A Supplementary Materials: Overview

These supplementary materials are organized as follows. In Sec. B we discuss general properties of pathwise gradient estimators derived from the transport equation. In Sec. C we give further details on Adaptive Velocity Fields. In Sec. D we give further details on our pathwise gradient estimators for mixture distributions, in particular describing velocity fields for four families of Normal mixtures. Finally, in Sec. E we describe the setup of the various experiments described in the main text.

B Pathwise Gradient Estimators and the Transport Equation

As discussed in the main text, a solution $\boldsymbol{v}^\theta$ to the transport equation allows us to form an unbiased pathwise gradient estimator via

$$\nabla_\theta \mathcal{L} = \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})} \left[\nabla_{\boldsymbol{z}} f \cdot \boldsymbol{v}^\theta \right] \tag{16}$$

In order for this to be a sensible Monte Carlo estimator, we require that the variance is finite, i.e

$$\mathbb{V}(\nabla_\theta \mathcal{L}) = \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})} \left[||\nabla_{\boldsymbol{z}} f \cdot \boldsymbol{v}^\theta||^2 \right] - ||\nabla_\theta \mathcal{L}||^2 < \infty \tag{17}$$

In order for the derivation given in the main text to hold, we also require for $\boldsymbol{v}^\theta$ to be everywhere continuously differentiable and that the surface integral $\oint_S (q_{\boldsymbol{\theta}} f \boldsymbol{v}^\theta) \cdot \hat{\mathbf{n}} \, \mathrm{d}S$ go to zero as $\mathrm{d}S$ tends towards the boundary at infinity. A natural way to ensure the latter condition for a large class of test functions is to require the boundary condition

$$q_{\boldsymbol{\theta}} \boldsymbol{v}^{\pi_j} \to 0 \qquad \text{as} \qquad ||\boldsymbol{z}|| \to \infty \;\; \text{(for all directions } \hat{\boldsymbol{z}}) \tag{18}$$

Note that this boundary condition is satisfied by all the gradient estimators proposed in this work. Much of the difficulty in using the transport equation to construct pathwise gradient estimators is in finding velocity fields that satisfy all these desiderata.

For example consider a mixture of products of univariate distributions of the form:

$$q_{\boldsymbol{\theta}}(\boldsymbol{z}) = \sum_{j=1}^{K} \pi_j q_{\boldsymbol{\theta}_j}(\boldsymbol{z}) \qquad \text{with} \qquad q_{\boldsymbol{\theta}_j}(\boldsymbol{z}) = \prod_{i=1}^{D} q_{\boldsymbol{\theta}_{ji}}(z_i) \tag{19}$$

Here j runs over the components and i runs over the dimensions of $\boldsymbol{z}$. Note that a mixture of diagonal Normal distributions is a special case of Eqn. 19. Suppose each $q_{\boldsymbol{\theta}_{ji}}$ has a CDF $F_{\boldsymbol{\theta}_{ji}}$ that we have analytic control over. Then we can form the velocity field

$$\boldsymbol{v}_i^{\pi_j} = -\frac{F_{\boldsymbol{\theta}_{ji}} q_{j,-i}}{D q_{\boldsymbol{\theta}}} \qquad \text{with} \qquad q_{j,-i} \equiv \prod_{k \neq i} q_{\boldsymbol{\theta}_{jk}} \tag{20}$$

This is a solution to the transport equation for the mixture weight π_j; however, it does not satisfy the boundary condition Eqn. 18 and so it is of limited practical use for estimating gradients.[14] Intuitively, the problem with Eqn. 20 is that $\boldsymbol{v}^{\pi_j}$ sends mass to infinity.

C Adaptive Velocity Fields for the Multivariate Normal Distribution

We show that the velocity field

$$\tilde{\boldsymbol{v}}_{\boldsymbol{A}}^{L_{ab}} = \boldsymbol{L} \boldsymbol{A}^{ab} \boldsymbol{L}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) \tag{21}$$

given in the main text is a solution to the corresponding null transport equation.[15] The transport equation can be written in the form

$$\frac{\partial}{\partial L_{ab}} \log q + \nabla \cdot \tilde{\boldsymbol{v}} + \tilde{\boldsymbol{v}} \cdot \nabla \log q = 0 \tag{22}$$

Transforming to whitened coordinates $\tilde{\boldsymbol{z}} = \boldsymbol{L}^{-1}(\boldsymbol{z} - \boldsymbol{\mu})$, the null equation is given by

$$\nabla \cdot \tilde{\boldsymbol{v}} = \tilde{\boldsymbol{v}} \cdot \tilde{\boldsymbol{z}} \tag{23}$$

We let $\tilde{\boldsymbol{v}} = \boldsymbol{A}^{ab} \tilde{\boldsymbol{z}}$ and compute

$$\nabla_{\tilde{\boldsymbol{z}}} \cdot \tilde{\boldsymbol{v}} = \mathrm{Tr}\, \boldsymbol{A}^{ab} = 0 = \sum_{ij} \tilde{z}_i \boldsymbol{A}_{ij}^{ab} \tilde{z}_j = \tilde{\boldsymbol{v}} \cdot \tilde{\boldsymbol{z}} \tag{24}$$

where we have used that $\boldsymbol{A}^{ab}$ is antisymmetric. Transforming $\tilde{\boldsymbol{v}}$ back to the given coordinates $\boldsymbol{z}$, we end up with Eqn. 21 (the factor of $\boldsymbol{L}$ enters when we transform the vector field).

For the 'reference solution' $\boldsymbol{v}_0^{L_{ab}}$ we simply use the velocity field furnished by the reparameterization trick, which is given by

$$(\boldsymbol{v}_0^{L_{ab}})_i = \delta_{ia} (\boldsymbol{L}^{-1} \boldsymbol{z})_b \tag{25}$$

Thus the complete specification of $\boldsymbol{v}_{\boldsymbol{A}}^{L_{ab}}$ is

$$(\boldsymbol{v}_{\boldsymbol{A}}^{L_{ab}})_i = \delta_{ia} (\boldsymbol{L}^{-1} \boldsymbol{z})_b + (\boldsymbol{L} \boldsymbol{A}^{ab} \boldsymbol{L}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}))_i \tag{26}$$

As mentioned in the main text, the computational complexity of using AVF gradients with this class of parameterized velocity fields (including the $\boldsymbol{A}$ update equations) is

$$\mathcal{O}(D^3 + MD^2) \tag{27}$$

[14] Note that since $\boldsymbol{\pi}$ is constrained to lie on the simplex, the relevant velocity fields to consider are defined w.r.t. an appropriate parameterization like softmax logits $\boldsymbol{\ell}$. It is for these velocity fields that the boundary condition needs to hold and not for $\boldsymbol{v}^{\pi_j}$ itself. This is why Eqn. 20 can be used for $D = 1$, where the boundary condition *does* hold.

[15] This derivation can also be found in reference [18].

This should be compared to the $\mathcal{O}(D^2)$ cost of the reparameterization trick gradient and the $\mathcal{O}(D^3)$ cost of the OMT gradient. However, the computational complexity in Eqn. 27 is somewhat misleading in that the $\mathcal{O}(D^3)$ term arises from matrix multiplications, which tend to be quite fast. By contrast the OMT gradient estimator involves a singular value decomposition, which tends to be substantially more expensive than a matrix multiplication on modern hardware. In cases where computing the test function involves expensive operations like Cholesky factorizations, the additional cost reflected in Eqn. 27 is limited (at least for $M \ll D$). For example, as reported in the GP experiment in Sec. 6.1.2 in the main text where $D = 468$, the AVF gradient estimator for $M = 1$ ($M = 5$) requires only $\sim$6% ($\sim$11%) more time per iteration.

C.1 Adaptive Velocity Fields for the Multivariate t-Distribution

We consider the multivariate t-distribution in D dimensions with probability density

$$\begin{aligned} q_{\boldsymbol{\theta}}(\boldsymbol{z}) &= \int_0^\infty q(\tau|\nu) q(\boldsymbol{z}|\boldsymbol{L}, \tau) d\tau \\ &\propto \frac{1}{|\boldsymbol{L}|} \left(1 + \tfrac{1}{\nu} \boldsymbol{z}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{z}\right)^{-\frac{\nu+D}{2}} \end{aligned} \tag{28}$$

where

$$q(\tau|\nu) = \mathrm{Ga}(\tau|\tfrac{\nu}{2}, \tfrac{\nu}{2}) \quad q(\boldsymbol{z}|\boldsymbol{L}, \tau) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \tau^{-\frac{1}{2}} \boldsymbol{L})$$

We want to compute derivatives w.r.t L_{ab}. We compute

$$\begin{aligned} &\frac{\partial \log q_{\boldsymbol{\theta}}(\boldsymbol{z})}{\partial L_{ab}} = \\ &\frac{\partial}{\partial L_{ab}} \left(-\log|\boldsymbol{L}| - \tfrac{\nu+D}{2} \log\left(1 + \tfrac{1}{\nu} \boldsymbol{z}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{z}\right)\right) = \\ &- L^{-1}_{ba} + \tfrac{\nu+D}{\nu} \left(1 + \tfrac{1}{\nu} \boldsymbol{z}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{z}\right)^{-1} (\boldsymbol{\Sigma}^{-1} \boldsymbol{z})_a (\boldsymbol{L}^{-1} \boldsymbol{z})_b \end{aligned} \tag{29}$$

Now suppose $\boldsymbol{v}_{\boldsymbol{A}}^{L_{ab}}$ is given as in Eqn. 26. Then we have

$$\nabla \cdot \boldsymbol{v}_{\boldsymbol{A}}^{L_{ab}} = \nabla \cdot \boldsymbol{v}_0^{L_{ab}} = L^{-1}_{ba} \tag{30}$$

and

$$\nabla q_{\boldsymbol{\theta}}(\boldsymbol{z}) = -\tfrac{\nu+D}{\nu} \left(1 + \tfrac{1}{\nu} \boldsymbol{z}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{z}\right)^{-1} (\boldsymbol{\Sigma}^{-1} \boldsymbol{z}) \tag{31}$$

It is easy to show that

$$\boldsymbol{v}_{\boldsymbol{A}}^{L_{ab}} \cdot \nabla q_{\boldsymbol{\theta}}(\boldsymbol{z}) = \boldsymbol{v}_0^{L_{ab}} \cdot \nabla q_{\boldsymbol{\theta}}(\boldsymbol{z}) \tag{32}$$

since the term containing $\boldsymbol{A}^{ab}$ vanishes due to the antisymmetry of $\boldsymbol{A}^{ab}$. Thus one has

$$\boldsymbol{v}_{\boldsymbol{A}}^{L_{ab}} \cdot \nabla q_{\boldsymbol{\theta}} = -\tfrac{\nu+D}{\nu} \left(1 + \tfrac{1}{\nu} \boldsymbol{z}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{z}\right)^{-1} (\boldsymbol{\Sigma}^{-1} \boldsymbol{z})_a (\boldsymbol{L}^{-1} \boldsymbol{z})_b \tag{33}$$

Gathering terms we see that $\boldsymbol{v}_{\boldsymbol{A}}^{L_{ab}}$ satisfies the relevant transport equation, namely

$$\frac{\partial}{\partial L_{ab}} \log q_{\boldsymbol{\theta}} + \nabla \cdot \boldsymbol{v}_{\boldsymbol{A}}^{L_{ab}} + \boldsymbol{v}_{\boldsymbol{A}}^{L_{ab}} \cdot \nabla \log q_{\boldsymbol{\theta}} = 0 \tag{34}$$

Consequently we have shown that the velocity fields given in Eqn. 26 can be used to a construct adaptive gradient estimators for the cholesky matrix of the multivariate t-distribution.

Note that the only property of $q_{\boldsymbol{\theta}}(\boldsymbol{z})$ that was used in the derivation was the fact that

$$q_{\boldsymbol{\theta}}(\boldsymbol{z}) \propto \frac{1}{|\boldsymbol{L}|} g(\boldsymbol{z}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{z}) \tag{35}$$

for some scalar density $g(\cdot)$. Thus the velocity fields in Eqn. 26 can in fact be used to construct adaptive gradient estimators for all distributions of the form given in Eqn. 35, i.e. for all elliptical distributions.

D Mixture distributions

In Table 3 we summarize the four families of mixture distributions for which we have found closed form solutions to the transport equation. The first one was presented in the main text; here we also present the solutions for the three other families of mixture distributions.

D.1 Pairwise Mass Transport

We begin with the transport equation for π_j, which reads

$$q_{\boldsymbol{\theta}_j} + \nabla_{\boldsymbol{z}} \cdot (q_{\boldsymbol{\theta}} \boldsymbol{v}^{\pi_j}) = 0 \tag{36}$$

Introducing softmax logits ℓ_j given by

$$\pi_j = \frac{e^{\ell_j}}{\sum_k e^{\ell_k}} \tag{37}$$

and using the fact that

$$\frac{\partial \pi_k}{\partial \ell_j} = \pi_j (\delta_{kj} - \pi_k) \tag{38}$$

we observe that the velocity field for ℓ_j satisfies the following transport equation

$$\pi_j \left(q_{\boldsymbol{\theta}_j} - \sum_k \pi_k q_{\boldsymbol{\theta}_k}\right) + \nabla_{\boldsymbol{z}} \cdot \left(q_{\boldsymbol{\theta}} \boldsymbol{v}^{\ell_j}\right) = 0 \tag{39}$$

We substitute Eqn. 14 for $\boldsymbol{v}^{\ell_j}$ and compute the divergence term, which yields

$$\begin{aligned} &\nabla_{\boldsymbol{z}} \cdot \left(q_{\boldsymbol{\theta}} \boldsymbol{v}^{\ell_j}\right) = \nabla_{\boldsymbol{z}} \cdot \left(q_{\boldsymbol{\theta}} \pi_j \sum_{k \neq j} \pi_k \tilde{\boldsymbol{v}}^{jk}\right) = \\ &- \pi_j \sum_{k \neq j} \pi_k \left(q_{\boldsymbol{\theta}_j} - q_{\boldsymbol{\theta}_k}\right) = -\pi_j \left(q_{\boldsymbol{\theta}_j} - \sum_k \pi_k q_{\boldsymbol{\theta}_k}\right) \end{aligned}$$

Distribution Name	Component Distributions	Velocity Field	Computational Complexity
DiagNormalsSharedCovariance	$\mathcal{N}(\boldsymbol{z}\|\boldsymbol{\mu}_j, \boldsymbol{\sigma})$	Eqn. 15	$\mathcal{O}(K^2D)$
ZeroMeanGSM	$\mathcal{N}(\boldsymbol{z}\|\mathbf{0}, \lambda_j\boldsymbol{\sigma})$	Eqn. 40	$\mathcal{O}(KD)$
GSM	$\mathcal{N}(\boldsymbol{z}\|\boldsymbol{\mu}_j, \lambda_j\boldsymbol{\sigma})$	Eqn. 54	$\mathcal{O}(K^2D)$
DiagNormals	$\mathcal{N}(\boldsymbol{z}\|\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j)$	Eqn. 51	$\mathcal{O}(KD)$

Table 3: Four families of mixture distributions for which we can compute pathwise derivatives. The names are those used in Fig. 1b in the main text.

Thus $\boldsymbol{v}^{\ell_j}$ satisfies the relevant transport equation Eqn. 39.

D.2 Zero Mean Discrete Gaussian Scale Mixture

Here each component distribution is specified by $q_{\boldsymbol{\theta}_j}(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\mathbf{0}, \lambda_j\boldsymbol{\sigma})$, where each λ_j is a positive scalar. Defining $\tilde{\boldsymbol{z}} = \boldsymbol{z} \odot \boldsymbol{\sigma}^{-1}$ and making use of radial coordinates with $r = ||\tilde{\boldsymbol{z}}||$ we find that a solution of the form in Eqn. 14 reduces to

$$\boldsymbol{v}^{\ell_j} = \pi_j \text{diag}(\boldsymbol{\sigma}) \left(\tilde{\boldsymbol{v}}^j - \sum_k \pi_k \tilde{\boldsymbol{v}}^k \right) \tag{40}$$

where

$$\begin{aligned} \tilde{\boldsymbol{v}}^j &= \frac{\tilde{\Phi}(\frac{r}{\lambda_j})}{q_{\boldsymbol{\theta}} \lambda_j^{D-1} \prod_{i=1}^D \sigma_i} \hat{\boldsymbol{r}} \qquad \text{and} \\ \tilde{\Phi}(z) &= \frac{z^{1-D}}{(2\pi)^{D/2}} \int_z^\infty \tilde{z}^{D-1} e^{-\tilde{z}^2/2} d\tilde{z} \end{aligned} \tag{41}$$

The 'radial CDF' $\tilde{\Phi}$ in Eqn. 41 can be computed analytically. In even dimensions we find[16]

$$\tilde{\Phi}(z) = \frac{e^{-z^2/2}}{(2\pi)^{D/2}} \sum_{k=0}^{\frac{D}{2}-1} \frac{(D-2)!!}{(2k)!!} z^{2k+1-D} \tag{42}$$

and in odd dimensions we find

$$\begin{aligned} \tilde{\Phi}(z) = &\frac{e^{-z^2/2}}{(2\pi)^{D/2}} \sum_{k=1}^{\frac{D-1}{2}} \frac{(D-2)!!}{(2k-1)!!} z^{2k-D} + \\ &(D-2)!! \sqrt{\tfrac{\pi}{2}} \frac{1 - \text{erf}(\frac{z}{\sqrt{2}})}{(2\pi)^{D/2} z^{D-1}} \end{aligned} \tag{43}$$

where $\text{erf}(\cdot)$ is the error function. Note that in contrast to all the other solutions in Table 3, Eqn. 40 for D even does not involve any error functions.

We show explicitly that Eqn. 40 is a solution of the corresponding transport equation. The derivations for the other families of mixture distributions are similar. It is enough to show that

$$\sum_i \frac{\partial}{\partial z_i} \left(q_{\boldsymbol{\theta}} \tilde{\boldsymbol{v}}_i^j \right) = \sum_i \frac{\partial}{\partial z_i} \left(\frac{\tilde{\Phi}(\frac{r}{\lambda_j})}{\lambda_j^{D-1} \prod_{i=1}^D \sigma_i} \hat{\boldsymbol{r}}_i \right) = -q_{\boldsymbol{\theta}_j}(\boldsymbol{z}) \tag{44}$$

Using the identities

$$\frac{\partial r}{\partial z_i} = \frac{z_i}{r\sigma_i^2} \qquad \frac{\partial}{\partial z_i} = \frac{\partial r}{\partial z_i} \frac{\partial}{\partial r} \qquad \hat{\boldsymbol{r}}_i = \frac{z_i}{r} \tag{45}$$

which follow from the definition $r^2 = \sum_i \frac{z_i^2}{\sigma_i^2}$, we have

$$\begin{aligned} &\sum_i \frac{\partial}{\partial z_i} \left(\tilde{\Phi}(\tfrac{r}{\lambda_j}) \hat{\boldsymbol{r}}_i \right) = \frac{1}{\lambda_j} \tilde{\Phi}'(\tfrac{r}{\lambda_j}) \sum_i \frac{z_i^2}{\sigma_i^2 r^2} = \\ &\frac{1}{\lambda_j} \tilde{\Phi}'(\tfrac{r}{\lambda_j}) \frac{r^2}{r^2} = \frac{1}{\lambda_j} \tilde{\Phi}'(\tfrac{r}{\lambda_j}) \end{aligned} \tag{46}$$

By construction we have that

$$\Phi'(\tfrac{r}{\lambda_j}) = -\frac{1}{(2\pi)^{D/2}} e^{-\frac{1}{2} r^2/\lambda_j^2} = -\left(\lambda_j^D \prod_{i=1}^D \sigma_i \right) q_{\boldsymbol{\theta}_j}(\boldsymbol{z}) \tag{47}$$

Comparing terms, we see that $\tilde{\boldsymbol{v}}^j$ is indeed a solution of the transport equation for π_j as desired.

D.3 Mixture of Diagonal Normal Distributions

Here each component distribution is given by

$$q_{\boldsymbol{\theta}_j}(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j) \qquad \text{for } \; j = 1, 2, ..., K \tag{48}$$

For $i = 1, ..., D$ define

$$\tilde{z}_{ji} = \frac{z_i - \mu_i}{\sigma_{ji}} \qquad \bar{z}_{ji} = \frac{z_i - \mu_i}{\sigma_i^0} \qquad r_{ji}^2 = \sum_{k<i} \tilde{z}_{jk}^2 + \sum_{k>i} \bar{z}_{jk}^2 \tag{49}$$

where $\boldsymbol{\sigma}^0$ is an arbitrary reference scale. We find that if we define[17]

$$\breve{\boldsymbol{v}}^j = \sum_i \frac{\left(\Phi(\tilde{z}_{ji}) - \Phi(\bar{z}_{ji}) \right) \phi(r_{ji}^2)}{q_{\boldsymbol{\theta}} \prod_{k<i} \sigma_{jk} \prod_{k>i} \sigma_k^0} \hat{\boldsymbol{z}}_i \tag{50}$$

[16] The notation $k!!$ refers to the double factorial of k, which occurs in this context through the identity $(2n-1)!! = 2^n \Gamma(n + \frac{1}{2})/\sqrt{\pi}$.

[17] Here we take the empty products $\prod_{k<1}$ and $\prod_{k>D}$ to be equal to unity.

then we can construct a solution of the form specified in Eqn. 14 that is given by

$$\boldsymbol{v}^{\ell_j} = \pi_j \left(\breve{\boldsymbol{v}}^j - \sum_k \pi_k \breve{\boldsymbol{v}}^k \right) \tag{51}$$

Since the reference scale $\boldsymbol{\sigma}^0$ is *arbitrary*, this is actually a parameterized family of solutions. Thus this solution is in principle amenable to the Adaptive Velocity Field approach described in Sec. 3 in the main text. In addition, the ordering of the dimensions $i = 1, ..., D$ that occurs implicitly in the telescopic structure of Eqn. 50 is also arbitrary. Thus Eqn. 50 corresponds to a very large family of solutions. In practice we use the fixed ordering given in Eqn. 50 and choose

$$\sigma_i^0 \equiv \min_{j \in [1,K]} \sigma_{ji} \tag{52}$$

We find this works pretty well empirically.

D.4 Discrete GSM

Here each component distribution is given by

$$q_{\boldsymbol{\theta}_j}(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z} | \boldsymbol{\mu}_j, \lambda_j \boldsymbol{\sigma}) \qquad \text{for } j = 1, 2, ..., K \tag{53}$$

where each λ_j is a positive scalar. We can solve the corresponding transport equation for the mixture weights by superimposing the solutions in Eqn. 15 and Eqn. 40. In more detail, in this case the solution to Eqn. 13 is given by

$$\bar{\boldsymbol{v}}^{jk} = \underbrace{\tilde{\boldsymbol{v}}^{jk;\lambda=\lambda_0}}_{\text{soln. from Eqn. 15}} + \tilde{\boldsymbol{w}}^j - \tilde{\boldsymbol{w}}^k \tag{54}$$

where

$$\tilde{\boldsymbol{w}}^j = \underbrace{\tilde{\boldsymbol{v}}^{j;\lambda_j} - \tilde{\boldsymbol{v}}^{j;\lambda=\lambda_0}}_{\text{solutions from Eqn. 40}} \tag{55}$$

Analogously to the reference scale $\boldsymbol{\sigma}^0$ in Sec. D.3, λ_0 is arbitrary. As such Eqn. 54 is a parametric family of solutions that is amenable to the Adaptive Velocity Field approach. Intuitively, we use the solutions from Eqn. 15 to effect mass transport between component means and solutions from Eqn. 40 to shrink/dilate covariances.

D.5 Mixture of Multivariate Normals with Shared Diagonal Covariance

To finish specifying the solution Eqn. 15 given in the main text, we define the following coordinates:

$$\tilde{\boldsymbol{z}} = \boldsymbol{z} \odot \boldsymbol{\sigma}^{-1} \qquad \tilde{\boldsymbol{\mu}}^j = \boldsymbol{\mu}^j \odot \boldsymbol{\sigma}^{-1} \qquad \hat{\boldsymbol{\mu}}^{jk} = \frac{\tilde{\boldsymbol{\mu}}^j - \tilde{\boldsymbol{\mu}}^k}{||\tilde{\boldsymbol{\mu}}^j - \tilde{\boldsymbol{\mu}}^k||}$$
$$\tilde{\mu}_{\parallel}^{jk} = \tilde{\boldsymbol{\mu}}^j \cdot \hat{\boldsymbol{\mu}}^{jk} \qquad \tilde{z}_{\parallel}^{jk} = \tilde{\boldsymbol{z}} \cdot \hat{\boldsymbol{\mu}}^{jk} \qquad \tilde{\boldsymbol{z}}_{\perp}^{jk} = \tilde{\boldsymbol{z}} - \tilde{z}_{\parallel}^{jk} \hat{\boldsymbol{\mu}}^{jk}$$

D.6 Velocity Fields for the Component Parameters of Multivariate Mixtures

Suppose $\boldsymbol{v}_{\text{single}}^{\boldsymbol{\theta}_j}$ is a solution of the single-component transport equation for the parameter $\boldsymbol{\theta}_j$, i.e.

$$\frac{\partial q_{\boldsymbol{\theta}_j}}{\partial \boldsymbol{\theta}_j} + \nabla \cdot (q_{\boldsymbol{\theta}_j} \boldsymbol{v}_{\text{single}}^{\boldsymbol{\theta}_j}) = 0 \tag{56}$$

Then

$$\boldsymbol{v}^{\boldsymbol{\theta}_j} = \frac{\pi_j q_{\boldsymbol{\theta}_j}}{q_{\boldsymbol{\theta}}} \boldsymbol{v}_{\text{single}}^{\boldsymbol{\theta}_j} \tag{57}$$

is a solution of the multi-component transport equation, since

$$\frac{\partial q_{\boldsymbol{\theta}}(\boldsymbol{z})}{\partial \boldsymbol{\theta}_j} + \nabla \cdot (q_{\boldsymbol{\theta}}(\boldsymbol{z}) \boldsymbol{v}^{\boldsymbol{\theta}_j}) = \pi_j \frac{\partial q_{\boldsymbol{\theta}_j}(\boldsymbol{z})}{\partial \boldsymbol{\theta}_j} + \nabla \cdot (q_{\boldsymbol{\theta}}(\boldsymbol{z}) \boldsymbol{v}^{\boldsymbol{\theta}_j})$$
$$= \pi_j \left(\frac{\partial q_{\boldsymbol{\theta}_j}(\boldsymbol{z})}{\partial \boldsymbol{\theta}_j} + \nabla \cdot (q_{\boldsymbol{\theta}_j}(\boldsymbol{z}) \boldsymbol{v}_{\text{single}}^{\boldsymbol{\theta}_j}) \right) = 0 \tag{58}$$

This completes the derivation for the claim about $\boldsymbol{v}^{\boldsymbol{\theta}_j}$ made at the beginning of Sec. 4 in the main text.

D.7 Pairwise Mass Transport and Control Variates

For $j = 1, ..., K$ define $K \times K$ square matrices A_{ik}^j such that all the rows and columns of each A_{ik}^j sum to zero. Then

$$\boldsymbol{w}^{\ell_j} = \sum_{i,k} A_{ik}^j \tilde{\boldsymbol{v}}^{ik} \tag{59}$$

is a null solution to the transport equation for $\boldsymbol{v}^{\ell_j}$, Eqn. 39. While we have not done so ourselves, these null solutions could be used to adaptively move mass among the K component distributions of a mixture instead of using the recipe in Eqn. 14, which takes mass from each component distribution j in proportion to its mass π_j (which is in general suboptimal).

E Experimental Details

E.1 Synthetic Test Function Experiments

We describe the setup for the experiments in Sec. 6.1.1 and Sec. 6.2.1 in the main text.

For the experiment in Sec. 6.1.1 the dimension is fixed to $D = 50$ and the mean of $q_{\boldsymbol{\theta}}$ is fixed to the zero vector. The Cholesky factor $\boldsymbol{L}$ that enters into $q_{\boldsymbol{\theta}}$ is constructed as follows. The diagonal of $\boldsymbol{L}$ consists of all ones. To construct the off-diagonal terms we proceed as follows. We populate the entries below the diagonal of a matrix $\Delta \boldsymbol{L}$ by drawing each entry from the uniform distribution on the unit interval. Then we define $\boldsymbol{L} = \mathbb{1}_D + r \Delta \boldsymbol{L}$. Here r controls the magnitude of

off-diagonal terms of $\boldsymbol{L}$ and appears on the horizontal axis of Fig. 1a in the main text. The three test functions are constructed as follows. First we construct a strictly lower diagonal matrix $\boldsymbol{Q}'$ by drawing each entry from a bernoulli distribution with probability 0.5. We then define $\boldsymbol{Q} = \boldsymbol{Q}' + \boldsymbol{Q}'^T$. The cosine test function is then given by

$$f(\boldsymbol{z}) = \cos\left(\sum_{i,j} Q_{ij} z_i / D\right) \tag{60}$$

The quadratic test function is given by

$$f(\boldsymbol{z}) = \boldsymbol{z}^T \boldsymbol{Q} \boldsymbol{z} \tag{61}$$

The quartic test function is given by

$$f(\boldsymbol{z}) = \left(\boldsymbol{z}^T \boldsymbol{Q} \boldsymbol{z}\right)^2 \tag{62}$$

The AVF gradient uses $M = 1$ and we train $\boldsymbol{A}^{ab}$ to (approximate) convergence before estimating the gradient variance.

For the experiment in Sec. 6.2.1 that is depicted in Fig. 1b the test function is fixed to $f(\boldsymbol{z}) = ||\boldsymbol{z}||^2$. The distributions $q_{\boldsymbol{\theta}}$ are constructed as follows. For the distributions that admit a parameter $\boldsymbol{\mu}_j$, each $\boldsymbol{\mu}_j$ is sampled from the sphere centered at $\boldsymbol{z} = \boldsymbol{0}$ with radius 2. For the distribution whose velocity field is given in Eqn. 40, the mean is fixed to $\boldsymbol{0}$. The covariance matrices are sampled from a narrow distribution centered at the identity matrix. Consequently the different mixture components have little overlap.

For the experiment in Sec. 6.2.1 that is depicted in Fig. 1c the test function is also fixed to $f(\boldsymbol{z}) = ||\boldsymbol{z}||^2$. The distributions $q_{\boldsymbol{\theta}}$ are constructed as follows. The K means are placed uniformly around the unit circle. The covariance of each component distribution is given by $\sigma^2 \mathbb{1}$, where σ is the parameter that is varied along the horizontal axis of the figure. For the gradient estimator derived from the transport equation, we use the estimator described in Sec. D.3, although in this particular case the estimator given by Eqn. 15 yields identical results.

In all cases the gradients can be computed analytically, which makes it easier to reliably estimate the variance of the gradient estimators.

E.2 Gaussian Process Regression

We use the Adam optimizer [21] to optimize the ELBO with single-sample gradient estimates. We chose the Adam hyperparameters by doing a grid search over the learning rate and β_1. For the AVF gradient estimator the learning rate and β_1 are allowed to differ between $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ gradient steps (the latter needs a much larger learning rate for good results). For each combination of hyperparameters we did 500 training iterations for five trials with different random seeds and then chose the combination that yielded the highest mean ELBO after 500 iterations. We then trained the model for 500 iterations, initializing with another random number seed. The figure in the main text shows the training curves for that single run. We confirmed that other random number seeds give similar results.

E.3 Baseball Experiment

There are 18 baseball players and the data consists of 45 hits/misses for each player. The model has two global latent random variables, ϕ and κ, with priors Uniform$(0, 1)$ and Pareto$(1, 1.5) \propto \kappa^{-5/2}$, respectively. There are 18 local latent random variables, θ_i for $i = 0, ..., 17$, with $p(\theta_i) = \text{Beta}(\theta_i | \alpha = \phi\kappa, \beta = (1-\phi)\kappa)$. The data likelihood factorizes into 45 Bernoulli observations with mean chance of success θ_i for each player i. All variational approximations are formed in the unconstrained space $\{\text{logit}(\phi), \log(\kappa - 1), \text{logit}(\theta_i)\}$. The mean field variational approximation consists of a diagonal Normal distribution in the unconstrained space, while the mixture variational approximation consists of K diagonal Normal distributions in the same space. We use the Adam optimizer for training with a learning rate of 5×10^{-3} [21].

E.4 Continuous State Space Model

We consider the following simple model with two dimensional observations $\boldsymbol{x}_t$ and two dimensional latent random variables $\boldsymbol{z}_t$:

$$p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}) = p(\boldsymbol{z}_1) p(\boldsymbol{x}_1 | \boldsymbol{z}_1) \prod_{t=2}^{T} p(\boldsymbol{z}_t | \boldsymbol{z}_{t-1}) p(\boldsymbol{x}_t | \boldsymbol{z}_t) \tag{63}$$

where

$$\begin{aligned} p(\boldsymbol{z}_1) &= \mathcal{N}(\boldsymbol{z}_1 | \boldsymbol{0}, \sigma_z \mathbb{1}_2) \\ p(\boldsymbol{z}_t | \boldsymbol{z}_{t-1}) &= \mathcal{N}(\boldsymbol{z}_t | \boldsymbol{T} \boldsymbol{z}_{t-1}, \sigma_z \mathbb{1}_2) \\ p(\boldsymbol{x}_t | \boldsymbol{z}_t) &= \mathcal{N}(\boldsymbol{x}_t | \boldsymbol{\mu}(\boldsymbol{z}_t), \sigma_x \mathbb{1}_2) \end{aligned} \tag{64}$$

and

$$\begin{aligned} &\boldsymbol{\mu}(\boldsymbol{z}_t) = (z_{1t}^2, 2 z_{2t}) \qquad \sigma_z = \frac{1}{2} \qquad \sigma_x = \frac{1}{4} \\ &\boldsymbol{T} = \frac{1}{2} \begin{pmatrix} -\sin(\theta) & \cos(\theta) \\ \cos(\theta) & \sin(\theta) \end{pmatrix} \qquad \theta = \frac{\pi}{4} \end{aligned} \tag{65}$$

The quadratic term z_{1t}^2 in $\boldsymbol{\mu}(\boldsymbol{z}_t)$ results in a highly multi-modal posterior. We generate 1000 sequences with $T = 10$ and use 800 for training and 200 for testing. The model dynamics are assumed to be known, and the variational family is constructed along the lines of the DKS inference network in [23]. We use the pathwise gradient estimators introduced in Sec. 4 in the main text.

E.5 Deep Markov Model

The training data consist of 229 sequences with a mean length of 60 time steps from the JSB Chorales polyphonic music dataset considered in [3]. Each time slice in a sequence spans a quarter note and is represented by an 88-dimensional binary vector. We use a Bernoulli likelihood. The dimension of the latent $\boldsymbol{z}_t$ at each time step is 32. The inference network a.k.a. variational family follows the DKS variant described in [23]. Similarly, the architecture of the various neural network components follows the architectures used in [23]. In particular the RNN dimension is fixed to be 600 and the dimension of the hidden layer in the neural network that parameterizes $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$ is 200; all other neural network hidden layers are 100-dimensional. We use a mini-batch size of 20. Following [23] we anneal the contribution of KL divergence-like terms over the course of optimization (we use a linear schedule). We use the Adam optimizer [21] with gradient clipping and an exponentially decaying learning rate and do up to 7000 epochs of learning. We do a grid search over optimization hyperparameters, which include the learning rate, β_1, the KL annealing schedule, and temperature (the latter only in the case of the Gumbel Softmax estimators). We use the validation set to fix the hyperparameters and then report results on the test set. The reported test ELBOs use a 200-sample estimator and are normalized per timestep.

E.5.1 Gradient Estimators

For $K = 2$ the mixture distributions have arbitrary diagonal covariance matrices; consequently the pathwise gradient estimator is of the form described in Sec. D.3. For $K = 3$ the mixture distributions share diagonal covariance matrices at each time step (we make this choice to limit the total number of parameters); consequently the pathwise gradient estimator is of the form described in Sec. 4.2 in the main text.

The two variants of the Gumbel Softmax estimator we use are more similar to the approach adopted in [17] than to the approach adopted in [24]. In particular we do *not* relax the objective function in the manner of [24] so that the resulting gradient estimators are biased. In GS-Soft we draw a K dimensional vector $\boldsymbol{y}$ from the Gumbel Softmax distribution so that $\boldsymbol{y}$ lies in the interior of the $K-1$ dimensional simplex. To generate a sample from the mixture $q(\boldsymbol{z}_t|\cdot)$ we draw a D-dimensional sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ and form the sample $\boldsymbol{z}_t$ via

$$z_{t,i} = \sum_{k=1}^{K} y_k \left(\mu_{ki} + \epsilon_i \sigma_{ki}\right) \qquad \text{for} \qquad i = 1, ..., D \tag{66}$$

In GS-Hard we adopt the same approach as in GS-Soft, except $\boldsymbol{y}$ is discretized via arg max, c.f. the straight-through estimator in [17]. We adopt the approach in Eqn. 66 so that we do not need to introduce variational distributions of the form $q(\boldsymbol{\pi}_t|\cdot)$, as we expect that this additional variational relaxation would lead to looser ELBO bounds (and would make a direct comparison to variational setups without ELBO terms of the form $\log q(\boldsymbol{\pi}_t|\cdot)$ more difficult).

We do not report numbers for the score function estimator, since the extremely high variance—$\mathcal{O}(10^5)$ times higher than for the pathwise gradient estimator—prevented us from obtaining competitive results. In particular we were unable to obtain test ELBOs above -9.0 nats; by contrast a mean field variational family with diagonal Normal distributions of the form $q(\boldsymbol{z}_t|\boldsymbol{x}_t)$ at each time step can achieve ~ -8.0 nats.

E.6 VAE

We train using MNIST 50k and fix the latent dimensionality to 50. The prior is Normal and the likelihood is Bernoulli. We fix the number of hidden units in the inference network to 400 and the number of hidden units in the decoder network to 200. We use the Adam optimizer and do a grid search over the following optimization hyperparameters: learning rate, β_1, and the temperature (for the Gumbel Softmax estimators). We train all models for 3000 epochs with a batch size of 256. Test ELBOs are computed with a 50-sample estimator. For details on the Gumbel Softmax estimators, see Sec. E.5.1.