

A Derivation of Result 2.1

In this Appendix, we detail the derivation of the tight ODE description (10) of the SGD training dynamics (4), as provided in Result 2.1. We sequentially examine the dynamics for the skip connection b and the weight matrix w .

A.1 SGD dynamics of the skip connection

We first derive a closed-form expression for the evolution of the skip connection strength b (3) over the SGD iterations. We recall that the latter read

$$b_{\mu+1} - b_\mu = -\frac{\eta}{d^2} \mathbb{E}_t \left[-2(1 - b\beta_t)\beta_t \|x_1^\mu\|^2 + 2b\alpha_t^2 \|x_0^\mu\|^2 + O(\sqrt{d}) \right], \quad (16)$$

keeping only leading order terms. Note that $\|x_1^\mu\|^2/d$ (resp. $\|x_0^\mu\|^2/d$) asymptotically concentrate to Λ (resp. 1) in the limit $d \rightarrow \infty$. Therefore, the increment $db = b_{\mu+1} - b_\mu$ self-averages as

$$\frac{d}{2\eta} db = \mathbb{E}_t [\beta_t(1 - b\beta_t)\Lambda - b\alpha_t^2]. \quad (17)$$

A.2 SGD dynamics of the weight matrix

A.2.1 SGD update

We now turn to deriving a similar tight asymptotic characterization for the evolution of the weight matrix w (3) under the SGD dynamics. Let us first write explicitly the SGD updates (4). Developing the derivative, and dropping the time index μ for readability, for $1 \leq \gamma \leq r, 1 \leq i \leq d$, the SGD update of the weight matrix reads

$$\begin{aligned} dw_{i\gamma} = & -\frac{2\eta}{d} \sum_{\delta=1}^r \mathbb{E}_t [\sigma(\omega_\gamma^t + v_\gamma p_t) \sigma(\omega_\delta^t + v_\delta p_t)] w_{i\delta} + \frac{2\eta}{\sqrt{d}} \mathbb{E}_t [((1 - b\beta_t)x_i^1 - b\alpha_t x_i^0) \sigma(\omega_\gamma^t + v_\gamma p_t)] \\ & - \frac{2\eta}{\sqrt{d}} \mathbb{E}_t \left[(\alpha_t x_i^0 + \beta_t x_i^1) \sum_{\delta=1}^r \sigma(\omega_\delta^t + v_\delta p_t) \left(\frac{w_\delta^\top w_\gamma}{d} \right) \sigma'(\omega_\gamma^t + v_\gamma p_t) \right] \\ & + \frac{2\eta}{\sqrt{d}} \mathbb{E}_t [(\alpha_t x_i^0 + \beta_t x_i^1) ((1 - b\beta_t)\lambda_\gamma^1 - b\alpha_t \lambda_\gamma^0) \sigma'(\omega_\gamma^t + v_\gamma p_t)] - \frac{\eta}{d} \lambda w_{i\gamma} \end{aligned} \quad (18)$$

We introduced the shorthands

$$\lambda_\gamma^1 \equiv \frac{w_\gamma^\top x^1}{\sqrt{d}}, \quad \lambda_\gamma^0 \equiv \frac{w_\gamma^\top x^0}{\sqrt{d}}, \quad \omega_\gamma^t = \alpha_t \lambda_\gamma^0 + \beta_t \lambda_\gamma^1. \quad (19)$$

A.2.2 Expected increment

In the likeness of the settings studied by e.g. [57, 58, 52], we expect the dynamics to asymptotically self-average. Let us accordingly evaluate the expectation $\mathbb{E}[dw_{i\gamma}]$ over the running data sample $x_\mu^{1,0}$. This can be compactly rewritten as

$$\mathbb{E}[dw_{i\gamma}] = \mathbb{E}_t \mathbb{E}_c [\mathbb{E}^c dw_{i\gamma}^{t,c}], \quad (20)$$

with the expectation \mathbb{E}_c bearing over the manifold coordinate $c \sim \pi$ (8). Conditional on the manifold coordinate c , the expectation \mathbb{E}^c bears over the Gaussian random variable associated to the c -indexed cluster in (8), distributed as $\mathcal{N}(\mu(c), \Sigma(c))$. We remind the reader that the covariances $\{\Sigma(c)\}_c$ are assumed jointly diagonalizable. Without loss of generality, we place ourselves in the basis in which they are directly diagonal, and denote in the following by ϱ_i^c the i -th eigenvalue of $\Sigma(c)$. The

$$\begin{aligned}
\mathbb{E}^c[dw_{i\gamma}^{t,c}] = & -\frac{2\eta}{d} \sum_{\delta=1}^r \underbrace{\mathbb{E}^c [\sigma(\omega_\gamma^t + v_\gamma p_t) \sigma(\omega_\delta^t + v_\delta p_t)]}_{\mathcal{A}_{i\gamma\delta}^{t,c}} w_{i\delta} + \frac{2\eta}{\sqrt{d}} (1 - b\beta_t) \underbrace{\mathbb{E}^c [x_i^1 \sigma(\omega_\gamma^t + v_\gamma p_t)]}_{\mathcal{B}_{i\gamma}^{1,t,c}} \\
& - \frac{2\eta}{\sqrt{d}} b\alpha_t \underbrace{\mathbb{E}^c [x_i^0 \sigma(\omega_\gamma^t + v_\gamma p_t)]}_{\mathcal{B}_{i\gamma}^{0,t,c}} \\
& - \frac{2\eta}{\sqrt{d}} \alpha_t \sum_{\delta=1}^r \left(\frac{w_\delta^\top w_\gamma}{d} \right) \underbrace{\mathbb{E}^c [x_i^0 \sigma(\omega_\delta^t + v_\delta p_t) \sigma'(\omega_\gamma^t + v_\gamma p_t)]}_{\mathcal{C}_{i\gamma\delta}^{0,t,c}} \\
& - \frac{2\eta}{\sqrt{d}} \beta_t \sum_{\delta=1}^r \left(\frac{w_\delta^\top w_\gamma}{d} \right) \underbrace{\mathbb{E}^c [x_i^1 \sigma(\omega_\delta^t + v_\delta p_t) \sigma'(\omega_\gamma^t + v_\gamma p_t)]}_{\mathcal{C}_{i\gamma\delta}^{1,t,c}} \\
& + \frac{2\eta}{\sqrt{d}} \alpha_t (1 - b\beta_t) \underbrace{\mathbb{E}^c [x_i^0 \lambda_\gamma^1 \sigma'(\omega_\gamma^t + v_\gamma p_t)]}_{\mathcal{D}_{i\gamma}^{01,t,c}} - \frac{2\eta}{\sqrt{d}} \alpha_t^2 b \underbrace{\mathbb{E}^c [x_i^0 \lambda_\gamma^0 \sigma'(\omega_\gamma^t + v_\gamma p_t)]}_{\mathcal{D}_{i\gamma}^{00,t,c}} \\
& + \frac{2\eta}{\sqrt{d}} \beta_t (1 - b\beta_t) \underbrace{\mathbb{E}^c [x_i^1 \lambda_\gamma^1 \sigma'(\omega_\gamma^t + v_\gamma p_t)]}_{\mathcal{D}_{i\gamma}^{11,t,c}} - \frac{2\eta}{\sqrt{d}} \alpha_t \beta_t b \underbrace{\mathbb{E}^c [x_i^1 \lambda_\gamma^0 \sigma'(\omega_\gamma^t + v_\gamma p_t)]}_{\mathcal{D}_{i\gamma}^{10,t,c}} - \frac{2\eta}{d} \lambda w_{i\gamma}.
\end{aligned} \tag{21}$$

588 The various coefficients $\mathcal{A}^{t,c}, \mathcal{B}^{t,c}, \mathcal{C}^{t,c}, \mathcal{D}^{t,c}$ can be evaluated leveraging the fact that the data com-
589 ponents $x_i^{1,0}$ are weakly correlated with the local fields $\omega, \lambda^{0,1}$, i.e. have $\Theta_d(1/\sqrt{d})$ covariance. Using
590 the expansions for weakly correlated Gaussian variables reported e.g. in [53] (Appendix B.1), we

591 reach

$$\mathcal{A}_{\gamma\delta}^{t,c} = I_{\sigma\sigma}^{t,c}(\gamma, \delta) \quad (22)$$

$$\mathcal{B}_{\gamma}^{1,t,c} = I_{\sigma}^{t,c}(\gamma)\mu_i^c + \frac{1}{\sqrt{d}} \frac{\beta_t w_{i\gamma} \varrho_i^c}{\Omega_{\gamma\gamma}^{t,c}} (I_{\sigma\omega}^{t,c}(\gamma, \gamma) - \beta_t M_{\gamma}^c I_{\sigma}^{t,c}(\gamma)) \quad (23)$$

$$\mathcal{B}_{\gamma}^{0,t,c} = \frac{1}{\sqrt{d}} \frac{\alpha_t w_{i\gamma}}{\Omega_{\gamma\gamma}^{t,c}} (I_{\sigma\omega}^{t,c}(\gamma, \gamma) - \beta_t M_{\gamma}^c I_{\sigma}^{t,c}(\gamma)) \quad (24)$$

$$\begin{aligned} \mathcal{C}_{i\gamma\delta}^{0,t,c} = & \frac{1 - \delta_{\gamma\delta}}{\sqrt{d}} \frac{\alpha_t}{\Omega_{\gamma\gamma}^{t,c} \Omega_{\delta\delta}^{t,c} - (\Omega_{\gamma\delta}^{t,c})^2} \left[(I_{\sigma'\sigma\omega}^{t,c}(\gamma, \delta, \gamma) - \beta_t M_{\gamma}^c I_{\sigma'\sigma}^{t,c}(\gamma, \delta)) (\Omega_{\delta\delta}^{t,c} w_{i\gamma} - \Omega_{\gamma\delta}^{t,c} w_{i\delta}) \right. \\ & \left. + (I_{\sigma'\sigma\omega}^{t,c}(\gamma, \delta, \delta) - \beta_t M_{\delta}^c I_{\sigma'\sigma}^{t,c}(\gamma, \delta)) (\Omega_{\gamma\gamma}^{t,c} w_{i\delta} - \Omega_{\gamma\delta}^{t,c} w_{i\gamma}) \right] \\ & + \frac{\delta_{\gamma\delta}}{\Omega_{\gamma\gamma}^{t,c}} \alpha_t w_{i\gamma} (I_{\sigma'\sigma\omega}^{t,c}(\gamma, \gamma, \gamma) - \beta_t M_{\gamma}^c I_{\sigma'\sigma}^{t,c}(\gamma, \gamma)) \end{aligned} \quad (25)$$

$$\mathcal{C}_{i\gamma\delta}^{1,t,c} = I_{\sigma'\sigma}^{t,c}(\gamma, \delta)\mu_i^c + \frac{\beta_t \varrho_i^c}{\alpha_t} \mathcal{C}_{i\gamma\delta}^{0,t,c} \quad (26)$$

$$\mathcal{D}_{i\gamma}^{01,t,c} = \frac{1}{\sqrt{d}} \frac{Q_{\gamma\gamma}^c \alpha_t w_{i\gamma}}{Q_{\gamma\gamma}^c \Omega_{\gamma\gamma}^{t,c} - \beta_t^2 (Q_{\gamma\gamma}^c)^2} \left[I_{\lambda^1 \sigma' \omega}^{t,c}(\gamma, \gamma, \gamma) - \beta_t I_{(\lambda^1)^2 \sigma'}^{t,c}(\gamma, \gamma) \right] \quad (27)$$

$$\mathcal{D}_{i\gamma}^{00,t,c} = \frac{1}{\sqrt{d}} \frac{w_{i\gamma}}{Q_{\gamma\gamma} \Omega_{\gamma\gamma}^{t,c} - \alpha_t^2 (Q_{\gamma\gamma})^2} \left[I_{(\lambda^0)^2 \sigma'}^{t,c}(\gamma, \gamma) (\Omega_{\gamma\gamma}^{t,c} - \alpha_t^2 Q_{\gamma\gamma}) \right] \quad (28)$$

$$\mathcal{D}_{i\gamma}^{10,t,c} = \mu_i^c I_{\lambda^0 \sigma'}^{t,c}(\gamma, \gamma) + \frac{1}{\sqrt{d}} \frac{Q_{\gamma\gamma} \beta_t \varrho_i^c w_{i\gamma}}{Q_{\gamma\gamma} \Omega_{\gamma\gamma}^{t,c} - \alpha_t^2 (Q_{\gamma\gamma})^2} \left[I_{\lambda^0 \sigma' \omega}^{t,c}(\gamma, \gamma, \gamma) - \alpha_t I_{(\lambda^0)^2 \sigma'}^{t,c}(\gamma, \gamma) - M_{\gamma}^c \beta_t I_{\lambda^0 \sigma'}^{t,c}(\gamma, \gamma) \right] \quad (29)$$

$$\mathcal{D}_{i\gamma}^{11,t,c} = \mu_i^c I_{\lambda^1 \sigma'}^{t,c}(\gamma, \gamma) + \frac{1}{\sqrt{d}} \frac{w_{i\gamma} \varrho_i^c}{Q_{\gamma\gamma}^c \Omega_{\gamma\gamma}^{t,c} - \beta_t^2 (Q_{\gamma\gamma}^c)^2} \left[(I_{(\lambda^1)^2 \sigma'}^{t,c}(\gamma, \gamma) - M_{\gamma}^c I_{\lambda^1 \sigma'}^{t,c}(\gamma, \gamma)) (\Omega_{\gamma\gamma}^{t,c} - \beta_t^2 Q_{\gamma\gamma}^c) \right]. \quad (30)$$

592 We introduced the summary statistics

$$M^c = \frac{w^\top \mu(c)}{\sqrt{d}}, \quad Q^c = \frac{w^\top \Sigma(c) w}{d}, \quad (31)$$

$$\mathcal{Q} = \frac{w^\top w}{d}, \quad \Omega^{t,c} = \alpha_t^2 \mathcal{Q} + \beta_t^2 Q^c, \quad T^{ck} = \mu(c)^\top \mu(k) \quad (32)$$

593 One also needs to introduce the further statistics

$$G = \frac{w^\top E}{\sqrt{d}}, \quad P^c = E^\top \mu(c), \quad (33)$$

594 where we remind that the columns of $E \in \mathbb{R}^{d \times R}$ constitute an orthonormal basis of the
595 R -dimensional subspace \mathcal{E} in which we aim to characterize the generated density. Finally, we
596 also used the shorthands:

$$I_{\sigma\sigma}^{t,c}(\gamma, \delta) = \mathbb{E}_{\omega_\gamma, \omega_\delta} [\sigma(\omega_\gamma + v_\gamma p_t) \sigma(\omega_\delta + v_\delta p_t)], \quad \omega_\gamma, \omega_\delta \sim \mathcal{N}(\beta_t M_{(\gamma,\delta)}^c, \Omega_{(\gamma,\delta)}^{t,c}) \quad (34)$$

$$I_{\sigma}^{t,c}(\gamma) = \mathbb{E}_{\omega_\gamma} [\sigma(\omega_\gamma + v_\gamma p_t)], \quad \omega_\gamma \sim \mathcal{N}(\beta_t M_{\gamma}^c, \Omega_{\gamma\gamma}^{t,c}) \quad (35)$$

$$I_{\sigma\omega}^{t,c}(\gamma, \delta) = \mathbb{E}_{\omega_\gamma, \omega_\delta} [\sigma(\omega_\gamma + v_\gamma p_t) \omega_\delta], \quad \omega_\gamma, \omega_\delta \sim \mathcal{N}(\beta_t M_{(\gamma,\delta)}^c, \Omega_{(\gamma,\delta)}^{t,c}) \quad (36)$$

$$I_{\sigma'\sigma\omega}^{t,c}(\gamma, \delta, \epsilon) = \mathbb{E}_{\omega_\gamma, \omega_\delta, \omega_\epsilon} [\sigma'(\omega_\gamma + v_\gamma p_t) \sigma(\omega_\delta + v_\delta p_t) \omega_\epsilon], \quad \omega_\gamma, \omega_\delta, \omega_\epsilon \sim \mathcal{N}(\beta_t M_{(\gamma,\delta,\epsilon)}^c, \Omega_{(\gamma,\delta,\epsilon)}^{t,c}) \quad (37)$$

$$I_{\sigma'\sigma}^{t,c}(\gamma, \delta) = \mathbb{E}_{\omega_\gamma, \omega_\delta} [\sigma'(\omega_\gamma + v_\gamma p_t) \sigma(\omega_\delta + v_\delta p_t)], \quad \omega_\gamma, \omega_\delta \sim \mathcal{N}(\beta_t M_{(\gamma,\delta)}^c, \Omega_{(\gamma,\delta)}^{t,c}), \quad (38)$$

597 and

$$I_{\lambda^1 \sigma' \omega}^{t,c}(\gamma, \delta, \epsilon) = \mathbb{E}_{\lambda_\gamma^1, \omega_\delta, \omega_\epsilon} [\lambda_\gamma^1 \sigma'(\omega_\delta + v_\delta p_t) \omega_\epsilon],$$

$$\lambda_\gamma^1, \omega_\delta, \omega_\epsilon \sim \mathcal{N} \left(\begin{pmatrix} M_\gamma^c \\ \beta_t M_{(\delta, \epsilon)}^c \end{pmatrix}, \begin{pmatrix} Q_{\gamma\gamma}^c & \beta_t Q_{\gamma, (\delta, \epsilon)}^c \\ \beta_t (Q_{\gamma, (\delta, \epsilon)}^c)^\top & \Omega_{(\delta, \epsilon)}^{t,c} \end{pmatrix} \right) \quad (39)$$

$$I_{(\lambda^1)^2 \sigma'}^{t,c}(\gamma, \delta) = \mathbb{E}_{\omega_\gamma, \omega_\delta} [(\lambda_\gamma^1)^2 \sigma'(\omega_\delta + v_\delta p_t)],$$

$$\lambda_\gamma^1, \omega_\delta \sim \mathcal{N} \left(\begin{pmatrix} M_\gamma^c \\ \beta_t M_\delta^c \end{pmatrix}, \begin{pmatrix} Q_{\gamma\gamma}^c & \beta_t Q_{\gamma, \delta}^c \\ \beta_t (Q_{\gamma, \delta}^c)^\top & \Omega_\delta^{t,c} \end{pmatrix} \right) \quad (40)$$

$$I_{\lambda^0 \sigma' \omega}^{t,c}(\gamma, \delta, \epsilon) = \mathbb{E}_{\lambda_\gamma^0, \omega_\delta, \omega_\epsilon} [\lambda_\gamma^0 \sigma'(\omega_\delta + v_\delta p_t) \omega_\epsilon],$$

$$\lambda_\gamma^0, \omega_\delta, \omega_\epsilon \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ \beta_t M_{(\delta, \epsilon)}^c \end{pmatrix}, \begin{pmatrix} Q_{\gamma\gamma} & \alpha_t Q_{\gamma, (\delta, \epsilon)} \\ \alpha_t (Q_{\gamma, (\delta, \epsilon)})^\top & \Omega_{(\delta, \epsilon)}^{t,c} \end{pmatrix} \right) \quad (41)$$

$$I_{(\lambda^0)^2 \sigma'}^{t,c}(\gamma, \delta) = \mathbb{E}_{\omega_\gamma, \omega_\delta} [(\lambda_\gamma^0)^2 \sigma'(\omega_\delta + v_\delta p_t)],$$

$$\lambda_\gamma^0, \omega_\delta \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ \beta_t M_\delta^c \end{pmatrix}, \begin{pmatrix} Q_{\gamma\gamma} & \alpha_t Q_{\gamma, \delta} \\ \alpha_t (Q_{\gamma, \delta})^\top & \Omega_\delta^{t,c} \end{pmatrix} \right) \quad (42)$$

$$(43)$$

598 A.2.3 Update equation for the summary statistics

599 The training dynamics of the DAE weights w are thus governed by set of finite-dimensional summary
600 statistics. To close the equations and reach a self-contained characterization, we now turn to deriving
601 the induced dynamics of the summary statistics. To that end, following e.g. [28], it proves convenient
602 to first introduce a new set of summary statistics *densities*. For any $\varrho : \mathbb{R}^\kappa \rightarrow \mathbb{R}$ —denoting a
603 joint sequence of eigenvalues $\{\varrho(c)\}$ —, let us assume the existence of the densities $m, p : \mathbb{R}^\kappa \times$
604 $\mathcal{F}(\mathbb{R}^\kappa, \mathbb{R}) \rightarrow \mathbb{R}, q, g : \times \mathcal{F}(\mathbb{R}^\kappa, \mathbb{R}) \rightarrow \mathbb{R}$ and $\theta : (\mathbb{R}^\kappa)^2 \times \mathcal{F}(\mathbb{R}^\kappa, \mathbb{R}) \rightarrow \mathbb{R}$ so that the summary
605 statistics $M^c, Q, Q^c, T, G, P(31)$ can be decomposed as

$$M^c = \int d\varrho m^c(\varrho), \quad (44)$$

$$Q^c = \int d\varrho q(\varrho) \varrho_c, \quad (45)$$

$$Q = \int d\varrho q(\varrho), \quad (46)$$

$$T^{ck} = \int d\varrho \theta^{ck}(\varrho), \quad (47)$$

$$G = \int d\varrho g(\varrho), \quad (48)$$

$$P^c = \int d\varrho p^c(\varrho), \quad (49)$$

606 The following subsections focus on deriving the updates of the summary statistic densities
607 $m(\cdot), q(\cdot), \theta(\cdot), g(\cdot)$ inherited from the SGD dynamics of the weight matrix w (21).

608 A.2.4 Overlap $m^k(\cdot)$

609 The expected increment for $m^k(\varrho)$ can also be decomposed as

$$\mathbb{E}[dm^k(\varrho)] = \mathbb{E}_t \mathbb{E}_c [\mathbb{E}^c dm_k^{t,c}(\varrho)], \quad (50)$$

610 with

$$\begin{aligned}
\frac{d}{2\eta} \mathbb{E}^c [dm_k^{t,c}(\varrho)_\gamma] &= - \sum_{\delta=1}^r I_{\sigma\sigma}^{t,c}(\gamma, \delta) m^k(\varrho)_\delta + (1 - b\beta_t) I_\sigma^{t,c}(\gamma) \theta^{ck}(\varrho) \\
&+ \frac{(1 - b\beta_t) \beta_t \varrho_c m^k(\varrho)_\gamma - b\alpha_t^2 m^k(\varrho)_\gamma}{\Omega_{\gamma\gamma}^{t,c}} (I_{\sigma\omega}^{t,c}(\gamma, \gamma) - \beta_t M_\gamma^c I_\sigma^{t,c}(\gamma)) \\
&- \sum_{\delta \neq \gamma} \frac{(\alpha_t^2 + \beta_t^2 \varrho_c) \mathcal{Q}_{\gamma\delta}}{\Omega_{\gamma\gamma}^{t,c} \Omega_{\delta\delta}^{t,c} - (\Omega_{\gamma\delta}^{t,c})^2} \left[(I_{\sigma'\sigma\omega}^{t,c}(\gamma, \delta, \gamma) - \beta_t M_\gamma^c I_{\sigma'\sigma}^{t,c}(\gamma, \delta)) (\Omega_{\delta\delta}^{t,c} m^k(\varrho)_\gamma - \Omega_{\gamma\delta}^{t,c} m^k(\varrho)_\delta) \right. \\
&\quad \left. + (I_{\sigma'\sigma\omega}^{t,c}(\gamma, \delta, \delta) - \beta_t M_\delta^c I_{\sigma'\sigma}^{t,c}(\gamma, \delta)) (\Omega_{\gamma\gamma}^{t,c} m^k(\varrho)_\delta - \Omega_{\gamma\delta}^{t,c} m^k(\varrho)_\gamma) \right] \\
&- \frac{\mathcal{Q}_{\gamma\gamma} (\alpha_t^2 + \beta_t^2 \varrho_c) m^k(\varrho)_\gamma}{\Omega_{\gamma\gamma}^{t,c}} (I_{\sigma'\sigma\omega}^{t,c}(\gamma, \gamma, \gamma) - \beta_t M_\gamma^c I_{\sigma'\sigma}^{t,c}(\gamma, \gamma)) - \beta_t \sum_{\delta=1}^r \mathcal{Q}_{\gamma\delta} I_{\sigma'\sigma}^{t,c}(\gamma, \delta) \theta^{ck}(\varrho) \\
&+ \frac{\alpha_t^2 (1 - b\beta_t) \mathcal{Q}_{\gamma\gamma}^c m^k(\varrho)_\gamma}{\mathcal{Q}_{\gamma\gamma}^c \Omega_{\gamma\gamma}^{t,c} - \beta_t^2 (\mathcal{Q}_{\gamma\gamma}^c)^2} \left[I_{\lambda^1 \sigma' \omega}^{t,c}(\gamma, \gamma, \gamma) - \beta_t I_{(\lambda^1)^2 \sigma'}^{t,c}(\gamma, \gamma) \right] \\
&- \frac{\alpha_t^2 b m^k(\varrho)_\gamma}{\mathcal{Q}_{\gamma\gamma} \Omega_{\gamma\gamma}^{t,c} - \alpha_t^2 (\mathcal{Q}_{\gamma\gamma})^2} \left[I_{(\lambda^0)^2 \sigma'}^{t,c}(\gamma, \gamma) (\Omega_{\gamma\gamma}^{t,c} - \alpha_t^2 \mathcal{Q}_{\gamma\gamma}) \right] \\
&+ \beta_t (1 - b\beta_t) I_{\lambda^1 \sigma'}^{t,c}(\gamma, \gamma) \theta^{ck}(\varrho) \\
&+ \frac{\beta_t (1 - b\beta_t) \varrho_c m^k(\varrho)_\gamma}{\mathcal{Q}_{\gamma\gamma}^c \Omega_{\gamma\gamma}^{t,c} - \beta_t^2 (\mathcal{Q}_{\gamma\gamma}^c)^2} \left[(I_{(\lambda^1)^2 \sigma'}^{t,c}(\gamma, \gamma) - M_\gamma^c I_{\lambda^1 \sigma'}^{t,c}(\gamma, \gamma)) (\Omega_{\gamma\gamma}^{t,c} - \beta_t^2 \mathcal{Q}_{\gamma\gamma}^c) \right] \\
&- \alpha_t \beta_t b I_{\lambda^0 \sigma'}(\gamma, \gamma) \theta^{ck}(\varrho) \\
&- \frac{\mathcal{Q}_{\gamma\gamma} b \beta_t \alpha_t^2 \varrho_c m^k(\varrho)_\gamma}{\mathcal{Q}_{\gamma\gamma} \Omega_{\gamma\gamma}^{t,c} - \alpha_t^2 (\mathcal{Q}_{\gamma\gamma})^2} \left[I_{\lambda^0 \sigma' \omega}^{t,c}(\gamma, \gamma, \gamma) - \alpha_t I_{(\lambda^0)^2 \sigma'}^{t,c}(\gamma, \gamma) - M_\gamma^c \beta_t I_{\lambda^0 \sigma'}^{t,c}(\gamma, \gamma) \right] \\
&- \lambda m^k(\varrho)_\gamma
\end{aligned} \tag{51}$$

611 A.2.5 Overlap $g(\cdot)$

612 The expected SGD for $g(\varrho)$ can be derived along nearly identical lines. By the same token, for
613 $1 \leq i \leq R$, the decomposition

$$\mathbb{E}[dg_i(\varrho)] = \mathbb{E}_t \mathbb{E}_c [\mathbb{E}^c dg_i^{t,c}(\varrho)] \tag{52}$$

614 holds with

$$\begin{aligned}
\frac{d}{2\eta} \mathbb{E}^c[dg^{t,c}(\varrho)_\gamma] &= - \sum_{\delta=1}^r I_{\sigma\sigma}^{t,c}(\gamma, \delta) g_i(\varrho)_\delta + (1 - b\beta_t) I_{\sigma}^{t,c}(\gamma) p_i^c(\varrho) \\
&\quad + \frac{(1 - b\beta_t) \beta_t \varrho_c g_i(\varrho)_\gamma - b\alpha_t^2 g_i(\varrho)_\gamma}{\Omega_{\gamma\gamma}^{t,c}} (I_{\sigma\omega}^{t,c}(\gamma, \gamma) - \beta_t M_\gamma^c I_{\sigma}^{t,c}(\gamma)) \\
&\quad - \sum_{\delta \neq \gamma} \frac{(\alpha_t^2 + \beta_t^2 \varrho_c) \mathcal{Q}_{\gamma\delta}}{\Omega_{\gamma\gamma}^{t,c} \Omega_{\delta\delta}^{t,c} - (\Omega_{\gamma\delta}^{t,c})^2} \left[(I_{\sigma'\sigma\omega}^{t,c}(\gamma, \delta, \gamma) - \beta_t M_\gamma^c I_{\sigma'\sigma}^{t,c}(\gamma, \delta)) (\Omega_{\delta\delta}^{t,c} g_i(\varrho)_\gamma - \Omega_{\gamma\delta}^{t,c} g_i(\varrho)_\delta) \right. \\
&\quad \left. + (I_{\sigma'\sigma\omega}^{t,c}(\gamma, \delta, \delta) - \beta_t M_\delta^c I_{\sigma'\sigma}^{t,c}(\gamma, \delta)) (\Omega_{\gamma\gamma}^{t,c} g(\varrho)_\delta - \Omega_{\gamma\delta}^{t,c} g(\varrho)_\gamma) \right] \\
&\quad - \frac{\mathcal{Q}_{\gamma\gamma}(\alpha_t^2 + \beta_t^2 \varrho_c) g_i(\varrho)_\gamma}{\Omega_{\gamma\gamma}^{t,c}} (I_{\sigma'\sigma\omega}^{t,c}(\gamma, \gamma, \gamma) - \beta_t M_\gamma^c I_{\sigma'\sigma}^{t,c}(\gamma, \gamma)) - \beta_t \sum_{\delta=1}^r \mathcal{Q}_{\gamma\delta} I_{\sigma'\sigma}^{t,c}(\gamma, \delta) p_i^c(\varrho) \\
&\quad + \frac{\alpha_t^2 (1 - b\beta_t) \mathcal{Q}_{\gamma\gamma}^c g_i(\varrho)_\gamma}{\mathcal{Q}_{\gamma\gamma}^c \Omega_{\gamma\gamma}^{t,c} - \beta_t^2 (\mathcal{Q}_{\gamma\gamma}^c)^2} \left[I_{\lambda^1 \sigma'\omega}^{t,c}(\gamma, \gamma, \gamma) - \beta_t I_{(\lambda^1)^2 \sigma'}^{t,c}(\gamma, \gamma) \right] \\
&\quad - \frac{\alpha_t^2 b g_i(\varrho)_\gamma}{\mathcal{Q}_{\gamma\gamma} \Omega_{\gamma\gamma}^{t,c} - \alpha_t^2 (\mathcal{Q}_{\gamma\gamma})^2} \left[I_{(\lambda^0)^2 \sigma'}^{t,c}(\gamma, \gamma) (\Omega_{\gamma\gamma}^{t,c} - \alpha_t^2 \mathcal{Q}_{\gamma\gamma}) \right] \\
&\quad + \beta_t (1 - b\beta_t) I_{\lambda^1 \sigma'}^{t,c}(\gamma, \gamma) p_i^c(\varrho) \\
&\quad + \frac{\beta_t (1 - b\beta_t) \varrho_c g_i(\varrho)_\gamma}{\mathcal{Q}_{\gamma\gamma}^c \Omega_{\gamma\gamma}^{t,c} - \beta_t^2 (\mathcal{Q}_{\gamma\gamma}^c)^2} \left[(I_{(\lambda^1)^2 \sigma'}^{t,c}(\gamma, \gamma) - M_\gamma^c I_{\lambda^1 \sigma'}^{t,c}(\gamma, \gamma)) (\Omega_{\gamma\gamma}^{t,c} - \beta_t^2 \mathcal{Q}_{\gamma\gamma}^c) \right] \\
&\quad - \alpha_t \beta_t b I_{\lambda^0 \sigma'}(\gamma, \gamma) p_i^c(\varrho) \\
&\quad - \frac{\mathcal{Q}_{\gamma\gamma} b \beta_t \alpha_t^2 \varrho_c g_i(\varrho)_\gamma}{\mathcal{Q}_{\gamma\gamma} \Omega_{\gamma\gamma}^{t,c} - \alpha_t^2 (\mathcal{Q}_{\gamma\gamma})^2} \left[I_{\lambda^0 \sigma'\omega}^{t,c}(\gamma, \gamma, \gamma) - \alpha_t I_{(\lambda^0)^2 \sigma'}^{t,c}(\gamma, \gamma) - M_\gamma^c \beta_t I_{\lambda^0 \sigma'}^{t,c}(\gamma, \gamma) \right] \\
&\quad - \lambda g_i(\varrho)_\gamma, \tag{53}
\end{aligned}$$

615 yielding the increment of $g(\cdot)$ under the SGD dynamics.

616 A.2.6 Overlap $q(\cdot)$

617 We now turn to the summary statistic $q(\varrho)$ (31). First note that

$$\begin{aligned}
\mathbb{E}[d\mathcal{Q}] &= \frac{1}{d} \sum_{i=1}^d \mathbb{E}_c \left[w_i \mathbb{E}_t[\mathbb{E}^c dw_i^{t,c}]^\top + \mathbb{E}_t[\mathbb{E}^c dw_i^{t,c}] w_i^\top + \mathbb{E}_{t,t'}[\mathbb{E}^c dw_i^{t,c} (dw_i^{t',c})^\top] \right] \\
&\equiv \mathbb{E}_t \mathbb{E}_c[\mathbb{E}^c d\mathcal{Q}_{(1)}^{t,c}(\varrho)] + \mathbb{E}_{t,t'} \mathbb{E}_c[\mathbb{E}^c d\mathcal{Q}_{(2)}^{t,t',c}(\varrho)]. \tag{54}
\end{aligned}$$

618 We have separated the linear term and the quadratic term. It follows that the density statistic $q(\cdot)$ can
619 be similarly decomposed as

$$\mathbb{E}[dq(\varrho)] = \mathbb{E}_t \mathbb{E}_c[\mathbb{E}^c dq_{(1)}^{t,c}(\varrho)] + \mathbb{E}_{t,t'} \mathbb{E}_c[\mathbb{E}^c dq_{(2)}^{t,t',c}(\varrho)]. \tag{55}$$

620 In the following, we sequentially examine the linear and quadratic terms. The expected increment for
 621 the linear term $dq_{(1)}^{t,c}(\cdot)$ can be read from (21) as

$$\begin{aligned}
& \frac{d}{2\eta} \mathbb{E}^c[(dq_{(1)}^{t,c}(\varrho))_{\gamma\delta}] = \\
& - \sum_{\epsilon=1}^r I_{\sigma\sigma}^{t,c}(\gamma, \epsilon) q(\varrho)_{\delta\epsilon} + (1 - b\beta_t) I_{\sigma}^{t,c}(\gamma) m^c(\varrho)_{\delta} \\
& + \frac{((1 - b\beta_t)\beta_t \varrho_c - b\alpha_t^2) q(\varrho)_{\gamma\delta}}{\Omega_{\gamma\gamma}^{t,c}} (I_{\sigma\omega}^{t,c}(\gamma, \gamma) - \beta_t M_{\gamma}^c I_{\sigma}^{t,c}(\gamma)) \\
& - \sum_{\epsilon \neq \gamma} \frac{(\alpha_t^2 + \beta_t^2 \varrho_c) \mathcal{Q}_{\epsilon\gamma}}{\Omega_{\gamma\gamma}^{t,c} \Omega_{\epsilon\epsilon}^{t,c} - (\Omega_{\gamma\epsilon}^{t,c})^2} \left[(I_{\sigma'\sigma\omega}^{t,c}(\gamma, \epsilon, \gamma) - \beta_t M_{\gamma}^c I_{\sigma'\sigma}^{t,c}(\gamma, \epsilon)) (\Omega_{\epsilon\epsilon}^{t,c} q(\varrho)_{\gamma\delta} - \Omega_{\gamma\epsilon}^{t,c} q(\varrho)_{\epsilon\delta}) \right. \\
& \quad \left. + (I_{\sigma'\sigma\omega}^{t,c}(\gamma, \epsilon, \epsilon) - \beta_t M_{\epsilon}^c I_{\sigma'\sigma}^{t,c}(\gamma, \epsilon)) (\Omega_{\gamma\gamma}^{t,c} q(\varrho)_{\delta\epsilon} - \Omega_{\gamma\epsilon}^{t,c} q(\varrho)_{\gamma\delta}) \right] \\
& - \frac{(\alpha_t^2 + \beta_t^2 \varrho_c) \mathcal{Q}_{\gamma\gamma} q(\varrho)_{\gamma\delta}}{\Omega_{\gamma\gamma}^{t,c}} (I_{\sigma'\sigma\omega}^{t,c}(\gamma, \gamma, \gamma) - \beta_t M_{\gamma}^c I_{\sigma'\sigma}^{t,c}(\gamma, \gamma)) - \beta_t \sum_{\epsilon=1}^r \mathcal{Q}_{\epsilon\gamma} I_{\sigma'\sigma}^{t,c}(\gamma, \epsilon) m^c(\varrho)_{\delta} \\
& + \frac{\mathcal{Q}_{\gamma\gamma}^c \alpha_t^2 (1 - b\beta_t) q(\varrho)_{\gamma\delta}}{\mathcal{Q}_{\gamma\gamma}^c \Omega_{\gamma\gamma}^{t,c} - \beta_t^2 (\mathcal{Q}_{\gamma\gamma}^c)^2} \left[I_{\lambda^1 \sigma'\omega}^{t,c}(\gamma, \gamma, \gamma) - \beta_t I_{(\lambda^1)^2 \sigma'}^{t,c}(\gamma, \gamma) \right] \\
& - \frac{\alpha_t^2 b q(\varrho)_{\gamma\delta}}{\mathcal{Q}_{\gamma\gamma} \Omega_{\gamma\gamma}^{t,c} - \alpha_t^2 (\mathcal{Q}_{\gamma\gamma})^2} \left[I_{(\lambda^0)^2 \sigma'}^{t,c}(\gamma, \gamma) (\Omega_{\gamma\gamma}^{t,c} - \alpha_t^2 \mathcal{Q}_{\gamma\gamma}) \right] + \beta_t (1 - b\beta_t) I_{\lambda^1 \sigma'}^{t,c}(\gamma, \gamma) m^c(\varrho)_{\delta} \\
& + \frac{\beta_t (1 - b\beta_t) q(\varrho)_{\gamma\delta} \varrho^c}{\mathcal{Q}_{\gamma\gamma}^c \Omega_{\gamma\gamma}^{t,c} - \beta_t^2 (\mathcal{Q}_{\gamma\gamma}^c)^2} \left[(I_{(\lambda^1)^2 \sigma'}^{t,c}(\gamma, \gamma) - M_{\gamma}^c I_{\lambda^1 \sigma'}^{t,c}(\gamma, \gamma)) (\Omega_{\gamma\gamma}^{t,c} - \beta_t^2 \mathcal{Q}_{\gamma\gamma}^c) \right] - \alpha_t \beta_t b I_{\lambda^0 \sigma'}^{t,c}(\gamma, \gamma) m^c(\varrho)_{\delta} \\
& - \frac{\alpha_t^2 \beta_t b \mathcal{Q}_{\gamma\gamma} \varrho^c q(\varrho)_{\gamma\delta}}{\mathcal{Q}_{\gamma\gamma} \Omega_{\gamma\gamma}^{t,c} - \alpha_t^2 (\mathcal{Q}_{\gamma\gamma})^2} \left[I_{\lambda^0 \sigma'\omega}^{t,c}(\gamma, \gamma, \gamma) - \alpha_t I_{(\lambda^0)^2 \sigma'}^{t,c}(\gamma, \gamma) - M_{\gamma}^c \beta_t I_{\lambda^0 \sigma'}^{t,c}(\gamma, \gamma) \right] \\
& - \lambda q(\varrho)_{\gamma\delta} \\
& + (\gamma \leftrightarrow \delta)
\end{aligned} \tag{56}$$

622 We now turn to the quadratic term $dq_{(2)}^{t,c}(\cdot)$. Keeping only leading order terms,

$$\begin{aligned}
& \frac{d}{4\eta^2} \mathbb{E}^c[(dq_{(2)}^{t,c}(\varrho))_{\gamma\delta}] = \frac{1}{2} \nu(\varrho) I_{\sigma\sigma}^{t,t',c}(\gamma, \delta) [(1 - b\beta_t)(1 - b\beta_{t'}) \varrho_c + b^2 \alpha_t \alpha_{t'}] \\
& - \nu(\varrho) \sum_{\epsilon=1}^r I_{\sigma\sigma\sigma'}^{t,t',c}(\gamma, \epsilon, \delta) \mathcal{Q}_{\epsilon\delta} [\beta_{t'} (1 - b\beta_t) \varrho_c - b\alpha_t \alpha_{t'}] \\
& + \nu(\varrho) \left((1 - b\beta_{t'}) I_{\sigma\sigma'\lambda^1}^{t,t',c}(\gamma, \delta, \delta) - b\alpha_{t'} I_{\sigma\sigma'\lambda^0}^{t,t',c}(\gamma, \delta, \delta) \right) [\beta_{t'} (1 - b\beta_t) \varrho_c - b\alpha_t \alpha_{t'}] \\
& + \frac{1}{2} \nu(\varrho) \sum_{\epsilon, \iota}^r I_{\sigma'\sigma\sigma'\sigma}^{t,t,t',c}(\gamma, \epsilon, \delta, \iota) \mathcal{Q}_{\gamma\epsilon} \mathcal{Q}_{\delta\iota} [\beta_t \beta_{t'} \varrho_c + \alpha_t \alpha_{t'}] \\
& - \nu(\varrho) \sum_{\epsilon=1}^r \left((1 - b\beta_{t'}) I_{\sigma'\sigma\sigma'\lambda^1}^{t,t,t',c}(\gamma, \epsilon, \delta, \delta) - b\alpha_{t'} I_{\sigma'\sigma\sigma'\lambda^0}^{t,t,t',c}(\gamma, \epsilon, \delta, \delta) \right) \mathcal{Q}_{\epsilon\gamma} [\beta_t \beta_{t'} \varrho_c + \alpha_t \alpha_{t'}] \\
& + \frac{1}{2} \left((1 - b\beta_{t'}) (1 - b\beta_t) I_{\sigma'\sigma'\lambda^1\lambda^1}^{t,t',t,t',c}(\gamma, \delta, \gamma, \delta) - (1 - b\beta_{t'}) b\alpha_t I_{\sigma'\sigma'\lambda^0\lambda^1}^{t,t',t,t',c}(\gamma, \delta, \gamma, \delta) \right. \\
& \quad \left. - b\alpha_{t'} (1 - b\beta_t) I_{\sigma'\sigma'\lambda^0\lambda^1}^{t,t',t,t',c}(\gamma, \delta, \delta, \gamma) + b^2 \alpha_t \alpha_{t'} I_{\sigma'\sigma'\lambda^0\lambda^0}^{t,t',t,t',c}(\gamma, \delta, \gamma, \delta) \right) \nu(\varrho) (\beta_t \beta_{t'} \varrho_c + \alpha_t \alpha_{t'}) \\
& + (\gamma \leftrightarrow \delta)
\end{aligned} \tag{57}$$

$$I_{\sigma\sigma}^{t,t',c}(\gamma, \delta) = \mathbb{E}_{\omega_\gamma, \omega_\delta} [\sigma(\omega_\gamma + v_\gamma p_t) \sigma(\omega_\delta + v_\delta p_t)]$$

$$\omega_\gamma, \omega_\delta \sim \mathcal{N} \left((\beta_t, \beta_{t'}) \odot M_{(\gamma, \delta)}^c, \Omega_{(\gamma, \delta)}^{t,t',c} \right) \quad (58)$$

$$I_{\sigma\sigma\sigma'}^{t_1, t_2, t_3, c}(\gamma, \epsilon, \delta) = \mathbb{E}_{\omega_\gamma, \omega_\epsilon, \omega_\delta} [\sigma(\omega_\gamma + v_\gamma p_t) \sigma(\omega_\epsilon + v_\epsilon p_t) \sigma'(\omega_\delta + v_\delta p_t)],$$

$$\omega_\gamma, \omega_\epsilon, \omega_\delta \sim \mathcal{N} \left((\beta_{t_1}, \beta_{t_2}, \beta_{t_3}) \odot M_{(\gamma, \epsilon, \delta)}^c, \Omega_{(\gamma, \epsilon, \delta)}^{(3), t_1, t_2, t_3, c} \right) \quad (59)$$

$$I_{\sigma\sigma', \lambda^1}^{t_1, t_2, t_3, c}(\gamma, \epsilon, \delta) = \mathbb{E}_{\omega_\gamma, \omega_\epsilon, \lambda_\delta^1} [\sigma(\omega_\gamma + v_\gamma p_t) \sigma'(\omega_\epsilon + v_\epsilon p_t) \lambda_\delta^1],$$

$$\omega_\gamma, \omega_\epsilon, \lambda_\delta^1 \sim \mathcal{N} \left((\beta_{t_1}, \beta_{t_2}, 1) \odot M_{(\gamma, \epsilon, \delta)}^c, \Phi_{(\gamma, \epsilon, \delta)}^{(3), t_1, t_2, t_3, c} \right) \quad (60)$$

$$I_{\sigma\sigma', \lambda^0}^{t_1, t_2, t_3, c}(\gamma, \epsilon, \delta) = \mathbb{E}_{\omega_\gamma, \omega_\epsilon, \lambda_\delta^0} [\sigma(\omega_\gamma + v_\gamma p_t) \sigma'(\omega_\epsilon + v_\epsilon p_t) \lambda_\delta^0],$$

$$\omega_\gamma, \omega_\epsilon, \lambda_\delta^0 \sim \mathcal{N} \left((\beta_{t_1}, \beta_{t_2}, 0) \odot M_{(\gamma, \epsilon, \delta)}^c, \Psi_{(\gamma, \epsilon, \delta)}^{(3), t_1, t_2, t_3, c} \right) \quad (61)$$

$$I_{\sigma'\sigma\sigma'}^{t_1, t_2, t_3, t_4, c}(\gamma, \epsilon, \delta, \ell) = \mathbb{E}_{\omega_\gamma, \omega_\epsilon, \omega_\delta, \omega_\ell} [\sigma'(\omega_\gamma + v_\gamma p_t) \sigma(\omega_\epsilon + v_\epsilon p_t) \sigma'(\omega_\delta) \sigma(\omega_\ell + v_\ell p_t)]$$

$$\omega_\gamma, \omega_\epsilon, \omega_\delta, \omega_\ell \sim \mathcal{N} \left((\beta_{t_1}, \beta_{t_2}, \beta_{t_3}, \beta_{t_4}) \odot M_{(\gamma, \epsilon, \delta, \ell)}^c, \Omega_{(\gamma, \epsilon, \delta, \ell)}^{(4), t_1, t_2, t_3, t_4, c} \right) \quad (62)$$

$$I_{\sigma'\sigma\sigma', \lambda^1}^{t_1, t_2, t_3, t_4, c}(\gamma, \epsilon, \delta, \ell) = \mathbb{E}_{\omega_\gamma, \omega_\epsilon, \omega_\delta, \lambda_\ell^1} [\sigma'(\omega_\gamma + v_\gamma p_t) \sigma(\omega_\epsilon + v_\epsilon p_t) \sigma'(\omega_\delta + v_\delta p_t) \lambda_\ell^1]$$

$$\omega_\gamma, \omega_\epsilon, \omega_\delta, \lambda_\ell^1 \sim \mathcal{N} \left((\beta_{t_1}, \beta_{t_2}, \beta_{t_3}, 1) \odot M_{(\gamma, \epsilon, \delta, \ell)}^c, \Phi_{(\gamma, \epsilon, \delta, \ell)}^{(4), t_1, t_2, t_3, t_4, c} \right) \quad (63)$$

$$I_{\sigma'\sigma\sigma', \lambda^0}^{t_1, t_2, t_3, t_4, c}(\gamma, \epsilon, \delta, \ell) = \mathbb{E}_{\omega_\gamma, \omega_\epsilon, \omega_\delta, \lambda_\ell^0} [\sigma'(\omega_\gamma + v_\gamma p_t) \sigma(\omega_\epsilon + v_\epsilon p_t) \sigma'(\omega_\delta + v_\delta p_t) \lambda_\ell^0]$$

$$\omega_\gamma, \omega_\epsilon, \omega_\delta, \lambda_\ell^0 \sim \mathcal{N} \left((\beta_{t_1}, \beta_{t_2}, \beta_{t_3}, 0) \odot M_{(\gamma, \epsilon, \delta, \ell)}^c, \Psi_{(\gamma, \epsilon, \delta, \ell)}^{(4), t_1, t_2, t_3, t_4, c} \right) \quad (64)$$

$$I_{\sigma'\sigma', \lambda^1 \lambda^1}^{t_1, t_2, t_3, t_4, c}(\gamma, \epsilon, \delta, \ell) = \mathbb{E}_{\omega_\gamma, \omega_\epsilon, \lambda_\delta^1, \lambda_\ell^1} [\sigma'(\omega_\gamma + v_\gamma p_t) \sigma'(\omega_\epsilon + v_\epsilon p_t) \lambda_\delta^1 \lambda_\ell^1]$$

$$\omega_\gamma, \omega_\epsilon, \lambda_\delta^1, \lambda_\ell^1 \sim \mathcal{N} \left((\beta_{t_1}, \beta_{t_2}, 1, 1) \odot M_{(\gamma, \epsilon, \delta, \ell)}^c, P_{(\gamma, \epsilon, \delta, \ell)}^{(4, 1, 1), t_1, t_2, t_3, t_4, c} \right) \quad (65)$$

$$I_{\sigma'\sigma', \lambda^0 \lambda^1}^{t_1, t_2, t_3, t_4, c}(\gamma, \epsilon, \delta, \ell) = \mathbb{E}_{\omega_\gamma, \omega_\epsilon, \lambda_\delta^0, \lambda_\ell^1} [\sigma'(\omega_\gamma + v_\gamma p_t) \sigma'(\omega_\epsilon + v_\epsilon p_t) \lambda_\delta^0 \lambda_\ell^1]$$

$$\omega_\gamma, \omega_\epsilon, \lambda_\delta^0, \lambda_\ell^1 \sim \mathcal{N} \left((\beta_{t_1}, \beta_{t_2}, 0, 1) \odot M_{(\gamma, \epsilon, \delta, \ell)}^c, P_{(\gamma, \epsilon, \delta, \ell)}^{(4, 0, 1), t_1, t_2, t_3, t_4, c} \right) \quad (66)$$

$$I_{\sigma'\sigma', \lambda^0 \lambda^0}^{t_1, t_2, t_3, t_4, c}(\gamma, \epsilon, \delta, \ell) = \mathbb{E}_{\omega_\gamma, \omega_\epsilon, \lambda_\delta^0, \lambda_\ell^0} [\sigma'(\omega_\gamma + v_\gamma p_t) \sigma'(\omega_\epsilon + v_\epsilon p_t) \lambda_\delta^0 \lambda_\ell^0]$$

$$\omega_\gamma, \omega_\epsilon, \lambda_\delta^0, \lambda_\ell^0 \sim \mathcal{N} \left((\beta_{t_1}, \beta_{t_2}, 0, 0) \odot M_{(\gamma, \epsilon, \delta, \ell)}^c, P_{(\gamma, \epsilon, \delta, \ell)}^{(4, 0, 0), t_1, t_2, t_3, t_4, c} \right). \quad (67)$$

624 We further denoted

$$\begin{aligned}
\Omega^{t,t',c} &= \alpha_t \alpha_{t'} \mathcal{Q} + \beta_t \beta_{t'} Q^c \\
\Omega_{(\gamma,\epsilon,\delta)}^{(3),t_1,t_2,t_3,c} &= \begin{pmatrix} \Omega_{\gamma\gamma}^{t_1,t_1,c} & \Omega_{\gamma\epsilon}^{t_1,t_2,c} & \Omega_{\gamma\delta}^{t_1,t_3,c} \\ \Omega_{\epsilon\gamma}^{t_2,t_1,c} & \Omega_{\epsilon\epsilon}^{t_2,t_2,c} & \Omega_{\epsilon\delta}^{t_2,t_3,c} \\ \Omega_{\delta\gamma}^{t_3,t_1,c} & \Omega_{\delta\epsilon}^{t_3,t_2,c} & \Omega_{\delta\delta}^{t_3,t_3,c} \end{pmatrix} \\
\Phi_{(\gamma,\epsilon,\delta)}^{(3),t_1,t_2,t_3,c} &= \begin{pmatrix} \Omega_{\gamma\gamma}^{t_1,t_1,c} & \Omega_{\gamma\epsilon}^{t_1,t_2,c} & \beta_{t_1} Q_{\gamma\delta}^c \\ \Omega_{\epsilon\gamma}^{t_2,t_1,c} & \Omega_{\epsilon\epsilon}^{t_2,t_2,c} & \beta_{t_2} Q_{\epsilon\delta}^c \\ \beta_{t_1} Q_{\gamma\delta}^c & \beta_{t_2} Q_{\epsilon\delta}^c & Q_{\delta\delta}^c \end{pmatrix} \\
\Psi_{(\gamma,\epsilon,\delta)}^{(3),t_1,t_2,t_3,c} &= \begin{pmatrix} \Omega_{\gamma\gamma}^{t_1,t_1,c} & \Omega_{\gamma\epsilon}^{t_1,t_2,c} & \alpha_{t_1} \mathcal{Q}_{\gamma\delta} \\ \Omega_{\epsilon\gamma}^{t_2,t_1,c} & \Omega_{\epsilon\epsilon}^{t_2,t_2,c} & \alpha_{t_2} \mathcal{Q}_{\epsilon\delta} \\ \alpha_{t_1} Q_{\gamma\delta}^c & \alpha_{t_2} \mathcal{Q}_{\epsilon\delta} & \mathcal{Q}_{\delta\delta} \end{pmatrix} \\
\Omega_{(\gamma,\epsilon,\delta,\iota)}^{(4),t_1,t_2,t_3,t_4,c} &= \begin{pmatrix} \Omega_{\gamma\gamma}^{t_1,t_1,c} & \Omega_{\gamma\epsilon}^{t_1,t_2,c} & \Omega_{\gamma\delta}^{t_1,t_3,c} & \Omega_{\gamma\iota}^{t_1,t_4,c} \\ \Omega_{\epsilon\gamma}^{t_2,t_1,c} & \Omega_{\epsilon\epsilon}^{t_2,t_2,c} & \Omega_{\epsilon\delta}^{t_2,t_3,c} & \Omega_{\epsilon\iota}^{t_2,t_4,c} \\ \Omega_{\delta\gamma}^{t_3,t_1,c} & \Omega_{\delta\epsilon}^{t_3,t_2,c} & \Omega_{\delta\delta}^{t_3,t_3,c} & \Omega_{\delta\iota}^{t_3,t_4,c} \\ \Omega_{\iota\gamma}^{t_4,t_1,c} & \Omega_{\iota\epsilon}^{t_4,t_2,c} & \Omega_{\iota\delta}^{t_4,t_3,c} & \Omega_{\iota\iota}^{t_4,t_4,c} \end{pmatrix} \\
\Phi_{(\gamma,\epsilon,\delta,\iota)}^{(4),t_1,t_2,t_3,t_4,c} &= \begin{pmatrix} \Omega_{\gamma\gamma}^{t_1,t_1,c} & \Omega_{\gamma\epsilon}^{t_1,t_2,c} & \Omega_{\gamma\delta}^{t_1,t_3,c} & \beta_{t_1} Q_{\gamma\iota}^c \\ \Omega_{\epsilon\gamma}^{t_2,t_1,c} & \Omega_{\epsilon\epsilon}^{t_2,t_2,c} & \Omega_{\epsilon\delta}^{t_2,t_3,c} & \beta_{t_2} Q_{\epsilon\iota}^c \\ \Omega_{\delta\gamma}^{t_3,t_1,c} & \Omega_{\delta\epsilon}^{t_3,t_2,c} & \Omega_{\delta\delta}^{t_3,t_3,c} & \beta_{t_3} Q_{\delta\iota}^c \\ \beta_{t_1} Q_{\gamma\iota}^c & \beta_{t_2} Q_{\epsilon\iota}^c & \beta_{t_3} Q_{\delta\iota}^c & Q_{\iota\iota}^c \end{pmatrix} \\
\Psi_{(\gamma,\epsilon,\delta,\iota)}^{(4),t_1,t_2,t_3,t_4,c} &= \begin{pmatrix} \Omega_{\gamma\gamma}^{t_1,t_1,c} & \Omega_{\gamma\epsilon}^{t_1,t_2,c} & \Omega_{\gamma\delta}^{t_1,t_3,c} & \alpha_{t_1} \mathcal{Q}_{\gamma\iota} \\ \Omega_{\epsilon\gamma}^{t_2,t_1,c} & \Omega_{\epsilon\epsilon}^{t_2,t_2,c} & \Omega_{\epsilon\delta}^{t_2,t_3,c} & \alpha_{t_2} \mathcal{Q}_{\epsilon\iota} \\ \Omega_{\delta\gamma}^{t_3,t_1,c} & \Omega_{\delta\epsilon}^{t_3,t_2,c} & \Omega_{\delta\delta}^{t_3,t_3,c} & \alpha_{t_3} \mathcal{Q}_{\delta\iota} \\ \alpha_{t_1} \mathcal{Q}_{\gamma\iota} & \alpha_{t_2} \mathcal{Q}_{\epsilon\iota} & \alpha_{t_3} \mathcal{Q}_{\delta\iota} & \mathcal{Q}_{\iota\iota} \end{pmatrix} \\
P_{(\gamma,\epsilon,\delta,\iota)}^{(4,1,1),t_1,t_2,t_3,t_4,c} &= \begin{pmatrix} \Omega_{\gamma\gamma}^{t_1,t_1,c} & \Omega_{\gamma\epsilon}^{t_1,t_2,c} & \beta_{t_1} Q_{\gamma\delta}^c & \beta_{t_1} Q_{\gamma\iota}^c \\ \Omega_{\epsilon\gamma}^{t_2,t_1,c} & \Omega_{\epsilon\epsilon}^{t_2,t_2,c} & \beta_{t_2} Q_{\epsilon\delta}^c & \beta_{t_2} Q_{\epsilon\iota}^c \\ \beta_{t_1} Q_{\gamma\delta}^c & \beta_{t_2} Q_{\epsilon\delta}^c & Q_{\delta\delta}^c & Q_{\delta\iota}^c \\ \beta_{t_1} Q_{\gamma\iota}^c & \beta_{t_2} Q_{\epsilon\iota}^c & Q_{\iota\delta}^c & Q_{\iota\iota}^c \end{pmatrix} \\
P_{(\gamma,\epsilon,\delta,\iota)}^{(4,0,1),t_1,t_2,t_3,t_4,c} &= \begin{pmatrix} \Omega_{\gamma\gamma}^{t_1,t_1,c} & \Omega_{\gamma\epsilon}^{t_1,t_2,c} & \alpha_{t_1} \mathcal{Q}_{\gamma\delta} & \beta_{t_1} Q_{\gamma\iota}^c \\ \Omega_{\epsilon\gamma}^{t_2,t_1,c} & \Omega_{\epsilon\epsilon}^{t_2,t_2,c} & \alpha_{t_2} \mathcal{Q}_{\epsilon\delta} & \beta_{t_2} Q_{\epsilon\iota}^c \\ \alpha_{t_1} \mathcal{Q}_{\gamma\delta} & \alpha_{t_2} \mathcal{Q}_{\epsilon\delta} & \mathcal{Q}_{\delta\delta} & 0 \\ \beta_{t_1} Q_{\gamma\iota}^c & \beta_{t_2} Q_{\epsilon\iota}^c & 0 & Q_{\iota\iota}^c \end{pmatrix} \\
P_{(\gamma,\epsilon,\delta,\iota)}^{(4,0,0),t_1,t_2,t_3,t_4,c} &= \begin{pmatrix} \Omega_{\gamma\gamma}^{t_1,t_1,c} & \Omega_{\gamma\epsilon}^{t_1,t_2,c} & \alpha_{t_1} \mathcal{Q}_{\gamma\delta} & \alpha_{t_1} \mathcal{Q}_{\gamma\iota} \\ \Omega_{\epsilon\gamma}^{t_2,t_1,c} & \Omega_{\epsilon\epsilon}^{t_2,t_2,c} & \alpha_{t_2} \mathcal{Q}_{\epsilon\delta} & \alpha_{t_2} \mathcal{Q}_{\epsilon\iota} \\ \alpha_{t_1} \mathcal{Q}_{\gamma\delta} & \alpha_{t_2} \mathcal{Q}_{\epsilon\delta} & \mathcal{Q}_{\delta\delta} & \mathcal{Q}_{\delta\iota} \\ \alpha_{t_1} \mathcal{Q}_{\gamma\iota} & \alpha_{t_2} \mathcal{Q}_{\epsilon\iota} & \mathcal{Q}_{\iota\delta} & \mathcal{Q}_{\iota\iota} \end{pmatrix} \tag{68}
\end{aligned}$$

625 A.3 Update for v

626 Finally, we can ascertain the asymptotic evolution of the time encoding weights v . In the considered
627 limit, the update dv again concentrates. As above, let us decompose the expected increment as

$$628 \quad \mathbb{E}[dv_\gamma] = \mathbb{E}_t \mathbb{E}_c [\mathbb{E}^c dv_\gamma^{t,c}], \tag{69}$$

with

$$\mathbb{E}^c dv_\gamma^{t,c} = -\frac{2\eta}{d} \left[\sum_{\delta} \mathcal{Q}_{\gamma\delta} I_{\sigma'\sigma}^{t,c}(\gamma, \delta) - (1 - b\beta_t) I_{\lambda^1\sigma'}^{t,c}(\gamma, \gamma) + b\alpha_t I_{\lambda^0\sigma'}^{t,c}(\gamma, \gamma) + \lambda v_\gamma \right] \tag{70}$$

629 A.4 Continuous time limit

630 Equations (51),(53),(56) and (57) provide the update equations for the summary statistic densities
 631 $m(\cdot), g(\cdot), q(\cdot)$ under SGD steps (4), which take the form

$$\begin{aligned}\frac{d}{2\eta}dm(\varrho) &= F_m(\varrho, m(\varrho), q(\varrho), M, Q, \mathcal{Q}, b), \\ \frac{d}{2\eta}dg &= F_g(\varrho, m(\varrho), q(\varrho), M, Q, \mathcal{Q}, b), \\ \frac{d}{2\eta}dq &= F_q(\varrho, m(\varrho), q(\varrho), M, Q, \mathcal{Q}, b),\end{aligned}\tag{71}$$

632 where the update functions $F_{m,q,g}$ denote the right hand sides of (51),(53),(56) and (57), and we have
 633 omitted the time step indices to ease the notations. From (44), these updates translate directly at the
 634 level of the summary statistics M, Q, \mathcal{Q}, G into

$$\begin{aligned}\frac{d}{2\eta}dM^c &= F_M(M, Q, \mathcal{Q}, b, v)^c, & F_M(\cdot)^c &= \int F_m(\varrho, m(\varrho), q(\varrho), \cdot)^c d\varrho, \\ \frac{d}{2\eta}dG &= F_G(M, Q, \mathcal{Q}, b, v), & F_G(\cdot) &= \int F_g(\varrho, m(\varrho), q(\varrho), \cdot) d\varrho, \\ \frac{d}{2\eta}dQ^c &= F_Q(M, Q, \mathcal{Q}, b, v)^c, & F_Q(\cdot)^c &= \int \varrho_c F_q(\varrho, m(\varrho), q(\varrho), \cdot) d\varrho, \\ \frac{d}{2\eta}d\mathcal{Q} &= F_{\mathcal{Q}}(M, Q, \mathcal{Q}, b, v), & F_{\mathcal{Q}}(\cdot) &= \int F_q(\varrho, m(\varrho), q(\varrho), \cdot) d\varrho.\end{aligned}\tag{72}$$

635 We remind that from (17), the skip connection strength similarly obeys

$$\frac{d}{2\eta}db = F_b(b),\tag{73}$$

636 where the update function F_b corresponds to the right hand side of equation (17). Similarly, from
 637 (70), the time encoding weights obey

$$\frac{d}{2\eta}dv = F_v(M, Q, \mathcal{Q}, b, v),\tag{74}$$

638 where F_v corresponds to the right-hand side of (70). Now remark that in the asymptotic limit $d \rightarrow \infty$,
 639 the coefficient $d/2\eta$ tends to zero. Introducing the time variable $\vartheta \equiv 2\eta\mu/d$, so that $d\vartheta = 2\eta/d$, the
 640 discrete processes (72) and (73) are thus asymptotically described by the limiting ODEs

$$\begin{aligned}\frac{dM}{d\vartheta} &= F_M(M, Q, \mathcal{Q}, b, v), \\ \frac{dQ}{d\vartheta} &= F_Q(M, Q, \mathcal{Q}, b, v), \\ \frac{d\mathcal{Q}}{d\vartheta} &= F_{\mathcal{Q}}(M, Q, \mathcal{Q}, b, v), \\ \frac{db}{d\vartheta} &= F_b(b), \\ \frac{dv}{d\vartheta} &= F_v(M, Q, \mathcal{Q}, b, v)\end{aligned}\tag{75}$$

641 Finally, the ODE for b , governing the dynamics of the skip connection strength b over the SGD
 642 optimization dynamics, can be solved in closed-form as

$$b(\vartheta) = \frac{\Lambda \mathbb{E}_t[\beta_t]}{\Lambda \mathbb{E}_t[\beta_t^2] + \mathbb{E}_t[\alpha_t^2]} \left[1 - e^{-(\Lambda \mathbb{E}_t[\beta_t^2] + \mathbb{E}_t[\alpha_t^2])\vartheta} \right] + b_0 e^{-(\Lambda \mathbb{E}_t[\beta_t^2] + \mathbb{E}_t[\alpha_t^2])\vartheta},\tag{76}$$

643 where b_0 designates the value of b at initialization. This completes the derivation of Result 2.1. \square

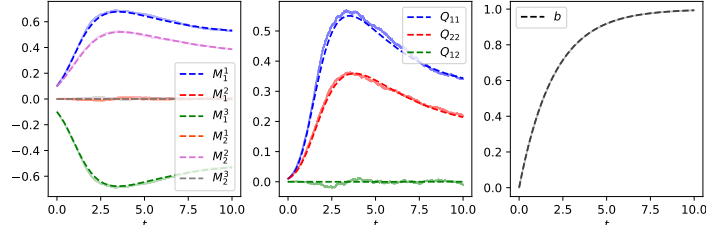


Figure 5: Evolution of the summary statistics (31) M (left), Q (middle) and skip connection strength b (right), characterizing the dynamics of the AE parameters (3) under SGD dynamics (4). Parameters $\sigma = \tanh, r = 2, \lambda = 0, \eta = 0.2, \mathcal{G} = \{1/2\}$ were used, and the target density ρ was taken to be a Gaussian mixture with three isotropic clusters (see also Fig. 7 in the main text). The weight vectors were initialized along the centroids of the target density, with norm 0.1, while the initial skip connection strength is $b_0 = 0$. Dashed lines: theoretical characterization of Result 2.1. Continuous lines: numerical experiments in $d = 1000$, for a single run.

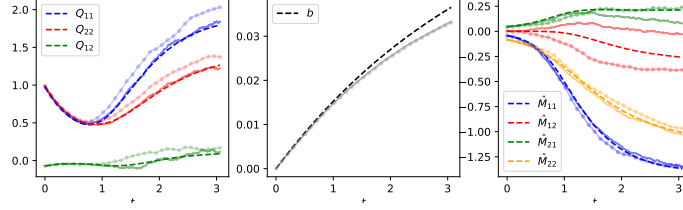


Figure 6: Evolution of the summary statistics (31) Q (left), b (middle) and skip connection strength M (right), characterizing the dynamics of the AE parameters (3) under SGD dynamics (4). Parameters $\sigma = \tanh, r = 2, \lambda = 0.784, \eta = 0.2, \mathcal{G} = \{1/2\}$ were used; weights were initialized with random independent Gaussian components and $b_0 = 0$. Dotted continuous lines : numerical experiments for a target density ρ given by the set of MNIST sevens. Continuous lines: numerical experiments for a unimodal Gaussian target distribution with covariance matching that of the set of MNIST sevens. Dashed lines: theoretical predictions of Result 2.1 for the latter Gaussian target density.

644 A.5 Numerical validation

645 We plot the theoretical predictions of Result 2.1 for the evolution of the summary statistics
646 M, Q, \mathcal{Q}, G, b (31) under the SGD dynamics (4) in Fig. 5 for a Gaussian mixture target density
647 ρ with three isotropic modes, learnt by an AE with $r = 2$ hidden units and \tanh activation, using
648 learning rate $\eta = 0.2$ and weight decay $\lambda = 0$. The centroids of the clusters were taken as $\pm e_1, e_2$
649 for two orthonormal vectors e_1, e_2 , and the columns of the weight matrix w were initialized with a
650 warm start as $0.1 \times e_{1,2}$. Finally, for simplicity, the expectation \mathbb{E}_t in (4) was chosen to bear over a
651 delta distribution around $\mathcal{G} = \{1/2\}$, instead of the full integral over $[0, 1]$. Including more points in
652 the grid \mathcal{G} was not found to significantly alter the qualitative aspect of the generated density. Fig. 5
653 reveals an overall good agreement between the theoretical predictions of Result 2.1 (dashed lines) and
654 numerical experiments (solid lines), obtained by simulating the model in large but finite dimension
655 $d = 1000$.

656 Fig. 6 similarly contrasts numerical experiments for a target distribution corresponding to MNIST
657 images of sevens (dotted lines), a Gaussian target density with matching covariance (solid lines), and
658 the theoretical predictions of Result 2.1 for the latter. All experimental details are specified in the
659 caption. Although the agreement between the three curves is overall good, discrepancies appear, in
660 particular due to the rather low dimensionality $d = 784$.

661 A.6 Extensions

662 We briefly describe, for completeness, how the analysis can be generalized to characterize the
 663 learning of more complex DAE architectures. Namely, we discuss how the derivation can be adapted
 664 to accommodate (a) untied weights and (b) time encodings.

665 **Untying the weights** – The analysis reported in the present appendix can be extended to untied
 666 DAE architectures of the form

$$f_{b,u,v}(x) = b \times x + \frac{u}{\sqrt{d}} \sigma \left(\frac{v^\top x}{\sqrt{d}} \right), \quad (77)$$

667 trained with online SGD

$$b_{\mu+1} - b_\mu = -\frac{\eta}{d^2} \left(\partial_b \mathbb{E}_t \|x_1^\mu - f_{b_\mu, u_\mu, v_\mu}(\alpha_t x_0^\mu + \beta_t x_1^\mu)\|^2 \right), \quad (78)$$

$$u_{\mu+1} - u_\mu = -\eta \nabla_u \mathbb{E}_t \|x_1^\mu - f_{b_\mu, u_\mu, v_\mu}(\alpha_t x_0^\mu + \beta_t x_1^\mu)\|^2 - \eta \frac{\lambda}{d} u_\mu, \quad (79)$$

$$v_{\mu+1} - v_\mu = -\eta \nabla_v \mathbb{E}_t \|x_1^\mu - f_{b_\mu, u_\mu, v_\mu}(\alpha_t x_0^\mu + \beta_t x_1^\mu)\|^2 - \eta \frac{\lambda}{d} v_\mu. \quad (80)$$

668 Such an extension, however, comes at the price of more cumbersome expressions, as the summary
 669 statistics (31) needs to be introduced for the two sets of weights u, v , in addition to cross-statistics
 670 of the form $u^\top \Sigma(c) v / d$ and $u^\top v / d$. We refer the interested reader to Appendix B of [52] where such a
 671 derivation is detailed, in a closely related setting. Experimentally, in the probed settings, we did not
 672 observe a significant effect of (un)tying the weights on the qualitative phenomenology discussed in
 673 the main text.

674 A.7 Example : linear model, Gaussian mixture

675 The previous analysis provides a tight characterization of the training dynamics of the non-linear
 676 DAE (3) under the SGD dynamics (4). For completeness and intuition, we conclude the present
 677 appendix by expounding a special simple case where the limiting ODEs of Result 2.1 admit a compact,
 678 closed-form expression, and consider the linear case $\sigma(x) = x, r = 1, p_t = 0$, when learning a
 679 target binary Gaussian mixture with isotropic clusters $\rho = 1/2 \mathcal{N}(\mu, I_d) + 1/2 \mathcal{N}(-\mu, I_d)$. We assume
 680 the squared norm $\|\mu\|^2$ asymptotically concentrates and denote by ρ its limiting value. Finally, we
 681 consider the limit of vanishing learning rate $\eta \rightarrow 0$. In this limit, the quadratic term $q_{(2)}^{t,c}$, which is of
 682 order $\Theta(\eta^2)$, can be neglected. The limiting ODEs (75) simplify in this case to

$$\begin{aligned} \frac{d}{d\vartheta} M &= -\mathbb{E}_t [\beta_t^2 M^2 + \mathcal{Q}(\beta_t^2 + \alpha_t^2) + \mathcal{Q}(\alpha_t^2 + \beta_t^2(1 + \rho)) - 2((1 - b\beta_t)\beta_t(\rho + 1) - \alpha_t^2 b) + \lambda] M \\ \frac{d}{d\vartheta} \mathcal{Q} &= -\mathbb{E}_t \left[2\mathcal{Q}(\beta_t^2 M^2 + \mathcal{Q}(\beta_t^2 + \alpha_t^2)) - 2((1 - b\beta_t)\beta_t(M^2 + \mathcal{Q}) - \alpha_t^2 b\mathcal{Q}) + \lambda\mathcal{Q} \right]. \end{aligned} \quad (81)$$

683 The evolution of the skip connection b , on the other hand, remains unchanged from (76). Let us now
 684 seek a solution of (81) at convergence, in the limit $\vartheta \rightarrow \infty$. It can be straightforwardly verified that,

$$M = \rho \mathcal{Q}, \quad \mathcal{Q} = \frac{\frac{\mathbb{E}_t[\beta_t] \mathbb{E}_t[\alpha_t^2]}{\mathbb{E}_t[\beta_t^2 + \alpha_t^2]} \rho - \frac{\lambda}{2}}{\mathbb{E}_t[\beta_t^2(1 + \rho) + \alpha_t^2]} \quad (82)$$

685 is a solution of (81) at convergence. Note that the identity $M = \rho \mathcal{Q}$, together with the definitions
 686 $M = w^\top \mu / \sqrt{d}$, $\mathcal{Q} = w^\top w / d$, implies that w lies entirely in $\text{span}(\mu)$. In other words, the weights w of
 687 the DAE recover perfectly the direction μ along which the data distribution ρ exhibits non-trivial
 688 structure.

689 Turning to the generative process, we aim at describing the generated density $\hat{\rho}$ in the one-dimensional
 690 subspace $\mathcal{E} = \text{span}(\mu)$, in which the original distribution ρ exhibits non-trivial structure. The SDE
 691 (12) describing the evolution of the projection of a sample in this subspace is for the considered
 692 setting linear, and can be solved in closed form as

$$Z_t = Z_0 e^{\int_0^t ds (\Delta_s^\infty + \Gamma_s \mathcal{Q})}, \quad Z_0 \sim \mathcal{N}(0, \mathcal{Q}). \quad (83)$$

694 In the orthogonal subspace $\text{span}(\mu)^\perp$, the law of a sample is still given by an isotropic Gaussian, as
 695 described by equation (13) in Result 2.2. Therefore, the law of a sample X_t remains Gaussian at all
 696 times, with the variance in $\text{span}(\mu)$ increasing over sampling time to adapt to – and approximate –
 697 the structure of the target distribution ρ in this subspace. This simple linear case sheds light on the
 698 workings of the DAE-parametrized diffusion models. Over training, the weights identify and learn
 699 the relevant structural features of the target distribution. Subsequently, the learned weights drive the
 700 generative process to reproduce the target structure in the identified subspace, while approximating
 701 the density in the orthogonal space by an isotropic Gaussian.

702 B Derivation of Result 2.2

703 In this section, we derive the tight characterization of Result 2.2 for the learnt generative transport
 704 process (7).

705 B.1 Generative SDE

706 We remind the generative SDE, leveraged to generate samples from $\hat{\rho}(t)$ starting from $X_0 \sim \mathcal{N}(0, \mathbb{I}_d)$:

$$\frac{dX_t}{dt} = \left(\dot{\beta}_t - \frac{\dot{\alpha}_t}{\alpha_t} \beta_t + \epsilon_t \frac{\beta_t}{\alpha_t^2} \right) f_{b_\tau, w_\tau, v_\tau}(X_t) + \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\epsilon_t}{\alpha_t^2} \right) X_t + \sqrt{2\epsilon_t} dW_t, \quad (84)$$

707 with W_t a Wiener process and ϵ_t the diffusion schedule. Introducing the shorthands

$$\Gamma_t = \dot{\beta}_t - \frac{\dot{\alpha}_t}{\alpha_t} \beta_t + \epsilon_t \frac{\beta_t}{\alpha_t^2} \quad (85)$$

$$\Delta_t^\tau = b_\tau \Gamma_t + \frac{\dot{\alpha}_t}{\alpha_t} - \frac{\epsilon_t}{\alpha_t^2}, \quad (86)$$

708 the generative SDE can be written more compactly as

$$\frac{dX_t}{dt} = \Delta_t^\tau X_t + \Gamma_t \frac{w_\tau}{\sqrt{d}} \sigma \left(\frac{w_\tau^\top X_t}{\sqrt{d}} + p_t v_\tau \right) + \sqrt{2\epsilon_t} dW_t. \quad (87)$$

709 Importantly, note that the non-linear term $\sigma(\cdot)$ acts on the projection of X_t in the space \mathcal{W}_τ spanned
 710 by the columns of the trained weights matrix w_τ . Furthermore, its image also resides in \mathcal{W}_τ . In
 711 contrast, the dynamics in the orthogonal space \mathcal{W}_τ^\perp is simply linear. This motivates one to examine
 712 in succession the variable $Z_t \equiv w_\tau^\top X_t / \sqrt{d}$ and the projection $Y_t \equiv \Pi_{\mathcal{W}_\tau^\perp}^\perp X_t$ of X_t in \mathcal{W}_τ^\perp .

713 B.2 Dynamics in \mathcal{W}_τ

714 Let us first ascertain the evolution of Z_t , which tracks the evolution of a sample X_t in the weight
 715 space \mathcal{W}_τ . It follows directly from (87) that Z_t obeys the r –dimensional SDE

$$\frac{d}{dt} Z_t = \Delta_t^\tau Z_t + \Gamma_t \mathcal{Q}_\tau \sigma(Z_t + p_t v_\tau) + \sqrt{2\epsilon_t} \mathcal{Q}^{1/2} dB_t, \quad (88)$$

716 with B_t a r –dimensional Wiener process, and \mathcal{Q}_τ the summary statistic sharply characterized in
 717 Result 2.1. This recovers equation (12) of Result 2.2.

718 B.3 Dynamics in \mathcal{W}_τ^\perp

719 In \mathcal{W}_τ^\perp , the transport induced by the SDE (87) is simply linear:

$$\frac{dY_t}{dt} = \Delta_t^\tau Y_t + \sqrt{2\epsilon_t} dH_t, \quad (89)$$

720 with H_t here a $(d - r)$ –dimensional Wiener process. This SDE admits a compact closed-form
 721 solution

$$Y_t = e^{\int_0^t ds \Delta_s^\tau} Y_0 + e^{\int_0^t ds \Delta_s^\tau} \int_0^t e^{-\int_0^s dh \Delta_h^\tau} \sqrt{2\epsilon_s} dW_s. \quad (90)$$

722 By Itô isometry, Y_t is Gaussian with law

$$Y_t \sim \mathcal{N} \left(0_{\mathcal{W}_\tau^\perp}, e^{2 \int_0^t ds \Delta_s^\tau} \left[1 + 2 \int_0^t e^{-2 \int_0^s dh \Delta_h^\tau} \epsilon_s ds \right] \Pi_{\mathcal{W}_\tau}^\perp \right), \quad (91)$$

723 which recovers equation (13). This completes the derivation of Result 2.2. \square

724 B.4 Discretized sampling

725 As a final remark, let us note that the derivation presented in the present Appendix can be carried
 726 out in completely unchanged fashion starting from any discretization of the generative SDE (1). Let
 727 $t_0 = 0, t_1, \dots, t_T \in (0, 1)$ and consider the Euler-Mayurama discretization of the stochastic process
 728 for $k \in \llbracket 0, T-1 \rrbracket$:

$$X_{k+1} - X_k = (t_{k+1} - t_k) \left[\left(\dot{\beta}_{t_k} - \frac{\dot{\alpha}_{t_k}}{\alpha_{t_k}} \beta_{t_k} + \epsilon_{t_k} \frac{\beta_{t_k}}{\alpha_{t_k}^2} \right) f_{b_\tau, w_\tau, v_\tau}(X_k) + \left(\frac{\dot{\alpha}_{t_k}}{\alpha_{t_k}} - \frac{\epsilon_{t_k}}{\alpha_{t_k}^2} \right) X_k \right. \\ \left. + \sqrt{2\epsilon_{t_k}(t_{k+1} - t_k)} \xi_k \right], \quad (92)$$

729 starting from $X_{t_0} \sim \mathcal{N}(0, \mathbb{I}_d)$. In (92), $\xi_k \sim \mathcal{N}(0, \mathbb{I}_d)$ independently for each step k . Then the
 730 following version of Result 2.2 holds:

731 **Result B.1. (Discrete dynamics)** Consider a discretization $t_1, \dots, t_T \in (0, 1)$ and the discretized
 732 sampling process $(X_k)_{k \in \llbracket 1, T \rrbracket}$ (92). Denote $Y_k = \Pi_{\mathcal{W}_\tau}^\perp X_k$ and $Z_k = w_\tau^\top X_k / \sqrt{d}$, for a process X_k
 733 satisfying the generative process (92) from an initialization $X_0 \sim \mathcal{N}(0, \mathbb{I}_d)$. Then Z_t follows the
 734 low-dimensional stochastic process

$$Z_{k+1} - Z_k = (t_{k+1} - t_k) [\Delta_{t_k}^\tau Z_t + \Gamma_{t_k} \mathcal{Q}_\tau \sigma(Z_k + p_{t_k} v_\tau)] + \sqrt{2\epsilon_{t_k}(t_{k+1} - t_k)} \mathcal{Q}_\tau^{1/2} \zeta_k, \quad (93)$$

735 from an initial condition $Z_0 \sim \mathcal{N}(0, \mathcal{Q}_\tau)$, with $\zeta_k \sim \mathcal{N}(0, \mathbb{I}_r)$ and $\mathbb{E}[\zeta_k \zeta_l^\top] = \delta_{kl} \mathbb{I}_r$. On the other
 736 hand, Y_k is independently Gaussian-distributed as

$$Y_k \sim \mathcal{N} \left(0_{\mathcal{W}_\tau^\perp}, \left[\prod_{j=0}^{k-1} (1 + (t_{j+1} - t_j) \Delta_{t_j}^\tau)^2 + \sum_{j=0}^{k-2} 2\epsilon_{t_j} (t_{j+1} - t_j) \prod_{l=j+1}^{k-1} (1 + (t_{j+1} - t_l) \Delta_{t_l}^\tau)^2 + 2\epsilon_{t_{k-1}} (t_k - t_{k-1}) \right] \Pi_{\mathcal{W}_\tau}^\perp \right). \quad (94)$$

737 C Derivation of Corollary 2.3

738 Result 2.2 already provides a tight asymptotic characterization of the law of a sample X_t in terms
 739 of its projection Z_t (12) in the weights space \mathcal{W}_τ (characterized by a r -dimensional ODE) and
 740 its Gaussian component Y_t (13) in the orthogonal space \mathcal{W}_τ^\perp . A weakness of this characterization,
 741 however, lies in that it relies on a *training-time dependent* space \mathcal{W}_τ , with respect to which the
 742 characterization is formulated. Intuitively, this space rotates and changes as the model is further
 743 trained, making the result rather unwieldy. To palliate this shortcoming, one would rather select
 744 a *fixed*, τ -independent, reference subspace \mathcal{E} of finite dimension $R = \Theta_d(1)$, and transfer the
 745 characterization of Result 2.2 to this fixed subspace. Formally, this means ascertaining the law of the
 746 projection of X_t in \mathcal{E} , from that of its projections in $\mathcal{W}_\tau, \mathcal{W}_\tau^\perp$. This constitutes the objective of the
 747 present Appendix.

748 Let us fix an orthonormal basis $\{e_j\}_{j=1}^R$ of \mathcal{E} , stacked vertically in the matrix $E \in \mathbb{R}^{d \times R}$. We remind
 749 that we aim at characterizing the law of $E^\top X_t$. To that end, for any $1 \leq j \leq R$, start from the
 750 decomposition

$$e_j^\top X_t = (\Pi_{\mathcal{W}_\tau} e_j)^\top (\Pi_{\mathcal{W}_\tau} X_t) + e_j^\top \Pi_{\mathcal{W}_\tau}^\perp Y_t, \quad (95)$$

751 where we decomposed X_t into its projections in $\mathcal{W}_\tau, \mathcal{W}_\tau^\perp$. Note that, from Result 2.2 the two terms
 752 of this decomposition are independent. In the following, we sequentially ascertain the distribution of
 753 each of the terms in the decomposition (95).

754 **C.1 Law of $(\Pi_{\mathcal{W}_\tau} e_j)^\top (\Pi_{\mathcal{W}_\tau} X_t)$**

755 To compute $(\Pi_{\mathcal{W}_\tau} e_j)^\top (\Pi_{\mathcal{W}_\tau} X_t)$, we first aim to decompose e_j, X_t in a basis of \mathcal{W}_τ . Let us consider
 756 the eigendecomposition of the summary statistic $\mathcal{Q}_\tau = w_\tau^\top w_\tau / d$ (characterized in Result 2.3) as

$$\mathcal{Q}_\tau = U_\tau S_\tau U_\tau^\top. \quad (96)$$

757 This means that $B_\tau = 1/\sqrt{d}(S_\tau^+)^{1/2} U_\tau^\top w_\tau^\top$ forms a set of r orthonormal vectors (or a set of orthonor-
 758 mal vectors plus zero vectors if \mathcal{Q}_τ is rank deficient), which we will use as a basis. We denoted S_τ^+
 759 the Moore-Penrose pseudo-inverse of S_τ . The components of the reference vectors $E \in \mathbb{R}^{d \times R}$ (with
 760 columns $\{e_j\}$) and X_t in this basis are then given by

$$B_\tau E = \frac{1}{\sqrt{d}}(S_\tau^+)^{1/2} U_\tau^\top w_\tau^\top E = (S_\tau^+)^{1/2} U_\tau^\top G_\tau^\top \quad (97)$$

$$B_\tau X_t = \frac{1}{\sqrt{d}}(S_\tau^+)^{1/2} U_\tau^\top w_\tau^\top X_t = (S_\tau^+)^{1/2} U_\tau^\top Z_t, \quad (98)$$

761 where Z_t is characterized in Result 2.2. Then, very simply, the decomposition of X_t in the reference
 762 basis E restricted to \mathcal{W}_τ reads

$$(\Pi_{\mathcal{W}_\tau} e_j)^\top (\Pi_{\mathcal{W}_\tau} X_t) = e_j^\top B_\tau^\top B_\tau X_t = G_\tau \mathcal{Q}_\tau^+ Z_t \quad (99)$$

763 **C.2 Law of $E^\top \Pi_{\mathcal{W}_\tau}^\perp Y_t$**

764 In distribution, $E^\top \Pi_{\mathcal{W}_\tau}^\perp Y_t$ inherits the Gaussianity of Y_t , as established in Result 2.2. It has mean
 765 zero and covariance

$$\begin{aligned} e^{2 \int_0^t ds \Delta_s^\tau} E^\top \Pi_{\mathcal{W}_\tau}^\perp E &= e^{2 \int_0^t ds \Delta_s^\tau} E^\top (\mathbb{I}_d - B_\tau^\top B_\tau) E \\ &= e^{2 \int_0^t ds \Delta_s^\tau} [\mathbb{I}_R - G_\tau \mathcal{Q}_\tau^+ G_\tau^\top]. \end{aligned} \quad (100)$$

766 **C.3 Law of $E^\top X_t$**

767 One is now in a position to ascertain the law of $E^\top X_t$. Putting the above results together, in
 768 distribution:

$$E^\top X_t \stackrel{d}{=} G_\tau \mathcal{Q}_\tau^+ Z_t + \mathcal{N} \left(0_R, e^{2 \int_0^t ds \Delta_s^\tau} [\mathbb{I}_r - G_\tau \mathcal{Q}_\tau^+ G_\tau^\top] \right), \quad (101)$$

769 which recovers Corollary 2.3. □

770 **D Additional experiments**

771 **D.1 Additional details on the numerical experiments**

772 In this Appendix, we provide further specifications on the numerical experiments illustrated in Fig. 7,
 773 3 and Fig. 4.

774 **Generative process—** In all the figures, the sampling was carried out by discretizing the interval
 775 $(0, 1)$ in N steps $t_k = 1/N$ for $k \in \llbracket 0, N \rrbracket$, and running the discretized SDE (92) in experiments, and
 776 the associated theoretical characterization of Results B.1 and 2.3 for the theoretical predictions, up to
 777 a stopping time $0 \leq t_f \leq 1$. In Fig. 7, $N = 100, t_f = 0.95$; in Fig. 8, $N = 100, t_f = 0.98$ and in
 778 Figs. 3, ?? and 4, $N = 50, t_f = 0.98$. Note that all choices for N, t_f are up to the experimentalist,
 779 and captured by the theoretical characterizations. Generically, one needs to opt for $t_f < 1$ due to the
 780 DAE-parametrized SDE (7) being ill-defined at $t = 1$, since $\alpha_1 = 0$. This is an artifact of the neural
 781 network parametrization; the ground-truth SDE (1) is on the hand well-defined even at $t = 1$.

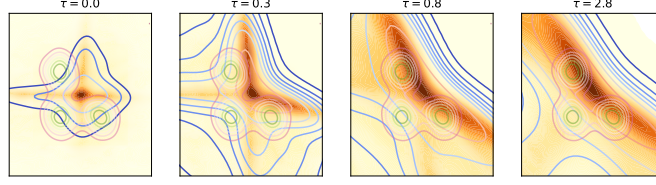


Figure 7: Evolution of the projected density $\Pi_{\mathcal{E}} \hat{\rho}_{\tau}$ generated by a DAE (3) with $r = 4$ hidden units and $\sigma = \text{ReLU}$ activation, trained on a trimodal Gaussian mixture, with $\eta = 0.2, \lambda = 1.5, \epsilon_t = 0.1, p_t = 0, \alpha_t = 1 - t, \beta_t = t, \mathcal{G} = \{0.7\}$, from a warm start. The generative SDE (7) was run up to $t = 0.98$, and the subspace \mathcal{E} is spanned by the centroids of the target density. Different panels correspond to different training times τ . Blue contours: contour levels of the theoretical prediction of Corollary 2.3 for the density $\Pi_{\mathcal{E}} \hat{\rho}_{\tau}$. Colormap: numerical experiments in large but finite dimension $d = 1000$. Green contours: contour levels of the target density ρ . Over training time, the four branches of the generated density rotate to align with the clusters of the target density, with two branches merging in the process.

Discretization of the manifold density π — In the generic case where $\pi(\cdot)$ (8) is not discrete, the ODE updates (10) still involve an integral over $d\pi(c)$, with c spanning \mathbb{R}^{κ} . For instance, in the setting of Fig. 4, at generation $g = 2$, $\kappa = r = 2$ and $\pi(c) = \Pi_{\mathcal{W}_2^{(1)}} \hat{\rho}^{(1)}(c)$. The latter is however still characterized in terms of a SDE (12), and not in closed-form. As a first step, we thus generated 4000 samples from π , using the theoretical characterization of Result 2.2, and approximated the density using the `scipy` [71] implementation of Gaussian kernel density estimation (KDE), in order to access a smooth estimation of π . The bandwidth was elected to be 1.5 times that determined using the Silverman method [63]. To perform the integral with measure $d\pi(c)$, we discretized π over a 10×10 grid, restricting the support to $[-1.5, 1.5] \times [-2.5, 2.5]$ where almost all of its mass was found to lie. The relative weights of the $10 \times 10 = 100$ discretized points were then evaluated from the KDE estimation, and overall normalization was finally enforced to ensure the relative weights sum to 1. Finally, this discretization was used in evaluating the theoretical characterization of Result 2.1, replacing the integrals over π by finite sums over the 100 points of the discretization. All results have been observed to be rather robust with respect to the choice of discretization, range, and bandwidth.

Preprocessing of the MNIST images— Finally, we detail the procedure used to evaluate the covariance of MNIST sevens used in Fig. 3. The total MNIST training set was used, retaining only sevens. The data was vectorized (flattened), centered, and normalized by 300. The empirical covariance was finally evaluated over the entire dataset, and used to generate the Gaussian target density considered in Fig. 3.

Evaluation of the Hellinger distance— To estimate the Hellinger distance between $\hat{\rho}$ and ρ (see Fig. 2 (right)), we first sample 5000 points from the trained diffusion model, project them in the considered subspace \mathcal{E} , and approximate the density using the `scipy` implementation of Gaussian KDE. In the case of Fig. 2 (right), we used $\mathcal{E} = \text{span}(\mu)$, and used a 1000-points grid discretization of the interval $[-10, 10]$ for the purpose of the KDE. The Hellinger distance between ρ and the (KDE of) $\hat{\rho}$ is then numerically estimated as

$$H(\Pi_{\mathcal{E}} \rho, \Pi_{\mathcal{E}} \hat{\rho}) = \int_{\mathcal{E}} dz \left[\sqrt{(\Pi_{\mathcal{E}} \rho)(z)} - \sqrt{(\Pi_{\mathcal{E}} \hat{\rho})(z)} \right]^2. \quad (102)$$

By the same token, the theoretical prediction of the Hellinger distance is obtained by sampling 5000 samples from the theoretical expression for $\hat{\rho}$, as described in Result 2.3, and repeating the same KDE procedure.

D.2 Additional example

For completeness, we conclude this Appendix by illustrating the theoretical results 2.1 2.3 on an additional example, namely a trimodal Gaussian mixture density with isotropic clusters.

We consider a generative model parametrized by a DAE (3) with $r = 4$ hidden units and ReLU activation. Fig. 7 illustrates, for different training times τ , the generated density $\hat{\rho}_{\tau}$ projected in the

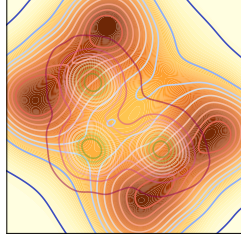


Figure 8: Density $\Pi_{\mathcal{E}}\hat{\rho}_{\tau}$ generated by a DAE (3) with $r = 2$ hidden units and $\sigma = \tanh$ activation, trained on a trimodal Gaussian mixture, with $\eta = 0.2, \lambda = 1.5, \epsilon_t = 0.1, p_t = 0, \alpha_t = 1 - t, \beta_t = t, \mathcal{G} = \{1/2\}, \tau = 2.8$. The generative SDE (7) was run up to $t = 0.98$, and the subspace \mathcal{E} is spanned by the centroids of the target density. Blue contours: contour levels of the theoretical prediction of Corollary 2.3 for the density $\Pi_{\mathcal{E}}\hat{\rho}_{\tau}$. Colormap: numerical experiments in large but finite dimension $d = 1000$. Green contours: contour levels of the target density ρ .

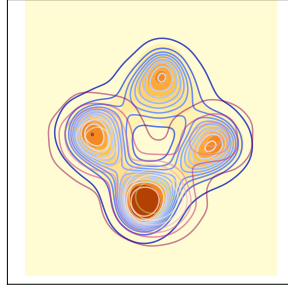


Figure 9: Evolution of the projected density $\Pi_{\mathcal{E}}\hat{\rho}_{\tau}$ generated by a DAE (3) with $r = 2$ hidden units and $\sigma = \tanh$ activation, trained on a trimodal Gaussian mixture, with $\eta = 0.5, \lambda = 0.1, \epsilon_t = 0.0, p_t = \cos(\pi t), \alpha_t = 1 - t, \beta_t = t, \mathcal{G} = \{0.2, 0.4, 0.6, 0.8\}$, from a random initialization. For two unit vectors e_1, e_2 , the target density is the mixture $\rho = 1/2\mathcal{N}(-3e_2, I_d) + 1/6\mathcal{N}(3e_1, I_d) + 1/3\mathcal{N}(-3e_1, I_d)$. The generative SDE (7) was run up to $t = 0.9$, and the subspace \mathcal{E} is spanned by e_1, e_2 . Different panels correspond to different training times τ . Blue contours: contour levels of the theoretical prediction of Corollary 2.3 for the density $\Pi_{\mathcal{E}}\hat{\rho}_{\tau}$. Colormap: numerical experiments in large but finite dimension $d = 1000$.

space \mathcal{E} spanned by the cluster centroids of the target density. A comparison between the theoretical predictions (blue contour levels) and numerical experiments in large but finite dimension $d = 1000$ (orange colormap) reveals a good agreement. Interestingly, the modes of the generated density $\hat{\rho}_{\tau}$ rotate over training time to align with the modes of the target density ρ , with two modes merging in the process. The resulting density $\hat{\rho}_{\tau}$ at large training time τ exhibits a similar geometry to the target density ρ , without however perfectly reproducing it – a sign of the architectural bias due to the limited expressivity of the model (3), which cannot perfectly generate the target distribution.

Perhaps unsurprisingly, this bias furthermore strongly depends on the architecture of the DAE. Fig. 8 represents the density generated by a DAE with $r = 2$ hidden units and \tanh activation, for the same target density ρ , with all parameters otherwise unchanged, revealing a very different geometry compared to the ReLU network. In particular, the model fails to generate a trimodal density, with four modes emerging instead. This instance of architectural bias can be easily rationalized. Observe indeed that from equation (12) of Result (2.2), for odd activations such as $\sigma = \tanh$, the transport process is equivariant with respect to the transformation $X \rightarrow -X$. In other words, the generated density $\hat{\rho}_{\tau}$ then necessarily exhibits a symmetry with respect to inversions around the origin – thus forbidding the existence of an odd number of modes. This provides a particularly simple yet telling

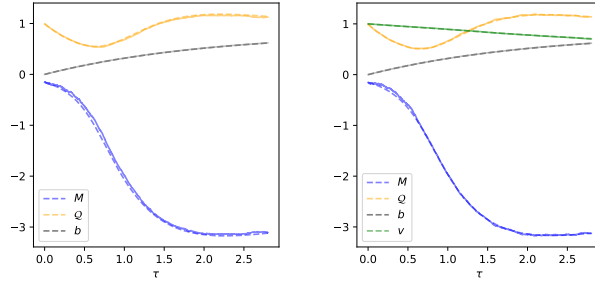


Figure 10: Evolution of the summary statistics M_τ , Q_τ and of the skip connection strength b_τ and time encoding weights v_τ as a function of the training time τ , for $\sigma = \tanh$, $r = 1$, $\alpha_t = 1 - t$, $\beta_t = t$, $\mathcal{G} = \{0.2, 0.4, 0.6, 0.8\}$. The target density is the same bimodal Gaussian mixture as Fig. 2. Solid lines: numerical experiments in dimension $d = 1000$. Dashed: theoretical characterization (10) of Result 2.1. **Right:** no time encoding $p_t = 0$. **Left:** with a time encoding $p_t = \cos(\pi t)$. The introduction of a time encoding leaves the training dynamics sensibly unchanged.

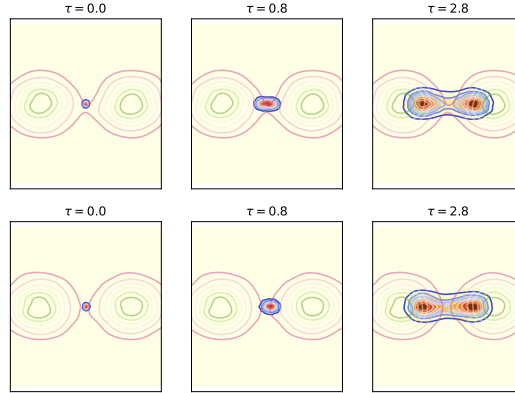


Figure 11: Evolution of the projected density $\Pi_{\mathcal{E}} \hat{\rho}_\tau$ generated by a DAE (3) with $r = 1$ hidden unit and $\sigma = \tanh$ activation, trained on a bimodal Gaussian mixture, with $\eta = 0.2$, $\lambda = 1.5$, $\epsilon_t = 0$, $\alpha_t = 1 - t$, $\beta_t = t$, $\mathcal{G} = \{0.2, 0.4, 0.6, 0.8\}$. The generative SDE (7) was run up to $t = 0.9$, and the subspace \mathcal{E} is a plane containing the centroid of the target density. Different panels correspond to different training times τ . Blue contours: contour levels of the theoretical prediction of Corollary 2.3 for the density $\Pi_{\mathcal{E}} \hat{\rho}_\tau$. Colormap: numerical experiments in large but finite dimension $d = 1000$. Green contours: contour levels of the target density ρ . **Top:** no time encoding $p_t = 0$. **Bottom:** with a time encoding $p_t = \cos(\pi t)$.

831 example of how the choice of architecture can strongly constrain the geometry of the generated
832 densities.

833 **Class imbalance** – The above example concerns a balanced Gaussian mixture, with all three
834 clusters sharing equal relative probability $1/3$. One may naturally wonder whether the model can
835 also adapt to class imbalance. Fig. 9 is set for the same target density as Fig. 8, with the difference
836 that clusters now have probabilities $1/2, 1/3, 1/6$. While the generated density still presents a spurious
837 mode (see discussion above), it correctly reproduces the clusters, and correctly gives higher mass to
838 the most probable clusters.

839 D.3 The effect of time encoding

840 We conclude this appendix by discussing the effect of including the time encoding p_t through its
841 associated set of weights v in the DAE model (3). For the binary target mixture described in the main
842 text (see Fig. 2), we plot in Fig. 2.1 the evolution of the summary statistics M , Q of Result 2.1 over
843 training time, alongside that of the skip connection strength b and encoding weights v . The two plots

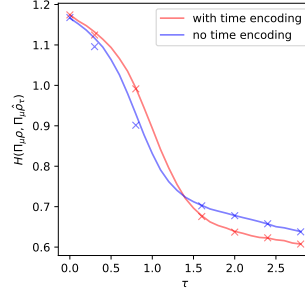


Figure 12: In the same setting as Fig. 11, Hellinger distance between the target and generated densities, projected in the space spanned by the centroid, as a function of the training time τ . **Red:** model with a time encoding $p_t = \cos(\pi t)$. **Blue:** without time encoding $p_t = 0$.

844 correspond to a model with no time encoding (i.e. $p_t = 0$) and a model endowed with a sinusoidal
845 time encoding $p_t = \cos(\pi t)$. As can be observed, the introduction of the time encoding has a very
846 small effect, and the curves are left sensibly unchanged. The generated densities, illustrated in Fig. 11,
847 are also strongly similar. A more quantitative viewpoint is displayed in Fig. 12, which shows how the
848 introduction of a time encoding yields a slightly lower, but overall very similar, Hellinger distance
849 between target and generated densities at large training times. These observations temptingly suggest
850 that, in all probed settings, for this simple model, the inclusion of a time encoding has an overall
851 small effect on the qualitative behaviour of the considered model.