Annealing in variational inference mitigates mode collapse: A theoretical study on Gaussian mixtures

Luigi Fogliani

Département d'informatique, École Normale Supérieure, Université PSL
Laboratoire de Physique de l'École Normale Supérieure, Université PSL
luigi.fogliani@di.ens.fr

Bruno Loureiro

Département d'informatique, École Normale Supérieure, Université PSL

Marylou Gabrié

Laboratoire de Physique de l'École Normale Supérieure, Université PSL

Abstract

Mode collapse — the failure to capture one or more modes when targetting a multimodal distribution — is a central challenge in modern variational inference. In this work, we provide a mathematical analysis of annealing-based strategies for mitigating mode collapse in a tractable setting: learning a Gaussian mixture, where mode collapse is known to arise. Leveraging a low-dimensional summary statistics description, we precisely characterize the interplay between the initial temperature and the annealing rate, and derive a sharp formula for the probability of mode collapse. Our analysis shows that an appropriately chosen annealing scheme can robustly prevent mode collapse. Finally, we present numerical evidence that these theoretical trade-offs qualitatively extend to neural network–based models, RealNVP normalizing flows, providing guidance for designing annealing strategies mitigating mode collapse in practical variational inference pipelines.

1 Introduction

Variational inference is a widely used framework for approximating complex probability distributions in settings where direct sampling is computationally prohibitive, with applications ranging from scalable Bayesian inference in machine learning to the sampling of Boltzmann distributions and free-energy landscapes in physics, chemistry, and materials science. Given a target distribution π, variational inference (VI) replaces the task of sampling from π with that of finding, within a family of candidate distributions $\mathcal{Q}$, the probabilistic model q that best approximates π. The variational family $\mathcal{Q}$ is restricted to distributions that are easy to sample from and whose probability density function, also denoted by q, is known, normalization included. These requirements make the reverse Kullback–Leibler (KL) divergence

$$\mathcal{D}_{\mathrm{KL}}(q|\pi) = \mathbb{E}_q\left[\log \frac{q(x)}{\pi(x)}\right] \tag{1.1}$$

a tractable criterion for selecting q, and allow samples from q to be used a posteriori as proxies for samples from π. Notably, optimizing $\mathcal{D}_{\mathrm{KL}}(q|\pi)$ with respect to q does not require having knowledge of the normalization of π. Rooted in statistical mechanics [Weiss, 1907], VI was later popularized in

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

Bayesian statistics [Jordan et al., 1999, Blei et al., 2017], in particular for its good scaling properties. Moreover, advances in deep generative modelling have greatly expanded the class of tractable variational distributions, with normalizing flows providing flexible parameterizations that retain both easy sampling and tractable probability densities [Rezende and Mohamed, 2015, Albergo et al., 2019] and autoregressive networks providing the equivalent for discrete variables [Wu et al., 2019].

However, a class of target distributions poses a significant challenge to VI, namely multimodal distributions. The reverse KL objective is well known to be mode seeking [Jordan et al., 1999, Minka, 2005], yielding approximations that concentrate on a subset of the target modes, a phenomenon commonly referred to as mode collapse. Increasing the expressivity of variational families through deep generative models does not, in general, alleviate this issue [Noé et al., 2019, Hackett et al., 2021, Nicoli et al., 2023, Blessing et al., 2024, Soletskyi et al., 2025, Grenioux et al., 2025].

Various empirical strategies have been proposed to mitigate mode collapse in VI in the context of different fields of applications including molecular simulations [Noé et al., 2019], classical and quantum field theories [Wu et al., 2019, Vaitl et al., 2022, Nicoli et al., 2023] and Bayesian inference [Wang et al., 2025]. Among them, a widely applicable approach when optimization of the reverse KL is performed iteratively, typically using stochastic gradient descent, is to combine VI with annealing. Rather than targeting π directly, the variational distribution q is first optimized to approximate a tempered distribution proportional to π^β with $\beta < 1$, for which modes are typically connected. The temperature is then progressively lowered during optimization until $\beta = 1$ is reached. The early phases of optimization at high "temperature" $1/\beta$ promote exploration and mitigate mode collapse. Annealed VI has been successfully applied to sampling from Boltzmann distributions in statistical mechanics [Wu et al., 2019, Schopmans and Friederich, 2025], to sampling from Bayesian posteriors [Huang et al., 2018, Wang et al., 2025], and to optimization tasks, including locating multiple modes [LeMinh et al., 2025] or identifying global minima [Hibat-Allah et al., 2021, Khandoker et al., 2025].

Despite the major challenge posed by mode-collapse and the practical successes that annealing has brought, little theoretical work has focused on gaining understanding around its behavior. In this work, we make a first step in closing this gap by investigating annealing in a tractable multimodal VI setting. The target is a bimodal Gaussian mixture with well-specified variational family whose parameters are trained by gradient flow. Our **main contributions** are:

- First, we show that mode collapse can be systematically avoided in this setting by employing a suitable annealing schedule, including in parameter regimes where collapse is otherwise unavoidable. These experiments reveal a critical trade-off between the initial temperature and the annealing rate, which must be carefully balanced to reliably recover all target modes.
- We show that in the high-dimensional limit the dynamics of the student variances decouple from those of the means and admit a closed-form solution: the variances track the temperature with a delay set by the annealing rate. Combined with a set of summary statistics for the means, this recasts the high-dimensional gradient flow into a low-dimensional dynamical system, allowing us to tractably identify distinct high- and low-temperature dynamical regimes and derive a closed-form expression for the mode collapse probability that quantitatively captures the aforementioned experimental trade-off.
- We show that the previous findings extend to a practical and expressive choice of variational family, RealNVP normalizing flows [Dinh et al., 2017]. This suggests that our theoretical results can provide guidance for the design of annealing schedules in practical pipelines.

Related works. Theoretical investigation of mode collapse in VI has received little attention to this day. By studying the gradient flow associated with VI over Gaussian mixtures with fixed weights, Huix et al. [2024] obtained theoretical guarantees in a setting related to the one considered below. Their analysis, however, was not concerned with multimodal target distributions and did not investigate the mode-collapse phenomenon. Petit-Talamon et al. [2025] investigated variational inference with Gaussian mixture variational families through gradient-flow-based algorithms, emphasizing computational and algorithmic aspects, while leaving the analysis of mode collapse unexamined. An exception is [Soletskyi et al., 2025], from which the present paper draws inspiration.

Building on ideas from statistical physics of learning [Saad and Solla, 1995], Soletskyi et al. [2025] derived a summary statistics description for VI on a bimodal Gaussian mixture. A key difference with respect to our work is that in [Soletskyi et al., 2025] the variance is kept fixed, making the description not amenable to the study of annealing. Further, Soletskyi et al. [2025] showed that the probability of

occurrence of mode-collapse is controlled by the distance between modes of the target mixture with a critical radius R_c, function of the class imbalance, above which the dynamics from an uninformed initialization is highly likely to collapse on one mode. Using a similar setup, keeping the weights fixed for simplicity, the present work extends the analysis to learnable variances whose dynamics we characterize exactly in the high-dimensional limit, a key ingredient to study annealing which changes the effective variance of the target model along the training dynamics.

Finally, annealing was previously identified as a mean to avoid collapse in VI powered by modern generative models by several authors as already cited in the introduction [Wu et al., 2019, Schopmans and Friederich, 2025, Huang et al., 2018, Wang et al., 2025, LeMinh et al., 2025, Hibat-Allah et al., 2021, Khandoker et al., 2025]. Used as a heuristic method, little guidance is provided in these works on how to adjust the annealing schedule. An exception is [Wang et al., 2025], which proposes an adaptive schedule for the annealing speed based on the instantaneous effective sample size of proposals from a normalizing flow along its training by VI. Comparatively, the adjustment of annealing methods has been extensively discussed in the context of Monte Carlo samplers such as parallel tempering [Syed et al., 2021] or sequential Monte Carlo [Barzegar et al., 2024] and optimizers with simulated annealing [Hajek, 1988, Nadler and Hansmann, 2007]. In a complementary direction, [Ichikawa and Hukushima, 2024, 2025] theoretically analyze the effects of KL annealing in variational autoencoders, deriving sharp thresholds for posterior collapse and characterizing the resulting learning dynamics.

2 Setup and preliminary experiments

2.1 Annealed VI with bimodal Gaussian mixtures

We consider a minimal multimodal model for mode collapse in VI, where both π and q are mixtures of two isotropic Gaussian distributions on $\mathbb{R}^d$:

$$\begin{aligned} \pi(x) =& w_* \mathcal{N}(x|\mu_*, I_d) + (1-w_*)\mathcal{N}(x|-\mu_*, I_d), \\ q_\theta(x) =& w_1 \mathcal{N}(x|\mu_1, \sigma_1^2 I_d) + (1-w_1)\mathcal{N}(x|\mu_2, \sigma_2^2 I_d), \end{aligned} \tag{2.1}$$

where $w_*, w_1 \in [0,1]$, $\mu_*, \mu_1, \mu_2 \in \mathbb{R}^d$ and $\sigma_1, \sigma_2 \in \mathbb{R}_+$. Here the trainable parameters of the variational distribution q_θ are the two means and the two standard deviations: $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2) \in \mathbb{R}^{2(d+1)}$. To simplify the analytical treatment, we fix the weights and choose $w_1 = w_2 = 0.5$ in order not to break the symmetry between the modes of the variational model. Following [Soletskyi et al., 2025] we constrain the means to lie on a sphere $\mu_\star, \mu_{1,2} \in \mathbb{S}^{d-1}(R)$, which further simplifies the analytical discussion without affecting the qualitative behavior of the model. In this setup, q_θ is referred to as the student distribution.

As discussed in the introduction, given the variational family q_θ, annealed-VI consists of learning θ by minimizing the reverse Kullback-Leibler divergence loss between q_θ and the tempered distribution proportional to π^β where $1 > \beta > 0$ is the inverse temperature. Denoting by $\mathcal{Z}_\beta$ the normalization constant of the tempered distribution, the loss is the free energy:

$$\begin{aligned} \mathcal{L}(\theta, \beta) &= \beta^{-1} \mathcal{D}_{\mathrm{KL}} \left(q_\theta \,\middle\|\, \frac{\pi^\beta}{\mathcal{Z}_\beta} \right) \\ &= \beta^{-1} \mathbb{E}_{q_\theta}[\log q_\theta(x)] - w_1 \mathbb{E}_{\mathcal{N}(\mu_1, \sigma_1^2 I_d)}[\log \pi(x)] \\ &\quad - w_2 \mathbb{E}_{\mathcal{N}(\mu_2, \sigma_2^2 I_d)}[\log \pi(x)] + \beta^{-1} \log \mathcal{Z}_\beta. \end{aligned} \tag{2.2}$$

The first term in eq. (2.2) is the entropy of q_θ, which favors exploration pushing μ_1 and μ_2 apart. The two following terms are the cross-entropy between q_θ and π, each of them attracting μ_1 and μ_2, separately, at $\pm\mu_*$. The last term is a constant with respect to θ and is omitted in the following. The parameter β tunes the relative importance between student entropy and cross-entropy with the target, higher temperatures (i.e. lower β) encouraging exploration. Multiplying the KL divergence between q_θ and π by β^{-1} ensures that the derivative of the loss stays of order one throughout annealing.

In practice, $\mathcal{L}(\theta, \beta)$ is minimized with a descent-based algorithm. For simplicity, here we will focus on the spherical gradient flow:

$$\begin{aligned} \dot{\mu}_c(t) &= -\nabla_{\mu_c}^{\mathbb{S}^{d-1}(R)} \mathcal{L}\left(\theta(t), \beta(t)\right), \\ \dot{\sigma}_c(t) &= -\frac{1}{d} \nabla_{\sigma_c} \mathcal{L}\left(\theta(t), \beta(t)\right), \qquad c \in \{1, 2\}, \end{aligned} \tag{2.3}$$

from an initial $\mu_{1,2}$ uniformly sampled from $\mathbb{S}^{d-1}(R)$ and equal initial variance $\sigma_1^2(0) = \sigma_2^2(0) = \beta^{-1}(0)$. $\beta(t)$ is progressively annealed throughout the trajectory from an initial high-temperature ($\beta < 1$) to the target temperature ($\beta = 1$). In particular, we employ the JKO scheme [Jordan et al., 1998] on a geometry adapted for isotropic Gaussian mixtures proposed by Petit-Talamon et al. [2025], which allows us to derive the $1/d$ factor difference between the gradient flow equations for the means and for the standard deviations (see Appendix A). This scaling ensures that time derivatives with respect to means and variances stay of order one as d grows.

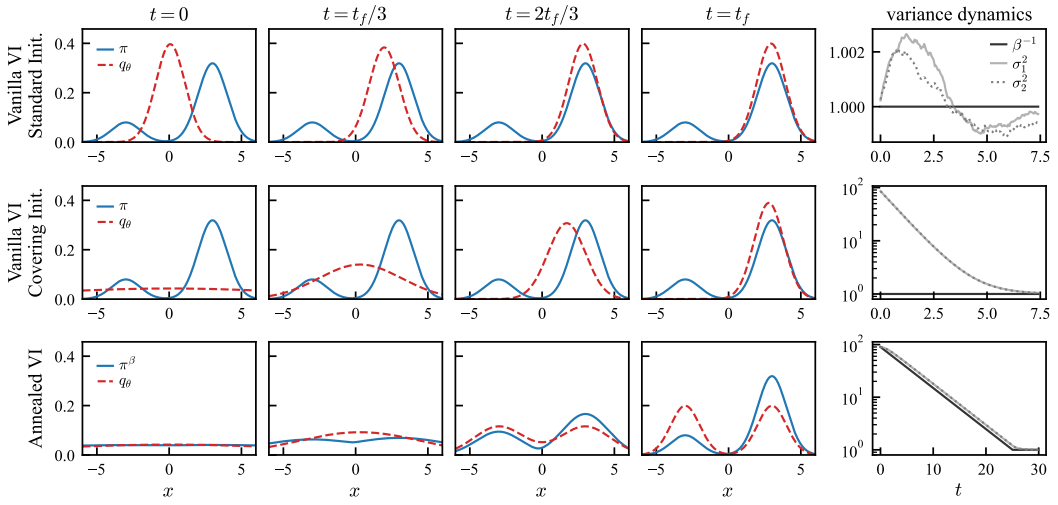


Figure 1: *Preliminary experiments: annealing mitigates mode collapse.* The first 4 columns show the marginal density along the direction of μ_* of both π and q_θ, at 4 different training stages. The fifth column shows the variance dynamics. Each row represents a different scenario; from top to bottom: no annealing with $\sigma_{1,2}$ initialized to one, no annealing with covering initial distribution (high initial variances $\sigma_{1,2} \gg R$), annealing with schedule given by eq. (2.4). Hyperparameters values are: $d = 512, R = 3, w_* = 0.8, w_1 = 0.5$ and time discretization step $\eta = 0.05$.

As shown by Soletskyi et al. [2025], despite its simplicity this setting displays a sufficiently rich phenomenology to investigate mode collapse, and offers the perfect playground to understand how annealing can mitigate it. As we will discuss in the following, our results considerably extend the scope of this work by considering trainable variances for the student. This ingredient is crucial for investigating annealing, as the annealing temperature directly acts on the effective variance of the tempered target π^β. More generally, it allows the student to have a wider distribution, which is believed to be important to avoid mode collapse in VI.

As a starting point, we consider a set of preliminary numerical experiments that will provide us guidance and illustrate interesting phenomenology for the discussion that follows.

2.2 Preliminary experiments

We consider a set up where the probability of mode collapse is high with vanilla VI ($\beta = 1$ fixed), setting the parameters of π to $R = 3$ (greater than the R_c, the onset radius for mode collapse as characterized in Soletskyi et al. [2025]), $w_* = 0.8$ and $w_1 = 0.5$ in dimension $d = 512$. Figure 1 illustrates the learning dynamics of q_θ with three different learning strategies.

Vanilla VI with an initial variance equal to 1 leads to a collapse of μ_1 and μ_2 to the same mode of π (first row). Furthermore, the variances σ_1 and σ_2 remain close to 1 throughout the descent, suggesting

that the learning dynamics are not very different from the fixed variance setup and, effectively, that trainable variances alone are not sufficient to avoid mode collapse.

A natural idea to avoid mode collapse is to improve the initialization of the gradient descent algorithm by increasing the initial variance of the modes of q_θ to make it overlap with the two modes of π. This can be done by setting the initial variances $\sigma_{1,2}^2 = \Omega(R^2)$ (second row of fig. 1). Yet, the two modes still collapse to the same mode, and variances rapidly converge to 1. Moreover, when fixing the values of $\sigma_{1,2}$ to their initial large value, we still observe mode collapse on the same mode (see Appendix fig. 4 first row) and interestingly the variances converge to the target value of 1 even if we freeze the means (see the second row of fig. 4 in the Appendix).

Finally, we find instead that the learning dynamics with an annealing schedule of the form [1]:

$$\beta(t) = \min(\beta_i^{1-t/t_0}, 1), \tag{2.4}$$

with $\beta_i^{-1} = 10R^2$ and $t_0 = 25$ (third row of fig. 1), avoids mode collapse. Variances are initialized to $1/\beta_i$ and interestingly, stay close to $1/\beta(t)$ throughout the learning dynamics, a behavior that we derive in section 3.1. The schedule in eq. (2.4) is a geometric progression in inverse temperature, which is a common baseline in tempering and has the advantage of remaining analytically tractable. This lets us isolate the trade-off between initial temperature and annealing time, which we discuss next.

These observations illustrate the typical phenomenology observed by practitioners performing VI on a multimodal target distribution with a flexible variational family. Despite a covering initialization, mode collapse can still occur and suitable annealing can be a mean to avoid it. However, choosing the right parameters — namely the initial temperature β_i and annealing time t_0 — is crucial to successfully avoid mode collapse. Next, we look closer at this dependency.

Influence of annealing schedule. While the efficacy of temperature annealing is well-documented in practice, the selection of an effective schedule remains a non-trivial challenge. To investigate this, we evaluate the impact of the annealing schedule in our controlled setting with an exponential form for $\beta(t)$ given in eq. (2.4). This choice is motivated by its parameter parsimony and the analytical findings that will follow, which suggest that preventing mode collapse requires the dynamics to remain in a high-temperature regime (specifically, $\beta < 1/R^2$) for a sufficient duration.[2]

Figure 2 (left) reports the probability of mode-collapse over a grid of annealing times (t_0) and initial inverse temperatures (β_i) for the same experimental set up as fig. 1. We observe a trade-off: if β_i is too high (low initial temperature), the system crosses the $1/R^2$ threshold too rapidly to resolve the target distribution's modes, leading to irreversible collapse. Conversely, while a very low β_i provides better exploration, it necessitates a significantly longer annealing time to avoid mode collapse leading to a higher computational budget. This implies the existence of an optimal β_i that balances modal resolution with convergence speed. Note that the same qualitative behavior holds with learnable student weights w_1, illustrated by fig. 6a in appendix E.1.

Our analytical analysis, presented next, leads to an estimation of the mode collapse probability consistent with these empirical observations (0.5 isoprobability line predicted by theory represented by the black dashed line on left panel of fig. 2). Further below, we also show that these observations hold qualitatively for more complex variational families, namely RealNVPs.

3 Analytical analysis

Our goal in this section is to give mathematical grounding to the numerical discussion of section 2.2. Our analysis builds on a low-dimensional reduction of the gradient flow dynamics in eq. (2.3), which we discuss in detail in section 3.1. Leveraging this description, we precisely characterize the probability of mode collapse for different choices of initial temperature and annealing rate, showing that a well-tempered annealing schedule is able to successfully mitigate mode collapse.

[1] In the experiments with Gaussian mixture students, the gradient flow is discretized with step $\eta = 0.05$, and times (including t_0) are expressed in gradient-flow units, i.e. $t = \eta\times$ (number of iterations). In the experiments using Adam (section 4 and Appendix E.1), t_0 is instead expressed in number of iterations.

[2] For instance, annealing schemes of the form $\beta(t) = \beta_i + (1-\beta_i)(1-e^{-t/t_0})$ comparatively stay longer in the low temperature regime, which does not help avoiding mode collapse.

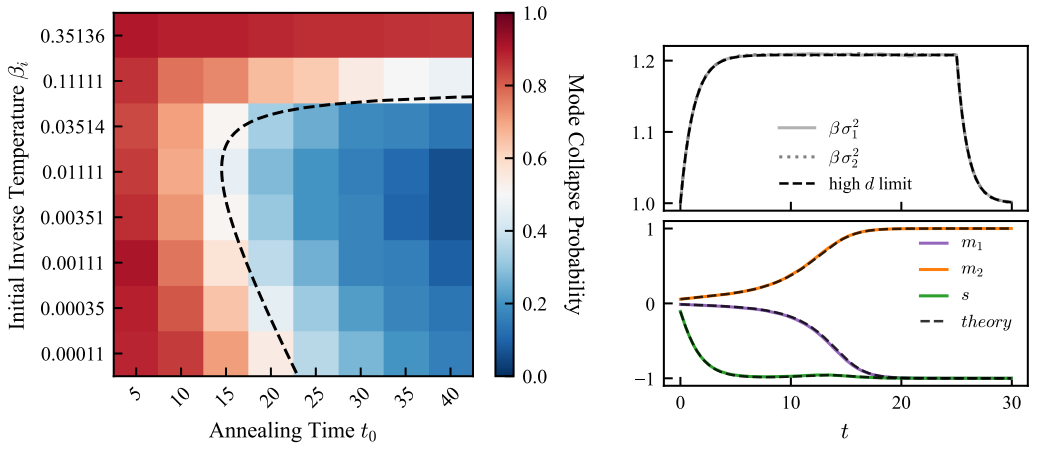


Figure 2: **(Left)** Annealed VI mode collapse probability with Gaussian mixture student. The annealing scheme is fixed by eq. (2.4) and different values of the initial inverse temperature β_i and annealing time t_0 are explored. Hyperparameters are the same as in fig. 1. For each (β_i, t_0), the mode collapse probability is computed using 100 different random student mean initializations. The dashed black line is the 0.5 mode collapse isoprobability line obtained with the analytical estimate of eq. (3.9). $\kappa < 1$ along this curve.
(Top Right) Dynamics of $\beta(t)\sigma^2(t)$. Grey curves correspond to SGD trajectories of the experiment corresponding to the third row of fig. 1. The black curve is the closed-form solution eq. (3.3). **(Bottom Right)** Dynamics of the summary statistics for the same experiment. Coloured curves correspond to SGD trajectories and black curves to the numerical integration of the eq. (3.6).

3.1 Summary statistics learning dynamics

Consider the gradient flow dynamics on $\mathcal{L}(\theta, \beta)$ introduced in eq. (2.3). Recall that our preliminary numerical experiments (fig. 1) showed that the student variances track the temperature during annealing, irrespective of the means. The following result, proven in Appendix C.1, explains this observation.

Proposition 1 (informal)**.** In the high-dimensional limit, the derivative of the loss with respect to the variances decouples from the means and from each other:

$$\lim_{d\to\infty} \frac{1}{d}\nabla_{\sigma_c}\mathcal{L}(\theta,\beta) = w_c\left(\sigma_c - \frac{1}{\beta\sigma_c}\right), \qquad c \in \{1,2\}. \tag{3.1}$$

Consequently, throughout our analysis, we consider the high-dimensional limit of the variance dynamics defined in eq. (2.3):

$$\dot{\sigma}_c(t) = -w_c\left(\sigma_c(t) - \frac{1}{\beta(t)\sigma_c(t)}\right), \qquad c \in \{1,2\}, \tag{3.2}$$

where the variances are initialized at the same value $\sigma_1^2(0) = \sigma_2^2(0) = \beta_i^{-1}$. Hence, equal weights ($w_1 = w_2 = 0.5$) and equal initialization imply $\sigma_1(t) = \sigma_2(t) := \sigma(t)$ at all times, and σ^2 relaxes towards β^{-1}. In particular, for the exponential schedule eq. (2.4), the above dynamics of eq. (3.2) yields

$$\beta(t)\sigma^2(t) = \begin{cases} \dfrac{1}{1-\kappa} - \dfrac{\kappa}{1-\kappa}e^{-(1-\kappa)t}, & t \le t_0, \\ 1 + \dfrac{\kappa(1-e^{-(1-\kappa)t_0})}{1-\kappa}e^{-(t-t_0)}, & t > t_0. \end{cases} \tag{3.3}$$

where κ is the annealing rate, defined as the derivative of $\log\beta(t)$: $\kappa = \log(1/\beta_i)/t_0 > 0$. eq. (3.3) holds for any $\kappa \neq 1$. When $\kappa < 1$, $\beta\sigma^2$ reaches a plateau; when $\kappa \geq 1$ (the fastest schedules of fig. 2), the variance cannot follow the temperature and $\beta\sigma^2$ grows. The analysis below assumes $\kappa < 1$. During the annealing ($t \leq t_0$), after a transient of order one, $\sigma^2(t) \approx \big((1-\kappa)\beta(t)\big)^{-1}$: the

variance tracks the temperature with a delay. After the temperature reaches its final value of 1, $\sigma^2(t)$ goes to one exponentially. The validity of the high-dimensional limit of the variance dynamics is illustrated by the excellent agreement of eq. (3.3) with the SGD trajectories of the variances in fig. 2 (top right) with $d = 512$.

We now focus on the trajectory of the means $\mu_{1,2}$ given $\sigma(t)$, which explicitly reads

$$\dot{\mu}_c(t) = -\left(I_d - \frac{\mu_c(t)\mu_c(t)^\top}{R^2}\right)\nabla_{\mu_c}\mathcal{L}(\theta(t), \beta(t)), \qquad c \in \{1, 2\}, \tag{3.4}$$

where the student's means $\mu_c(0)$ are initialized uniformly on the sphere $\mathbb{S}^{d-1}(R)$.

The key idea of our analysis is to note that the reverse-KL loss $\mathcal{L}(\mu_{1,2}, \sigma_{1,2}, \beta)$ can be equivalently described in terms of the following three scalar quantities, akin to *order parameters* in statistical physics

$$m_1 = \frac{\mu_1^\top \mu_*}{R^2}, \qquad m_2 = \frac{\mu_2^\top \mu_*}{R^2}, \qquad s = \frac{\mu_1^\top \mu_2}{R^2}. \tag{3.5}$$

For $c \in \{1, 2\}$, m_c represents the alignment of μ_c with the target modes $\pm\mu_*$, and s measures the alignment between the two student means. Crucially, the time evolution of these correlations recast the complex $2d$-dimensional flow into a 3-dimensional ordinary differential equation:

$$\begin{aligned}
\dot{m}_1 &= -\left[\frac{(m_2 - m_1 s)}{\beta(t)\sigma^2(t)} f(s, \sigma(t)) + w_1(1 - m_1^2) g(m_1, \sigma(t))\right], \\
\dot{m}_2 &= -\left[\frac{(m_1 - m_2 s)}{\beta(t)\sigma^2(t)} f(s, \sigma(t)) + w_2(1 - m_2^2) g(m_2, \sigma(t))\right], \\
\dot{s} &= -\left[\frac{2(1 - s^2)}{\beta(t)\sigma^2(t)} f(s, \sigma(t)) + w_1(m_2 - m_1 s) g(m_1, \sigma(t)) + w_2(m_1 - m_2 s) g(m_2, \sigma(t))\right].
\end{aligned} \tag{3.6}$$

Detailed computations and expressions for f and g are given in Appendix C.2. Here, $f(s, \sigma)$ represents the repulsive force arising from the student entropy term, while $g(m, \sigma)$ encapsulates the attractive force from the cross-entropy term. The validity of the gradient flow approximation is confirmed by the agreement of the numerical integration of eq. (3.6) with the SGD trajectories on the means (fig. 2 (bottom right)).

Notably, the above dynamical system does not depend on d, which only plays a role through the typical values of the summary statistics at initialization. In high dimensions, randomly initialized means concentrate near the equator orthogonal to the target axis: $m_c \sim \mathcal{N}(0, 1/d)$, with a small overlap $s \sim \mathcal{N}(0, 1/d)$. To better understand the learning dynamics described in eq. (3.6) and the interplay with the dimension, we analyze two instructive limiting cases of the dynamical system defined by the scaling of the temperature β^{-1} with respect to R^2.

3.2 Limiting cases of the dynamics

The low *temperature regime* ($\beta \gg R^{-2}$) is observed in later stages of the annealing schedule. Recall that at any time, the factor $\beta(t)\sigma^2(t)$ is of order 1, and that after a transient of order 1, $\sigma^2(t) \approx ((1 - \kappa)\beta(t))^{-1}$. Hence, in this regime, $\sigma \ll R$. In Appendix C.3, we show that, in this regime, the cross-entropy term g dominates, while the entropic repulsion f vanishes exponentially:

$$\begin{aligned}
f(s, \sigma) &\sim \frac{\sqrt{\pi}}{2} \frac{\sqrt{w_1 w_2}}{\sqrt{1 - s}} \frac{\sigma}{R} e^{\frac{-R^2(1-s)}{4\sigma^2}}, \\
g(m, \sigma) &= -\tanh\left(R^2(m + \epsilon)\right) + \mathcal{O}\left(\sigma^2/R^2\right),
\end{aligned}$$

where the bias $\epsilon = \frac{1}{2R^2} \log \frac{w_*}{1 - w_*}$ is an increasing function of w_*. Therefore, the evolution of the means according to eq. (3.6) becomes independent: $\dot{m}_c = w_c(1 - m_c^2) \tanh\left(R^2(m_c + \epsilon)\right)$ for $c \in \{1, 2\}$. This creates two distinct basins of attraction for each mean separated by the hyperplane $m = -\epsilon$. The asymptotic fate of each mean is determined solely by whether its value at the onset of this regime is greater or lesser than $-\epsilon$.

A particular case is when no annealing is used ($\beta(t) = 1$, with a sufficiently large R), as in Soletskyi et al. [2025], which leads to inevitable mode collapse in high dimensional settings. If $1/\sqrt{d} \ll |\epsilon|$ (as in the preliminary experiments), the two student means will fall within the basin of attraction of the heaviest target mode with very high probability, resulting in mode collapse.

Instead, in the *high temperature* regime ($\beta \ll R^{-2}$), as discussed next, the student means are driven away from each other, and consequently, one of them escapes the basin of attraction of the heaviest mode. Namely, when the temperature is high, $\sigma \gg R$ (because $\beta(t)\sigma^2(t)$ is of order 1) and the entropy term dominates the cross-entropy term:

$$f(s, \sigma) = w_1 w_2 + \mathcal{O}(R^2/\sigma^2)\,, \tag{3.7}$$
$$g(m, \sigma) = -\sqrt{\frac{2}{\pi}}\frac{R}{\sigma}(m + \epsilon) + \mathcal{O}\left(R^2/\sigma^2\right).$$

Again, see Appendix C.3 for a detailed discussion. Here, the dynamics drive the overlap s to -1, therefore effectively projecting the system onto the symmetric subspace where $\mu_1 = -\mu_2$ with both scaling as $1/\sqrt{d}$.

To analyze the stability of this configuration in high dimensions, we linearize the dynamics around the symmetric point ($s = -1$, $m_1 = -m_2 = 0$). We assume that $\beta(t)\sigma^2(t)$ has reached its plateau $(1 - \kappa)^{-1}$. The time evolution of the sum and difference of m_1 and m_2 decouple from s:

$$\begin{aligned}
\dot{m}_1 + \dot{m}_2 &= -\left[2(1-\kappa)f(-1) + \frac{1}{2}g'(0)\right](m_1 + m_2) - g(0), \\
\dot{m}_1 - \dot{m}_2 &= -\frac{1}{2}g'(0)(m_1 - m_2),
\end{aligned} \tag{3.8}$$

where the second argument of f and g equal to $((1-\kappa)\beta)^{-1/2}$ has been omitted for clarity. The system exhibits a saddle point structure. From the asymptotics of eq. (3.7), the sum is stable because $2(1-\kappa)f(-1) + \frac{1}{2}g'(0) \approx 2(1-\kappa)w_1 w_2 > 0$, however the difference is unstable since $g'(0) \approx -\sqrt{\frac{2}{\pi}}\sqrt{(1-\kappa)\beta}\,R < 0$. The latter instability is the key mechanism to mitigate mode collapse as it drives m_1 and m_2 away from 0 and in opposite directions, each of the student mean specializing into a different target mode.

Interestingly, the escape speed from the saddle point $-g'(0) \propto \sqrt{(1-\kappa)\beta}\,R$ is a decreasing function of the temperature and the annealing rate κ. This implies that arbitrarily high temperatures and annealing rates are comparatively less efficient at driving the specialization of the student means into different modes than temperatures closer (but still superior) to R^2. Leveraging this high-temperature dynamics, we derive next an analytical estimate for the probability of mode collapse.

3.3 Mode collapse probability

Mode collapse occurs if the student means have not been sufficiently driven away during the high temperature regime to fall into distinct basin of attraction of the low temperature regime. Assuming the transition from the high temperature regime to low temperature regime is sharp and occurs at inverse temperature $\beta = \alpha R^{-2}$ — where α is a constant of order 1 — yields the following mode collapse probability $p(\beta)$ for an annealing schedule β and random initialization of the student means:

$$p(\beta) = \int_{-2\epsilon e^{-I(\beta)}}^{2\epsilon e^{-I(\beta)}} \mathrm{d}x \sqrt{\frac{d}{4\pi}} e^{-dx^2/4}, \tag{3.9}$$

where $I(\beta) = \int_0^{t_1} ds \sqrt{\frac{R^2(1-\kappa)\beta(s)}{2\pi}}$ and t_1 is the transition time from the high to low temperature regime, defined implicitly by $\beta(t_1) = \alpha R^{-2}$. Details for the derivation of this estimate are provided in Appendix D. For the exponential schedule eq. (2.4), neglecting the initial transient of the variance, $\sigma^2 = ((1-\kappa)\beta)^{-1}$ and

$$I(\beta_i, t_0) = \sqrt{\frac{2}{\pi}}\frac{\sqrt{1-\kappa}}{\kappa}(\sqrt{\alpha} - \sqrt{R^2\beta_i}), \tag{3.10}$$

from which we compute a theoretical mode collapse probability that exhibits excellent agreement with the experimental results for $\alpha \approx 0.85$ (see isoprobability line of fig. 2 (left) and fig. 5 in the

Appendix). In the limit of high initial temperature $R^2\beta_i \to 0$, the integral — and consequently the mode collapse probability — depends solely on the annealing rate $\kappa = \log(1/\beta_i)/t_0$. Namely, the probability scales as the inverse of the annealing rate: $I(\beta_i, t_0) \sim \sqrt{\frac{2}{\pi}}\frac{\sqrt{(1-\kappa)\alpha}}{\kappa}$. This implies that for sufficiently high initial temperatures, the annealing rate emerges as the relevant parameter to determine annealing success. Figure 7 in Appendix confirms that this asymptotic behavior is already at play in the preliminary experiments: as the initial temperature increases, the curves rapidly converge toward this asymptotic regime where the mode collapse probability is dictated purely by the annealing rate.

The success of annealing depends on maximizing the integral $I(\beta_i, t_0)$, which has one global maximum (more details in Appendix section D.3). Important deviation from this optimum leads to failure in two different ways. If the initial temperature is set too low ($\beta_i > \beta_i^*$), the duration of the high-temperature regime is too short to generate sufficient mode separation and prevent collapse. Conversely, for a fixed annealing time t_0 , raising the initial temperature ($\beta_i < \beta_i^*$) can end up being detrimental because the integrand — equal to the escape speed $\sqrt{R^2(1-\kappa)\beta(s)}$ — remains negligible for a comparatively longer time. Consequently, when the true mode separation R is unknown, the most robust strategy for tuning the annealing schedule is to scale t_0 and β_i in successive trials in a way that ensures that the annealing rate is not increasing.

4 Numerical experiments with normalizing flows

Experimental setup. In this section, we validate our theoretical findings in a practical setting by learning a multimodal target using a highly expressive variational family: RealNVPs normalizing flows [Dinh et al., 2017]. While generative models like RealNVPs offer significant flexibility, high expressivity alone does not resolve mode collapse for multimodal targets. Consequently, annealing is an increasingly common strategy for mitigating it in practice.

To isolate the influence of the annealing schedule, we retain the bimodal Gaussian target distribution (Equation 2.1) in dimension $d = 128$ and employ the exponential annealing schedule defined in eq. (2.4). The student distribution q_θ is parameterized using a RealNVP.

A normalizing flow defines a parametric invertible map $\Phi_\theta : \mathbb{R}^d \to \mathbb{R}^d$ with a tractable Jacobian, optimized to transport samples from a base distribution ρ_0 to the target π. In this implementation, we employ a standard Gaussian base distribution $\rho_0(z) = \mathcal{N}(0, I_d)$ in dimension $d = 128$. Our RealNVP architecture consists of 8 affine coupling layers. The internal scale and translation networks within each coupling block are multi-layer perceptrons (MLPs) with 4 hidden layers and a hidden dimension of $4d = 512$.

The target modes are separated by $R = 3$ with the weight imbalance $w_* = 0.8$. Using the change of variable formula, the loss is estimated as a Monte Carlo estimator of:

$$\mathcal{L}(\theta, \beta) = \beta^{-1} \int_{\mathbb{R}^d} dz \rho_0(z) \log \left[\frac{\rho_0(z) \left| \det \frac{\partial \Phi_\theta}{\partial z}(z) \right|^{-1}}{\pi(\Phi_\theta(z))^\beta} \right]. \tag{4.1}$$

Unlike the setting of a Gaussian mixture student, where the overlap s serves as a direct order parameter, detecting mode collapse in high-dimensional flows requires a derived metric. We utilize the variance of the generated samples projected onto $\hat{\mu} = \mu_*/R$. Specifically, we monitor $\mathcal{V} = \mathrm{Var}_{x\sim q_\theta}[x \cdot \hat{\mu}]$. A collapsed model capturing only a single mode (variance ≈ 1) yields $\mathcal{V} \approx 1$, whereas a model that correctly learns the target mixture yields a significantly higher variance ($\mathcal{V} \approx 2.6$ for our parameters). Accordingly, we define a collapse threshold of $\mathcal{V} < 1.6$.

Results. Figure 3 (left) reports the probability of mode collapse across a grid of initial inverse temperatures β_i and annealing times t_0 on a bidodal Gaussian mixture target. The results exhibit the same phenomenology observed in the Gaussian mixture student analysis in fig. 2 (note that t_0 is measured in iterations here). If the initial temperature is too low (high β_i) or no annealing is applied, mode collapse is inevitable regardless of the model's expressivity. Also, for a fixed computational budget (fixed t_0), there exists an optimal initial temperature that maximizes the probability of success; increasing the temperature beyond this point becomes detrimental, as the annealing rate becomes too fast to resolve mode collapse. These results also hold for targets with more modes: fig. 3 (right) and

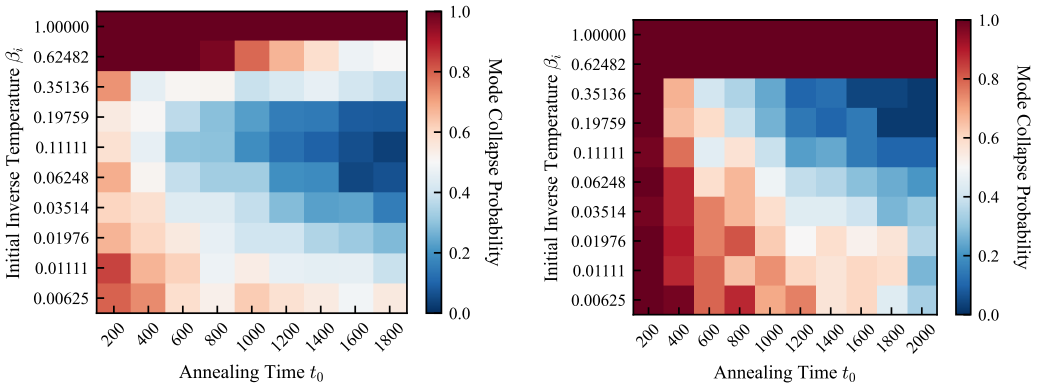


Figure 3: **(Left)** Annealed VI mode collapse probability with RealNVP student. We employ the exponential annealing schedule eq. (2.4) on a bimodal Gaussian mixture target with hyperparameters $d = 128$, $R = 3$ and $w_* = 0.8$. For each (β_i, t_0) mode collapse probabilities are computed on 200 different runs (t_0 in Adam iterations). The generative model is a RealNVP with 8 coupling layers, each containing two MLPs of depth 4 and hidden dimension $4d = 512$ (more details in appendix E.2). **(Right)** RealNVP student on a 5-mode target. The target is a Gaussian mixture with 5 isotropic components. All other hyperparameters are the same as in the left panel. Additional details in Appendix E.3.

appendix E.3 show that the same trade-off between initial inverse temperature and annealing time governs collapse on a 5-mode Gaussian mixture, providing evidence that the phenomenology is not specific to bimodal targets (while remaining within the Gaussian-mixture family).

Moreover, fig. 7 in appendix E.4 confirms that for sufficiently high initial temperatures, the mode collapse probability depends essentially on the annealing rate alone. This recovers for a complex generative model architecture a behavior predicted by our theoretical analysis, which further confirms the relevance of our theoretical insights in practical settings.

These results are of interest for practitioners dealing with multimodal targets, even when the mode separation R is unknown. To reliably enter the mode collapse mitigation parameter region, one cannot simply increase the initial temperature indefinitely. Instead, our findings indicate that a better strategy is to increase the annealing duration t_0 alongside the initial temperature, ensuring that the annealing rate remains sufficiently small to resolve the modes.

5 Conclusion

In this work, we have presented a theoretical investigation of annealing in VI on bimodal Gaussian mixture target distribution. Based on the high-dimensional limit of the variance dynamics and a low-dimensional description of the means, we show that simulated annealing can robustly mitigate mode collapse by controlling the balance between the entropy and cross-entropy terms of the reverse Kullback-Leibler divergence. Characterizing the interplay between initial temperature and annealing time, we show that an increase in the initial temperature should be accompanied by an increase in the annealing time when searching for a combination avoiding mode collapse. Our numerical experiments show that this strategy works with more practical variational families such as RealNVPs.

Overall, our theoretical findings, in agreement with our numerical observations, highlight the role of the annealing rate that controls the probability of mode collapse when starting at high enough temperature. Our work considers a geometric annealing scheme, an interesting perspective for future work is the comparison with other schemes such as Effective Sample Size or χ^2-adapted schemes.

Another interesting direction is to replicate this study for VI with diffusion models, a recent promising framework [Tzen and Raginsky, 2019, Zhang and Chen, 2021, Vargas et al., 2023, Richter and Berner, 2023, Phillips et al., 2024, Noble et al., 2025]. In this approach, VI is performed on a path measure space which may be more robust to mode collapse.

Data availability

A running implementation of the algorithms and all code to reproduce results presented in this work are publicly available at https://github.com/Luijios/annealed_vi_gmms.

Acknowledgments and Disclosure of Funding

This work was supported by the PR[AI]RIE-PSAI project, a French government grant managed by the Agence Nationale de la Recherche under the France 2030 program, reference ANR-23-IACL-0008. BL also acknowledges funding from the Choose France - CNRS AI Rising Talents program.

References

M. S. Albergo, G. Kanwar, and P. E. Shanahan. Flow-based generative models for Markov chain Monte Carlo in lattice field theory. *Physical Review D*, 100(3):034515, Aug. 2019. ISSN 2470-0010, 2470-0029. doi: 10.1103/PhysRevD.100.034515. arXiv:1904.12072 [hep-lat].

A. Barzegar, F. Hamze, C. Amey, and J. Machta. Optimal schedules for annealing algorithms. *Physical Review E*, 109(6):065301, June 2024. doi: 10.1103/PhysRevE.109.065301.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, Apr. 2017. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2017.1285773.

D. Blessing, X. Jia, J. Esslinger, F. Vargas, and G. Neumann. Beyond elbos: a large-scale evaluation of variational methods for sampling. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HkpbnH9lx.

L. Grenioux, M. Noble, and M. Gabrié. Improving the evaluation of samplers on multi-modal targets. In *ICLR Workshop on Frontiers in Probabilistic Inference: Learning Meets Sampling*, Apr. 2025.

D. C. Hackett, C.-C. Hsieh, M. S. Albergo, D. Boyda, J.-W. Chen, K.-F. Chen, K. Cranmer, G. Kanwar, and P. E. Shanahan. Flow-based sampling for multimodal distributions in lattice field theory, July 2021.

B. Hajek. Cooling schedules for optimal annealing. *Mathematics of Operations Research*, 13(2):311–329, 1988. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/3689827.

M. Hibat-Allah, E. M. Inack, R. Wiersema, R. G. Melko, and J. Carrasquilla. Variational Neural Annealing. *Nature Machine Intelligence*, 3(11):952–961, Oct. 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00401-3. arXiv:2101.10154 [cond-mat].

C.-W. Huang, S. Tan, A. Lacoste, and A. C. Courville. Improving Explorability in Variational Inference with Annealed Variational Objectives. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

T. Huix, A. Korba, A. O. Durmus, and E. Moulines. Theoretical Guarantees for Variational Inference with Fixed-Variance Mixture of Gaussians. In *Proceedings of the 41st International Conference on Machine Learning*, pages 20700–20721. PMLR, July 2024. URL https://proceedings.mlr.press/v235/huix24a.html. ISSN: 2640-3498.

Y. Ichikawa and K. Hukushima. Learning dynamics in linear vae: Posterior collapse threshold, superfluous latent space pitfalls, and speedup with kl annealing. In *International Conference on Artificial Intelligence and Statistics*, pages 1936–1944. PMLR, 2024.

Y. Ichikawa and K. Hukushima. High-dimensional asymptotics of vaes: threshold of posterior collapse and dataset-size dependence of rate-distortion curve. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(7):073402, 2025.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, Nov. 1999. ISSN 1573-0565. doi: 10.1023/A:1007665907178.

R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. doi: 10.1137/S0036141096303359. URL https://doi.org/10.1137/S0036141096303359.

S. A. Khandoker, E. M. Inack, and M. Hibat-Allah. Lattice protein folding with variational annealing. *Machine Learning: Science and Technology*, 6(3):035023, Aug. 2025. ISSN 2632-2153. doi: 10.1088/2632-2153/adf376.

T. LeMinh, J. Arbel, T. Möllenhoff, M. E. Khan, and F. Forbes. Natural variational annealing for multimodal optimization. *arXiv preprint arXiv:2501.04667*, 2025.

T. Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Technical report, Microsoft Research, 2005.

W. Nadler and U. H. E. Hansmann. Dynamics and optimal number of replicas in parallel tempering simulations. *Physical Review E*, 76(6):065701, Dec. 2007. doi: 10.1103/PhysRevE.76.065701.

K. A. Nicoli, C. J. Anders, T. Hartung, K. Jansen, P. Kessel, and S. Nakajima. Detecting and mitigating mode-collapse for flow-based sampling of lattice field theories. *Physical Review D*, 108(11):114501, Dec. 2023. doi: 10.1103/PhysRevD.108.114501.

M. Noble, L. Grenioux, M. Gabrié, and A. O. Durmus. Learned Reference-based Diffusion Sampler for multi-modal distributions. In *The Thirteenth International Conference on Learning Representations*, Apr. 2025.

F. Noé, S. Olsson, J. Köhler, and H. Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, Sept. 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaw1147.

M. Petit-Talamon, M. Lambert, and A. Korba. Variational Inference with Mixtures of Isotropic Gaussians. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, Oct. 2025.

A. Phillips, H.-D. Dau, M. J. Hutchinson, V. D. Bortoli, G. Deligiannidis, and A. Doucet. Particle Denoising Diffusion Sampler. In *Proceedings of the 41st International Conference on Machine Learning*, pages 40688–40724. PMLR, July 2024.

D. Rezende and S. Mohamed. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1530–1538. PMLR, June 2015.

L. Richter and J. Berner. Improved sampling via learned diffusions. In *The Twelfth International Conference on Learning Representations*, Oct. 2023.

D. Saad and S. Solla. Dynamics of on-line gradient descent learning for multilayer neural networks. *Advances in neural information processing systems*, 8, 1995.

H. Schopmans and P. Friederich. Temperature-Annealed Boltzmann Generators. In *Forty-Second International Conference on Machine Learning*, June 2025.

R. Soletskyi, M. Gabrié, and B. Loureiro. A theoretical perspective on mode collapse in variational inference. *Machine Learning: Science and Technology*, 6(2):025056, June 2025. ISSN 2632-2153. doi: 10.1088/2632-2153/adde2a.

S. Syed, V. Romaniello, T. Campbell, and A. Bouchard-Cote. Parallel tempering on optimized paths. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10033–10042. PMLR, July 2021.

B. Tzen and M. Raginsky. Theoretical guarantees for sampling and inference in generative models with latent diffusions. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 3084–3114. PMLR, June 2019.

L. Vaitl, K. A. Nicoli, S. Nakajima, and P. Kessel. Gradients should stay on path: Better estimators of the reverse- and forward KL divergence for normalizing flows. *Machine Learning: Science and Technology*, 3(4):045006, Dec. 2022. ISSN 2632-2153. doi: 10.1088/2632-2153/ac9455.

F. Vargas, W. S. Grathwohl, and A. Doucet. Denoising Diffusion Samplers. In *The Eleventh International Conference on Learning Representations*, Sept. 2023.

Y. Wang, C. Chi, and A. R. Dinner. Mitigating mode collapse in normalizing flows by annealing with an adaptive schedule: Application to parameter estimation. *arXiv preprint arXiv:2505.03652*, 2025.

P. Weiss. L'hypothèse du champ moléculaire et la propriété ferromagnétique. *Journal de Physique Théorique et Appliquée*, 6(1):661–690, 1907. ISSN 0368-3893. doi: 10.1051/jphystap:019070060066100.

D. Wu, L. Wang, and P. Zhang. Solving Statistical Mechanics Using Variational Autoregressive Networks. *Physical Review Letters*, 122(8):080602, Feb. 2019. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.122.080602. arXiv:1809.10606 [cond-mat].

Q. Zhang and Y. Chen. Path Integral Sampler: A Stochastic Control Approach For Sampling. In *International Conference on Learning Representations*, Oct. 2021.

First, we recall the general context of Variational Inference. Given a potentially unnormalized target probability distribution π on $\mathbb{R}^d$, and a parametric probability distribution q_θ, the goal is to minimize the reverse Kullback-Leibler objective $\mathcal{D}_{KL}(q_\theta||\pi)$ with respect to θ. In other words, the goal is to find:

$$\theta_* = \underset{\theta\in\Theta}{\text{argmin}}\{\mathcal{D}_{KL}(q_\theta||\pi)\}. \tag{.1}$$

In order to do so, a common strategy is annealing. Introducing an inverse temperature $\beta < 1$, the goal is to relax the objective by minimizing $\mathcal{D}_{KL}(q_\theta||\pi^\beta)$. The inverse temperature is then progressively increased to the target value $\beta = 1$.

Here we considered for q_θ either isotropic Gaussian mixtures, or the probability distribution of a RealNVP normalizing flow. For uniform-weighted Gaussian mixtures,

$$q_\theta = \sum_{i=1}^{K} \frac{1}{K}\mathcal{N}(\mu_i, \sigma_i^2 I_d), \tag{.2}$$

where $\theta = (\mu_1, \dots, \mu_K, \sigma_1, \dots, \sigma_K) \in \mathbb{R}^{K(d+1)}$. In the following section we describe the update rule used to optimize θ.

A JKO scheme for isotropic Gaussian mixtures

Recall the free energy loss defined in eq. (2.2):

$$\mathcal{L}(\theta, \beta) = \beta^{-1}\,\mathcal{D}_{\mathrm{KL}}\left(q_\theta \,\middle\|\, \frac{\pi^\beta}{\mathcal{Z}_\beta}\right). \tag{A.1}$$

In numerical experiments with Gaussian mixtures, we use a JKO scheme [Jordan et al., 1998] with a Wasserstein distance adapted to isotropic Gaussians mixtures that has beein introduced by [Petit-Talamon et al., 2025]. It results in the following discrete update rules, when minimizing the free energy with learning rate η:

$$\mu_{k+1} = \mu_k - \eta\beta^{-1}\nabla_{\mu_k}\mathcal{D}_{KL}(q_{\theta_k}||\pi^\beta), \tag{A.2}$$

$$\sigma_{k+1}^2 = \left(1 - \frac{2\eta\beta^{-1}}{d}\nabla_{\sigma_k^2}\mathcal{D}_{KL}(q_{\theta_k}||\pi^\beta)\right)^2 \sigma_k^2, \tag{A.3}$$

where μ and σ respectively denote μ_1 or μ_2, and σ_1 or σ_2. This update rule is equivalent to standard gradient descent update on μ_k and σ_k (not σ_k^2) with a specific scaling between the learning rate of the means and the learning rate of the standard deviations. In fact:

$$\begin{aligned}\sigma_{k+1} &= \left(1 - \frac{2\eta\beta^{-1}}{d}\nabla_{\sigma_k^2}\mathcal{D}_{KL}(q_\theta||\pi^\beta)\right)\sigma_k \\ &= \sigma_k - \frac{\eta\beta^{-1}}{d}2\sigma_k\nabla_{\sigma_k^2}\mathcal{D}_{KL}(q_\theta||\pi^\beta) = \sigma_k - \frac{\eta\beta^{-1}}{d}\nabla_{\sigma_k}\mathcal{D}_{KL}(q_\theta||\pi^\beta).\end{aligned} \tag{A.4}$$

The JKO scheme implies the $1/d$ factor in front of the derivative with respect to the variances. In terms of the loss the update rule is:

$$\mu_{k+1} = \mu_k - \eta\nabla_{\mu_k}\mathcal{L}(\theta_k, \beta), \tag{A.5}$$

$$\sigma_{k+1} = \sigma_k - \frac{\eta}{d}\nabla_{\sigma_k}\mathcal{L}(\theta_k, \beta). \tag{A.6}$$

The corresponding gradient flow limit is:

$$\dot{\mu}(t) = -\nabla_\mu\mathcal{L}(\theta(t), \beta), \tag{A.7}$$

$$\dot{\sigma}(t) = -\frac{1}{d}\nabla_\sigma\mathcal{L}(\theta(t), \beta), \tag{A.8}$$

with $t = \eta k$.

B Additional preliminary experiments

fig. 4 completes fig. 1 and illustrates additional learning strategies. Initial student means are the same as the one of fig. 1. Frozen variances (first row) lead to mode collapse, even if the variance is set to $1/10R^2$.

Interestingly, the variances rapidly converge to the target value of 1 even if we freeze the means (second row), in agreement with the high-dimensional decoupling of the variance dynamics (proposition 1).

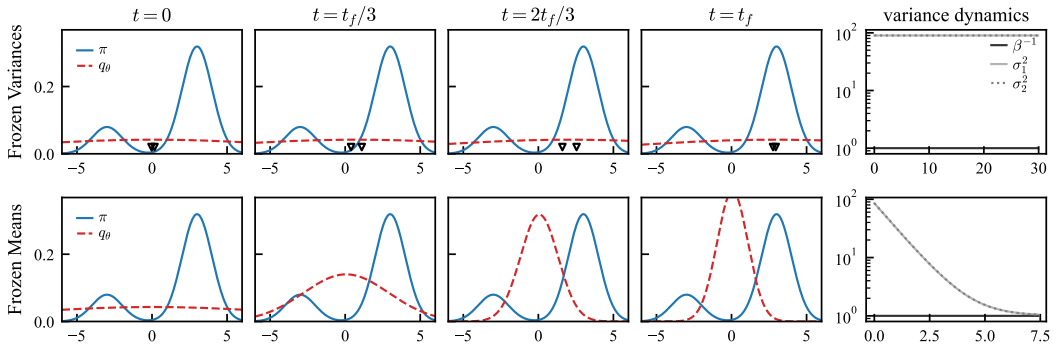


Figure 4: *Additional preliminary experiments.* The first 4 columns show the marginal density along the direction of μ_* of both π and q_θ, at 4 different training stages. The fifth column shows the variance dynamics. Each row represents a different scenario; from top to bottom: non learned variances with $\sigma_{1,2}^2 = 1/10R^2$ fixed, non learned means with $\sigma_{1,2}^2$ initialized to $1/10R^2$. Black triangles on the first row indicate the projection of $\mu_{1,2}$ along the axis of μ_*. Hyperparameters are: $d = 512, R = 3, w_* = 0.8, w_1 = 0.5$ and time discretization step $\eta = 0.05$.

C Gradient flow dynamics and low dimensional dynamical system

In the following, we derive the analytical description of annealed Variational Inference with bimodal Gaussians. By definition, the reverse KL divergence between an isotropic Gaussian mixture with K components and an arbitrary target distribution π at inverse temperature β is:

$$\mathcal{D}_{KL}(q_\theta||\pi^\beta) = \sum_{i=1}^{K} \frac{1}{K} \int_{\mathbb{R}^d} dx \frac{e^{-(x-\mu_i)^2/2\sigma_i^2}}{(2\pi\sigma_i^2)^{d/2}} \log\left(\frac{\sum_{j=1}^{K} \frac{1}{K} \frac{e^{-(x-\mu_j)^2/2\sigma_j^2}}{(2\pi\sigma_j^2)^{d/2}}}{\pi(x)^\beta} \right). \tag{C.1}$$

The goal of this section is to study the gradient flow dynamics related to this loss in the particular case when q_θ and π are bimodal gaussian mixtures.

$$\pi(x) = w_* \mathcal{N}(x|\mu_*, I_d) + (1-w_*)\mathcal{N}(x| - \mu_*, I_d) \tag{C.2}$$

$$q_\theta(x) = w_1 \mathcal{N}(x|\mu_1, \sigma_1^2 I_d) + (1-w_1)\mathcal{N}(x|\mu_2, \sigma_2^2 I_d) \tag{C.3}$$

In what follows, we derive the high-dimensional limit of the variance dynamics (section C.1) and the reduced dynamics of the means (section C.2).

C.1 Variance dynamics in high dimensions

We consider the following gradient flow for the variances, given in eq. (2.3) and derived from the JKO scheme (Appendix A):

$$\dot{\sigma}_c(t) = -\frac{1}{d} \nabla_{\sigma_c} \mathcal{L}(\theta(t), \beta(t)). \tag{C.4}$$

This dynamics is untractable in general, however, it takes a much simpler form in high dimensions, as shown by the following propositions.

Proposition 2. Assume $\|\mu_1\| = \|\mu_2\| = \|\mu_*\| = R$, with R independent of d. Then the loss $\mathcal{L}$ only depends on the variances $\sigma_{1,2}$, the overlap $s = \langle \mu_1, \mu_2 \rangle / R^2$ between the two student means and their respective overlaps $m_i = \langle \mu_i, \mu_* \rangle / R^2$ with the teacher mean. In the high-dimensional limit,

$$\lim_{d\to\infty} \frac{1}{d}\mathcal{L}(\sigma_1, \sigma_2, s, m_1, m_2, \beta) = \sum_{i\in\{1,2\}} w_i \left[-\frac{\beta^{-1}}{2} \log(2\pi e \sigma_i^2) + \frac{\beta^{-1}}{2} \log(2\pi\beta^{-1}) + \frac{1}{2}\sigma_i^2 \right] \tag{C.5}$$

Moreover, the above convergence is uniform:

$$\sup \left| \frac{1}{d}\mathcal{L}(\sigma_1, \sigma_2, s, m_1, m_2, \beta) - \sum_{i\in\{1,2\}} w_i \left[-\frac{\beta^{-1}}{2} \log(2\pi e \sigma_i^2) + \frac{\beta^{-1}}{2} \log(2\pi\beta^{-1}) + \frac{1}{2}\sigma_i^2 \right] \right| = O\left(\frac{1}{d}\right), \tag{C.6}$$

where the supremum is taken over $(\sigma_1, \sigma_2, s, m_1, m_2) \in [\sigma_{\min}, \sigma_{\max}]^2 \times [-1, 1]^3$, and $\beta \in [\beta_{\min}, \beta_{\max}]$. $0 < \sigma_{\min} < \sigma_{\max}$ and $0 < \beta_{\min} < \beta_{\max}$ can be arbitrary.

Proof. We consider here the reverse KL divergence loss, where $Z(\beta) = \int_{\mathbb{R}^d} \pi(x)^\beta \, \mathrm{d}x$ normalizes π^β:

$$\mathcal{L}(\sigma_1, \sigma_2, s, m_1, m_2) = \beta^{-1}\mathbb{E}_{Y\sim q_\theta}[\log q_\theta(Y)] - \mathbb{E}_{Y\sim q_\theta}[\log \pi(Y)] + \beta^{-1} \log Z(\beta). \tag{C.7}$$

We treat the three terms separately. Denote $k_\sigma(x) = (2\pi\sigma^2)^{-d/2} e^{-\frac{\|x\|^2}{2\sigma^2}}$, so that $q_\theta(y) = \sum_{j\in\{1,2\}} w_j k_{\sigma_j}(y - \mu_j)$.

Entropy term. We have:

$$\mathbb{E}_{Y\sim q_\theta}[\log q_\theta(Y)] = \sum_{i\in\{1,2\}} w_i \mathbb{E}_{\mathcal{N}(\mu_i, \sigma_i^2 I_d)} \left[\log \sum_{j\in\{1,2\}} w_j k_{\sigma_j}(Y - \mu_j) \right] \tag{C.8}$$

$$= \sum_{i\in\{1,2\}} w_i \mathbb{E}_{\mathcal{N}(\mu_i, \sigma_i^2 I_d)} \left[\log k_{\sigma_i}(Y - \mu_i) \right] + \sum_{i\in\{1,2\}} w_i \mathbb{E}_{\mathcal{N}(\mu_i, \sigma_i^2 I_d)} \left[\log \sum_{j\in\{1,2\}} w_j \frac{k_{\sigma_j}(Y - \mu_j)}{k_{\sigma_i}(Y - \mu_i)} \right] \tag{C.9}$$

$$= -\sum_{i\in\{1,2\}} \frac{w_i \, d}{2} \log(2\pi e \sigma_i^2) + \sum_{i\in\{1,2\}} w_i \mathbb{E}_{\mathcal{N}(\mu_i, \sigma_i^2 I_d)} \left[\log \sum_{j\in\{1,2\}} w_j \frac{k_{\sigma_j}(Y - \mu_j)}{k_{\sigma_i}(Y - \mu_i)} \right]. \tag{C.10}$$

The first term is proportional to d and gives a contribution to the high-dimensional limit. The second term is bounded by a constant, and hence vanishes once divided by d. Indeed, keeping only the term $j = i$ in the sum gives the lower bound, and Jensen's inequality gives the upper bound:

$$\sum_{i\in\{1,2\}} w_i \log(w_i) < \sum_{i\in\{1,2\}} w_i \mathbb{E}_{\mathcal{N}(\mu_i, \sigma_i^2 I_d)} \left[\log \sum_{j\in\{1,2\}} w_j \frac{k_{\sigma_j}(Y - \mu_j)}{k_{\sigma_i}(Y - \mu_i)} \right]$$

$$\leq \sum_{i\in\{1,2\}} w_i \log \mathbb{E}_{\mathcal{N}(\mu_i, \sigma_i^2 I_d)} \left[\sum_{j\in\{1,2\}} w_j \frac{k_{\sigma_j}(Y - \mu_j)}{k_{\sigma_i}(Y - \mu_i)} \right] \tag{C.11}$$

$$= \sum_{i\in\{1,2\}} w_i \log \int_{\mathbb{R}^d} k_{\sigma_i}(Y - \mu_i) \sum_{j\in\{1,2\}} w_j \frac{k_{\sigma_j}(Y - \mu_j)}{k_{\sigma_i}(Y - \mu_i)} \, \mathrm{d}Y \tag{C.12}$$

$$= 0. \tag{C.13}$$

Combining the terms, the high-dimensional limit of the entropy term is

$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}_{Y\sim q_\theta}[\log q_\theta(Y)] = -\sum_{i\in\{1,2\}} \frac{w_i}{2} \log(2\pi e \sigma_i^2). \tag{C.14}$$

Cross-entropy term. We decompose $y = y^{\parallel} + y^{\perp}$, with $y^{\parallel}$ colinear to μ_* and $y^{\perp}$ in the $(d-1)$-dimensional space orthogonal to μ_*, and we identify $y^{\parallel}$ with its coordinate $\langle y, \mu_* \rangle / R$ along μ_*. We have:

$$\log \pi(y) = -\frac{d}{2}\log(2\pi) - \frac{\|y\|^2}{2} - \frac{R^2}{2} + h(y^{\parallel}), \tag{C.15}$$

where $h(y^{\parallel}) = \log(w_* e^{y^{\parallel} R} + (1-w_*)e^{-y^{\parallel} R})$. For $y \sim \mathcal{N}(\mu_i, \sigma_i^2 I_d)$, $y^{\parallel}$ is a one-dimensional gaussian with variance σ_i^2 whose law does not depend on d, and h grows at most linearly:

$$\log \min(w_*, 1-w_*) - |z|R \le h(z) \le |z|R, \tag{C.16}$$

so that $\mathbb{E}[h(y^{\parallel})]$ is bounded by a constant independent on d (depending on $\sigma_{\max}, R, w_*$). Therefore,

$$\mathbb{E}_{Y\sim q_\theta}[\log \pi(Y)] = \sum_{i\in\{1,2\}} w_i \mathbb{E}_{\mathcal{N}(\mu_i,\sigma_i^2 I_d)}[\log \pi(Y)] = -\frac{d}{2}\log(2\pi) - \sum_{i\in\{1,2\}} \frac{w_i\, d}{2}\sigma_i^2 + O(1), \tag{C.17}$$

and

$$\lim_{d\to\infty} -\frac{1}{d}\mathbb{E}_{Y\sim q_\theta}[\log \pi(Y)] = \sum_{i\in\{1,2\}} \frac{w_i}{2}\left(\log(2\pi) + \sigma_i^2\right). \tag{C.18}$$

The orthogonal component $y^{\perp}$ dominates, and the cross-entropy behaves as if π were the standard Gaussian.

Normalization term. Using the same decomposition,

$$Z(\beta) = (2\pi)^{-\frac{\beta d}{2}} \int_{\mathbb{R}^{d-1}} e^{-\beta\frac{\|y^{\perp}\|^2}{2}}\, \mathrm{d}y^{\perp} \int_{\mathbb{R}} \left(w_* e^{-\frac{(y^{\parallel}-R)^2}{2}} + (1-w_*)e^{-\frac{(y^{\parallel}+R)^2}{2}}\right)^{\beta} \mathrm{d}y^{\parallel} \tag{C.19}$$

$$= (2\pi)^{-\frac{\beta d}{2}} \left(\frac{2\pi}{\beta}\right)^{\frac{d-1}{2}} \int_{\mathbb{R}} \left(w_* e^{-\frac{(y^{\parallel}-R)^2}{2}} + (1-w_*)e^{-\frac{(y^{\parallel}+R)^2}{2}}\right)^{\beta} \mathrm{d}y^{\parallel}. \tag{C.20}$$

By the upper bound on h, the last integral is finite, positive and independent of d. Hence,

$$\lim_{d\to\infty} \frac{1}{d}\log Z(\beta) = \frac{1-\beta}{2}\log(2\pi) - \frac{1}{2}\log\beta. \tag{C.21}$$

Summing the three contributions gives the result. □

The gradient flow eq. (2.3) involves the derivative of the loss with respect to the variances, so we also need the convergence of this derivative. It holds uniformly on compact sets of variances.

Proposition 3. Let $0 < \sigma_{\min} \le \sigma_{\max}$. Then, for $c \in \{1, 2\}$,

$$\sup \left| \frac{1}{d}\frac{\partial \mathcal{L}}{\partial \sigma_c}(\sigma_1, \sigma_2, s, m_1, m_2) - w_c\left(\sigma_c - \frac{1}{\beta\sigma_c}\right)\right| = O\left(\frac{1}{\sqrt{d}}\right), \tag{C.22}$$

where the supremum is taken over $\sigma_1, \sigma_2 \in [\sigma_{\min}, \sigma_{\max}]$, all admissible overlaps s, m_1, m_2 and all values of $\beta \in [\beta_{\min}, \beta_{\max}]$. Note that $w_c\left(\sigma_c - \frac{1}{\beta\sigma_c}\right)$ is the derivative with respect to σ_c of the limit in Proposition 2.

Proof. By symmetry, we take $c = 1$. The normalization term $\log Z(\beta)$ does not depend on σ_1, so we only need to treat the entropy and the cross-entropy terms. Throughout, $\xi \sim \mathcal{N}(0, I_d)$ and $\zeta \sim \mathcal{N}(0, 1)$.

Cross-entropy term. Recall the definition of h in the the proof of proposition 2: $h(z) = \log\left(w_* e^{Rz} + (1-w_*)e^{-Rz}\right)$, so that

$$\log \pi(y) = -\frac{d}{2}\log(2\pi) - \frac{\|y\|^2}{2} - \frac{R^2}{2} + h(y^{\parallel}). \tag{C.23}$$

Since $\mathbb{E}_{\mathcal{N}(\mu_c,\sigma_c^2 I_d)}\|y\|^2 = R^2 + d\sigma_c^2$ and $y^{\parallel} \sim \mathcal{N}(m_c/R, \sigma_c^2)$, this gives the exact expression

$$\mathbb{E}_{y\sim q_\theta}[\log \pi(y)] = -\frac{d}{2}\log(2\pi) - R^2 - \sum_{i\in\{1,2\}} \frac{w_i\, d}{2}\sigma_i^2 + \sum_{i\in\{1,2\}} w_i\, \mathbb{E}\left[h\left(Rm_i + \sigma_i\zeta\right)\right]. \tag{C.24}$$

Since $|h'| \le R$, we can differentiate under the expectation, and

$$\left|\frac{\partial}{\partial\sigma_1}\mathbb{E}\left[h\left(\frac{m_1}{R} + \sigma_1\zeta\right)\right]\right| = \left|\mathbb{E}\left[h'\left(\frac{m_1}{R} + \sigma_1\zeta\right)\zeta\right]\right| \le R\,\mathbb{E}|\zeta| = R\sqrt{\frac{2}{\pi}}. \tag{C.25}$$

Hence, uniformly over all $\sigma_1, \sigma_2 > 0$ and all overlaps,

$$\left|-\frac{1}{d}\frac{\partial}{\partial\sigma_1}\mathbb{E}_{Y\sim q_\theta}[\log \pi(Y)] - w_1\sigma_1\right| \le \frac{R}{\sqrt{2\pi}\, d}. \tag{C.26}$$

Entropy term. We write $\mathbb{E}_{y\sim q_\theta}[\log q_\theta(y)] = w_1\mathbb{E}[\log q_\theta(\mu_1 + \sigma_1\xi)] + w_2\mathbb{E}[\log q_\theta(\mu_2 + \sigma_2\xi)]$ and differentiate. The terms coming from the explicit dependence of q_θ on σ_1 sum to

$$\mathbb{E}_{y\sim q_\theta}[\partial_{\sigma_1}\log q_\theta(y)] = \partial_{\sigma_1}\int q_\theta = 0, \tag{C.27}$$

so that

$$\frac{\partial}{\partial\sigma_1}\mathbb{E}_{y\sim q_\theta}[\log q_\theta(y)] = w_1\mathbb{E}_\xi\left[\xi\cdot\nabla\log q_\theta(\mu_1 + \sigma_1\xi)\right]. \tag{C.28}$$

Denote by

$$p_j(y) = \frac{w_j\, k_{\sigma_j}(y-\mu_j)}{w_1\, k_{\sigma_1}(y-\mu_1) + w_2\, k_{\sigma_2}(y-\mu_2)} \tag{C.29}$$

the responsibilities, so that $p_1 + p_2 = 1$ and $\nabla\log q_\theta(y) = -\sum_{j\in\{1,2\}} p_j(y)\frac{y-\mu_j}{\sigma_j^2}$. With $y = \mu_1 + \sigma_1\xi$ and $\Delta\mu = \mu_1 - \mu_2$, we get

$$\frac{\partial}{\partial\sigma_1}\mathbb{E}_{y\sim q_\theta}[\log q_\theta(y)] = -\frac{w_1 d}{\sigma_1} + w_1\left(\frac{1}{\sigma_1} - \frac{\sigma_1}{\sigma_2^2}\right)\mathbb{E}_\xi\left[p_2(y)\|\xi\|^2\right] - \frac{w_1}{\sigma_2^2}\mathbb{E}_\xi\left[p_2(y)\,\xi\cdot\Delta\mu\right]. \tag{C.30}$$

The first term is the derivative of $-\frac{w_1 d}{2}\log(2\pi e\sigma_1^2)$, and gives the limit $-\frac{w_1}{\beta\sigma_1}$ once multiplied by β^{-1} and divided by d. Since $0 \le p_2(y) \le 1$ and $\|\Delta\mu\| \le 2R$, the last term is bounded by $\frac{R}{\sigma_2^2}\sqrt{\frac{2}{\pi}}$, and vanishes once divided by d.

The second term is of order d a priori: the components cannot be told apart when $\sigma_1 \approx \sigma_2$, but then the prefactor $\frac{1}{\sigma_1} - \frac{\sigma_1}{\sigma_2^2}$ is small. To quantify this, note that $p_2(y) \le \min\left(1, \frac{w_2\, k_{\sigma_2}(y-\mu_2)}{w_1\, k_{\sigma_1}(y-\mu_1)}\right)$, so that, with $\delta = |\sigma_1 - \sigma_2|$ and $S = \sigma_1^2 + \sigma_2^2$,

$$\begin{aligned}\mathbb{E}_{y\sim\mathcal{N}(\mu_1,\sigma_1^2 I_d)}[p_2(y)] &\le \int_{\mathbb{R}^d}\min\left(k_{\sigma_1}(y-\mu_1), \frac{w_2}{w_1}k_{\sigma_2}(y-\mu_2)\right)\mathrm{d}y \le \sqrt{\frac{w_2}{w_1}}\int_{\mathbb{R}^d}\sqrt{k_{\sigma_1}(y-\mu_1)\,k_{\sigma_2}(y-\mu_2)}\,\mathrm{d}y \\ &\le \sqrt{\frac{w_2}{w_1}}\left(\frac{2\sigma_1\sigma_2}{S}\right)^{\frac{d}{2}} \le \sqrt{\frac{w_2}{w_1}}e^{-\frac{d\delta^2}{2S}},\end{aligned} \tag{C.31}$$

where we computed the Gaussian integral and used $\frac{2\sigma_1\sigma_2}{S} = 1 - \frac{\delta^2}{S} \le e^{-\frac{\delta^2}{S}}$. By the Cauchy–Schwarz inequality and $p_2(y)^2 \le p_2(y)$,

$$\mathbb{E}[p_2(y)\|\xi\|^2] \le \sqrt{\mathbb{E}[p_2]}\sqrt{\mathbb{E}\|\xi\|^4} \le (d+1)e^{-\frac{d\delta^2}{4S}}. \tag{C.32}$$

Since $\left|\frac{1}{\sigma_1} - \frac{\sigma_1}{\sigma_2^2}\right| = \frac{\delta(\sigma_1+\sigma_2)}{\sigma_1\sigma_2^2}$ and $\sup_{\delta\geq 0} \delta e^{-\frac{d\delta^2}{4S}} = \sqrt{\frac{2S}{ed}}$, the second term divided by d is bounded by

$$\frac{d+1}{2d}\frac{\sigma_1+\sigma_2}{\sigma_1\sigma_2^2}\sqrt{\frac{2S}{ed}} \leq \frac{4\sigma_{\max}^2}{\sigma_{\min}^3\sqrt{ed}}. \tag{C.33}$$

Combining the three bounds, all of which are uniform over $\sigma_1, \sigma_2 \in [\sigma_{\min}, \sigma_{\max}]$ and all overlaps, gives the result. □

Remark 1. The rate $d^{-1/2}$ comes from the second term of the entropy derivative, and is only attained when $|\sigma_1 - \sigma_2|$ is of order $d^{-1/2}$. Otherwise, this term vanishes (when $\sigma_1 = \sigma_2$) or is exponentially small in d (when $|\sigma_1 - \sigma_2|$ is bounded away from 0), and the rate is d^{-1}. Note also that the bound on the cross-entropy part is uniform over all $\sigma_1, \sigma_2 > 0$, while the bound on the entropy part uses $\sigma_1, \sigma_2 \in [\sigma_{\min}, \sigma_{\max}]$.

C.2 Gradient flow equation for the means

In this section we derive the learning dynamics of the means presented in the analytical analysis section of the main body. As a preliminary, we need to compute and simplify $\mathcal{D}_{KL}(q_\theta||\pi^\beta)$ when q_θ is a Gaussian mixture of the form eq. (.2), in the particular case of *equal variances*: $\sigma = \sigma_1 = \cdots = \sigma_K$. — We restrict to this case since we will eventually consider the high dimensional limit with equal initialization. — The loss simplifies :

$$\beta^{-1}\mathcal{D}_{KL}(q_\theta||\pi^\beta) = -\beta^{-1}\frac{d}{2}\log(2\pi e\sigma^2) + \beta^{-1}\int_{\mathbb{R}^d} dx \frac{e^{-x^2/2}}{(2\pi)^{d/2}} \sum_i \frac{1}{K}\log\left(\sum_{j=1}^K \frac{1}{K} e^{-\frac{x^T(\mu_i-\mu_j)}{\sigma} - \frac{|\mu_i-\mu_j|^2}{2\sigma^2}}\right) \tag{C.34}$$

$$- \int_{\mathbb{R}^d} dx \frac{e^{-x^2/2}}{(2\pi)^{d/2}} \sum_i \frac{1}{K}\log(\pi(\sigma x + \mu_i)). \tag{C.35}$$

$$\tag{C.36}$$

Now, we set $K = 2$ and π as a bimodal isotropic Gaussian mixture and use notations of eq. (2.1). We fix the separation radius $|\mu_*| = R$ and the radii of the student means $= |\mu_k|$. This yields:

$$\beta^{-1}\mathcal{D}_{KL}(q_\theta||\pi^\beta) = -\beta^{-1}\frac{d}{2}\log(2\pi e\sigma^2) - \beta^{-1}\frac{R^2}{\sigma^2} + \beta^{-1}\int_{\mathbb{R}^d} dx \frac{e^{-x^2/2}}{(2\pi)^{d/2}} \sum_{i=1}^2 w_i \log\left(\sum_{j=1}^2 w_j e^{\frac{x^T\mu_j}{\sigma} + \frac{\mu_i^T\mu_j}{\sigma^2}}\right) \tag{C.37}$$

$$+ \frac{d\sigma^2}{2} - \int_{\mathbb{R}^d} dx \frac{e^{-x^2/2}}{(2\pi)^{d/2}} \sum_{i=1}^2 w_i \log\left(w_* e^{(\sigma x+\mu_i)\mu_*} + (1-w_*)e^{-(\sigma x+\mu_i)\mu_*}\right) \tag{C.38}$$

up to an additional constant that does not depend on μ_i nor on σ. Note that the loss only depends on the vectors $\mu_*, \mu_{1,2}$ through the following statistics:

$$m_i = \frac{\mu_i.\mu_x}{R^2}, \quad s = \frac{\mu_1.\mu_2}{R^2}. \tag{C.39}$$

In the following, we alleviate notations by replacing $\int_{\mathbb{R}^d} dx \frac{e^{-x^2/2}}{(2\pi)^{d/2}}[\dots]$ with $\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[\dots]$

Let's first compute the dynamics of the means. By definition, the spherical gradient flow is given by:

$$\dot{\mu_c}(t) = \left(I_d - \frac{\mu_c(t)\mu_c(t)^T}{R^2}\right)\nabla_{\mu_c}\mathcal{L}(\theta(t),\beta(t)) \quad c = 1,2. \tag{C.40}$$

Let's first compute the Euclidean gradient of the KL divergence eq. (C.38) with respect to $\mu_{1,2}$.

$$\begin{aligned}\nabla_{\mu_1}\mathcal{D}_{KL} = w_1\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\frac{1}{\sigma}\frac{w_1(x+2\frac{\mu_1}{\sigma})e^{(x+\frac{\mu_1}{\sigma}).\frac{\mu_1}{\sigma}} + w_2\frac{\mu_2}{\sigma}e^{(x+\frac{\mu_1}{\sigma}).\frac{\mu_2}{\sigma}}}{w_1e^{(x+\frac{\mu_1}{\sigma}).\frac{\mu_1}{\sigma}} + w_2e^{(x+\frac{\mu_1}{\sigma}).\frac{\mu_2}{\sigma}}}\right]\\ +w_2\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\frac{1}{\sigma}\frac{w_1(x+\frac{\mu_2}{\sigma})e^{(x+\frac{\mu_2}{\sigma}).\frac{\mu_1}{\sigma}}}{w_1e^{(x+\frac{\mu_2}{\sigma}).\frac{\mu_1}{\sigma}} + w_2e^{(x+\frac{\mu_2}{\sigma}).\frac{\mu_2}{\sigma}}}\right]\\ -\beta w_1\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\frac{w_*\mu_*e^{(\sigma x+\mu_1).\mu_*} - (1-w_*)\mu_*e^{-(\sigma x+\mu_1).\mu_*}}{w_*e^{(\sigma x+\mu_1).\mu_*} + (1-w_*)e^{-(\sigma x+\mu_1).\mu_*}}\right]\end{aligned} \tag{C.41}$$

Defining the sigmoid function $\text{expit}(t) = (1+e^{-x})^{-1}$, we can rewrite:

$$\begin{aligned}\nabla_{\mu_1}\mathcal{D}_{KL} = \frac{w_1}{\sigma}\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[(x+2\frac{\mu_1}{\sigma})\text{expit}\left((\frac{\mu_1}{\sigma}-\frac{\mu_2}{\sigma}).(x+\frac{\mu_1}{\sigma})+\log\frac{w_1}{w_2}\right)\right.\\ \left.+\frac{\mu_2}{\sigma}\text{expit}\left((\frac{\mu_2}{\sigma}-\frac{\mu_1}{\sigma}).(x+\frac{\mu_1}{\sigma})+\log\frac{w_2}{w_1}\right)\right]\\ +\frac{w_2}{\sigma}\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}\left[(x+\frac{\mu_2}{\sigma})\text{expit}\left((\frac{\mu_1}{\sigma}-\frac{\mu_2}{\sigma}).(x+\frac{\mu_2}{\sigma})+\log\frac{w_1}{w_2}\right)\right]\\ +\beta w_1\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}\left[\mu_*\left(1-2\text{expit}\left(2\mu_*.(\sigma x+\mu_1)+\log\frac{w_*}{1-w_*}\right)\right)\right]\end{aligned} \tag{C.42}$$

It's convenient to eliminate $xg(x)$-like terms using Stein's lemma. Since $x\sim\mathcal{N}(0,I_d)$, we have $\mathbb{E}_x[xg(x)] = \mathbb{E}_x[\nabla g(x)]$, and:

$$\begin{aligned}\nabla_{\mu_1}\mathcal{D}_{KL} = w_1\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\frac{\mu_1}{\sigma^2}\text{expit}\left((\frac{\mu_1}{\sigma}-\frac{\mu_2}{\sigma}).(x+\frac{\mu_1}{\sigma})+\log\frac{w_1}{w_2}\right)\left(3-\text{expit}\left((\frac{\mu_1}{\sigma}-\frac{\mu_2}{\sigma}).(x+\frac{\mu_1}{\sigma})+\log\frac{w_1}{w_2}\right)\right)\right.\\ \left.+\frac{\mu_2}{\sigma^2}\text{expit}\left((\frac{\mu_2}{\sigma}-\frac{\mu_1}{\sigma}).(x+\frac{\mu_1}{\sigma})+\log\frac{w_2}{w_1}\right)^2+\beta\mu_*\left(1-2\text{expit}\left(2\mu_*.(\sigma x+\mu_1)+\log\frac{w_*}{1-w_*}\right)\right)\right]\\ +w_2\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\frac{\mu_1}{\sigma^2}\text{expit}\left((\frac{\mu_1}{\sigma}-\frac{\mu_2}{\sigma}).(x+\frac{\mu_2}{\sigma})+\log\frac{w_1}{w_2}\right)\left(1-\text{expit}\left((\frac{\mu_1}{\sigma}-\frac{\mu_2}{\sigma}).(x+\frac{\mu_2}{\sigma})+\log\frac{w_1}{w_2}\right)\right)\right.\\ \left.+\frac{\mu_2}{\sigma^2}\text{expit}\left((\frac{\mu_1}{\sigma}-\frac{\mu_2}{\sigma}).(x+\frac{\mu_2}{\sigma})+\log\frac{w_1}{w_2}\right)^2\right]\end{aligned} \tag{C.43}$$

Projecting the gradient with eq. (3.4) results in:

$$\begin{aligned}\nabla^{\mathbb{S}}_{\mu_1}\mathcal{D}_{KL} = \frac{\mu_2-\mu_1 s}{\sigma^2}\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[w_1\text{expit}\left((\frac{\mu_2}{\sigma}-\frac{\mu_1}{\sigma}).(x+\frac{\mu_1}{\sigma})+\log\frac{w_2}{w_1}\right)^2\right.\\ \left.+w_2\text{expit}\left((\frac{\mu_1}{\sigma}-\frac{\mu_2}{\sigma}).(x+\frac{\mu_2}{\sigma})+\log\frac{w_1}{w_2}\right)^2\right]\\ +\beta w_1(\mu_*-m_1\mu_1)\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[1-2\text{expit}\left(2\mu_*(\sigma x+\mu_1)+\log\frac{w_*}{(1-w_*)}\right)\right].\end{aligned} \tag{C.44}$$

We notice that the functions inside the expectations depend only on the projections of x in the direction of two axis: $\mu_*.x$ wich has variance R^2, and $(\mu_2-\mu_1).x$ wich has variance $2R^2(1-s)$.

Introducing the functions:

$$\begin{aligned}f(s,\sigma) =&\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[w_1\text{expit}\left(\frac{R}{\sigma}\sqrt{2(1-s)}x+\frac{R^2}{\sigma^2}(s-1)+\log\frac{w_2}{w_1}\right)^2+\right.\\ &\left.+w_2\text{expit}\left(\frac{R}{\sigma}\sqrt{2(1-s)}x+\frac{R^2}{\sigma^2}(s-1)+\log\frac{w_1}{w_2}\right)^2\right]\\ g(m,\sigma) =& \mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[1-2\text{expit}\left(2\sigma Rx+2R^2m+\log\frac{w_*}{1-w_*}\right)\right]\end{aligned} \tag{C.45}$$

we can conveniently write the gradient flow as:

$$\nabla^{\mathbb{S}}_{\mu_1}\mathcal{D}_{KL} = \frac{\mu_2 - \mu_1 s}{\sigma^2} f(s,\sigma) + \beta w_1(\mu_* - m_1\mu_1)g(m_1,\sigma). \tag{C.46}$$

Using the symmetry $\mu_1 \to \mu_2$, $w_1 \to w_2$, we obtain the gradient with respect to μ_2.

$$\nabla^{\mathbb{S}}_{\mu_2}\mathcal{D}_{KL} = \frac{\mu_1 - \mu_2 s}{\sigma^2} f(s,\sigma) + \beta w_2(\mu_* - m_2\mu_2)g(m_2,\sigma). \tag{C.47}$$

Finally, we obtain the learning dynamics of the summary statistics by writing:

$$\begin{aligned}
\dot{m_1} &= \frac{\mu_*.\dot{\mu_1}}{R^2} = -\frac{\mu_*}{R^2}.\beta^{-1}\nabla^{\mathbb{S}}_{\mu_1}\mathcal{D}_{KL} \\
&= -\left[\frac{(m_2 - m_1 s)}{\beta\sigma^2} f(s,\sigma) + w_1(1-m_1^2)g(m_1,\sigma)\right] \qquad \text{(C.48)} \\
\dot{s} &= -\frac{1}{R^2}[\mu_2.\beta^{-1}\nabla^{\mathbb{S}}_{\mu_1}\mathcal{D}_{KL} + \mu_1.\beta^{-1}\nabla^{\mathbb{S}}_{\mu_2}\mathcal{D}_{KL}] \\
&= -\left[\frac{2(1-s^2)}{\beta\sigma^2} f(s,\sigma) + w_1(m_2 - m_1 s)g(m_1,\sigma) + w_2(m_1 - m_2 s)g(m_2,\sigma)\right] \qquad \text{(C.49)}
\end{aligned}$$

Finally, we get:

$$\begin{aligned}
\dot{m}_1 &= -\left[\frac{(m_2 - m_1 s)}{\beta(t)\sigma^2(t)} f(s,\sigma(t)) + w_1(1-m_1^2)g(m_1,\sigma(t))\right] \\
\dot{m}_2 &= -\left[\frac{(m_1 - m_2 s)}{\beta(t)\sigma^2(t)} f(s,\sigma(t)) + w_2(1-m_2^2)g(m_2,\sigma(t))\right] \\
\dot{s} &= -\left[\frac{2(1-s^2)}{\beta(t)\sigma^2(t)} f(s,\sigma(t)) + w_1(m_2 - m_1 s)g(m_1,\sigma(t)) + w_2(m_1 - m_2 s)g(m_2,\sigma(t))\right]
\end{aligned} \tag{C.50}$$

Numerical integration of these coupled ODEs are in excellent agreement with trajectories obtained by SGD on both means and variances (bottom right panel of fig. 2).

C.3 Low- and High-variance limits

Recall that since $\beta(t)\sigma^2(t)$ is of order 1 at any time (eq. (3.3)), high and low variance limits correspond respectively to high and low temperature limits. Precisely, after an order 1 transient, the variance and the inverse temperature are related by an order one constant: $\beta(t)\sigma^2(t) \approx (1-\kappa)^{-1}$.

C.3.1 Low variance limit ($\sigma \ll R$)

Let's begin with f. We denote $r = \frac{R}{\sigma}$, $A = \sqrt{2(1-s)}$, $B = s - 1 = -(1-s) = -\frac{A^2}{2}$, and $C = \log\left(\frac{w_2}{w_1}\right)$. We consider:

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\text{expit}\left(rAx - r^2A^2/2 + C\right)^2\right] \\
=&\mathbb{E}_{x\sim\mathcal{N}(-r^2A^2/2+C,rA)}\left[\text{expit}\left(x\right)^2\right] \\
=&\int_{\mathbb{R}} \frac{dx}{\sqrt{2\pi}rA} e^{-\frac{(x+r^2A^2/2-C)^2}{2r^2A^2}} \text{expit}(x)^2 \\
=&\frac{1}{\sqrt{2\pi}rA} e^{-\frac{(r^2A^2/2-C)^2}{2r^2A^2}} \int_{\mathbb{R}} dx e^{-\frac{x^2}{2r^2A^2}} e^{-x(1/2-C/r^2A^2)} \text{expit}(x)^2 \\
\sim&\frac{1}{\sqrt{2\pi}rA} e^{-r^2A^2/8} e^{C/2} \int_{\mathbb{R}} dx \frac{e^{-x/2}}{(1+e^{-x})^2} \sim \frac{\sqrt{\pi}}{4}\sqrt{\frac{w_2}{w_1}}\frac{1}{r\sqrt{1-s}} e^{\frac{-r^2(1-s)}{4}}
\end{aligned} \tag{C.51}$$

Recombining the different terms, this implies $f(s,\sigma) \sim \frac{\sqrt{\pi}}{2}\frac{\sqrt{w_1 w_2}}{r\sqrt{1-s}} e^{\frac{-r^2(1-s)}{4}}$

Let's now tackle g. Recall $\epsilon = \frac{1}{2R^2}\log\frac{w_*}{1-w_*}$.

$$\begin{aligned} g(m,\sigma) &= \mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[1 - 2\text{expit}\left(2\sigma R x + 2R^2 m + \log\frac{w_*}{1-w_*}\right)\right] \\ &= 1 - 2\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\text{expit}\left(2R^2(\frac{\sigma}{R}x + m + \epsilon)\right)\right] \\ &= 1 - 2\mathbb{E}_{x\sim\mathcal{N}(0,1/r)}\left[\text{expit}\left(2R^2(x + m + \epsilon)\right)\right] \\ &= 1 - 2\text{expit}(2R^2(m+\epsilon)) + \mathcal{O}(r^{-2}) \qquad \text{(C.52)} \\ & \qquad \text{(C.53)} \end{aligned}$$

Comparing the order of magnitude of f and g, the low variance dynamics are dominated by g which comes from the cross-entropy which attracts the means at $\pm\mu_*$.

C.3.2 High variance limit ($\sigma \gg R$)

We derive the high variance asymptotics of the entropic force f and the cross-entropic force g. We denote $r = R/\sigma$ and consider $r \to 0$. Let's begin with f and decompose it in the following way:

$$\begin{aligned} f(s,\sigma) = \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}&\left[w_1\text{expit}\left(r\sqrt{2(1-s)}x + \log\frac{w_2}{w_1}\right)^2 + w_2\text{expit}\left(r\sqrt{2(1-s)}x + \log\frac{w_1}{w_2}\right)^2\right] \\ + \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}&\left[w_1\text{expit}\left(r\sqrt{2(1-s)}x + r^2(s-1) + \log\frac{w_2}{w_1}\right)^2\right. \\ &\left. + w_2\text{expit}\left(r\sqrt{2(1-s)}x + r^2(s-1) + \log\frac{w_1}{w_2}\right)^2\right] \\ - \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}&\left[w_1\text{expit}\left(r\sqrt{2(1-s)}x + \log\frac{w_2}{w_1}\right)^2 + w_2\text{expit}\left(r\sqrt{2(1-s)}x + \log\frac{w_1}{w_2}\right)^2\right] \end{aligned} \qquad \text{(C.54)}$$

For the first line, we make the change of variable $z = rx$. As $r \to 0$, we can use the asymptotic development of the convolution of a function with a gaussian of vanishing variance. As a result, the first line is equal to $w_1\text{expit}(\frac{w_2}{w_1})^2 + w_2\text{expit}(\frac{w_1}{w_2})^2 + \mathcal{O}(r^2) = w_1 w_2 + \mathcal{O}(r^2)$.

The difference between the second line and the third is bounded by Taylor equality at order 1. Using $\left|\text{expit}\left(r\sqrt{2(1-s)}x + r^2(s-1) + \log\frac{w_2}{w_1}\right)^2 - \text{expit}\left(r\sqrt{2(1-s)}x + \log\frac{w_2}{w_1}\right)^2\right| < ||(\text{expit}^2)'||_\infty r^2(s-1) < r^2$, we obtain that the difference is $\mathcal{O}(r^2)$.

As a results, $f(s,\sigma) = w_1 w_2 + \mathcal{O}\left(\frac{R^2}{\sigma^2}\right)$ as $\frac{R}{\sigma} \to 0$, uniformly in s.

Let's continue with g. Recall that $\epsilon = \frac{1}{2R^2}\log\frac{w_*}{1-w_*}$ We decompose it in the following way:

$$\begin{aligned} g(m,\sigma) =& \mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[1 - 2\mathbb{1}_{\{2\sigma R x + 2R^2 m + \log\frac{w_*}{1-w_*} > 0\}}\right] \\ &+ \mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[1 - 2\text{expit}\left(2\sigma R x + 2R^2 m + \log\frac{w_*}{1-w_*}\right)\right] - \mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[1 - 2\mathbb{1}_{\{2\sigma R x + 2R^2 m + \log\frac{w_*}{1-w_*} > 0\}}\right] \\ =& \mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[1 - 2\mathbb{1}_{\{x + r(m+\epsilon) > 0\}}\right] \\ &+ \mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[1 - 2\text{expit}\left(2\sigma R x + 2R^2 m + \log\frac{w_*}{1-w_*}\right)\right] - \mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[1 - 2\mathbb{1}_{\{x + r(m+\epsilon) > 0\}}\right] \end{aligned} \qquad \text{(C.55)}$$

The first term is straightforward and equal to $-\sqrt{\frac{2}{\pi}}r(m+\epsilon)+\mathcal{O}(r^3)$, by Taylor expanding the Gaussian density in 0. The second and the third term can be rewritten as:

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[1-2\text{expit}\left(2\sigma Rx+2R^2m+\log\frac{w_*}{1-w_*}\right)\right]-\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[1-2\mathbb{1}_{\{x+r(m+\epsilon)>0\}}\right]\\
=&2\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[-\text{expit}\left(2\sigma Rx+2R^2m+\log\frac{w_*}{1-w_*}\right)+\mathbb{1}_{\{x+r(m+\epsilon)>0\}}\right]\\
=&2\mathbb{E}_{x\sim\mathcal{N}(2R^2(m+\epsilon),2\sigma R)}\left[-\text{expit}\left(x\right)+\mathbb{1}_{\{x>0\}}\right]
\end{aligned} \tag{C.56}$$

Note that the function $x\to-\text{expit}\left(x\right)+\mathbb{1}_{\{x>0\}}$ is odd an decays exponentially as $|x|\to\infty$. Thus, we can Taylor expand the Gaussian density as a function of $\frac{1}{2\sigma R}$ near 0. Because even terms of this expansion are 0, we get that the first non vanishing term is of order $\mathcal{O}(\frac{r}{\sigma^2})$.

Comparing the order of magnitude of f and g, the high variance dynamics are dominated by f which comes from the entropy and drive μ_1 and μ_2 away.

D Mode collapse probability estimate

D.1 Linearization around $(m_1=m_2=0, s=-1)$

We study here the high temperature ($\beta\ll R^{-2}$) dynamical regime of the dynamics. At initialization, in high dimensions, $m_1\sim\mathcal{N}(0,1/d)$ and $m_2\sim\mathcal{N}(0,1/d)$ are independent.

From the above high variance limits of f and g,

$$f(s,\sigma(t))=w_1w_2+\mathcal{O}(R^2/\sigma^2) \tag{D.1}$$

$$g(m,\sigma(t))=-\sqrt{\frac{2}{\pi}}\frac{R}{\sigma}(m+\epsilon)+\mathcal{O}\left(R^2/\sigma^2\right). \tag{D.2}$$

We deduce that the high temperature dynamics are dominated by the entropic force f, driving the means apart and effectively projecting them onto the symmetric subspace $s=-1$ on a typical time scale $\mathcal{O}(w_1w_2)=\mathcal{O}(1)$. As m_1 and m_2 typical initial values scale as $1/\sqrt{d}$ (considering$\sqrt{d}\gg1$), once $s\approx-1$, the dynamics can be understood by linearizing the coupled ODEs of eq. (3.6) around the point $(m_1=0,m_2=0,s=-1)$. The time evolution of the sum and difference of m_1 and m_2 decouple from s:

$$\begin{aligned}
\dot{m_1}+\dot{m_2}&=-\left[2(1-\kappa)f(-1,\sigma(t))+\frac{1}{2}g'(0,\sigma(t))\right](m_1+m_2)-g(0,\sigma(t))\\
\dot{m_1}-\dot{m_2}&=-\frac{1}{2}g'(0,\sigma(t))(m_1-m_2)\\
\dot{\delta s}&=-\left[4\delta s(1-\kappa)f(-1,\sigma(t))+(m_1+m_2)g(0,\sigma(t))\right].
\end{aligned} \tag{D.3}$$

where $\delta s=s+1$. We can explicitly solve the above linearized dynamical system for an arbitrary time dependant $\beta(t)$.

$$m_1(t)+m_2(t)=(m_1(0)+m_2(0))\,e^{-\int_0^t ds\,a(s)}-\int_0^t ds\,b(s)e^{-\int_s^t du\,a(u)} \tag{D.4}$$

$$m_1(t)-m_2(t)=(m_1(0)-m_2(0))\,e^{-\int_0^t ds\,\frac{1}{2}g'(0,\sigma(s))} \tag{D.5}$$

where $a(s)=2(1-\kappa)f(-1,\sigma(s))+\frac{1}{2}g'(0,\sigma(s))$ and $b(s)=g(0,\sigma(s))$.

Assuming the high temperature limit, $a(s)=w_1w_2+\mathcal{O}(R/\sigma(s))$ and $b(s)=\mathcal{O}(R/\sigma(s))$. Hence, the sum $m_1(t)+m_2(t)$ is of order $\mathcal{O}(R/\sigma(t))$.

On the other hand, the difference m_1-m_2 is unstable since $-g'(0,\sigma(t))=\sqrt{\frac{2}{\pi}}\frac{R}{\sigma(t)}+\mathcal{O}\left(R^2/\sigma(t)^2\right)>0$. Assuming that the transient regime of the variance is over, we have $\beta(t)\sigma^2(t)=(1-\kappa)^{-1}$

$$m_1(t)-m_2(t)=(m_1(0)-m_2(0))\,e^{\int_0^t \mathrm{d}s\,\sqrt{\frac{(1-\kappa)\beta(s)R^2}{2\pi}}}+\mathcal{O}\left(\beta R^2\right). \tag{D.6}$$

D.2 Mode collapse probability

We derive the theoretical estimate of the mode collapse probability, assuming that the transition from the high-temperature regime to the low temperature is really abrupt and occurs at $\beta = \alpha R^{-2}$.

Given an annealing schedule $\beta(t)$, we denote t_1 such that $\beta(t_1) = \alpha R^{-2}$. In other words, we suppose that for $t < t_1$, the dynamics can be described by the high temperature limit $\beta \ll R^{-2}$, and that for $t > t_1$, the dynamics can be described by the low temperature limit. We stress that the transition at $\beta = \alpha R^{-2}$ is a modelling approximation: the crossover between regimes is smooth in the full ODE eq. (3.6), while the sharp-transition formula makes the derivation tractable.

Leveraging the high temperature linearization of the dynamics until $t = t_1$, with a good approximation we have:

$$m_1(t_1) + m_2(t_1) \ll 1 \tag{D.7}$$

$$m_1(t_1) - m_2(t_1) = (m_1(0) - m_2(0))\, e^{\int_0^{t_1} ds \sqrt{\frac{(1-\kappa)\beta(s)R^2}{2\pi}}}. \tag{D.8}$$

For $t > t_1$, the dynamics is described by the low temperature limit, which means that the asymptotic fate of m_1 and m_2 are governed by the sign of $m_1(t_1)+\epsilon$ and $m_2(t_1)+\epsilon$ respectively. Mode collapse is avoided if $\min(m_1(t_1), m_2(t_1)) < -\epsilon$ and $\max(m_1(t_1), m_2(t_1)) > -\epsilon$. The previous condition is achieved if (recall eq. (D.8)):

$$|m_1(t_1) - m_2(t_1)| > 2\epsilon \quad \iff \quad |m_1(0) - m_2(0)| > 2\epsilon e^{-\int_0^{t_1} ds \sqrt{\frac{(1-\kappa)\beta(s)R^2}{2\pi}}}. \tag{D.9}$$

Since $m_1(0) - m_2(0)$ is the sum of two Gaussians of variance $1/d$ (in high dimensions), it is a Gaussian of variance $2/d$. Hence, we can compute the mode collapse probability $p(\beta)$:

$$p(\beta) = \int_{-2\epsilon e^{-I(t_1)}}^{2\epsilon e^{-I(t_1)}} \mathrm{d}x \sqrt{\frac{d}{4\pi}} e^{-\frac{dx^2}{4}} \tag{D.10}$$

where we define:

$$I(\beta) = \int_0^{t_1} \mathrm{d}s \sqrt{\frac{(1-\kappa)\beta(s)R^2}{2\pi}}.$$

D.3 Focus on the exponential annealing schedule

Let's focus on the particular case of an exponential annealing schedule of the form:

$$\beta(t) = \min(\beta_i^{1-t/t_0}, 1). \tag{D.11}$$

Then,

$$\int_0^{t_1} \mathrm{d}s \sqrt{\beta(s)} = \sqrt{\beta_i} \int_0^{t_1} \mathrm{d}s e^{-\frac{s}{2t_0}\log\beta_i} = \frac{2t_0}{\log 1/\beta_i}\left[\frac{\sqrt{\alpha}}{R} - \sqrt{\beta_i}\right]. \tag{D.12}$$

This yields

$$I(\beta_i, t_0) = \sqrt{\frac{2}{\pi}} \frac{\sqrt{(1-\kappa)}}{\kappa}\left[\sqrt{\alpha} - \sqrt{R^2\beta_i}\right], \tag{D.13}$$

where $\kappa = \frac{\log(1/\beta_i)}{t_0}$ is the annealing rate. Reinjecting this integral into eq. (D.10) we obtain a family of theoretical estimates of the mode collapse probability parametrized by α. Since α/R^2 corresponds to the end of the high temperature dynamical regime, we expect it to be of order one. fig. 5 shows the agreement with the experimental curves when setting $\alpha \approx 0.85$. The agreement for high initial temperature (low β_i) is excellent. The agreement worsens as β_i is increased to $1/R^2$, as expected since the high temperature dynamical regime description becomes less and less valid. When β_i is already close to the crossover scale, the high-temperature phase is too short or absent, so $I(\beta_i, t_0)$ remains small even as t_0 changes. Equation (3.9) then predicts collapse probability close to one, explaining the black dashed line in the last panels.

The success of annealing depends on maximazing the integral $I(\beta_i, t_0)$. The dependence with respect to t_0 is increasing and intuitive: the longer the annealing time, the better. The integral has a unique

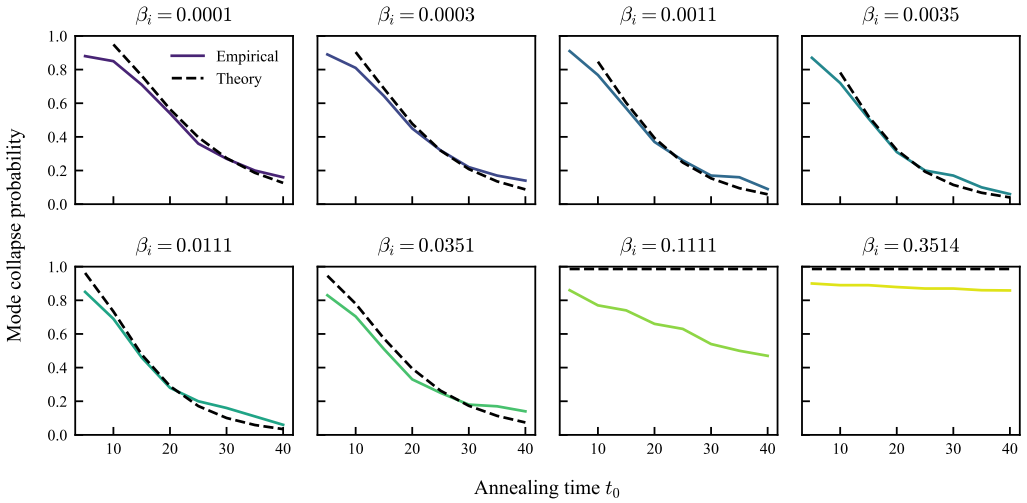


Figure 5: *Collapse probability for different initial temperatures, for Gaussian mixture student with an exponential annealing schedule.* Coloured curves are experimental mode collapse probability computed overs 100 random student mean initialization. The experimental data are the same as in fig. 2. The dashed black line are the theoretical mode collapse probability eq. (3.9), with $\alpha = 0.85$, and are only drawn for $\kappa < 1$. Hyperparameters are $d = 512$, $R = 3$, and $w_* = 0.8$.

global optimum with respect to β_i. In the regime $\kappa \ll 1$,

$$\begin{aligned} \frac{\partial I(\beta_i, t_0)}{\partial \beta_i} = 0 \Longleftrightarrow & \frac{1}{2} R - \frac{R\sqrt{\beta_i} - \sqrt{\alpha}}{\sqrt{\beta_i} \log \beta_i} = 0 \\ & \log(\beta_i/e) \frac{\beta_i}{e} = \frac{-\sqrt{\alpha}}{eR} \\ & \beta_i = e^2 W_0^2 \left(-\frac{\sqrt{\alpha}}{eR} \right) \end{aligned} \tag{D.14}$$

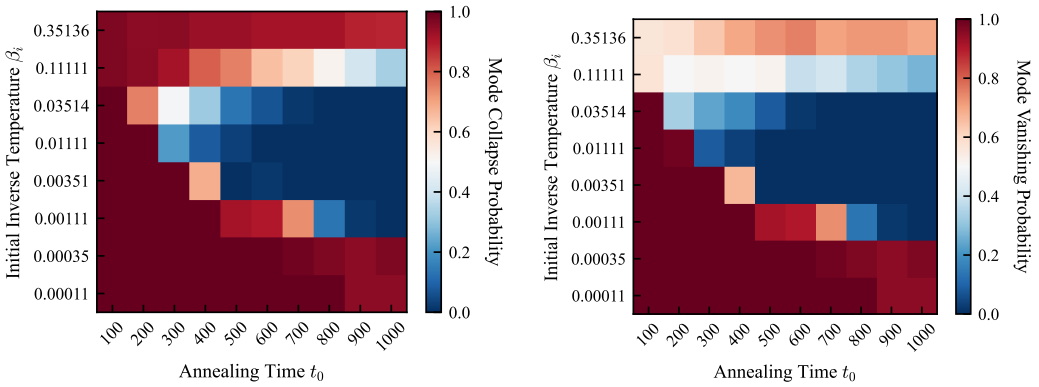
where W_0 is the Lambert W_0 function defined by $W_0(z) e^{W_0(z)} = z$ On the real axis $W_0(x) < x$, and $W_0(x) \sim x$ as $x \to 0$.

Hence the optimal value of the initial temperature β_i is such that $\beta_i < \alpha/R^2$ and $\beta_i \sim \alpha/R^2$ as $R \to \infty$.

We now focus on the limit where the initial temperature is high ($\beta_i \ll R^{-2}$) —which is not optimal but relevant to practical cases when the true R is a priori unknown. We have:

$$I(\beta_i, t_0) \sim \sqrt{\frac{2}{\pi}} \frac{\sqrt{(1-\kappa)\alpha}}{\kappa}. \tag{D.15}$$

Asymptotically, the value of the integral only depends on the annealing rate $\kappa = \log(1/\beta_i)/t_0$. This is expected since the higher the annealing rate, the faster the temperature will go to 1. Notably, it implies that if the initial temperature is high enough, the relevant quantity for the success of annealing is the annealing rate. fig. 7 reports the probability of mode collapse with respect to the annealing rate. Increasing temperature curves rapidly collapse towards the asymptotic regime where mode collapse probability depends only on the annealing rate.



(a) (b)

Figure 6: *Annealed VI mode collapse probability with Gaussian mixture student with learnable weights.* Both the student and the teacher are bimodal gaussian mixtures. The form of the annealing scheme is fixed by eq. (2.4) and different values of the initial temperature β_i and annealing time t_0 are explored. Hyperparamters are $d = 512$, $R = 3$ and $w_* = 0.8$. For each values of (β_i, t_0), the mode collapse probability is obtained using 50 different random student mean initialization (student means are initialized randomly on the sphere of radius R, variances are initialized at $1/\beta_i$ and weights at 0.5). We use Adam optimizer with learning rates $\eta_\mu = 0.003$ for the means, $\eta_{\sigma^2} = 0.02\beta^{-1}$ for the variances and $\eta_w = 0.01$ for the weights.
fig. 6a reports the probability of mode collapse (due to mean alignment or mode vanishing), fig. 6b reports the probability mode vanishing.

E Supplementary experimental results

E.1 Learnable weights

Recall the minimal setting for the study of mode collapse in variational inference:

$$\pi(x) = w_* \mathcal{N}(x|\mu_*, I_d) + (1 - w_*)\mathcal{N}(x| - \mu_*, I_d), \tag{E.1}$$

$$q_\theta(x) = w_1 \mathcal{N}(x|\mu_1, \sigma_1^2 I_d) + w_2 \mathcal{N}(x|\mu_2, \sigma_2^2 I_d). \tag{E.2}$$

To simplify the analytical treatment in the main body, we fix the student weights to be $w_1 = w_2 = 0.5$. In this section we provide experimental results when weights are learnable, that is to say $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, w_1, w_2)$ with constrain $w_1 + w_2 = 1$.

Learnable weights introduce a supplementary failure mode, as mode collapse can also be driven by vanishing weights ([Soletskyi et al., 2025]). Weight vanishing happens when either of one of the 2 student weights irreversibly converge to 0, meaning that the student distribution becomes effectively unimodal. Figure 6a shows the mode collapse probability over a range of initial inverse temperature β_i and annealing time t_0 (t_0 in Adam iterations). In this figure, mode collapse due to both vanishing weights and mean alignment (μ_1 and μ_2 converging to the same target mode) are taken into account. Compared to the fixed weight case (fig. 2), we observe that the model has the same qualitative behavior, with a more abrupt transition between the success or failure of annealing.

The right panel fig. 6b shows the probability of observing vanishing weights. We observe that the low t_0 and low β_i region is prone to mode vanishing. In fact, comparing with fig. 6a, we deduce that mode collapse in this parameter region is mostly due to vanishing weights. We stress that the vanishing weight probability undergoes the same tradeoff identified in this work: the higher the initial temperature, the longer the annealing time has to be in order to prevent it.

E.2 RealNVP training

The architecture used for normalizing flows are RealNVP [Dinh et al., 2017], with 8 affine coupling layers. In each layer, the neural networks parametrizing the scale and the translation are MLPs with

depth 4 and hidden dimension $h = 4d$, if d is the latent space dimension. We use Adam optimizer with learning rate $\eta = 0.002$. Our implementation is provided in the code made available.

E.3 5-mode Gaussian mixture target

This section provides details about the 5-mode Gaussian mixture target distribution used in fig. 3 (right panel). We generated a 5-mode target following this procedure. First, we drew the 5 mode locations uniformly on the sphere of radius $R = 3\sqrt{2}$. Because those 5 modes will be approximately orthogonal in high dimensions, this ensures that the distance between any pair of target mode location is ≈ 6, consistent with the bimodal target used in fig. 3. Then, one of the modes was assigned probability 0.5 while the remaining 4 were assigned probability 0.125.

E.4 Dependence on the annealing rate

fig. 7 shows that the experimental mode collapse probabilities as a function of the annealing rate collapse to a single curve as $\beta_i \to 0$. This confirms that for sufficiently high initial temperatures, the mode collapse probability depends essentially on the annealing rate alone. The dotted black line on the left column is the theoretical ($\beta_i \to 0$) asymptote of the mode collapse probability eq. (3.9). The right panel recovers for a complex generative model architecture the behavior predicted by our theoretical analysis, which further confirms the relevance of our theoretical insights in practical settings. Furthermore, fig. 8 extends this results to a more complex target (5-mode Gaussian mixture).

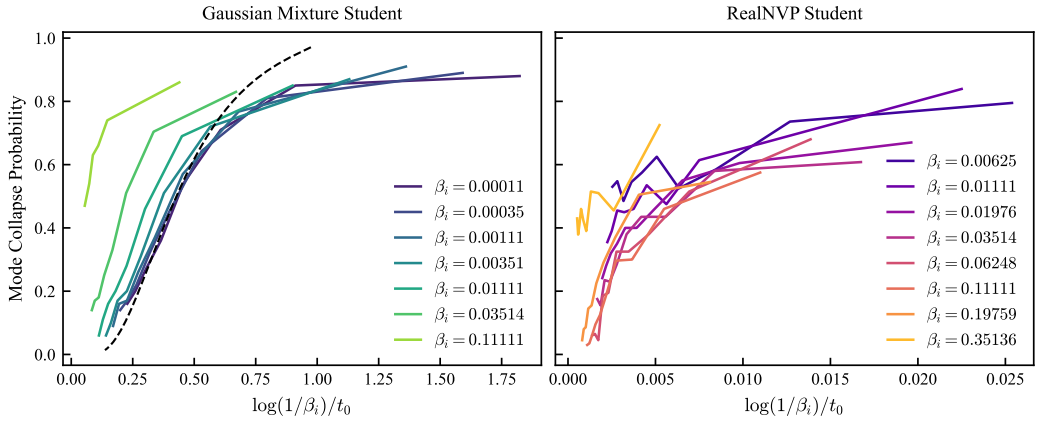


Figure 7: Probability of mode collapse as a function of the annealing rate $\kappa = \log(1/\beta_i)/t_0$, for an exponential annealing schedule given by eq. (2.4). For the Gaussian mixture student the data are the same as fig. 2, hyperparameters are $d = 512, R = 3, w_* = 0.8$. The dashed black line is the high initial temperature ($\beta_i \to 0$) asymptote of the mode collapse probability eq. (3.9) ($\alpha = 0.85$), only drawn for $\kappa < 1$. For the realNVP student the data are the same as fig. 3 (left), hyperparameters are $d = 128, R = 3, w_* = 0.8$. For the Gaussian mixture student t_0 is in gradient-flow time units, for the RealNVP student in Adam iterations, so the values of κ are not comparable between the two panels.

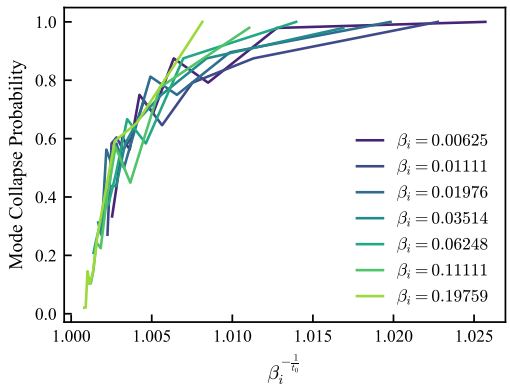


Figure 8: Annealed VI mode collapse probability with RealNVP student on a 5-mode target as a function of e^{κ}. Experimental setting and data are the same as in fig. 3 (right).

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: The abstract and introduction accurately outline the claims regarding mode collapse mitigation, which are rigorously supported by the theoretical analysis in Section 3 and the numerical experiments in Section 4.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: The conclusion acknowledges a limitation by noting that the current analysis is restricted to variational families where the number of modes matches the target modes, and proposes extending this as future work.

   Guidelines:

   - The answer [N/A] 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: The high-dimensional limit of the variance dynamics (Proposition 1) is stated in the main text and proven in Appendix C.1; the derivations of the reduced dynamics and the mode-collapse probability are in Appendices C and D.

   Guidelines:

   - The answer [N/A] 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 experimental setup, including target distribution configurations, hyperparameters, and the RealNVP architecture, is thoroughly detailed in Section 4 and Appendix E.

   Guidelines:

   - The answer [N/A] 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: [Yes]

   Justification: The paper states that code for reproducing the numerical experiments available at https://github.com/Luijios/annealed_vi_gmms.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: Training details, including learning rates, dimensions ($d = 128$ and $d = 512$), and neural network architectures (8 coupling layers, 4-layer MLPs), are specified in the text and appendix E.

Guidelines:

- The answer [N/A] 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: Statistical significance is addressed by reporting mode collapse probabilities aggregated over 100 random initializations for the Gaussian mixture student and 200 runs for the RealNVP student.

   Guidelines:

   - The answer [N/A] 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: [No]

   Justification: While architectural details and hyperparameters are listed, the paper does not explicitly detail the specific compute infrastructure (e.g., GPU types, execution times, or memory) used to run the experiments.

   Guidelines:

   - The answer [N/A] 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: The foundational and theoretical nature of this research complies with the NeurIPS Code of Ethics without raising specialized ethical concerns.

Guidelines:

- The answer [N/A] 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: This work constitutes foundational theoretical research on variational inference and does not present direct, foreseeable negative societal impacts or applications that require specific discussion.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release datasets or models with a high risk for misuse, focusing instead on theoretical derivations and basic generative architectures.

Guidelines:

- The answer [N/A] 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: Existing models and methods, such as RealNVPs and the JKO scheme, are properly cited and credited.

    Guidelines:

    - The answer [N/A] 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: The paper does not introduce novel, standalone assets (like large datasets or pretrained foundation models) that require specialized documentation.

    Guidelines:

    - The answer [N/A] 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 research methodology relies purely on theoretical derivations and computational simulations, involving no crowdsourcing or human subjects.

Guidelines:

- The answer [N/A] 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 research does not involve human subjects, making IRB approval unnecessary.

Guidelines:

- The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Large Language Models were not utilized as a core, original, or non-standard component of the research methodology.

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.