A DERIVATION OF RESULT 2.1

In this appendix, we detail the heuristic derivation of Result (2.1), which provides a sharp asymptotic characterization of the parameters $\hat{c}_t, \hat{w}_t$ of the DAE (9) minimizing the time t risk $\hat{\mathcal{R}}_t$ (10). In the following, the time index $t \in [0,1]$ is considered fixed.

In order to characterize observables depending on the minimizers $\hat{c}_t, \hat{w}_t$ of the risk $\hat{\mathcal{R}}_t$ (10), observe that for any test function $\phi(c_t, w_t)$:

$$\phi(\hat{c}_t, \hat{w}_t) = \lim_{\gamma\to\infty} \frac{1}{Z} \int dw dc \phi(c,w) e^{-\gamma \hat{\mathcal{R}}_t(w,c)}, \tag{28}$$

where the normalization Z is

$$\mathcal{Z}(\mathcal{D}) = \int dc\, dw e^{-\frac{\gamma}{2}\left\|x_1^\mu - \left[c\times(\beta(t)x_1^\mu+\alpha(t)x_0^\mu)+w\varphi\left(w^\top(\beta(t)x_1^\mu+\alpha(t)x_0^\mu)\right)\right]\right\|^2 - \frac{\gamma\lambda}{2}\|w\|^2}. \tag{29}$$

We emphasized the dependence on the train set $\mathcal{D} = \{x_0^\mu, x_t^\mu\}_{\mu=1}^n$. $\ln Z(\mathcal{D})$ can then be studied as a moment generating function, and integrals of the form (28) deduced therefrom. In the following, we therefore seek to establish an asymptotic characterization of $\ln Z(\mathcal{D})$.

An important observation lies in the fact that the argument $w^\top x$ of the activation φ of the DAE is expected in high dimensions $d \to \infty$ to be very large. In particular, we shall self-consistently establish that it is more precisely scaling like $\Theta_d(d)$. As a result, only the asymptotic behaviour in $\pm\infty$ of φ matters, and by assumption $\varphi(w^\top x) \approx \mathrm{sign}(w^\top x)$ asymptotically. We shall therefore self-consistently take $\varphi = \mathrm{sign}$ in the following.

A.1 COMPUTATION OF THE PARTITION FUNCTION

In the following, for clarity, we use the decomposition $x_1^\mu = s^\mu \mu + z^\mu$, introduced below (8) in the main text, with $s^\mu \in \{-1,+1\}$ and $z^\mu \sim \mathcal{N}(0,\sigma^2\mathbb{I}_d)$. Under these notations, the partition function reads:

$$\begin{aligned}\mathcal{Z}(\mathcal{D}) = \int dc\, dw e^{-\frac{\gamma d}{2}\sum_{\mu=1}^n \frac{\|w\|^2}{d}\mathrm{sign}\left(w^\top(\beta(t)(s^\mu\mu+z^\mu)+\alpha(t)x_0^\mu)\right)^2} \\ \times e^{\gamma d \sum_{\mu=1}^n \mathrm{sign}\left(w^\top(\beta(t)(s^\mu\mu+z^\mu)+\alpha(t)x_0^\mu)\right)\frac{w^\top((1-c\beta(t))(s^\mu\mu+z^\mu)-c\alpha(t)x_0^\mu)}{d}} \times e^{-\frac{\gamma\lambda}{2}\|w\|^2} \\ \times e^{-\frac{\gamma d}{2}\sum_{\mu=1}^n\left[(1-\beta(t)c)^2\frac{\|\mu^\mu\|^2+\|z^\mu\|^2+2s^\mu\mu^\top z^\mu}{d}+c^2\frac{\|\alpha(t)x_0^\mu\|^2}{d}+2b(\beta(t)c-1)\frac{s^\mu\mu^\top\alpha(t)x_0^\mu+\eta^{\mu\top}\alpha(t)x_0^\mu}{d}\right]}\end{aligned} \tag{30}$$

Note that we benignly introduced a $1/d$ factor inside the sign function sign. One is now in position to introduce the overlaps

$$q \equiv \frac{\|w\|^2}{2}, \qquad q_\xi^\mu \equiv s^\mu \frac{w^\top x_0^\mu}{d}, \qquad q_\eta^\mu \equiv s^\mu \frac{w^\top z^\mu}{d}, \qquad m \equiv \frac{w^\top \mu}{d}. \tag{31}$$

Note that because of $n = \Theta(1)$ there is a finite number of these such overlaps. Besides, note that our starting assumption that the argument $w^\top x$ of the activation φ is $\Theta_d(d)$ translates into the fact that all these order parameters should be $\Theta_d(1)$, which we shall self-consistently show to be indeed the case. The partition function then reads

$$\begin{aligned}\mathcal{Z}(\mathcal{D}) = \int dc\, dm d\hat{m}\, dq d\hat{q} \prod_{\mu=1}^n d\hat{q}_\eta^\mu d\hat{q}_\eta^\mu\, dq_\xi^\mu e^{\frac{d}{2}\hat{q}q + d\hat{m}m + d\sum_{\mu=1}^n(\hat{q}_\xi^\mu q_\xi^\mu + \hat{q}_\eta^\mu q_\eta^\mu)} \\ \int dw e^{-\frac{\gamma\lambda}{2}\|w\|^2 - \frac{\hat{q}}{2}\|w\|^2 - \left(\hat{m}\mu + \sum_{\mu=1}^n(\hat{q}_\xi^\mu s^\mu x_0^\mu + \hat{q}_\eta^\mu s^\mu z^\mu)\right)^\top w} e^{-\frac{\gamma d}{2}\sum_{\mu=1}^n\left[(1+\sigma^2)(1-\beta(t)c)^2+c^2\alpha(t)^2\right]} \\ e^{-\frac{\gamma d}{2}\sum_{\mu=1}^n\left[q\,\mathrm{sign}\left(\beta(t)(m+q_\eta^\mu)+\alpha(t)q_\xi^\mu\right)^2 - 2\,\mathrm{sign}\left(\beta(t)(m+q_\eta^\mu)+\alpha(t)q_\xi^\mu\right)\left[(1-c\beta(t))(m+q_\eta^\mu)-c\alpha(t)q_\xi^\mu\right]\right]}.\end{aligned} \tag{32}$$

Therefore

$$\mathcal{Z}(\mathcal{D}) = \int dc\, dm d\hat{m}\, dq d\hat{q} \prod_{\mu=1}^{n} d\hat{q}_\xi^\mu dq_\eta^\mu d\hat{q}_\eta^\mu\, dq_\xi^\mu$$
$$e^{\frac{d}{2}\hat{q}q + d\hat{m}m + d\sum_{\mu=1}^{n}(\hat{q}_\xi^\mu q_\xi^\mu + \hat{q}_\eta^\mu q_\eta^\mu) + \frac{d}{2(\gamma\lambda+\hat{q})}\frac{1}{d}\left\|\hat{m}\mu + \sum_{\mu=1}^{n}(\hat{q}_\xi^\mu s^\mu x_0^\mu + \hat{q}_\eta^\mu s^\mu z^\mu)\right\|^2}$$
$$e^{-\frac{\gamma d}{2}\sum_{\mu=1}^{n}\left[q\,\mathrm{sign}\left(\beta(t)(m+q_\eta^\mu)+\alpha(t)q_\xi^\mu\right)^2 - 2\,\mathrm{sign}\left(\beta(t)(m+q_\eta^\mu)+\alpha(t)q_\xi^\mu\right)\left[(1-c\beta(t))(m+q_\eta^\mu) - c\alpha(t)q_\xi^\mu\right]\right]}$$
$$e^{-\frac{\gamma d}{2}\sum_{\mu=1}^{n}\left[(1+\sigma^2)(1-\beta(t)c)^2 + c^2\alpha(t)^2\right] - \frac{d}{2}\ln(\gamma\lambda+\hat{q})}. \tag{33}$$

The last term in the first exponent can be further simplified as

$$\begin{aligned}\frac{1}{d}\left\|\hat{m}\mu + \sum_{\mu=1}^{n}(\hat{q}_\xi^\mu s^\mu x_0^\mu + \hat{q}_\eta^\mu s^\mu z^\mu)\right\|^2 &= \hat{m}^2 + \sum_{\mu=1}^{n}\left[(\hat{q}_\xi^\mu)^2 + (\hat{q}_\eta^\mu)^2\sigma^2\right] \\ &\quad + 2\sum_{\mu=1}^{n}\left[\hat{q}_\xi^\mu s^\mu \frac{\mu^\top x_0^\mu}{d} + \hat{q}_\eta^\mu s^\mu \frac{\mu^\top z^\mu}{d}\right] \\ &\quad + \sum_{\mu,\nu=1}^{n} s^\mu s^\nu \left[\hat{q}_\xi^\mu \hat{q}_\xi^\nu \frac{(x_0^\nu)^\top x_0^\mu}{d} + \hat{q}_\eta^\nu \hat{q}_\eta^\mu \frac{(z^\nu)^\top z^\mu}{d} + \hat{q}_\xi^\mu \hat{q}_\eta^\nu \frac{(z^\nu)^\top x_0^\mu}{d}\right] \\ &= \hat{m}^2 + \sum_{\mu=1}^{n}\left[(\hat{q}_\xi^\mu)^2 + (\hat{q}_\eta^\mu)^2\sigma^2\right] + \mathcal{O}\left(1/\sqrt{d}\right),\end{aligned} \tag{34}$$

with the last line holding with high probability, using the fact that since z, x_0 are two independently drawn standard Gaussian vectors $z^\top x_0/d = \Theta_d(1/\sqrt{d})$ with high probability. Finally,

$$\mathcal{Z}(\mathcal{D}) = \int dc\, dm d\hat{m}\, dq d\hat{q} \prod_{\mu=1}^{n} d\hat{q}_\xi^\mu dq_\eta^\mu d\hat{q}_\eta^\mu\, dq_\xi^\mu$$
$$e^{\frac{d}{2}\hat{q}q + d\hat{m}m + d\sum_{\mu=1}^{n}(\hat{q}_\xi^\mu q_\xi^\mu + \hat{q}_\eta^\mu q_\eta^\mu) + \frac{d}{2(\gamma\lambda+\hat{q})}\left[\hat{m}^2 + \sum_{\mu=1}^{n}\left[(\hat{q}_\xi^\mu)^2 + (\hat{q}_\eta^\mu)^2\sigma^2\right]\right]}$$
$$e^{-\frac{\gamma d}{2}\sum_{\mu=1}^{n}\left[q\,\mathrm{sign}\left(\beta(t)(m+q_\eta^\mu)+\alpha(t)q_\xi^\mu\right)^2 - 2\,\mathrm{sign}\left(\beta(t)(m+q_\eta^\mu)+\alpha(t)q_\xi^\mu\right)\left[(1-c\beta(t))(m+q_\eta^\mu) - c\alpha(t)q_\xi^\mu\right]\right]}$$
$$e^{-\frac{\gamma d}{2}\sum_{\mu=1}^{n}\left[(1+\sigma^2)(1-\beta(t)c)^2 + c^2\alpha(t)^2\right] - \frac{d}{2}\ln(\gamma\lambda+\hat{q})} \tag{35}$$

Since all the terms in the exponent of the integrand scale like d, in the asymptotic limit $d \to \infty$ the integral can be computed using a Laplace approximation.

A.2 Sample-symmetric ansatz

The partition function is given by taking the saddle point in (35). This involves a maximization problem over $4n + 5$ variables. Note that since $n = \Theta_d(1)$, this is a low dimensional – thus a priori tractable – optimization problem, but which nevertheless remains cumbersome. However, the symmetries of the problem make it possible to determine the form of the maximizer, and thus drastically simplify the optimization problem. Note that indeed, asymptotically, the vectors $\mu, \{x_0^\mu, z^\mu\}_{\mu=1}^n$ involved in the definition of the overlaps $m, \{q_\xi^\mu, q_\eta^\mu\}_{\mu=1}^n$ (31) are all mutually asymptotically orthogonal – i.e. they have vanishing cosine similarity. Therefore, the parameters $m, \{q_\xi^\mu, q_\eta^\mu\}_{\mu=1}^n$ can be considered as independent variables. Since furthermore all the samples play interchangeable roles in high dimensions – in that all data points are asymptotically at the same angle with the cluster mean μ, which is the only relevant direction of the problem–, one can look for the saddle point assuming the symmetric ansatz

$$\forall \mu,\ q_\xi^\mu = q_\xi,\ \ \hat{q}_\xi^\mu = \hat{q}_\xi, \tag{36}$$
$$\forall \mu,\ \ q_\eta^\mu = q_\eta,\ \ \hat{q}_\eta^\mu = \hat{q}_\eta. \tag{37}$$

This ansatz is further validated in numerical experiments, when training a DAE with the `Pytorch` implementation of the Adam optimizer. Under this ansatz, the partition function reduces to

$$\mathcal{Z}(\mathcal{D}) = \int dc\, dm d\hat{m}\, dq d\hat{q} d\hat{q}_\xi dq_\eta d\hat{q}_\eta dq_\xi e^{\frac{d}{2}\hat{q}q + d\hat{m}m + dn(\hat{q}_\xi q_\xi + \hat{q}_\eta q_\eta) + \frac{d}{2(\gamma\lambda+\hat{q})}\left[\hat{m}^2 + n(\hat{q}_\xi^2 + \hat{q}_\eta^2\sigma^2)\right]}$$
$$e^{-\frac{\gamma d}{2}n\left[q\,\mathrm{sign}\left(\beta(t)(m+q_\eta)+\alpha(t)q_\xi\right)^2 - 2\,\mathrm{sign}\left(\beta(t)(m+q_\eta)+\alpha(t)q_\xi\right)\left[(1-c\beta(t))(m+q_\eta)-c\alpha(t)q_\xi\right] + (1+\sigma^2)(1-\beta(t)c)^2 + c^2\alpha(t)^2\right]}$$
$$e^{-\frac{d}{2}\ln(\gamma\lambda+\hat{q})} \tag{38}$$

Note that the exponent is now *independent of the dataset* $\mathcal{D}$. In other words, in the regime $d \to \infty, n = \Theta_d(1)$, the log partition function concentrates with respect to the randomness associated with the sampling of the training set. The effective action (log partition function) therefore reads

$$\begin{aligned}\ln \mathcal{Z}(\mathcal{D}) = \underset{c,\hat{q},q,\hat{m},m,\hat{q}_{\eta,\xi},q_{\eta,\xi}}{\mathrm{extr}} & \frac{1}{2}\hat{q}q + \hat{m}m + n(\hat{q}_\xi q_\xi + \hat{q}_\eta q_\eta) + \frac{1}{2(\gamma\lambda+\hat{q})}\left[\hat{m}^2 + n(\hat{q}_\xi^2 + \hat{q}_\eta^2\sigma^2)\right] \\ & - \frac{\alpha}{2}n\left[q\,\mathrm{sign}(\beta(t)(m+q_\eta)+\alpha(t)q_\xi)^2 - 2\,\mathrm{sign}(\beta(t)(m+q_\eta)+\alpha(t)q_\xi)[(1-c\beta(t))(m+q_\eta)-c\alpha(t)q_\xi]\right] \\ & - \frac{\gamma n}{2}\left[(1+\sigma^2)(1-\beta(t)c)^2 + c^2\alpha(t)^2\right] - \frac{1}{2}\ln(\gamma\lambda+\hat{q})\end{aligned} \tag{39}$$

This expression has to be extremized with respect to $c, \hat{q}, q, \hat{m}, m, \hat{q}_{\eta,\xi}, q_{\eta,\xi}$ in the $\gamma \to \infty$ limit. Rescaling the conjugate variables as

$$\gamma\hat{q} \leftarrow \hat{q}, \qquad \gamma\hat{q}_{\eta,\xi} \leftarrow \hat{q}_{\eta,\xi}, \qquad \gamma\hat{m} \leftarrow \hat{m} \tag{40}$$

the action becomes, in the $\gamma \to \infty$ limit (changing for readability the conjugates $\hat{m}, \hat{q}_{\eta,\xi} \to -\hat{m}, -\hat{q}_{\eta,\xi}$):

$$\begin{aligned}\ln \mathcal{Z}(\mathcal{D}) = \underset{c,\hat{q},q,\hat{m},m,\hat{q}_{\eta,\xi},q_{\eta,\xi}}{\mathrm{extr}} & \frac{1}{2}\hat{q}q - \hat{m}m - n(\hat{q}_\xi q_\xi + \hat{q}_\eta q_\eta) + \frac{1}{2(\lambda+\hat{q})}\left[\hat{m}^2 + n(\hat{q}_\xi^2 + \hat{q}_\eta^2\sigma^2)\right] \\ & - \frac{n}{2}\left[q\,\mathrm{sign}(\beta(t)(m+q_\eta)+\alpha(t)q_\xi)^2 - 2\,\mathrm{sign}(\beta(t)(m+q_\eta)+\alpha(t)q_\xi)[(1-c\beta(t))(m+q_\eta)-c\alpha(t)q_\xi]\right] \\ & - \frac{n}{2}\left[(1+\sigma^2)(1-\beta(t)c)^2 + c^2\alpha(t)^2\right]\end{aligned} \tag{41}$$

A.3 Saddle-point equations

The extremization of $\ln \mathcal{Z}(\mathcal{D})$ can be alternatively written as zero-gradient equations on each of the parameters the extremization is carried over, yielding

$$\begin{cases} q = \frac{\hat{m}^2 + n(\hat{q}_\xi^2 + \hat{q}_\eta^2\sigma^2)}{(\lambda+\hat{q})^2} \\ m = \frac{\hat{m}}{\lambda+\hat{q}} \\ q_\xi = \frac{\hat{q}_\xi}{\lambda+\hat{q}} \\ q_\eta = \frac{\hat{q}_\eta\sigma^2}{\lambda+\hat{q}} \end{cases} \qquad \begin{cases} \nu \equiv \beta(t)(m+q_\eta) + \alpha(t)q_\xi \\ \hat{q} = n \\ \hat{m} = n\,\mathrm{sign}(\nu)(1-c\beta(t)) \\ \hat{q}_\eta = \frac{\hat{m}}{n} \\ \hat{q}_\xi = -c\alpha(t)\,\mathrm{sign}(\nu) \\ c = \frac{(1+\sigma^2)\beta(t) - \mathrm{sign}(\nu)(\beta(t)(m+q_\eta)+\alpha(t)q_\xi)}{\alpha(t)^2 + \beta(t)^2(1+\sigma^2)} \end{cases} \tag{42}$$

Note the identity

$$q = m^2 + n(q_\xi^2 + q_\eta^2/\sigma^2) \tag{43}$$

which follows from the asymptotic orthogonality of the vectors and the Pythagorean theorem. This implies in particular that the square norm of $\hat{w}$, as measure by q, is only the sum of its projections along μ (corresponding to m) ξ (corresponding to q_ξ) and η (q_η). Therefore, the norm of the orthogonal projection of $\hat{w}$ with respect to $\mathrm{span}(\mu, \eta, \xi)$ is asymptotically vanishing. In other words, $\hat{w}$ is asymptotically contained in $\mathrm{span}(\mu, \eta, \xi)$.

Remark that the symmetry $m, \hat{m}, q_{\eta,\xi}, \hat{q}_{\eta,\xi} \to -m, -\hat{m}, -q_{\eta,\xi}, -\hat{q}_{\eta,\xi}$ leaves equations (42) unchanged, meaning that if $m, \hat{m}, q_{\eta,\xi}, \hat{q}_{\eta,\xi}$ is a solution to the saddle-point equations, so is $m, \hat{m}, q_{\eta,\xi}, \hat{q}_{\eta,\xi}$. This is due to the symmetry between the clusters in the target density (8), as

$\mu \to -\mu$ yields the same model. As a convention, we can thus suppose without loss of generality $\nu \geq 0$ in (42). (42) then simplifies to

$$\begin{cases} q = \frac{\hat{m}^2 + n(\hat{q}_\xi^2 + \hat{q}_\eta^2 \sigma^2)}{(\lambda + \hat{q})^2} \\ m = \frac{\hat{m}}{\lambda + \hat{q}} \\ q_\xi = \frac{\hat{q}_\xi}{\lambda + \hat{q}} \\ q_\eta = \frac{\hat{q}_\eta \sigma^2}{\lambda + \hat{q}} \end{cases} \qquad \begin{cases} \hat{q} = n \\ \hat{m} = n(1 - c\beta(t)) \\ \hat{q}_\eta = \frac{\hat{m}}{n} \\ \hat{q}_\xi = -\alpha(t)c \\ c = \frac{(1+\sigma^2)\beta(t) - (\beta(t)(m + q_\eta) + \alpha(t) q_\xi)}{\alpha(t)^2 + \beta(t)^2(1+\sigma^2)} \end{cases} . \tag{44}$$

The skip connection strength c thus satisfies the self-consistent equation

$$c = \frac{(1+\sigma^2)\beta(t) - \frac{1}{\lambda+n}(\beta(t)(1 - \beta(t)c)(n + \sigma^2) - \alpha(t)^2 c)}{\alpha(t)^2 + \beta(t)^2(1+\sigma^2)}, \tag{45}$$

which can be solved as

$$c = \frac{\beta(t)(\lambda(1+\sigma^2) + (n-1)\sigma^2)}{\alpha(t)^2(\lambda + n - 1) + \beta(t)^2(\lambda(1+\sigma^2) + (n-1)\sigma^2)} \tag{46}$$

which recovers equation (11) of Result 2.1. Plugging this expression back to (44), and redefining $\sigma^2 q_\eta \leftarrow q_\eta$, yields

$$\begin{cases} m_t = \frac{n}{\lambda+n} \frac{\alpha(t)^2(\lambda+n-1)}{\alpha(t)^2(\lambda+n-1) + \beta(t)^2(\lambda(1+\sigma^2)+(n-1)\sigma^2)} \\ q_t^\eta = \frac{\sigma^2}{\lambda+n} \frac{\alpha(t)^2(\lambda+n-1)}{\alpha(t)^2(\lambda+n-1) + \beta(t)^2(\lambda(1+\sigma^2)+(n-1)\sigma^2)} \\ q_t^\xi = -\frac{1}{\lambda+n} \frac{\alpha(t)\beta(t)(\lambda(1+\sigma^2)+(n-1)\sigma^2)}{\alpha(t)^2(\lambda+n-1) + \beta(t)^2(\lambda(1+\sigma^2)+(n-1)\sigma^2)} \end{cases} \tag{47}$$

We have added subscripts t to emphasize the dependence on the time index t. Note that for $t > 0$, $\nu > 0$, which is self-consistent. For $t = 0$, $\nu = 0$ and the sign function in equation (42) becomes ill-defined, signalling that the extremum of equation (41) ceases to be a critical point (i.e. differentiable). However, one expects the extremum to still be given by the $t = 0$ limit of equation (42), as there is a priori no singularity in the learning problem for $t = 0$. This remark, together with (47), recovers equation (14) from Result 2.1. □

A.4 METRICS

Result 2.1 provides a tight characterization of the skip conneciton strength $\hat{c}_t$ and of the vector $\hat{w}_t$. The performance of the trained DAE $f_{\hat{c}_t, \hat{w}_t}$ (9) as a denoiser can be further quantified with a number of metrics, for which we also provide sharp asymptotic characterizations below, for completeness.

Result A.1. ***(MSE)*** *The test MSE of the learnt denoiser $f_{\hat{c}_t, \hat{w}_t}$ is defined as the test error associated to the risk $\hat{R}_t$ (10)*

$$\mathrm{mse}_t \equiv \mathbb{E}_{x_1 \sim \rho_1, x_0 \sim \rho_0} \left\| f_{\hat{c}_t, \hat{w}_t}(\alpha(t)x_0 + \beta(t)x_1) - x_1 \right\|^2 . \tag{48}$$

In the same asymptotic limit as Result 2.1 in the main text, this metric is sharply characterized by the closed-form formula

$$\mathrm{mse}_t = m_t^2 + n((q_t^\xi)^2 + (q_t^\eta)^2\sigma^2) - 2(1 - \hat{c}_t\beta(t))m_t + (1 - \hat{c}_t\beta(t))^2(1+\sigma^2) + \hat{c}_t^2\alpha(t)^2 \tag{49}$$

where $\hat{c}_t, m_t, q_t^\xi, q_t^\eta$ were defined in Result 2.1. Furthermore, the MSE (48) is lower-bounded by the oracle MSE

$$\mathrm{mse}_t^\star \equiv \mathbb{E}_{x_1 \sim \rho_1, x_0 \sim \rho_0} \left\| f_t^\star(\alpha(t)x_0 + \beta(t)x_1) - x_1 \right\|^2 , \tag{50}$$

where the oracle denoiser follows from an application of Tweedie's formula Efron (2011); Albergo et al. (2023) as

$$f_t^\star(x) = \frac{\beta(t)\sigma^2}{\alpha(t)^2 + \beta(t)^2\sigma^2} x + \frac{\alpha(t)^2}{\alpha(t)^2 + \beta(t)^2\sigma^2} \mu \times \tanh\left(\frac{\beta(t)}{\alpha(t)^2 + \beta(t)^2\sigma^2} \mu^\top x \right). \tag{51}$$

Finally, the oracle MSE $\mathrm{mse}_t^\star$ admits the following asymptotic characterization:

$$\mathrm{mse}_t^\star = \alpha(t)^4\sigma^2 \frac{\alpha(t)^2 + \sigma^2(1 - \alpha(t)^2)}{(\sigma^2\beta(t)^2 + \alpha(t)^2)^2} \tag{52}$$

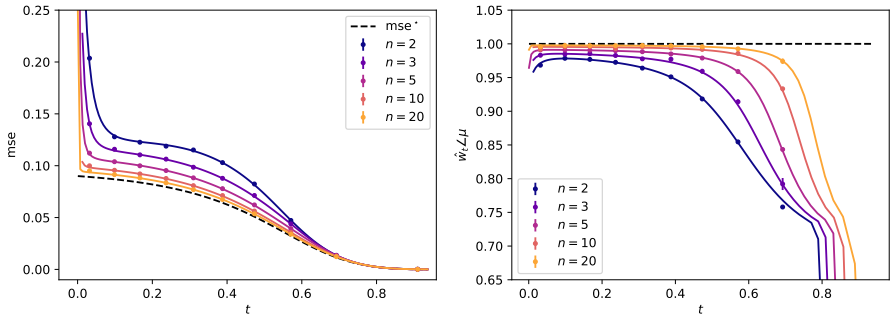

Figure 4: $\sigma = 0.3, \lambda = 0.1, \alpha(t) = \cos(\pi t/2), \beta(t) = \sin(\pi t/2)$. Solid lines: theoretical predictions for the MSE of Result A.1 (left) and the cosine similarity of Result A.2 (right). Different colors correspond to different number of samples n. Dots: numerical simulations, corresponding to training the DAE (9) on the risk (10) using the `Pytorch` implementation of full-batch Adam, with learning rate 0.01 over 2000 epochs and weight decay $\lambda = 0.1$. The experimental points correspond to a single instance of the model, and were collected in dimension $d = 500$. In the left plot, the dashed line represent the oracle baseline (52).

Derivation of Result A.1 We begin by detailing the characterizaiton of the DAE MSE (49):

$$\begin{aligned}
\text{mse}_t &= \frac{1}{d}\mathbb{E}_{x_1,x_0}\langle \left\| x_1 - \left[\hat{c}_t \times (\beta(t)x_1 + \alpha(t)x_0) + \hat{w}_t \,\text{sign}\left(\hat{w}_t^\top(\beta(t)x_1 + \alpha(t)x_0)\right)\right]\right\|^2 \\
&= m_t^2 + n((q_t^\xi)^2 + (q_t^\eta)^2\sigma^2) - 2\,\text{sign}(\beta(t)m_t)(1 - \hat{c}_t\beta(t))m_t + (1 - \hat{c}_t\beta(t))^2(1+\sigma^2) + \hat{c}_t^2\alpha(t)^2 \\
&= m_t^2 + n((q_t^\xi)^2 + (q_t^\eta)^2\sigma^2) - 2(1 - \hat{c}_t\beta(t))m_t + (1 - \hat{c}_t\beta(t))^2(1+\sigma^2) + \hat{c}_t^2\alpha(t)^2 \qquad (53)
\end{aligned}$$

which recovers (49) of Result A.1. (51) follows directly from an application of Tweedie's formula Efron (2011). The associated MSE (52) can be derived as

$$\begin{aligned}
\text{mse}^\star &= \frac{\alpha(t)^4(1+\sigma^2) + \sigma^4\alpha(t)^2(1-\alpha(t)^2)}{(\sigma^2\beta(t)^2 + \alpha(t)^2)^2} + \frac{\alpha(t)^4}{(\sigma^2\beta(t)^2 + \alpha(t)^2)^2}\left[\text{sign}\left(\frac{\beta(t)^2}{\sigma^2\beta(t)^2 + \alpha(t)^2}\right)^2\right] \\
&\quad - \frac{2\alpha(t)^2}{\sigma^2\beta(t)^2 + \alpha(t)^2}\left[\text{sign}\left(\frac{\beta(t)^2}{\sigma^2\beta(t)^2 + \alpha(t)^2}\right)\right] \times \frac{\alpha(t)^2}{\sigma^2 + \alpha(t)^2 - \sigma^2\alpha(t)^2} \\
&= \frac{\alpha(t)^4(1+\sigma^2) + \sigma^4\alpha(t)^2(1-\alpha(t)^2)}{(\sigma^2\beta(t)^2 + \alpha(t)^2)^2} - \frac{\alpha(t)^4}{(\sigma^2\beta(t)^2 + \alpha(t)^2)^2} = \frac{\alpha(t)^4\sigma^2 + \sigma^4\alpha(t)^2(1-\alpha(t)^2)}{(\sigma^2\beta(t)^2 + \alpha(t)^2)^2} \qquad (54)
\end{aligned}$$

which concludes the derivation of Result A.1 □

Result A.2. *The cosine similarity* $\hat{w}_t\angle\mu \equiv \hat{w}_t^\top\mu/\|\hat{w}_t\|\|\mu\|$ *admits the asymptotic characterization*

$$\hat{w}_t\angle\mu = \frac{m_t}{\sqrt{m_t^2 + n((q_t^\xi)^2 + (q_t^\eta)^2\sigma^2}} \qquad (55)$$

where m_t, q_t^ξ, q_t^η *are characterized in Result 2.1.*

Result A.2 follows directly from the definition of the summary statistics (13).

These metrics are plotted in Fig. 4 and contrasted to numerical simulations, corresponding to training the network (9) using the `Pytorch` implementation of full-batch Adam.

B Derivation of Result 3.1

In this Appendix, we detail the heuristic derivation of Result 3.1. Given an initial condition $X_0 \sim \rho_0$, a sample follows the transport (7)

$$\frac{d}{dt}X_t = \left(\dot{\beta}(t)\hat{c}_t + \frac{\dot{\alpha}(t)}{\alpha(t)}(1 - \hat{c}_t\beta(t))\right)X_t + \left(\dot{\beta}(t) - \frac{\dot{\alpha}(t)}{\alpha(t)}\beta(t)\right)\hat{w}_t\,\text{sign}(\hat{w}_t^\top X_t) \qquad (56)$$

driven by the learnt velocity field $\hat{b}$ (6). This follows from Result 2.1 and (7). Taking scalar products with μ, ξ, η,

$$\begin{cases} \frac{d}{dt}\frac{X_t^\top \mu}{d} = \left(\dot\beta(t)\hat c_t + \frac{\dot\alpha(t)}{\alpha(t)}(1-\hat c_t\beta(t))\right)\frac{X_t^\top \mu}{d} + \left(\dot\beta(t) - \frac{\dot\alpha(t)}{\alpha(t)}\beta(t)\right)\operatorname{sign}(\hat w_t^\top X_t)\frac{\hat w_t^\top \mu}{d} \\ \frac{d}{dt}\frac{X_t^\top \xi}{nd} = \left(\dot\beta(t)\hat c_t + \frac{\dot\alpha(t)}{\alpha(t)}(1-\hat c_t\beta(t))\right)\frac{X_t^\top \xi}{nd} + \left(\dot\beta(t) - \frac{\dot\alpha(t)}{\alpha(t)}\beta(t)\right)\operatorname{sign}(\hat w_t^\top X_t)\frac{\hat w_t^\top \xi}{nd} \\ \frac{d}{dt}\frac{X_t^\top \eta}{nd\sigma^2} = \left(\dot\beta(t)\hat c_t + \frac{\dot\alpha(t)}{\alpha(t)}(1-\hat c_t\beta(t))\right)\frac{X_t^\top \eta}{nd\sigma^2} + \left(\dot\beta(t) - \frac{\dot\alpha(t)}{\alpha(t)}\beta(t)\right)\operatorname{sign}(\hat w_t^\top X_t)\frac{\hat w_t^\top \eta}{nd\sigma^2} \end{cases} . \quad (57)$$

It is reasonable to assume the sign $\operatorname{sign}(\hat w_t^\top X_t)$ stays constant during the transport, and therefore takes value ± 1 with equal probability $1/2$, according to the initial condition X_0. This is an heuristic assumption which is further confirmed numerically. Finally, plugging the definitions (16) and (13) in (57), one reaches

$$\begin{cases} \frac{d}{dt}M_t = \left(\dot\beta(t)\hat c_t + \frac{\dot\alpha(t)}{\alpha(t)}(1-\hat c_t\beta(t))\right)M_t + \left(\dot\beta(t) - \frac{\dot\alpha(t)}{\alpha(t)}\beta(t)\right)m_t \\ \frac{d}{dt}Q_t^\xi = \left(\dot\beta(t)\hat c_t + \frac{\dot\alpha(t)}{\alpha(t)}(1-\hat c_t\beta(t))\right)Q_t^\xi + \left(\dot\beta(t) - \frac{\dot\alpha(t)}{\alpha(t)}\beta(t)\right)q_t^\xi \\ \frac{d}{dt}Q_t^\eta = \left(\dot\beta(t)\hat c_t + \frac{\dot\alpha(t)}{\alpha(t)}(1-\hat c_t\beta(t))\right)Q_t^\eta + \left(\dot\beta(t) - \frac{\dot\alpha(t)}{\alpha(t)}\beta(t)\right)q_t^\eta \end{cases} , \quad (58)$$

which recovers equation (15) of Result 3.1. Noting that $\hat w_t \in \operatorname{span}(\mu, \xi, \eta)$ (see Result 2.1), the differential equation (56) becomes, for the orthogonal component $X_t^\perp \in \operatorname{span}(\mu, \xi, \eta)^\perp$

$$\frac{d}{dt}X_t^\perp = \left(\dot\beta(t)\hat c_t + \frac{\dot\alpha(t)}{\alpha(t)}(1-\hat c_t\beta(t))\right)X_t^\perp \quad (59)$$

which recovers (17) of Result 3.1. This can be explicitly solved as

$$X_t^\perp = X_0^\perp e^{\int_0^t \left(\dot\beta(t)\hat c_t + \frac{\dot\alpha(t)}{\alpha(t)}(1-\hat c_t\beta(t))\right)dt} \quad (60)$$

Finally,

$$\begin{aligned} Q_t &\equiv \|X_t\|^2 \\ &= M_t^2 + n(Q_t^\xi)^2 + n\sigma^2(Q_t^\eta)^2 + \|X_t^\perp\|^2 \\ &= M_t^2 + n(Q_t^\xi)^2 + n\sigma^2(Q_t^\eta)^2 + e^{2\int_0^t \left(\dot\beta(t)\hat c_t + \frac{\dot\alpha(t)}{\alpha(t)}(1-\hat c_t\beta(t))\right)dt} \end{aligned} \quad (61)$$

which concludes the derivation of Result 3.1 □

B.1 Derivation of Remark 3.2

The derivation of Remark 3.2 follows identical steps, building on the observation that the discretized flow 19 is explicitly expressed as

$$\begin{aligned} X_{t_{k+1}} = X_{t_k} &+ \delta t_k \left(\dot\beta(t_k)\hat c_{t_k} + \frac{\dot\alpha(t_k)}{\alpha(t_k)}(1-\hat c_{t_k}\beta(t_k))\right)X_{t_k} \\ &+ \delta t_k \left(\dot\beta(t_k) - \frac{\dot\alpha(t_k)}{\alpha(t_k)}\beta(t_k)\right)\hat w_{t_k}\operatorname{sign}(\hat w_{t_k}^\top X_{t_k}) \end{aligned} \quad (62)$$

Taking overlaps with μ, ξ, η yields

$$\begin{cases} \frac{\mu^\top X_{t_{k+1}}}{d} = \frac{\mu^\top X_{t_k}}{d} + \delta t_k\left(\dot\beta(t_k)\hat c_{t_k} + \frac{\dot\alpha(t_k)}{\alpha(t_k)}(1-\hat c_{t_k}\beta(t_k))\right)\frac{\mu^\top X_{t_k}}{d} + \delta t_k\left(\dot\beta(t_k) - \frac{\dot\alpha(t_k)}{\alpha(t_k)}\beta(t_k)\right)\operatorname{sign}(\hat w_{t_k}^\top X_{t_k})\frac{\mu^\top \hat w_{t_k}}{d} \\ \frac{\xi^\top X_{t_{k+1}}}{nd} = \frac{\xi^\top X_{t_k}}{nd} + \delta t_k\left(\dot\beta(t_k)\hat c_{t_k} + \frac{\dot\alpha(t_k)}{\alpha(t_k)}(1-\hat c_{t_k}\beta(t_k))\right)\frac{\xi^\top X_{t_k}}{nd} + \delta t_k\left(\dot\beta(t_k) - \frac{\dot\alpha(t_k)}{\alpha(t_k)}\beta(t_k)\right)\operatorname{sign}(\hat w_{t_k}^\top X_{t_k})\frac{\xi^\top \hat w_{t_k}}{nd} \\ \frac{\eta^\top X_{t_{k+1}}}{nd\sigma^2} = \frac{\eta^\top X_{t_k}}{nd\sigma^2} + \delta t_k\left(\dot\beta(t_k)\hat c_{t_k} + \frac{\dot\alpha(t_k)}{\alpha(t_k)}(1-\hat c_{t_k}\beta(t_k))\right)\frac{\eta^\top X_{t_k}}{nd\sigma^2} + \delta t_k\left(\dot\beta(t_k) - \frac{\dot\alpha(t_k)}{\alpha(t_k)}\beta(t_k)\right)\operatorname{sign}(\hat w_{t_k}^\top X_{t_k})\frac{\eta^\top \hat w_{t_k}}{nd\sigma^2} \end{cases} \quad (63)$$

Like in the continuous case, one makes the heuristic assumption that $\operatorname{sign}(\hat w_{t_k}^\top X_{t_k})$ stays constant along the flow, taking value ± 1 with equal probability, depending on the initial condition. Doing so

yields

$$\begin{cases} M_{t_{k+1}} = M_{t_k} + \delta t_k \left(\dot\beta(t_k)\hat c_{t_k} + \frac{\dot\alpha(t_k)}{\alpha(t_k)}(1-\hat c_{t_k}\beta(t_k))\right) M_{t_k} + \delta t_k \left(\dot\beta(t_k) - \frac{\dot\alpha(t_k)}{\alpha(t_k)}\beta(t_k)\right) m_{t_k} \\ Q^\xi_{t_{k+1}} = Q^\xi_{t_k} + \delta t_k \left(\dot\beta(t_k)\hat c_{t_k} + \frac{\dot\alpha(t_k)}{\alpha(t_k)}(1-\hat c_{t_k}\beta(t_k))\right) Q^\xi_{t_k} + \delta t_k \left(\dot\beta(t_k) - \frac{\dot\alpha(t_k)}{\alpha(t_k)}\beta(t_k)\right) q^\xi_{t_k} \\ Q^\eta_{t_{k+1}} = Q^\eta_{t_k} + \delta t_k \left(\dot\beta(t_k)\hat c_{t_k} + \frac{\dot\alpha(t_k)}{\alpha(t_k)}(1-\hat c_{t_k}\beta(t_k))\right) Q^\eta_{t_k} + \delta t_k \left(\dot\beta(t_k) - \frac{\dot\alpha(t_k)}{\alpha(t_k)}\beta(t_k)\right) q^\eta_{t_k} \end{cases}, \tag{64}$$

which recovers equation (20). Equation (21) follows from equation (62) and the fact that $\hat w_t \in \mathrm{span}(\mu, \xi, \eta)$, see Result 2.1. This recursion can be explicitly solved as

$$X^\perp_{t_{k+1}} = X^\perp_{t_0} \prod_{\ell=0}^{k} \left(1 + \left(\dot\beta(t_\ell) + \tfrac{\dot\alpha(t_\ell)}{\alpha(t_\ell)}(1-\hat c_{t_\ell}\beta(t_\ell))\right)\delta t_\ell\right)^2. \tag{65}$$

Using the fact that $\|X^\perp_{t_0}\|/d = 1$ with high probability and the definition of the summary statistics Q finally yields equation (22). □

B.2 Derivation of Corollary 3.3

As implied by Result 3.1, the mean of the generated mixture is contained in $\mathrm{span}(\mu, \xi, \eta)$ and characterized by the summary statistics M_1, Q^η_1, Q^ξ_1 at time $t = 1$. Furthermore

$$\begin{aligned} \frac{1}{d}\|\hat\mu - \mu\|^2 &= \frac{1}{d}\|\hat\mu\|^2 - 2\frac{\hat\mu^\top \mu}{d} + 1 \\ &= M_1^2 + n(Q^\xi_1)^2 + n\sigma^2(Q^\eta_1)^2 - 2R_1 + 1. \end{aligned} \tag{66}$$

This recovers (23). Equation (24) follows from the definition of the cosine similarity.

The derivation of the $\Theta_n(1/n)$ decay of this distance require more work. The first step lies in the analysis of the exact flow (1).

Remark B.1. ***(exact velocity field)*** *For the target density ρ_1 (8), b is given by Efron (2011); Albergo et al. (2023) as*

$$b(x,t) = \left(\dot\beta(t) - \frac{\dot\alpha(t)}{\alpha(t)}\beta(t)\right)\left(\frac{\beta(t)\sigma^2}{\alpha(t)^2 + \beta(t)^2\sigma^2}x + \frac{\alpha(t)^2}{\alpha(t)^2 + \beta(t)^2\sigma^2}\mu \times \tanh\left(\mu^\top x\right)\right) + \frac{\dot\alpha(t)}{\alpha(t)}x. \tag{67}$$

The formula (67) follows from an application of Tweedie's formula Efron (2011) for the the density (8). Note that with high probability for $x \sim \rho_0$, or for any x such that $\mu^\top x \gg 1$,

$$\tanh\left(\mu^\top x\right) = \mathrm{sign}(\mu^\top x) + o_d(1). \tag{68}$$

One is now in a position to characterize the exact flow (1).

Corollary B.2. ***(Summary statistics for the exact flow)*** *Let $X^\star_t$ be a solution of the exact flow (1) from an initialization $X^\star_0 \sim \rho_0$. Consider the summary statistic*

$$M^\star_t \equiv \frac{\mu^\top X^\star_t}{d}. \tag{69}$$

Asymptotically, $M^\star_t$ is equal with probability $1/2$ to the solution of the differential equation

$$\frac{d}{dt}M^\star_t = \left(\dot\beta(t)c_t + \frac{\dot\alpha(t)}{\alpha(t)}(1 - c_t\beta(t))\right) M^\star_t + \left(\dot\beta(t) - \frac{\dot\alpha(t)}{\alpha(t)}\beta(t)\right)\frac{\alpha(t)^2}{\alpha(t)^2 + \beta(t)^2\sigma^2} \tag{70}$$

and with probability $1/2$ to the opposite thereof. We have introduced

$$c_t \equiv \frac{\beta(t)\sigma^2}{\alpha(t)^2 + \beta(t)^2\sigma^2}. \tag{71}$$

Corollary B.2 follows from equation (67) using a derivation identical to that of Result 3.1, presented in Appendix B, provided the heuristic assumption is made that the tanh can always be approximated by a sign (68) along the flow. To show that the learnt flow 3.1 converges to the exact flow, observe the following scalings:

Remark B.3. *Let $t > 0$ and m_t, q_t^ξ, q_t^η be defined by result 2.1. Then*

$$\left| m_t - \frac{\alpha(t)^2}{\alpha(t)^2 + \beta(t)^2\sigma^2} \right| = \Theta_n(1/n), \quad |\hat{c}_t - c_t| = \Theta_n(1/n), \quad q_t^\xi = \Theta_n(1/n), \quad q_t^\eta = \Theta_n(1/n). \tag{72}$$

These observations immediately imply the following asymptotics, characterizing the difference between the learnt flow (7) and the exact flow (1):

Corollary B.4. ***(Convergence of the learnt flow)*** *Let $X_t^\star$ (resp. X_t) be a solution of the exact flow (1) (resp. learnt flow (7)), from a common initialization $X_0 \sim \rho_0$. Define the following summary statistics:*

$$\epsilon_t^m \equiv \frac{1}{d}\mu^\top(X_t - X_t^\star), \qquad \epsilon_t^\xi \equiv \frac{1}{nd}\xi^\top(X_t - X_t^\star), \qquad \epsilon_t^\eta \equiv \frac{1}{nd\sigma^2}\eta^\top(X_t - X_t^\star) \tag{73}$$

Then with high probability these statistics obey the differential equations

$$\begin{cases} \frac{d}{dt}\epsilon_t^m = \left(\dot\beta(t)c_t + \frac{\dot\alpha(t)}{\alpha(t)}(1 - c_t\beta(t))\right)\epsilon_t^m + \Theta_n(1/n) \\ \frac{d}{dt}\epsilon_t^\xi = \left(\dot\beta(t)c_t + \frac{\dot\alpha(t)}{\alpha(t)}(1 - c_t\beta(t))\right)\epsilon_t^\xi + \Theta_n(1/n) \\ \frac{d}{dt}\epsilon_t^\eta = \left(\dot\beta(t)c_t + \frac{\dot\alpha(t)}{\alpha(t)}(1 - c_t\beta(t))\right)\epsilon_t^\eta + \Theta_n(1/n) \end{cases}, \tag{74}$$

from the initial condition $\epsilon_0^{m,\xi,\eta} = 0$. Therefore at time $t = 1$

$$\epsilon_1^m = \Theta_n(1/n), \qquad \epsilon_1^\xi = \Theta_n(1/n), \qquad \epsilon_1^\eta = \Theta_n(1/n). \tag{75}$$

Corollary B.4 follows from substracting the differential equations governing the *learnt* flow of Result 3.1 and the *true* flow of Corollary B.2, using the scaling derived in Remark B.3. Finally, noting that $M_1^\star = 1$ by definition of the exact flow,

$$\begin{aligned} \frac{1}{d}\|\hat\mu - \mu\|^2 &= \frac{1}{d}\|\epsilon_1^m\mu + \epsilon_1^\xi\xi + \epsilon_1^\eta\eta\|^2 \\ &= (\epsilon_1^m)^2 + n(\epsilon_1^\xi)^2 + n\sigma^2(\epsilon_1^\eta)^2 + O_d(1/\sqrt{d}) \\ &= \Theta_n(1/n). \end{aligned} \tag{76}$$

In the last line, we used Corollary B.4. This concludes the derivation of Corollary 3.3. Fig. 2 (right) gives a PCA visualization of the convergence of the generated density $\hat\rho_1$ to the target density ρ_1 as the number available training samples n accrues. □

C Derivation of Remark 4.1

In this appendix, we analyze the performance of the Bayes-optimal estimator of the cluster mean, defined as the estimator minimizing the average MSE knowing the train set $\mathcal{D} = \{x_0^\mu, x_1^\mu\}_{\mu=1}^n$, the clusters variance σ, but *not* the mean μ. This estimator yields the information-theoretically minimal achievable MSE, and is known to be given by the mean of the posterior distribution over the estimate w of the true mean μ :

$$\begin{aligned} \mathbb{P}(w|\mathcal{D},\sigma) &= e^{-\frac{1}{2}\|w\|^2}\prod_{\mu=1}^n\left[\frac{1}{2}e^{-\frac{1}{2\sigma^2}\|x_1^\mu - w\|^2} + \frac{1}{2}e^{-\frac{1}{2\sigma^2}\|x_1^\mu + w\|^2}\right] \\ &\equiv \frac{1}{Z}e^{-\frac{1}{2\hat\sigma^2}\|w\|^2 + \sum_{\mu=1}^n \ln\cosh\left(\frac{w^\top x_1^\mu}{\sigma^2}\right)}, \end{aligned} \tag{77}$$

where

$$\hat\sigma^2 \equiv \frac{\sigma^2}{n + \sigma^2}. \tag{78}$$

We remind the reader that the prior distribution over the cluster mean is supposed to be the standard Gaussian prior $\mathcal{N}(0, \mathbb{I}_d)$. In high dimensions, statistics associated to the posterior distribution (77) are expected to concentrate. Again, it is useful to study the partition function (normalization) Z to

access some key summary statistics, which will in turn provide a sharp characterization of the vector $\hat{\mu}^\star(\mathcal{D})$ extremizing the posterior $\mathbb{P}(w|\mathcal{D},\sigma)$.

The partition function reads

$$\begin{aligned} Z &= \int dw e^{-\frac{1}{2\hat{\sigma}^2}\|w\|^2+\sum_{\mu=1}^{n}\ln\cosh\left(\frac{w^\top x_1^\mu}{\sigma^2}\right)} \\ &= \int dq d\hat{q} dm d\hat{m} \prod_{\mu=1}^{n} dq_\eta^\mu d\hat{q}_\eta^\mu e^{\frac{d}{2}q\hat{q}+d\sum_{\mu=1}^{n} q_\eta^\mu \hat{q}_\eta^\mu + dm\hat{m}} \\ &\int dw e^{-\frac{\hat{q}}{2}\|w\|^2-\frac{1}{2\hat{\sigma}^2}\|w\|^2-\left(\hat{m}\mu+\sum_{\mu=1}^{n}\hat{q}_\eta^\mu s^\mu z^\mu\right)^\top w} e^{-d\sum_{\mu=1}^{n}\ln\cosh^{1/d}\left[\frac{d}{\sigma^2}\left(m+q_\eta^\mu\right)\right]} \end{aligned} \tag{79}$$

As in Appendix A, we have introduced the summary statistics

$$q_\eta^\mu \equiv s^\mu \frac{w^\top z^\mu}{d}, \qquad\qquad m = \frac{w^\top \mu}{d}. \tag{80}$$

The integral of (79) can be evaluated using a Laplace approximation. Again, we assume the extremizer is realized at the sample-symmetric point

$$\begin{aligned} &\forall 1 \le \mu \le n,\ q_\eta^\mu = q_\eta, \\ &\forall 1 \le \mu \le n,\ \hat{q}_\eta^\mu = \hat{q}_\eta. \end{aligned} \tag{81}$$

The partition function (79) then reduces to

$$\begin{aligned} Z &= \int dq d\hat{q} dq_\eta d\hat{q}_\eta dm d\hat{m} e^{\frac{d}{2}q\hat{q}+dnq_\eta\hat{q}_\eta+dm\hat{m}} \\ &\int dw e^{-\frac{\hat{q}}{2}\|w\|^2-\frac{1}{2\hat{\sigma}^2}\|w\|^2-(\hat{m}\mu+\hat{q}_\eta\eta)^\top w} e^{-d\sum_{\mu=1}^{n}\ln\cosh^{1/d}\left[\frac{d}{\sigma^2}(m+q_\eta)\right]} \\ &= \int dq d\hat{q} dq_\eta d\hat{q}_\eta dm d\hat{m} e^{\frac{d}{2}q\hat{q}+dnq_\eta\hat{q}_\eta+dm\hat{m}} e^{-d\sum_{\mu=1}^{n}\ln\cosh^{1/d}\left[\frac{d}{\sigma^2}(m+q_\eta)\right]} \\ &\frac{1}{(1+\hat{\sigma}^2\hat{q})^{d/2}} e^{\frac{d}{2}\frac{\hat{\sigma}^2}{1+\hat{\sigma}^2\hat{q}}(\hat{m}^2+n\sigma^2\hat{q}_\eta^2)}. \end{aligned} \tag{82}$$

Therefore $\hat{q}_\eta, q_\eta, \hat{m}, m$ must extremize the effective action

$$\Phi = \frac{q\hat{q}}{2} + nq_\eta\hat{q}_\eta + m\hat{m} - \frac{1}{2}\ln\left(1+\hat{\sigma}^2\hat{q}\right) + \frac{\hat{\sigma}^2}{2(1+\hat{\sigma}^2\hat{q})}\left(\hat{m}^2+n\sigma^2\hat{q}_\eta^2\right) + \frac{n}{\sigma^2}|m+q_\eta|, \tag{83}$$

leading to

$$\begin{cases} \hat{q}_\eta = -\frac{1}{\sigma^2} \\ \hat{m} = -\frac{n}{\sigma^2} \end{cases}, \qquad\qquad \begin{cases} q_\eta = -\hat{q}_\eta\sigma^2\hat{\sigma}^2 = \frac{\sigma^2}{n+\sigma^2} \\ m = -\hat{m}\hat{\sigma}^2 = \frac{n}{n+\sigma^2} \end{cases}. \tag{84}$$

Refining $\sigma^2 q_\eta \leftarrow q_\eta$ so that

$$q_\eta \equiv \frac{w^\top \eta}{nd\sigma^2}, \tag{85}$$

as in Remark 4.1, one finally reaches

$$q_\eta = \frac{1}{n+\sigma^2}, \qquad\qquad m = \frac{n}{n+\sigma^2}, \tag{86}$$

Thus, remembering $\hat{\mu}^\star(\mathcal{D}) = \langle w \rangle$ (where the bracket notation denotes averages with respect to the posterior $P(\cdot|\mathcal{D},\sigma)$:

$$\frac{\mu^\top \hat{\mu}^\star(\mathcal{D})}{d} = \left\langle \frac{w^\top \mu}{d} \right\rangle = \frac{n}{n+\sigma^2}, \qquad\qquad \frac{\eta^\top \hat{\mu}^\star(\mathcal{D})}{nd\sigma^2} = \left\langle \frac{w^\top \eta}{nd\sigma^2} \right\rangle = \frac{1}{n+\sigma^2}, \tag{87}$$

using the concentration of the bracketed quantities. Furthermore,

$$\frac{\|\hat{\mu}^\star(\mathcal{D})\|^2}{d} = \frac{1}{d}\|\langle w\rangle\|^2 = \frac{\mu^\top \langle w\rangle}{d} = m. \tag{88}$$

We employed the Nishimori identity (Nishimori, 2001; Iba, 1999).Note further the identity:

$$\frac{\|\hat{\mu}^\star(\mathcal{D})\|^2}{d} = m = m^2 + n\sigma^2 q_\eta^2 = \frac{1}{d}\|m\mu + q_\eta \eta\|^2, \tag{89}$$

which implies that the norm of $\hat{\mu}^\star(\mathcal{D})$ is equal to the norm of its projection in $\mathrm{span}(\mu, \eta)$, which means that asymptotically the former is contained in the latter. One is now in a position to derive the Bayes-optimal MSE of Remark 4.1. With high probability

$$\frac{1}{d}\|\hat{\mu}^\star(\mathcal{D}) - \mu\|^2 = m + 1 - 2m = 1 - m = \frac{\sigma^2}{n + \sigma^2}. \tag{90}$$

This completes the derivation of Remark 4.1 □

D Further settings

D.1 Imbalanced clusters

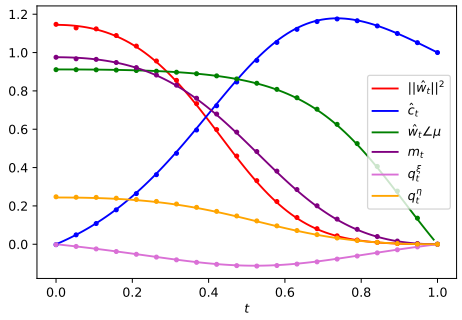

Figure 5: $n = 4, \sigma = 0.9, \lambda = 0.1, \alpha(t) = 1 - t, \beta(t) = t$. Imbalanced mixture with relative weights $\rho = 0.24$ and $\mathbf{1} - \rho = 0.76$. Solid lines: theoretical predictions of Result 2.1: squared norm of the DAE weight vector $\|\hat{\boldsymbol{w}}_t\|^2$ (red), skip connection strength $\hat{c}_t$ (blue) cosine similarity between the weight vector $\hat{\boldsymbol{w}}_t$ and the target cluster mean $\boldsymbol{\mu}$, $\hat{\boldsymbol{w}}_t \angle \boldsymbol{\mu} \equiv \hat{\boldsymbol{w}}_t^\top \boldsymbol{\mu}/\|\boldsymbol{\mu}\|\|\hat{\boldsymbol{w}}_t\|$ (green), components m_t, q_t^ξ, q_t^η of $\hat{\boldsymbol{w}}_t$ along the vectors $\boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\eta}$ (purple, pink, orange). Dots: numerical simulations in $d = 5 \times 10^4$, corresponding to training the DAE (9) on the risk (10) uisng the `Pytorch` implementation of full-batch Adam, with learning rate 0.001 over 20000 epochs and weight decay $\lambda = 0.1$. The experimental points correspond to a single instance of the model.

In this appendix, we address the case of a binary homoscedastic but *imbalanced* mixture

$$\rho_1 = \rho\mathcal{N}(\mu, \sigma^2\mathbb{I}_d) + (1 - \rho)\mathcal{N}(-\mu, \sigma^2\mathbb{I}_d), \tag{91}$$

where $\rho \in (0, 1)$ controls the relative weights of the two clusters. The target density considered in the main text (8) thus corresponds to the special case $\rho = 1/2$.

It is immediate to verify that the derivations presented in Appendices A, B carry through. In other words, Result 2.1, Result 3.1 and Corollary 3.3 still exactly hold. Figure 5 shows that the sharp characterization of Result 2.1 indeed still tightly captures the learning curves of a DAE, trained on *imbalanced* clusters, using the `Pytorch` (Paszke et al., 2019) implementation of the Adam (Kingma & Ba, 2014) optimizer.

An important consequence of this observation is that the generative model will generate a *balanced* density $\hat{\rho}_1$, failing to capture the asymmetry of the target distribution ρ_1. This echoes the findings of Biroli & Mézard (2023) in the related setting of a target ferromagnetic Curie-Weiss model, where they argue that the asymmetry of the ground state can only be learnt for $n \gg d$ samples.

D.2 DAE WITHOUT SKIP CONNECTION

In this appendix, we examine the importance of the skip connection in the DAE architecture (9). More precisely, we consider the generative model parameterized by the DAE *without* skip connection

$$g_{w_t}(x) = w_t \varphi(w_t^\top x) \tag{92}$$

where φ is an activation admitting horizontal asymptots at $+1$ (-1) in $+\infty$ $(-\infty)$. A tight characterization of the learnt weight $\hat{w}_t$ can also straightfowardly be accessed, and is summarized in the following result, which is the equivalent of Result 2.1 for the DAE without skip connection (92)

Result D.1. ***(Sharp characterization of the trained weight of (92))*** *For any given activation φ satisfying $\varphi(x) \xrightarrow{x\to\pm\infty} \pm 1$ and any $t \in [0,1]$, in the limit $d \to \infty$, $n, \|\boldsymbol{\mu}\|^2/d, \sigma = \Theta_d(1)$, the learnt weight vector $\hat{\boldsymbol{w}}_t$ of the DAE without skip connection (92) trained on the loss (10) is asymptotically contained in* $\mathrm{span}(\boldsymbol{\mu}, \boldsymbol{\eta})$ *(in the sense that its projection on the orthogonal space* $\mathrm{span}(\boldsymbol{\mu}, \boldsymbol{\eta})^\perp$ *has asymptotically vanishing norm). The components of $\hat{\boldsymbol{w}}_t$ along each of these two vectors is given by the summary statistics*

$$m_t = \frac{\boldsymbol{\mu}^\top \hat{\boldsymbol{w}}_t}{d}, \qquad\qquad q_t^\eta = \frac{\hat{\boldsymbol{w}}_t^\top \boldsymbol{\eta}}{nd\sigma^2}, \tag{93}$$

which concentrate as $d \to \infty$ to the time-constant quantities characterized by the closed-form formulae

$$m = \frac{n}{\lambda + n}, \qquad\qquad q^\eta = \frac{1}{\lambda + n}. \tag{94}$$

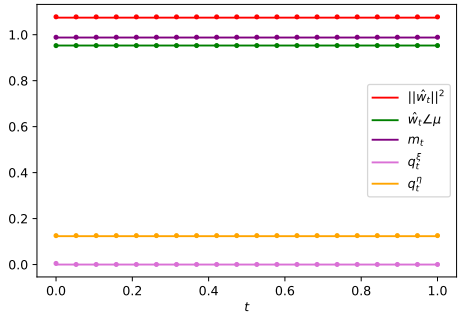

Figure 6: $n = 4, \sigma = 0.9, \lambda = 0.1, \alpha(t) = 1 - t, \beta(t) = t$. Solid lines: theoretical predictions of Result 2.1: squared norm of the weight vector $\|\hat{\boldsymbol{w}}_t\|^2$ of the DAE without skip connection (92) (red), skip connection strength $\hat{c}_t$ (blue) cosine similarity between the weight vector $\hat{\boldsymbol{w}}_t$ and the target cluster mean $\boldsymbol{\mu}$, $\hat{\boldsymbol{w}}_t \angle \boldsymbol{\mu} \equiv \hat{\boldsymbol{w}}_t^\top \boldsymbol{\mu} / \|\boldsymbol{\mu}\| \|\hat{\boldsymbol{w}}_t\|$ (green), components m_t, q_t^ξ, q_t^η of $\hat{\boldsymbol{w}}_t$ along the vectors $\boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\eta}$ (purple, pink, orange). Dots: numerical simulations in $d = 5 \times 10^4$, corresponding to training the DAE without skip connection (92) on the risk (10) uisng the `Pytorch` implementation of full-batch Adam, with learning rate 0.001 over 20000 epochs and weight decay $\lambda = 0.1$. The experimental points correspond to a single instance of the model.

Result D.1 follows from a straightforward adapatation of the derivation of Result 2.1 as presented in Appendix A. In fact, it naturally corresponds to setting the skip connection strength c to 0 in the expression of the log partition function (41). equation (93) corresponds to the zero-gradient condition thereof.

A striking consequence of Result D.1 is that asymptotically the trained vector $\hat{w}$ of the DAE (92) *does not depend on the time t*. Fig. 6 provides further support of this fact, as the summary statistics measured in simulations – training the DAE (92) using the `Pytorch` implementation of full-batch Adam – are also observed to be constant in time, and furthermore to agree well with the theoretical prediction. As for the analysis presented in the main text, it is possible to track the generative flow with a finite set of summary statistics. This is the object of the following result:

Result D.2. ***(Summary statistics for the no-skip connection case)*** *Le $\boldsymbol{X}_t$ be a solution of the ordinary differential equation (7) with initial condition $\boldsymbol{X}_0$, when parametrized by the DAE without skip connection (92). For a given t, the projection of $\boldsymbol{X}_t$ on $\mathrm{span}(\boldsymbol{\mu}, , \boldsymbol{\eta})$ is characterized by the summary statistics*

$$M_t \equiv \frac{\boldsymbol{X}_t^\top \boldsymbol{\mu}}{d}, \qquad\qquad Q_t^\eta \equiv \frac{\boldsymbol{X}_t^\top \boldsymbol{\eta}}{nd\sigma^2}. \tag{95}$$

With probability asymptotically $1/2$ the summary statistics M_t, Q_t^η (15) concentrate for all t to the solution of the ordinary differential equations

$$\begin{cases} \frac{d}{dt} M_t = \frac{\dot\alpha(t)}{\alpha(t)} M_t + \left(\dot\beta(t) - \frac{\dot\alpha(t)}{\alpha(t)}\beta(t)\right) \frac{n}{\lambda+n} \\ \frac{d}{dt} Q_t^\eta = \frac{\dot\alpha(t)}{\alpha(t)} Q_t^\eta + \left(\dot\beta(t) - \frac{\dot\alpha(t)}{\alpha(t)}\beta(t)\right) \frac{1}{\lambda+n} \end{cases} . \tag{96}$$

The derivation of Result D.2 can be made along the exact same lines as the one for Result 3.1, presented in Appendix B. An important observation is that the flows (96) are actually the *exact* flows corresponding to a particular Gaussian mixture, as explicited in the following remark:

Remark D.3. ***(Generated density)*** *The summary statistics evolution (96) are the same evolutions that would follow from the velocity field*

$$b(x,t) = \left(\dot\beta(t) - \frac{\dot\alpha(t)}{\alpha(t)}\beta(t)\right)\left(\hat\mu \times \tanh\left(\hat\mu^\top x\right)\right) + \frac{\dot\alpha(t)}{\alpha(t)} x. \tag{97}$$

where

$$\hat\mu \equiv \frac{n}{\lambda+n}\mu + \frac{1}{\lambda+n}\eta. \tag{98}$$

Comparing with equation (67), this is the exact velocity field associated to the singular Gaussian mixture

$$\hat\rho_1(x) = \frac{1}{2}\delta(x-\hat\mu) + \frac{1}{2}\delta(x+\hat\mu). \tag{99}$$

The generative model parameterized by the DAE *without* skip connection (92) thus learns a singular density, which is a sum of two Dirac atoms, centered at $\pm\hat\mu$. It thus fails to generate a good approximation of the target ρ_1 (8). Note however that interestingly $\hat\mu$ remains a good approximation of the true mean μ, and actually converges thereto as $n \to \infty$. This is made more precise by the following result:

Remark D.4. ***(mse in the no-skip-connection case)*** *Let $\hat\mu$ be the cluster mean of the estimated density $\hat\rho_1$, as defined in Remark D.3. Then its squared distance to the true mean μ is*

$$\frac{1}{d}\|\hat\mu - \mu\|^2 = \frac{\lambda^2 + n\sigma^2}{(\lambda+n)^2} \tag{100}$$

The minimum is achieved for $\lambda = \sigma^2$ and is equal to the Bayes MSE 4.1. In particular, this MSE decays as $\Theta_n(1/n)$.

Strikingly, the generative model parametrized by (92) manages to achieve the Bayes optimal $\Theta_n(1/n)$ rate in terms of the *estimation MSE* over the cluster means, but completely fails to accurately estimate the true variance.

E A FULLY EXPRESSIVE MODEL MEMORIZES

In this appendix, we show that the absence of memorization – defined as the ability of the generative model to generate new samples, and not just retrieve the training samples– is enabled by the network parametrization of the generative model. In fact, a fully expressive (flexible) model would in fact *memorize* the train set. Consider the network-parametrized minimization problem over the parameter space $\{\theta_t\}_{t\in[0,1]}$

$$\hat{\mathcal{R}}(\{\theta_t\}_{t\in[0,1]}) = \frac{1}{n}\int_0^1 \sum_{\mu=1}^n \mathbb{E}_{x_0} \left\| \boldsymbol{f}_{\theta_t}(\boldsymbol{x}_t^\mu) - \boldsymbol{x}_1^\mu \right\|^2 dt. \tag{101}$$

For ease of discussion, compared to equation (5), we consider the case where for each sample x_1^μ of the target ρ_1, we sample an infinity of noises x_0 from the easy-to-sample base Gaussian distribution ρ_0, which corresponds to averaging over x_0 in equation (101). Note that doing so, compared to the case where only one x_0^μ is sampled for very x_1^μ, is expected to prevent the model from overfitting the noise and should only improve the performance. Now consider replacing the minimization equation (101) by the minimization over the space of *all* denoising functions

$$\hat{\mathcal{R}}[f] = \frac{1}{n}\int_0^1 \sum_{\mu=1}^{n} \mathbb{E}_{x_0} \left\| \boldsymbol{f}(\boldsymbol{x}_t^\mu, t) - \boldsymbol{x}_1^\mu \right\|^2 dt = \int_0^1 \mathbb{E}_{x_1\sim\tilde{\rho}_1} \mathbb{E}_{x_0} \left\| \boldsymbol{f}(\boldsymbol{x}_t, t) - \boldsymbol{x}_1 \right\|^2 dt. \tag{102}$$

In equation (102) we denoted $\tilde{\rho}_1$ the empirical distribution supported on the training samples

$$\tilde{\rho}_1(x_1) = \frac{1}{n}\sum_{\mu=1}^{n} \delta(x_1 - x_1^\mu) \tag{103}$$

and remind that the distribution of the variable x_t follows from its definition as $x_t = \alpha(t)x_0 + \beta(t)x_1$. Finally, in equation (102), instead of minimizing a function $f_t : \mathbb{R}^d \to \mathbb{R}^d$ for each $t \in [0, 1]$, we have without loss of generality rewritten $f_t(\cdot) = f(\cdot, t)$. The objective equation (102) can be seen as the limit of equation (5) when the network is infinitely flexible and can express *any* denoising function f. Comparing equation (102) to equation (4), it follows from Albergo et al. (2023) that the minimizer of equation (102) leads to a flow mapping the base Gaussian distribution ρ_0 to the density $\tilde{\rho}_1$ equation (103), since equation (102) is the *population* objective for $\tilde{\rho}_1$. Since ρ_0 and $\tilde{\rho}_1$ are both Gaussian mixtures (with the clusters of the latter being of vanishing variance), the corresponding velocity field is furthermore explicitly given in Appendix A of Albergo et al. (2023). Therefore, an infinitely expressive model minimizing the empirical risk equation (101) leads to a generated density $\tilde{\rho}_1$. In other words, it only allows to generate samples x_1^μ from the training set, and the generated mixture has n clusters with 0 variance. In contrast, the DAE-parametrized model equation (9) learns a bimodal mixture with non-zero variance.

F WASSERSETEIN DISTANCE

In this appendix, we derive a precise description of the generated distribution $\hat{\rho}_1$ and the target ρ_1. We remind that the distribution of its projection in $\mathrm{span}(\xi, \mu_{\mathrm{emp.}})^\perp$ follows the Gaussian distribution

$$X_1^\perp \sim \mathcal{N}\left(0, \underbrace{e^{2\int_0^1 \left(\dot{\beta}(t)\hat{c}_t + \frac{\dot{\alpha}(t)}{\alpha(t)}(1-\hat{c}_t\beta(t))\right)dt}}_{\hat{\sigma}^2} \mathbb{I}_{d-2}\right). \tag{104}$$

Observe that from Result 2.1,

$$\hat{c}_t = \frac{\sigma^2\beta(t)}{\alpha(t)^2 + \beta(t)^2\sigma^2} + \Theta_n(1/n). \tag{105}$$

Thus,

$$\ln\hat{\sigma}^2 = 2\int_0^1 \frac{\dot{\alpha}(t)\alpha(t) + \dot{\beta}(t)\beta(t)\sigma^2}{\alpha(t)^2 + \beta(t)^2\sigma^2} + \Theta_n(1/n) = \left[\ln\left(\alpha(t)^2 + \beta(t)^2\sigma^2\right)\right]_0^1 + \Theta_n(1/n) = \ln\sigma^2 + \Theta_n(1/n). \tag{106}$$

Thus

$$\hat{\sigma}^2 = \sigma^2 + \Theta_n(1/n) \tag{107}$$

We are now in a position to compute the Mixture Wasserstein distance between the generated density $\hat{\rho}_1$ and the target ρ_1. Because these are $d-$ dimensional distributions, we normalize this distance by $1/d$ so as to have order 1 metrics in the considered asymptotic limit $d \to \infty$. Because of this, the precise distribution of the clusters of $\hat{\rho}_1$ in the two dimensional space $\mathrm{span}(\xi, \mu_{\mathrm{emp.}})$ does *not* matter, provided it does not involve moments diverging with d. We will show that this very reasonable assumption is indeed verified, after the computation of the distance.

F.1 WASSERSTEIN DISTANCE

We aim at evaluating a distance metric in the space of distributions to quantify the discrepancy between the true density ρ_1 and the generated $\hat{\rho}_1$. Many natural metrics (e.g. the KL divergence) are however intractable in our setting. We take inspiration from the Gaussian mixture Wasserstein distance $M\mathcal{W}_2$, proposed in Delon & Desolneux (2020) as variant of the $\mathcal{W}_2$ distance for Gaussian mixtures. Because $\hat{\rho}_1$ is a mixture, but the clusters in $\mathrm{span}(\xi, \mu_{\mathrm{emp.}})$ are only halves of Gaussian clusters, we define and employ a very similar metric for arbitrary mixtures:

Definition F.1. (*mixture Wasserstein distance*) Given two mixtures $\sum_{i=1}^{K} \rho_i \mu_i$ and $\sum_{i=1}^{J} \tau_i \nu_i$ (with μ_i, ν_i not necessarily Gaussian densities), the MW_2 distance is defined as

$$M\mathcal{W}_2^2 = \min_{w \in \mathbb{R}^{K \times J} | w 1_J = (\rho_1, \dots, \rho_K), w^\top 1_K = (\tau_1, \dots, \tau_J)} \frac{1}{d} \sum_{k=1}^{K} \sum_{j=1}^{J} w_{kj} \mathcal{W}_2^2(\mu_k, \nu_j). \tag{108}$$

This is the same definition as Delon & Desolneux (2020), except that we allow for non-Gaussian μ_i, ν_i. Note that we introduced without loss of generality a normalization $1/d$, since we are comparing $d-$dimensional densities, and expect the distance to scale with d. With the normalization, the metric stays $\Theta_d(1)$ as $d \to \infty$. In the present setting, this evaluates to

$$M\mathcal{W}_2^2[\rho_1, \hat{\rho}_1] = \frac{1}{d}\mathcal{W}_2^2(\hat{\rho}_1^+, \mathcal{N}(\mu, \sigma^2)) + \frac{1}{d}\mathcal{W}_2^2(\hat{\rho}_1^-, \mathcal{N}(-\mu, \sigma^2)) \tag{109}$$

where we introduced the densities $\hat{\rho}_1 = 1/2 \hat{\rho}_1^+ + 1/2 \hat{\rho}_1^-$ for the two clusters of $\hat{\rho}_1$ centered at $\pm\hat{\mu}$. We denote further decompose $\rho_1^\pm = \rho_1^{\pm\parallel} \otimes \rho_1^{\pm\perp}$ into the product of the distribution $\rho_1^{\pm\parallel}$ in $\mathrm{span}(\xi, \mu_{\mathrm{emp.}})$ and the Gaussian $d-2$ dimensional density $\rho_1^{\pm\perp}$ in $\mathrm{span}(\xi, \mu_{\mathrm{emp.}})^\perp$. We can similarly decompose the target Gaussian density $\mathcal{N}(\pm\mu, \sigma^2 \mathbb{I}_d) = \mathcal{N}(\pm\mu, \sigma^2 \mathbb{I}_2) \otimes \mathcal{N}(0, \sigma^2 \mathbb{I}_{d-2})$. Using the the properties of Wasserstein distances between product measures Panaretos & Zemel (2019),

$$\frac{1}{d}\mathcal{W}_2^2(\hat{\rho}_1^+, \mathcal{N}(\mu, \sigma^2)) = \frac{1}{d}\mathcal{W}_2^2(\rho_1^{+\parallel}, \mathcal{N}(\pm\mu^\parallel, \sigma^2)) + \frac{1}{d}\mathcal{W}_2^2(\rho_1^{+\perp}, \mathcal{N}(0, \sigma^2)) \tag{110}$$

Note that since $\rho_1^{\pm\perp}$ is Gaussian with variance $\hat{\sigma}^2$, the second term corresponds to the Wasserstein distance between two Gaussian distributions and read

$$\frac{1}{d}\mathcal{W}_2^2(\rho_1^{+\perp}, \mathcal{N}(0, \sigma^2)) = (\sigma - \hat{\sigma})^2 = \Theta_n(1/n). \tag{111}$$

We now bound $\frac{1}{d}\mathcal{W}_2^2(\rho_1^{+\parallel}, \mathcal{N}(\pm\mu^\parallel, \sigma^2))$. Note that the two densities are centered around $\hat{\mu}$ and μ. The discrepancy between these means will provide the dominant term in the distance. To see this, we introduce $\nu_1(x) = \delta(x - \hat{\mu})$ and $\nu_2(x) = \delta(x - \mu)$, two Diracs centered at the means, and upper-bound $\frac{1}{d}\mathcal{W}_2^2(\rho_1^{+\parallel}, \mathcal{N}(\pm\mu^\parallel, \sigma^2))$ using the triangular inequality

$$\frac{1}{d}\mathcal{W}_2^2(\rho_1^{+\parallel}, \mathcal{N}(\pm\mu^\parallel, \sigma^2)) \leq \frac{1}{d}\mathcal{W}_2^2(\rho_1^{+\parallel}, \nu_1) + \frac{1}{d}\mathcal{W}_2^2(\nu_1, \nu_2) + \frac{1}{d}\mathcal{W}_2^2(\nu_2, \mathcal{N}(\pm\mu^\parallel, \sigma^2)). \tag{112}$$

The last term is asymptotically vanishing as $\Theta_d(1/d)$. Under very mild assumption on $\rho_1^{+\parallel}$ (which we show are verified in the next subsection), the Wasserstein distance between two-dimensional distributions $\mathcal{W}_2^2(\rho_1^{+\parallel}, \nu_1)$ should be $\Theta_d(1)$, so the first term also vanishes as $\Theta_d(1/d)$. The second term is equal to

$$\frac{1}{d}\mathcal{W}_2^2(\nu_1, \nu_2) = \frac{1}{d}\|\mu - \hat{\mu}\|^2 = \Theta_n(1/n), \tag{113}$$

using result B.3. The derivation proceeds identically for the other pair of clusters $\frac{1}{d}\mathcal{W}_2^2(\hat{\rho}_1^-, \mathcal{N}(-\mu, \sigma^2))$. Putting everything together, we reach

$$M\mathcal{W}_2^2[\rho_1, \hat{\rho}_1] \leq \Theta_n(1/n) \tag{114}$$

Bayes optimal rate We briefly consider the mixture corresponding to the bimodal mixture centered at the Bayes optimal mean estimator $\pm\hat{\mu}(\mathcal{D})$, and assuming perfect knowledge of the cluster covariances. Again, this is provided as an insightful baseline, and does *not* constitute a generative model, since exact oracle knowledge of the form of ρ_1 and of σ^2 is assumed. For the Bayes estimator:

$$M\mathcal{W}_2^2[\rho_1, \hat{\rho}_1] = \frac{2}{d}\|\mu - \hat{\mu}\|^2 = \Theta_n(1/n), \tag{115}$$

using Result 4.1.

We now briefly give two other examples for generative models differently parametrized, for which the generated density does not converge to the target ρ_1 in $M\mathcal{W}_2$ distance.

Auto-encoder without skip connection As derived in D, the generated density when the model is parametrized by a DAE *without* skip connection is a degenerate mixture $1/2\delta(\cdot - \hat{\mu}) + 1/2\delta(\cdot + \hat{\mu})$, which corresponds to setting $\hat{\sigma} = 0$ in the above derivation. Thus, it follows that

$$M\mathcal{W}_2^2[\rho_1, \hat{\rho}_1] = \Theta_n(1), \tag{116}$$

i.e. without skip connection the generative model fails to learn to generate the target mixture.

Fully expressive model We now consider the case of a model which memorizes the train set, as discussed in Appendix E. In this case

$$\hat{\rho}_1(x) = \frac{1}{n}\sum_{\mu=1}^{n} \delta(x - x_1^\mu), \tag{117}$$

which is a (degenerate) Gaussian mixture. It is straightforward to see that for any x_1^μ, $\mathcal{W}_2^2[\mathcal{N}(\pm\mu, \sigma^2\mathbb{I}_d, \delta(\cdot - x_1^\mu)] \geq \sigma^2 = \Theta_n(1)$, and therefore

$$M\mathcal{W}_2^2[\rho_1, \hat{\rho}_1] \geq \sigma^2 = \Theta_n(1). \tag{118}$$

Thus $\hat{\rho}_1$ is bounded away from the target ρ_1.

We close the appendix by deriving the precise form of $\hat{\rho}_1^{\pm\parallel}$, although the precise distribution in this two-dimensional space is asymptotically irrelevant for all the considered metrics, as we showed.

F.2 Distribution in $\mathrm{span}(\xi, \mu_{\mathrm{emp.}})$

We study in more detail the dynamics of $X_t^\parallel$, defined as the projection of X_t in $\mathrm{span}(\xi, \mu_{\mathrm{emp.}})$. Since the initial $X_0 \sim \rho_0$ is Gaussian, so is its projection $X_0^\parallel \sim \mathcal{N}(0, \mathbb{I}_2)$. Let us also call $\hat{w}_t^\parallel$ the projection of $\hat{w}_t$ in $\mathrm{span}(\xi, \mu_{\mathrm{emp.}})$. Projecting the dynamics equation (7) into $\mathrm{span}(\xi, \mu_{\mathrm{emp.}})$,

$$\dot{X}_t^\parallel = \left(\dot{\beta}(t)\hat{c}_t + \frac{\dot{\alpha}(t)}{\alpha(t)}(1 - \hat{c}_t\beta(t))\right) X_t^\parallel \pm \left(\dot{\beta}(t) - \frac{\dot{\alpha}(t)}{\alpha(t)}\beta(t)\right)\hat{w}_t^\parallel, \tag{119}$$

where the sign of the drift term is given by $\mathrm{sign}(X_0^{\parallel\top}\hat{w}_0^\parallel)$. Like in Appendix B, we assumed that $\mathrm{sign}(\hat{w}_t^\top X_t)$ stays constant during the transport. This can be solved in closed form for $t = 1$ as

$$X_1^\parallel = X_0^\parallel e^{\int_0^1 \gamma(t)dt} + \mathrm{sign}(X_0^{\parallel\top}\hat{w}_0^\parallel) e^{\int_0^1 \gamma(t)dt} \int_0^1 e^{-\int_0^t \gamma(s)ds} \left(\dot{\beta}(t) - \frac{\dot{\alpha}(t)}{\alpha(t)}\beta(t)\right)\hat{w}_t^\parallel dt \tag{120}$$

where we used the shorthand

$$\gamma(t) \equiv \left(\dot{\beta}(t)\hat{c}_t + \frac{\dot{\alpha}(t)}{\alpha(t)}(1 - \hat{c}_t\beta(t))\right). \tag{121}$$

The second term multiplied by the sign corresponds to $\hat{\mu} \in \mathrm{span}(\xi, \mu_{\mathrm{emp.}})$ as characterized by Result 3.1. The distribution of $X_1^\parallel$ follows from that of the Gaussian $X_0^\parallel$. If $X_0^\parallel$ is in the half-space $\{x \in \mathbb{R}^2 | x^\top \hat{w}_0^\parallel \geq 0\}$ then $X_1^\parallel = \hat{\sigma}X_0^\parallel + \hat{\mu}$; If $X_0^\parallel$ is in the half-space $\{x \in \mathbb{R}^2 | x^\top \hat{w}_0^\parallel \leq 0\}$ then $X_1^\parallel = \hat{\sigma}X_0^\parallel - \hat{\mu}$. In other words, the distribution of $X_1^\perp$ is a mixture of two clusters, at $\pm\hat{\mu}$. Each cluster corresponds to half a Gaussian cluster of variance $\hat{\sigma}^2\mathbb{I}$, i.e. a Gaussian cluster cleft along a hyperplane whose othogonal vector is $\hat{w}_0^\parallel$ as characterized by Result 2.1.