Training Variational Autoencoders with Buffered Stochastic Variational Inference

Rui Shu
Stanford University

Hung H. Bui
VinAI

Jay Whang
Stanford University

Stefano Ermon
Stanford University

Abstract

The recognition network in deep latent variable models such as variational autoencoders (VAEs) relies on amortized inference for efficient posterior approximation that can scale up to large datasets. However, this technique has also been demonstrated to select suboptimal variational parameters, often resulting in considerable additional error called the *amortization gap*. To close the amortization gap and improve the training of the generative model, recent works have introduced an additional refinement step that applies stochastic variational inference (SVI) to improve upon the variational parameters returned by the amortized inference model. In this paper, we propose the *Buffered Stochastic Variational Inference* (BSVI), a new refinement procedure that makes use of SVI's sequence of intermediate variational proposal distributions and their corresponding importance weights to construct a new generalized importance-weighted lower bound. We demonstrate empirically that training the variational autoencoders with BSVI consistently out-performs SVI, yielding an improved training procedure for VAEs.

1 Introduction

Deep generative latent-variable models are important building blocks in current approaches to a host of challenging high-dimensional problems including density estimation [1, 2, 3], semi-supervised learning [4, 5] and representation learning for downstream tasks [6, 7, 8, 9]. To train these models, the principle of maximum likelihood is often employed. However, maximum likelihood is often intractable due to the difficulty of marginalizing the latent variables. Variational Bayes addresses this by instead providing a tractable lower bound of the log-likelihood, which serves as a surrogate target for maximization. Variational Bayes, however, introduces a *per sample* optimization subroutine to find the variational proposal distribution that best matches the true posterior distribution (of the latent variable given an input observation). To amortize the cost of this optimization subroutine, the variational autoencoder introduces an amortized inference model that learns to predict the best proposal distribution given an input observation [1, 10, 11, 12].

Although the computational efficiency of amortized inference has enabled latent variable models to be trained at scale on large datasets [13, 14], amortization introduces an additional source of error in the approximation of the posterior distributions if the amortized inference model fails to predict the optimal proposal distribution. This additional source of error, referred to as the amortization gap [15], causes variational autoencoder training to further deviate from maximum likelihood training [15, 16].

To improve training, numerous methods have been developed to reduce the amortization gap. In this paper, we focus on a class of methods [17, 18, 19] that takes an initial proposal distribution predicted by the amortized inference model and refines this initial distribution with the application of *Stochastic Variational Inference* (SVI) [20]. Since SVI applies gradient ascent to iteratively update the proposal distribution, a by-product of this procedure is a trajectory of proposal distributions $(q_0, \dots, q_k)$ and their corresponding importance weights $(w_0, \dots w_k)$. The intermediate distributions are discarded, and only the last distribution q_k is retained for updating the generative model. Our key insight is that the intermediate importance weights can be *repurposed* to further improve training. Our contributions are as follows

1. We propose a new method, *Buffered Stochastic*

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

Variational Inference (BSVI), that takes advantage of the intermediate importance weights and constructs a new lower bound (the BSVI bound).

2. We show that the BSVI bound is a special instance of a family of generalized importance-weighted lower bounds.
3. We show that training variational autoencoders with BSVI consistently outperforms SVI, demonstrating the effectiveness of leveraging the intermediate weights.

Our paper shows that BSVI is an attractive replacement of SVI with minimal development and computational overhead.

2 Background and Notation

We consider a latent-variable generative model $p_\theta(x, z)$ where $x \in \mathcal{X}$ is observed, $z \in \mathcal{Z}$ is latent, and θ are the model's parameters. The marginal likelihood $p_\theta(x)$ is intractable but can be lower bounded by the evidence lower bound (ELBO)

$$\ln p_\theta(x) \geq \mathbb{E}_{q(z)}\left[\ln \frac{p_\theta(x,z)}{q(z)}\right] = \mathbb{E}_{q(z)} \ln w(z), \quad (1)$$

which holds for any distribution $q(z)$. Since the gap of this bound is exactly the Kullback-Leibler divergence $D(q(z) \parallel p_\theta(z \mid x))$, $q(z)$ is thus the variational approximation of the posterior. Furthermore, by viewing q as a proposal distribution in an importance sampler, we refer to $w(z) = \frac{p_\theta(x,z)}{q(z)}$ as an unnormalized importance weight. Since $w(z)$ is a random variable, the variance can be reduced by averaging the importance weights derived from i.i.d samples from $q(z)$. This yields the Importance-Weighted Autonenocder (IWAE) bound [21],

$$\ln p_\theta(x) \geq \mathbb{E}_{z_1 \dots z_k \overset{\text{i.i.d.}}{\sim} q}\left[\ln \frac{1}{k}\sum_{i=1}^{k} w(z_i)\right] \geq \text{ELBO}, \quad (2)$$

which admits a tighter lower bound than the ELBO [21, 22].

2.1 Stochastic Variational Inference

The generative model can be trained by jointly optimizing q and θ to maximize the lower bound over the data distribution $\hat{p}(x)$. Supposing the variational family $\mathcal{Q} = \{q(z\,;\lambda)\}_{\lambda\in\Lambda}$ is parametric and indexed by the parameter space Λ (e.g. a Gaussian variational family indexed by mean and covariance parameters), the optimization problem becomes

$$\max_\theta \mathbb{E}_{\hat{p}(x)}\left[\max_\lambda \mathbb{E}_{q(z;\lambda)} \ln w(z\,;\lambda,\theta)\right]. \quad (3)$$

where importance weight w is now

$$w(z\,;\lambda,\theta) = \frac{p_\theta(x,z)}{q(z\,;\lambda)}. \quad (4)$$

For notational simplicity, we omit the dependency on x. For a fixed choice of θ and x, [17] proposed to optimize λ via gradient ascent, where one initializes with λ_0 and takes successive steps of

$$\lambda_{i+1} \leftarrow \lambda_i + \eta \nabla_{\lambda_i} \text{ELBO}, \quad (5)$$

for which the ELBO gradient with respect to λ_i can be approximated via Monte Carlo sampling as

$$\nabla_{\lambda_i} \text{ELBO} \approx \frac{1}{m}\sum_{j=1}^{m} \nabla_{\lambda_i} \ln w(z_{\lambda_i}(\epsilon_i^{(j)})\,;\lambda_i,\theta) \quad (6)$$

where $z_i^{(j)} = z_{\lambda_i}(\epsilon_i^{(j)}) \sim q(z\,;\lambda_i)$ is reparameterized as a function of λ_i and a base distribution $p_0(\epsilon)$. We note that k applications gradient ascent generates a trajectory of variational parameters $(\lambda_0, \dots, \lambda_k)$, where we use the final parameter λ_k for the approximation. Following the convention in [20], we refer to this procedure as *Stochastic Variational Inference* (SVI).

2.2 Amortized Inference Suboptimality

The SVI procedure introduces an inference *subroutine* that optimizes the proposal distribution $q(z\,;\lambda)$ *per sample*, which is computationally costly. [1, 10] observed that the computational cost of inference can be *amortized* by introducing an inference model $f_\phi : \mathcal{X} \to \Lambda$, parameterized by ϕ, that directly seeks to learn the mapping $x \mapsto \lambda^*$ from each sample x to an optimal λ^* that solves the maximization problem

$$\lambda^* = \arg\max_\lambda \mathbb{E}_{q(z;\lambda)} \ln \frac{p_\theta(x,z)}{q(z\,;\lambda)}. \quad (7)$$

This yields the amortized ELBO optimization problem

$$\max_{\theta,\phi} \mathbb{E}_{\hat{p}(x)}\left[\mathbb{E}_{q(z;f_\phi(x))} \ln \frac{p_\theta(x,z)}{q(z\,;f_\phi(x))}\right], \quad (8)$$

where $q(z\,;f_\phi(x))$ can be concisely rewritten (with a slight abuse of notation) as $q_\phi(z \mid x)$ to yield the standard variational autoencoder objective [1].

While computationally efficient, the influence of the amortized inference model on the training dynamics of the generative model has recently come under scrutiny [15, 17, 18, 16]. A notable consequence of amortization is the *amortization gap*

$$D(q_\phi(z \mid x) \parallel p_\theta(z \mid x)) - D(q(z\,;\lambda^*) \parallel p_\theta(z \mid x)) \quad (9)$$

which measures the additional error incurred when the amortized inference model is used instead of the optimal λ^* for approximating the posterior [15]. A large amortization gap can present a potential source of concern since it introduces further deviation from the maximum likelihood objective [16].

2.3 Amortization-SVI Hybrids

To close the amortization gap, [17] proposed to blend amortized inference with SVI. Since SVI requires one to initialize λ_0, a natural solution is to set $\lambda_0 = f_\phi(x)$. Thus, SVI is allowed to fine-tune the initial proposal distribution found by the amortized inference model and reduce the amortization gap. Rather than optimizing θ, ϕ jointly with the amortized ELBO objective Eq. (8), the training of the inference and generative models is now decoupled; ϕ is trained to optimize the amortized ELBO objective, but θ is trained to approximately optimize Eq. (3), where $\lambda^* \approx \lambda_k$ is approximated via SVI. To enable end-to-end training of the inference and generative models, [18] proposed to backpropagate through the SVI steps via a finite-difference estimation of the necessary Hessian-vector products. Alternatively, [19] adopts a learning-to-learn framework where an inference model iteratively outputs λ_{i+1} as a function of λ_i and the ELBO gradient.

3 Buffered Stochastic Variational Inference

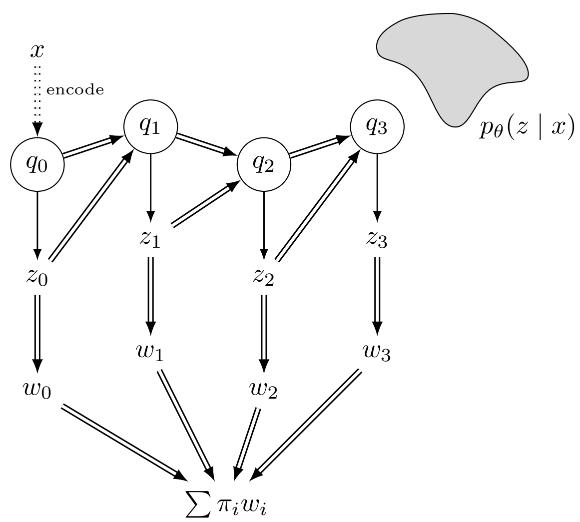


Figure 1: Idealized visualization of Buffered Stochastic Variational Inference. Double arrows indicate deterministic links, and single arrows indicate stochastic links that involve sampling. The dotted arrow from x to q_0 denotes that the initial variational parameters are given by the encoder. For notational simplicity, we omitted the dependence of $q_{1:k}$ on x and the model parameters ϕ, θ.

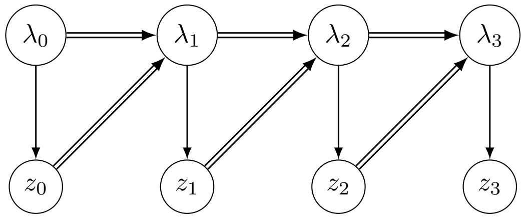
In this paper, we focus on the simpler, decoupled training procedure described by [17] and identify a new way of improving the SVI training procedure (orthogonal to the end-to-end approaches in [18, 19]). Our key observation is that, as part of the gradient ascent estimation in Eq. (6), the SVI procedure necessarily generates a sequence of importance weights $(w_0, \ldots, w_k)$, where $w_i = w(z_i;\lambda_i, \theta)$. Since $(\ln w_k)$ likely achieves the highest ELBO, the intermediate weights $(w_0, \ldots, w_{k-1})$ are subsequently discarded in the SVI training procedure, and only $\nabla_\theta \ln w_k$ is retained for updating the generative model parameters. However, if the preceding proposal distributions $(q_{k-1}, q_{k-2}, \ldots)$ are also reasonable approximations of the posterior, then it is potentially wasteful to discard their corresponding importance weights. A natural question to ask then is whether the full trajectory of weights $(w_0, \ldots, w_k)$ can be leveraged to further improve the training of the generative model.

Taking inspiration from IWAE's weight-averaging mechanism, we propose a modification to the SVI procedure where we simply keep a *buffer* of the entire importance weight trajectory and use an average of the importance weights $\sum_i \pi_i w_i$ as the objective in training the generative model.[1] The generative model is then updated with the gradient $\nabla_\theta \ln \sum_i \pi_i w_i$. We call this procedure *Buffered Stochastic Variational Inference* (BSVI) and denote $\ln \sum_i \pi_i w_i$ as the BSVI objective. We describe the BSVI training procedure in Algorithm 1 and contrast it with SVI training. For notational simplicity, we shall always imply initialization with an amortized inference model when referring to SVI and BSVI.

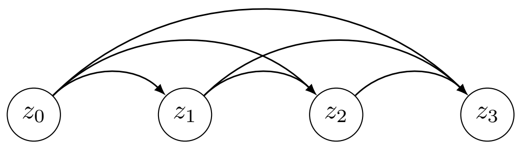
Algorithm 1 Training with Buffered Stochastic Variational Inference. We contrast training with SVI versus BSVI. We denote the stop-gradient operation with $\lceil \cdot \rceil$, reflecting that we do not backpropagate through the SVI steps.

Inputs: $\mathcal{D} = \{x^{(1)}, \ldots, x^{(n)}\}$.
for $t = 1 \ldots T$ **do**
  $x \sim \mathcal{D}$
  $\lambda_0 \leftarrow f_{\phi_t}(x)$
  for $i = 0 \ldots k$ **do**
    $z_i \sim q(z \,;\, \lambda_i)$ $\triangleright$ reparameterize as $z_{\lambda_i}(\epsilon)$
    $w(z \,;\, \lambda_i, \theta) \leftarrow p_\theta(x, z_i)/q(z_i \,;\, \lambda_i)$
    if $i < k$ **then**
      $\lambda_{i+1} \leftarrow \lceil \lambda_i + \eta \nabla_{\lambda_i} \ln w(z \,;\, \lambda_i, \theta) \rceil$
    end if
  end for
  $\phi_{t+1} \leftarrow \phi_t + \nabla_{\phi_t} \ln w(z_0 \,;\, \lambda_0, \theta_t)$
  if Train with SVI **then**
    $\theta_{t+1} \leftarrow \theta_t + \nabla_{\theta_t} \ln w(z_k \,;\, \lambda_k, \theta_t)$
  else if Train with BSVI **then**
    $\theta_{t+1} \leftarrow \theta_t + \nabla_{\theta_t} \ln \sum_i \pi_i w(z_i \,;\, \lambda_i, \theta_t)$
  end if
end for

[1] For simplicity, we use the uniform-weighting $\pi_i = 1/(k+1)$ in our base implementation of BSVI. In Section 4.1, we discuss how to optimize π during training.



(a) Dependent proposal distributions



(b) Dependent samples

Figure 2: Graphical model for dependent proposal distributions and samples. When $\lambda_{1:k}$ is marginalized, the result is a joint distribution of dependent samples. For notational simplicity, the dependency on θ is omitted.

4 Theoretical Analysis

An important consideration is whether the BSVI objective serves as a valid lower bound to the log-likelihood $\ln p_\theta(x)$. A critical challenge in the analysis of the BSVI objective is that the trajectory of variational parameters $(\lambda_0, \ldots, \lambda_k)$ is actually a sequence of *statistically-dependent random variables*. This statistical dependency is a consequence of SVI's stochastic gradient approximation in Eq. (6). We capture this dependency structure in Figure 2a, which shows that each λ_{i+1} is only deterministically generated after z_i is sampled. When the proposal distribution parameters $\lambda_{0:k}$ are marginalized, the resulting graphical model is a joint distribution over $q(z_{0:k} \mid x)$. To reason about such a joint distribution, we introduce the following generalization of the IWAE bound.

Theorem 1. *Let $p(x, z)$ be a distribution where $z \in \mathcal{Z}$. Consider a joint proposal distribution $q(z_{0:k})$ over $\mathcal{Z}^k$. Let $v(i) \subset \{0, \ldots, k\} \setminus \{i\}$ for all i, and π be a categorical distribution over $\{0, \ldots, k\}$. The following construction, which we denote the Generalized IWAE Bound, is a valid lower bound of the log-marginal-likelihood*

$$\mathbb{E}_{q(z_{0:k})} \ln \sum_{i=0}^{k} \pi_i \frac{p(x, z_i)}{q(z_i \mid z_{v(i)})} \leq \ln p(x), \qquad (10)$$

The proof follows directly from the linearity of expectation when using $q(z_{0:k})$ for importance-sampling to construct an unbiased estimate of $p_\theta(x)$, followed by application of Jensen's inequality. A detailed proof is provided in Appendix A.

Notably, if $q(z_{0:k}) = \prod_i q(z_i)$, then Theorem 1 reduces to the IWAE bound. Theorem 1 thus provides a generalization of IWAE, where the samples drawn are potentially *non-independently and non-identically* distributed. Theorem 1 thus provides a way to construct new lower bounds on the log-likelihood whenever one has access to a set of non-independent samples.

In this paper, we focus on a special instance where a chain of samples is constructed from the SVI trajectory. We note that the BSVI objective can be expressed as

$$\mathbb{E}_{q(z_{0:k} \mid x)} \ln \sum_{i=0}^{k} \pi_i w_i = \mathbb{E}_{q(z_{0:k} \mid x)} \ln \sum_{i=0}^{k} \pi_i \frac{p_\theta(x, z_i)}{q(z_i \mid z_{<i}, x)}. \qquad (11)$$

Note that since λ_i can be deterministically computed given $(x, z_{<i})$, it is therefore admissible to interchange the distributions $q(z_i \mid z_{<i}, x) = q(z_i \mid \lambda_i)$. The BSVI objective is thus a special case of the Generalized IWAE bound, where $z_{v(i)} = z_{<i}$ with auxiliary conditioning on x. Hence, the BSVI objective is a valid lower bound of $\ln p_\theta(x)$; we now refer to it as the BSVI bound where appropriate.

In the following two subsections, we address two additional aspects of the BSVI bound. First, we propose a method for ensuring that the BSVI bound is tighter than the Evidence Lower Bound achievable via SVI. Second, we provide an initial characterization of BSVIs implicit *sampling-importance-resampling* distribution.

4.1 Buffer Weight Optimization

Stochastic variational inference uses a series of gradient ascent steps to generate a final proposal distribution $q(z \mid \lambda_k)$. As evident from Figure 2a, the parameter λ_k is in fact a random variable. The ELBO achieved via SVI, *in expectation*, is thus

$$\mathbb{E}_{q(z, \lambda_k \mid x)} \ln \frac{p_\theta(x, z)}{q_\phi(z \mid \lambda_k)} = \mathbb{E}_{q(z_{0:k} \mid x)} \ln w_k, \qquad (12)$$

where the RHS re-expresses it in notation consistent with Eq. (11). We denote Eq. (12) as the SVI bound. In general, the BSVI bound with uniform-weighting $\pi_i = 1/(k+1)$ is not necessarily tighter than the SVI bound. For example, if SVI's last proposal distribution exactly matches posterior $q_k(z) = p_\theta(z \mid x)$, then assigning equal weighting to across $(w_0, \ldots w_k)$ would make the BSVI bound looser.

In practice, we observe the BSVI bound with uniform-weighting to consistently achieve a tighter lower bound than SVI's last proposal distribution. We attribute this phenomenon to the effectiveness of variance-reduction from averaging multiple importance weights—even when these importance weights are generated from dependent and non-identical proposal distributions.

To guarantee that the BSVI is tighter than the SVI bound, we propose to optimize the buffer weight π. This guarantees a tighter bound,

$$\max_{\pi} \mathbb{E}_{q(z_{0:k}|x)} \ln \sum_{i=0}^{k} \pi_i w_i \geq \mathbb{E}_{q(z_{0:k}|x)} \ln w_k, \tag{13}$$

since the SVI bound is itself a special case of the BSVI bound when $\pi = (0, \dots, 0, 1)$. It is worth noting that Eq. (13) is concave with respect to π, allowing for easy optimization of π.

Although π is a local variational parameter, we shall, for simplicity, optimize only a single global π that we update with gradient ascent throughout the course of training. As such, π is jointly optimized with θ and ϕ.

4.2 Dependence-Breaking via Double-Sampling

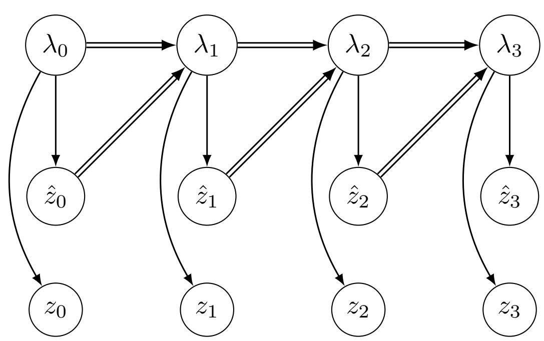


Figure 3: Graphical model for double sampling. Notice that the samples $z_{0:k}$ are now independent given $\lambda_{0:k}$ and x. Again the dependence on θ is omitted for notational simplicity.

As observed in [20], taking the gradient of the log-likelihood with respect to θ results in the expression

$$\nabla_\theta \ln p_\theta(x) = \mathbb{E}_{p_\theta(z|x)} \nabla_\theta \ln p_\theta(x, z). \tag{14}$$

We note that gradient of the ELBO with respect to θ results in a similar expression

$$\nabla_\theta \text{ELBO}(x) = \mathbb{E}_{q_\phi(z|x)} \nabla_\theta \ln p_\theta(x, z). \tag{15}$$

As such, the ELBO gradient differs from log-likelihood gradient only in terms of the distribution applied by the expectation operator. To approximate the log-likelihood gradient, we wish to set $q_\phi(z \mid x)$ close to $p_\theta(z \mid x)$ under some divergence.

We now show what results from computing the gradient of the BSVI objective.

Lemma 1. *The BSVI gradient with θ is*

$$\nabla_\theta \mathit{BSVI}(x) = \mathbb{E}_{q_{\text{sir}}(z|x)} \nabla_\theta \ln p_\theta(x, z), \tag{16}$$

where q_{sir} is a sampling-importance-resampling procedure defined by the generative process

$$z_{0:k} \sim q(z_{0:k} \mid x) \tag{17}$$
$$i \sim r(i \mid z_{0:k}) \tag{18}$$
$$z \leftarrow z_i, \tag{19}$$

and $r(i \mid z_{0:k}) = (\pi_i w_i)/(\sum_j \pi_j w_j)$ is a probability mass function over $\{0, \dots, k\}$.

A detailed proof is provided in Appendix A.

A natural question to ask is whether BSVI's q_{sir} is closer to the posterior than q_k in expectation. To assist in this analysis, we first characterize a particular instance of the Generalized IWAE bound when $(z_1, \dots, z_k)$ are independent but *non-identically distributed.*

Theorem 2. *When $q(z_{0:k}) = \prod_i q_i(z_i)$, the implicit distribution $q_{\text{sir}}(z)$ admits the inequality*

$$\mathbb{E}_{q_{\text{sir}}(z)} \ln \frac{p_\theta(x, z)}{q_{\text{sir}}(z)} \geq \mathbb{E}_{q(z_{0:k})} \ln \sum_{i=0}^{k} \pi_i w_i \tag{20}$$

$$= \mathbb{E}_{q(z_{0:k})} \ln \sum_{i=0}^{k} \pi_i \frac{p_\theta(x, z)}{q_i(z_i)}. \tag{21}$$

Theorem 2 extends the analysis by [23] from the i.i.d. case (i.e. the standard IWAE bound) to the non-identical case (proof in Appendix A). It remains an open question whether the inequality holds for the non-independent case.

Since the BSVI objective employs dependent samples, it does not fulfill the conditions for Theorem 2. To address this issue, we propose a variant, *BSVI with double-sampling* (BSVI-DS), that breaks dependency by drawing two samples at each SVI step: $\hat{z}_i$ for computing the SVI gradient update and z_i for computing the BSVI importance weight w_i. The BSVI-DS bound is thus

$$\mathbb{E}_{q(\hat{z}_{<k}|x)} \left(\mathbb{E}_{q(z_{0:k}|\hat{z}_{<k}, x)} \ln \sum_{i=0}^{k} \pi_i \frac{p_\theta(x, z)}{q(z_i \mid \hat{z}_{<k}, x)} \right), \tag{22}$$

where $q(z_{0:k} \mid \hat{z}_{<k}, x) = \prod_i q(z_i \mid \hat{z}_{<k}, x)$ is a product of independent but non-identical distributions when conditioned on $(\hat{z}_{<k}, x)$. Double-sampling now allows us to make the following comparison.

Corollary 1. *Let $q_k = q(z_k \mid \hat{z}_{<i}, x)$ denote the proposal distribution found by SVI. For any choice of*

$(\hat{z}_{<i}, x)$, the distribution q_{sir} implied by BSVI-DS (with optimal weighting π^*) is at least as close to $p_\theta(z \mid x)$ as q_k,

$$D(q_{\text{sir}} \parallel p_\theta(z \mid x)) \leq D(q_k \parallel p_\theta(z \mid x)), \tag{23}$$

as measured by the Kullback-Leibler divergence.

Corollary 1 follows from Theorem 2 and that the BSVI-DS bound under optimal π^* is no worse than the SVI bound. Although the double-sampling procedure seems necessary for inequality in Corollary 1 to hold, in practice we do not observe any appreciable difference between BSVI and BSVI-DS.

5 Computational Considerations

Another important consideration is the speed of training the generative model with BSVI versus SVI. Since BSVI reuses the trajectory of weights $(w_0, \dots, w_k)$ generated by SVI, the forward pass incurs the same cost. The backwards pass for BSVI, however, is $O(k)$ for k SVI steps—in contrast to SVI's $O(1)$ cost. To make the cost of BSVI's backwards pass $O(1)$, we highlight a similar observation from the original IWAE study [21] that the gradient can be approximated via Monte Carlo sampling

$$\nabla_\theta \text{BSVI}(x) \approx \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \ln p_\theta(x, z^{(i)}), \tag{24}$$

where $z^{(i)}$ is sampled from BSVI's implicit distribution $q_{\text{sir}}(z \mid x)$. We denote this as training *BSVI with sample-importance-resampling* (BSVI-SIR). Setting $m = 1$ allows variational autoencoder training with BSVI-SIR to have the same wall-clock speed as training with SVI.

6 Experiments

6.1 Setup

We evaluated the performance of our method by training variational autoencoders with BSVI-SIR with buffer weight optimization (BSVI-SIR-π)) on the dynamically-binarized Omniglot, grayscale SVHN datasets, and FashionMNIST (a complete evaluation of all BSVI variants is available in Appendix B). Our main comparison is against the SVI training procedure (as described in Algorithm 1). We also show the performance of the standard VAE and IWAE training procedures. Importantly, we note that we have chosen to compare SVI-k' and IWAE-k' trained with $k' = 10$ against BSVI-k-SIR trained with $k = 9$ SVI steps. This is because that BSVI-k-SIR generates $k+1$ importance weights.

For all our experiments, we use the same architecture as [18] (where the decoder is a PixelCNN) and train with the AMSGrad optimizer [24]. For grayscale SVHN, we follow [25] and replaced [18]'s bernoulli observation model with a discretized logistic distribution model with a global scale parameter. Each model was trained for up to 200k steps with early-stopping based on validation set performance. For the Omniglot experiment, we followed the training procedure in [18] and annealed the KL term multiplier [2, 26] during the first 5000 iterations. We replicated all experiments four times and report the mean and standard deviation of all relevant metrics. For additional details, refer to Appendix D

6.2 Log-Likelihood Performance

For all models, we report the log-likelihood (as measured by BSVI-500). We additionally report the SVI-500 (ELBO*) bound along with its decomposition into rate (KL*) and distortion (Reconstruction*) components [27]. We highlight that KL* provides a *fair* comparison of the rate achieved by each model without concern of misrepresentation caused by the amortized inference suboptimality.

Omniglot. Table 1 shows that BSVI-SIR outperforms SVI on the test set log-likelihood. BSVI-SIR also makes greater usage of the latent space (as measured by the lower Reconstruction*). Interestingly, BSVI-SIR's log-likelihoods are noticeably higher than its corresponding ELBO*, suggesting that BSVI-SIR has learned posterior distributions not easily approximated by the Gaussian variational family when trained on Omniglot.

SVHN. Table 2 shows that BSVI-SIR outperforms SVI on test set log-likelihood. We observe that both BSVI-SIR and SVI significantly outperform both VAE and IWAE on log-likelihood, ELBO*, and Reconstruction*, demonstrating the efficacy of iteratively refining the proposal distributions found by amortized inference model during training.

FashionMNIST. Table 3 similarly show that BSVI-SIR outperforms SVI on test set log-likelihood. Here, BSVI achieves significantly better Reconstruction* as well as achieving higher ELBO* compared to VAE, IWAE, and SVI.

In Tables 4 to 6 (Appendix B), we also observe that the use of double sampling and buffer weight optimization does not make an appreciable difference than their appropriate counterparts, demonstrating the efficacy of BSVI even when the samples $(z_{0:k})$ are statistically dependent and the buffer weight is simply uniform.

Table 1: Test set performance on the Omniglot dataset. Note that $k = 9$ and $k' = 10$ (see Section 6.1). We approximate the log-likelihood with BSVI-500 bound (Appendix C). We additionally report the SVI-500 bound (denoted ELBO*) along with its KL and reconstruction decomposition.

Model	**Log-likelihood**	**ELBO***	**KL***	**Reconstruction***
VAE	-89.83 ± 0.03	-89.88 ± 0.02	0.97 ± 0.13	88.91 ± 0.15
IWAE-k'	-89.02 ± 0.05	-89.89 ± 0.06	4.02 ± 0.18	85.87 ± 0.15
SVI-k'	-89.65 ± 0.06	**-89.73 ± 0.05**	1.37 ± 0.15	88.36 ± 0.20
BSVI-k-SIR	**-88.80 ± 0.03**	-90.24 ± 0.06	7.52 ± 0.21	**82.72 ± 0.22**

Table 2: Test set performance on the grayscale SVHN dataset.

Model	**Log-likelihood**	**ELBO***	**KL***	**Reconstruction***
VAE	-2202.90 ± 14.95	-2203.01 ± 14.96	0.40 ± 0.07	2202.62 ± 14.96
IWAE-k'	-2148.67 ± 10.11	-2153.69 ± 10.94	2.03 ± 0.08	2151.66 ± 10.86
SVI-k'	-2074.43 ± 10.46	-2079.26 ± 9.99	45.28 ± 5.01	2033.98 ± 13.38
BSVI-k-SIR	**-2059.62 ± 3.54**	**-2066.12 ± 3.63**	51.24 ± 5.03	**2014.88 ± 5.30**

Table 3: Test set performance on the FashionMNIST dataset.

Model	**Log-likelihood**	**ELBO***	**KL***	**Reconstruction***
VAE	-1733.86 ± 0.84	-1736.49 ± 0.73	11.62 ± 1.01	1724.87 ± 1.70
IWAE-k'	-1705.28 ± 0.66	-1710.11 ± 0.72	33.04 ± 0.36	1677.08 ± 0.70
SVI-k'	-1710.15 ± 2.51	-1718.39 ± 2.13	26.05 ± 1.90	1692.34 ± 4.03
BSVI-k-SIR	**-1699.44 ± 0.45**	**-1707.00 ± 0.49**	41.48 ± 0.12	**1665.52 ± 0.41**

6.3 Stochastic Gradient as Regularizer

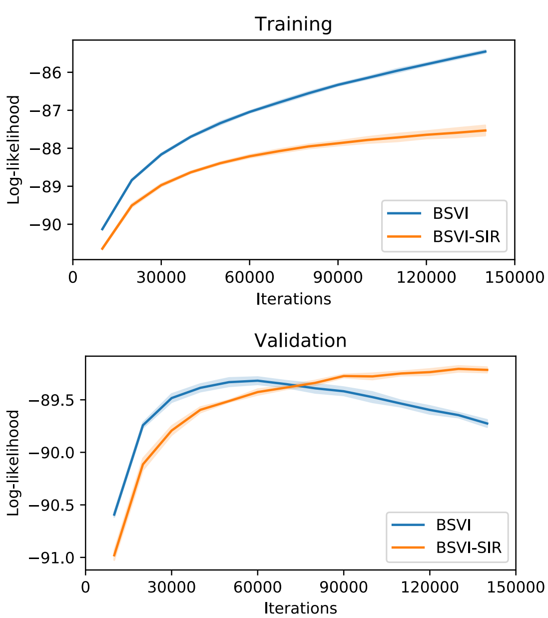


Figure 4: Performance comparison between BSVI and BSVI-SIR on training (top) and validation (bottom) sets for Omniglot. Although BSVI achieves lower training loss, BSVI-SIR avoids overfitting and performs better on the test set.

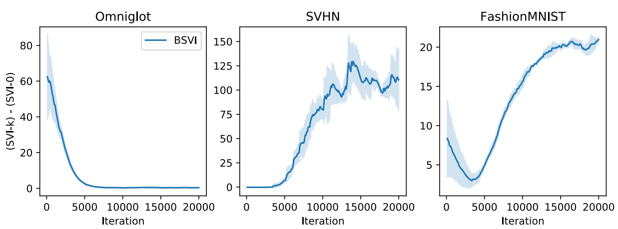
Interestingly, Table 4 shows that BSVI-SIR can outperform BSVI on the test set despite having a higher variance gradient. We show in Figure 4 that this is the result of BSVI overfitting the training set. The results demonstrate the regularizing effect of having noisier gradients and thus provide informative empirical evidence to the on-going discussion about the relationship between generalization and the gradient signal-to-noise ratio in variational autoencoders [28, 16].

6.4 Latent Space Visualization

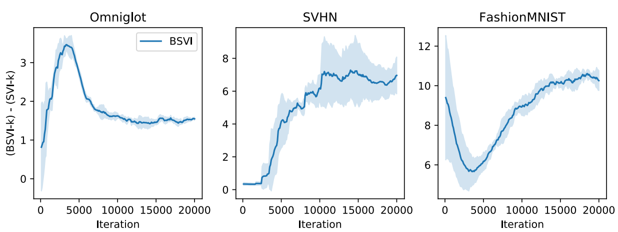
Figure 5: Visualization of images sampled from decoder trained using SVI (top) and BSVI-SIR (bottom). Each row represents a different z sampled from the prior. Conditioned on z, 20 images $x^{(1:20)} \sim p_\theta(x \mid z)$ are then sampled from the PixelCNN decoder.

Table 1 shows that the model learned by BSVI-SIR training has better Reconstruction* than SVI, indicating greater usage of the latent variable for encoding information about the input image. We provide a visualization of the difference in latent space usage in Figure 5. Here, we sample multiple images conditioned on a fixed z. Since BSVI encoded more information into z than SVI on the Omniglot dataset, we see that the conditional distribution $p_\theta(x \mid z)$ of the model learned by BSVI has lower entropy (i.e. less diverse) than SVI.

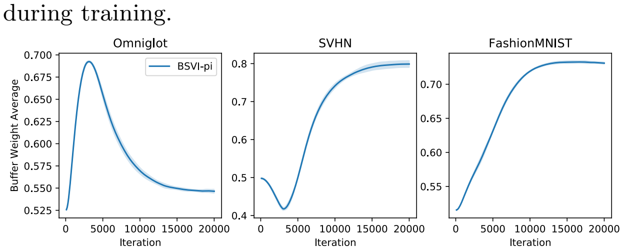
6.5 Analysis of Training Metrics



(a) Difference between lower bounds achieved by q_k (SVI-k) and q_0 (SVI-0) during training.



(b) Difference between the BSVI-k bound and SVI-k bound during training.



(c) Plot of the buffer weight average (defined as $\mathbb{E}_{\pi(i)} i/k$) during training when the buffer weight is optimized.

Figure 6: Plots of metrics during BSVI-k training, where $k = 9$. Since BSVI-k uses SVI-k as a subroutine, it is easy to check how the BSVI-k bound compares against the SVI-k and the amortized ELBO (SVI-0) bounds on a random mini-batch at every iteration during training.

Recall that the BSVI-k training procedure runs SVI-k as a subroutine, and therefore generates the trajectory of importance weights $(w_0, \ldots, w_k)$. Note that $\ln w_0$ and $\ln w_k$ are unbiased estimates of the ELBO achieved by the proposal distribution q_0 (SVI-0 bound) and q_k (SVI-k bound) respectively. It is thus possible to monitor the health of the BSVI training procedure by checking whether the bounds adhere to the ordering

$$\text{BSVI-}k \geq \text{SVI-}k \geq \text{SVI-0} \tag{25}$$

in expectation. Figures 6a and 6b show that this is indeed the case. Since Omniglot was trained with KL-annealing [18], we see in Figure 6a that SVI plays a negligible role once the warm-up phase (first 5000 iterations) is over. In contrast, SVI plays an increasingly large role when training on the more complex SVHN and FashionMNIST datasets, demonstrating that the amortization gap is a significantly bigger issue in the generative modeling of SVHN and FashionMNIST. Figure 6b further shows that BSVI-k consistently achieves a better bound than SVI-k. When the buffer weight is also optimized, we see in Figure 6c that π learns to upweight the later proposal distributions in $(q_0, \ldots, q_k)$, as measured by the buffer weight average $\mathbb{E}_{\pi(i)} i/k$. For SVHN, the significant improvement of SVI-k over SVI-0 results in π being biased significantly toward the later proposal distributions. Interestingly, although Figure 6c suggests that the optimal buffer weight π^* can differ significantly from naive uniform-weighting, we see from Tables 1 and 2 that buffer weight optimization has a negligible effect on the overall model performance.

7 Conclusion

In this paper, we proposed *Buffered Stochastic Variational Inference* (BSVI), a novel way to leverage the intermediate importance weights generated by stochastic variational inference. We showed that BSVI is effective at alleviating inference suboptimality and that training variational autoencoders with BSVI consistently outperforms its SVI counterpart, making BSVI an attractive and simple drop-in replacement for models that employ SVI. One promising line of future work is to extend the BSVI training procedure with end-to-end learning approaches in [18, 19]. Additionally, we showed that BSVI procedure is a valid lower bound and belongs to general class of importance-weighted (Generalized IWAE) bounds where the importance weights are statistically dependent. Thus, it would be of interest to study the implications of this bound for certain MCMC procedures such as Annealed Importance Sampling [29] and others.

Acknowledgements

We would like to thank Matthew D. Hoffman for his insightful comments and discussions during this project. This research was supported by NSF (#1651565, #1522054, #1733686), ONR (N00014-19-1-2145), AFOSR (FA9550-19-1-0024), and FLI.

References

[1] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[2] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder Variational Autoencoders. In *Advances In Neural Information Processing Systems*, pages 3738–3746, 2016.

[3] Rui Shu, Hung H Bui, and Mohammad Ghavamzadeh. Bottleneck conditional density estimation. *International Conference on Machine Learning*, 2017.

[4] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-Supervised Learning With Deep Generative Models. In *Advances In Neural Information Processing Systems*, pages 3581–3589, 2014.

[5] Volodymyr Kuleshov and Stefano Ermon. Deep hybrid models: Bridging discriminative and generative approaches. UAI, 2017.

[6] Tian Qi Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating Sources Of Disentanglement In Variational Autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

[7] Manuel Watter, Jost Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in neural information processing systems*, pages 2746–2754, 2015.

[8] Ershad Banijamali, Rui Shu, Mohammad Ghavamzadeh, Hung Bui, and Ali Ghodsi. Robust locally-linear controllable embedding. *Artificial Intelligence And Statistics*, 2018.

[9] Yunzhu Li, Jiaming Song, and Stefano Ermon. Infogail: Interpretable imitation learning from visual demonstrations. In *Advances in Neural Information Processing Systems*, pages 3812–3822, 2017.

[10] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation And Approximate Inference In Deep Generative Models. *arXiv preprint arXiv:1401.4082*, 2014.

[11] Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

[12] Shengjia Zhao, Jiaming Song, and Stefano Ermon. A lagrangian perspective on latent variable generative models. In *Proc. 34th Conference on Uncertainty in Artificial Intelligence*, 2018.

[13] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems*, pages 2352–2360, 2016.

[14] Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. Pixelvae: A latent variable model for natural images. *arXiv preprint arXiv:1611.05013*, 2016.

[15] Chris Cremer, Xuechen Li, and David Duvenaud. Inference Suboptimality In Variational Autoencoders. *arXiv preprint arXiv:1801.03558*, 2018.

[16] Rui Shu, Hung H Bui, Shengjia Zhao, Mykel J Kochenderfer, and Stefano Ermon. Amortized inference regularization. *Advances in Neural Information Processing Systems*, 2018.

[17] Rahul G Krishnan, Dawen Liang, and Matthew Hoffman. On the challenges of learning with inference networks on sparse, high-dimensional data. *arXiv preprint arXiv:1710.06085*, 2017.

[18] Yoon Kim, Sam Wiseman, Andrew C Miller, David Sontag, and Alexander M Rush. Semi-Amortized Variational Autoencoders. *arXiv preprint arXiv:1802.02550*, 2018.

[19] Joseph Marino, Yisong Yue, and Stephan Mandt. Iterative amortized inference. *arXiv preprint arXiv:1807.09356*, 2018.

[20] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic Variational Inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[21] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance Weighted Autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[22] Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. In *Advances in Neural Information Processing Systems*, pages 4475–4484, 2018.

[23] Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting Importance-Weighted Autoencoders. *arXiv preprint arXiv:1704.02916*, 2017.

[24] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[25] Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling.

Improved Variational Inference With Inverse Autoregressive Flow. In *Advances In Neural Information Processing Systems*, pages 4743–4751, 2016.

[26] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

[27] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pages 159–168, 2018.

[28] Tom Rainforth, Adam R Kosiorek, Tuan Anh Le, Chris J Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter Variational Bounds Are Not Necessarily Better. *arXiv preprint arXiv:1802.04537*, 2018.

[29] Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

[30] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *CoRR*, abs/1701.05517, 2017.

[31] Jakub M Tomczak and Max Welling. VAE With A Vampprior. *arXiv preprint arXiv:1705.07120*, 2017.