Nested Variational Inference

Heiko Zimmermann
zimmermann.h@northeastern.edu

Hao Wu
wu.hao10@northeastern.edu

Babak Esmaeili
esmaeili.b@northeastern.edu

Sam Stites
stites.s@northeastern.edu

Jan-Willem van de Meent
j.vandemeent@northeastern.edu

Northeastern University

Abstract

We develop nested variational inference (NVI), a family of methods that learn proposals for nested importance samplers by minimizing an inclusive or exclusive KL divergence at each level of nesting. NVI is applicable to many commonly-used importance sampling strategies and additionally provides a mechanism for learning intermediate densities, which can serve as heuristics to guide the sampler. Our experiments apply NVI to learn samplers targeting (a) an unnormalized density using annealing and (b) the posterior of a hidden Markov model. We observe improved sample quality in terms of log average weight and effective sample size.

1. Introduction

Deep generative models provide a mechanism for incorporating priors into methods for unsupervised representation learning. This is particularly useful in settings where the prior defines a meaningful inductive bias that reflects a known structure of the underlying domain.

Training models with structured priors, however, poses some challenges. A standard strategy for training these models is to maximize a reparameterized variational lower bound with respect to a generative model and an inference model that approximates its posterior (Kingma and Welling, 2013; Rezende et al., 2014). This approach works well in variational autoencoders (VAEs) with isotropic Gaussian priors, but often fails in models with high-dimensional and correlated latent variables.

In recent years, a range of strategies for improving upon standard reparameterized variational inference have been put forward. These include wake-sleep style variational methods that minimize the inclusive KL-divergence (Bornschein and Bengio, 2014; Le et al., 2019), as well as sampling schemes that incorporate annealing (Huang et al., 2018), Sequential Monte Carlo (Le et al., 2017; Naesseth et al., 2017; Maddison et al., 2017), Gibbs sampling (Wu et al., 2019; Wang et al., 2018), and MCMC updates (Salimans et al., 2015; Hoffman, 2017; Li et al., 2017). While these methods offer flexible inference, typically resulting in better approximations to the posterior compared to traditional variational inference methods, they are also model-specific, requiring specialized sampling schemes and gradient estimators, and can not be easily composed with other techniques.

In this paper, we propose nested variational inference (NVI), a framework for combining nested importance sampling and variational inference. Nested importance sampling formalizes the construction of proposals by way of nested calls to other importance samplers.
Nested Variational Inference

(Naeseth et al., 2015, 2019), and admits many existing importance sampling strategies as special cases, including methods based on annealing (Neal, 2001) and sequential Monte Carlo (Del Moral et al., 2006). NVI learns proposals by optimizing an inclusive or exclusive KL divergence at each level of nesting. Combining nested variational objectives with importance resampling allows us to compute gradient estimates based on incremental weights, which depend only on variables that are sampled locally, rather than on all variables in the model. Doing so yields lower variance weights, whilst maintaining a high sample diversity relative to existing variational methods based on sequential Monte Carlo.

NVI extends beyond existing methods in that it defines objectives for learning intermediate densities in addition to proposals. In a nested importance sampler, the target density at each level of nesting defines the proposal at the next level of nesting. These learned intermediate densities can serve as heuristics that guide the sampler.

To demonstrate the potential of this approach. We compare NVI to existing techniques in two settings. The first is an annealed sampling task, where we use NVI to learn proposals and optimize the annealing schedule which specifies the intermediate densities. The second task is performing posterior inference on a Hidden Markov model, in which NVI is used to learn proposals and a heuristic factor that incorporates future observations. In both cases, NVI results in substantial improvements to sample quality.

2. Nested Variational Inference

Problem setting. Nested variational inference makes use of nested importance samplers (Naesseth et al., 2015, 2019), which provide a means of reasoning about methods that recursively use importance samplers to generate proposals. Here, we will consider a sequence of intractable densities \( \{\pi_k(z_k; \theta_k)\}_{k=1}^K \) and corresponding unnormalized densities \( \{\gamma_k(z_k; \theta_k)\}_{k=1}^K \) with intractible normalizing constant such that

\[
\pi_k(z_k; \theta_k) = \gamma_k(z_k; \theta_k)/Z_k(\theta_k), \quad Z_k(\theta_k) = \int dz_k \gamma_k(z_k; \theta_k).
\]

We are typically interested in the case where the final density \( \gamma_K(z; \theta) \) corresponds to the posterior distribution \( p_\theta(z|x) \). To simplify notation, we will in the following omit parameters when they are not needed for context (i.e. we write \( \gamma(z) \) instead of \( \gamma(z; \theta) \)).

We define the sequence of intermediate densities to interpolate between a density \( \pi_1(z_1) \), for which sampling is easy, to the final density \( \pi_K(z_K) \) for which sampling is difficult. Two standard strategies are to define variables \( z_k \in Z_k \) on (1) a fixed sample space \( Z_1 = \cdots = Z_K \) with increasingly tightly-peaked densities, e.g. when performing annealing, or (2) on sample spaces with increasing dimensionality \( Z_1 = Z_1', Z_2 = Z_1' \times Z_2', \ldots Z_K = Z_1' \times Z_2' \times \ldots \times Z_K' \), which is common when sampling from state-space models.

In the first case, we introduce a forward kernel \( q_k(z_k | z_{k-1}; \phi_k) \), and a reverse kernel \( r_k(z_{k-1} | z_k; \phi_k) \). This yields a forward and reverse density on the extended space \( Z_k \times Z_{k-1} \),

\[
\gamma_k(z_k, z_{k-1}) = q_k(z_k | z_{k-1}) \gamma_{k-1}(z_{k-1}), \quad \hat{\gamma}_k(z_k, z_{k-1}) = \gamma_k(z_k) r_k(z_{k-1} | z_k).
\]

In the latter case, we can omit the construction of a reverse kernel as the reverse density at every step is fully specified by the corresponding intermediate density on \( Z_k' \times Z_{k-1} \)

\[
\gamma_k(z_{1:k}) = q_k(z_k' | z_{k-1}) \gamma_k(z_{k-1}) = q_k(z_k' | z_{1:k-1}) \gamma_k(z_{1:k-1}) \quad \hat{\gamma}_k(z_k) = \gamma_k(z_k) = \gamma_k(z_{1:k}).
\]
Nested Importance Samplers. Nested importance samplers (Naesseth et al., 2015, 2019) define a sampling problem recursively by generalizing from standard sequential importance sampling (SIS) methods. Suppose that we have a mechanism by which we can generate weighted samples \((w_{k-1}, z_{k-1})\) from the target density \(\pi_{k-1}(z_{k-1})\). We can construct samples \((w_k, z_k)\) that target the next density by the following construction,

\[
z_k \sim q_k(\cdot \mid z_{k-1}), \quad w_k = v_k w_{k-1}, \quad v_k = \frac{z_k(z_k, z_{k-1})}{\gamma_k(z_k, z_{k-1})}, \quad \tilde{v}_k = \frac{Z_{k-1}}{Z_k} v_k,
\]

where \(\tilde{v}_k\) denotes the incremental weight computed w.r.t. the normalized densities. This notion of compositionality can be formalized by introducing the concept of proper weighting.

**Definition 1 (Proper weighting)** Let \(\pi\) be a probability density. A random pair \((w, z)\sim \Pi\) for an unnormalized probability density \(\gamma = Z\pi\) if \(w \geq 0\) and for all measurable functions \(g\) it holds that

\[
\mathbb{E}_{z, w \sim \Pi}[wg(z)] = \int dz \gamma(z) g(z) = cZ \mathbb{E}_{z \sim \pi}[g(z)]
\]

for some constant \(c > 0\).

Hence, given properly weighted samples \((z', w')\sim \Pi\) this ensures that the we can compute strongly consistent self-normalized estimates

\[
\frac{1}{L} \sum_{l=1}^{L} w' g(z') \overset{a.s.}{\rightarrow} \frac{cZ}{cZ} \mathbb{E}_{z \sim \pi}[g(z)] = \mathbb{E}_{z \sim \pi}[g(z)].
\]

In other words, proper weighting ensures that the bias of our self-normalized estimators vanishes in the limit of infinite samples.

**Compositionality of Proper Weighting.** The proper weighting property allows us to reason about the validity of compositions of sampling operations in a straightforward manner. If we can show that individual operations preserve proper weighting, then any composition of these operations also preserves proper weighting. Importance sampling and hence the sequential construction in Equation 1 preserves proper weighting, which is to say that if \((w_{k-1}, z_{k-1})\) is properly weighted w.r.t. \(\gamma_{k-1}(z_{k-1})\), then \((w_k, z_k)\) is properly weighted w.r.t. \(\gamma_k(z_k)\). Other operations that preserve proper weighting are rejection sampling, the application of an MCMC transition operator, and most notably importance resampling, which forms the basis for sequential Monte Carlo methods. Composition of these operations leads to a broad class of verifiably correct importance samplers that admit many existing methods as special cases.

**Nested Variational Objectives.** We are interested in defining variational objectives that can be used to optimize the parameters of nested importance samplers. As before, we will for purposes of exposition restrict ourselves to samplers that follow the sequential construction in Equation 1. Given an initial target density \(\gamma_1(z_1)\) and initial proposal \(q_1(z_1)\), we define objectives which minimizes an \(f\)-divergence based term at each level of nesting

\[
\mathcal{D} = D_f(\pi_1 \mid \mid q_1) + \sum_{k=2}^{K} D_f(\tilde{\pi}_k \mid \mid \hat{\pi}_k).
\]
An $f$-divergence $D_f(p \parallel q) = \mathbb{E}_{z \sim q}[f(p(z)/q(z))]$ is parameterized by a convex function $f$ that satisfies $f(1) = 0$. When $f_{\text{exc}}(w) := f(w) = -\log(w)$ we recover the exclusive KL divergence, whereas $f_{\text{inc}}(w) := f(w) = w \log(w)$ recovers the inclusive KL divergence. We can write each $f$-divergence in terms of the incremental weight $v_k$

$$D_f(\tilde{\pi}_k \parallel \hat{\pi}_k) = \mathbb{E}_{\hat{\pi}_k} \left[ f \left( \frac{\hat{\pi}_k(z_k, z_{k-1}; \theta_k, \phi_k)}{\tilde{\pi}_k(z_k, z_{k-1}; \theta_k, \phi_k)} \right) \right] = \mathbb{E}_{\hat{\pi}_k} \left[ f \left( v_k \frac{Z_{k-1}}{Z_k} \right) \right] = \mathbb{E}_{\hat{\pi}_k} \left[ f \left( \tilde{v}_k \right) \right]$$

In the following we assume that all parameters $\theta_K$ of the final target density are known or estimated by way of maximum likelihood estimation. Our goal is to minimize $D$ with respect to parameters $\{\theta_k\}_{k=1}^{K-1}$ of the intermediate densities and parameters $\{\hat{\phi}_k\}_{k=2}^K$ and $\{\hat{\phi}_k\}_{k=2}^K$ of the forward and reverse kernels at each level of nesting. Intuitively, placing consecutive intermediate densities closer to each other should result in an easier learning problem for the corresponding forward and reverse kernels, while bringing the forward and reverse densities closer should result in an easier sampling problem, e.g optimising a Pearson $\chi^2$-divergence can be shown to minimize the variance of the importance weight (Müller et al., 2019). Here we consider KL-divergences only but still take motivation from this intuition. To the best of our knowledge, optimizing a sequence of divergences combined with the ability to learn the parameters of the intermediate densities is novel to NVI.

**Gradient Computation** Building on the NIS framework described above, we are able to compute consistent self-normalized gradient estimators as shown in Equation 2. While the computation of gradient estimates for the parameters $\hat{\phi}_k$ of the reverse kernel and reparameterized gradients estimates for parameters $\hat{\phi}_k$ of the forward kernel is straightforward, estimating the gradients w.r.t. parameters of the intermediate densities $\theta_{k-1}$ and $\theta_k$ is less convenient. E.g. in case of the exclusive KL-divergence, the gradient w.r.t. $\theta_{k-1}$ is computed using a score function estimator, which uses an additional baseline, and both, the estimators for $\theta_{k-1}$ and $\theta_k$ require to estimate the gradient of the respective log normalizing constants. Detailed deviation of the gradient estimators can be found in Appendix D.

3. Experiments

We evaluate NVI on two tasks, (1) learning to sample form an unnormalized target density where intermediate densities are generated along a geometric annealing path, and (2) learning intermediate densities to generate posterior samples of a hidden Markov model.

3.1. Sampling from an unnormalized target density via annealing

We are targeting a 2-dimensional unnormalized Gaussian mixture model (GMM) $\gamma_K$ with $M = 8$ equidistantly spaced modes along a circle a round the origin. Starting from a initial proposal $q_1(z_1) = \gamma_1(z_1)$ we construct a sequence of $K$ annealed densities

$$\gamma_k(z) = q_1(z)^{1-\beta_k} \gamma_K(z)^{\beta_k}, \quad \beta_k = \frac{k - 1}{K - 1}, \quad \text{for } k = 1 \ldots K$$

equi-distantly scheduled along a geometric annealing path. We define the corresponding forward and reverse densities using forward and reverse kernel,

$$\tilde{\gamma}_k(z_k, z_{k-1}) = q_k(z_k \mid z_{k-1}) \gamma_{k-1}(z_{k-1}), \quad \tilde{\gamma}_k(z_k, z_{k-1}) = \gamma_k(z_k) r_k(z_{k-1} \mid z_k).$$
Figure 1: (Top) Exemplary samples from forward kernels trained with AVO, and NVI*. The rightmost column shows ground truth samples from the GMM target. (Bottom-Left) Annealing schedules learned by NVI* and NVIR* and the equi-distant annealing schedule (EQUI) used by AVO, NVI, and NVIR. Results are averaged over 10 independent restarts, error bars indicate two standard deviations. (Bottom-Right) The KL-divergences (computed by numeric integration) between consecutive intermediate distributions based on equi-distant and learned schedules.

We compare 4 different variants of Nested Variational Inference (NVI, NVIR, NVI*, and NVIR*), which optimize an exclusive KL-divergence at every step, and Annealed Variational objectives (AVO) (Huang et al., 2018). NVIR employs additional resampling after every step, NVI* additionally learns the annealing schedule of the intermediate densities, and NVIR* combines both. All methods use the architecture described above for the forward and reverse kernels and are trained for 20,000 iteration using Adam with a learning rate of $10^{-3}$. A detailed description of the model and architecture can be found in Appendix C.1.

We report the sample quality of the learned samplers in terms of the log average weight and effective sample size in Table 1 and show the learned annealing schedules and samples from the intermediate densities in Figure 1. Our results show that samplers trained with NVI are able to more accurately estimate the log normalizing constant whilst maintaining a higher effective sample sizes compared to AVO. Moreover, NVI* and NVIR* learn more equi-distantly spaced annealing schedules in terms of KL-divergence. Both learning the annealing schedule and resampling empirically helps to learn better samplers.

### 3.2. Learning Intermediate Heuristic for Hidden Markov Models

Here, we are considering a Hidden Markov Model (HMM) with a GMM likelihood consisting of $M$ cluster (see Appendix C.2) with data points $x_{1:T}$, global variables $\eta$, and hidden states $z_{1:T}$. We can define a sequence of unnormalized target densities $\{\gamma_t\}_{t=0}^T$ as

$$
\gamma_0(x_{1:T}, \eta) = p(\eta), \quad \gamma_t(z_{1:t}, x_{1:T}, \eta) = p(z_{1:t}, x_{1:t}, \eta).
$$
Table 1: (Left) Experiment 1: AVO and NVI-variants trained for different numbers of annealing steps $K$ and particles per step $L$ for fixed budget of $K \cdot L = 288$ samples. We report the log average weight ($\log \hat{Z}$) and effective sample size (ESS) for 1000 samples per step at test time. All numbers are averages over 100 batches across 10 independent restarts. (Right) Experiment 2: NVIR for different types of heuristics.

Intuitively, this evaluates the utility of a sample for the global variables $\eta$ based on the probability under the HMM up to current step $t$, neglecting future observations $x_{t:T}$. If the observations up to step $t$ do only contain a small subset of possible states this can lead to mode collapse. Here we design a sequence of target densities by learning a heuristic factor $\Psi(\cdot; \theta)$ as part of our target densities

$$
\gamma_0(x_{1:T}, \eta; \theta) = p(\eta)\Psi(x_{1:T} \mid \eta; \theta), \quad \gamma_t(z_{1:t}, x_{1:T}, \eta; \theta) = p(z_{1:t}, x_{1:t}, \eta)\Psi(x_{t+1:T} \mid \eta; \theta).
$$

The heuristic factor evaluates the likelihood of future data points given the global variables. We consider two heuristic factors which enumerate over individual clusters: (1) a GMM-style heuristic factor and (2) a neural heuristic factor

$$
\Psi^{\text{GMM}}(x_{t:T}; \eta) = \prod_{t=t}^{T} \sum_{m=1}^{M} p(x_t \mid z_t = m, \eta)p(z_t = m),
$$

$$
\Psi^{\text{NEURAL}}(x_{t:T}; \eta, \theta) = \prod_{t=t}^{T} \sum_{m=1}^{M} p(x_t \mid z_t = m, \eta)\psi(z_t = m; x_t, \eta, \theta),
$$

We compare four different variants of NVI, NVIR without a heuristic factor, NVIR-GMM using the GMM-style heuristic factor, NVIR* and NVIR*, which both use the neural heuristic factor. We found NVIR*, a variant of NVIR* which detaches a part of the gradient, described in appendix C.2, using an inclusive KL-divergence on a hidden

Table 1 shows that, as expected, employing a heuristic factor results in better sample quality in terms of log average importance weight and effective sample size. Moreover the neural heuristic factor outperforms the GMM-style heuristic factor.

4. Conclusion

We develop NVI, a framework that combines nested importance sampling and variational inference by optimizing a variational objective at every level of nesting. The formulation allows to learn proposals and intermediate densities for a general class of samplers, which admit most commonly used importance sampling strategies as special cases. Our experiments demonstrate that samplers, targeting (a) an unnormalized GMM and (b) the posterior of a HMM, trained with NVI are able to outperform baselines in terms of log average weight and effective sampling size. Moreover, we found that learning intermediate distributions based on the inclusive and exclusive KL-divergence results in better samplers in our experiments.
Acknowledgments

We would like to thank our reviewers for their thoughtful comments. This work was supported by the Intel Corporation, the 3M Corporation, NSF award 1835309, startup funds from Northeastern University, the Air Force Research Laboratory (AFRL), and DARPA.

References


Appendix A. Notation

\[ u_k = \sum_{k' = 1}^{k-1} \pi_k(z_{k-1}) q_{k'}(z_{k-1} | z_k, \phi_k) \]

\[ \hat{\gamma}_k(z_k) = Z_k \hat{\pi}_k(z_k) \]

\[ \hat{\pi}_k(z_k, z_{k-1}) = \pi_k(z_k) r_k(z_{k-1} | z_k, \phi_k) \]

\[ \hat{\gamma}_k(z_k, z_{k-1}) = Z_k \hat{\pi}_k(z_k, z_{k-1}) \]

Appendix B. Important Identities

Thermodynamic Identity:

\[ \frac{d}{d\theta} \log Z_\theta = \frac{1}{Z_\theta} \frac{d}{d\theta} \int z \gamma(z; \theta) dz = \int z \frac{\gamma(z; \theta)}{Z_\theta} \frac{d}{d\theta} \log \gamma(z; \theta) = \mathbb{E}_{z \sim \pi(z; \theta)} \left[ \frac{\partial}{\partial \theta} \log \gamma \right]. \]

Log-derivative trick a.k.a. reinforce trick:

\[ \frac{d}{d\theta} \pi(z; \theta) = \pi(z) \frac{1}{\pi(z; \theta)} \frac{d}{d\theta} \pi(z; \theta) = \pi(z) \frac{d}{d\theta} \log \pi(z; \theta) \]

(3)

Consequently, it holds that

\[ \mathbb{E}_{z \sim \pi(z; \theta)} \left[ \frac{d}{d\theta} \log \pi(z; \theta) \right] = \int z \pi(z; \theta) \frac{d}{d\theta} \log \pi(z; \theta) = \int z \frac{d}{d\theta} \pi(z; \theta) = \frac{d}{d\theta} \int z \pi(z; \theta) = 0 \]

Fisher’s Identity:

\[ \nabla_{\theta} \log p_\theta(x) = \int dz p_\theta(z | x) \frac{d}{d\theta} \log p_\theta(x, z) \]

Appendix C. Experiment Details

C.1. Experiment 1: Annealing

We are targeting an unnormalized Gaussian mixture model (GMM) \( \gamma_K \) with \( M = 8 \) equidistantly spaced modes along a circle with radius \( r = 10 \),

\[ \gamma_K(z_K) = \sum_{k=1}^{M} \mathcal{N}(z_K; \mu_m, \sigma^2 I_{2 \times 2}), \quad \mu_m = \left( r \sin \left( \frac{2m\pi}{M} \right), r \cos \left( \frac{2m\pi}{M} \right) \right), \]

for \( m = 1, 2, \ldots, M \) and \( \sigma = 0.5 \). The model the forward and backward kernel as conditional Gaussian, where the mappings for the means \( \mu_k \) and standard deviations \( \sigma_k \) consist of a multilayer perceptron with a single share hidden layer of 50 neurons with sigmoid activation functions

\[ q_k(z_k | z_{k-1}) = \mathcal{N}(z_k; z_{k-1} + \mu_k(z_{k-1}), \Sigma_k(z_{k-1})), \]

\[ r_k(z_{k-1} | z_k) = \mathcal{N}(z_{k-1}; z_k + \mu_k(z_k), \Sigma_k(z_k)). \]
C.2. Experiment 2: Hidden Markov Models

In the second set of experiments, we evaluate NVIR on a hidden Markov model with a GMM likelihood over data points $x_{1:T}$, global variables for the GMM components $\eta = \{\tau_1:M, \mu_1:M\}$, and local states $z_{1:T}$,

$$
\begin{align*}
\tau_m, \mu_m &\sim \text{NormGamma}(\alpha_0, \beta_0, \mu_0, \nu_0), & m = 1, 2, ..., M, \\
z_1 &\sim \text{Cat}(\pi), \\
z_t | z_{t-1} = m &\sim \text{Cat}(A_m) & t = 1, 2, ..., T, \\
x_t | z_t = m &\sim \text{Norm}(\mu_m, \sigma_m), & t = 1, 2, ..., T.
\end{align*}
$$

We then construct forward densities by defining proposals $q_0(\eta|x_{1:T}; \phi), q_1(z_1|x_1, \eta; \phi)$ and $q_t(z_t|z_{t-1}, x_t, \eta; \phi)$ for $t = 2, \ldots T$. At each step, we define a inclusive KL divergence

$$
\begin{align*}
\mathcal{L}_0(\phi) &= \text{KL}(\pi(x_{1:T}, \eta) \| q_0(\eta|x_{1:T}; \phi)), \\
\mathcal{L}_1(\phi, \theta_0) &= \text{KL}(\pi(z_1, x_{1:T}, \eta) \| \pi(x_{1:T}, \eta; \theta_0)q_1(z_1|x_1, \eta; \phi)) \\
\mathcal{L}_t(\phi, \theta_{t-1}) &= \text{KL}(\pi(z_{t:t-1}, x_{1:T}, \eta; \theta_{t-1})q(z_t|z_{t-1}, x_t, \eta; \phi)), & t=2, 3, ..., T.
\end{align*}
$$

We empirically found that we achieve better results when detaching the target density (i.e. the left-hand-side density in each KL) in the objective. This version of NVIR* is
denoted NVIR\(^*\). We compute self-normalized gradient estimates

\[-\nabla_\phi \mathcal{L}_0(\phi) = \mathbb{E}_{\pi(x_1:T, \eta)} [\nabla_\phi \log q_0(\eta|x_1:T; \phi)]
\]

\[\simeq \sum_{s=1}^S w_0^s \sum_{s'} w_{0}^{s'} \nabla_\phi \log q_0(\eta^s|x_1:T; \phi)\]

\[-\nabla_\phi,\theta_0 \mathcal{L}_1(\phi, \theta_0) = \mathbb{E}_{\pi(z_1,x_1:T, \eta)} [\nabla_\phi,\theta_0 (\log \pi(x_1:T, \eta; \theta_0) + \log q_1(z_1|x_1, \eta; \phi))]
\]

\[\simeq \sum_{s=1}^S w_1^s \sum_{s'} w_{1}^{s'} \nabla_\phi,\theta_0 (\log \pi(x_1:T, \eta^s; \theta_0) + \log q_1(z_1^s|x_1, \eta^s; \phi))\]

\[-\nabla_\phi,\theta_{t-1} \mathcal{L}_t(\phi, \theta_{t-1}) = \mathbb{E}_{\pi(z_1,t,x_1:T, \eta)} [\nabla_\phi,\theta_{t-1} (\log \pi(z_1:t-1, x_1:T, \eta; \theta_{t-1} + \log q(z_t|z_{t-1}, x_t, \eta; \phi))]
\]

\[\simeq \sum_{s=1}^S w_t^s \sum_{s'} w_{t}^{s'} \nabla_\phi,\theta_{t-1} (\log \pi(z_1^{s}, x_1:T, \eta^s; \theta_{t-1}) + \log q(z_1^s|z_{t-1}^s, x_t, \eta^s; \phi))\]

where the importance weights are defined as

\[w_0^s = v_0^s = \frac{\gamma(x_1:T, \eta^s)}{q_0(\eta^s|x_1:T)}, \quad (4)\]

\[w_1^s = v_1^s w_0^s = \frac{\gamma(z_1^s, x_1:T, \eta^s)}{\gamma(x_1:T, \eta^s; \theta_0) q_1(z_1^s|x_1, \eta^s; \phi)} w_0^s, \quad (5)\]

\[w_t^s = v_t^s w_{t-1}^s = \frac{\gamma(z_1^{s}, x_1:T, \eta^s)}{\gamma(z_1^{t-1}, x_1:T, \eta^s; \theta_{t-1}) q(z_1^{s}|z_{t-1}^{s}, x_t, \eta^s; \phi)} w_{t-1}^s. \quad (6)\]

Figure 3: Qualitative results of the HMM experiment. Visualization of single samples of predicted means (horizontal lines) and standard deviations (error bars) for one test HMM instance with 100 data points.

C.3. Architectures of the Proposals

We model the the proposal for global variables \(\eta := \{\tau_1:M, \mu_1:M\}\) as Normal-Gamma distribution, where the mapping for its parameters \(\alpha_1:M, \beta_1:M, \mu_1:M, \nu_1:M\) consists of a LSTM.
with 2 layers, followed by two individual multilayer perceptron with a single hidden layer of 128 neurons,

$$q_0(\tau_{1:M}, \mu_{1:M}|x_{1:T}) = \prod_{m=1}^{M} \text{NormGamma}(\tau_m, \mu_m; \alpha_m, \beta_m, \nu_m)$$  \hspace{1cm} (7)

We model the initial state proposal and consecutive state proposal as Categorical distributions, where the mapping for its parameters \(\pi_t\) consists of a multilayer perceptron with a single hidden layer of 128 neurons and Tanh activation function,

$$q_1(z_1|x_1, \tau_{1:M}, \mu_{1:M}) = \text{Cat}(z_1; \pi_1),$$  \hspace{1cm} (8)

$$q(z_t|z_{t-1}, x_t, \tau_{1:M}, \mu_{1:M}) = \text{Cat}(z_t; \pi_t), \hspace{0.5cm} t = 2, 3, ..., T.$$  \hspace{1cm} (9)
Appendix D. Gradient estimation

To compute the gradient of the nested variational objective (NVO) we need to compute the gradients of the individual terms $D_f(\hat{\pi}_k \parallel \tilde{\pi}_k)$ w.r.t. parameters $\hat{\phi}_k, \tilde{\phi}_k, \theta_k,$ and $\theta_{k-1}$.

$$
\frac{dD}{d\hat{\phi}_k} = \frac{dD}{d\hat{\phi}_k} (\frac{\hat{\pi}_k}{\tilde{\pi}_k}) = \frac{dD}{d\hat{\phi}_k} (\hat{\pi}_k) \cdot \frac{\hat{\pi}_k}{\tilde{\pi}_k}
$$

$$
\frac{dD}{d\tilde{\phi}_k} = \frac{dD}{d\tilde{\phi}_k} (\frac{\hat{\pi}_k}{\tilde{\pi}_k}) = \frac{dD}{d\tilde{\phi}_k} (\hat{\pi}_k) \cdot \frac{\hat{\pi}_k}{\tilde{\pi}_k}
$$

$$
\frac{dD}{d\theta_k} = \frac{dD}{d\theta_k} (\frac{\hat{\pi}_k}{\tilde{\pi}_k}) + \frac{dD}{d\theta_k} (\hat{\pi}_k) \cdot \frac{\hat{\pi}_k}{\tilde{\pi}_k}
$$

$$
\frac{dD}{d\theta_{k-1}} = \frac{dD}{d\theta_{k-1}} (\frac{\hat{\pi}_k}{\tilde{\pi}_k}) + \frac{dD}{d\theta_{k-1}} (\hat{\pi}_k) \cdot \frac{\hat{\pi}_k}{\tilde{\pi}_k}
$$

In the following we are deriving the relevant gradients for the general f-divergence, exclusive KL-divergence ($f(w) = -\log w$), and inclusive KL-divergence ($f(w) = w \log w$).

D.1. Gradients for general f-divergences

Gradient w.r.t. parameters $\hat{\phi}_k$ of the forward kernel: Reparameterizing the sample $z_k \equiv z_k(\epsilon_k; \hat{\phi}_k)$ allows us, under mild conditions $^1$, to interchange the order of integration and differentiation and compute path-wise derivatives.

$$
\frac{d}{d\hat{\phi}_k} D_f (\hat{\pi}_k \parallel \tilde{\pi}_k) = \mathbb{E}_{z_{k-1} \sim \hat{\pi}_{k-1}} \left[ \mathbb{E}_{\epsilon_k \sim p_k} \left[ \frac{d}{d\hat{\phi}_k} f \left( v_k \frac{Z_{k-1}}{Z_k} \right) \right] \right]
$$

$$
= \mathbb{E}_{z_{k-1} \sim \hat{\pi}_{k-1}} \left[ \mathbb{E}_{\epsilon_k \sim p_k} \left[ \frac{\partial}{\partial \hat{\phi}_k} f \left( \frac{Z_{k-1}}{Z_k} \frac{Z_{k-1} \partial v_k \partial z_k}{Z_k \partial z_k \partial \hat{\phi}_k} \right) \right] \right]
$$

$$
= \mathbb{E}_{z_{k-1} \sim \hat{\pi}_{k-1}} \left[ \mathbb{E}_{\epsilon_k \sim p_k} \left[ \frac{\partial}{\partial \hat{\phi}_k} f \left( \frac{Z_{k-1}}{Z_k} \frac{Z_{k-1} \partial v_k \partial z_k}{Z_k \partial z_k \partial \hat{\phi}_k} \right) \right] \right]
$$

$$
= \mathbb{E}_{z_{k-1} \sim \hat{\pi}_{k-1}} \left[ \mathbb{E}_{\epsilon_k \sim p_k} \left[ \frac{\partial}{\partial \hat{\phi}_k} f \left( \frac{Z_{k-1}}{Z_k} \frac{Z_{k-1} \partial v_k \partial z_k}{Z_k \partial z_k \partial \hat{\phi}_k} \right) \right] \right]
$$

$$
= \mathbb{E}_{w_{k-1} \sim Z_{k-1} \sim \hat{\pi}_{k-1}} \left[ \mathbb{E}_{\epsilon_k \sim p_k} \left[ \frac{\partial}{\partial \hat{\phi}_k} f \left( \frac{Z_{k-1}}{Z_k} \frac{Z_{k-1} \partial v_k \partial z_k}{Z_k \partial z_k \partial \hat{\phi}_k} \right) \right] \right]
$$

Alternatively, we can compute a score function gradient which does not require the target density $\gamma_k$ to be differentiable w.r.t. the sample $z_k$ and hence can also be computed for

$^1$ Leibniz Integration Rules
discrete variable models.

\[
\frac{d}{d\phi_k} \mathcal{D}_f \left( \tilde{\pi}_k \mid \hat{\pi}_k \right) = \mathbb{E}_{z_{k-1} \sim \pi_{k-1}} \left[ \int_{Z_k} dz_k \frac{d}{d\phi_k} \left( q_k(z_k \mid z_{k-1}, \hat{\phi}_k) f \left( v_k \frac{Z_{k-1}}{Z_k} \right) \right) \right]
\]

\[
= \mathbb{E}_{z_{k-1} \sim \pi_{k-1}} \mathbb{E}_{z_k \sim q_k(\cdot \mid z_{k-1}, \hat{\phi}_k)} \left[ f \left( v_k \frac{Z_{k-1}}{Z_k} \right) \frac{\partial \log q_k}{\partial \phi_k} + \frac{\partial f}{\partial w} \bigg|_{w=v_k} \frac{z_{k-1}}{z_k} \frac{Z_{k-1}}{Z_k} \frac{\partial \log v_k}{\partial \phi_k} \right]
\]

\[
= \mathbb{E}_{z_{k-1} \sim \pi_{k-1}} \mathbb{E}_{z_k \sim q_k(\cdot \mid z_{k-1}, \hat{\phi}_k)} \left[ \left( f \left( v_k \frac{Z_{k-1}}{Z_k} \right) - \frac{\partial f}{\partial w} \bigg|_{w=v_k} \frac{z_{k-1}}{z_k} \frac{Z_{k-1}}{Z_k} \right) \frac{\partial \log q_k}{\partial \phi_k} \right]
\]

Gradient w.r.t. parameters \( \hat{\phi}_k \) of the reverse kernel:

\[
\frac{d}{d\phi_k} \mathcal{D}_f \left( \tilde{\pi}_k \mid \hat{\pi}_k \right) = \mathbb{E}_{z_{k-1}, z_k \sim \hat{\pi}_k} \left[ \frac{d}{d\phi_k} f \left( v_k \frac{Z_{k-1}}{Z_k} \right) \right]
\]

\[
= \mathbb{E}_{z_{k-1}, z_k \sim \hat{\pi}_k} \left[ \frac{\partial f}{\partial w} \bigg|_{w=v_k} \frac{z_{k-1}}{z_k} \frac{Z_{k-1}}{Z_k} \frac{\partial \log v_k}{\partial \phi_k} \right]
\]

\[
= \mathbb{E}_{z_{k-1}, z_k \sim \hat{\pi}_k} \left[ \frac{\partial f}{\partial w} \bigg|_{w=v_k} \frac{z_{k-1}}{z_k} \frac{Z_{k-1}}{Z_k} \frac{\partial \log r_k}{\partial \phi_k} \right]
\]

\[
p.u. \quad \mathbb{E}_{w_{k-1}, z_{k-1} \sim \Pi_{k-1}} \left[ \frac{w_{k-1}}{cZ_{k-1}} \mathbb{E}_{z_k \sim q_k(\cdot \mid z_{k-1}, \hat{\phi}_k)} \left[ \frac{\partial f}{\partial w} \bigg|_{w=v_k} \frac{z_{k-1}}{z_k} \frac{Z_{k-1}}{Z_k} \frac{\partial \log r_k}{\partial \phi_k} \right] \right]
\]
Gradient w.r.t. parameters $\theta_k$ of the current target

\[
\frac{d}{d\theta_k} D_f (\hat{\pi}_k \mid \hat{\pi}_k) = \mathbb{E}_{Z_{k-1} \sim \hat{\pi}_k} \left[ \frac{d}{d\theta_k} f \left( v_k \frac{Z_{k-1}}{Z_k} \right) \right] \\
= \mathbb{E}_{Z_{k-1} \sim \hat{\pi}_k} \left[ \frac{\partial f}{\partial w} \bigg|_{w=v_k} \frac{Z_{k-1}}{Z_k} \frac{\partial v_k}{\partial \theta_k} \right] \\
= \mathbb{E}_{Z_{k-1} \sim \hat{\pi}_k} \left[ \frac{\partial f}{\partial w} \bigg|_{w=v_k} \frac{Z_{k-1}}{Z_k} \frac{\partial \log v_k}{\partial \theta_k} \right] \\
= \mathbb{E}_{Z_{k-1} \sim \hat{\pi}_k} \left[ \frac{\partial f}{\partial w} \bigg|_{w=v_k} \frac{Z_{k-1}}{Z_k} \frac{\partial \log \pi_k}{\partial \theta_k} \right] \\
= \mathbb{E}_{Z_{k-1} \sim \hat{\pi}_k} \left[ \frac{\partial f}{\partial w} \bigg|_{w=v_k} \frac{Z_{k-1}}{Z_k} \frac{\partial \log \gamma_k}{\partial \theta_k} \right] \\
= \mathbb{E}_{Z_{k-1} \sim \hat{\pi}_k} \left[ \frac{\partial f}{\partial w} \bigg|_{w=v_k} \frac{Z_{k-1}}{Z_k}, \frac{\partial \log \gamma_k}{\partial \theta_k} \right] \\
= \text{Cov}_{\hat{\pi}_k} \left[ \frac{\partial f}{\partial w} \bigg|_{w=v_k} \frac{Z_{k-1}}{Z_k}, \frac{\partial \log \gamma_k}{\partial \theta_k} \right] \\
p.w. \quad \mathbb{E}_{v_k \mid \phi_k \sim \Pi_{k-1}} \left[ \frac{w_{k-1}}{c Z_{k-1}} \frac{\partial \log \gamma_k}{\partial \theta_k} \right] \\
= \mathbb{E}_{v_k \mid \phi_k \sim \Pi_{k-1}} \left[ \frac{w_{k-1}}{c Z_{k-1}} \frac{\partial \log \gamma_k}{\partial \theta_k} \right]
\]
Gradient w.r.t. parameters $\theta_{k-1}$ of the current proposal

$$
\frac{d}{d\theta_{k-1}} D_f (\hat{\pi}_k \parallel \hat{\pi}_k) = \frac{d}{d\theta_{k-1} z_{k-1}, z_k \sim \hat{\pi}_k} \mathbb{E} \left[ f \left( \frac{v_k Z_{k-1}}{Z_k} \right) \right] = \frac{d}{d\theta_{k-1} z_{k-1} \sim \pi_{k-1}} \mathbb{E} \left[ z_k \sim q_k (\cdot | z_{k-1}, \hat{\phi}_k) \mathbb{E} \left[ f \left( \frac{v_k Z_{k-1}}{Z_k} \right) \right] \right]
$$

$$
= \int_{z_{k-1}} dz_{k-1} \frac{d}{d\theta_{k-1}} \left( \pi_{k-1}(z_{k-1}; \theta_{k-1}) \mathbb{E} \left[ z_k \sim q_k (\cdot | z_{k-1}, \hat{\phi}_k) \mathbb{E} \left[ f \left( \frac{v_k Z_{k-1}}{Z_k} \right) \right] \right] \right)
$$

$$
= \int_{z_{k-1}} dz_{k-1} \frac{d}{d\theta_{k-1}} \left( \pi_{k-1}(z_{k-1}; \theta_{k-1}) \frac{\partial \log \pi_{k-1}}{\partial \theta_{k-1}} \mathbb{E} \left[ z_k \sim q_k (\cdot | z_{k-1}, \hat{\phi}_k) \mathbb{E} \left[ f \left( \frac{v_k Z_{k-1}}{Z_k} \right) \right] \right] \right)
$$

$$
+ \pi_{k-1}(z_{k-1}; \theta_{k-1}) \frac{\partial}{\partial \theta_{k-1}} \mathbb{E} \left[ z_k \sim q_k (\cdot | z_{k-1}, \hat{\phi}_k) \mathbb{E} \left[ f \left( \frac{v_k Z_{k-1}}{Z_k} \right) \right] \right]
$$

$$
= \mathbb{E}_{z_{k-1} \sim \pi_{k-1}} \left[ \frac{\partial \log \pi_{k-1}}{\partial \theta_{k-1}} \mathbb{E}_{z_k \sim q_k (\cdot | z_{k-1}, \hat{\phi}_k)} \left[ f \left( \frac{v_k Z_{k-1}}{Z_k} \right) \right] \right] + \mathbb{E}_{z_{k-1} \sim \pi_{k-1}} \left[ \frac{\partial}{\partial \theta_{k-1}} \mathbb{E}_{z_k \sim q_k (\cdot | z_{k-1}, \hat{\phi}_k)} \left[ f \left( \frac{v_k Z_{k-1}}{Z_k} \right) \right] \right]
$$

$$
= \mathbb{E}_{z_{k-1}, z_k \sim \hat{\pi}_k} \left[ \frac{\partial \log \pi_{k-1}}{\partial \theta_{k-1}} \mathbb{E}_{z_k \sim q_k (\cdot | z_{k-1}, \hat{\phi}_k)} \left[ f \left( \frac{v_k Z_{k-1}}{Z_k} \right) \right] \right] - \mathbb{E}_{z_{k-1}, z_k \sim \hat{\pi}_k} \left[ \frac{\partial}{\partial \theta_{k-1}} \mathbb{E}_{z_k \sim q_k (\cdot | z_{k-1}, \hat{\phi}_k)} \left[ f \left( \frac{v_k Z_{k-1}}{Z_k} \right) \right] \right]
$$

$$
= \text{Cov}_{\hat{\theta}_k} \left[ f \left( \frac{v_k Z_{k-1}}{Z_k} \right), \frac{\partial \log \gamma_{k-1}}{\partial \theta_{k-1}} \right] - \text{Cov}_{\hat{\theta}_k} \left[ \frac{\partial}{\partial \theta_{k-1}} \mathbb{E}_{z_k \sim q_k (\cdot | z_{k-1}, \hat{\phi}_k)} \left[ f \left( \frac{v_k Z_{k-1}}{Z_k} \right) \right] \right]
$$

$$
= \frac{1}{w_{k-1} \sim \Pi_{k-1}} \mathbb{E} \left[ w_{k-1} \frac{\partial}{\partial \theta_{k-1}} \mathbb{E}_{z_k \sim q_k (\cdot | z_{k-1}, \hat{\phi}_k)} \left[ f \left( \frac{v_k Z_{k-1}}{Z_k} \right) \right] \right] - \frac{1}{w_{k-1} \sim \Pi_{k-1}} \mathbb{E} \left[ w_{k-1} \frac{\partial}{\partial \theta_{k-1}} \mathbb{E}_{z_k \sim q_k (\cdot | z_{k-1}, \hat{\phi}_k)} \left[ f \left( \frac{v_k Z_{k-1}}{Z_k} \right) \right] \right]
$$
D.2. Gradients for the exclusive KL-divergence \((f(w) = -\log(w))\)

Building on the deviations for the general case derived in D.1 we derive the gradients for the exclusive KL-divergence as special cases by substituting \(f(w) = -\log(w)\).

**Gradient w.r.t. parameters \(\hat{\phi}_k\) of the forward kernel:**

The reparameterized gradient takes the form

\[
\frac{d}{d\hat{\phi}_k} D_{-\log w} (\tilde{\pi}_k \| \hat{\pi}_k) = \mathbb{E}_{z_{k-1} \sim \pi_{k-1}} \left[ \mathbb{E}_{\epsilon_k \sim p_k} \left[ -\frac{\partial \log v_k}{\partial z_k} \frac{\partial z_k}{\partial \hat{\phi}_k} - \frac{\partial \log q_k}{\partial \hat{\phi}_k} \right] \right]
\]  

(10)

whereas the score function gradient takes the form

\[
\frac{d}{d\hat{\phi}_k} D_{-\log w} (\tilde{\pi}_k \| \hat{\pi}_k) = \mathbb{E}_{z_{k-1} \sim \pi_{k-1}} \left[ \mathbb{E}_{\epsilon_k \sim p_k} \left[ -\log v_k \frac{\partial \log q_k}{\partial \hat{\phi}_k} \right] \right]
\]  

(11)

The final equalities holds due to the reinforce property (Appendix B Equation 3)

\[
\mathbb{E}_{\epsilon_k \sim p_k} \left[ \frac{\partial \log q_k}{\partial \hat{\phi}_k} \bigg|_{z_k = z_k(\epsilon, \phi)} \right] = \mathbb{E}_{z_k \sim q_k(\cdot \| z_{k-1}, \hat{\phi}_k)} \left[ \frac{\partial \log q_k}{\partial \hat{\phi}_k} \right] = 0.
\]

**Gradient w.r.t. parameters \(\hat{\phi}\) of the reverse kernel**

\[
\frac{d}{d\hat{\phi}_k} D_{-\log w} (\tilde{\pi}_k \| \hat{\pi}_k) = \mathbb{E}_{z_{k-1} \sim \hat{\pi}_k} \left[ -\frac{\partial \log r_k}{\partial \hat{\phi}_k} \right].
\]

(14)

**Gradient w.r.t. parameters \(\theta_k\) of the current target**

\[
\frac{d}{d\theta_k} D_{-\log w} (\tilde{\pi}_k \| \hat{\pi}_k) = \mathbb{E}_{z_{k-1} \sim \hat{\pi}_k} \left[ -\frac{\partial \log \gamma_k}{\partial \theta_k} \right] + \mathbb{E}_{z_k \sim \pi_k} \left[ \frac{\partial \log \gamma_k}{\partial \theta_k} \right].
\]

(15)

**Gradient w.r.t. parameters \(\theta_{k-1}\) of the current proposal**

\[
\frac{d}{d\theta_{k-1}} D_{-\log w} (\tilde{\pi}_k \| \hat{\pi}_k) = \text{Cov}_{\hat{\pi}_k} \left[ -\log v_k, -\frac{\partial \log \gamma_{k-1}}{\partial \theta_{k-1}} \right]
\]

(16)
### D.3. Gradients for the inclusive KL-divergence \((f(w) = w \log(w))\)

First notice that

\[
D_{w \log w}(\hat{\pi}_k \| \hat{\pi}_k) = \mathbb{E}_{z_{k-1}, z_k \sim \hat{\pi}_k} [w \log w] \\
= \mathbb{E}_{z_{k-1}, z_k \sim \hat{\pi}_k} [\log w] \\
= \mathbb{E}_{z_{k-1}, z_k \sim \hat{\pi}_k} [- \log w^{-1}] \\
= D_{- \log w}(\hat{\pi}_k \| \hat{\pi}_k). \tag{20}
\]

Hence the gradients for the inclusive KL-divergence follow by symmetry from the gradient of the exclusive KL-divergence by swapping the arguments and identifying the components \(r_k, \pi_k\) and corresponding parameters \(\phi_k, \theta_k\) with the components of the forward density \(q_k, \pi_k^{-1}\) and parameters \(\hat{\phi}_k, \theta_k^{-1}\) respectively.

**Gradient w.r.t. parameters \(\hat{\phi}_k\) of the forward kernel:**

\[
\frac{d}{d\phi_k} D_{w \log w} (\hat{\pi}_k \| \hat{\pi}_k) = \mathbb{E}_{z_{k-1}, z_k \sim \hat{\pi}_k} \left[ - \frac{\partial \log q_k}{\partial \phi_k} \right]. \tag{21}
\]

**Gradient w.r.t. parameters \(\hat{\phi}_k\) of the reverse kernel:** Note that the sample \(z_{k-1}\) is assumed to be not reparameterized. Hence we only state the score-function gradient for the inclusive KL-divergence.

\[
\frac{d}{d\phi_k} D_{w \log w} (\hat{\pi}_k \| \hat{\pi}_k) = \mathbb{E}_{z_{k-1}, z_k \sim \hat{\pi}_k} \left[ \log v_k \frac{\partial \log r_k}{\partial \phi_k} \right]. \tag{21}
\]

**Gradient w.r.t. parameters \(\theta_k\) of the current target**

\[
\frac{d}{d\theta_k} D_{w \log w} (\hat{\pi}_k \| \hat{\pi}_k) = \text{Cov}_{\hat{\pi}_k} \left[ \log v_k, \frac{\partial \log \gamma_k}{\partial \theta_k} \right]. \tag{22}
\]

**Gradient w.r.t. parameters \(\theta_{k-1}\) of the current proposal**

\[
\frac{d}{d\theta_{k-1}} D_{w \log w} (\hat{\pi}_k \| \hat{\pi}_k) = \mathbb{E}_{z_{k-1}, z_k \sim \hat{\pi}_k} \left[ - \frac{\partial \log \gamma_{k-1}}{\partial \theta_{k-1}} \right] + \mathbb{E}_{z_k \sim \pi_{k-1}} \left[ \frac{\partial \log \gamma_{k-1}}{\partial \theta_{k-1}} \right]. \tag{23}
\]