Variational Inference for Sequential Data with Future Likelihood Estimates

Geon-Hyeong Kim [1] **Youngsoo Jang** [1] **Hongseok Yang** [1] **Kee-Eung Kim** [1 2]

Abstract

The recent development of flexible and scalable variational inference algorithms has popularized the use of deep probabilistic models in a wide range of applications. However, learning and reasoning about high-dimensional models with non-differentiable densities are still a challenge. For such a model, inference algorithms struggle to estimate the gradients of variational objectives accurately, due to high variance in their estimates. To tackle this challenge, we present a novel variational inference algorithm for sequential data, which performs well even when the density from the model is not differentiable, for instance, due to the use of discrete random variables. The key feature of our algorithm is that it estimates future likelihoods at all time steps. The estimated future likelihoods form the core of our new low-variance gradient estimator. We formally analyze our gradient estimator from the perspective of variational objective, and show the effectiveness of our algorithm with synthetic and real datasets.

1. Introduction

Learning a deep probabilistic model for complex data with latent variables is one of the most important tasks in machine learning. Such models are typically built using neural networks, and applied to analyze a wide variety of data, including high-dimensional ones such as images (Kingma & Welling, 2014; Rezende & Mohamed, 2015; Chen et al., 2016; Gulrajani et al., 2016), speech and music (Chung et al., 2015; Fraccaro et al., 2016), and videos (Babaeizadeh et al., 2017; Denton & Fergus, 2018). However, when these models do not have differentiable densities (for instance, due to discrete latent variables), learning and analyzing the models for high-dimensional data is still a challenge despite impressive progress in the past few years, including, but not limited to, tighter variational bounds (Burda et al., 2016; Maddison et al., 2017a; Naesseth et al., 2018; Le et al., 2018; Lawson et al., 2018; Masrani et al., 2019), low-variance gradient estimators (Mnih & Gregor, 2014; Mnih & Rezende, 2016; Tucker et al., 2017; Grathwohl et al., 2018), and techniques for preventing the diminishing gradient due to the multiple samples (Rainforth et al., 2018; Tucker et al., 2019).

In this paper, we tackle the challenge for models for sequential data in the context of black-box variational inference (Ranganath et al., 2014). For these models, the noise from the data or the algorithm accumulates over time, and the variance of the algorithm's estimates typically grows exponentially with the number of time steps. As a result, parameter learning and posterior inference via black-box variational inference are particularly difficult for these models. While there have been a large amount of prior work on these models, such as tighter variational bound derived from sequential Monte Carlo (Maddison et al., 2017a; Naesseth et al., 2018; Le et al., 2018; Lawson et al., 2018), the current variational inference algorithms struggle to approximate the gradients of variational bounds accurately, due to high variance in their estimates, especially when the models use discrete latent variables and thus do not have smooth densities; in such a case, the techniques that leverage the smoothness of the model, such as reparameterization trick, are no longer straightforwardly applicable.

We propose a novel gradient estimator for a sequential variant of the *importance-weighted-autoencoder (IWAE)* objective (Burda et al., 2016), which is a provably tighter lower bound to the log marginal likelihood than the more traditional *evidence-lower-bound (ELBO)* objective. In order to approximate the gradient accurately, our estimator computes the estimates of future likelihoods at each time step, by exploiting the recursive nature of sequential models. The estimated future likelihoods are then used to improve the standard score-function-based gradient estimator of the IWAE objective. Specifically, they serve as baselines (i.e. control variates), and also enable to replace a high-variance part of this standard estimator with a low-variance counterpart. Our estimator does not require the differentiability of a model's density, and can be applied to sequential models with both

[1]School of Computing, KAIST, Daejeon, Republic of Korea [2]Graduate School of AI, KAIST, Daejeon, Republic of Korea. Correspondence to: Geon-Hyeong Kim <ghkim@ai.kaist.ac.kr>, Youngsoo Jang <ysjang@ai.kaist.ac.kr>, Hongseok Yang <hongseok.yang@kaist.ac.kr>, Kee-Eung Kim <keeeung.kim@kaist.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

discrete and continuous latent variables. Our use of future likelihoods is inspired by the use of the value functions in reinforcement learning to estimate future rewards (Sutton et al., 1998).

Based on the preliminaries provided in Section 2, we formally introduce our estimator in Section 3. In Section 4, we show that the estimator computes an unbiased estimate of not the IWAE objective, but its new lower bound, although our gradient estimator is derived from the IWAE objective and its gradient estimator. Effectively, it trades off the tightness of the IWAE objective with the reduction of variance. In Section 5, we position our estimator in the context of relevant prior work on variational inference. In Section 6, we report experimental results of our estimator with synthetic and polyphonic music datasets, to show its effectiveness compared to the state-of-the-art algorithms.

2. Preliminaries

2.1. Model learning via variational inference

Consider a probabilistic model defined by a parameterized density function $p_\theta(x, z)$ over random variables $z \in \mathcal{Z}$ and $x \in \mathcal{X}$. Here z is a latent variable, x an observed variable, and θ a model parameter. The goal of maximum likelihood estimation (MLE) is to find a value of θ that maximizes the log marginal likelihood $\log p_\theta(x)$ for a given x. Achieving this goal is difficult because it involves computing the integral $p_\theta(x) = \int p_\theta(x, z)\,\mathrm{d}z$ which is intractable except for very simple models.

Instead of the intractable integral, the variational approach optimizes a lower bound for $\log p_\theta(x)$, typically using an approximate posterior $q_\phi(z \mid x)$. Two well-known lower bounds are ELBO (Jordan et al., 1999) and IWAE with N (independent) particles (Burda et al., 2016; Domke & Sheldon, 2018):

$$\mathcal{L}_{\mathrm{ELBO}}(\theta, \phi; x) = \mathbb{E}\left[\log \frac{p_\theta(x, z)}{q_\phi(z \mid x)}\right], \tag{1}$$

$$\mathcal{L}_{\mathrm{IWAE}}(\theta, \phi; x) = \mathbb{E}\left[\log\left\{\frac{1}{N}\sum_{i=1}^{N} \frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(i)} \mid x)}\right\}\right], \tag{2}$$

where the expectations are taken with respect to $q_\phi(z \mid x)$ and $q_\phi(z^{(1:N)} \mid x) = \prod_{i=1}^{N} q_\phi(z^{(i)} \mid x)$, respectively. The fact that $\mathcal{L}_{\mathrm{ELBO}}$ and $\mathcal{L}_{\mathrm{IWAE}}$ are indeed lower bounds of $\log p_\theta(x)$ follows from Jensen's inequality. Note that $\mathcal{L}_{\mathrm{IWAE}}$ has the parameter N for controlling the number of particles inside the log. For $N = 1$, $\mathcal{L}_{\mathrm{IWAE}}$ coincides with $\mathcal{L}_{\mathrm{ELBO}}$, but as N increases, it becomes tighter, eventually converging to $\log p_\theta(x)$. In the paper, we focus on $\mathcal{L}_{\mathrm{IWAE}}$.

In the variational approach, we typically optimize a target variational lower bound by stochastic gradient descent (SGD). The key step in this optimization is to estimate the gradient of the lower bound using samples. Two widely-used unbiased estimators are *score-function estimator*, also called REINFORCE (Williams, 1992) and likelihood ratio estimator (Glynn, 1990), and *reparameterization estimator*, also known as pathwise estimator (Kingma & Welling, 2014). The score-function estimator does not require differentiable densities and can be used for continuous as well as discrete latent variable models, but it is known to have high variance. The reparameterization estimator is, on the other hand, applicable only to differentiable models, but it typically has much lower variance than the score-function estimator. Our goal is to fix this high-variance problem in the score-function estimator, and develop a new gradient estimator that keeps the wide applicability of the score-function estimator but does not suffer from the high-variance issue.

We start with the score-function estimator derived from the gradient of the IWAE objective:

$$\begin{aligned}
&\nabla_\phi \mathbb{E}\left[\log\left\{\frac{1}{N}\sum_{i=1}^{N} \frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(i)} \mid x)}\right\}\right] \\
&= \mathbb{E}\left[\nabla_\phi \log\left\{\frac{1}{N}\sum_{i=1}^{N} \frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(i)} \mid x)}\right\}\right] \\
&\quad + \mathbb{E}\left[\log\left\{\frac{1}{N}\sum_{i=1}^{N} \frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(i)}|x)}\right\} \nabla_\phi \log q_\phi(z^{(1:N)} \mid x)\right],
\end{aligned}$$

where all the expectations are taken with respect to the distribution $q_\phi(z^{(1:N)} \mid x)$. The score-function estimator approximates the two expectation terms from above using samples. Specifically, it computes the following sum of two sample-based estimates:

$$g_{\mathrm{low}} + g_{\mathrm{high}}, \tag{3}$$

where

$$g_{\mathrm{low}} = \nabla_\phi \log\left\{\frac{1}{N}\sum_{i=1}^{N} \frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(i)}|x)}\right\},$$

$$g_{\mathrm{high}} = \log\left\{\frac{1}{N}\sum_{i=1}^{N} \frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(i)}|x)}\right\} \nabla_\phi \log q_\phi(z^{(1:N)}|x),$$

and each $z^{(i)}$ is sampled from $q_\phi(z^{(i)}|x)$. In most cases, the estimate g_{high} has significantly higher variance than g_{low}, and it is the root cause of the high-variance issue for the score-function estimator. (More detailed discussion about score estimator is provided in Appendix B). Throughout the paper, we will delve into finding an alternative to g_{high} with significantly lower variance.

2.2. State-space model

The state-space model is a standard tool for modeling the dynamics behind sequential data. In the state-space model,

both the latent variable z and the observed x have the sequence form: $z = (z_1, \ldots, z_T)$ and $x = (x_1, \ldots, x_T)$ where T is the length of the sequence. The joint density of x and z is represented by an initial density $p_\theta(x_1, z_1)$ and a transition density $p_\theta(x_{t+1}, z_{t+1} \mid x_1, \ldots, x_t, z_1, \ldots, z_t)$ as follows:

$$p_\theta(x, z) = p_\theta(x_1, z_1) \prod_{t=2}^{T} p_\theta(x_t, z_t | x_{1:t-1}, z_{1:t-1}). \quad (4)$$

Here we use the subscript notation $x_{i:j}$ to denote the subsequence $(x_i, x_{i+1}, \ldots, x_j)$. Note that this setup permits time-dependent transition densities on x_t and z_t.

When variational inference is applied to a state-space model using an approximate posterior $q_\phi(z \mid x)$ of the form

$$q_\phi(z \mid x) = q_\phi(z_1 \mid x) \prod_{t=2}^{T} q_\phi(z_t \mid z_{1:t-1}, x), \quad (5)$$

the IWAE objective in (2) and the g_{high} in (3) yield particularly convenient formulas. To see this, let $w_t^{(i)}$ and $w_{t:T}^{(i)}$ be the following importance weights:

$$w_0^{(i)} = 1, \qquad w_t^{(i)} = \frac{p_\theta(x_t, z_t^{(i)} \mid x_{1:t-1}, z_{1:t-1}^{(i)})}{q_\phi(z_t^{(i)} \mid z_{1:t-1}^{(i)}, x)},$$

$$w_{t:T}^{(i)} = \prod_{t'=t}^{T} w_{t'}^{(i)}, \quad w^{(i)} = w_{1:T}^{(i)}$$

Then, the IWAE in (2) can be expressed as follows:

$$\mathcal{L}_{\text{IWAE}}(\theta, \phi; x) = \mathbb{E}\left[\log \left\{ \frac{1}{N} \sum_{i=1}^{N} w^{(i)} \right\} \right]$$

where the expectation is taken with respect to the product of approximate posteriors:

$$\prod_{i=1}^{N} q_\phi(z^{(i)} \mid x) = \prod_{i=1}^{N} \left\{ q_\phi(z_1^{(i)} \mid x) \prod_{t=2}^{T} q_\phi(z_t^{(i)} \mid z_{1:t-1}^{(i)}, x) \right\}.$$

Also, the high-variance term g_{high} in (3) can be written in the factored form below:

$$g_{\text{high}} = \sum_{t=1}^{T} \sum_{i=1}^{N} \left(\log \left(\frac{1}{N} \sum_{j=1}^{N} w^{(j)} \right) \times \nabla_\phi \log q_\phi(z_t^{(i)} \mid z_{1:t-1}^{(i)}, x) \right). \quad (6)$$

We will assume these structured formulas for the rest of the paper.

3. Variational Inference using Future Likelihood Estimates

In this section, we describe our gradient estimator, which approximates future likelihoods at all time steps and uses this approximation to replace g_{high} by a low-variance alternative.

3.1. Future likelihood function

We first define a *future likelihood function*, which plays an important role in our new gradient estimator.

Definition 1 (Future likelihood function)**.** *For given parameters θ and ϕ, we define the **future likelihood function** $\Gamma_{\theta,\phi}(z_{1:t}, x)$ **at step** t to be the following function of latent variables $z_{1:t}$ and observation $x = x_{1:T}$:*

$$\Gamma_{\theta,\phi}(z_{1:t}, x) = \mathbb{E}\left[w_{t+1:T} \right] = \mathbb{E}\left[\prod_{t'=t+1}^{T} w_{t'} \right],$$

where the expectations are taken with respect to

$$q_\phi(z_{t+1:T} \mid z_{1:t}, x) = \prod_{t'=t+1}^{T} q_\phi(z_{t'} \mid z_{1:t'-1}, x).$$

Note that $\Gamma_{\theta,\phi}(z_{1:t}, x)$ is an importance-sampling formulation of the future marginal likelihood $p_\theta(x_{t+1:T} \mid x_{1:t}, z_{1:t})$. Thus, it coincides with the future marginal likelihood, under the mild condition that the support of $p_\theta(z_{t+1:T} \mid x_{1:t}, z_{1:t})$ should be covered by that of $q_\phi(z_{t+1:T} \mid z_{1:t}, x)$. In the remainder of the paper, we will denote $\Gamma_{\theta,\phi}(z_{1:t}, x)$ and $\Gamma_{\theta,\phi}(z_{1:t}^{(i)}, x)$ simply by Γ_t and $\Gamma_t^{(i)}$, respectively, whenever the relevant latent variables and observation are clear from the context.

An important part of our gradient estimator is to utilize an approximator of the future likelihood functions at all time steps. Note that the definition of the future likelihood function immediately gives rise to a Monte-Carlo approximation scheme, which computes the average of the term inside the expectation of Γ_t using independent samples from $q_\phi(z_{t+1:T} \mid z_{1:t}, x)$. However, this scheme is not practical because of high variance. Thus, we instead approximate the future likelihood function by making use of the recursive nature of the state-space model, that is, the following recurrent relation satisfied by the future likelihood functions at consecutive time steps:

$$\Gamma_{t-1} = \mathbb{E}_{z_t \sim q_\phi(z_t \mid z_{1:t-1}, x)} \left[w_t \Gamma_t \right]. \quad (7)$$

Concretely, we fix a parameterized function $\hat{\Gamma}_{\psi,t}$ to approximate Γ_t for every t. Usually, $\hat{\Gamma}_{\psi,t}$ is defined in terms of a neural network parameterized by ψ. Then, we learn the value of the parameter ψ by optimizing the following

objective which is derived from the recurrent relation (7), analogous to temporal difference learning in RL (Sutton et al., 1998):

$$\min_{\psi} \mathbb{E}\left[\sum_{t=2}^{T}\left(\hat{\Gamma}_{\psi,t-1}(z_{1:t-1},x) - \mathbb{E}\left[w_t\hat{\Gamma}_{\psi,t}(z_{1:t},x)\right]\right)^2\right]$$

where the outer and inner expectations are taken respectively by $q_\phi(z_{1:T}\,|\,x)$ and $q_\phi(z_t\,|\,z_{1:t-1},x)$. The optimization is done by stochastic gradient descent: From N samples $z_{1:T}^{(1:N)}$ of $q_\phi(z_{1:T}\,|\,x)$, update ψ via:

$$\psi \leftarrow \psi - \eta\cdot\frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T}\Bigg(\nabla_\psi\hat{\Gamma}_\psi(z_{1:t}^{(i)},x) \times\left(\hat{\Gamma}_{\psi,t-1}(z_{1:t-1}^{(i)},x) - w_t\cdot\hat{\Gamma}_{\psi,t}(z_{1:t}^{(i)},x)\right)\Bigg).$$

The future likelihood function is a good candidate for a *baseline*, which helps reduce the variance of a gradient estimator. To explain this, we recall a defining property of the baseline (Weber et al., 2019):

Lemma 1. *For every distribution $q(z)$ and quantity B that does not depend on z, we have*

$$\mathbb{E}_{z\sim q(z)}[B\nabla\log q(z)] = 0.$$

*We call B a **baseline** for $q(z)$.*

Since neither Γ_t nor its approximation $\hat{\Gamma}_{\psi,t}$ depends on $z_{t+1:T}$ (i.e. marginalized out by definition), they can serve as a baseline for the distribution $q_\phi(z_{t+1:T}\,|\,z_{1:t},x)$.

3.2. Gradient estimator

We are now ready to present our inference algorithm, which we name *VIFLE* (Variational Inference with Future Likelihood Estimates). At the core of the algorithm lies a highly-effective combination of the future likelihood function and the baseline, which replaces the high-variance term g_{high} in (3) by the low-variance gradient estimator g_{VIFLE}.

Concretely, we derive g_{VIFLE} from g_{high} in two steps. First, we take the characterization of g_{high} for the state-space model in (6), and form a baseline for $q(z_t^{(i)}\,|\,z_{1:t-1}^{(i)},x)$ for each particle $z^{(i)}$ and each time step t, defined as follows:

$$\log\frac{1}{N}\left(w_{1:t-1}^{(i)}\Gamma_{t-1}^{(i)} + \sum_{j\neq i}w^{(j)}\right). \tag{8}$$

This term is just the coefficient of the gradient term for the particle i and the time step t in (6) except that the $w^{(i)}$ part in the coefficient is replaced by $w_{1:t-1}^{(i)}\Gamma_{t-1}^{(i)}$. The replacement makes the term independent of $z_t^{(i)}$, unlike the original coefficient, so that the term becomes a baseline for $q_\phi(z_t^{(i)}\,|\,z_{1:t-1}^{(i)},x)$. We subtract this baseline from the coefficient of the gradient term in (6) for the particle i and the time step t.

Our baseline is chosen to correlate the coefficient of the gradient term $\nabla_\phi\log q_\phi(z_t^{(i)}\,|\,z_{1:t-1}^{(i)},x)$ in (6) well. We can see this by subtracting the baseline from the coefficient and simplifying the result slightly, as shown below:

$$\begin{aligned}
&\log\frac{1}{N}\sum_{j=1}^{N}w^{(j)} - \log\frac{1}{N}\left(w_{1:t-1}^{(i)}\Gamma_{t-1}^{(i)} + \sum_{j\neq i}w^{(j)}\right)\\
&\quad= \log\frac{\frac{1}{N}\sum_{j=1}^{N}w^{(j)}}{\frac{1}{N}\left(w_{1:t-1}^{(i)}\Gamma_{t-1}^{(i)} + \sum_{j\neq i}w^{(j)}\right)}\\
&\quad= \log\frac{w^{(i)} + \sum_{j\neq i}w^{(j)}}{w_{1:t-1}^{(i)}\Gamma_{t-1}^{(i)} + \sum_{j\neq i}w^{(j)}}
\end{aligned} \tag{9}$$

The only difference between the numerator and the denominator here is whether we use $w^{(i)}$ and $w_{1:t-1}^{(i)}\Gamma_{t-1}^{(i)}$. Note that these two terms are closely related: the latter is the expectation of the former over the distribution $q_\phi(z_{t:T}^{(i)}\,|\,z_{1:t-1}^{(i)},x)$. By sharing large parts of their definitions, these two options are highly correlated, sometimes being close to each other, which mean that the outcome of the subtraction in (9) has smaller variance than the original coefficient $\log(\sum_{j=1}^{N}w^{(j)}/N)$.

Before moving on to the second step of our derivation of g_{VIFLE}, we summarize the outcome of the first, which we denote by g_{VIFLE}^u:

$$g_{\text{VIFLE}}^u = \sum_{t=1}^{T}\sum_{i=1}^{N}\Bigg(\log\frac{w^{(i)} + \sum_{j\neq i}^{N}w^{(j)}}{w_{1:t-1}^{(i)}\Gamma_{t-1}^{(i)} + \sum_{j\neq i}w^{(j)}} \times\nabla_\phi\log q_\phi(z_t^{(i)}\,|\,z_{1:t-1}^{(i)},x)\Bigg). \tag{10}$$

We point out that g_{VIFLE}^u and g_{high} have the same expectation by Lemma 1, and that this relationship continues to hold even when we replace the future likelihood function Γ by its approximation $\hat{\Gamma}_\psi$ (The superscript u emphasizes that it is an unbiased estimator).

Second, we approximate $w^{(i)}$ in the numerator of (10) by $w_{1:t}^{(i)}\Gamma_t^{(i)}$, which gives the final formula g_{VIFLE} of our estimator:

$$g_{\text{VIFLE}} = \sum_{t=1}^{T}\sum_{i=1}^{N}\Bigg(\log\frac{w_{1:t}^{(i)}\Gamma_t^{(i)} + \sum_{j\neq i}w^{(j)}}{w_{1:t-1}^{(i)}\Gamma_{t-1}^{(i)} + \sum_{j\neq i}w^{(j)}} \times\nabla_\phi\log q_\phi(z_t^{(i)}\,|\,z_{1:t-1}^{(i)},x)\Bigg). \tag{11}$$

Note that both the numerator and the denominator in (11) are the expectations of the same formula $\sum_{j=1}^{N}w^{(j)}$ but

taken over different distributions. The former is over that of the random variables $z^{(i)}_{t+1:T}$ and the latter over that of $z^{(i)}_{t:T}$. As a result, they tend to be highly correlated, making g_{VIFLE} have low variance. Another reason for the reduction of variance is that g_{VIFLE} involves far fewer random variables than g^u_{VIFLE}. Replacing $w^{(i)}$ by $w^{(i)}_{1:t}\Gamma^{(i)}_t$ stops the coefficient for the i-th particle and the step t from depending on the sequence of samples $z^{(i)}_{t+1:T}$. In so doing, it at least partially addresses the issue of the accumulation of model or algorithm noise over time steps, which often makes learning and inference particularly difficult for the state-space models.

The second step makes our gradient estimator depart from the score-function estimator. The new g_{VIFLE} and the old g_{high} no longer have the same expectation. This means that as an estimator for the gradient of the IWAE objective, our estimator is biased. However, in Section 4, we will identify a new variational lower bound and show that our estimator is unbiased with respect to this bound.

3.3. Gradient estimator in a general form

The main recipe for deriving g_{VIFLE} in our estimator is to replace the $w^{(i)}$ term in the definition of g_{high} by its expectation over $z^{(i)}_{t+1:T}$ or $z^{(i)}_{t:T}$. Our derivation restricts the application of this recipe to the current particle i (i.e. particle for which we compute $\nabla_\phi \log q_\phi(z^{(i)}_t \mid z^{(i)}_{1:t-1}, x)$ in g_{high}), but this restriction can be lifted and we can derive a new gradient estimator in doing so. For instance, we may apply the recipe to every particle, and get the following alternative to g_{VIFLE} called *FR*, which stands for *Full Replacement*:

$$g_{\mathrm{FR}} = \sum_{t=1}^{T}\sum_{i=1}^{N}\left(\log \frac{\sum_{j=1}^{N} w^{(j)}_{1:t}\Gamma^{(j)}_t}{\sum_{j=1}^{N} w^{(j)}_{1:t-1}\Gamma^{(j)}_{t-1}} \times \nabla_\phi \log q_\phi(z^{(i)}_t \mid z^{(i)}_{1:t-1}, x)\right). \quad (12)$$

More generally, for every particle i, we can pick a set S_i of particles with $i \in S_i$. Let $\mathcal{S} = \{S_i \mid 1 \le i \le N\}$. Then, we can apply our recipe to all the particles in S_i when we consider the coefficient of the gradient term for the particle i in g_{high} (i.e., $\nabla_\phi \log q_\phi(z^{(i)}_t \mid z^{(i)}_{1:t-1}, x)$ in (6)). This gives the following new variant of g_{high}:

$$g_{\mathcal{S}} = \sum_{t=1}^{T}\sum_{i=1}^{N}\left(\log \frac{\sum_{j\in S_i} w^{(j)}_{1:t}\Gamma^{(j)}_t + \sum_{j\notin S_i} w^{(j)}}{\sum_{j\in S_i} w^{(j)}_{1:t-1}\Gamma^{(j)}_{t-1} + \sum_{j\notin S_i} w^{(j)}} \times \nabla_\phi \log q_\phi(z^{(i)}_t \mid z^{(i)}_{1:t-1}, x)\right). \quad (13)$$

As in the case of g_{VIFLE}, the gradient estimator with g_{FR} and $g_{\mathcal{S}}$ no longer compute unbiased estimates of the IWAE objective, because they approximate the $w^{(j)}$ term in the numerator by $w^{(j)}_{1:t}\Gamma^{(j)}_t$. In Section 4, we formally show that for all $\mathcal{S}$, the estimator $g_{\mathcal{S}}$ implicitly optimizes an alternative variational lower bound. We also establish the relationship between the IWAE objective and the lower bound for $g_{\mathcal{S}}$, and also between $g_{\mathcal{S}}$ and $g_{\mathcal{S}'}$ for different $\mathcal{S}$ and $\mathcal{S}'$. We leave as future work for the design of an algorithm that optimizes $\mathcal{S}$.

3.4. Relationship with the VIMCO estimator

We can relate our gradient estimator with g_{VIFLE} with VIMCO (Mnih & Rezende, 2016). Essentially, VIMCO is a score-function estimator with the following baseline for each particle distribution $q_\phi(z^{(i)}_t \mid z^{(i)}_{1:t-1}, x)$ in g_{high}:

$$\log\left(\widetilde{w}^{(i)} + \sum_{j\neq i} w^{(j)}\right) \text{ where } \widetilde{w}^{(i)} = \left(\prod_{j\neq i} w^{(j)}\right)^{\frac{1}{N-1}}.$$

Here, the term $\widetilde{w}^{(i)}$ is the geometric mean of the $w^{(j)}$ with $j \in \{0, 1, \ldots, N\} \setminus \{i\}$. The idea is to approximate the $w^{(i)}$ for the particle i by the geometric mean over other particles. Thus, VIMCO is the score-function estimator (3) with g_{high} replaced by:

$$g_{\mathrm{VIMCO}} = \sum_{t=1}^{T}\sum_{i=1}^{N}\left(\log \frac{\sum_{j=1}^{N} w^{(j)}}{\widetilde{w}^{(i)} + \sum_{j\neq i} w^{(j)}} \times \nabla_\phi \log q_\phi(z^{(i)}_t | z^{(i)}_{1:t-1}, x)\right). \quad (14)$$

Both VIMCO and our estimator share the idea of using information from other particles to form a baseline, but with following substantial differences: VIMCO does not approximate any terms in the numerator, and it remains unbiased with respect to the IWAE objective. On the other hand, our estimator does approximate a term there, and in so doing, it changes the target variational lower bound. In addition, our estimator utilizes the future likelihood function approximator to further reduce its variance.

4. Theoretical Analysis

We analyze theoretical properties of our gradient estimator with g_{VIFLE}, g_{FR}, and more generally $g_{\mathcal{S}}$ for a family of index sets $\mathcal{S}$. As we have stated earlier, none of these estimators computes an unbiased estimate of the IWAE objective, although they all originate from an unbiased estimator for the objective. Our first result says that in fact, these estimators target at different variational lower bounds, and for these new bounds, the estimators are unbiased.

Theorem 1. *Fix θ and x. Then, for every parameter $\phi = \phi_0$,*

there are variational objectives $\mathcal{L}_{\mathrm{VIFLE}}$ and $\mathcal{L}_{FR}$ such that

$$\mathcal{L}_{\mathrm{IWAE}}(\theta, \phi_0; x) \geq \mathcal{L}_{\mathrm{VIFLE}}(\theta, \phi_0; x) \geq \mathcal{L}_{FR}(\theta, \phi_0; x),$$
$$\nabla_\phi \mathcal{L}_{\mathrm{VIFLE}}(\theta, \phi; x)|_{\phi=\phi_0} = \mathbb{E}[g_{\mathrm{low}} + g_{\mathrm{VIFLE}}],$$
$$\nabla_\phi \mathcal{L}_{FR}(\theta, \phi; x)|_{\phi=\phi_0} = \mathbb{E}[g_{\mathrm{low}} + g_{\mathrm{FR}}].$$

The expectations in the second and third equations are taken with respect to the distribution $q_{\phi_0}(z^{(1:N)}|x)$, and the gradients inside g_{low}, g_{VIFLE} and g_{FR} are computed at the point $\phi = \phi_0$. More generally, for all families of index sets, $\mathcal{S}$ and $\mathcal{T}$, if $\mathcal{S}_i \subseteq \mathcal{T}_i$ for every i, then there are variational objectives $\mathcal{L}_{\mathcal{S}}$ and $\mathcal{L}_{\mathcal{T}}$ such that

$$\mathcal{L}_{\mathrm{IWAE}}(\theta, \phi_0; x) \geq \mathcal{L}_{\mathcal{S}}(\theta, \phi_0; x) \geq \mathcal{L}_{\mathcal{T}}(\theta, \phi_0; x).$$

The next theorem shows that the target lower bound $\mathcal{L}_{\mathrm{VIFLE}}$ of our estimator enjoys the same convergence property of the IWAE objective.

Theorem 2. *The bound $\mathcal{L}_{\mathrm{VIFLE}}$ converges to $\log p_\theta(x)$ almost surely, as the number of particles N goes to infinity.*

All of these results support the use of our gradient estimator in variational inference. The detailed proofs for the results can be found in the supplementary material due to the page limit.

5. Related Works

In the context of variational inference, there have been a number of prior works to estimate quantities involving future time steps in the state-space models and exploit these estimates for approximating gradients of variational objectives accurately (Weber et al., 2015; Levine, 2018; Grathwohl et al., 2018), sometimes using the connection between variational inference and reinforcement learning. However, these works are based on the standard ELBO objective, not the IWAE objective, and they compute the future ELBO, not the future likelihood, unlike our gradient estimator. If we naively adjusted those works to the IWAE objective, we would have to estimate the future IWAE objective, which is apparently a more difficult task than estimating the future likelihood because of the use of multiple particles in the IWAE.

The existing approaches for handling discrete latent variables in variational inference can be roughly classified into three categories. The first category of approaches use the score-function estimator with a carefully-designed variance reduction technique. Mnih & Gregor (2014) developed such techniques for the ELBO objective, and Mnih & Rezende (2016) developed VIMCO for the IWAE objective. The second class of approaches relax discrete variables to continuous ones, and approximate the gradient using the reparameterization estimator. A well-known relaxation for the categorical distribution is a technique called Gumbel-Softmax (Jang et al., 2017) and also Concrete distribution (Maddison et al., 2017b). The approaches in the last categories, such as REBAR (Tucker et al., 2017) and RELAX (Grathwohl et al., 2018), combine the techniques of the other two groups and use both the score-function estimator and the reparameterization estimator with relaxation, so as to remove the bias issue in the approaches in the second group. Our gradient estimator falls into the first group, and shows for the first time how to use future likelihood estimates to obtain a low-variance gradient estimator for an IWAE-like multi-particle variational objective.

The multi-particle variational objectives, such as IWAE, are usually tighter lower bounds for the marginal likelihood than the standard ELBO objective. It has been observed this tightening helps learning the generative model, but it has a side-effect of hindering learning the variational distribution (Rainforth et al., 2018). The current solution for this issue (Tucker et al., 2019) is based on the repeated application of the reparameterization idea, and is applicable only to models with differentiable densities. Since our gradient estimator is designed for a multi-particle objective and is not limited to models with differentiable densities, an interesting future direction is to investigate whether and how it can be used for that line of research.

Some previous studies (Maddison et al., 2017a; Naesseth et al., 2018; Le et al., 2018; Lawson et al., 2018) proposed tighter variational bounds for the state-space models, by exploiting sequential Monte Carlo (SMC) and its ability for estimating the marginal likelihood without any bias. However, the gradient estimator for the bound suffers from the high-variance issue, and these studies simply drop the high-variance term from the estimator. Our idea of using future likelihood estimates and applying approximation only to some particles may lead to a new approach for addressing this issue in these studies. We also want to point out several studies about reducing the variance of the gradient estimator for the state-space models (Weber et al., 2015; Ahmed et al., 2019). As explained earlier, these estimators target at the gradient of the ELBO, not the IWAE, and they do not use the notion of future likelihood estimates.

6. Experiments

6.1. Synthetic datasets from two dynamical systems

Our first set of experiments uses synthetic datasets from two dynamical systems, and has the goal of computing posterior distributions.

The first dynamical system models a simple continuous linear dynamics with Gaussian noises, and it is defined as

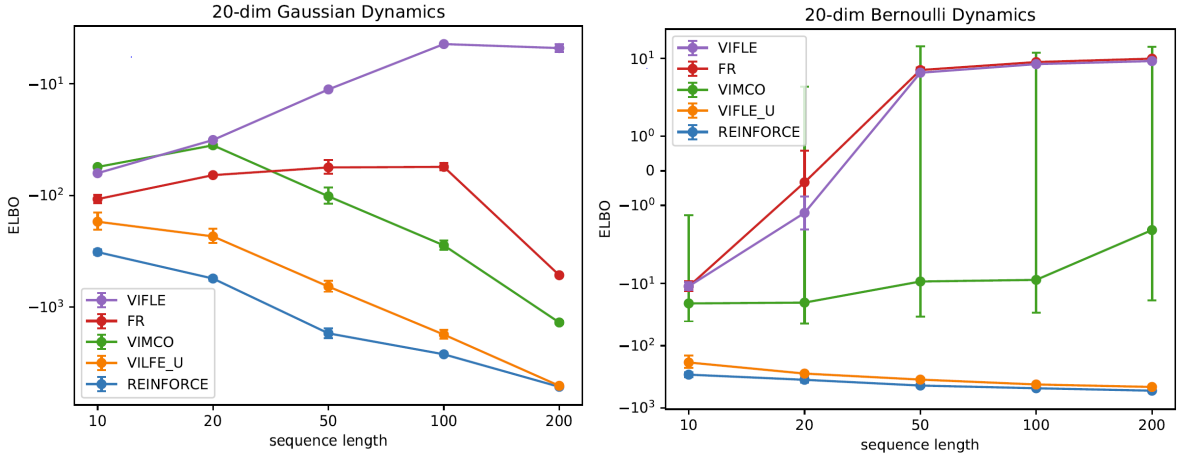

Figure 1. In this figure, the x-axis indicates the length of observed sequences in each dataset, and the y-axis indicates in log scale the ELBO per time step to measure the trends of negative KL divergence between proposal distribution and true posterior distribution. In Gaussian dynamics, each error bar represents the sample standard error with 5 random samples. In Bernoulli dynamics, each error bar represents the sample standard error with 10 random samples.

follows:

$$z_0 = v_0, \qquad z_t = Az_{t-1} + v_t, \qquad x_t = Bz_t + w_t,$$

where variables z_t, x_t, v_t, and w_t are 20-dimensional vectors, A and B are randomly-generated 20×20 matrices with determinant 1, and v_t and w_t denote Gaussian noises with covariance $0.1I$. We use an approximate posterior distribution of the form $q_\phi(z_t \mid z_{t-1}, x_t)$, where q_ϕ is a Gaussian distribution.

The second dynamical system models a dynamics on a discrete latent state space and continuous observations, given by:

$$z_1 \sim \mathrm{Bern}(0.5),\ z_t = F(z_{t-1}),\ x_t = Az_t + \sin(10z_t) + w_t.$$

Here variables z_t, x_t, and w_t are 20-dimensional vectors, A is a randomly-generated 20×20 matrix with determinant 1, and F is a transition function on the boolean vectors that flips each component independently with probability 0.1. Also, w_t denotes a Gaussian noise with covariance $0.1I$. For this model, we use an approximate posterior distribution of the form $q_\phi(z_t \mid z_{t-1}, x_t)$ where q_ϕ is a Bernoulli distribution. The detailed settings for the two dynamical systems are provided in Appendix H.

We performed posterior inference on both dynamical systems, each on finite different datasets with different maximum time steps, 10, 20, 50, 100, and 200. In all of these cases, we tried five variational inference algorithms that optimize multi-particle variational objectives, such as IWAE. The tried algorithms are (a) REINFORCE that estimates the gradient by the score-function estimator $g_{\text{low}} + g_{\text{high}}$; (b) VIMCO based on $g_{\text{low}} + g_{\text{VIMCO}}$; (c) VIFLE_U that uses the gradient estimator $g_{\text{low}} + g^u_{\text{VIFLE}}$; (d) VIFLE based on $g_{\text{low}} + g_{\text{VIFLE}}$; (f) FR with the gradient estimator $g_{\text{low}} + g_{\text{FR}}$. In our experiments, we trained the parameters ϕ of approximate posterior distributions q_ϕ using these algorithms, and measured the qualities of the trained q_ϕ's by computing the ELBO which is equivalent to negative KL divergence between true posterior $p(z|x)$ and $q_\phi(z|x)$ plus log evidence, which is a constant in this case.

The results of our experiments are given in Figure 1. They show a general trend that as the length of data increases, the performance of VIFLE and FR improves, while other three algorithms deteriorates. The only exception to this trend is the 200-step case for FR in the continuous dynamical system. One possible explanation for the behaviour of FR is that it relies too much on the future likelihood approximator, and its error is manifested in particularly long sequences. Otherwise, the use of approximate future likelihood functions generally helped reduce the variance of the gradient estimator. Furthermore, we can see that VIFLE and FR generally performs better as we increase the length of data, which is actually due to the increase in the amount of training data for training the future likelihood functions. We can also notice the unstable performance of VIMCO in the discrete dynamical system, which was highly dependent on how the initial samples were generated.

VIFLE uses the approximate future likelihood function in two places, one for a baseline and the other for replacing a part of the coefficient of a particle-specific gradient term. Our experimental results show the impact of each of these two uses. Comparing the results for REINFORCE, VIMCO, and VIFLE_U reveals that using future likelihood estimates as baselines help, but it is not as effective as using the baselines of VIMCO. We hypothesize that this is due to

N	Training algorithm	JSB	Nottingham	MuseData	Piano-midi.de
4	REPARAM (Kingma & Welling, 2014)	-7.47	-3.59	-6.47	-8.47
	VIMCO (Mnih & Rezende, 2016)	-7.22	**-2.92**	-6.34	-8.42
	VIFLE	**-7.11**	-3.03	**-5.94**	**-8.06**
8	REPARAM (Kingma & Welling, 2014)	-7.49	-3.84	-6.48	-8.49
	VIMCO (Mnih & Rezende, 2016)	-7.25	**-2.84**	-5.95	-8.30
	VIFLE	**-7.23**	-3.01	**-5.92**	**-8.02**

Table 1. Test-set marginal log-likelihood bounds for the trained models with continuous latent variables by three algorithms. For each algorithm, we pick a model by its validation performance, and report the $\mathcal{L}_{\mathrm{IWAE}}^{(128)}$ for the model.

N	Training algorithm	JSB	Nottingham	MuseData	Piano-midi.de
4	CONCRETE (Maddison et al., 2017b)	-8.58	-3.55	-6.47	-8.26
	VIMCO (Mnih & Rezende, 2016)	-8.54	-3.66	-7.18	-8.44
	VIFLE	**-8.40**	**-3.16**	**-6.10**	**-8.17**
8	CONCRETE (Maddison et al., 2017b)	-8.71	-3.83	-6.41	-8.58
	VIMCO (Mnih & Rezende, 2016)	-8.56	-3.92	-6.42	-8.36
	VIFLE	**-8.53**	**-3.15**	**-6.13**	**-8.29**

Table 2. Test-set marginal log-likelihood bounds for the trained models with discrete latent variables by three algorithms. For each algorithm, we pick a model by its validation performance, and report the $\mathcal{L}_{\mathrm{IWAE}}^{(128)}$ for the model.

the fact that the baselines for VIMCO do not require any further estimation, whereas those for VIFLE_U do require the estimation of future likelihoods. The benefit of the other use of future likelihood estimates is more prominent in our experimental results. In Figure 1, the graphs for VIFLE clearly dominate those for VIFLE_U and VIMCO in both dynamical systems, and the domination gets more prominent as the length of data grows. This means that for these two dynamical systems, dramatic performance gain can be attained by trading off the tightness of a variational bound over the reduction of variance in gradient estimation.

6.2. Polyphonic music datasets

We further conducted experiments on real-world datasets using four polyphonic music datasets (Boulanger-lewandowski et al., 2012). Here, the goal is to learn the generative model parameter p_θ as well as the approximate posterior parameter q_ϕ as in standard variational inference. We considered two types of models, one with a continuous latent state space and the other one with a discrete latent state space. The model parameters θ are learned by the standard stochastic gradient ascent with respect to the IWAE objective. On the other hand, the parameters ϕ of approximate distributions are found by different variational inference algorithms. The model distributions are factorized as in (4), and the approximate posteriors assume the form $q_\phi(z_t \mid z_{1:t-1}, x) = q_\phi(z_t \mid z_{1:t-1}, x_{1:t})$. We used four and eight particles for each algorithm. As for the learning rate, we report the best results among the choices in $\{3 \times 10^{-4}, 1 \times 10^{-4}, 3 \times 10^{-5}, 1 \times 10^{-5}\}$. For more details on the experimental setting, we refer the readers to Appendix H.

Table 1 shows the results on the model with continuous state variables, and Table 2 the results on the model with discrete state variables. VIFLE shows the state-of-the-art performance except for the continuous model on the Nottingham dataset. This result suggests that our algorithm can be used to solve real-world problems.

7. Conclusion

We have presented a new variational inference algorithm for sequential data, which is demonstrated to be highly effective even when the models have non-differentiable densities, for instance, due to the use of discrete latent variables. Our algorithm optimizes a multi-particle variational lower bound of the marginal likelihood, and uses a gradient estimator for this bound that achieves a low variance by using an approximator of the future likelihoods at all time steps. Our gradient estimator originates from the score-function estimator for the IWAE objective, but it computes an approximation of a new multi-particle lower bound, as shown in our theoretical analysis of the estimator. Our experiments with synthetic and real datasets show that our algorithm is highly effective for models with discrete latent variables and also for real-world data.

As for the future work, we hope to provide a more principled approach to the trade off between variance and tightness, based on our generalized form presented in Section 3.3. Although we mostly focused on g_{VIFLE} and g_{FR}, finding a good index set $\mathcal{S}$ may further improve the performance of the posterior inference. Another promising approach is to

generalize VIFLE. We only focused on the IWAE objective, but VIFLE can be extended to other objectives with multiple particles such as *Filtering Variational Objective (FIVO)* (Maddison et al., 2017a). The gradient estimator of FIVO is known to have a high variance term from resampling, and VIFLE may provide an approach to address this issue.

Acknowledgements

This work was supported by the National Research Foundation (NRF) of Korea (NRF-2019R1A2C1087634 and NRF-2019M3F2A1072238), the Ministry of Science and Information communication Technology (MSIT) of Korea (IITP No. 2020-0-00940, IITP 2019-0-00075 and IITP No. 2017-0-01779 XAI), and POSCO. Yang was supported by the Engineering Research Center Program through the National Research Foundation of Korea (NRF) funded by the Korean Government MSIT (NRF-2018R1A5A1059921), and also by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT (2017M3C4A7068177).

References

Ahmed, Z., Karuvally, A., Precup, D., and Gravel, S. Learning proposals for sequential importance samplers using reinforced variational inference. In *International Conference on Learning Representations Workshops*, 2019.

Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R. H., and Levine, S. Stochastic variational video prediction. In *International Conference on Learning Representations*, 2017.

Boulanger-lewandowski, N., Bengio, Y., and Vincent, P. Modeling temporal dependencies in highdimensional sequences: Application to polyphonic music generation and transcription. In *In ICML 29*. Citeseer, 2012.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016.

Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., and Abbeel, P. Variational lossy autoencoder. In *International Conference on Learning Representations*, 2016.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., and Bengio, Y. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pp. 2980–2988, 2015.

Denton, E. and Fergus, R. Stochastic video generation with a learned prior. In *International Conference on Machine Learning*, pp. 1182–1191, 2018.

Domke, J. and Sheldon, D. R. Importance weighting and variational inference. In *Advances in neural information processing systems*, pp. 4470–4479, 2018.

Fraccaro, M., Sønderby, S. K., Paquet, U., and Winther, O. Sequential neural models with stochastic layers. In *Advances in neural information processing systems*, pp. 2199–2207, 2016.

Glynn, P. W. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33 (10):75–84, 1990.

Grathwohl, W., Choi, D., Wu, Y., Roeder, G., and Duvenaud, D. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. In *ICLR*, 2018.

Gulrajani, I., Kumar, K., Ahmed, F., Taiga, A. A., Visin, F., Vazquez, D., and Courville, A. Pixelvae: A latent variable model for natural images. In *International Conference on Learning Representations*, 2016.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Lawson, D., Tucker, G., Naesseth, C. A., Maddison, C. J., Adams, R. P., and Teh, Y. W. Twisted variational sequential monte carlo. In *Third workshop on Bayesian Deep Learning (NeurIPS)*, 2018.

Le, T. A., Igl, M., Rainforth, T., Jin, T., and Wood, F. Auto-encoding sequential monte carlo. In *International Conference on Learning Representations*, 2018.

Levine, S. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Maddison, C. J., Lawson, J., Tucker, G., Heess, N., Norouzi, M., Mnih, A., Doucet, A., and Teh, Y. Filtering variational objectives. In *Advances in Neural Information Processing Systems*, pp. 6576–6586, 2017a.

Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR*, 2017b.

Masrani, V., Le, T. A., and Wood, F. The thermodynamic variational objective. In *Advances in Neural Information Processing Systems*, pp. 11521–11530, 2019.

Mnih, A. and Gregor, K. Neural variational inference and learning in belief networks. In *Proceedings of The 31st International Conference on Machine Learning*, pp. 1791–1799, 2014.

Mnih, A. and Rezende, D. J. Variational inference for monte carlo objectives. In *Proceedings of The 33rd5 International Conference on Machine Learning*, pp. 2188–2196, 2016.

Naesseth, C., Linderman, S., Ranganath, R., and Blei, D. Variational sequential monte carlo. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pp. 968–977, 2018.

Rainforth, T., Kosiorek, A., Le, T. A., Maddison, C., Igl, M., Wood, F., and Teh, Y. W. Tighter variational bounds are not necessarily better. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4277–4285. PMLR, 2018.

Ranganath, R., Gerrish, S., and Blei, D. M. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 814–822, 2014.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pp. 1530–1538. JMLR. org, 2015.

Sutton, R. S., Barto, A. G., et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

Tucker, G., Mnih, A., Maddison, C. J., Lawson, J., and Sohl-Dickstein, J. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. In *Advances in Neural Information Processing Systems*, pp. 2627–2636, 2017.

Tucker, G., Lawson, D., Gu, S., and Maddison, C. J. Doubly reparameterized gradient estimators for monte carlo objectives. In *International Conference on Learning Representations*, 2019.

Weber, T., Heess, N., Eslami, A., Schulman, J., Wingate, D., and Silver, D. Reinforced variational inference. In *Advances in Neural Information Processing Systems (NIPS) Workshops*, 2015.

Weber, T., Heess, N., Buesing, L., and Silver, D. Credit assignment techniques in stochastic computation graphs. *arXiv preprint arXiv:1901.01761*, 2019.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

A. Cumulative Noise Example

In this section, we show the example of cumulative noise in sequential structure. Suppose that the true posterior is given by $p(z|x) = \prod_{t=1}^{T} \mathcal{N}(z_t|0, 1^2)$ and it is approximated by $q(z|x) = \prod_{t=1}^{T} \mathcal{N}(z_t|0, 2^2)$. Then,

$$
\begin{aligned}
\mathrm{Var}\left(\frac{\frac{1}{N}\sum_{i=1}^{N} w^{(i)}}{p(x)}\right) &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{p(z^{(i)}|x)}{q(z^{(i)}|x)}\right) \\
&= \mathbb{E}_{z^{(1:N)}\sim q(\_|x)}\left[\left\{\frac{1}{N}\sum_{i=1}^{N}\frac{p(z^{(i)}|x)}{q(z^{(i)}|x)}\right\}^2\right] - \mathbb{E}_{z^{(1:N)}\sim q(\_|x)}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{p(z^{(i)}|x)}{q(z^{(i)}|x)}\right]^2 \\
&= \mathbb{E}_{z^{(1:N)}\sim q(\_|x)}\left[\left\{\frac{1}{N}\sum_{i=1}^{N}\frac{p(z^{(i)}|x)}{q(z^{(i)}|x)}\right\}^2\right] - 1 \\
&= \mathbb{E}_{z^{(1:N)}\sim q(\_|x)}\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{p(z^{(i)}|x)}{q(z^{(i)}|x)}\frac{p(z^{(j)}|x)}{q(z^{(j)}|x)}\right] - 1 \\
&= \mathbb{E}_{z^{(1:N)}\sim q(\_|x)}\left[\frac{1}{N^2}\sum_{i=1}^{N}\left\{\frac{p(z^{(i)}|x)}{q(z^{(i)}|x)}\right\}^2 + \frac{1}{N^2}\sum_{i\neq j}\frac{p(z^{(i)}|x)}{q(z^{(i)}|x)}\frac{p(z^{(j)}|x)}{q(z^{(j)}|x)}\right] - 1 \\
&= \mathbb{E}_{z^{(1:N)}\sim q(\_|x)}\left[\frac{1}{N^2}\sum_{i=1}^{N}\left\{\frac{p(z^{(i)}|x)}{q(z^{(i)}|x)}\right\}^2\right] + \frac{N^2-N}{N^2} - 1 \\
&= \frac{1}{N}\mathbb{E}_{z\sim q(\_|x)}\left[\left\{\frac{p(z|x)}{q(z|x)}\right\}^2\right] - \frac{1}{N} \\
&= \frac{1}{N}\left[\int_z \frac{p(z|x)^2}{q(z|x)}\mathrm{d}z - 1\right] \\
&= \frac{1}{N}\left[\int_{z_{1:T}}\prod_{t=1}^{T}\sqrt{\frac{4}{3}}\left(\frac{1}{\sqrt{2\pi(2/3)}}\exp\left(-\frac{3}{4}z_t^2\right)\right)\mathrm{d}z_{1:T} - 1\right] \\
&= \frac{1}{N}\left[\left(\frac{4}{3}\right)^{T/2} - 1\right],
\end{aligned}
$$

where $\frac{1}{N}\sum_i w^{(i)}$ is an unbiased estimator of $p(x)$. (i.e. $\mathbb{E}_q[\frac{1}{N}\sum_i w^{(i)}] = p(x)$) This example shows that the variance can be exponentially increased over time T.

B. Details for Score Estimator

First we provide the detailed derivation of (3):

$$
\begin{aligned}
&\nabla_\phi \mathbb{E}_{z^{(1:N)}\sim q_\phi(\_|x)}\left[\log\left\{\frac{1}{N}\sum_{i=1}^N w^{(i)}\right\}\right] \\
&= \nabla_\phi \int_{z^{(1:N)}} q_\phi(z^{(1:N)}|x)\log\left\{\frac{1}{N}\sum_{i=1}^N w^{(i)}\right\}\mathrm{d}z^{(1:N)} \\
&= \int_{z^{(1:N)}} \log\left\{\frac{1}{N}\sum_{i=1}^N w^{(i)}\right\}\nabla_\phi q_\phi(z^{(1:N)}|x) + q_\phi(z^{(1:N)}|x)\nabla_\phi \log\left\{\frac{1}{N}\sum_{i=1}^N w^{(i)}\right\}\mathrm{d}z^{(1:N)} \\
&= \int_{z^{(1:N)}} q_\phi(z^{(1:N)}|x)\log\left\{\frac{1}{N}\sum_{i=1}^N w^{(i)}\right\}\nabla_\phi \log q_\phi(z^{(1:N)}|x) + q_\phi(z^{(1:N)}|x)\nabla_\phi \log\left\{\frac{1}{N}\sum_{i=1}^N w^{(i)}\right\}\mathrm{d}z^{(1:N)} \\
&= \int_{z^{(1:N)}} q_\phi(z^{(1:N)}|x)\left[\log\left\{\frac{1}{N}\sum_{i=1}^N w^{(i)}\right\}\nabla_\phi \log q_\phi(z^{(1:N)}|x) + \nabla_\phi \log\left\{\frac{1}{N}\sum_{i=1}^N w^{(i)}\right\}\right]\mathrm{d}z^{(1:N)} \\
&= \mathbb{E}_{z^{(1:N)}\sim q_\phi(\_|x)}\left[\log\left\{\frac{1}{N}\sum_{i=1}^N w^{(i)}\right\}\nabla_\phi \log q_\phi(z^{(1:N)}|x) + \nabla_\phi \log\left\{\frac{1}{N}\sum_{i=1}^N w^{(i)}\right\}\right] \\
&= \mathbb{E}_{z^{(1:N)}\sim q_\phi(\_|x)}[g_{\mathrm{high}} + g_{\mathrm{low}}].
\end{aligned}
$$

In this equation, the fourth line is obtained by log derivative trick. Now, we rearrange g_{low} as:

$$
\begin{aligned}
g_{\mathrm{low}} &= \nabla_\phi \log\left\{\frac{1}{N}\sum_{i=1}^N w^{(i)}\right\} \\
&= \frac{\frac{1}{N}\sum_{i=1}^N \nabla_\phi w^{(i)}}{\frac{1}{N}\sum_{j=1}^N w^{(j)}} \\
&= \frac{\sum_{i=1}^N w^{(i)}\nabla_\phi \log w^{(i)}}{\sum_{j=1}^N w^{(j)}} \\
&= \sum_{i=1}^N \frac{w^{(i)}}{\sum_{j=1}^N w^{(j)}}\nabla_\phi \log w^{(i)} \\
&= -\sum_{i=1}^N \frac{w^{(i)}}{\sum_{j=1}^N w^{(j)}}\nabla_\phi \log q_\phi(z^{(i)}|x).
\end{aligned}
$$

In this equation, the weight of each gradient $\nabla_\phi \log q_\phi(z^{(i)}|x)$

$$
\frac{w^{(i)}}{\sum_{j=1}^N w^{(j)}},
$$

is bounded in $(0, 1)$ and sum of these weights are just 1. However, g_{high} has following formula

$$
g_{\mathrm{high}} = \log\left\{\frac{1}{N}\sum_{i=1}^N w^{(i)}\right\}\nabla_\phi \log q_\phi(z^{(1:N)}|x).
$$

Here, the weight of each gradient $\nabla_\phi \log q_\phi(z^{(i)}|x)$

$$
\log\left\{\frac{1}{N}\sum_{i=1}^N w^{(i)}\right\}
$$

has large magnitude, which causes large variance.

C. Proof of Lemma 1

Proofs for Lemma 1. Firstly, we can easily derive the following equation:

$$\begin{aligned}\mathbb{E}_{z\sim q(z)}[\nabla \log q(z)] &= \int_z q(z)\nabla \log q(z) dz \\ &= \int_z \nabla q(z) dz \\ &= \nabla \int_z q(z) dz \\ &= \nabla 1 = 0.\end{aligned}$$

More generally, for any B which does not depend on z,

$$\begin{aligned}\mathbb{E}_{z\sim q(z)}[B\nabla \log q(z)] &= \int_z q(z)B\nabla \log q(z) dz \\ &= B\int_z \nabla q(z) dz \\ &= B\nabla \int_z q(z) dz \\ &= B\nabla 1 = 0.\end{aligned}$$

□

D. Proof of Theorem 1

$$\widetilde{w}_t^{(i)} = \frac{p(x_t, z_t^{(i)})}{q_{\tilde{\phi}}(z_t^{(i)}|x)}, \quad \widetilde{w}^{(i)} = \widetilde{w}_{1:T}^{(i)} = \prod_{t=1}^{T} \widetilde{w}_t^{(i)}$$
$$\widetilde{\Gamma}_t^{(i)} = \mathbb{E}_{z^{(i)}\sim q_{\tilde{\phi}}(\cdot|x)}[\widetilde{w}_{t+1:T}^{(i)}],$$

$$\mathcal{L}_{\mathrm{VIFLE}}^{(N)}(\theta,\phi,\tilde{\phi};x) = \mathbb{E}_{z^{(1:N)}\sim q_{\tilde{\phi}}(\cdot|x)}\left[\log\left(\frac{1}{N}\sum_{i=1}^{N} w^{(i)}\right) + \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{q_\phi(z_t^{(i)}|x)}{q_{\tilde{\phi}}(z_t^{(i)}|x)}\log\frac{\widetilde{w}_{1:t}^{(i)}\widetilde{\Gamma}_t^{(i)} + \sum_{j\neq i}\widetilde{w}^{(j)}}{\widetilde{w}_{1:t-1}^{(i)}\widetilde{\Gamma}_{t-1}^{(i)} + \sum_{j\neq i}\widetilde{w}^{(j)}}\right],$$

$$\mathcal{L}_{\mathrm{FR}}^{(N)}(\theta,\phi,\tilde{\phi};x) = \mathbb{E}_{z^{(1:N)}\sim q_{\tilde{\phi}}(\cdot|x)}\left[\log\left(\frac{1}{N}\sum_{i=1}^{N} w^{(i)}\right) + \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{q_\phi(z_t^{(i)}|x)}{q_{\tilde{\phi}}(z_t^{(i)}|x)}\log\frac{\sum_{j=1}^{N}\widetilde{w}_{1:t}^{(j)}\widetilde{\Gamma}_t^{(j)}}{\sum_{j=1}^{N}\widetilde{w}_{1:t-1}^{(j)}\widetilde{\Gamma}_{t-1}^{(j)}}\right],$$

$$\mathcal{L}_{\mathcal{S}}^{(N)}(\theta,\phi,\tilde{\phi};x) = \mathbb{E}_{z^{(1:N)}\sim q_{\tilde{\phi}}(\cdot|x)}\left[\log\left(\frac{1}{N}\sum_{i=1}^{N} w^{(i)}\right) + \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{q_\phi(z_t^{(i)}|x)}{q_{\tilde{\phi}}(z_t^{(i)}|x)}\log\frac{\sum_{j\in\mathcal{S}_i}\widetilde{w}_{1:t}^{(j)}\widetilde{\Gamma}_t^{(j)} + \sum_{j\notin\mathcal{S}_i}\widetilde{w}^{(j)}}{\sum_{j\in\mathcal{S}_i}\widetilde{w}_{1:t-1}^{(j)}\widetilde{\Gamma}_{t-1}^{(j)} + \sum_{j\notin\mathcal{S}_i}\widetilde{w}^{(j)}}\right].$$

We simply denote $\mathcal{L}_{\mathrm{VIFLE}}^{(N)}(\theta,\phi_0,\phi_0;x)$ as $\mathcal{L}_{\mathrm{VIFLE}}^{(N)}(\theta,\phi_0;x)$, $\mathcal{L}_{\mathrm{FR}}^{(N)}(\theta,\phi_0,\phi_0;x)$ as $\mathcal{L}_{\mathrm{FR}}^{(N)}(\theta,\phi_0;x)$, and $\mathcal{L}_{\mathcal{S}}^{(N)}(\theta,\phi_0,\phi_0;x)$ as $\mathcal{L}_{\mathcal{S}}^{(N)}(\theta,\phi_0;x)$. More generally, we denote $\mathcal{L}_{\mathrm{VIFLE}}^{(N)}(\theta,\phi,\tilde{\phi};x)$ as $\mathcal{L}_{\mathrm{VIFLE}}^{(N)}(\theta,\phi;x)$, $\mathcal{L}_{\mathrm{FR}}^{(N)}(\theta,\phi,\tilde{\phi};x)$ as $\mathcal{L}_{\mathrm{FR}}^{(N)}(\theta,\phi;x)$, and $\mathcal{L}_{\mathcal{S}}^{(N)}(\theta,\phi,\tilde{\phi};x)$ as $\mathcal{L}_{\mathcal{S}}^{(N)}(\theta,\phi;x)$ for $\tilde{\phi} = \mathrm{stop\_grad}(\phi)$. Using these definitions, we provide the proof of gradients in Theorem 1.

Lemma 2. *Note that*

$$\nabla_\phi \mathcal{L}_{\mathrm{VIFLE}}^{(N)}(\theta,\phi;x)|_{\phi=\phi_0} = \mathbb{E}[g_{\mathrm{low}} + g_{\mathrm{VIFLE}}],$$
$$\nabla_\phi \mathcal{L}_{FR}^{(N)}(\theta,\phi;x)|_{\phi=\phi_0} = \mathbb{E}[g_{\mathrm{low}} + g_{\mathrm{FR}}],$$

where the expectations are taken with respect to $q_{\phi_0}(z^{(1:N)}|x)$ and the gradients at $\phi = \phi_0$.

Proof. Using the definition of $\mathcal{L}_{\mathrm{VIFLE}}$ and $\tilde{\phi} = \text{stop\_grad}(\phi)$, we can easily derive the following equation:

$$\begin{aligned}\nabla_\phi \mathcal{L}^{(N)}_{\mathrm{VIFLE}}(\theta,\phi,\tilde{\phi};x) &= \mathbb{E}_{q_{\tilde{\phi}}(z^{(1:N)}|x)}\left[\nabla_\phi \log\left(\frac{1}{N}\sum_{i=1}^N w^{(i)}\right) + \sum_{i=1}^N\sum_{t=1}^T \log\frac{\sum_{j\neq i}\widetilde{w}^{(j)} + \widetilde{w}^{(i)}_{1:t}\widetilde{\Gamma}^{(i)}_t}{\sum_{j\neq i}\widetilde{w}^{(j)} + \widetilde{w}^{(i)}_{1:t-1}\widetilde{\Gamma}^{(i)}_{t-1}}\frac{\nabla_\phi q_\phi(z^{(i)}_t|x)}{q_{\tilde{\phi}}(z^{(i)}_t|x)}\right] \\ &= \mathbb{E}_{q_{\phi}(z^{(1:N)}|x)}\left[\nabla_\phi \log\left(\frac{1}{N}\sum_{i=1}^N w^{(i)}\right) + \sum_{i=1}^N\sum_{t=1}^T \log\frac{\sum_{j\neq i} w^{(j)} + w^{(i)}_{1:t}\Gamma^{(i)}_t}{\sum_{j\neq i} w^{(j)} + w^{(i)}_{1:t-1}\Gamma^{(i)}_{t-1}}\frac{\nabla_\phi q_\phi(z^{(i)}_t|x)}{q_{\phi}(z^{(i)}_t|x)}\right].\end{aligned}$$

This equation leads $\nabla_\phi \mathcal{L}^{(N)}_{\mathrm{VIFLE}}(\theta,\phi;x)|_{\phi=\phi_0} = \mathbb{E}[g_{\mathrm{low}} + g_{\mathrm{VIFLE}}]$. Similarly, we can get the other equation $\nabla_\phi \mathcal{L}^{(N)}_{\mathrm{FR}}(\theta,\phi;x)|_{\phi=\phi_0} = \mathbb{E}[g_{\mathrm{low}} + g_{\mathrm{FR}}]$. □

Now, we prove the proof of inequalities in Theorem 1.

Lemma 3. *The inequality*

$$\mathcal{L}^{(N)}_{\mathrm{IWAE}}(\theta,\phi;x) \geq \mathcal{L}^{(N)}_{\mathrm{VIFLE}}(\theta,\phi;x)$$

holds for any θ, ϕ, and x.

Proof. From the definitions of $\mathcal{L}_{\mathrm{IWAE}}$ and $\mathcal{L}_{\mathrm{VIFLE}}$,

$$\begin{aligned}\mathcal{L}^{(N)}_{\mathrm{IWAE}}(\theta,\phi;x) - \mathcal{L}^{(N)}_{\mathrm{VIFLE}}(\theta,\phi;x) &= \mathbb{E}_{z^{(1:N)}\sim q_\phi(\cdot|x)}\left[-\sum_{i=1}^N\sum_{t=1}^T \log\frac{\sum_{j\neq i} w^{(j)} + w^{(i)}_{1:t}\Gamma^{(i)}_t}{\sum_{j\neq i} w^{(j)} + w^{(i)}_{1:t-1}\Gamma^{(i)}_{t-1}}\right] \\ &= \mathbb{E}_{z^{(1:N)}\sim q_\phi(\cdot|x)}\left[\sum_{i=1}^N \log\frac{\sum_{j\neq i} w^{(j)} + \Gamma^{(i)}_0}{\sum_{j\neq i} w^{(j)} + w^{(i)}}\right] \\ &= \mathbb{E}_{z^{(1:N)}\sim q_\phi(\cdot|x)}\left[\sum_{i=1}^N \log\frac{\sum_{j\neq i} w^{(j)} + \mathbb{E}_{z^{(i)}\sim q_\phi(\cdot|x)}[w^{(i)}]}{\sum_{j=1}^N w^{(j)}}\right] \\ &\geq \mathbb{E}_{z^{(1:N)}\sim q_\phi(\cdot|x)}\left[\sum_{i=1}^N \log\frac{\sum_{j\neq i} w^{(j)} + w^{(i)}}{\sum_{j=1}^N w^{(j)}}\right] = 0.\end{aligned}$$

Here, the last inequality is derived from the Jensen's inequality. □

Lemma 4. *The inequality*

$$\mathcal{L}^{(N)}_{\mathrm{VIFLE}}(\theta,\phi;x) \geq \mathcal{L}^{(N)}_{FR}(\theta,\phi;x)$$

holds for any fixed θ, ϕ, and x.

Proof. Using definitions, we can derive the following inequality:

$$\begin{aligned}&\mathcal{L}^{(N)}_{\mathrm{VIFLE}}(\theta,\phi;x) - \mathcal{L}^{(N)}_{\mathrm{FR}}(\theta,\phi;x) \\ &= \mathbb{E}_{z^{(1:N)}\sim q_\phi(\cdot|x)}\left[\sum_{i=1}^N\sum_{t=1}^T\left(\log\frac{\sum_{j\neq i} w^{(j)} + W^{(i)}_t\Gamma^{(i)}_t}{\sum_{j\neq i} w^{(j)} + w^{(i)}_{1:t-1}\Gamma^{(i)}_{t-1}} - \log\frac{\sum_{j=1}^N w^{(j)}_{1:t}\Gamma^{(j)}_t}{\sum_{j=1}^N w^{(j)}_{1:t-1}\Gamma^{(j)}_{t-1}}\right)\right] \\ &= \mathbb{E}_{z^{(1:N)}\sim q_\phi(\cdot|x)}\left[\sum_{i=1}^N\left(\log\frac{\sum_{j\neq i} w^{(j)} + w^{(i)}}{\sum_{j\neq i} w^{(j)} + \Gamma^{(i)}_0} - \log\frac{\sum_{j=1}^N w^{(j)}}{\sum_{j=1}^N \Gamma^{(j)}_0}\right)\right] \\ &= \mathbb{E}_{z^{(1:N)}\sim q_\phi(\cdot|x)}\left[\sum_{i=1}^N \log\frac{\sum_{j=1}^N \Gamma^{(j)}_0}{\sum_{j\neq i} w^{(j)} + \Gamma^{(i)}_0}\right] \\ &= \mathbb{E}_{z^{(1:N)}\sim q_\phi(\cdot|x)}\left[\sum_{i=1}^N \log\frac{\sum_{j\neq i}\mathbb{E}_{z^{(j)}\sim q_\phi(\cdot|x)}[w^{(j)}] + \Gamma^{(i)}_0}{\sum_{j\neq i} w^{(j)} + \Gamma^{(i)}_0}\right] \\ &\geq \mathbb{E}_{z^{(1:N)}\sim q_\phi(\cdot|x)}\left[\sum_{i=1}^N \log\frac{\sum_{j\neq i} w^{(j)} + \Gamma^{(i)}_0}{\sum_{j\neq i} w^{(j)} + \Gamma^{(i)}_0}\right] = 0.\end{aligned}$$

To derive the inequality in the last line, we use Jensen's inequality. □

Lemma 5. *For all index sets $\mathcal{S}$ and $\mathcal{T}$, if $\mathcal{S}_i \subseteq \mathcal{T}_i$ for every i, then there are variational objectives $\mathcal{L}_\mathcal{S}$ and $\mathcal{L}_\mathcal{T}$ have following relation:*

$$\mathcal{L}_\mathcal{S}(\theta, \phi_0; x) \geq \mathcal{L}_\mathcal{T}(\theta, \phi_0; x)$$

for any fixed θ, ϕ, and x.

Proof. From the inequality

$$\begin{aligned}
&\mathcal{L}_\mathcal{S}^{(N)}(\theta, \phi; x) - \mathcal{L}_\mathcal{T}^{(N)}(\theta, \phi; x) \\
&= \mathbb{E}_{z^{(1:N)} \sim q_\phi(_|x)} \left[\sum_{i=1}^{N} \sum_{t=1}^{T} \left(\log \frac{\sum_{j \in \mathcal{S}_i} w_{1:t}^{(j)} \Gamma_t^{(j)} + \sum_{j \notin \mathcal{S}_i} w^{(j)}}{\sum_{j \in \mathcal{S}_i} w_{1:t-1}^{(j)} \Gamma_{t-1}^{(j)} + \sum_{j \notin \mathcal{S}_i} w^{(j)}} - \log \frac{\sum_{j \in \mathcal{T}_i} w_{1:t}^{(j)} \Gamma_t^{(j)} + \sum_{j \notin \mathcal{T}_i} w^{(j)}}{\sum_{j \in \mathcal{T}_i} w_{1:t-1}^{(j)} \Gamma_{t-1}^{(j)} + \sum_{j \notin \mathcal{T}_i} w^{(j)}} \right) \right] \\
&= \mathbb{E}_{z^{(1:N)} \sim q_\phi(_|x)} \left[\sum_{i=1}^{N} \left(\log \frac{\sum_{j \in \mathcal{S}_i} w^{(j)} + \sum_{j \notin \mathcal{S}_i} w^{(j)}}{\sum_{j \in \mathcal{S}_i} \Gamma_0^{(j)} + \sum_{j \notin \mathcal{S}_i} w^{(j)}} - \log \frac{\sum_{j \in \mathcal{T}_i} w^{(j)} + \sum_{j \notin \mathcal{T}_i} w^{(j)}}{\sum_{j \in \mathcal{T}_i} \Gamma_0^{(j)} + \sum_{j \notin \mathcal{T}_i} w^{(j)}} \right) \right] \\
&= \mathbb{E}_{z^{(1:N)} \sim q_\phi(_|x)} \left[\sum_{i=1}^{N} \log \frac{\sum_{j \in \mathcal{T}_i} \Gamma_0^{(j)} + \sum_{j \notin \mathcal{T}_i} w^{(j)}}{\sum_{j \in \mathcal{S}_i} \Gamma_0^{(j)} + \sum_{j \notin \mathcal{S}_i} w^{(j)}} \right] \\
&= \mathbb{E}_{z^{(1:N)} \sim q_\phi(_|x)} \left[\sum_{i=1}^{N} \log \frac{\sum_{j \in \mathcal{S}_i} \Gamma_0^{(j)} + \sum_{j \in \mathcal{T}_i \setminus \mathcal{S}_i} \mathbb{E}[w^{(j)}] + \sum_{j \notin \mathcal{T}_i} w^{(j)}}{\sum_{j \in \mathcal{S}_i} \Gamma_0^{(j)} + \sum_{j \notin \mathcal{S}_i} w^{(j)}} \right] \\
&\geq \mathbb{E}_{z^{(1:N)} \sim q_\phi(_|x)} \left[\sum_{i=1}^{N} \log \frac{\sum_{j \in \mathcal{S}_i} \Gamma_0^{(j)} + \sum_{j \in \mathcal{T}_i \setminus \mathcal{S}_i} w^{(j)} + \sum_{j \notin \mathcal{T}_i} w^{(j)}}{\sum_{j \in \mathcal{S}_i} \Gamma_0^{(j)} + \sum_{j \notin \mathcal{S}_i} w^{(j)}} \right] \\
&= \mathbb{E}_{z^{(1:N)} \sim q_\phi(_|x)} \left[\sum_{i=1}^{N} \log \frac{\sum_{j \in \mathcal{S}_i} \Gamma_0^{(j)} + \sum_{j \notin \mathcal{S}_i} w^{(j)}}{\sum_{j \in \mathcal{S}_i} \Gamma_0^{(j)} + \sum_{j \notin \mathcal{S}_i} w^{(j)}} \right] = 0.
\end{aligned}$$

To derive the inequality in the last line, we use Jensen's inequality. □

Now, we are ready to prove the Theorem 1.

Proof of Theorem 1. Lemma 2 states that gradient estimators are indeed unbiased gradient estimators of surrogate objectives $\mathcal{L}_{\text{VIFLE}}$ and $\mathcal{L}_{\text{FR}}$. Moreover, Lemma 3, Lemma 4 prove the remaining statement of theorem. □

E. Proof of Theorem 2

Proof. Using the mean value theorem, we obtain the following equation:

$$\log \left\{ \frac{1}{N} \sum_{j=1}^{N} w^{(j)} \right\} - \log \left\{ \frac{1}{N} \left(\sum_{j \neq i} w^{(j)} + p_\theta(x) \right) \right\} = \frac{1}{m} \left\{ \frac{1}{N} \sum_{j=1}^{N} w^{(j)} - \frac{1}{N} \left(\sum_{j \neq i} w^{(j)} + p_\theta(x) \right) \right\},$$

where the value m lies between $\frac{1}{N} \sum_{j=1}^{N} w^{(j)}$ and $\frac{1}{N} \left(\sum_{j \neq i} w^{(j)} + p_\theta(x) \right)$

From the fact that $\mathbb{E}_{z^{(i)} \sim q_\phi(\_|x)}[w^{(i)}] = p_\theta(x)$, the strong law of large number states that

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} w^{(i)} \xrightarrow{a.s.} p_\theta(x). \tag{15}$$

We notice that for sufficient large N, $\frac{1}{2}p_\theta(x) \leq \frac{1}{N}\sum_{i=1}^N w^{(i)}$ and $\frac{1}{2}p_\theta(x) \leq \frac{1}{N}\left(\sum_{j\neq i} w^{(j)} + p_\theta(x)\right)$ from (15). Thus,

$$-M\left|\frac{1}{N}\sum_{j=1}^N w^{(j)} - \frac{1}{N}\left(\sum_{j\neq i} w^{(j)} + p_\theta(x)\right)\right| \leq \log\frac{1}{N}\sum_{j=1}^N w^{(j)} - \log\frac{1}{N}\left(\sum_{j\neq i} w^{(j)} + p_\theta(x)\right).$$

where $M = \frac{2}{p_\theta(x)}$.

In addition, (15) guarantees that

$$\lim_{N\to\infty} \mathcal{L}_{\text{IWAE}}^{(N)} \xrightarrow{a.s.} \log p_\theta(x)$$

as shown by Burda et al. (2016).

$$\begin{aligned}
\mathcal{L}_{\text{VIFLE}}^{(N)}(\theta,\phi;x) &= \mathcal{L}_{\text{IWAE}}^{(N)}(\theta,\phi;x) + \mathbb{E}_{z^{(1:N)}\sim q_\phi(\cdot|x)}\left[\sum_{i=1}^N \log\frac{\sum_{j=1}^N w^{(j)}}{\sum_{j\neq i} w^{(j)} + p_\theta(x)}\right] \\
&= \mathcal{L}_{\text{IWAE}}^{(N)}(\theta,\phi;x) + \mathbb{E}_{z^{(1:N)}\sim q_\phi(\cdot|x)}\left[\sum_{i=1}^N \log\frac{\frac{1}{N}\sum_{j=1}^N w^{(j)}}{\frac{1}{N}\left(\sum_{j\neq i} w^{(j)} + p_\theta(x)\right)}\right] \\
&\geq \mathcal{L}_{\text{IWAE}}^{(N)}(\theta,\phi;x) - \mathbb{E}_{z^{(1:N)}\sim q_\phi(\cdot|x)}\left[M\left|\sum_{i=1}^N \frac{1}{N}\left(\sum_{j=1}^N w^{(j)} - \sum_{j\neq i} w^{(j)} - p_\theta(x)\right)\right|\right] \\
&= \mathcal{L}_{\text{IWAE}}^{(N)}(\theta,\phi;x) - M\mathbb{E}_{z^{(1:N)}\sim q_\phi(\cdot|x)}\left[\left|\frac{1}{N}\sum_{i=1}^N w^{(i)} - p_\theta(x)\right|\right] \\
&\xrightarrow{a.s.} \log p_\theta(x) + 0
\end{aligned}$$

Since the inequality $\mathcal{L}_{\text{IWAE}}^{(N)} \geq \mathcal{L}_{\text{VIFLE}}^{(N)}$ always holds, we can conclude that $\mathcal{L}_{\text{VIFLE}}^{(N)} \xrightarrow{a.s.} \log p_\theta(x)$. □

F. Variance of Gradient

Figure 2 shows the variance of gradient of VIFLE, VIMCO, and reparameterization trick in 20-dim Gaussian linear dynamics. The variance of gradient is measured by trace of sample covariance matrix during training. Note that the variance of REINFORCE is above 500, so does not appear in the figure.

G. Generalization of VIFLE

For sequential model learning, there are tighter objectives based on SMC (Maddison et al., 2017a; Naesseth et al., 2018; Le et al., 2018). For these objectives, we extend our work. Suppose that our algorithm resamples S times at $0 = t_1 < t_2 < \cdots < t_{S+1} = T$. Then, FIVO objective becomes

$$\mathcal{L}_{\text{FIVO}} = \mathbb{E}\left[\sum_{s=1}^S \log\left(\frac{1}{N}\sum_{i=1}^N \prod_{t=t_s+1}^{t_{s+1}} w_t^{(i)}\right)\right].$$

Here, the gradient of FIVO w.r.t. ϕ is given by

$$\nabla_\phi \mathcal{L}_{\text{FIVO}} = \mathbb{E}\left[\nabla_\phi \log\hat{p}_{\text{FIVO}} + \sum_{t=1}^T\sum_{i=1}^N\left\{\log\hat{p}_{\text{FIVO}}\nabla_\phi \log q_\phi(z_t^{(i)}|z_{1:t-1}^{(i)},x) + \sum_{s=1}^S \log\hat{p}_{\text{FIVO}}\nabla_\phi \log r_s^{(i)}\right\}\right]$$

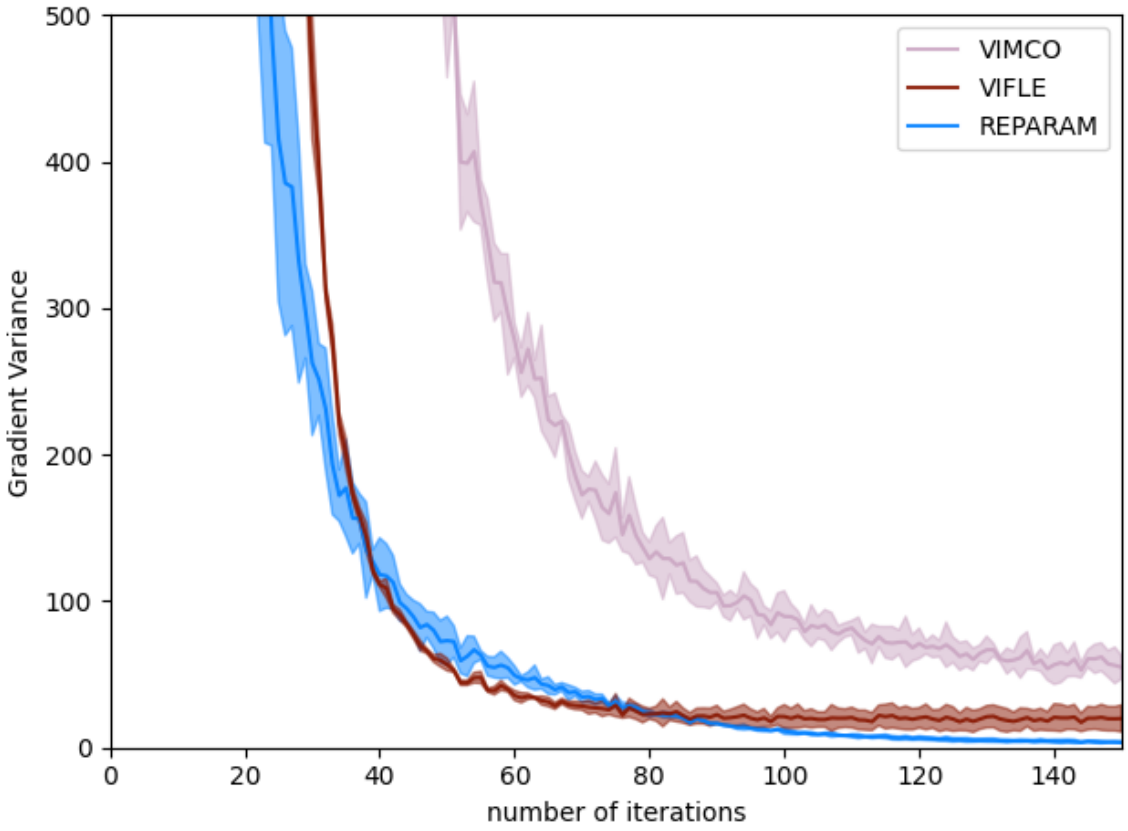

Figure 2. In this figure, the x-axis indicates the number of training iterations, and the y-axis indicates the variance of gradient measured by trace of sample covariance matrix.

where

$$\hat{p}_{\text{FIVO}} = \sum_{s=1}^{S} \log\left(\frac{1}{N}\sum_{i=1}^{N}\prod_{t=t_s+1}^{t_{s+1}} w_t^{(i)}\right),$$
$$r_s^{(i)} = \frac{w_{t_s+1:t_{s+1}}^{(i)}}{\sum_{j=1}^{N} w_{t_s+1:t_{s+1}}^{(j)}}.$$

Due to the high variance, we ignore gradient term from resampling as same as previous works.

To compute the baselines for $q_\phi(z_{t'}^{(i)}|z_{1:t'-1}^{(i)}, x)$, suppose that $t' \in [t_{s'}+1, t_{s'+1}]$. Then,

$$b(z_{1:t'-1}^{(i)}, x; \phi) = \sum_{s=0}^{s'-1} \log\left(\frac{1}{N}\sum_{j=1}^{N}\prod_{t=t_s+1}^{t_{s+1}} w_t^{(j)}\right) + \log\left[\frac{1}{N}\left(\sum_{j\neq i} w_{t_{s'}+1:t_{s'+1}}^{(j)} \Gamma(z_{1:t_{s'+1}-1}^{(j)}, x) + w_{t_{s'}+1:t'-1}^{(i)} \Gamma(z_{1:t'-1}^{(i)}, x)\right)\right]$$

is not depend on $z_{t'}^{(i)}$, so it is a baselines for $q_\phi(z_{t'}^{(i)}|z_{1:t'-1}^{(i)}, x)$. Now, we will derive the biased but low-variance objective:

$$g_{\text{VIFLE}} = \sum_{t=1}^{T}\sum_{i=1}^{N}(b'(z_{1:t}^{(i)}, x; \phi) - b(z_{1:t}^{(i)}, x; \phi))\nabla_\phi \log q_\phi(z_t^{(i)}|z_{1:t-1}^{(i)}, x)$$

where

$$b'(z^{(j)}_{1:t'}, x; \phi) = \sum_{s=0}^{s'-1} \log\left(\frac{1}{N}\sum_{i=1}^{N}\prod_{t=t_s+1}^{t_{s+1}} w_t^{(i)}\right) + \log\left[\frac{1}{N}\left(\sum_{i\neq j} w^{(i)}_{t_{s'}+1:t_{s'+1}}\Gamma(z^{(i)}_{1:t_{s'+1}}, x) + w^{(j)}_{t_{s'}+1:t'}\Gamma(z^{(j)}_{1:t'}, x)\right)\right].$$

When $S = 0$, it reduces to IWAE case.

H. Experiment Details

All gradient estimators show similar running times for all tasks. For synthetic datasets, the proposal distributions are of the form

$$q_\phi(z_t|z_{t-1}, x_t) = \text{Bernoulli}(f_\phi(z_{t-1}, x_t)),$$
$$q_\phi(z_t|z_{t-1}, x_t) = \mathcal{N}(\mu_\phi(z_{t-1}, x_t), \sigma^2_\phi(z_{t-1}, x_t))$$

with respect to Bernoulli dynamics and Gaussian dynamics. Here, μ_θ and σ^2_θ are implemented by a neural network to generate the mean and diagonal of covariance matrix for Gaussian distribution respectively. Also, f_ϕ is implemented by neural networks to generate logit for Bernoulli distribution. During the posterior inference in synthetic domains, we use Adam optimizer with learning rate 10^{-3}. Each neural network is composed of one fully connected layer.

For polyphonic music datasets, our algorithm is implemented based on *variational recurrent neural network (VRNN)* (Chung et al., 2015) and especially, each joint distribution $p_\theta(x_t, z_t|x_{1:t-1}, z_{1:t-1})$ is factorized by

$$p_\theta(z_t|z_{1:t-1}, x_{1:t-1})p_\theta(x_t|z_{1:t}, x_{1:t-1}).$$

Here, it is factorized by $p_\theta(z_1)p_\theta(x_1|z_1)$ for $t = 1$ case. Also, all distributions are of the form

$$p_\theta(z_t|z_{1:t-1}, x_{1:t-1}) = \mathcal{N}(\mu_\theta(z_{1:t-1}, x_{1:t-1}), \sigma^2_\theta(z_{1:t-1}, x_{1:t-1})),$$
$$p_\theta(x_t|z_{1:t}, x_{1:t-1}) = \text{Bernoulli}(f_\theta(z_{1:t}, x_{1:t-1})),$$
$$q_\phi(z_t|z_{1:t-1}, x_{1:t}) = \mathcal{N}(\mu_\phi(z_{1:t-1}, x_{1:t}), \sigma^2_\phi(z_{1:t-1}, x_{1:t}))$$

for continuous latent variable model. Here, μ_θ and σ^2_θ are implemented by a neural network to generate the mean and diagonal of covariance matrix for Gaussian distribution respectively. Also, f_θ and f_ϕ are implemented by neural networks, which generate logit for Bernoulli distribution. For discrete latent variable model,

$$p_\theta(z_t|z_{1:t-1}, x_{1:t-1}) = \text{Bernoulli}(f^{(1)}_\theta(z_{1:t-1}, x_{1:t-1})),$$
$$p_\theta(x_t|z_{1:t}, x_{1:t-1}) = \text{Bernoulli}(f^{(2)}_\theta(z_{1:t}, x_{1:t-1})),$$
$$q_\phi(z_t|z_{1:t-1}, x_{1:t}) = \text{Bernoulli}(f_\phi(z_{1:t-1}, x_{1:t}))$$

for discrete latent variables. Here, all f_θ and f_ϕ are implemented by neural networks, which generate logit for Bernoulli distribution. All neural networks use a fully connected layer and share a recurrent neural network to summarize $z_{1:t-1}, x_{1:t-1}$. In addition, for both discrete and continuous latent variable models, we use 32-dimensional latent variables for JSB dataset and 64-dimensional latent variables for the others.