Variational Inference and Model Selection with Generalized Evidence Bounds

Chenyang Tao * **Liqun Chen** * **Ruiyi Zhang** **Ricardo Henao** **Lawrence Carin**

Abstract

Recent advances on the scalability and flexibility of variational inference have made it successful at unravelling hidden patterns in complex data. In this work we propose a new variational bound formulation, yielding an estimator that extends beyond the conventional variational bound. It naturally subsumes the importance-weighted and Rényi bounds as special cases, and it is provably sharper than these counterparts. We also present an improved estimator for variational learning, and advocate a novel high signal-to-variance ratio update rule for the variational parameters. We discuss model-selection issues associated with existing evidence-lower-bound-based variational inference procedures, and show how to leverage the flexibility of our new formulation to address them. Empirical evidence is provided to validate our claims.

1. Introduction

One of the key challenges in modern machine learning is to approximate complex distributions. Due to recent advances on learning scalability (Hoffman et al., 2013) and flexibility (Kingma et al., 2016), and the development of automated inference procedures (Ranganath et al., 2014), variational inference (VI) has become a popular approach for general latent variable models (Blei et al., 2017). Variational inference leverages a posterior approximation to derive a lower bound on the log-evidence of the observed samples, which can be efficiently optimized. This bound, commonly known as the evidence lower bound (ELBO), serves as a surrogate objective for maximum likelihood estimation (MLE) of the model parameters. Successful applications of VI have been reported in document analysis (Blei et al., 2003), neuroscience (Friston et al., 2007), generative modeling (Kingma & Welling, 2014), among many others.

It has been widely recognized that tightening the variational bound, in general, significantly improves model performance. Consequently, considerable research has been directed toward this goal. The most direct approach seeks to boost the expressive power of the approximate posterior. Normalizing flows (Rezende & Mohamed, 2015; Kingma et al., 2016) exploited invertible transformations on latent codes in latent variable models, Ranganath et al. (2016) and Gregor et al. (2015) explored the hierarchical structure of the latent code generation, and Miller et al. (2016) modeled the posterior as a mixture of Gaussians. Adversarial variational Bayes (Mescheder et al., 2017) employed a neural generator to produce posterior samples, and further leveraged a density ratio estimator to compute the ELBO. Notably, the matching between the true and approximate posterior can be made implicit (Pu et al., 2017a) via an application of Stein's lemma (Liu & Wang, 2016).

An alternative direction seeks a modification of the variational objective. The importance-weighted autoencoder (Burda et al., 2016) showed that the bound can be sharpened by leveraging multiply-weighted posterior samples. Further, χ-VI (Dieng et al., 2017), also known as Rényi-VI (Li & Turner, 2016), derived an alternative bound that is sharper than the ELBO. More generally, a sandwich formula holds for the χ bound, thus tightening this gap improves performance. Bamler et al. (2017) developed a more general view on variational bounds and presented a low-variance estimator based on a perturbative argument. It is important to note that sharpening the variational bound may unexpectedly hurt learning of the inference arm of the model (approximate posterior) (Rainforth et al., 2017), thereby compromising performance.

While most studies have focused on the scalability and flexibility of VI, a less-studied issue is that different models can be equally plausible in terms of of the mean evidence lower bound wrt a finite number of samples (Blei et al., 2017). This is a fundamental problem *inherited* by VI, associated with doing an empirical estimation of the expected log like-

*Equal contribution Affiliation: Electrical & Computer Engineering, Duke University, Durham, NC 27708, USA. Correspondence to: Chenyang Tao <chenyang.tao@duke.edu>, Liqun Chen <liqun.chen@duke.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

lihood. To make this more clear, recall that variational inference optimizes a lower bound to the expected log-evidence $\mathbb{E}_{X\sim p_d(x)}[\log p_\alpha(X)]$, where $p_d(x)$ is the unknown data distribution and distribution $p_\alpha(x)$ is our model parameterized by α. Consider a set of samples $\mathcal{D} = \{x_i\}_{i=1}^n$ drawn from a ground truth model $p_d(x)$. A model $p_\alpha(x)$ can achieve the same empirical expected log-evidence score as that of $p_d(x)$ by yielding lower log-evidence score on an underfitted subset of samples $\{x_i; p_\alpha(x_i) < p_d(x_i)\}$, while compensating its losses on another subset of overfitted samples $\{x_i; p_\alpha(x_i) > p_d(x_i)\}$. Neither underfitting nor overfitting are desirable when learning a probabilistic representation of the data. However, such behavior is not explicitly penalized by the standard variational objective, and each may be a consequence of approximating $\mathbb{E}_{X\sim p_d(x)}[\log p_\alpha(X)]$ with the finite det of samples in $\mathcal{D}$. Consequently, when doing MLE or VI with a flexible model family on finite samples, the maximizer may not be unique (Mäkeläinen et al., 1981; Dharmadhikari & Joag-Dev, 1985).

This paper addresses model selection issues in variational inference. Our key contributions are: (i) An extension of the concept of evidence score and its use as a new model-selection criterion. (ii) Derivation of novel importance-weighted evidence bounds, and proof of their theoretical properties. (iii) Development of a novel variational inference procedure that favors more plausible models, compared to existing VI-based approaches. (iv) A new low bias-variance variational estimator with an improved update rule for optimization.

2. ELBO and Variational Inference

Let $p_d(x)$ be the true and *unknown* data-generating distribution, which we seek to model with $p_\alpha(x) = \int p_\alpha(x, z)dz$ where $p_\alpha(x, z) = p_\alpha(x|z)p(z)$. Here $z \in \mathbb{R}^d$ represents latent variables responsible for $x \in \mathbb{R}^p$, $p(z)$ is a specified prior on z, and $p_\alpha(x|z)$ is the conditional distribution of the data with model parameters α. The joint distribution may also be expressed $p_\alpha(x, z) = p_\alpha(x)p_\alpha(z|x)$, where $p_\alpha(z|x)$ is the model conditional distribution of latent z given x. The posterior $p_\alpha(z|x)$ is typically difficult to compute, and therefore in variational inference it is approximated by $q_\beta(z|x)$, a distribution with parameters β. The evidence lower bound (ELBO) is defined

$$\text{ELBO}(p_\alpha(x, z), q_\beta(z|x)) = \mathbb{E}_{q_\beta(z|x)} \log \big[\frac{p_\alpha(x, z)}{q_\beta(z|x)}\big]. \quad (1)$$

It is well known that $\text{ELBO}(p_\alpha(x, z), q_\beta(z|x)) = \log p_\alpha(x) - \text{KL}(q_\beta(z|x)\|p_\alpha(z|x)) \leq \log p_\alpha(x)$, where $\text{KL}(p\|q)$ is the Kullback-Leibler divergence between distributions p and q. Hence, the ELBO, characterized by cumulative parameters $\theta = (\alpha, \beta)$, serves as a lower bound on evidence $\log p_\alpha(x)$.

When learning, we seek to maximize the ELBO wrt parameters θ. One may also readily show that $\text{KL}(p_d\|p_\alpha) = -h(p_d) - \mathbb{E}_{x\sim p_d}[\text{KL}(q_\beta(z|x)\|p_\alpha(z|x))] - \mathbb{E}_{x\sim p_d}[\text{ELBO}(p_\alpha(x, z), q_\beta(z|x))]$, where the differential entropy $h(p_d)$ is an unknown constant. Hence, minimization of $\text{KL}(p_\alpha\|p_d)$ corresponds to maximizing $\mathbb{E}_{x\sim p_d}[\text{ELBO}(p_\alpha(x, z), q_\beta(z|x))]$, which has the commensurate goal of pushing $\mathbb{E}_{x\sim p_d}[\text{KL}(q_\beta(z|x)\|p_\alpha(z|x))] \to 0$. In practice, variational inference (VI) learns the "best" model by optimizing θ wrt the expected ELBO.

A variational model is defined by $p_\alpha(x, z)$ and $q_\beta(z|x)$, with $\theta = (\alpha, \beta)$. For notational convenience, the corresponding model is denoted $\mathcal{M}_\theta$. In the context of variational auto-encoder, α and β are also respectively known as the generator and inference parameters.

3. Generalized Evidence Bound

To motivate our formal development below, we first provide some intuitions. Our key observation is that existing bounds are almost exclusively based on the Jensen inequality, which implies the variational gap can be improved with a *less convex* transform. On the other hand, it is desirable to prioritize underfitted samples when adjusting the model. We now describe a principled framework to address these two points.

Let $\phi(u) : \mathbb{R}^+ \to \mathbb{R}$ be a *non-decreasing* function defined on the non-negative real line, referred to as the *evidence function*. Further, assume (i) $\phi(u)$ is concave, (ii) $\psi(u)$ is a convex and monotonically increasing function, and (iii) $h(u) \triangleq \psi(\phi(u))$ is concave. For notational clarity, we omit dependence on $\phi(u)$, $\psi(u)$, $p_\alpha(x, z)$ and $q_\beta(z|x)$ when the context is clear. We will refer to $\phi(p_\alpha(x))$ as the ϕ-evidence score of sample x wrt model $p_\alpha(x)$.

Definition 1. *The K-sample Generalized Evidence Lower Bound (GLBO) is defined as*

$$\textit{GLBO}(x; K) \triangleq \psi^{-1}\left(\mathbb{E}_{Z_{1:K}\sim q_\beta}\left[h\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p_\alpha(x, Z_k)}{q_\beta(Z_k|x)}\right)\right]\right), \quad (2)$$

where $Z_{1:K} = \{Z_k\}_{k=1}^K$ are K iid samples from $q_\beta(z|x)$.

Note that (2) is closely related to importance sampling (Liu, 2008), where the approximate posterior $q_\beta(z|x)$ is understood as the proposal distribution and the term $\frac{1}{K}\sum_{k=1}^{K}\frac{p_\alpha(x,Z_k)}{q_\beta(Z_k|x)}$ is the K-sample importance-weighted estimate of $p_\alpha(x)$. When $q_\beta(z|x)$ equals $p_\alpha(z|x)$, the GLBO exactly recovers the ϕ-evidence score.

Concerning intuitions for the above assumptions, the concavity of $\phi(u)$ in (i) reflects that, in general, we want to use a $\phi(u)$ that is monotonically increasing. Importantly, we also want $\phi(u)$ to saturate for large values of u, to reduce the influence of the high-evidence region (well-fit samples)

in the objective, allowing the model to focus on less-well-fit samples (the saturation also minimizes the desire to over-fit when learning). Assumption (ii) from above introduces a convex auxiliary function $\psi(u)$, and (iii) states that the concavity of $\phi(u)$ dominates the convexity of $\psi(u)$. As discussed below, the additional convexity from $\psi(u)$ generally improves our variational bound.

Theorem 2. *Under assumptions (i)-(iii),*

$$GLBO(x;1) \leq GLBO(x;2) \leq \cdots \overset{K\to\infty}{\longrightarrow} \phi(p_\alpha(x)).$$

All proofs are provided in the Supplementary Material (SM). We denote

$$\begin{aligned}\text{NLBO}(x;K) &\triangleq \text{GLBO}(x;\psi(u)=u)\\ &= \mathbb{E}_{Z_{1:K}\sim q_\beta}\left[\phi\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p_\alpha(x,Z_k)}{q_\beta(Z_k|x)}\right)\right],\end{aligned}$$

as the *naïve bound* for the K-sample ϕ-evidence score. The following theorem shows the concavity introduced via $\psi(u)$ in GLBO improves the NLBO bound.

Theorem 3. *Under assumptions (i)-(iii),*

$$NLBO(x;K) \leq GLBO(x;K).$$

As a particular case, recall the K-sample importance-weighted log-evidence lower bound (ELBO) is defined as

$$\begin{aligned}&\text{ELBO}(x;K) \triangleq\\ &\qquad \mathbb{E}_{Z_{1:K}\sim q_\beta}\left[\log\left(\tfrac{1}{K}\textstyle\sum_k \frac{p_\alpha(x,Z_k)}{q_\beta(Z_k|x)}\right)\right],\end{aligned} \tag{3}$$

which is the variational objective proposed by Burda et al. (2016). The K-sample GLBO in (2) recovers the importance-weighted ELBO in (3), when $(\psi(u),\phi(u)) \to (u,\log(u))$. From Theorems 2 and 3, GLBO is generally a tighter bound relative to the importance-weighted ELBO. This is formalized in the following corollary.

Corollary 4. *Under assumptions (i)-(iii),*

$$ELBO(x;K) \leq GLBO(x;K,\phi=\log(u)) \leq \log p_\alpha(x).$$

Note this is for the special case where $\phi = \log(u)$.

3.1. χ-evidence bounds

We now consider a few concrete examples. Letting $T \geq 1$ be a temperature parameter, we define the K-sample χ-Evidence Lower Bound (CLBO) to be

$$\begin{aligned}&\text{CLBO}(x;K,T) \triangleq\\ &\qquad \text{GLBO}(x;K,\phi=\log(u),\psi=\exp(T^{-1}u)).\end{aligned} \tag{4}$$

CLBO recovers the χ-evidence bound (Dieng et al., 2017), or Rényi variational bound (RVB) (Li & Turner, 2016) when $K=1$. Further, our K-sample CLBO is superior to the K-sample bound used in (Dieng et al., 2017; Li & Turner, 2016). Specifically, our lower bound is guaranteed to be sharper than RVB (see Section 5.1, Theorem 8), and our upper bound is guaranteed to be an upper bound (see SM), while for RVB this property only holds in the asymptotic limit. See additional discussion in Section 5.1 and experimental results in Figure 1.

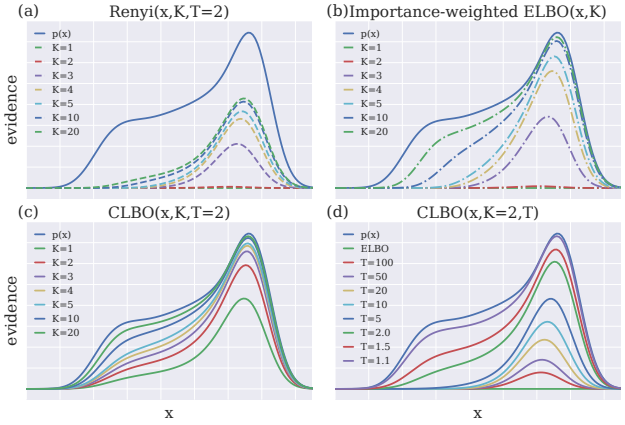

Figure 1: Comparison of theoretical bounds on a toy distribution. We compare the results in the original scale of $p(x)$ so all evidence bounds are exponentially transformed, see Section 7.1 for details.

Stronger results can be established for the CLBO in (4). The following theorem proves that $\text{CLBO}(x;K,T)$ is non-increasing wrt T.

Theorem 5. *Let $1 \leq T_1 \leq T_2$, then*

$$CLBO(x;K,T_2) \leq CLBO(x;K,T_1).$$

While in theory as $T \to 1$ the bound gets sharper, we note that in practice the empirical estimator becomes more unstable as the bound gets sharper. See SM for details on the effects of T related to empirical performance. We further establish asymptotic results for CLBO, as follows.

Theorem 6. *The following asymptotic results hold for CLBO:*

$$\begin{aligned}\lim_{T\to 1} CLBO(x;K,T) &\to \log p_\alpha(x),\\ \lim_{T\to\infty} CLBO(x;K,T) &\to ELBO(x;K).\end{aligned}$$

This implies that asymptotically,

$$\text{ELBO}(x;K) \leq \text{CLBO}(x;K,T) \leq \log p_\alpha(x),$$

for $T \in (1,\infty)$. Further, for $K=1$ it can be shown that

Corollary 7. *When T is sufficiently large,*

$$CLBO(x;1,T) \approx ELBO + \frac{1}{2T} var[f(x,y)],$$

where $f(x,z) = \log p_\alpha(x,z) - \log q_\beta(z|x)$.

4. Model Selection with ϕ-evidence Score

Conventional VI picks a variational model $\mathcal{M}_\theta$ that maximizes the expected ELBO wrt data. As discussed in the

Introduction, when choosing from a flexible family of variational models, the maximizer may not be unique. Therefore, we need to be more specific about what is a good model to select from these candidate models, which all maximize the variational objective.

A straightforward strategy is to employ the minimax criteria: select a model $\mathcal{M}_\theta^*$ that has highest value for its worst evidence bound wrt the data samples, *i.e.*

$$\mathcal{M}_\theta^* = \arg\max_{\mathcal{M}_\theta \in \mathcal{C}} \{\min_{x \in \mathcal{D}} \{\text{ELBO}(p_\alpha(x,z), q_\beta(z|x))\}\}, \quad (5)$$

where $\mathcal{C}$ denotes the collection of all models that maximizes the variational objective. Intuitively, this ensures that the selected model gives a reasonable explanation for the sample least consistent with it.

Unfortunately (5) does not readily translate into a differentiable objective wrt the variational parameters θ, and therefore cannot be directly optimized within stochastic gradient descent framework. Instead, we can relax (5) by reweighting the data with a weight function $w(x)$, and optimize the ELBO wrt the weighted distribution. Following the spirit of minimax criteria, we want our model to improve its fit on the low evidence (underfitted) samples, and thus we put larger weights on those low evidence samples.

Now we show optimizing a reweighted ELBO objective is equivalent to optimizing GLBO, with the weighting strategy implicitly implied by $\phi(u)$. First consider the gradient of ϕ-evidence wrt model parameters α

$$\nabla_\alpha \phi(p_\alpha(x)) = \underbrace{\phi'(p_\alpha(x))p_\alpha(x)}_{(a)} \underbrace{\nabla_\alpha \log p_\alpha(x)}_{(b)},$$

where $\phi'(u)$ denotes the derivative of $\phi(u)$. Here term (a) can be identified as the weight $w_\alpha(x)$ to term (b), the gradient of the log-evidence. So each gradient update can be considered as an infinitesimal attempt to improve match wrt the reweighted data distribution $\tilde{p}_d(x) \propto w_\alpha(x)p_d(x)$, where the weight is determined by the current model $p_\alpha(x)$.

To assign larger weights to the low evidence samples, we can specify a $\phi(u)$ with faster growth rate in the low evidence region and saturation in the high evidence region, which we call a saturating ϕ-evidence function. Figure 2 compares the standard log-evidence function with an example saturating ϕ-evidence function (on the log-scale). With a saturating score function, the model update strives to improve its fit on the samples that are less consistent with the current model. This also helps to prevent the optimization from entering a state of greedy improvement of already well-fitted samples, that may result in overfitting, thus in overoptimistic evidence scores.

Note that optimizing the GLBO objective with an evidence function other than $\log(u)$ is framed as a model regularizer, rather than a primary objective for model fitting. The

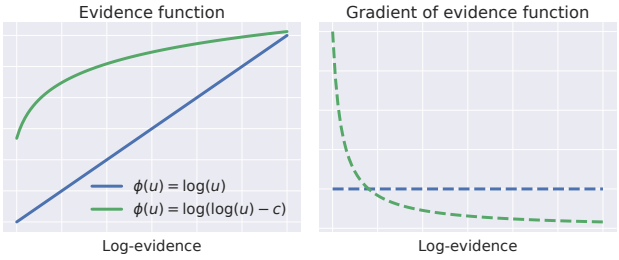

Figure 2: Model selection with ϕ-evidence score.

GLBO does not compromise the primary objective, such as maximizing the expected log-evidence bound, in the model selection phase. Second, a saturating $\phi(u)$ encourages a low variance evidence distribution. This closely connects to the maximal entropy principle for model selection (Jaynes, 2003), and we provide an informal argument in the SM to support this view.

5. Related work

χ^2 / Rényi variational inference The Rényi variational bound proposed by Li & Turner (2016) is a special case of the GLBO. The authors also investigated a K-sample importance-weighted variational objective of the form

$$\text{RVB}(x; K, T) \triangleq \mathbb{E}_{Z_{1:K} \sim q} \left[T \log \left(\frac{1}{K} \sum_{k=1}^{K} \left(\frac{p_\alpha(x, Z_k)}{q_\beta(Z_k|x)} \right)^{1/T} \right) \right]. \quad (6)$$

However, $\text{RVB}(x; K, T)$ is problematic because: (i) it is a loose lower bound when $T > 1$, (ii) when $T < 1$ it approaches the upper bound from below as K grows, which means that it may not hold as upper bound until K is sufficiently large (see Figure 2(a) of Li & Turner (2016)). In fact, Li & Turner (2016) maximized a K-sample estimate of an upper bound in their experiments, while with their particular choice of K, the upper bound estimator turns out to be a lower bound. The following theorem states that our CLBO is guaranteed to be sharper than RVB.

Theorem 8. *When $T \geq 1$,*

$$\textit{RVB}(x; K, T) \leq \textit{CLBO}(x; K, T) \leq \log p_\alpha(x).$$

Webb & Teh (2016) hypothesized a better form of importance-weighted estimator for the Rényi variational bound in (6) and validated their hypothesis with some empirical experiments. However, they were unable to provide a theoretical justification for their estimator, thus they left it as future work, which we address here.

Motivated by the issue of the posterior variance underestimation suffered by ELBO-based VI procedures, Dieng et al. (2017) proposed to minimize the variational upper

bound rather than the lower bound, which in turn over estimates the posterior variance. The authors focused on the χ^2 variational bound, a special case of GLBO's upper bound (see the SM for more details). They proved the equivalence between the χ^2 variational upper bound minimization and the minimization of χ^2-divergence between the true and approximate posteriors. However, we note that optimization of variational upper bounds is considerably more numerically unstable relative to its lower bound counterpart. Consequently, χ^2-VI is more appropriate for relatively simple problems. The estimator they proposed is essentially the exponential of the estimator used in Rényi variational inference (Li & Turner, 2016). However, this estimator cannot be used to construct a variational auto-encoder. A minimal variance argument was made to establish a connection to importance sampling. However, Dieng et al. (2017) did not consider using importance sampling technique to improve their estimator.

Efficiency of importance-weighted VI Rainforth et al. (2017) recently analyzed the trade-off of using an importance-weighted estimator in VI. In particular, the authors considered the signal-to-noise ratio (SNR) of a gradient estimator as defined by

$$\text{SNR}(\hat{\nabla}\ell(\theta)) \triangleq \frac{\mathbb{E}[\hat{\nabla}\ell(\theta)]}{\sqrt{\text{var}(\hat{\nabla}\ell(\theta))}}, \tag{7}$$

where $\hat{\nabla}\ell(\theta)$ is the gradient of the variational objective $\ell(\theta)$ wrt θ, and $\mathbb{E}[\cdot]$ and $\text{var}(\cdot)$ are approximated with samples. Rainforth et al. (2017) showed that (7) converges with rates $\mathcal{O}(\sqrt{K})$ and $\mathcal{O}(1/\sqrt{K})$ for the generator parameters α and inference parameters β, respectively. This raises the concern that the gains in the improved bound by increasing K may not be feasible due to unstable updates for β.

While also using an importance-weighted estimator, our GLBO explores some orthogonal directions. We consider the problem of deriving variational bounds for more generalized ϕ-evidence functions, which can be designed to encourage desirable properties of a solution. Additionally, in our framework the improvement for the bound also comes from the ψ-transformation. We also advocate a new update rule for the parameters to mitigate the SNR issue; see Section 6.2 for details.

Regularized variational inference While not directly motivated from a model-selection perspective, recent developments in regularized variational inference shares similarities with our approach. In generative modeling, traditional VI has been criticized for producing unrealistic samples. This issue traces back to the fact that the aggregated approximate posterior does not match the prior (Makhzani et al., 2016), as ELBO based inference tend to underestimate posterior variance (Pu et al., 2018). A number of solutions have been proposed to alleviate this problem, most notably, adversarially regularized solutions (Dumoulin et al., 2016; Pu et al., 2017b; Li et al., 2017). These methods introduced an adversarial loss, penalizing the mismatch between the marginal latent distributions $p(z)$ and $q_\beta(z) = \int p_d(x)q_\beta(z|x)\,\mathrm{d}x$, or the joint distributions $p_\alpha(x,z)$ and $q_\beta(x,z) = p_d(x)q_\beta(z|x)$. These regularization approaches favor models that output more realistic samples.

Robust variational inference Our work also complements recent developments in *robust variational inference* (Wang et al., 2017; Figurnov et al., 2016). In Wang et al. (2017), the authors hypothesized that the cause of instability in VI comes from the presence of "bad" observations. To address this, they proposed to dynamically reweight samples, constrained by a prior distribution on the weight vector. This effectively down-weights "bad" observations and relieves the learner from modeling nonconforming examples. However, this method does not follow a standard probabilistic approach, and the results are sensitive to the choice of prior and other hyper parameters. In the work of Figurnov et al. (2016) the authors heuristically applied a soft-thresholded $\log(u)$ as the evidence function in ELBO. Similar to the reweighting strategy, this effectively eliminates any signal from low evidence samples during training.

6. Optimization of GLBO

In this section we first describe an easy-to-implement low-variance estimator that improves GLBO training, then discuss a new update rule based on theoretical insights. We also detail the state-of-the-art VI models we tested with, the fact that GLBO improves upon these models demonstrates its wide applicability (see Section 7).

6.1. Improving the bound with moving average

A naïve estimator for the GLBO in the stochastic gradient descent setting is

$$V_{L,K}(x, \{Z_{l,k}\}; \alpha, \beta) = \psi^{-1}\left(\frac{1}{L}\sum_{l=1}^{L} h\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p_\alpha(x, Z_{l,k})}{q_\beta(Z_{l,k}|x)}\right)\right), \tag{8}$$

where the expectation over $Z_{1:K} \sim q_\beta(z|x)$ is replaced with the average of L empirical samples $\{Z_{l,1:K}\}_{l=1}^{L}$ drawn from $q_\beta(z|x)$. However, this potentially introduces a negative bias, as one can readily derive from the Jensen's inequality that $\mathbb{E}_{\{Z_{l,k}\}\sim q_\beta(z|x)}\left[V_{L,K}(x, \{Z_{l,k}\})\right] \leq \text{GLBO}(x; K)$. To reduce this bias, we note that our stochastic objective can be rewritten in a more general form as $\psi^{-1}(\hat{h}(x,p,q))$, where $\hat{h}(x,p,q)$ is an estimator for the term $\mathbb{E}[h]$ in the definition of the GLBO. Interestingly, this bias can be ameliorated by reducing the variance of estimator $\hat{h}(x,p,q)$. We provide an asymptotic argument to support this claim in the SM.

Given the insights from above, we propose to replace the naïve estimator $\hat{h}$ with a moving average estimator $\hat{h}_{\text{ema}}$, which in principle should reduce the variance and provide tighter estimate for the bound. Specifically, our empirical estimator for GLBO at iteration t is computed as

$$V_K^{\text{ema}}(x,t) = \psi^{-1}(\hat{h}_{\text{ema}}(x,t)),$$
$$\hat{h}_{\text{ema}}(x,t) = (1-w_t)\hat{h}_{\text{ema}}(x,t-1) + w_t h\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p_\alpha(x,Z_{t,k})}{q_\beta(Z_{t,k}|x)}\right),$$

where $Z_{t,1:K} \sim q_\beta(z|x)$ are approximate posterior samples and $0 \le w_t \le 1$ is the update weight for the averaging estimator. At iteration t, the historical estimate $\hat{h}_{\text{ema}}(x, t-1)$ is treated as a constant baseline and the gradients are only propagated through the evaluation on current posterior samples $Z_{t,1:K}$.

6.2. A high SNR update rule

The GLBO estimator is also vulnerable to the SNR issue associated with importance-weighted estimators analyzed by Rainforth et al. (2017) (see discussion in Section 5 above). Motivated by their theoretical insights, we propose to update the generator parameter α and inference parameter β with objective functions based on estimators of different variational bounds. Consider the naive estimator $V_{L,K}(x, \{Z_{l,k}\}; \theta)$ in (8), and assume we have a budget of S posterior samples for each parameter update. In the SGD setting, we propose to update the parameters with

$$\alpha_{t+1} \leftarrow \alpha_t + \eta_t \nabla_\alpha V_{S,1}(x_t, \{Z_{l,k}\}; \alpha_t, \beta_t),$$
$$\beta_{t+1} \leftarrow \beta_t + \eta_t \nabla_\beta V_{1,S}(x_t, \{Z_{l,k}\}; \alpha_t, \beta_t), \quad (9)$$

where x_t is the data sampled at iteration t and η_t is the learning rate. This combines the best of two worlds, as α and β are respectively updated by a high SNR gradient estimator. Generalization to the moving average estimator discussed above is straightforward. We note that the extra computational cost for using (9) instead of vanilla update rule $\theta_{t+1} \leftarrow \theta_t + \eta_t \nabla_\theta V(x_t; \theta_t)$ is neglectable in the way modern differentiable learning algorithms are implemented.

6.3. Local ELBO with flexible posterior approximation

To allow for more flexible posterior representation, we consider nonparametric posterior $q_\beta(z|x)$ implicitly defined by a latent code generator $z = G(x, \xi; \beta)$, where $\xi \sim q(\xi)$ is a simple of randomness for the posterior, *e.g.*, Gaussian.

Tractable posterior approximation If $G(x, \xi; \beta)$ is invertible wrt ξ, then $q_\beta(z|x) = q(\xi)|\det(\nabla_\xi G^{-1}(x, \xi; \beta))|$. A well known example of such invertible generators is the normalizing flow (NF) (Tabak et al., 2010; Rezende & Mohamed, 2015), which considers $G(x, \xi; \beta) = z_M$ recursively defined by $z_m = f_m(z_{m-1}, x; \beta), \forall m = 1, \cdots, M$. Here $z_0 = \xi$ and $\{f_m(z, x; \beta)\}_{m=1}^M$ is a chain of transformations invertible wrt to z parameterized by β. This allows a flexible posterior approximation $\log q_\beta(z|x)$, with a tractable log density that can be explicitly computed by back tracing the Jacobians of $\{f_m\}$, *e.g.* $\log q_\beta(z|x) = \log q(\xi) - \sum_{m=1}^M \log\left|\det(\nabla_{z_{m-1}} f_m)\right|$.

Intractable posterior approximation Using an unconstrained transformation $G(x, \xi; \beta)$ allows more expressive posterior approximation at the cost of no explicit expression for $q_\beta(z|x)$. To overcome this difficulty, we use the adversarial approach proposed by Mescheder et al. (2017). Specifically, we can decompose the local ELBO into the sum of a tractable log-likelihood term and an intractable log-likelihood ratio term (also known as the local KL) $\log f(x, z) = \log p_\alpha(x|z) + \log \frac{p(z)}{q_\beta(z|x)}$. Here we learn $r_\beta(x, z) \triangleq \log \frac{p(z)}{q_\beta(z|x)}$ by training an optimal discriminator $\sigma(r(x, z))$ between samples drawn from $p(z)$ and $q_\beta(z|x)$

$$r_\beta(x,z) = \arg\max_{r(x,z;\phi)} \{\mathbb{E}_{Z\sim p(z)}[\log\sigma(r(x,Z;\phi))] + \mathbb{E}_{Z'\sim q_\beta(z|x)}[\log(1-\sigma(r(x,Z';\phi)))]\},$$

where $\sigma(u) = (1+e^{-u})^{-1}$ is the sigmoid function. To circumvent the numerical difficulties associated with vanishing likelihood ratios, we further leverage the *adaptive contrast* (AC) technique proposed in Mescheder et al. (2017), introducing an auxiliary distribution to improve the estimate of $r_\beta(x, z)$. See the SM for details.

7. Experiments

To compare the performance of our new bound and its predecessors, we empirically evaluate the sharpness of these bounds on a toy distribution, and benchmark them on a series of VI tasks. In all K-sample experiments we use the K-sample ELBO estimator to make the comparisons fair wrt computational cost, and report the vanilla ELBO as the log-evidence bound for all models in quantitative evaluations. We use the proposed moving average estimator except for the Bayesian regression experiment. Details of the experimental setup are in the SM, and source code is available (upon publication) from `https://www.github.com/LiqunChen0606/glbo`.

7.1. Bound sharpness

We consider the following toy distribution to quantitatively evaluate the performance of different bounds $X = \sin(Z) + \mathcal{N}(0, 0.01)$, $Z \sim \mathcal{U}[0, \pi]$, where $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian with mean μ and variance σ^2, and $\mathcal{U}[a, b]$ to denote a uniform distribution on interval $[a, b]$. This specifies a simple two-dimensional distribution $p(x, z)$, and we specify a simple unit variance Gaussian $q(z|x) = \mathcal{N}(\pi/2, 1)$ centered at $\pi/2$ as our posterior approximation to estimate a bound on marginal $p(x)$ (See Figure SM-1(a-b)). The ground truth $p(x)$ is estimated using a naive Monte Carlo

Table 1: Average test log-likelihood for variational auto-encoder. † Results collected from Burda et al. (2016); Li & Turner (2016).

Dataset	L	K	VAE†	IW-VAE†	RVB†	CLBO
Frey Face	1	5	1322.96	1380.30	1377.40	**1381.23**
Caltech 101 Silhouettes	1	5	-119.69	**-117.89**	-118.01	**-117.90**
		50	-119.61	-117.21	-117.10	**-116.95**
MNIST	1	5	-86.47	-85.41	-85.42	**-84.71**
		50	-86.35	-84.80	-84.81	**-84.30**
	2	5	-85.01	-83.92	-84.04	**-83.45**
		50	-84.75	-83.12	-83.44	**-82.94**
OMNIGLOT	1	5	-107.62	**-106.30**	**-106.33**	**-106.31**
		50	-107.80	-104.68	-105.05	**-104.58**
	2	5	-106.31	-104.64	-104.71	**-104.52**
		50	-106.30	**-103.25**	-103.72	-103.30

estimator $\hat{p}(x) = \frac{1}{S}\sum_{s=1}^{S} p(x, z_l),\ z_{1:S} \sim \mathcal{U}[0, \pi]$, where we set $S = 10,000$.

Figure 1(c,d) summarize the estimated GLBO with growing K and decreasing T, respectively. As our theory predicts, GLBO gets sharpened as K increases or as T decreases. Vanilla ELBO does not provide a reasonable bound in this experiment. Figure 1(a-c) compare the K-sample RVB, ELBO and CLBO. Our results verify that RVB does not necessarily improve importance-weighted ELBO, which is consistent with the empirical results reported by the original RVI paper (Li & Turner, 2016). GLBO on the other hand, is guaranteed to improve its ELBO counterpart. Notably, the performance boost is especially significant in the low-sample regime ($K < 5$) under our experimental conditions.

7.2. Variational autoencoder

Our next experiment considers applying GLBO to variational autoencoders for unsupervised learning. To make comparisons fair, we focus on modifying publicly available implementations with our GLBO objective[1]. All experimental results are produced with the recommended settings from the original implementations.

We first compare GLBO with the vanilla variational autoencoder, importance-weighted VAE and RVB under the experimental setups from Li & Turner (2016) [2], on the MNIST dataset. The encoders and decoders are implemented with $L \in \{1, 2\}$ neural network layers and leveraging $K \in \{5, 50\}$ posterior samples. We choose the VR-Max estimator for RVB and set GLBO to CLBO$(x; T, K)$ with $T = 200$. To evaluate performance, we estimate the true log-likelihood with $S = 5,000$ importance-weighted posterior samples, and report the average of test set log-likelihood in Table 1. Our GLBO consistently improves performance, and the gain is more pronounced in the low posterior sample regime. We have also observed that our GLBO leads to faster convergence (not shown). We have varied our experimental settings and the results are qualitatively similar.

Table 2: ELBO, AIS and reconstruction error on MNIST for different models. ‡ Results collected from Mescheder et al. (2017).

Model	ELBO	AIS	Recon. Err.
AVB+GLBO	**-79.97 ± 0.15**	**-81.2**	**57.2 ± 0.12**
AVB‡	-82.7 ± 0.2	-81.7	**57. ± 0.2**
VAE‡	-85.7 ± 0.2	-81.9	59.4 ± 0.2
AuxiliaryVAE ‡	-85.6 ± 0.2	-81.6	59.6 ± 0.2
VAE/IAF ‡	-85.5 ± 0.2	-82.1	59.6 ± 0.2

We also examined if GLBO can enhance models with more flexible posterior distribution. We follow the experimental setup used in AVB (Mescheder et al., 2017) [3]. In the MNIST experiment, we compare a GLBO version AVB with vanilla VAE, inverse autoregressive flow (IAF) (Kingma et al., 2016), auxiliary VAE (Maaløe et al., 2016) and AVB. No importance sampling is used, as in the original implementation, and we choose the best performing adaptive contrast AVB for comparison. We summarize the ELBO, AIS score (Wu et al., 2017) and reconstruction error in Table 2. Both GLBO and AVB achieved better reconstruction than other competitors, and our GLBO leads the performance on ELBO and AIS by a large margin.

We further evaluate GLBO on the more complex CelebA face dataset (Liu et al., 2015). We benchmark K-sample CLBO-VAE against ELBO, IW-ELBO and Rényi VAEs, using a convolutional neural net encoder and deconvolution-layer based decoder as our architecture (Radford et al., 2016). The training and testing set log-evidence bounds as a function of training epochs are shown in Figure 4. CLBO-VAE showed both better evidence score and more stable training dynamics compared with its counterparts. In Figure 4 we also show the learning curves of each model augmented with NF posterior approximation, trained on the MNIST data. All models except for ELBO performed similarly, possibly because of the highly expressive NF approximation. Additionally, the high SNR update rule failed to improve the vanilla ELBO-VAE on CelebA, but provided slightly better performance compared with all other methods on MNIST+NF.

For the last experiment on VAE, we explore the idea of instantiating model-selection with GLBO. We first train the regular log-evidence objective to convergence with the AVB model on MNIST, and then switch to optimize ϕ-evidence with GLBO to prioritize more plausible models. Here we use the shifted log-log function $\phi(u) = \log\left(\log(u) - \ell_{\text{lower}}\right)$ as our evidence function, so that we can vary ℓ_{lower} to manipulate the shape of ϕ (see SM for details of our choice). Figure 3 compares the log-evidence dis-

[1] In this work we use the results reported by the original papers, and we are able to reproduce these results with the publicly available code.

[2] https://github.com/YingzhenLi/vae_renyi_divergence

[3] https://github.com/LMescheder/AdversarialVariationalBayes

Table 3: Test RMSE and log-likelihood results for Bayesian neural net regression.

	Test RMSE (lower is better)				Test log-likelihood (higher is better)			
Dataset	VI	PBP	Rényi	CLBO	VI	PBP	Rényi	CLBO
Boston	$4.32 \pm .29$	$3.01 \pm .18$	$2.86 \pm .40$	$\mathbf{2.71 \pm .29}$	$-2.90 \pm .07$	$-2.57 \pm .09$	$-2.46 \pm .16$	$\mathbf{-2.40 \pm .09}$
Concrete	$7.19 \pm .12$	$5.67 \pm .09$	$5.15 \pm .25$	$\mathbf{5.04 \pm .27}$	$-3.39 \pm .02$	$-3.16 \pm .02$	$-3.04 \pm .07$	$\mathbf{-3.02 \pm .05}$
Energy	$2.65 \pm .08$	$1.80 \pm .05$	$1.00 \pm .18$	$\mathbf{0.95 \pm .15}$	$-2.39 \pm .03$	$-2.04 \pm .02$	$-1.67 \pm .05$	$\mathbf{-1.65 \pm .04}$
Kin8nm	$0.10 \pm .00$	$0.10 \pm .00$	$\mathbf{0.08 \pm .00}$	$\mathbf{0.08 \pm .00}$	$0.90 \pm .01$	$0.90 \pm .01$	$\mathbf{1.14 \pm .02}$	$\mathbf{1.14 \pm .02}$
Naval	$0.01 \pm .00$	$0.01 \pm .00$	$\mathbf{0.00 \pm .00}$	$\mathbf{0.00 \pm .00}$	$3.73 \pm .12$	$3.73 \pm .01$	$4.11 \pm .11$	$\mathbf{4.17 \pm .10}$
CCPP	$4.33 \pm .04$	$4.12 \pm .03$	$4.13 \pm .04$	$\mathbf{4.03 \pm .06}$	$-2.89 \pm .02$	$-2.85 \pm .05$	$-2.84 \pm .04$	$\mathbf{-2.81 \pm .02}$
Winequality	$0.65 \pm .01$	$0.64 \pm .02$	$0.62 \pm .03$	$\mathbf{0.61 \pm .03}$	$-0.98 \pm .01$	$-0.97 \pm .01$	$-0.94 \pm .04$	$\mathbf{-0.93 \pm .04}$
Yacht	$6.89 \pm .67$	$1.02 \pm .05$	$0.94 \pm .23$	$\mathbf{0.87 \pm .18}$	$-3.43 \pm .16$	$-1.63 \pm .02$	$-1.61 \pm .00$	$\mathbf{-1.52 \pm .00}$
Protein	$4.84 \pm .03$	$4.73 \pm .01$	$4.65 \pm .07$	$\mathbf{4.43 \pm .05}$	$-2.99 \pm .01$	$-2.97 \pm .00$	$-2.93 \pm .00$	$\mathbf{-2.89 \pm .01}$
Year	$9.03 \pm$ NA	$8.88 \pm$ NA	$8.80 \pm$ NA	$\mathbf{8.78 \pm NA}$	$-3.62 \pm$ NA	$-3.60 \pm$ NA	$-3.60 \pm$ NA	$\mathbf{-3.57 \pm NA}$

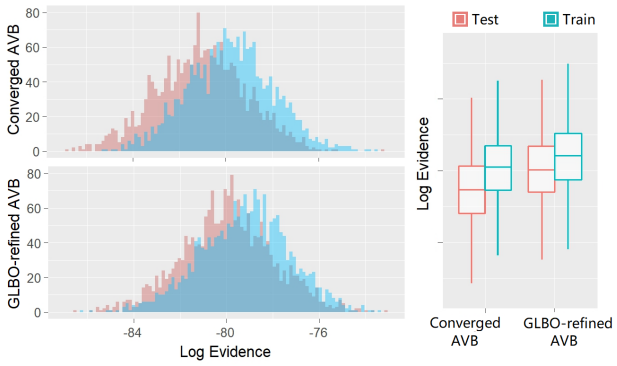

Figure 3: Model-selection result on MNIST. log-evidence histogram plot (left), box plot of mean and quantile (right) for a converged and refined AVB model.

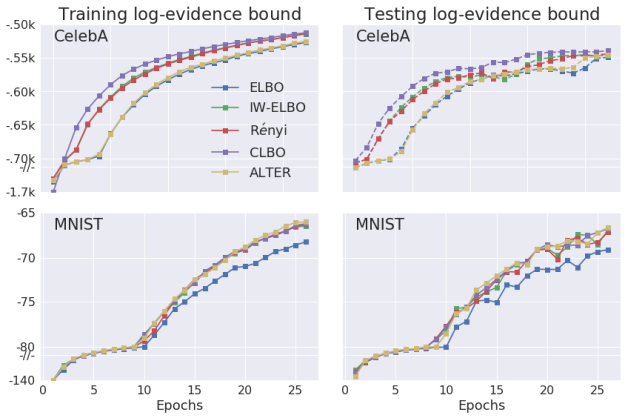

Figure 4: log-evidence bound evolution wrt training epochs on CelebA (upper panel) and MNIST (lower panel). Low evidence scores rescaled for better visualization. Normalizing flow is used for the MNIST variational models. ALTER denotes CLBO learned with the high SNR update rule proposed in Sec 6.2.

tribution with and without model-selection on the MNIST dataset. The refinement phase improves performance on both the training and testing set, and the boost in generalization is more pronounced (testing $+1.47$ vs training $+0.58$ nats). This validates our hypothesis that applying the maximum entropy heuristic favors more plausible models.

8. Bayesian Neural Net Regression

Finally we consider the problem of Bayesian regression with neural nets. We use ten datasets from the UCI Machine Learning Repository (Lichman, 2013) and followed the experimental setup from Li & Turner (2016); see SM for details. We use a random $90\%/10\%$ split for training and testing, and use test root mean squared error (RMSE) and log-likelihood (LL) for evaluation.

We compared CLBO with ELBO, IW-ELBO, Rényi-VI and probabilistic backpropagation (PBP) (Hernández-Lobato & Adams, 2015) in this experiment. For CLBO and Rényi we fixed $T = 2$. The results are summarized in Table 3.[4] The proposed CLBO in general improves over its counterparts. This provides evidence that CLBO learns a better model rather than simply bumping up the evidence bound.

[4]The results for IW-ELBO is quantitatively similar to those of Rényi-VI, we therefore report it in the SM to save space.

9. Conclusion

We have considered generalization of the evidence score, and have proposed a new family of evidence bounds and improved estimators. Our work subsumes many existing bounds as special cases, while also being provably sharper. We carried out experiments to validate our claims, and the results are consistent with our theoretical predictions. We provided empirical evidence that our method improves state-of-the-art approaches. We also investigated the issue of model-selection in variational inference, and proved, empirically, that our theoretically-inspired strategy leads to an improvement in generalization performance.

In future work, we intend to build on automated inference procedures with generalized evidence bounds. This involves further understanding of ϕ-evidence bounds, and designing principled strategies that are guaranteed to achieve desired optimality conditions. Adaptive hyper-parameter tuning is also desirable to simplify ϕ-evidence based VI.

Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments. This research was supported in part by DARPA, DOE, NIH, ONR and NSF. The authors would also like to thank S Dai, Dr. Y Li and Dr. Y Pu for fruitful discussions.

References

Bamler, R., Zhang, C., Opper, M., and Mandt, S. Perturbative black box variational inference. In *NIPS*, 2017.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3 (Jan):993–1022, 2003.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. In *ICLR*, 2016.

Dharmadhikari, S. and Joag-Dev, K. Examples of nonunique maximum likelihood estimators. *The American Statistician*, 39(3):199–200, 1985.

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. M. Variational inference via chi upper bound minimization. In *NIPS*, 2017.

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., and Courville, A. Adversarially learned inference. In *ICLR*, 2016.

Figurnov, M., Struminsky, K., and Vetrov, D. Robust variational inference. *arXiv preprint arXiv:1611.09226*, 2016.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., and Penny, W. Variational free energy and the laplace approximation. *NeuroImage*, 34(1):220–234, 2007.

Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., and Wierstra, D. Draw: A recurrent neural network for image generation. In *ICML*, 2015.

Hernández-Lobato, J. M. and Adams, R. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, pp. 1861–1869, 2015.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Jaynes, E. T. *Probability theory: The logic of science*. Cambridge university press, 2003.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *ICLR*, 2014.

Kingma, D. P., Salimans, T., and Welling, M. Improving variational inference with inverse autoregressive flow. In *NIPS*, 2016.

Li, C., Liu, H., Chen, C., Pu, Y., Chen, L., Henao, R., and Carin, L. Alice: Towards understanding adversarial learning for joint distribution matching. In *NIPS*, 2017.

Li, Y. and Turner, R. E. Rényi divergence variational inference. In *NIPS*, 2016.

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Liu, J. S. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *NIPS*, 2016.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *ICCV*, 2015.

Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O. Auxiliary deep generative models. *arXiv preprint arXiv:1602.05473*, 2016.

Mäkeläinen, T., Schmidt, K., and Styan, G. P. On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples. *The Annals of Statistics*, pp. 758–767, 1981.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. Adversarial autoencoders. In *ICLR*, 2016.

Mescheder, L., Nowozin, S., and Geiger, A. Adversarial variational Bayes: unifying variational autoencoders and generative adversarial networks. In *ICML*, 2017.

Miller, A. C., Foti, N., and Adams, R. P. Variational boosting: Iteratively refining posterior approximations. In *ICML*, 2016.

Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., and Carin, L. Variational autoencoder for deep learning of images, labels and captions. In *NIPS*, 2016.

Pu, Y., Gan, Z., Henao, R., Li, C., Han, S., and Carin, L. VAE learning via Stein variational gradient descent. In *NIPS*, 2017a.

Pu, Y., Wang, W., Henao, R., Chen, L., Gan, Z., Li, C., and Carin, L. Adversarial symmetric variational autoencoder. In *NIPS*, 2017b.

Pu, Y., Chen, L., Dai, S., Wang, W., Li, C., and Carin, L. Symmetric variational autoencoder and connections to adversarial learning. In *AISTATS*, 2018.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

Rainforth, T., Le, T. A., Maddison, M. I. C. J., and Wood, Y. W. T. F. Tighter variational bounds are not necessarily better. In *NIPS workshop*. 2017.

Ranganath, R., Gerrish, S., and Blei, D. Black box variational inference. In *AISTATS*, 2014.

Ranganath, R., Tran, D., and Blei, D. Hierarchical variational models. In *ICML*, 2016.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. In *ICML*, 2015.

Tabak, E. G., Vanden-Eijnden, E., et al. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

Wang, Y., Kucukelbir, A., and Blei, D. M. Reweighted data for robust probabilistic models. In *ICML*, 2017.

Webb, S. and Teh, Y. W. A tighter monte carlo objective with rényi α-divergence measures. In *NIPS Workshop*, 2016.

Wu, Y., Burda, Y., Salakhutdinov, R., and Grosse, R. On the quantitative analysis of decoder-based generative models. In *ICLR*. 2017.

Zhang, R., Li, C., Chen, C., and Carin, L. Learning structural weight uncertainty for sequential decision-making. In *AISTATS*, 2018.