E}_\theta\|\nabla_\theta \log p_t(\theta \mid x_i) - \nabla_\theta \log p_{t+1}(\theta \mid x_i)\|^2}\Big), \\
\mathcal{W}_2(\pi^L_{t+1}, \pi^L_t) \leq \frac{1}{m^L_t}\Big((n-1)\sqrt{\mathrm{E}_\theta\big\|\nabla_\theta \log \lambda_t(\theta) - B_{t,\lambda}\nabla_\theta \log \lambda_{t+1}(\theta)\|^2_{\Psi^{-1}_{t,\lambda}}} \\
+ \sum_{i=1}^n \sqrt{\mathrm{E}_\theta\|\nabla_\theta \log p_t(\theta \mid x_i) - B_{t,i}\nabla_\theta \log p_{t+1}(\theta \mid x_i)\|^2_{\Psi^{-1}_{t,i}}}\Big),
\end{aligned}$$

where $\Psi_{t,k} = \Sigma_{t,k}\Lambda_t\Lambda_t^T\Sigma_{t,k}^T$ and $B_{t,k} = \Sigma_{t,k}\Lambda_t\Lambda_{t+1}^{-1}\Sigma_{t+1,k}^{-1}$ with $k \in \{\lambda, 0, \ldots, n\}$.

Proof: We replace the classical level index t_p by t for readability concerns, with the abuse of notation that π_{t+1} denotes the consecutive density to π_t, usually called $\pi_{t_{p+1}}$. In Proposition 2, we require intermediate densities π_t to be m_t-log-concave for any $t = t_0, \ldots, t_T$. They thus satisfy a log-Sobolev inequality and the Talagrand inequality as per:

$$\begin{aligned}
KL(\pi_{t+1}\|\pi_t) &\leq \frac{1}{2m_t}\mathrm{E}_{\pi_{t+1}}[\|\nabla_\theta \log \pi_{t+1}(\theta) - \nabla_\theta \log \pi_t(\theta)\|^2], \\
\mathcal{W}_2^2(\pi_{t+1}, \pi_t) &\leq \frac{2}{m_t} KL(\pi_{t+1}\|\pi_t),
\end{aligned}$$

yielding therefore

$$\mathcal{W}_2^2(\pi_{t+1}, \pi_t) \leq \frac{1}{m_t^2}\underbrace{\mathrm{E}_{\pi_{t+1}}[\|\nabla_\theta \log \pi_{t+1}(\theta) - \nabla_\theta \log \pi_t(\theta)\|^2]}_{\text{Fisher divergence}}.$$

Let's now bound the Fisher divergence between two consecutive intermediate densities by the Fisher divergence between individual posteriors at two consecutive annealing levels. We denote by $\|\cdot\|_{A^{-1}}$ the Mahalanobis norm induced by a SPD matrix A, defined as $\|x\|_{A^{-1}} = \sqrt{x^T A^{-1} x}$. We start with

the densities defined by Geffner et al. [13]:

$$
\begin{aligned}
&\mathrm{E}_\theta\left[\|\nabla_\theta \log \pi_t^G(\theta \mid x_{1:n}) - \nabla_\theta \log \pi_{t+1}^G(\theta \mid x_{1:n})\|^2\right] \\
&\quad= \mathrm{E}_\theta\left[\|(1-n)\left(\nabla_\theta \log \lambda_t(\theta) - \nabla_\theta \log \lambda_{t+1}(\theta)\right) + \sum_{i=1}^n \nabla_\theta \log p_t(\theta \mid x_i) - \nabla_\theta \log p_{t+1}(\theta \mid x_i)\|^2\right] \\
&\quad= \mathrm{E}_\theta\|(1-n)(\nabla_\theta \log \lambda_t(\theta) - \nabla_\theta \log \lambda_{t+1}(\theta))\|^2 + \sum_{i=1}^n \mathrm{E}_\theta\|\nabla_\theta \log p_t(\theta \mid x_i) - \nabla_\theta \log p_{t+1}(\theta \mid x_i)\|^2 \\
&\qquad+ 2\mathrm{E}_\theta\big\langle(1-n)(\nabla_\theta \log \lambda_t(\theta) - \nabla_\theta \log \lambda_{t+1}(\theta)), \sum_{i=1}^n \nabla_\theta \log p_t(\theta \mid x_i) - \nabla_\theta \log p_{t+1}(\theta \mid x_i)\big\rangle \\
&\qquad+ \mathrm{E}_\theta \sum_{i\neq j}\langle\nabla_\theta \log p_t(\theta \mid x_i) - \nabla_\theta \log p_{t+1}(\theta \mid x_i), \nabla_\theta \log p_t(\theta \mid x_j) - \nabla_\theta \log p_{t+1}(\theta \mid x_j)\rangle \\
&\quad\leq (1-n)^2\mathrm{E}_\theta\|\nabla_\theta \log \lambda_t(\theta) - \nabla_\theta \log \lambda_{t+1}(\theta)\|^2 + \sum_{i=1}^n \mathrm{E}_\theta\|\nabla_\theta \log p_t(\theta \mid x_i) - \nabla_\theta \log p_{t+1}(\theta \mid x_i)\|^2 \\
&\qquad+ 2(n-1)\sum_{i=1}^n \mathrm{E}_\theta\big\|\nabla_\theta \log \lambda_t(\theta) - \log \lambda_{t+1}(\theta)\|\|\nabla_\theta \log p_t(\theta \mid x_i) - \nabla_\theta \log p_{t+1}(\theta \mid x_i)\big\| \\
&\qquad+ \sum_{i\neq j} \mathrm{E}_\theta\|\nabla_\theta \log p_t(\theta \mid x_i) - \nabla_\theta \log p_{t+1}(\theta \mid x_i)\|\|\nabla_\theta \log p_t(\theta \mid x_j) - \nabla_\theta \log p_{t+1}(\theta \mid x_j)\| \\
&\qquad\text{by Cauchy-Schwarz and } n \geq 1 \\
&\quad\leq (1-n)^2\mathrm{E}_\theta\|\nabla_\theta \log \lambda_t(\theta) - \nabla_\theta \log \lambda_{t+1}(\theta)\|^2 + \sum_{i=1}^n \mathrm{E}_\theta\|\nabla_\theta \log p_t(\theta \mid x_i) - \nabla_\theta \log p_{t+1}(\theta \mid x_i)\|^2 \\
&\qquad+ 2(n-1)\sum_{i=1}^n \sqrt{\mathrm{E}_\theta\big\|\nabla_\theta \log \lambda_t(\theta) - \nabla_\theta \log \lambda_{t+1}(\theta)\|^2}\sqrt{\mathrm{E}_\theta\|\nabla_\theta \log p_t(\theta \mid x_i) - \nabla_\theta \log p_{t+1}(\theta \mid x_i)\big\|^2} \\
&\qquad+ \sum_{i\neq j} \sqrt{\mathrm{E}_\theta\|\nabla_\theta \log p_t(\theta \mid x_i) - \nabla_\theta \log p_{t+1}(\theta \mid x_i)\|^2}\sqrt{\mathrm{E}_\theta\|\nabla_\theta \log p_t(\theta \mid x_j) - \nabla_\theta \log p_{t+1}(\theta \mid x_j)\|^2} \\
&\qquad\text{by Cauchy-Schwarz a second time} \\
&\quad= \Big((n-1)\sqrt{\mathrm{E}_\theta\big\|\nabla_\theta \log \lambda_t(\theta) - \nabla_\theta \log \lambda_{t+1}(\theta)\|^2} + \sum_{i=1}^n \sqrt{\mathrm{E}_\theta\|\nabla_\theta \log p_t(\theta \mid x_i) - \nabla_\theta \log p_{t+1}(\theta \mid x_i)\|^2}\Big)^2.
\end{aligned}
$$

For Linhart et al. [16], we get the following bound:

$$
\begin{aligned}
&\mathrm{E}_\theta\left[\| \nabla_\theta \log \pi_t^L(\theta \mid x_{1:n}) - \nabla_\theta \log \pi_{t+1}^L(\theta \mid x_{1:n})\|^2\right] \\
&= \mathrm{E}_\theta \Bigg[\|(1-n)\Lambda_t^{-1}\Sigma_{t,\lambda}^{-1}\left(\nabla_\theta \log \lambda_t(\theta) - B_{t,\lambda}\nabla_\theta \log \lambda_{t+1}(\theta)\right) \\
&\quad + \sum_{i=1}^n \Lambda_t^{-1}\Sigma_{t,i}^{-1}(\nabla_\theta \log p_t(\theta \mid x_i) - B_{t,i}\nabla_\theta \log p_{t+1}(\theta \mid x_i))\|^2 \Bigg] \\
&\leq \Bigg((n-1)\sqrt{\mathrm{E}_\theta \left\|\Lambda_t^{-1}\Sigma_{t,\lambda}^{-1}(\nabla_\theta \log \lambda_t(\theta) - B_{t,\lambda}\nabla_\theta \log \lambda_{t+1}(\theta))\right\|^2} \\
&\quad + \sum_{i=1}^n \sqrt{\mathrm{E}_\theta \|\Lambda_t^{-1}\Sigma_{t,i}^{-1}(\nabla_\theta \log p_t(\theta \mid x_i) - B_{t,i}\nabla_\theta \log p_{t+1}(\theta \mid x_i))\|^2} \Bigg)^2 \\
&\quad \text{using the previous result on Geffner's densities} \\
&= \Bigg((n-1)\sqrt{\mathrm{E}_\theta \left\|\nabla_\theta \log \lambda_t(\theta) - B_{t,\lambda}\nabla_\theta \log \lambda_{t+1}(\theta)\right\|^2_{\Psi_{t,\lambda}^{-1}}} \\
&\quad + \sum_{i=1}^n \sqrt{\mathrm{E}_\theta \|\nabla_\theta \log p_t(\theta \mid x_i) - B_{t,i}\nabla_\theta \log p_{t+1}(\theta \mid x_i)\|^2_{\Psi_{t,i}^{-1}}} \Bigg)^2,
\end{aligned}
$$

where $\Psi_{t,k} = \Sigma_{t,k}\Lambda_t\Lambda_t^T\Sigma_{t,k}^T$ and $B_{t,k} = \Sigma_{t,k}\Lambda_t\Lambda_{t+1}^{-1}\Sigma_{t+1,k}^{-1}$ with $k \in \{\lambda, 0, \ldots, n\}$.

A.4 Fine-tuning of hyperparameters

We want to find hyperparameters $\{h_t\}_t$ and $\{k_t\}_t$ for each level $t \in \{t_0, \ldots, t_T\}$ such that the sampling quality is lower (in Wasserstein distance) than a given threshold $\gamma > 0$. We first assume that the score error is bounded by ϵ_{DSM}^2 for all levels t. We will use the intermediate bound in Proposition 1 and assume that the sampling quality of the previous Langevin process is controlled by γ, i.e. $\mathcal{W}_2(\mathcal{L}(\theta_{t_{p+1}}^{(k_{p+1})}), \pi_{t_{p+1}}) \leq \gamma$. Then we get:

$$
\begin{aligned}
&\mathcal{W}_2(\mathcal{L}(\theta_{t_p}^{(k_p)}), \pi_{t_p}) \\
&\quad \leq (1 - m_p h_p)^{k_p}\left(\mathcal{W}_2(\mathcal{L}(\theta_{t_{p+1}}^{(k_{p+1})}), \pi_{t_{p+1}}) + \mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p})\right) + 1.65\frac{M_p}{m_p}(h_p d)^{0.5} + \frac{\epsilon_{\text{DSM}}}{m_p} \\
&\quad \leq (1 - m_p h_p)^{k_p}\left(\gamma + \mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p})\right) + 1.65\frac{M_p}{m_p}(h_p d)^{0.5} + \frac{\epsilon_{\text{DSM}}}{m_p}.
\end{aligned}
$$

To ensure that $\mathcal{W}_2(\mathcal{L}(\theta_{t_p}^{(k_p)}), \pi_{t_p})$ is also lower than γ, it is sufficient to choose a step size h_p such that the red part is lower than $\omega\gamma$ with weight $\omega \in (0, 1)$, then choose a number of Langevin steps k_p such that the blue part is lower than $(1-\omega)\gamma$. This yields

$$
\begin{aligned}
0 \leq h_p &\leq \min\left(\frac{(\omega\gamma - \frac{\epsilon_{\text{DSM}}}{m_p})^2}{d(1.65)^2}\frac{m_p^2}{M_p^2}, \frac{2}{m_p + M_p}\right), \\
k_p &\geq \log\left(\frac{(1-\omega)\gamma}{\gamma + \mathcal{W}_2(\pi_{t_{p+1}}, \pi_{t_p})}\right)\frac{1}{\log(1 - m_p h_p)} \geq 0,
\end{aligned}
$$

where we replace the index t_p by p for readability concerns. Note that the step size is increasing in ω, while the number of steps is decreasing in $(1-\omega)$. This means that we can achieve the same Wasserstein accuracy γ by choosing ω sufficiently large (e.g. 0.8) so that the number of steps becomes relatively small and the step size relatively high. However, we need to ensure that $\omega\gamma > \frac{\epsilon_{\text{DSM}}}{m_p}$ to find a valid positive step size h_p.

A.5 Proof of Lemma 1 - Smoothness and log-concavity preservation

This section is split into two parts: the first one shows how smoothness is preserved over time for individual posteriors and intermediate densities under some mild assumptions; the second one studies the preservation of log-concavity.

A.5.1 Smoothness

Let p denote a density measure sufficiently smooth. We first recall the following equivalence:

$$\begin{aligned} p \text{ is (log) } L\text{-smooth} &\Leftrightarrow \|\nabla \log p(\theta) - \nabla \log p(\tilde{\theta})\| \leq L\|\theta - \tilde{\theta}\| \\ &\Leftrightarrow -LI_d \preceq -\nabla^2 \log p(\theta) \preceq LI_d, \\ p \text{ is log } L\text{-Lipschiz} &\Leftrightarrow \|\log p(\theta) - \log p(\tilde{\theta})\| \leq L\|\theta - \tilde{\theta}\| \\ &\Leftrightarrow -LI_d \preceq -\nabla \log p(\theta) \preceq LI_d, \end{aligned}$$

where $L > 0$ and θ belongs to the support of p. We also say that p is Gaussian log-Lipschitz if it admits a density with respect to the Gaussian measure that is log-Lipschitz.

Lemma 5 *Assume that each individual posterior distribution $p(\cdot \mid x_i)$ (resp. the prior distribution $\lambda(\cdot)$) admits a density with respect to the Gaussian measure that is L_i-log-Lipschitz (resp. L_λ-log-Lipschitz). Then, their diffused version along the VP forward scheme $p_t(\cdot \mid x_i)$ (resp. $\lambda_t(\cdot)$) become $L_{t,i}$ (resp. $L_{t,\lambda}$) -smooth where*

$$\begin{aligned} L_{t,i} &= (R_{t,i} + L_i^2\alpha_t + 1) \quad \textit{with} \quad R_{t,i} = 5L_i\alpha_t^{\frac{3}{2}}\left(L_i + \frac{1}{\sqrt{\frac{1}{2}\log(\frac{1}{\alpha_t})}}\right), \\ L_{t,\lambda} &= (R_{t,\lambda} + L_\lambda^2\alpha_t + 1) \quad \textit{with} \quad R_{t,\lambda} = 5L_\lambda\alpha_t^{\frac{3}{2}}\left(L_\lambda + \frac{1}{\sqrt{\frac{1}{2}\log(\frac{1}{\alpha_t})}}\right). \end{aligned}$$

Moreover, the bridging densities π_t^G and π_t^L (under additional symmetry conditions) are also smooth with constants depending on $L_{t,\lambda}$ and $L_{t,i}$ and defined in the proof.

Recall that individual posterior scores are learned with denoising score-matching before being used in compositional scores. Let ν denote one individual posterior and p_0 be its density with respect to the Lebesgue measure, i.e. $\frac{d\nu}{d\theta} = p_0$. Samples θ_0 from ν are noised during a VP forward diffusion process. This defines a Markov chain $(\theta_t)_{t\geq 0}$ such that

$$\begin{aligned} &\theta_0 \sim \nu, \\ &\theta_t = r_t\theta_0 + \sqrt{1 - r_t^2}Z \quad \text{with} \quad Z \sim \mathcal{N}(0, I), \end{aligned}$$

where $r_t^2 := \alpha_t = e^{-\int_0^t \beta(s)ds} \in [0, 1]$ is a VP diffusion coefficient and a strictly decreasing function of time such that $\alpha_0 = 1$ and $\alpha_t \to_{t\to 1} 0$. The forward transition kernel can thus be written as $q_{t|0}(\theta_t \mid \theta_0) = \mathcal{N}(r_t\theta_0, (1 - r_t^2)I)$. Let $f(\theta) = \frac{d\nu}{d\gamma_d}(\theta)$ denote the density of ν with respect to the

standard d-dimensional Gaussian measure. We notice that :

$$
\begin{aligned}
p_t(\theta_t) &= \int q_{t|0}(\theta_t \mid \theta_0) p_0(\theta_0) d\theta_0 \\
&= (1-r_t^2)^{-\frac{d}{2}} \int \phi\left(\frac{\theta_t - r_t\theta_0}{\sqrt{1-r_t^2}}\right) p_0(\theta_0) d\theta_0 \quad \text{where } \phi \text{ is the density of a standard Gaussian,} \\
&= (1-r_t^2)^{-\frac{d}{2}} \int \phi\left(\frac{\theta_t - r_t\theta_0}{\sqrt{1-r_t^2}}\right) d\nu(\theta_0) \\
&= (1-r_t^2)^{-\frac{d}{2}} \int \phi\left(\frac{\theta_t - r_t\theta_0}{\sqrt{1-r_t^2}}\right) \frac{d\nu}{d\gamma_d}(\theta_0) d\gamma_d(\theta_0) \\
&= (1-r_t^2)^{-\frac{d}{2}} \int \phi\left(\frac{\theta_t - r_t\theta_0}{\sqrt{1-r_t^2}}\right) f(\theta_0)\phi(\theta_0) d\theta_0 \quad \text{since } \phi(\theta) = \frac{d\gamma_d}{d\theta}(\theta) \\
&= (1-r_t^2)^{-\frac{d}{2}} C^{-2d} \int f(\theta_0) e^{-\frac{1}{2}\frac{(\theta_t - r_t\theta_0)^2}{1-r_t^2}} e^{-\frac{1}{2}\theta_0^2} d\theta_0 \\
&\quad \text{where } C = \sqrt{2\pi} \text{ is the normalizing constant of } \phi, \\
&= (1-r_t^2)^{-\frac{d}{2}} C^{-2d} \int f(\theta_0) e^{-\frac{1}{2}\left(\frac{\theta_t^2 - 2r_t\theta_t\theta_0 + r_t^2\theta_0^2}{1-r_t^2} + \theta_0^2\right)} d\theta_0 \\
&= (1-r_t^2)^{-\frac{d}{2}} C^{-2d} \int f(\theta_0) e^{-\frac{1}{2}\left(\frac{\theta_t^2 - 2r_t\theta_t\theta_0 + \theta_0^2}{1-r_t^2}\right)} d\theta_0 \\
&= (1-r_t^2)^{-\frac{d}{2}} C^{-2d} e^{-\frac{1}{2}\frac{\theta_t^2}{1-r_t^2}} e^{\frac{1}{2}\frac{r_t^2\theta_t^2}{1-r_t^2}} \int f(\theta_0) e^{-\frac{1}{2}\left(\frac{(\theta_0 - r_t\theta_t)^2}{1-r_t^2}\right)} d\theta_0 \\
&= (1-r_t^2)^{-\frac{d}{2}} C^{-2d} e^{-\frac{1}{2}\theta_t^2} \int f(\theta_0) e^{-\frac{1}{2}\left(\frac{(\theta_0 - r_t\theta_t)^2}{1-r_t^2}\right)} d\theta_0 \\
&= C^{-2d} e^{-\frac{1}{2}\theta_t^2} \int f\left(r_t\theta_t + \sqrt{1-r_t^2}u\right) e^{-\frac{1}{2}u^2} du \\
&\quad \text{by the change of variable } u = \frac{\theta_0 - r_t\theta_t}{\sqrt{1-r_t^2}}, \\
&= \phi(\theta_t) \int f\left(r_t\theta_t + \sqrt{1-r_t^2}u\right) \phi(u) du \\
&= \phi(\theta_t) \int f\left(r_t\theta_t + \sqrt{1-r_t^2}u\right) d\gamma_d(u) \\
&= \phi(\theta_t) \mathrm{E}_{Z\sim\gamma_d}\left[f\left(r_t\theta_t + \sqrt{1-r_t^2}Z\right)\right] \\
&= \phi(\theta_t) \mathcal{Q}_t f(\theta_t),
\end{aligned}
$$

where $\mathcal{Q}_t h(x) := \mathrm{E}_{Z\sim N(0,I)}\left[h(r_t x + \sqrt{1-r_t^2}Z)\right]$. We can use Corollary C.1 from Gao et al. [12] with their coefficients as per $\beta_t = r_t \geq 0$ and $\alpha_t = \sqrt{1-r_t^2} \in [0,1]$. As f is assumed to be L-Lipschitz, then we get:

$$(-R_t - L^2 r_t^2) I_d \preceq \nabla^2 \log \mathcal{Q}_t f(\theta_t) \preceq R_t I_d,$$

where $R_t := 5Lr_t^3\left(L + \frac{1}{\sqrt{\log(\frac{1}{r_t})}}\right)$.

Then we note that:

$$
\begin{aligned}
-\nabla_\theta^2 \log p_t(\theta_t) &= -\nabla_\theta^2 \log \phi(\theta_t) - \nabla_\theta^2 \log \mathcal{Q}_t f(\theta_t) \\
&= I_d - \nabla_\theta^2 \log \mathcal{Q}_t f(\theta_t).
\end{aligned}
$$

Using the previous bounds, we get :

$$(1 - R_t)I_d \preceq -\nabla^2_\theta \log p_t(\theta_t) \preceq (R_t + L^2 r_t^2 + 1)I_d.$$

Replacing r_t by $\sqrt{\alpha_t}$ in the context of VP diffusion process, we get bounds for the Hessian of all diffused individual posteriors $p_t(\cdot \mid x_i)$, meaning that they are all $L_{t,i}$ smooth where $L_{t,i}$ depends on L_i as follows:

$$R_{t,i} = 5L_i\alpha_t^{\frac{3}{2}}\left(L_i + \frac{1}{\sqrt{\frac{1}{2}\log(\frac{1}{\alpha_t})}}\right),$$
$$L_{t,i} = (R_{t,i} + L_i^2 r_t^2 + 1) := (R_{t,i} + L_i^2\alpha_t + 1).$$

Let's now recall that the bridging densities π_t^G defined by Geffner are of the form $\pi_t^G(\theta \mid x_{1:n}) \propto e^{\int s_t^G(\theta|x_{1:n})d\theta}$ where s_t^G is defined as follows:

$$s_t^G(\theta \mid x_{1:n}) = (1-n)\nabla_\theta \log \lambda_t(\theta) + \sum_{i=1}^n \nabla_\theta \log p_t(\theta \mid x_i).$$

Consequently, we have:

$$-\nabla^2_\theta \log \pi_t^G(\theta \mid x_{1:n}) = -\nabla_\theta s_t^G(\theta \mid x_{1:n}) = -(1-n)\nabla^2_\theta \log \lambda_t(\theta) - \sum_{i=1}^n \nabla^2_\theta \log p_t(\theta \mid x_i).$$

Then, as $1 - n \leq 0$, we get the following bounds:

$$-\nabla^2_\theta \log \pi_t^G(\theta \mid x_{1:n}) \succeq (1-n)L_{t,\lambda} - \sum_{i=1}^n L_{t,i},$$
$$-\nabla^2_\theta \log \pi_t^G(\theta \mid x_{1:n}) \preceq (n-1)L_{t,\lambda} + \sum_{i=1}^n L_{t,i}.$$

So π_t^G is $\tilde{L}_t^G$-smooth with constant $\tilde{L}_t^G = (n-1)L_{t,\lambda} + \sum_{i=1}^n L_{t,i}$. Let's now turn to Linhart's bridging densities $\pi_t^L(\theta \mid x_{1:n})$ that underlie the compositional score s_t^L defined as follows:

$$s_t^L(\theta \mid x_{1:n}) = \Lambda^{-1}\left(\sum_{i=1}^n \Sigma_{t,i}^{-1}\nabla_\theta \log p_t(\theta \mid x_i) + (1-n)\Sigma_{t,\lambda}^{-1}\nabla_\theta \log \lambda_t(\theta)\right),$$

where Λ, $\Sigma_{t,\lambda}$ and $\Sigma_{t,i}$ for all i are symmetric positive definite matrices. We can then bound each of them as follows :

$$\frac{1}{\sigma_{\max}(\Lambda)}I \preceq \Lambda^{-1} \preceq \frac{1}{\sigma_{\min}(\Lambda)}I,$$
$$\frac{1}{\sigma_{\max}(\Sigma_{t,i})}I \preceq \Sigma_{t,i}^{-1} \preceq \frac{1}{\sigma_{\min}(\Sigma_{t,i})}I \quad \text{for all } i,$$
$$\frac{1}{\sigma_{\max}(\Sigma_{t,\lambda})}I \preceq \Sigma_{t,\lambda}^{-1} \preceq \frac{1}{\sigma_{\min}(\Sigma_{t,\lambda})}I,$$

where $\sigma_{\min}(A)$ (resp. $\sigma_{\max}(A)$) denotes the minimal (resp. maximal) eigenvalue of A. Then,

$$-\nabla^2_\theta \log \pi_t^L(\theta \mid x_{1:n}) = -\nabla_\theta s_t^L(\theta \mid x_{1:n}) = -\Lambda^{-1}\left(\sum_{i=1}^n \Sigma_{t,i}^{-1}\nabla^2_\theta \log p_t(\theta \mid x_i) + (1-n)\Sigma_{t,\lambda}^{-1}\nabla^2_\theta \log \lambda_t(\theta)\right).$$

If we assume that $\Sigma_{t,i}^{-1}\nabla^2_\theta \log p_t(\theta \mid x_i)$ is symmetric for all i (which is equivalent to say that both matrices commute) and as $1 - n \leq 0$, we get:

$$-\nabla^2_\theta \log \pi_t^L(\theta) \succeq -\frac{1}{\sigma_{\min}(\Lambda)}\left(\sum_{i=1}^n \frac{1}{\sigma_{\min}(\Sigma_{t,i})}L_{t,i} + (n-1)\frac{1}{\sigma_{\min}(\Sigma_{t,\lambda})}L_{t,\lambda}\right)I,$$
$$-\nabla^2_\theta \log \pi_t^L(\theta) \preceq \frac{1}{\sigma_{\min}(\Lambda)}\left(\sum_{i=1}^n \frac{1}{\sigma_{\min}(\Sigma_{t,i})}L_{t,i} + (n-1)\frac{1}{\sigma_{\min}(\Sigma_{t,\lambda})}L_{t,\lambda}\right)I.$$

So π_t^L is $\tilde{L}_t^L$-smooth with $\tilde{L}_t^L = \frac{1}{\sigma_{\min}(\Lambda)}\left(\sum_{i=1}^n \frac{1}{\sigma_{\min}(\Sigma_{t,i})}L_{t,i} + (n-1)\frac{1}{\sigma_{\min}(\Sigma_{t,\lambda})}L_{t,\lambda}\right)$ for all t.

A.5.2 Log-concavity

Lemma 6 *Assume that each individual posterior (resp. the prior) is Gaussian log-Lipschitz, then it becomes smooth along the forward diffusion scheme, which is equivalent to:*

$$m_{t,\lambda} I_d \preceq -\nabla^2_\theta \log \lambda_t(\theta) \preceq M_{t,\lambda} I_d,$$
$$m_{t,i} I_d \preceq -\nabla^2_\theta \log p_t(\theta \mid x_i) \preceq M_{t,i} I_d,$$

where $m_{t,\lambda}, m_{t,i} \in \mathbb{R}$ and $M_{t,\lambda}, M_{t,i} \in \mathbb{R}_+$. Let $\Sigma_{t,i}$ (resp. $\Sigma_{t,\lambda}$) denote the covariance matrices of the Gaussian backward kernels in Linhart's assumption. Then, assuming $\frac{1}{n}\sum_{i=1}^n \frac{1}{\gamma_{t,i}} m_{t,i} \geq \frac{n-1}{n}\frac{1}{\gamma_{t,\lambda}} M_{t,\lambda}$ is sufficient for Geffner's bridging densities to be log-concave with $\gamma_{t,i} = \gamma_{t,\lambda} = 1$. Under additional symmetry conditions, it is also sufficient for Linhart's bridging densities to be log-concave with $\gamma_{t,i} = \sigma_{min}(\Sigma_{t,i})$ (resp. $\sigma_{max}(\Sigma_{t,i})$) if $m_{t,i} \leq 0$ (resp. ≥ 0) and $\gamma_{t,\lambda} = \sigma_{min}(\Sigma_{t,\lambda})$.

Let's first assume that the prior $\lambda(\cdot)$ and all individual posteriors $p(\cdot \mid x_i)$ are log-concave with constant m_λ and m_i. Then their diffused versions correspond to the convolution of the scaled target density with the Gaussian transition kernel $q_{t|0}(\theta_t \mid \theta_0)$:

$$\begin{aligned}
\lambda_t(\theta_t) &= \int q_{t|0}(\theta_t \mid \theta_0)\lambda(\theta_0) d\theta_0 \\
&= \int e^{-\frac{1}{2}\frac{(\theta_t - \sqrt{\alpha_t}\theta_0)^2}{1-\alpha_t}} \lambda(\theta_0) d\theta_0 \\
&= \alpha_t^{-\frac{d}{2}} \int e^{-\frac{1}{2}\frac{(\theta_t - u)^2}{1-\alpha_t}} \lambda\left(\frac{u}{\sqrt{\alpha_t}}\right) du \quad \text{by the change of variables } u = \sqrt{\alpha_t}\theta_0.
\end{aligned}$$

The scaling function $x \to x\alpha_t^p$ is an affine transformation (for any $p \in \mathbb{R}$) that preserves the log-concavity of λ. Then, the convolution of two log-concave densities remains log-concave [21]. So λ_t is log-concave for all $t \in [0, 1]$. The same reasoning holds for each individual posterior $p(\cdot \mid x_i)$. So, log-concavity is preserved along the forward path. Silveri and Ocello [23] also show that the weak log-concavity of the data distribution evolves into log-concavity over time during the forward diffusion scheme.

Assume now that, for a given time t, some individual posteriors are only smooth, meaning that some eigenvalues of the Hessian of their scores are still **negative**: the previous lemma indeed shows that smoothness is strengthened along the forward diffusion scheme, which translates as per:

$$m_{t,\lambda} I_d \preceq -\nabla^2_\theta \log \lambda_t(\theta) \preceq M_{t,\lambda} I_d,$$
$$m_{t,i} I_d \preceq -\nabla^2_\theta \log p_t(\theta \mid x_i) \preceq M_{t,i} I_d,$$

where $m_{t,i} \geq 0$ for $i \in P$, $m_{t,i} \leq 0$ for $i \in N$ with $N \cup P = \{1, \dots, n\}$, $P \cap N = \emptyset$ and $m_{t,\lambda} \in \mathbb{R}$. If we assume that $\sum_{i=1}^n m_{t,i} \geq (n-1) M_{t,\lambda}$ then using the bounds derived to prove the previous lemma, we get

$$-\nabla^2_\theta \log \pi^G_t(\theta \mid x_{1:n}) \succeq \left(\underbrace{(1-n) M_{t,\lambda} + \sum_{i=1}^n m_{t,i}}_{\geq 0} \right) I_d,$$

which shows that Geffner's bridging densities are log-concave over time. Note that the previous condition does not require all individual posteriors to be log-concave: for n sufficiently large, positive constants $m_{t,i}$ may catch up for negative ones in the sum, so that their mean remain positive.
Let's now take a look at Linhart's intermediate densities with the same smoothness conditions on each individual posterior and the prior. Let's assume that

$$\sum_{i \in P} \frac{1}{\sigma_{\max}(\Sigma_{t,i})} m_{t,i} + \sum_{i \in N} \frac{1}{\sigma_{\min}(\Sigma_{t,i})} m_{t,i} \geq \frac{n-1}{\sigma_{\min}(\Sigma_{t,\lambda})} M_{t,\lambda},$$

then we have:

$$-\nabla_\theta \log \pi_t^L(\theta) \succeq \frac{1}{\sigma_{\max}(\Lambda)} \left(\underbrace{\sum_{i \in P} \frac{m_{t,i}}{\sigma_{\max}(\Sigma_{t,i})} + \sum_{i \in N} \frac{m_{t,i}}{\sigma_{\min}(\Sigma_{t,i})} + (1-n)\frac{M_{t,\lambda}}{\sigma_{\min}(\Sigma_{t,\lambda})}}_{\geq 0} \right) I_d.$$

Again, the previous condition does not require all individual posteriors to be log-concave (meaning that $m_{t,i}$ are positive) for Linhart's bridging density to be log-concave: negative constants may catch up for negative ones in the sum.

A.6 Analytical bridging densities in a Gaussian setting

We derive here the analytical formula of the bridging densities in a Gaussian setting provided in Section 4.1. Recall that we use a VP forward process to train the individual scores: therefore, all annealing levels $t = t_0, \dots, t_T$ live in $[0, 1]$ with the convention that $t_0 = 0$ and $t_T = 1$. We start with the intermediate densities defined by Linhart et al. [16]. We first recall the formula of the individual posteriors $p(\theta \mid x_i) = \mathcal{N}(\theta; \mu_{\text{post}}(x_i), \Sigma_{\text{post}})$ and the multi-observation posterior $p(\theta \mid x_{1:n}) = \mathcal{N}(\theta; \mu_{\text{post},n}(x_{1:n}), \Sigma_{\text{post},n})$ in the Gaussian setting [3], where:

$$\Sigma_{\text{post}} = (\Sigma^{-1} + I_d)^{-1} \quad \text{and} \quad \Sigma_{\text{post},n} = (n\Sigma^{-1} + I_d)^{-1},$$
$$\mu_{\text{post}}(x_i) = \Sigma_{\text{post}}\Sigma^{-1}x_i \quad \text{and} \quad \mu_{\text{post},n}(x_{1:n}) = \Sigma_{\text{post},n}\Sigma^{-1}\sum_{i=1}^n x_i.$$

In the Gaussian setting, as all the assumptions of Linhart's work are fulfilled, their compositional score exactly corresponds to the diffused scores arising from a classical diffusion process. Thus, the underlying distributions π_t^L are the diffused versions of the target multi-observation posterior, i.e. $\pi_t^L(\theta \mid x_{1:n}) = p_t(\theta \mid x_{1:n}) = \mathcal{N}(\theta, \mu_t^L, \Sigma_t^L)$ where:

$$\Sigma_t^L = \alpha_t \Sigma_{\text{post,n}} + v_t I_d = \alpha_t(n\Sigma^{-1} + I_d)^{-1} + v_t I_d,$$
$$\mu_t^L = \sqrt{\alpha_t}\mu_{\text{post}}(x_{1:n}) = \sqrt{\alpha_t}\Sigma_{\text{post,n}}\Sigma^{-1}\sum_{i=1}^n x_i = \sqrt{\alpha_t}(nI_d + \Sigma)^{-1}\sum_{i=1}^n x_i.$$

We then consider the bridging densities defined by Geffner et al. [13] :

$$\pi_t^G(\theta) = \lambda_t(\theta)^{1-n}\prod_{i=1}^n p_t(\theta \mid x_i),$$

where the diffused distributions are Gaussian of the form

$$\lambda_t(\theta) = \mathcal{N}(\theta; 0, I_d) \quad \forall t \in (0, 1),$$
$$p_t(\theta \mid x_i) = \mathcal{N}(\theta; \sqrt{\alpha_t}\mu_{\text{post}}(x_i), \underbrace{\alpha_t(\Sigma^{-1} + I_d)^{-1} + v_t I_d)}_{\Sigma_t} \quad \text{diffused posterior under VP.}$$

Then all the bridging densities are also Gaussian of the form $\pi_t^G(\theta) = \mathcal{N}(\theta; \mu_t^G, \Sigma_t^G)$ where

$$\Sigma_t^G = \left((1-n)I_d + \sum_{i=1}^n \Sigma_t^{-1}\right)^{-1} = \left((1-n)I_d + n\Sigma_t^{-1}\right)^{-1},$$
$$\mu_t^G = \Sigma_t^G\left(\sum_{i=1}^n \Sigma_t^{-1}\sqrt{\alpha_t}\mu_{\text{post}}(x_i)\right) = \sqrt{\alpha_t}\Sigma_t^G\Sigma_t^{-1}\Sigma_{\text{post}}\Sigma^{-1}\left(\sum_{i=1}^n x_i\right)$$
$$= \sqrt{\alpha_t}\Sigma_t^G\Sigma_t^{-1}(I_d + \Sigma)^{-1}\left(\sum_{i=1}^n x_i\right).$$

For a well-defined Gaussian density, we need to be sure that the covariance matrix Σ_t^G is positive definite. In our case, Σ_t and I_d commute, so do Σ and I_d. Thus, we can diagonalize all matrices

in a common eigenbasis and compute the eigenvalues of any linear combination of them. Let $\sigma(A)$ denotes one eigenvalue of the matrix A and recall that $\sigma(A^{-1}) = \frac{1}{\sigma(A)}$. Then,

$$\begin{aligned}
\sigma((\Sigma_t^G)^{-1}) = (1-n) + n\sigma(\Sigma_t^{-1}) &= 1 - n + \frac{n}{\alpha_t \frac{\sigma(\Sigma)}{\sigma(\Sigma)+1} + v_t} \\
&= 1 - n + \frac{n(\sigma(\Sigma)+1)}{\sigma(\Sigma) + v_t} \\
&= 1 + n\left(\frac{1-v_t}{v_t + \sigma(\Sigma)}\right) \\
&\geq 1 \quad \text{since } v_t \in (0,1) \text{ and } \sigma(\Sigma) > 0.
\end{aligned}$$

When $t \to 1$, $v_t \to 1$ and the eigenvalues of Σ_t^G tend to 1, which is natural since π_t^G tends to a standard Gaussian. When $t \to 0$, then $v_t \to 0$ and the eigenvalue of Σ_t^G are equal to those of the multi-observation target posterior π_0, namely $\frac{\sigma(\Sigma)}{\sigma(\Sigma)+n}$. The covariance matrix of π_t^G remains well defined if one chooses another prior distribution with a covariance matrix that commutes with the likelihood one.

A.7 Proof of Lemma 2

Recall that the log-concavity constant m_t corresponds to the minimal eigenvalue of the precision matrix $(\Sigma_t)^{-1}$ of the intermediate density π_t while the smoothness constant M_t is the maximal eigenvalue of the same matrix. We start with the densities defined by Linhart.

$$\begin{aligned}
m_0 &= \min_\sigma \frac{n+\sigma}{\sigma} = \frac{n+\sigma_{\max}}{\sigma_{\max}} \quad \text{as the function decreases in } \sigma, \\
m_t^L &= \min_\sigma \frac{n+\sigma}{\sigma + nv_t} = \frac{n+\sigma_{\max}}{\sigma_{\max} + nv_t} \quad \text{as the function decreases in } \sigma \\
&= \frac{n+\sigma_{\max}}{\sigma_{\max}} \left(\frac{1}{1+\frac{nv_t}{\sigma_{\max}}}\right) \\
&= m_0 \left(\frac{1}{1+\frac{nv_t}{\sigma_{\max}}}\right), \\
M_0 &= \max_\sigma \frac{n+\sigma}{\sigma} = \frac{n+\sigma_{\min}}{\sigma_{\min}}, \\
M_t^L &= M_0 \left(\frac{1}{1+\frac{nv_t}{\sigma_{\min}}}\right) \quad \text{with the same reasoning.}
\end{aligned}$$

Then in Geffner's case we have

$$\begin{aligned}
m_t^G &= \min_\sigma \frac{\sigma + n + (1-n)v_t}{\sigma + v_t} = \frac{\sigma_{\max} + n + (1-n)v_t}{\sigma_{\max} + v_t} \quad \text{since the function decreases in } \sigma, \\
&= \frac{n+\sigma_{\max}}{\sigma_{\max}} \left(\frac{1}{1+\frac{v_t}{\sigma_{\max}}}\right) + \frac{(1-n)v_t}{\sigma_{\max} + v_t} \\
&= m_0 \left(\frac{1}{1+\frac{v_t}{\sigma_{\max}}}\right) + \frac{(1-n)v_t}{\sigma_{\max} + v_t}, \\
M_t^G &= M_0 \left(\frac{1}{1+\frac{v_t}{\sigma_{\min}}}\right) + \frac{(1-n)v_t}{\sigma_{\min} + v_t} \quad \text{with the same reasoning.}
\end{aligned}$$

A.8 Proof of Proposition 3

For Gaussian distributions, the log-concavity constant m_t^G or m_t^L (resp. smoothness constant M_t^G or M_t^L) corresponds to the minimal (resp. maximal) eigenvalue of its precision matrix $(\Sigma_t^G)^{-1}$ or $(\Sigma_t^L)^{-1}$. We first note that Σ, Σ_t^L and Σ_t^G commute: therefore, they share an eigen basis in which

they are diagonal. This allows us to easily derive the eigenvalues of any linear combination of them, particularly the ones of $(\Sigma_t^G)^{-1} - (\Sigma_t^L)^{-1}$. Let $\sigma(A)$ denote one eigenvalue of the matrix A and σ denote one eigenvalue of the likelihood covariance.

$$\begin{aligned}
\sigma((\Sigma_t^G)^{-1}) - \sigma((\Sigma_t^L)^{-1}) &= \frac{\sigma + v_t + n\alpha_t}{\sigma + v_t} - \frac{n + \sigma}{\sigma + nv_t} \\
&= \frac{(\sigma + v_t + n(1 - v_t))(\sigma + nv_t) - (\sigma + v_t)(n + \sigma)}{(\sigma + v_t)(\sigma + nv_t)} \\
&= \frac{nv_t^2 - nv_t + n^2(v_t - v_t^2)}{(\sigma + v_t)(\sigma + nv_t)} \\
&= \frac{n(n-1)v_t(1 - v_t)}{(\sigma + v_t)(\sigma + nv_t)} \\
&= \frac{n(n-1)v_t\alpha_t}{(\sigma + v_t)(\sigma + nv_t)} \geq 0 \quad \text{since } v_t = 1 - \alpha_t \\
&> 0 \quad \text{for } t \in (0, 1).
\end{aligned}$$

This shows that $(\Sigma_t^G)^{-1} - (\Sigma_t^L)^{-1} \succeq 0 \Leftrightarrow (\Sigma_t^G)^{-1} \succeq (\Sigma_t^L)^{-1}$, with equality only if $t = 0$ or $t = 1$. This also shows that the eigenvalues of $(\Sigma_t^G)^{-1}$ (denoted λ_i^G) are always higher than the ones of $(\Sigma_t^L)^{-1}$) (denoted λ_i^L) and that we can relate them analytically as per : $\lambda_i^G = \lambda_i^L + \frac{n(n-1)v_t\alpha_t}{(\sigma_i + v_t)(\sigma_i + nv_t)}$. Let's now study the denominator in the previous computation $D(\sigma, v_t) = (\sigma + v_t)(\sigma + nv_t)$ that is a polynomial of degree 2 in σ:

$$\begin{aligned}
D = (\sigma + v_t)(\sigma + nv_t) &= \sigma^2 + (n+1)v_t\sigma + nv_t^2, \\
\Delta_D = (n+1)^2v_t^2 - 4nv_t^2 &= (n-1)^2v_t^2 \geq 0, \\
r_D = \frac{-v_t(n+1) \pm v_t(n-1)}{2} &= \begin{cases} -nv_t \leq 0 \\ -v_t \leq 0 \end{cases}.
\end{aligned}$$

As σ should be positive, this means that D is positive and increases over $[0, \infty[$. So the function $g(\sigma) := \frac{n(n-1)\alpha_t v_t}{D(\sigma, v_t)}$ reaches its minimum in the maximal eigenvalue of the likelihood covariance σ_{max} and its maximum in the minimal eigenvalue of the likelihood covariance σ_{min}. We can also check that the eigenvalues of $(\Sigma_t^L)^{-1}$ are a decreasing function $f(\sigma) := \frac{n+\sigma}{\sigma + nv_t}$ of $\sigma \in]0, +\infty[$:

$$\frac{\partial}{\partial \sigma} f(\sigma) = \frac{\sigma + nv_t - n - \sigma}{(\sigma + nv_t)^2} = \frac{n(v_t - 1)}{(\sigma + nv_t)^2} \leq 0 \quad \text{since } v_t \in [0, 1].$$

f reaches its maximal and minimal values at the same points that g. Therefore,

$$\begin{aligned}
M_t^G &= \max_\sigma \left(\frac{n + \sigma}{\sigma + nv_t} + \frac{n(n-1)v_t\alpha_t}{D(\sigma, v_t)} \right) \\
&= \max_\sigma \frac{n + \sigma}{\sigma + nv_t} + \max_\sigma \frac{n(n-1)v_t\alpha_t}{D(\sigma, v_t)} \quad \text{since the functions reach their maximal value at the same point} \\
&= M_t^L + \frac{n(n-1)\alpha_t v_t}{D(\sigma_{\text{min}}, v_t)} \quad \text{since } D \text{ is decreasing in } \sigma.
\end{aligned}$$

Similarly,

$$\begin{aligned}
m_t^G &= \min_\sigma \left(\frac{n + \sigma}{\sigma + nv_t} + \frac{n(n-1)v_t\alpha_t}{D(\sigma, v_t)} \right) \\
&= \min_\sigma \frac{n + \sigma}{\sigma + nv_t} + \min_\sigma \frac{n(n-1)v_t\alpha_t}{D(\sigma, v_t)} \\
&\quad \text{since the functions reach their minimal value at the same point} \\
&= m_t^L + \frac{n(n-1)\alpha_t v_t}{D(\sigma_{\text{max}}, v_t)} \quad \text{since } D \text{ is decreasing in } \sigma.
\end{aligned}$$

A.9 Proof of Corollary 1

Proposition 3 shows that Geffner's constants m_t^G and M_t^G are linked to their Linhart's counterparts. Let $N_t = n(n-1)\alpha_t v_t$, $D_{\min} = D(\sigma_{\min})$ and $D_{\max} = D(\sigma_{\max})$. Then, we have:

$$\begin{aligned}
\frac{M_t^G}{m_t^G} &= \frac{M_t^L + \frac{N_t}{D_{\min}}}{m_t^L + \frac{N_t}{D_{\max}}} \\
&= \frac{M_t^L}{m_t^L + \frac{N_t}{D_{\max}}} + \frac{\frac{N_t}{D_{\min}}}{m_t^L + \frac{N_t}{D_{\max}}} \\
&= \frac{M_t^L}{m_t^L} \frac{1}{1 + \frac{N_t}{D_{\max} m_t^L}} + \frac{1}{m_t^L} \frac{\frac{N_t}{D_{\min}}}{1 + \frac{N_t}{m_t^L D_{\max}}}, \\
\frac{\frac{M_t^G}{m_t^G}}{\frac{M_t^L}{m_t^L}} &= \frac{1}{1 + \frac{N_t}{m_t^L D_{\max}}} + \frac{1}{M_t^L} \frac{\frac{N_t}{D_{\min}}}{1 + \frac{N_t}{m_t^L D_{\max}}} \\
&= \frac{1}{1 + \frac{N_t}{m_t^L D_{\max}}} \left(1 + \frac{N_t}{M_t^L D_{\min}}\right), \\
\frac{\frac{M_t^G}{m_t^G}}{\frac{M_t^L}{m_t^L}} \geq 1 &\Leftrightarrow 1 + \frac{N_t}{m_t^L D_{\max}} \leq 1 + \frac{N_t}{M_t^L D_{\min}} \\
&\Leftrightarrow m_t^L D_{\max} \geq M_t^L D_{\min}.
\end{aligned}$$

Let's now prove that the last condition is always true for any time t. For a fixed time t, we consider the function $l(\sigma) = \frac{n+\sigma}{\sigma+nv_t} D(\sigma, v_t) = (n+\sigma)(\sigma+v_t)$ after simplification.

$$\begin{aligned}
l(\sigma) &= (n+\sigma)(\sigma+v_t) = \sigma^2 + \sigma(v_t+n) + nv_t, \\
\frac{\partial}{\partial\sigma} l(\sigma) &= 2\sigma + n + v_t \geq 0 \; \forall \sigma \in [0, +\infty[.
\end{aligned}$$

So l is increasing in σ over $[0, +\infty[$. Recall that $m_t^L = \min_\sigma \frac{n+\sigma}{\sigma+nv_t} = \frac{n+\sigma_{\max}}{\sigma_{\max}+nv_t}$ and $M_t^L = \max_\sigma \frac{n+\sigma}{\sigma+nv_t} = \frac{n+\sigma_{\min}}{\sigma_{\min}+nv_t}$ since the function $f : \sigma \mapsto \frac{n+\sigma}{\sigma+nv_t}$ is decreasing over $[0, +\infty[$ (see Appendix A.8 for details). Thus, $m_t^L D(\sigma_{\max}) = l(\sigma_{\max})$ and $M_t^L D(\sigma_{\min}) = l(\sigma_{\min})$. As l is increasing then the condition is always true. Therefore, $\frac{M_t^G}{m_t^G}$ is higher than $\frac{M_t^L}{m_t^L}$ for any time $t \in [0, 1]$. and the inverse ratios are conversely ordered. As h_t^G and h_t^L (chosen as in Section 3) linearly depend on these inverse ratios, it is obvious that $h_t^G \leq h_t^L$ for any time $t \in [0, 1]$.

A.10 Proof of Proposition 4

Recall that the bridging densities π_t^G and π_t^L are Gaussian for all times $t \in [0, 1]$. The Wasserstein distance between two consecutive densities is analytically tractable and involves a mean difference term and the Bures distance between covariance matrices, that simplifies to a Fröbenius norm as prior and likelihood covariance matrices commute:

$$\begin{aligned}
\mathcal{W}_2^2(\mathcal{N}(\mu_{t_p}, \Sigma_{t_p}), \mathcal{N}(\mu_{t_{p+1}}, \Sigma_{t_{p+1}})) &= \underbrace{\|\mu_{t_p} - \mu_{t_{p+1}}\|_2^2}_{\text{mean difference}} + \underbrace{\operatorname{tr}(\Sigma_{t_p} + \Sigma_{t_{p+1}} - 2(\Sigma_{t_p}^{\frac{1}{2}} \Sigma_{t_{p+1}} \Sigma_{t_p}^{\frac{1}{2}})^{\frac{1}{2}})}_{\text{Bures distance}} \\
&= \|\mu_{t_p} - \mu_{t_{p+1}}\|_2^2 + \underbrace{\|\Sigma_{t_p}^{\frac{1}{2}} - \Sigma_{t_{p+1}}^{\frac{1}{2}}\|_F^2}_{\text{Fröbenius norm}}.
\end{aligned}$$

To demonstrate Proposition 4, we will first prove two intermediate lemmas: Lemma 7 compares the Bures distance between covariance matrices for both algorithms, while Lemma 8 studies the mean difference terms between algorithms. The reader is referred to Appendix A.10.3 for the final proof of Proposition 4.

Notation convention: in the rest of Appendix A.10, we use the abuse of notation π_{t+1} (instead of the usual $\pi_{t_{p+1}}$) to denote the intermediate density following π_t during the annealed sampling scheme for one given $t \in \{t_0, \ldots, t_T\} \cap [0, 1]$.

A.10.1 Comparison of Bures distances

Lemma 7 (Comparison of Bures distances) *Let Σ_t^L (resp. Σ_t^G) denote the covariance of the intermediate density defined by Linhart (resp. Geffner) for a given time t and d_B denote the Bures distance. Then, we have*

$$\begin{aligned} &\text{As } t \to 0, \quad d_B(\Sigma_t^G, \Sigma_{t+1}^G) \leq d_B(\Sigma_t^L, \Sigma_{t+1}^L). \\ &\text{As } t \to 1, \quad d_B(\Sigma_t^G, \Sigma_{t+1}^G) \geq d_B(\Sigma_t^L, \Sigma_{t+1}^L). \end{aligned}$$

We first note that in the Gaussian setting, the intermediate densities π_t and π_{t+1} are Gaussian with commuting covariance matrices, regardless of the algorithm (Linhart or Geffner). Thus, the Bures distance between their corresponding covariance matrices simplify to a Fröbenius norm as per:

$$\begin{aligned} Tr(\Sigma_t + \Sigma_{t+1} - 2(\Sigma_t^{\frac{1}{2}}\Sigma_{t+1}\Sigma_t^{\frac{1}{2}})^{\frac{1}{2}}) &= Tr(\Sigma_t + \Sigma_{t+1} - 2(\Sigma_t\Sigma_{t+1})^{\frac{1}{2}}) \\ &= Tr((\Sigma_t^{\frac{1}{2}} - \Sigma_{t+1}^{\frac{1}{2}})^T(\Sigma_t^{\frac{1}{2}} - \Sigma_{t+1}^{\frac{1}{2}})) \\ &= \|\Sigma_t^{\frac{1}{2}} - \Sigma_{t+1}^{\frac{1}{2}}\|_F^2 \\ &= \sum_{i=1}^{d} \left(\sqrt{\sigma_i(\Sigma_t)} - \sqrt{\sigma_i(\Sigma_{t+1})}\right)^2, \end{aligned}$$

where d is the dimension of the space and $\sigma_i(A)$ denotes the ith eigenvalue of the matrix A in the common eigenbasis. To compare these Bures distances, we study the sign of their difference over time as follows:

$$\begin{aligned} &\sum_{i=1}^{d} \left(\sqrt{\sigma_i(\Sigma_t^G)} - \sqrt{\sigma_i(\Sigma_{t+1}^G)}\right)^2 - \sum_{i=1}^{p} \left(\sqrt{\sigma_i(\Sigma_t^L)} - \sqrt{\sigma_i(\Sigma_{t+1}^L)}\right)^2 \\ &= \sum_{i=1}^{d} \left(\sqrt{\sigma_i(\Sigma_t^G)} - \sqrt{\sigma_i(\Sigma_{t+1}^G)}\right)^2 - \left(\sqrt{\sigma_i(\Sigma_t^L)} - \sqrt{\sigma_i(\Sigma_{t+1}^L)}\right)^2 \\ &= \sum_{i=1}^{d} \left(\sqrt{\sigma_i(\Sigma_t^G)} - \sqrt{\sigma_i(\Sigma_{t+1}^G)} + \sqrt{\sigma_i(\Sigma_t^L)} - \sqrt{\sigma_i(\Sigma_{t+1}^L)}\right) \\ &\quad \cdot \left(\sqrt{\sigma_i(\Sigma_t^G)} - \sqrt{\sigma_i(\Sigma_{t+1}^G)} - \sqrt{\sigma_i(\Sigma_t^L)} + \sqrt{\sigma_i(\Sigma_{t+1}^L)}\right) \\ &= \sum_{i=1}^{d} \left(\sqrt{\sigma_i(\Sigma_t^G)} + \sqrt{\sigma_i(\Sigma_t^L)} - \left(\sqrt{\sigma_i(\Sigma_{t+1}^G)} + \sqrt{\sigma_i(\Sigma_{t+1}^L)}\right)\right) \\ &\quad \cdot \left(\sqrt{\sigma_i(\Sigma_t^G)} - \sqrt{\sigma_i(\Sigma_t^L)} - \left(\sqrt{\sigma_i(\Sigma_{t+1}^G)} - \sqrt{\sigma_i(\Sigma_{t+1}^L)}\right)\right) \\ &= \sum_{i=1}^{d} \left(\sqrt{\sigma_i(\Sigma_t^L)}\left(1 + \frac{1}{\sqrt{1+\sigma_i(\Sigma_t^L)\Omega_{t,i}}}\right) - \sqrt{\sigma_i(\Sigma_{t+1}^L)}\left(1 + \frac{1}{\sqrt{1+\sigma_i(\Sigma_{t+1}^L)\Omega_{t+1,i}}}\right)\right) \\ &\quad \cdot \left(\sqrt{\sigma_i(\Sigma_t^L)}\left(\frac{1}{\sqrt{1+\sigma_i(\Sigma_t^L)\Omega_{t,i}}} - 1\right) - \sqrt{\sigma_i(\Sigma_{t+1}^L)}\left(\frac{1}{\sqrt{1+\sigma_i(\Sigma_{t+1}^L)\Omega_{t+1,i}}} - 1\right)\right), \end{aligned} \tag{*}$$

where $\Omega_{t,i} = \frac{n(n-1)\alpha_t v_t}{(\sigma_i + v_t)(\sigma_i + nv_t)} \geq 0$. Recall that all matrices commute, their eigenvalues are then ordered in the common eigen basis.

Behavior when $t \to 1$. We will first show that the red term is negative whatever the value of σ_i and Ω_i from a specific time step t_0. To do so, we will consider the function of time $t \mapsto \sqrt{\sigma(\Sigma_t^L)}\left(1 + \frac{1}{\sqrt{1+\sigma(\Sigma_t^L)\Omega_t}}\right)$ and show that it increases in time on $[t_0, 1]$.

Recall that $\sigma(\Sigma_t^L) = \frac{\sigma+nv_t}{n+\sigma}$. Thus for $n \geq 1$, the eigenvalues of the posterior ($\frac{\sigma}{n+\sigma}$) are lower than 1 and $\sigma(\Sigma_t^L)$ is **increasing** over time from $\frac{\sigma}{n+\sigma}$ to 1. Indeed

$$\frac{\partial}{\partial t}\sigma(\Sigma_t^L) = \frac{n\dot{v}_t}{n+\sigma} \geq 0 \text{ since } \dot{v}_t \geq 0 \; \forall t.$$

Concerning the function $\Omega_t = \frac{N_t}{D_t}$ we note that :

$$\frac{N_t}{D_t} = \frac{n(n-1)\alpha_t v_t}{(\sigma+v_t)(\sigma+nv_t)} > 0 \; \forall t \in (0,1),$$
$$\frac{N_1}{D_1} = \frac{N_0}{D_0} = 0 \text{ since } v_0 = 0 \text{ and } \alpha_1 = 0.$$

Let $\boxed{f(t) := \sigma(\Sigma_t^L)\Omega_t}$. f is a continuous and derivable function of time over the closed interval $[0,1]$, it reaches a global maximum at some time point t_0. Let's show that f is decreasing from this time point.

$$\begin{aligned}
\frac{\partial}{\partial t}f(t) &= \frac{\partial}{\partial t}\left(\sigma(\Sigma_t^L)\Omega_t\right) \\
&= \frac{\partial}{\partial t}\frac{n(n-1)\alpha_t v_t}{(n+\sigma)(v_t+\sigma)} \\
&= \frac{n(n-1)}{n+\sigma}\frac{\partial}{\partial t}\frac{\alpha_t v_t}{v_t+\sigma} \\
&= \frac{n(n-1)}{n+\sigma}\frac{(\dot{\alpha}_t v_t + \alpha_t \dot{v}_t)(v_t+\sigma) - \dot{v}_t\alpha_t v_t}{(v_t+\sigma)^2} \\
&= \frac{n(n-1)}{(n+\sigma)(v_t+\sigma)^2}\left[(\dot{\alpha}_t v_t - \alpha_t\dot{\alpha}_t)(v_t+\sigma) + \dot{\alpha}_t\alpha_t v_t\right] \quad \text{since } \dot{\alpha}_t = -\dot{v}_t \\
&= \frac{\dot{\alpha}_t n(n-1)}{(n+\sigma)(v_t+\sigma)^2}\left[(v_t-\alpha_t)(v_t+\sigma) + \alpha_t v_t\right] \\
&= \underbrace{\frac{\dot{\alpha}_t n(n-1)}{(n+\sigma)(v_t+\sigma)^2}}_{\leq 0}\underbrace{\left[v_t^2 + 2\sigma v_t - \sigma\right]}_{R(v_t)} \quad \text{since } \alpha_t = 1 - v_t \text{ and } \dot{\alpha}_t \leq 0,
\end{aligned}$$

$$\begin{aligned}
\Delta_R &= 4\sigma^2 + 4\sigma \geq 0, \\
r_1 &= -\sigma - \sqrt{\sigma(\sigma+1)} \leq 0 \text{ and } r_2 = -\sigma + \sqrt{\sigma(\sigma+1)} \in [0, \frac{1}{2}] \; \forall \sigma > 0 \\
&\Rightarrow R(v_t) \geq 0 \text{ for } v_t \geq r_2 \\
&\Rightarrow \frac{\partial}{\partial t}f(t) \begin{cases} \leq 0 & \text{if} \quad 1 \geq v_t \geq r_2 \\ \geq 0 & \text{if} \quad 0 \leq v_t \leq r_2 \end{cases}.
\end{aligned}$$

As v_t is strictly increasing there exists a bijection between $t \in [0,1]$ and $v_t \in [0,1]$, therefore there exists a time step t_0 such that $v_{t_0} = r_2$. $f(t)$ is **non negative** and thus **increases** on $[0,t_0]$ then **decreases** on $[t_0,1]$. Therefore, $t \mapsto \frac{1}{\sqrt{1+f(t)}} + 1 = \frac{1}{\sqrt{1+\sigma(\Sigma_t^L)\Omega_t}} + 1$ is **increasing** and **non negative** on $[t_0,1]$. We can then deduce that for $t \in [t_0,1]$:

$$0 \leq \frac{1}{\sqrt{1+\sigma(\Sigma_t^L)\Omega_t}} + 1 \leq \frac{1}{\sqrt{1+\sigma(\Sigma_{t+1}^L)\Omega_{t+1}}} + 1,$$

$$\begin{aligned}
\sqrt{\sigma(\Sigma_t^L)}\left(\frac{1}{\sqrt{1+\sigma(\Sigma_t^L)\Omega_t}} + 1\right) &\leq \sqrt{\sigma(\Sigma_t^L)}\left(\frac{1}{\sqrt{1+\sigma(\Sigma_{t+1}^L)\Omega_{t+1}}} + 1\right) \quad \text{since } \sigma(\Sigma_t^L) \geq 0 \; \forall t \\
&\leq \sqrt{\sigma(\Sigma_{t+1}^L)}\left(\frac{1}{\sqrt{1+\sigma(\Sigma_{t+1}^L)\Omega_{t+1}}} + 1\right) \\
&\quad \text{since } \sigma(\Sigma_{t+1}^L) \text{ is increasing over } [0,1].
\end{aligned}$$

Thus the red term is **negative** whatever the item i of the sum on the interval $[t_0, 1]$.

Let's now study the blue term. First recall that for $t \to 1$, $\sigma(\Sigma_t^L) \to 1$ and $\Omega_t \to 0$. We can derive a first order expansion at a given point t close to 1:

$$\begin{aligned}\sqrt{\sigma(\Sigma_t^L)}\left(\frac{1}{\sqrt{1+\sigma(\Sigma_t^L)\Omega_t}}-1\right) &= \sqrt{\sigma(\Sigma_t^L)}\left(1-\frac{1}{2}\sigma(\Sigma_t^L)\Omega_t - 1 + o(\sigma(\Sigma_t^L)\Omega_t)\right)\\ &= -\frac{1}{2}\sigma(\Sigma_t^L)^{\frac{3}{2}}\Omega_t + o(\sigma(\Sigma_t^L)^{\frac{3}{2}}\Omega_t).\end{aligned}$$

Locally around t sufficiently close to 1, the function $\sqrt{\sigma(\Sigma_t^L)}\left(\frac{1}{\sqrt{1+\sigma(\Sigma_t^L)\Omega_t}}-1\right)$ behaves like $\boxed{-\frac{1}{2}g(t) := -\frac{1}{2}\sigma(\Sigma_t^L)^{\frac{3}{2}}\Omega_t}$. Let's study g for time point near 1.

$$\begin{aligned}g(t) &:= \sigma(\Sigma_t^L)^{\frac{3}{2}}\Omega_t\\ &= \frac{n(n-1)\alpha_t v_t}{(\sigma+v_t)(\sigma+nv_t)}\frac{(\sigma+nv_t)^{\frac{3}{2}}}{(\sigma+n)^{\frac{3}{2}}}\\ &= \frac{n(n-1)\alpha_t v_t\sqrt{\sigma+v_t n}}{(\sigma+v_t)(\sigma+n)^{\frac{3}{2}}}\\ &= \underbrace{\frac{n(n-1)}{(\sigma+n)^{\frac{3}{2}}}}_{:=C\geq 0}\frac{(\alpha_t-\alpha_t^2)\sqrt{\sigma+v_t n}}{\sigma+v_t},\end{aligned}$$

$$\begin{aligned}C^{-1}\frac{\partial}{\partial t}g(t) &= \frac{(\sigma+v_t)\left[(\dot\alpha_t - 2\dot\alpha_t\alpha_t)\sqrt{\sigma+nv_t} + (\alpha_t-\alpha_t^2)\frac{n\dot v_t}{2\sqrt{\sigma+nv_t}}\right] - \dot v_t(\alpha_t v_t\sqrt{\sigma+nv_t})}{(\sigma+v_t)^2}\\ &= \frac{1}{2\sqrt{\sigma+nv_t}(\sigma+v_t)^2}\left[2(\dot\alpha_t-2\dot\alpha_t\alpha_t)(\sigma+nv_t)(\sigma+v_t)+\alpha_t v_t n\dot v_t(\sigma+v_t) - 2\alpha_t v_t\dot v_t(\sigma+nv_t)\right]\\ &= \underbrace{\frac{\dot\alpha_t}{2\sqrt{\sigma+nv_t}(\sigma+v_t)^2}}_{\leq 0}\underbrace{\left[2(1-2\alpha_t)(\sigma+nv_t)(\sigma+v_t)-\alpha_t v_t n(\sigma+v_t)+2\alpha_t v_t(\sigma+nv_t)\right]}_{Q(v_t)}.\end{aligned}$$

Let's study the polynmial Q over time:

$$\begin{aligned}Q(v_t) &= 2(2v_t-1)(\sigma+nv_t)(\sigma+v_t) - v_t(1-v_t)n(\sigma+v_t) + 2v_t(1-v_t)(\sigma+nv_t)\\ &= 3nv_t^3 + v_t^2(5n\sigma+2\sigma-n) + v_t(4\sigma^2-3n\sigma) - 2\sigma^2,\\ Q(0) &= -2\sigma^2 \leq 0 \quad \text{and} \quad Q(1) = 2n+2n\sigma+2\sigma^2+2\sigma \geq 0.\end{aligned}$$

As Q is a polynomial in v_t of order 3 it has at most 3 roots. Moreover, with regard to its values at the extreme points of the interval $[0,1]$, Q has exactly 1 or 3 roots in $[0,1]$. Let q_3 be the largest root in $[0,1]$ of Q. Q is necessarily positive from this root, otherwise, there would exist a largest root in $[0,1]$. Let $t_1 \in [0,1]$ such that $v_{t_1} = q_3$. Then, Q is positive on $[t_1,1]$, i.e. the function g is decreasing on $[t_1,1]$ and $g'(t) \neq 0$ on $(t_1,1)$. Thus, $t \mapsto -\frac{1}{2}g(t)$ increases on $[t_1,1]$. So for t sufficiently close to 1, we have $t \geq t_1$ and the function $t \mapsto \sqrt{\sigma(\Sigma_t^L)}\left(\frac{1}{\sqrt{1+\sigma(\Sigma_t^L)\Omega_t}}-1\right)$ locally behaves like $t \mapsto -\frac{1}{2}g(t)$ thus increases locally around t. As this reasoning is valid for any point in a neighborhood of 1, denoted $[t_2,1]$, we can deduce that the latter function follows an increasing trend on $[t_2,1]$. Therefore, the blue term is **negative** on $[t_2,1]$. This means that there exists a time point $\max(t_0,t_2) \leq 1$ such that all terms in the sum of equation $(*)$ are non negative. Therefore, **the Bures distance for Geffner's covariance matrices $d_B(\Sigma_t^G, \Sigma_{t+1}^G)$ is always greater than the Bures distance for Linhart's covariance matrices $d_B(\Sigma_t^L, \Sigma_{t+1}^L)$ as $t \to 1$.**

Behavior when $t \to 0$. Let's look at the function $\boxed{h(t) = \frac{\sigma(\Sigma_t^L)}{1+\sigma(\Sigma_t^L)\Omega_t}}$ on $[0,1]$.

$$
\begin{aligned}
h(t) &= \frac{\frac{\sigma+v_t n}{\sigma+n}}{1+\frac{n(n-1)\alpha_t v_t}{(\sigma+v_t)(\sigma+nv_t)}\frac{\sigma+v_t n}{\sigma+n}} \\
&= \frac{(\sigma+v_t n)(\sigma+v_t)}{(\sigma+n)(\sigma+v_t)+n(n-1)\alpha_t v_t} \\
&= \frac{\sigma^2+v_t\sigma+v_t n\sigma+nv_t^2}{\sigma^2+\sigma v_t+n\sigma+nv_t+n(n-1)\alpha_t v_t}, \\
\frac{\partial}{\partial t}h(t) &= \underbrace{\frac{\dot{v_t}}{\cdots}}_{\geq 0}(\sigma+n\sigma+2nv_t)(\sigma^2+\sigma v_t+n\sigma+nv_t+n(n-1)\alpha_t v_t) \\
&\quad - \underbrace{\frac{\dot{v_t}}{\cdots}}_{\geq 0}(\sigma^2+v_t\sigma+v_t n\sigma+nv_t^2)(\sigma+n-n(n-1)v_t+n(n-1)\alpha_t) \\
&= \underbrace{\frac{\dot{v_t}}{\cdots}}_{\geq 0}(v_t^2 n^3(1+\sigma)+2v_t\sigma n^2(1+\sigma)+n\sigma^2(1+\sigma)) \\
&\geq 0 \quad \forall t \in [0,1] \text{ since } \sigma > 0, n \geq 1, v_t \in [0,1].
\end{aligned}
$$

We voluntarily omit the denominator in the previous computations as it is always positive and does not change the sign of the derivative. The function h is positive and increases over the interval $[0,1]$, as does $\sqrt{h(t)}$. As the function $\sigma(\Sigma_t^L)$ also increases over $[0,1]$, the red term is always negative for any value σ_i and Ω_i of the sum. Let's now look at the blue term when $t \to 0$. We know that the function $f(t) = \sigma(\Sigma_t^L)\Omega_t$ increases on $[0, t_0]$ (see previous analysis when $t \to 1$). Thus, the function $t \mapsto \frac{1}{\sqrt{1+\sigma(\Sigma_t^L)\Omega_t}} - 1$ is decreasing and negative on $[0, t_0]$. We can then derive on $[0, t_0]$:

$$
\begin{aligned}
\frac{1}{\sqrt{1+\sigma(\Sigma_{t+1}^L)\Omega_{t+1}}} - 1 &\leq \frac{1}{\sqrt{1+\sigma(\Sigma_t^L)\Omega_t}} - 1 \leq 0, \\
\sqrt{\sigma(\Sigma_{t+1}^L)}\left(\frac{1}{\sqrt{1+\sigma(\Sigma_{t+1}^L)\Omega_{t+1}}} - 1\right) &\leq \sqrt{\sigma(\Sigma_{t+1}^L)}\left(\frac{1}{\sqrt{1+\sigma(\Sigma_t^L)\Omega_t}} - 1\right) \\
&\quad \text{since } \sigma(\Sigma_{t+1}^L) \geq 0 \; \forall t \\
&\leq \sqrt{\sigma(\Sigma_t^L)}\left(\frac{1}{\sqrt{1+\sigma(\Sigma_t^L)\Omega_t}} - 1\right) \\
&\quad \text{since } \sigma(\Sigma_t^L) \text{ increases on } [0,1].
\end{aligned}
$$

Then the blue term is **positive** in the sum on $[0, t_0]$ whatever the value of σ_i and Ω_i. This means that all terms in the sum are negative for $t \in [0, t_0]$, i.e. **as $t \to 0$, the Bures distance between Geffner's covariance matrices $d_B(\Sigma_t^G, \Sigma_{t+1}^G)$ is lower than the same distance between Linhart's matrices** $d_B(\Sigma_t^L, \Sigma_{t+1}^L)$.

A.10.2 Comparison of mean differences

Lemma 8 (Comparison of mean differences) *Let μ_t^L (resp. μ_t^G) denote the mean of the intermediate density defined by Linhart (resp. Geffner) for a given time t. Then, we have*

$$
\begin{aligned}
&\textit{As } t \to 0, \quad \|\mu_t^G - \mu_{t+1}^G\|_2^2 \leq \|\mu_t^L - \mu_{t+1}^L\|_2^2. \\
&\textit{As } t \to 1, \quad \|\mu_t^G - \mu_{t+1}^G\|_2^2 \geq \|\mu_t^L - \mu_{t+1}^L\|_2^2.
\end{aligned}
$$

Let $A_t = \sqrt{\alpha_t}(n\Sigma^{-1} + I_d)^{-1}$ and $B_t = \sqrt{\alpha_t}\Sigma_t^G\Sigma_t^{-1}(\Sigma^{-1} + I_d)^{-1}$ be respectively the matrices appearing in the computation of the mean differences derived in Section 4.1 and Appendix A.6.

$$\begin{aligned}\|\mu_t^L - \mu_{t+1}^L\|^2 - \|\mu_t^G - \mu_{t+1}^G\|^2 &= \|(A_t - A_{t+1})\underbrace{\Sigma^{-1}\sum_{i=1}^n x_i}_{:=y}\|^2 - \|(B_t - B_{t+1})\underbrace{\Sigma^{-1}\sum_{i=1}^n x_i}_{:=y}\|^2 \\ &= y^T(A_t - A_{t+1})^T(A_t - A_{t+1})y - y^T(B_t - B_{t+1})^T(B_t - B_{t+1})y \\ &= y^T\left((A_t - A_{t+1})^T(A_t - A_{t+1}) - (B_t - B_{t+1})^T(B_t - B_{t+1})\right)y.\end{aligned}$$

We will show that the matrix $(A_t - A_{t+1})^T(A_t - A_{t+1})$ is greater (in the sense of eigenvalues comparison) than $(B_t - B_{t+1})^T(B_t - B_{t+1})$ when $t \to 0$ and lower when $t \to 1$. Recall that:

$$\sigma(A_t) = \sqrt{\alpha_t}\frac{\sigma}{\sigma + n} \quad \text{and} \quad \sigma(B_t) = \sqrt{\alpha_t}\frac{\sigma}{\sigma + n + v_t(1-n)}.$$

These eigenvalues can be easily derived as all matrices commute and can be diagonalized in the same eigen basis. If we take the derivative with respect to time, we get:

$$\begin{aligned}\frac{\partial}{\partial t}\sigma(A_t) &= \frac{\dot{\alpha_t}}{2\sqrt{\alpha_t}}\frac{\sigma}{\sigma + n} \leq 0 \; \forall t \quad \text{since } \dot{\alpha}_t \leq 0 \; \forall t \\ &\Leftrightarrow A_t - A_{t+1} \succeq 0 \; \forall t \text{ (decreasing trend in time)}, \\ \frac{\partial}{\partial t}\sigma(B_t) &= \underbrace{\frac{\dot{\alpha_t}\sigma}{2\sqrt{\alpha_t}(\sigma + n + v_t(1-n))^2}}_{\leq 0 \; \forall t}(\sigma + 2 - n + v_t(n-1)) \\ &\leq 0 \text{ if } v_t \geq \frac{n-2-\sigma}{n-1} \\ &\Leftrightarrow \begin{cases} B_t - B_{t+1} \preceq 0 \text{ if } 0 \leq v_t \leq \max(0, \frac{n-2-\sigma}{n-1}) \\ B_t - B_{t+1} \succeq 0 \text{ if } 1 \geq v_t \geq \frac{n-2-\sigma}{n-1} \end{cases}.\end{aligned}$$

Note that $\frac{n-2-\sigma}{n-1} \leq 1$ for all $n \geq 1$ and $\sigma > 0$. In the following, we will study the sign of $\sigma(A_t - A_{t+1}) - \sigma(B_t - B_{t+1})$ over time to determine the behavior of the mean terms over time.

Behavior when $t \to 1$. In this case, $v_t \to 1$, thus $A_t - A_{t+1} \succeq 0$ and $B_t - B_{t+1} \succeq 0$. As A_t and B_t commute, we have

$$\begin{aligned}
\sigma(A_t - B_t) &= \sqrt{\alpha_t}\frac{\sigma}{\sigma+n} - \sqrt{\alpha_t}\frac{\alpha_t\sigma + v_t(\sigma+1)}{v_t(\sigma+1) + (1-n)\alpha_t\sigma + n\alpha_t(\sigma+1)}\frac{\sigma+1}{\alpha_t\sigma + v_t(\sigma+1)}\frac{\sigma}{\sigma+1} \\
&= \sqrt{\alpha_t}\frac{\sigma}{\sigma+n} - \sqrt{\alpha_t}\frac{\sigma+v_t}{\sigma+v_t+n\alpha_t}\frac{\sigma}{\sigma+v_t} \\
&= \sqrt{\alpha_t}\frac{\sigma}{\sigma+n} - \sqrt{\alpha_t}\frac{\sigma}{\sigma+v_t+n\alpha_t} \\
&= \frac{\sigma}{\sigma+n}\underbrace{\sqrt{\alpha_t}\left(1 - \frac{\sigma+n}{\sigma+v_t+n\alpha_t}\right)}_{:=e(t)}, \\
\frac{\partial}{\partial t}e(t) &= \frac{\dot{\alpha_t}}{2\sqrt{\alpha_t}}\left(1 - \frac{\sigma+n}{\sigma+v_t+n\alpha_t}\right) + \sqrt{\alpha_t}(\sigma+n)\frac{\dot{v_t}+n\dot{\alpha_t}}{(\sigma+v_t+n\alpha_t)^2} \\
&= \underbrace{\frac{\dot{\alpha_t}}{2\sqrt{\alpha_t}(\sigma+v_t+n\alpha_t)^2}}_{\leq 0}\left(\underbrace{(\sigma+v_t+n\alpha_t)^2 - (\sigma+n)(\sigma+v_t+n\alpha_t) + 2\alpha_t(\sigma+n)(-1+n)}_{S(v_t)}\right), \\
S(v_t) &= v_t^2(n-1)^2 + v_t(3n - 3n^2 + 3\sigma - 3n\sigma) + 2n^2 - 2n + 2n\sigma - 2\sigma \\
&= v_t^2(n-1)^2 - 3v_t(n-1)(\sigma+n) + 2(n-1)(\sigma+n) \\
&= (n-1)(v_t^2(n-1) - 3v_t(\sigma+n) + 2(\sigma+n)), \\
S(0) &= 2(n-1)(\sigma+n) \geq 0, \\
S(1) &= (1+\sigma)(1-n) \leq 0.
\end{aligned}$$

S is a polynomial of degree 2 in v_t. Given the values reached by S on the bounds of $[0,1]$, it has an unique root in this interval, denoted r_0. Let t_3 such that $v_{t_3} = r_0$. Then, S is positive on $[0, r_0]$ and negative on $[r_0, 1]$. Therefore, the function e decreases on $[0, t_3]$, then increases on $[t_3, 1]$. This is equivalent to say that :

$$\begin{cases} \sigma(A_t - B_t) \geq \sigma(A_{t+1} - B_{t+1}) & \text{if } t \in [0, t_3] \\ \sigma(A_t - B_t) \leq \sigma(A_{t+1} - B_{t+1}) & \text{if } t \in [t_3, 1] \end{cases}$$

Since all matrices commute, this translates as per

$$\Leftrightarrow \begin{cases} \sigma(A_t - A_{t+1}) \geq \sigma(B_t - B_{t+1}) & \text{if } t \in [0, t_3] \\ \sigma(A_t - A_{t+1}) \leq \sigma(B_t - B_{t+1}) & \text{if } t \in [t_3, 1] \end{cases}.$$

Thus, when $t \in [\max(t_3, \frac{n-2-\sigma}{n-1}), 1]$, we have $0 \leq \sigma(A_t - A_{t+1}) \leq \sigma(B_t - B_{t+1})$, therefore

$$\begin{aligned}
0 \leq \underbrace{\sigma((A_t - A_{t+1})^T(A_t - A_{t+1}))}_{=(\sigma(A_t - A_{t+1}))^2} &\leq \underbrace{\sigma((B_t - B_{t+1})^T(B_t - B_{t+1}))}_{=(\sigma(B_t - B_{t+1}))^2} \\
\Leftrightarrow\ 0 \preceq (A_t - A_{t+1})^T(A_t - A_{t+1}) &\preceq (B_t - B_{t+1})^T(B_t - B_{t+1}) \\
\Leftrightarrow\ \|\mu_t^L - \mu_{t+1}^L\|^2 &\leq \|\mu_t^G - \mu_{t+1}^G\|^2.
\end{aligned}$$

In the neighborhood of $t = 1$, Linhart's bridging densities always have a lower mean difference $\|\mu_t^L - \mu_{t+1}^L\|_2^2$ than Geffner's counterparts $\|\mu_t^G - \mu_{t+1}^G\|_2^2$.

Behavior when $t \to 0$. We consider the time steps t such that $0 \leq v_t \leq \max(0, \frac{n-2-\sigma}{n-1})$. Let's denote this time interval $[0, \tilde{t}_4]$. In this case, $A_t - A_{t+1} \succeq 0$ and $B_t - B_{t+1} \preceq 0$. Note that if $\max(0, \frac{n-2-\sigma}{n-1}) = 0$, then we turn back to the previous paragraph and study the case when $t \in [0, t_3]$. We have

$$|\sigma(A_t - A_{t+1})| - |\sigma(B_t - B_{t+1})| = \sigma(A_t - A_{t+1}) + \sigma(B_t - B_{t+1}).$$

Let's study the function of time $t \mapsto \sigma(A_t + B_t)$ and show that it decreases in the neighborhood of $t = 0$.

$$\begin{aligned}
\sigma(A_t + B_t) &= \sqrt{\alpha_t}\frac{\sigma}{\sigma + n} + \sqrt{\alpha_t}\frac{\sigma}{\sigma + v_t + n\alpha_t} \\
&= \frac{\sigma}{\sigma + n}\underbrace{\sqrt{\alpha_t}\left(1 + \frac{\sigma + n}{\sigma + v_t + n\alpha_t}\right)}_{=d(t)} \geq 0 \; \forall t, \\
\frac{\partial}{\partial t}d(t) &= \frac{\dot{\alpha_t}}{2\sqrt{\alpha_t}}\left(1 + \frac{\sigma + n}{\sigma + v_t + n\alpha_t}\right) - \sqrt{\alpha_t}(\sigma + n)\frac{\dot{v_t} + n\dot{\alpha_t}}{(\sigma + v_t + n\alpha_t)^2} \\
&= \frac{\dot{\alpha_t}}{2\sqrt{\alpha_t}}\left(1 + \frac{\sigma + n}{\sigma + v_t + n\alpha_t} - 2\alpha_t(\sigma + n)\frac{-1 + n}{(\sigma + v_t + n\alpha_t)^2}\right) \\
&= \frac{\dot{\alpha_t}}{2\sqrt{\alpha_t}(\sigma + v_t + n\alpha_t)^2}\left((\sigma + v_t + n\alpha_t)^2 + (\sigma + n)(\sigma + v_t + n\alpha_t) - 2\alpha_t(n-1)(\sigma + n)\right) \\
&= \underbrace{\frac{\dot{\alpha_t}}{2\sqrt{\alpha_t}(\sigma + v_t + n\alpha_t)^2}}_{\leq 0}(\underbrace{v_t^2(n-1)^2 + v_t(\sigma + n)(1 - n) + 2(\sigma + n)(\sigma + 1)}_{T(v_t)}), \\
T(0) &= 2(\sigma + n)(\sigma + 1) \geq 0, \\
T(1) &= 1 + 3\sigma + n + n\sigma + 2\sigma^2 \geq 0, \\
\Delta_T &= (n-1)^2(\sigma + n)(n - 7\sigma - 8).
\end{aligned}$$

T is a polynomial of degree 2 in v_t. If $n - 7\sigma - 8 \leq 0$, then T has at most one root and stays positive on $[0, 1]$. Otherwise, there exists 2 roots r_1, r_2; either both belong to the interval $[0, 1]$, or neither does. In the second case, T remains positive over $[0, 1]$. In the first case, T is positive for $0 \leq v_t \leq r_1$ and $1 \geq v_t \geq r_2$. In both cases, if $r_1 \geq 0$, T is positive on $[0, \min(r_1, 1)]$. So there exists a time steps t_4 such that $v_{t_4} = \min(r_1, 1)$ and the function d is decreasing on $[0, t_4]$. This is equivalent to say that

$$\begin{aligned}
&\sigma(A_t + B_t) \geq \sigma(A_{t+1} + B_{t+1}) \geq 0 \text{ if } t \in [0, t_4] \\
\Leftrightarrow\; &\sigma(A_t - A_{t+1}) \geq \sigma(B_{t+1} - B_t) \geq 0 \quad \text{since all matrices commute} \\
\Leftrightarrow\; &|\sigma(A_t - A_{t+1})| \geq |\sigma(B_t - B_{t+1})| \geq 0 \text{ if } t \in [0, \min(t_4, \tilde{t}_4)] \\
\Leftrightarrow\; &\underbrace{\sigma((A_t - A_{t+1})^T(A_t - A_{t+1}))}_{=|\sigma(A_t - A_{t+1})|^2} \geq \underbrace{\sigma((B_t - B_{t+1})^T(B_t - B_{t+1}))}_{=|\sigma(B_t - B_{t+1})|^2} \geq 0 \text{ if } t \in [0, \min(t_4, \tilde{t}_4)] \\
\Leftrightarrow\; &\|\mu_t^L - \mu_{t+1}^L\|^2 \geq \|\mu_t^G - \mu_{t+1}^G\|^2.
\end{aligned}$$

In the neighborhood of $t = 0$, Linhart's bridging densities always have a greater mean difference $\|\mu_t^L - \mu_{t+1}^L\|_2^2$ than Geffner's counterparts $\|\mu_t^G - \mu_{t+1}^G\|_2^2$.

A.10.3 Final derivation

We first state again the proposition before deriving its proof. The proof uses the same notation convention as in the previous subsections for readability concerns: π_{t+1} will denote $\pi_{t_{p+1}}$ and π_t stands for π_{t_p}, for one given $p = 0, \dots, T$.

Proposition 4 (Comparison of Wasserstein distances between intermediate densities) *Let $\pi_{t_p}^L$ (resp. $\pi_{t_p}^G$) denote the (Gaussian) intermediate density defined by Linhart (resp. by Geffner) at level t_p. Then, we have :*

$$\begin{aligned}
&\textit{As } t_p \to 0, \quad \mathcal{W}_2^2(\pi_{t_p}^G, \pi_{t_{p+1}}^G) \leq \mathcal{W}_2^2(\pi_{t_p}^L, \pi_{t_{p+1}}^L). \\
&\textit{As } t_p \to 1, \quad \mathcal{W}_2^2(\pi_{t_p}^G, \pi_{t_{p+1}}^G) \geq \mathcal{W}_2^2(\pi_{t_p}^L, \pi_{t_{p+1}}^L).
\end{aligned}$$

Therefore, the number of Langevin steps $k_{t_p}^G$ (chosen as in Section 3) used with Geffner's composite score in annealed Langevin is greater than its counterpart $k_{t_p}^L$ used with Linhart's score as $t_p \to 1$, and conversely as $t_p \to 0$.

Proof:
Behavior when $t \to 1$:

From Lemma 7, there exists a time step $\max(t_0, t_2) \leq 1$ such that $d_B(\Sigma_t^G, \Sigma_{t+1}^G) \geq d_B(\Sigma_t^L, \Sigma_{t+1}^L)$ on $[\max(t_0, t_2), 1]$.
From Lemma 8, there exists a time step $t_3 \leq 1$ such that $\|\mu_t^G - \mu_{t+1}^G\|_2^2 \geq \|\mu_t^L - \mu_{t+1}^L\|_2^2$ on $[t_3, 1]$.
Therefore, as we work with Gaussian distributions, for $t \in [\max(t_0, t_2, t_3), 1]$, $\mathcal{W}_2^2(\pi_t^G, \pi_{t+1}^G) \geq \mathcal{W}_2^2(\pi_t^L, \pi_{t+1}^L)$.

Behavior when $t \to 0$:

From Lemma 7, there exists a time step $0 \leq t_0 \leq 1$ such that $d_B(\Sigma_t^G, \Sigma_{t+1}^G) \leq d_B(\Sigma_t^L, \Sigma_{t+1}^L)$ on $[0, t_0]$.
From Lemma 8, there exists a time step $0 \leq \min(\tilde{t}_4, t_4) \leq 1$ such that $\|\mu_t^G - \mu_{t+1}^G\|_2^2 \leq \|\mu_t^L - \mu_{t+1}^L\|_2^2$ on $[0, \min(\tilde{t}_4, t_4)]$.
Therefore, as we work with Gaussian distributions, for $t \in [0, \min(t_0, \tilde{t}_4, t_4)]$, $\mathcal{W}_2^2(\pi_t^G, \pi_{t+1}^G) \leq \mathcal{W}_2^2(\pi_t^L, \pi_{t+1}^L)$.

When $t \to 0$ and $t \to 1$, the choice of the number of steps prescribed in Section 3 mainly depends on the intermediate Wasserstein distances, as the log-concavity constants m_t^G and m_t^L tend to have similar values in these regimes. Suppose that we choose the same step size for both algorithms, denoted h_t. Let's now show that the number of Langevin steps follows the growth trend of the Wasserstein distance between bridging densities. When $t \to 0$, then $0 \leq \mathcal{W}_2(\pi_t^G, \pi_{t+1}^G) \leq \mathcal{W}_2(\pi_t^L, \pi_{t+1}^L)$ and we have

$$
\begin{aligned}
0 \leq \frac{1}{2}\frac{\gamma}{\gamma + \mathcal{W}_2(\pi_{t+1}^L, \pi_t^L)} &\leq \frac{1}{2}\frac{\gamma}{\gamma + \mathcal{W}_2(\pi_{t+1}^G, \pi_t^G)} \\
\Leftrightarrow \log\left(\frac{1}{2}\frac{\gamma}{\gamma + \mathcal{W}_2(\pi_{t+1}^L, \pi_t^L)}\right) &\leq \log\left(\frac{1}{2}\frac{\gamma}{\gamma + \mathcal{W}_2(\pi_{t+1}^G, \pi_t^G)}\right) \leq 0 \\
\Leftrightarrow \log\left(\frac{1}{2}\frac{\gamma}{\gamma + \mathcal{W}_2(\pi_{t+1}^L, \pi_t^L)}\right) \underbrace{\frac{1}{\log(1 - m_t^G h_t)}}_{\leq 0} &\geq \log\left(\frac{1}{2}\frac{\gamma}{\gamma + \mathcal{W}_2(\pi_{t+1}^G, \pi_t^G)}\right) \underbrace{\frac{1}{\log(1 - m_t^G h_t)}}_{\leq 0} \geq 0.
\end{aligned}
$$

Note that $m_t^G \geq m_t^L$ for all time t. Thus,

$$
\begin{aligned}
0 \leq 1 - m_t^G h_t \leq 1 - m_t^L h_t \leq 1, \\
0 \geq \frac{1}{\log(1 - m_t^G h_t)} \geq \frac{1}{\log(1 - m_t^L h_t)}.
\end{aligned}
$$

Thus, we get

$$
\begin{aligned}
&\underbrace{\log\left(\frac{1}{2}\frac{\gamma}{\gamma + \mathcal{W}_2(\pi_{t+1}^L, \pi_t^L)}\right)}_{\leq 0} \frac{1}{\log(1 - m_t^L h_t)} \geq \underbrace{\log\left(\frac{1}{2}\frac{\gamma}{\gamma + \mathcal{W}_2(\pi_{t+1}^L, \pi_t^L)}\right)}_{\leq 0} \frac{1}{\log(1 - m_t^G h_t)} \geq 0 \\
\Rightarrow &\log\left(\frac{1}{2}\frac{\gamma}{\gamma + \mathcal{W}_2(\pi_{t+1}^L, \pi_t^L)}\right) \frac{1}{\log(1 - m_t^L h_t)} \geq \log\left(\frac{1}{2}\frac{\gamma}{\gamma + \mathcal{W}_2(\pi_{t+1}^G, \pi_t^G)}\right) \frac{1}{\log(1 - m_t^G h_t)} \\
&\Rightarrow k_t^L \geq k_t^G.
\end{aligned}
$$

Then, when $t \to 1$, both log-concavity constants tend to 1, so consider $m_t = \min(m_t^G, m_t^L)$ so that the choice of the number of steps only depends on the intermediate Wasserstein distances. As $\mathcal{W}_2(\pi_t^G, \pi_{t+1}^G) \geq \mathcal{W}_2(\pi_t^L, \pi_{t+1}^L)$ when $t \to 1$, then we have :

$$
\begin{aligned}
&\log\left(\frac{1}{2}\frac{\gamma}{\gamma + \mathcal{W}_2(\pi_{t+1}^L, \pi_t^L)}\right) \underbrace{\frac{1}{\log(1 - m_t h_t)}}_{\leq 0} \leq \log\left(\frac{1}{2}\frac{\gamma}{\gamma + \mathcal{W}_2(\pi_{t+1}^G, \pi_t^G)}\right) \underbrace{\frac{1}{\log(1 - m_t h_t)}}_{\leq 0} \\
&\Rightarrow k_t^G \geq k_t^L.
\end{aligned}
$$

A.11 Proof of Lemma 3

Let s_ϕ (resp. s_λ) be an amortized neural-based approximation of the score of the individual posteriors $p(\theta \mid x_i)$ (resp. the prior $\lambda(\theta)$). Assume that $\mathrm{E}_\theta\left[\|\nabla_\theta \log \lambda_t(\theta) - s_\lambda(\theta,t)\|^2\right] \leq \epsilon^2_{\mathrm{DSM},\lambda}$ and $\mathrm{E}_\theta\left[\|\nabla_\theta \log p_t(\theta \mid x_i) - s_\phi(\theta, x_i, t)\|^2\right] \leq \epsilon^2_{\mathrm{DSM}}$ for all $i = 1, \ldots, n$ and a given $t \in [0,1]$. We denote by $\tilde{s}^G_t$, the composite score defined in Equation 1, where individual posterior scores are inexact. Then, we have:

$$
\begin{aligned}
&\mathrm{E}_\theta\Big(\big\|\nabla_\theta \log \pi_t(\theta \mid x_{1:n}) - \tilde{s}^G_t(\theta \mid x_{1:n})\big\|^2\Big) \\
&\quad= \mathrm{E}_\theta\left(\left\|(1-n)\left(\nabla_\theta \log \lambda_t(\theta) - s_\lambda(\theta,t)\right) + \sum_{i=1}^n \left(\nabla_\theta \log p_t(\theta \mid x_i) - s_\phi(\theta, x_i, t)\right)\right\|^2\right) \\
&\quad= \mathrm{E}_\theta\big\|(1-n)(\nabla_\theta \log \lambda_t(\theta) - s_\lambda(\theta,t))\big\|^2 + \sum_{i=1}^n \mathrm{E}_\theta\big\|\nabla_\theta \log p_t(\theta \mid x_i) - s_\phi(\theta, x_i, t)\big\|^2 \\
&\qquad+ 2\mathrm{E}_\theta\left\langle (1-n)(\nabla_\theta \log \lambda_t(\theta) - s_\lambda(\theta,t)), \sum_{i=1}^n \left(\nabla_\theta \log p_t(\theta \mid x_i) - s_\phi(\theta, x_i, t)\right)\right\rangle \\
&\qquad+ \mathrm{E}_\theta \sum_{i\neq j} \left\langle \nabla_\theta \log p_t(\theta \mid x_i) - s_\phi(\theta, x_i, t), \nabla_\theta \log p_t(\theta \mid x_j) - s_\phi(\theta, x_j, t)\right\rangle \\
&\quad\leq (1-n)^2 \mathrm{E}_\theta\big\|\nabla_\theta \log \lambda_t(\theta) - s_\lambda(\theta,t)\big\|^2 + \sum_{i=1}^n \mathrm{E}_\theta\big\|\nabla_\theta \log p_t(\theta \mid x_i) - s_\phi(\theta, x_i, t)\big\|^2 \\
&\qquad+ 2(n-1)\sum_{i=1}^n \mathrm{E}_\theta\big\|\nabla_\theta \log \lambda_t(\theta) - s_\lambda(\theta,t)\big\|\big\|\nabla_\theta \log p_t(\theta \mid x_i) - s_\phi(\theta, x_i, t)\big\| \\
&\qquad+ \sum_{i\neq j} \mathrm{E}_\theta\big\|\nabla_\theta \log p_t(\theta \mid x_i) - s_\phi(\theta, x_i, t)\big\|\big\|\nabla_\theta \log p_t(\theta \mid x_j) - s_\phi(\theta, x_j, t)\big\| \\
&\qquad\text{by Cauchy–Schwarz and } n \geq 1 \\
&\quad\leq (1-n)^2 \epsilon^2_{\mathrm{DSM},\lambda} + \sum_{i=1}^n \epsilon^2_{\mathrm{DSM},i} \\
&\qquad+ 2(n-1)\sum_{i=1}^n \sqrt{\mathrm{E}_\theta\big\|\nabla_\theta \log \lambda_t(\theta) - s_\lambda(\theta,t)\big\|^2}\sqrt{\mathrm{E}_\theta\big\|\nabla_\theta \log p_t(\theta \mid x_i) - s_\phi(\theta, x_i, t)\big\|^2} \\
&\qquad+ \sum_{i\neq j} \sqrt{\mathrm{E}_\theta\big\|\nabla_\theta \log p_t(\theta \mid x_i) - s_\phi(\theta, x_i, t)\big\|^2}\sqrt{\mathrm{E}_\theta\big\|\nabla_\theta \log p_t(\theta \mid x_j) - s_\phi(\theta, x_j, t)\big\|^2} \\
&\qquad\text{by Cauchy–Schwarz} \\
&\quad\leq (1-n)^2 \epsilon^2_{\mathrm{DSM},\lambda} + \sum_{i=1}^n \epsilon^2_{\mathrm{DSM},i} + 2(n-1)\epsilon_{\mathrm{DSM},\lambda}\sum_{i=1}^n \epsilon_{\mathrm{DSM},i} + \sum_{i\neq j} \epsilon_{\mathrm{DSM},i}\epsilon_{\mathrm{DSM},j} \\
&\quad= \Big((n-1)\epsilon_{\mathrm{DSM},\lambda} + \sum_{i=1}^n \epsilon_{\mathrm{DSM},i}\Big)^2 \\
&\quad\leq \Big((n-1)\epsilon_{\mathrm{DSM},\lambda} + n\epsilon_{\mathrm{DSM}}\Big)^2 \quad \text{if } \epsilon_{\mathrm{DSM},i} \leq \epsilon_{\mathrm{DSM}}\ \forall i.
\end{aligned}
$$

A.12 Analytical derivation of prior and individual posterior scores

Multivariate Gaussian. The analytical formulas of individual Gaussian posteriors diffused along the VP process are given in Appendix A.6. This allows us to derive analytical scores of diffused individual posteriors:

$$
\nabla_\theta \log p_t(\theta \mid x_i) = -(\alpha_t \Sigma_{\text{post}} + v_t I_d)^{-1}(\theta - \sqrt{\alpha_t}\mu_{\text{post}}(x_i)) \quad \forall t \in [0,1],
$$

where Σ_{post} and $\mu_{\text{post}}(x_i)$ are explicitly defined in Appendix A.6. For Log-Normal priors, we first apply a log-transform on the samples before computing the score: therefore, it amounts to deriving the score of a Gaussian distribution along the VP process.

GMM prior. The prior is a mixture of Gaussians of the form $\lambda(\theta) = \frac{1}{2}\mathcal{N}(\theta; \mu_1, \sigma_1^2 I_d) + \frac{1}{2}\mathcal{N}(\theta; \mu_2, \sigma_2^2 I_d)$ and the likelihood is $p(x \mid \theta) = \mathcal{N}(x; \theta, \Sigma)$. Then, the score of the diffused prior along the VP process is equal to:

$$\begin{aligned}
\nabla_\theta \log \lambda_t(\theta) &= -\omega_{t,1}(\theta)\Sigma_{t,1}^{-1}(\theta - \sqrt{\alpha_t}\mu_1) - \omega_{t,2}(\theta)\Sigma_{t,2}^{-1}(\theta - \sqrt{\alpha_t}\mu_2), \\
\text{where } \Sigma_{t,j} &= (\alpha_t \sigma_j^2 + v_t) I_d \quad \forall j = 1, 2 \\
\text{and } \tilde{\omega}_{t,j}(\theta) &= \frac{1}{2}\mathcal{N}(\theta; \sqrt{\alpha_t}\mu_j, \Sigma_{t,j}), \quad \omega_{t,j}(\theta) = \tilde{\omega}_{t,j}(\theta) / \sum_{k=1}^{2} \tilde{\omega}_{t,k}(\theta).
\end{aligned}$$

The corresponding posterior is a mixture of 2 Gaussians:

$$p(\theta \mid x) = \omega_{\text{post},1}(x)\mathcal{N}(\mu_{\text{post},1}, \Sigma_{\text{post},1}) + \omega_{\text{post},2}(x)\mathcal{N}(\mu_{\text{post},2}, \Sigma_{\text{post},2}),$$

where

$$\begin{aligned}
\Sigma_{\text{post},j} &= (\Sigma^{-1} + \frac{1}{\sigma_j^2} I_d)^{-1}, \\
\mu_{\text{post},j} &= \Sigma_{\text{post},j}(\mu_j / \sigma_j^2 + \Sigma^{-1} x), \\
\tilde{\omega}_{\text{post},j}(x) &= \frac{1}{2}\mathcal{N}(x; \mu_j, \sigma_j^2 I_d + \Sigma), \\
\omega_{\text{post},j}(x) &= \tilde{\omega}_{\text{post},j}(x) / \sum_{k=1}^{2} \tilde{\omega}_{\text{post},k}(x) \quad \forall j = 1, 2.
\end{aligned}$$

We have the analytical formula of the score of the diffused posterior along the VP scheme:

$$\nabla_\theta \log p_t(\theta \mid x) = -\omega_{\text{post},t,1}(\theta)\Sigma_{\text{post},t,1}^{-1}(\theta - \sqrt{\alpha_t}\mu_{\text{post},1}) - \omega_{\text{post},t,2}(\theta)\Sigma_{\text{post},t,2}^{-1}(\theta - \sqrt{\alpha_t}\mu_{\text{post},2}),$$

where

$$\begin{aligned}
\Sigma_{\text{post},t,j} &= (\alpha_t \Sigma_{\text{post},j} + v_t I_d), \\
\tilde{\omega}_{\text{post},t,j}(\theta) &= \omega_{\text{post},j}(x)\mathcal{N}(\theta; \sqrt{\alpha_t}\mu_{\text{post},j}, \alpha_t \Sigma_{\text{post},j} + v_t I_d), \\
\omega_{\text{post},t,j}(\theta) &= \tilde{\omega}_{\text{post},t,j}(\theta) / \sum_{k=1}^{2} \omega_{\text{post},t,k}(\theta).
\end{aligned}$$

GMM likelihood. We consider a standard Gaussian prior $\lambda(\theta) = \mathcal{N}(0, I_d)$ and a mixture of two Gaussians as likelihood of the form $p(x \mid \theta) = \frac{1}{2}\mathcal{N}(x; \theta, 1/9\Sigma) + \frac{1}{2}\mathcal{N}(x; \theta, 2.25\Sigma)$ and Σ is a diagonal matrix with values increasing linearly between 0.6 and 1.4. Let $\Sigma_1 = 2.25\Sigma$ and $\Sigma_2 = 1/9\Sigma$. The individual posterior is a mixture of two Gaussians of the form:

$$\begin{aligned}
p(\theta \mid x) &= \omega_1(x)\mathcal{N}(\theta; \mu_1, \Sigma_{1,p}) + \omega_2(x)\mathcal{N}(\theta; \mu_2, \Sigma_{2,p}), \\
\text{where} \quad \Sigma_{j,p} &= (\Sigma_j^{-1} + I_d)^{-1}, \\
\mu_j &= \Sigma_{j,p}\Sigma_j^{-1} x, \\
\tilde{\omega}_j(x) &= \frac{1}{2}\mathcal{N}(x; 0, I_d + \Sigma_j), \\
\omega_j(x) &= \tilde{\omega}_j(x) / \sum_{k=1}^{2} \tilde{\omega}_k(x).
\end{aligned}$$

The score of the diffused individual posterior is as follows:

$$\begin{aligned}
&\nabla_\theta \log p_t(\theta \mid x) = -\omega_{t,1}(\theta)(\alpha_t \Sigma_{1,p} + v_t I_d)^{-1}(\theta - \sqrt{\alpha_t}\mu_1) - \omega_{t,2}(\theta)(\alpha_t \Sigma_{2,p} + v_t I_d)^{-1}(\theta - \sqrt{\alpha_t}\mu_2), \\
&\text{where} \quad \tilde{\omega}_{t,j}(\theta) = \omega_j(x)\mathcal{N}(\theta; \sqrt{\alpha_t}\mu_j, \alpha_t \Sigma_{j,p} + v_t I_d), \\
&\omega_{t,j}(\theta) = \tilde{\omega}_{t,j}(\theta) / \sum_{k=1}^{2} \tilde{\omega}_{t,k}(\theta).
\end{aligned}$$

B Empirical results and technical details

B.1 Empirical validation of the Wasserstein error - Gaussian setting

We provide here the final Wasserstein errors obtained with the prescribed hyperparameters (step sizes and number of steps per level) for the d-dimensional Gaussian case. Figure 4 presents an empirical validation of our decision rule for annealed Langevin hyperparameters: indeed, the final Wasserstein error obtained with our tuned annealed Langevin algorithm remains controlled by the chosen threshold γ, whatever the dimension d or the number of conditional observations n.

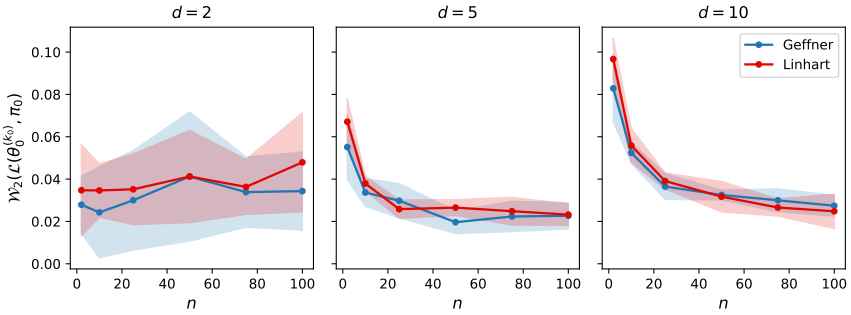

Figure 4: Mean final Wasserstein errors obtained in the Gaussian setting with annealed Langevin dynamics and prescribed hyperparameters from Section 3. Blue (resp. red) curves stand for Geffner's (resp. Linhart's) compositional score used in the sampling algorithm. We use $T = 10$ annealing levels, $\gamma = 0.5$ and $\omega = 0.5$ (resp. 0.8) for $d = 2, 5$ (resp. $d = 10$). The empirical Wasserstein error indeed remains below the fixed threshold γ whatever the dimension d and the number of conditional observations n. As theoretically expected, it is also similar for both choices of compositional scores. The error does not vary a lot whatever the dimension d. Mean and std are computed over 5 seeds.

B.2 Comparison of the number of steps - Gaussian setting

In the Gaussian setting, we test our results for several parameter dimensions $d = 2, 5, 10$. Figure 5 shows the evolution of the number of steps over time. We observe that there are two regimes separated by a switch point $\tilde{t}$: on $[0, \tilde{t}]$, $k_t^L \geq k_t^G$ but this regime is very short; on $[\tilde{t}, 1]$, $k_t^L \leq k_t^G$ and this regime covers most of the annealing levels.

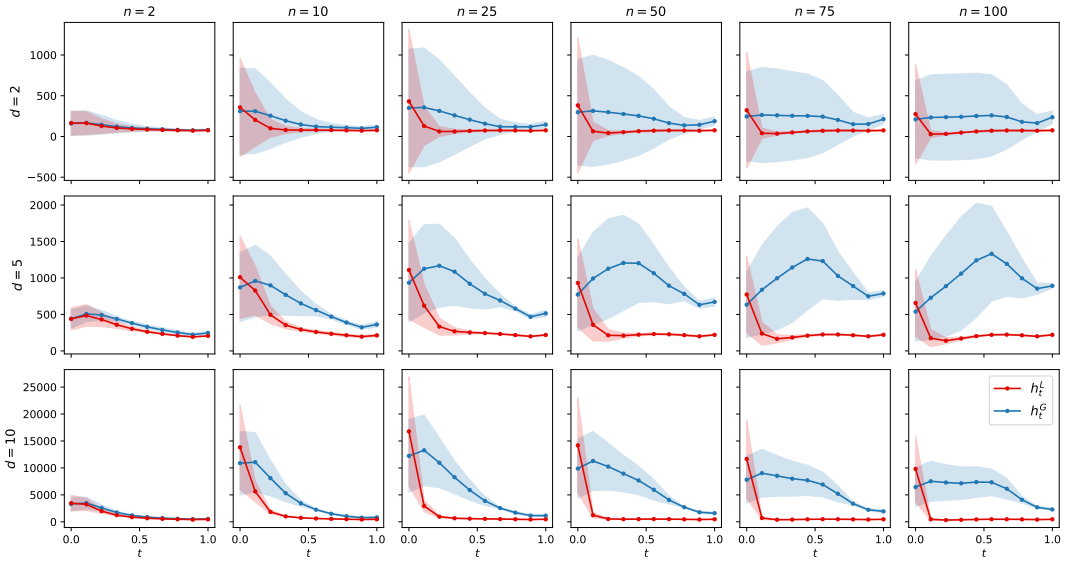

Figure 5: Evolution of k_t^G and k_t^L over time for the Gaussian case. The switch point $\tilde{t}$ (where blue and red curves cross) is unique and often appears from the second annealing level near $t = 0$, whatever the dimension of the space d or the number of conditional observations n. Mean and std are computed over 5 seeds, each model differs in its likelihood covariance matrix.

Figure 6 shows one example in dimension 2, where the switch point $\tilde{t}$ appears later than the second annealing level: in this case, $k_t^L \geq k_t^G$ for more annealing levels than in most cases. Note that this behavior becomes even more rare when the dimension increases.

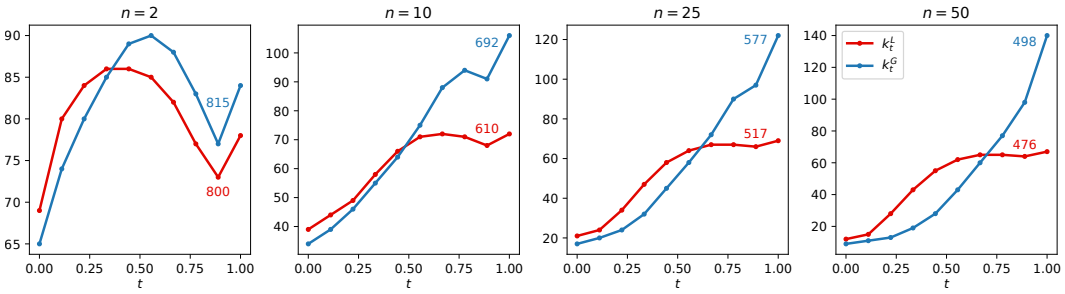

Figure 6: Evolution of k_t^G and k_t^L over time: Example of switch point appearing later when $d = 2$. The colored figures on the right of each subplot correspond to the total number of steps $\sum_{p=0}^{T} k_{t_p}$. In this specific case, blue and red curves cross later than the second annealing level. However, the complexity (in terms of total number of steps) remains lower when choosing Linhart's composite score than that of Geffner.

B.3 Additional figures - GMM Prior

For this specific task, the goal is to deviate from Gaussianity in a controlled manner. The prior is defined as a 2-dimensional mixture of two Gaussians: $\lambda(\theta) = \frac{1}{2}\mathcal{N}(\mu_1, \sigma_1^2 I_d) + \frac{1}{2}\mathcal{N}(\mu_2, \sigma_2^2 I_d)$. Instead of moving the means, we fix them to $\mu_1 = (0, 0)$ and $\mu_2 = (1, 1)$ and tune the couple of scales (σ_1, σ_2) to control non Gaussianity. We use the following scale values $(0.05, 0.05)$, $(0.05, 0.5)$ and $(0.5, 0.5)$. As the scales decrease, the individual posteriors become more bimodal. The likelihood is a Gaussian of the form $p(x \mid \theta) = \mathcal{N}(x; \theta, \Sigma)$, with Σ a positive definite matrix. Figure 7 shows the complexity and final Wasserstein error when the individual posterior is strongly bimodal (prior scales are chosen small as per $(0.05, 0.05)$ such that prior means are well separated) and less bimodal (with greater prior scales $(0.5, 0.5)$). Our hyperparameter decision rule allows one to reach a similar prescribed low Wasserstein error for both choices of algorithm. As n increases, the multi-observation posterior becomes more Gaussian, thus making the sampling task easier and requiring a lower total number of steps. Overall, Linhart's compositional score yields a more efficient sampler in terms of score function evaluations, especially for large n and less multimodal target posterior (larger scales σ_1, σ_2). The results for the last scale choice, $(0.5, 0.05)$, are provided in Figure 3 (total number of steps) and the following Appendix (final Wasserstein error): this model strikes a balance between Gaussianity and strong non-Gaussianity since the posterior rapidly converges to a Gaussian as n increases.

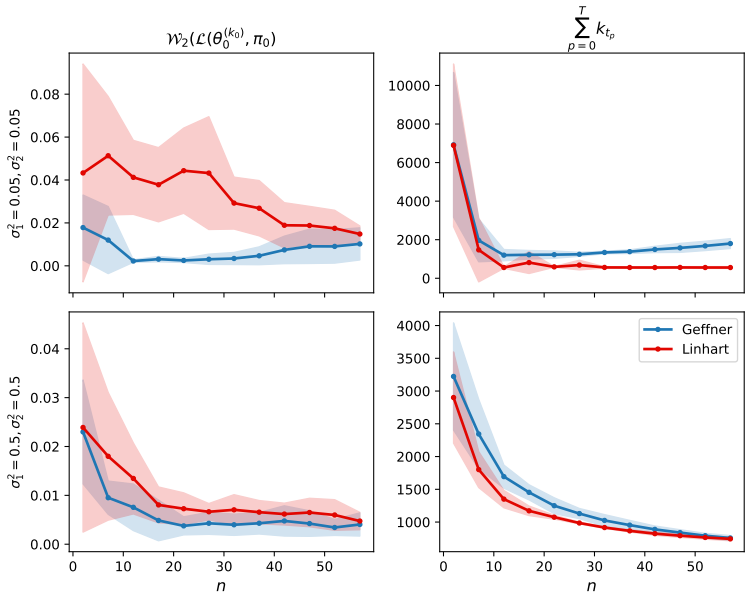

Figure 7: The top two subplots show the mean final Wasserstein error (left) and total number of steps (right) for a strong non Gaussian **GMM prior** model (scales σ_1 and σ_2 are very small). The bottom two subplots present the same results (mean Wasserstein error on the left and total number of steps on the right) for a less non Gaussian model (with greater prior scales). We use $T = 10$ annealing levels, $\gamma = 0.5$ and $\omega = 0.5$. The Wasserstein error and the complexity are presented for different numbers of conditional observations n. We observe that the empirical Wasserstein error is well controlled by γ on both tasks. Mean and std are computed over 5 seeds.

B.4 Technical details - Non Gaussian setting

Prior scores can be explicitly derived for all tasks. For **GMM prior** and **GMM likelihood** tasks, individual analytical scores can also be derived (see Appendix A.12). For **SIR** and **Lotka Volterra**, we train an amortized neural score network $s_\phi(\theta, x, t)$ using the `sbi` package (training loop, network architecture and default training parameters) [8]. The network is a MLP composed of 3 hidden layers of 256 hidden features and normalization layers. A Gaussian Fourier embedding is used for time.

For these two tasks, we train the network on log θ's for more stability, then apply the reverse transform during sampling. An additional embedding network (MLP with 2 hidden layers of 64 hidden features) is used to encode observations $x_1, \dots, x_n$ for the **Lotka Volterra** task. We use a training set of 25000 pairs (θ, x) for each task with a validation split $90\% - 10\%$. We train the score network for maximum 1000 epochs with early stopping (implemented by default in the `sbi` package).

We sample from the true target posteriors using MCMC with `numpyro` [18] and `jax` [4], except for **GMM prior** where analytical formula can be derived. We compute the Wasserstein error using the `POT` Python package [11] using 3000 samples obtained from the annealed Langevin algorithm and the true distribution. For Linhart's composite score we use their `GAUSS` algorithm to estimate covariance matrices using a small number (≈ 100) of diffusion steps. Importantly, we do not include this Linhart-specific cost (in terms of score evaluations) into the computational comparison: indeed, estimating such matrices is also necessary during fine-tuning (especially to compute m_t, M_t and $\mathcal{W}_2(\pi_t, \pi_{t+1})$), *regardless* of the choice of bridging densities. We thus treat this as a *shared cost* between both composite scores that we do not include in our comparison: the complexity only includes total score evaluations needed during sampling, i.e. $\sum_{p=0}^{T} k_{t_p}$.

We use $\gamma = 0.5$ as threshold for the Wasserstein error for all tasks and $\omega = 0.5$ during hyperparameter selection, except when the parameter dimension $d \geq 5$: in this case, we instead use $\omega = 0.8$ to reduce the number of steps while reaching a similar Wasserstein error. With respect to our training, we use a maximal value of the compositional error $\epsilon_{\text{DSM}} = 0.1$ for **SIR** and **Lotka Volterra**.

To run the experiments, we use a machine with an Intel CPU and 8 cores. We run the hyperparameter tuning followed by annealed Langevin sampling for different values of n at once: in practice we choose n IID observations $x_1, \dots, x_n$, then iteratively sample from the posteriors $p(\theta \mid x_{1:j})$ for $j \in [2, \dots, n]$. Finally, we consider $T = 10$ annealing levels (as in Song and Ermon [24]) uniformly spaced on the interval $[0, 1]$: for computational stability, we set $t_0 = 1e-4$ instead of 0 for sampling and fine-tuning the hyperparameters.

Figure 8 shows that hyperparameters chosen according to our decision rule indeed allow one to control the Wasserstein error of annealed Langevin algorithm for the different inference tasks.

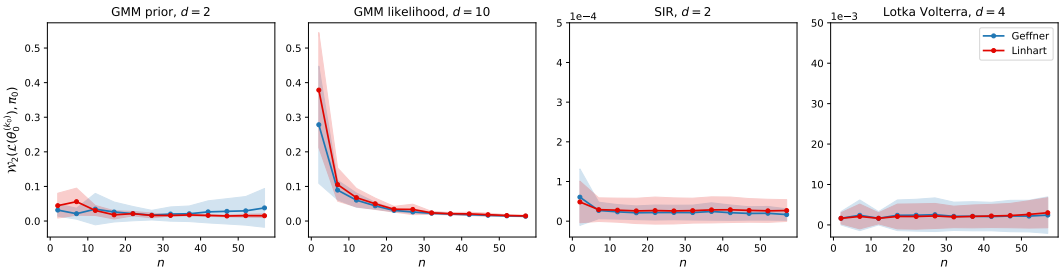

Figure 8: Evolution of the (mean) final Wasserstein error $\mathcal{W}_2(\mathcal{L}(\theta_0^{(k_0)}), \pi_0)$ of annealed Langevin algorithm run with the prescribed hyperparameters for the tasks presented in Section 5. We use $T = 10$ (as in Song and Ermon [24]) and $\gamma = 0.5$. These plots show that the prescribed hyperparameters allow one to control the final empirical Wasserstein error by γ whatever the number of conditional observations n. The Wasserstein error is also similar for both compositional scores, as expected with our fine-tuning. Importantly, the sampling does not include intermediate clipping steps. Mean and std are computed over 5 seeds.

As n grows, our Gaussian approximation of the multi-observation posterior (in Section 5) becomes more accurate, even when each individual posterior is approximated through moment matching (as discussed in Appendix B.6). Moreover, the necessary condition on the smoothness constants in

Lemma 1 is more likely to hold for larger values of n, making the bridging densities π_t^L and π_t^G satisfy the (smoothness and log-concavity) assumptions required for our decision rule to apply properly. In other words, increasing the number of conditioning variables makes these assumptions more likely to be satisfied, enabling our decision rule to yield well-tuned hyperparameters: this may explain why the Wasserstein error tends to decrease or stabilize in Figure 8 for all tasks. However, for large n, a slight increase may come from the accumulation of errors when aggregating individual scores into a compositional score, meaning that the compositional error ϵ_{DSM} may be slightly underestimated. Moreover, for strongly multimodal distributions (**GMM likelihood**), the Wasserstein error remains controlled but closer to the threshold γ than for large n where multimodality is progressively smoothed. Overall, we observe that the Wasserstein error remains controlled and little variable for large n, which manual tuning does not allow in general.

B.5 Comparison with score-based diffusion sampling and alternative Langevin tuning

In this section, we propose empirical evidence that diffusion sampling based on a compositional score fails to produce samples from the target distribution π_0, especially as n grows. Sampling quality is measured in Wasserstein distance. For diffusion sampling, we use a reverse VP diffusion scheme with 1000 steps (as prescribed in Song et al. [26]), $\beta_{\text{min}} = 0.01$, $\beta_{\text{max}} = 10$ and $t_0 = 1e-4$.

We also compare the final Wasserstein quality when choosing annealed Langevin hyperparameters according to the empirical tuning in Geffner et al. [13]: they use $T = 400$ levels and the same number of steps per annealing level while adjusting the step size according to a specific taming rule. We adapt their empirical tuning (constant number of step and taming rule for the step size) while keeping the same number of score evaluations as used with our decision rule. For our tuning, we keep using the same values: $\gamma = 0.5$, $T = 10$ and $t_0 = 1e-4$.

Tables 1 and 2 show that diffusion sampling indeed yields a greater empirical Wasserstein error than annealed Langevin sampling especially for $n > 2$, as theoretically expected. They also show that we achieve a better Wasserstein quality with our tuned annealed Langevin sampling than with Geffner's empirical tuning for both compositional scores, especially for $n > 2$.

Table 1: Mean final Wasserstein error when using *diffusion* sampling, annealed Langevin sampling with Geffner's *empirical* tuning or *our* fine-tuning on the **SIR** task for different numbers of conditional observations n. Mean and std are computed over 5 seeds. Bold figures indicate the lowest Wasserstein error across sampling methods.

Geffner's compositional score s_t^G

n	ours	empirical	diffusion
2	1.6e−04 ± 1.8e−04	6.4e−02 ± 6.4e−02	**8.0e−05 ± 8.0e−05**
12	5.3e−05 ± 5.0e−05	**4.2e−05 ± 4.2e−05**	1.2e−02 ± 1.2e−02
22	**5.0e−05 ± 5.4e−05**	1.1e−04 ± 1.1e−04	3.5e−01 ± 3.5e−01
32	**4.9e−05 ± 5.2e−05**	3.7e−04 ± 3.7e−04	2.9e+00 ± 2.9e+00
42	**5.1e−05 ± 5.6e−05**	1.0e−03 ± 1.0e−03	2.6e+01 ± 2.6e+01
52	**4.8e−05 ± 5.2e−05**	2.1e−03 ± 2.1e−03	8.1e+01 ± 8.1e+01

Linhart's compositional score s_t^L

n	ours	empirical	diffusion
2	**5.5e−05 ± 6.9e−05**	3.5e−01 ± 1.7e−01	8.7e−05 ± 9.0e−05
12	**3.0e−05 ± 2.5e−05**	3.9e−01 ± 6.5e−02	7.2e−04 ± 1.0e−03
22	**3.2e−05 ± 2.3e−05**	3.8e−01 ± 6.4e−02	2.2e−03 ± 1.5e−03
32	**3.1e−05 ± 2.3e−05**	3.8e−01 ± 6.4e−02	8.4e−03 ± 1.0e−02
42	**4.1e−05 ± 2.8e−05**	3.8e−01 ± 6.5e−02	7.0e−02 ± 1.6e−01
52	**5.2e−05 ± 4.3e−05**	3.7e−01 ± 6.5e−02	7.7e−01 ± 1.8e+00

Table 2: Mean final Wasserstein error when using *diffusion* sampling, annealed Langevin sampling with Geffner's *empirical* tuning or *our* fine-tuning on **Lotka Volterra** task for different numbers of conditional observations n. Mean and std are computed over 5 seeds. Bold figures indicate the lowest Wasserstein error across sampling methods.

Geffner's compositional score s_t^G

n	ours	empirical	diffusion
2	5.7e−03 ± 8.4e−03	**5.2e−03 ± 5.2e−03**	7.2e−03 ± 7.2e−03
12	1.6e−03 ± 1.6e−03	1.4e−03 ± 1.4e−03	**1.3e−03 ± 1.3e−03**
22	**2.4e−03 ± 3.9e−03**	9.5e−03 ± 9.5e−03	1.3e+00 ± 1.3e+00
32	**2.2e−03 ± 3.7e−03**	3.7e−02 ± 3.7e−02	4.7e+01 ± 4.7e+01
42	**2.2e−03 ± 3.7e−03**	9.6e−02 ± 9.6e−02	1.6e+04 ± 1.6e+04
52	**2.3e−03 ± 4.1e−03**	1.6e−01 ± 1.6e−01	1.7e+05 ± 1.7e+05

Linhart's compositional score s_t^L

n	ours	empirical	diffusion
2	5.2e−03 ± 7.4e−03	**4.7e−03 ± 6.1e−03**	6.2e−03 ± 5.3e−03
12	**1.7e−03 ± 1.4e−03**	8.5e−03 ± 9.7e−03	2.0e−03 ± 1.2e−03
22	**2.1e−03 ± 3.0e−03**	2.6e−02 ± 1.2e−02	3.0e−03 ± 1.5e−03
32	**2.2e−03 ± 2.5e−03**	2.5e−02 ± 9.7e−03	5.7e−03 ± 3.7e−03
42	**2.8e−03 ± 2.7e−03**	2.5e−02 ± 8.4e−03	8.2e−03 ± 5.7e−03
52	**4.0e−03 ± 3.6e−03**	2.7e−02 ± 1.2e−02	1.1e−02 ± 7.8e−03

B.6 Discussion on the Gaussian approximations

To get analytical expressions of bridging densities π_t^G and π_t^L and apply our decision rule in practice, we resort to a Gaussian approximation of individual posteriors $p(\theta \mid x_i)$ (see Section 5). We do not make any additional assumptions here but use the ones of `GAUSS` algorithm from Linhart et al. [16]: we thus choose a *moment matching* estimation of each individual posterior, where the first two moments are empirically computed.

As π_t^G and π_t^L coincide with the target multi-observation posterior π_0 at time $t = 0$, we approximate π_0 by composing the n Gauss-approximated individual posteriors as per $\pi_0(\theta \mid x_{1:n}) \propto \lambda(\theta)^{1-n} \prod_{i=1}^n p(\theta \mid x_i)$. This approximation at time $t = 0$ is used to get an analytical expression of π_t^L for all $t > 0$ and is easy to implement. However, beyond the pure Gaussian case, it does not ensure that the product of Gauss-approximated individual posteriors coincides with the asymptotic Gaussian distribution π_0 should converge to as n grows (according to the Bernstein von Mises theorem). Table 3 shows that there is indeed a mismatch with our moment-matching estimation but the Wasserstein error between π_0 (considered as Gaussian with the first two moments computed on MCMC samples) and our approximation remains low across different tasks, especially for unimodal posteriors and even for small n.

Table 3: Wasserstein error made when approximating the target multi-observation posterior by a product of Gauss-approximated individual posteriors. Mean is computed over 5 seeds.

	$n = 2$	$n = 12$	$n = 22$	$n = 32$	$n = 42$	$n = 52$
GMM likelihood	2.3e−01	1.2e+00	3.1e−01	1.9e−01	2.1e−01	1.3e−01
Lotka Volterra	5.5e−04	2.7e−04	7.8e−05	1.0e−04	2.1e−04	1.6e−04
SIR	9.5e−06	3.9e−05	4.4e−05	6.9e−05	6.2e−05	7.8e−05

Using instead a *Laplace/curvature matching* estimation of each individual posterior would be an interesting alternative as it would yield a multi-observation posterior closer to the one provided by the Bernstein von Mises theorem; this approximation is however harder to obtain as we need to estimate the curvature of each individual posterior.

B.7 Empirical estimation of the score error ϵ_{DSM}

The score error ϵ_{DSM} that appears in our decision rule denotes the compositional score matching error taken on average over the *iterates of the annealed Langevin process*. In practice, this error is difficult to estimate and we resort to approximations: first, we learn an amortized individual score by minimizing the denoising score matching loss and thus get an estimate of the score error ϵ_{indiv} on average over *samples from the diffused posteriors* $p_t(\theta \mid x_i)$; second, we upperbound the compositional score error ϵ_{DSM} as $O(n\epsilon_{\text{indiv}})$ according to Lemma 3 for both algorithms, even if Linhart's compositional score error seems to stabilize as n increases (see Touron et al. [27]).

This way of learning the score does not guarantee that the learned (individual or compositional) score provides an accurate estimate along the whole annealed Langevin sampling trajectory, since the Langevin sampler may visit regions that differ from the diffusion path, especially in middle to low noise regimes: this is not necessarily the case in (Gaussian) high noise regimes, where both samplers should have close trajectories or visit similar regions. Importantly, we empirically observe that miss estimating the compositional error ϵ_{DSM} does not have a huge impact on the choice of hyper-parameters for early annealing levels, as it is compensated by log-concavity constants m_t. This is even true when n grows as the log-concavity constant increases as $O(n)$ (see Appendix A.7). Tables 4 and 5 present the evolution of the log-concavity constants m_t^G and m_t^L for the tasks **SIR** and **Lotka Volterra**.

Table 4: Evolution of m_t^G and m_t^L for **SIR**. Mean and std are computed over 5 seeds.

m_t^G	$t = 0.001$	$t = 0.25$	$t = 0.5$	$t = 1$
$n = 2$	1524.19 ± 1078.99	2.81 ± 0.01	1.15 ± 0.00	1.01 ± 0.00
$n = 22$	16838.38 ± 12186.26	12.51 ± 0.13	1.67 ± 0.01	1.03 ± 0.00
$n = 52$	39721.55 ± 28684.69	27.05 ± 0.31	2.45 ± 0.01	1.08 ± 0.00

m_t^L	$t = 0.001$	$t = 0.25$	$t = 0.5$	$t = 1$
$n = 2$	1514.96 ± 1066.81	2.33 ± 0.00	1.12 ± 0.00	1.01 ± 0.00
$n = 22$	14785.96 ± 9764.15	2.33 ± 0.00	1.12 ± 0.00	1.01 ± 0.00
$n = 52$	29945.66 ± 17734.49	2.33 ± 0.00	1.12 ± 0.00	1.01 ± 0.00

Table 5: Evolution of m_t^G and m_t^L for **Lotka Volterra**. Mean and std are computed over 5 seeds.

m_t^G	$t = 0.001$	$t = 0.25$	$t = 0.5$	$t = 1$
$n = 2$	41.37 ± 6.11	2.65 ± 0.04	1.14 ± 0.00	1.01 ± 0.00
$n = 22$	440.34 ± 72.71	11.71 ± 0.44	1.68 ± 0.02	1.04 ± 0.00
$n = 52$	1040.12 ± 171.19	25.28 ± 0.99	2.49 ± 0.05	1.08 ± 0.00

m_t^L	$t = 0.001$	$t = 0.25$	$t = 0.5$	$t = 1$
$n = 2$	41.38 ± 6.09	2.26 ± 0.01	1.12 ± 0.00	1.01 ± 0.00
$n = 22$	440.57 ± 74.18	2.32 ± 0.00	1.12 ± 0.00	1.01 ± 0.00
$n = 52$	1032.74 ± 172.97	2.33 ± 0.00	1.12 ± 0.00	1.01 ± 0.00

B.8 Additional results on the choice of γ

Theoretical results indicate that decreasing γ yields a better Wasserstein quality but requires *lower* steps sizes and *more* steps per level. Indeed, for $\gamma \in [\frac{\epsilon_{\text{DSM}}}{\omega m_t}, +\infty[$ (which is the case in the experiments we run), h_t is a second order polynomial with unique root $\frac{\epsilon_{\text{DSM}}}{\omega m_t}$ and positive second order derivative, according to our decision rule in Section 3: it is thus an increasing function of γ. This means that $\gamma \leq \tilde{\gamma}$ implies $h_t \leq \tilde{h}_t$. Let $f(\gamma) = \frac{(1-\omega)\gamma}{\gamma + \mathcal{W}_2(\pi_t, \pi_{t+1})}$: it is an increasing function of $\gamma > 0$ that takes values in $[0, 1]$ since

$$f'(\gamma) = \frac{(1-\omega)\mathcal{W}_2(\pi_t, \pi_{t+1})}{(\gamma + \mathcal{W}_2(\pi_t, \pi_{t+1}))^2} \geq 0.$$

As we choose $k_t = \log\left(f(\gamma)\right) \frac{1}{\log(1-m_t h_t)}$ according to our decision rule,

$$\begin{aligned}
\gamma \leq \tilde{\gamma} &\Rightarrow \log f(\gamma) \leq \log f(\tilde{\gamma}) \leq 0 \quad \text{since } f \text{ and } \log \text{ are increasing} \\
&\Rightarrow \log\left(f(\gamma)\right) \frac{1}{\log(1-m_t h_t)} \geq \log\left(f(\tilde{\gamma})\right) \frac{1}{\log(1-m_t h_t)} \\
&\Rightarrow k_t \geq \tilde{k}_t.
\end{aligned}$$

We use the same γ for all the experiments for consistency and notice that $\gamma = 0.5$ yields good empirical Wasserstein quality for multimodal tasks (e.g. **GMM likelihood**) and a reasonable computational cost with respect to lower or higher γ. Concerning other tasks, a slightly greater γ would also be a proper choice. Indeed, Tables 6 and 7 seem to indicate that the final Wasserstein quality in these tasks decreases slower than $O(\gamma)$ for small n: this means that a greater γ yields a lower total computational cost without degrading the final accuracy too rapidly.

Table 6: Results for **GMM likelihood**. The first table presents the final Wasserstein accuracy; the second one shows the corresponding total computational cost. Mean and std are computed over 5 seeds.

Wasserstein accuracy

	Geffner's compositional score s_t^G		**Linhart's compositional score** s_t^L	
n	$\gamma = 0.5$	$\gamma = 1.5$	$\gamma = 0.5$	$\gamma = 1.5$
2	0.28 ± 0.17	0.30 ± 0.20	0.38 ± 0.17	0.47 ± 0.20
12	0.06 ± 0.02	0.07 ± 0.02	0.07 ± 0.03	0.11 ± 0.09
22	0.03 ± 0.00	0.04 ± 0.01	0.03 ± 0.01	0.07 ± 0.05
32	0.02 ± 0.00	0.04 ± 0.01	0.02 ± 0.00	0.07 ± 0.05
42	0.02 ± 0.00	0.06 ± 0.06	0.02 ± 0.00	0.10 ± 0.10
52	0.01 ± 0.00	0.07 ± 0.07	0.02 ± 0.00	0.11 ± 0.11

Computational cost

	Geffner's compositional score s_t^G		**Linhart's compositional score** s_t^L	
n	$\gamma = 0.5$	$\gamma = 1.5$	$\gamma = 0.5$	$\gamma = 1.5$
2	9548 ± 7379	906 ± 710	8152 ± 5787	766 ± 544
12	9014 ± 11693	781 ± 1006	3858 ± 2643	356 ± 238
22	5292 ± 3801	452 ± 325	2701 ± 562	249 ± 48
32	4711 ± 2638	400 ± 224	2510 ± 371	232 ± 30
42	4438 ± 2498	375 ± 213	2432 ± 328	226 ± 26
52	4606 ± 3059	389 ± 261	2351 ± 242	219 ± 19

Table 7: Results for **Lotka Volterra**. The first table presents the final Wasserstein accuracy; the second one shows the corresponding total computational cost. Mean and std are computed over 5 seeds.

Wasserstein accuracy

	Geffner's compositional score s_t^G		**Linhart's compositional score** s_t^L	
n	$\gamma = 0.5$	$\gamma = 1.5$	$\gamma = 0.5$	$\gamma = 1.5$
2	$5.7\text{e}{-03} \pm 8.4\text{e}{-03}$	$5.7\text{e}{-03} \pm 8.0\text{e}{-03}$	$5.2\text{e}{-03} \pm 7.4\text{e}{-03}$	$5.1\text{e}{-03} \pm 7.0\text{e}{-03}$
12	$1.6\text{e}{-03} \pm 1.6\text{e}{-03}$	$1.7\text{e}{-03} \pm 1.7\text{e}{-03}$	$1.7\text{e}{-03} \pm 1.4\text{e}{-03}$	$2.2\text{e}{-03} \pm 2.2\text{e}{-03}$
22	$2.4\text{e}{-03} \pm 3.9\text{e}{-03}$	$2.5\text{e}{-03} \pm 4.0\text{e}{-03}$	$2.1\text{e}{-03} \pm 3.0\text{e}{-03}$	$3.7\text{e}{-03} \pm 3.3\text{e}{-03}$
32	$2.2\text{e}{-03} \pm 3.7\text{e}{-03}$	$2.4\text{e}{-03} \pm 3.7\text{e}{-03}$	$2.2\text{e}{-03} \pm 2.5\text{e}{-03}$	$5.8\text{e}{-03} \pm 3.0\text{e}{-03}$
42	$2.2\text{e}{-03} \pm 3.7\text{e}{-03}$	$2.3\text{e}{-03} \pm 3.8\text{e}{-03}$	$2.8\text{e}{-03} \pm 2.7\text{e}{-03}$	$1.2\text{e}{-02} \pm 5.6\text{e}{-03}$
52	$2.3\text{e}{-03} \pm 4.1\text{e}{-03}$	$2.4\text{e}{-03} \pm 4.3\text{e}{-03}$	$4.0\text{e}{-03} \pm 3.6\text{e}{-03}$	$1.8\text{e}{-02} \pm 1.2\text{e}{-02}$

Computational cost

	Geffner's compositional score s_t^G		**Linhart's compositional score** s_t^L	
n	$\gamma = 0.5$	$\gamma = 1.5$	$\gamma = 0.5$	$\gamma = 1.5$
2	31048 ± 18609	14717 ± 8856	16926 ± 9567	5218 ± 2955
12	5805 ± 3771	2828 ± 1826	411 ± 85	122 ± 26
22	3249 ± 2124	1587 ± 1031	299 ± 16	94 ± 3
32	2268 ± 1406	1106 ± 683	283 ± 8	91 ± 2
42	1758 ± 1089	856 ± 529	278 ± 7	88 ± 2
52	1442 ± 855	702 ± 415	276 ± 7	87 ± 1

B.9 Additional results on the choice of T

According to theoretical results, the choice of T should not affect the final Wasserstein accuracy: indeed, we choose hyperparameters at each annealing level so that the intermediate Wasserstein quality is kept controlled by γ (see Section 3). However, as T increases, the Wasserstein distance between consecutive bridging densities, $\mathcal{W}_2(\pi_{t_p}, \pi_{t_{p+1}})$, which enters the selection of k_t, decreases: in the *pure Gaussian* setting, we observe that numerical instabilities may arise when T becomes sufficiently large (e.g. $T = 50$ or 100), although this is not systematic. The experiments presented in the paper always use $T = 10$ as originally proposed in Song and Ermon [24] for image generation. We test different values of T on non Gaussian tasks but this does not provide any significant empirical Wasserstein improvement (especially for Linhart's score). Table 8 presents the results for the **SIR** task: on this task, numerical instabilities may appear for larger T than in the pure Gaussian case (e.g. $T \geq 1000$).

Table 8: Wasserstein error obtained for $T = 10, 20, 50$ on the **SIR** task for different values of n. Mean and std are computed over 5 seeds.

Geffner's compositional score s_t^G

n	$T = 10$	$T = 20$	$T = 50$
2	1.6e−04 ± 1.8e−04	6.9e−05 ± 7.6e−05	4.6e−05 ± 4.5e−05
12	5.3e−05 ± 5.0e−05	2.3e−05 ± 1.8e−05	1.8e−05 ± 1.8e−05
22	5.0e−05 ± 5.4e−05	1.7e−05 ± 1.2e−05	1.2e−05 ± 9.1e−06
32	4.9e−05 ± 5.2e−05	1.5e−05 ± 1.1e−05	1.2e−05 ± 9.7e−06
42	5.1e−05 ± 5.6e−05	1.5e−05 ± 1.2e−05	1.3e−05 ± 1.1e−05
52	4.8e−05 ± 5.2e−05	1.3e−05 ± 1.1e−05	1.0e−05 ± 7.0e−06

Linhart's compositional score s_t^L

n	$T = 10$	$T = 20$	$T = 50$
2	5.5e−05 ± 6.9e−05	5.4e−05 ± 6.5e−05	4.8e−05 ± 4.8e−05
12	3.0e−05 ± 2.5e−05	2.9e−05 ± 2.5e−05	2.8e−05 ± 2.8e−05
22	3.2e−05 ± 2.3e−05	2.8e−05 ± 2.6e−05	2.8e−05 ± 3.1e−05
32	3.1e−05 ± 2.3e−05	2.9e−05 ± 2.4e−05	2.7e−05 ± 2.8e−05
42	4.1e−05 ± 2.8e−05	3.3e−05 ± 2.3e−05	2.8e−05 ± 2.8e−05
52	5.2e−05 ± 4.3e−05	3.5e−05 ± 2.5e−05	2.5e−05 ± 2.6e−05

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly define our contributions in the introduction and detail them throughout the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in the paper and try to weaken the assumptions of our main theoretical results.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All proofs are given in Appendix. The assumptions and notations used are detailed in the main sections.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The technical details are provided in the corresponding Appendix section for reproducibility. We detail the Python packages used as well as the score network architecture.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We do not release code or data, as we mainly use existing and open source packages and code for the experiments. All technical details are provided in the corresponding Appendix section.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All necessary details are provided in the corresponding Appendix section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide empirical evaluations of the proposed guidelines in the Appendix sections for all the considered tasks.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the requested details in the corresponding Appendix section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our work does not have any specific societal impacts. It only provides guidelines to tune parameters of a specific sampling method.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not pose such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The main compositional scores studied in this paper are cited several times. All the intermediate results used to prove the results are properly cited. The open source packages used to run the experiments are properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release new assets. All details are well documented in the Appendix sections.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowd sourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core development of our research does not involve LLMs as any important components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.