Energy-Inspired Models: Learning with Sampler-Induced Distributions

Dieterich Lawson[*,†]
Stanford University
jdlawson@stanford.edu

George Tucker[*], **Bo Dai**
Google Research, Brain Team
{gjt, bodai}@google.com

Rajesh Ranganath
New York University
rajeshr@cims.nyu.edu

Abstract

Energy-based models (EBMs) are powerful probabilistic models [8, 43], but suffer from intractable sampling and density evaluation due to the partition function. As a result, inference in EBMs relies on approximate sampling algorithms, leading to a mismatch between the model and inference. Motivated by this, we consider the sampler-induced distribution as the model of interest and maximize the likelihood of this model. This yields a class of *energy-inspired models* (EIMs) that incorporate learned energy functions while still providing exact samples and tractable log-likelihood lower bounds. We describe and evaluate three instantiations of such models based on truncated rejection sampling, self-normalized importance sampling, and Hamiltonian importance sampling. These models outperform or perform comparably to the recently proposed Learned Accept/Reject Sampling algorithm [5] and provide new insights on ranking Noise Contrastive Estimation [33, 45] and Contrastive Predictive Coding [55]. Moreover, EIMs allow us to generalize a recent connection between multi-sample variational lower bounds [9] and auxiliary variable variational inference [1, 61, 57, 46]. We show how recent variational bounds [9, 48, 51, 41, 68, 50, 63] can be unified with EIMs as the variational family.

1 Introduction

Energy-based models (EBMs) have a long history in statistics and machine learning [16, 70, 43]. EBMs score configurations of variables with an energy function, which induces a distribution on the variables in the form of a Gibbs distribution. Different choices of energy function recover well-known probabilistic models including Markov random fields [35], (restricted) Boltzmann machines [62, 24, 29], and conditional random fields [40]. However, this flexibility comes at the cost of challenging inference and learning: both sampling and density evaluation of EBMs are generally intractable, which hinders the applications of EBMs in practice.

Because of the intractability of general EBMs, practical implementations rely on approximate sampling procedures (*e.g.*, Markov chain Monte Carlo (MCMC)) for inference. This creates a mismatch between the model and the approximate inference procedure, and can lead to suboptimal performance and unstable training when approximate samples are used in the training procedure. Currently, most attempts to fix the mismatch lie in designing better sampling algorithms (*e.g.*, Hamiltonian Monte Carlo [53], annealed importance sampling [52]) or exploiting variational techniques [34, 15, 14] to reduce the inference approximation error.

Instead, we bridge the gap between the model and inference by directly treating the sampling procedure as the model of interest and optimizing the log-likelihood of the the sampling procedure.

[*]Equal contributions. [†]Research performed while at New York University.
Code and image samples: sites.google.com/view/energy-inspired-models.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

We call these models *energy-inspired models* (EIMs) because they incorporate a learned energy function while providing tractable, exact samples. This shift in perspective aligns the training and sampling procedure, leading to principled and consistent training and inference.

To accomplish this, we cast the sampling procedure as a latent variable model. This allows us to maximize variational lower bounds [32, 7] on the log-likelihood (c.f., Kingma and Welling [37], Rezende et al. [59]). To illustrate this, we develop and evaluate energy-inspired models based on truncated rejection sampling (Algorithm 1), self-normalized importance sampling (Algorithm 2), and Hamiltonian importance sampling (Algorithm 3). Interestingly, the model based on self-normalized importance sampling is closely related to *ranking* NCE [33, 45], suggesting a principled objective for training the "noise" distribution.

Our second contribution is to show that EIMs provide a unifying conceptual framework to explain many advances in constructing tighter variational lower bounds for latent variable models (*e.g.*, [9, 48, 51, 41, 68, 50, 63]). Previously, each bound required a separate derivation and evaluation, and their relationship was unclear. We show that these bounds can be viewed as specific instances of auxiliary variable variational inference [1, 61, 57, 46] with different EIMs as the variational family. Based on general results for auxiliary latent variables, this immediately gives rise to a variational lower bound with a characterization of the tightness of the bound. Furthermore, this unified view highlights the implicit (potentially suboptimal) choices made and exposes the reusable components that can be combined to form novel variational lower bounds. Concurrently, Domke and Sheldon [19] note a similar connection, however, their focus is on the use of the variational distribution for posterior inference.

In summary, our contributions are:

- The construction of a tractable class of *energy-inspired models* (EIMs), which lead to *consistent* learning and inference. To illustrate this, we build models with truncated rejection sampling, self-normalized importance sampling, and Hamiltonian importance sampling and evaluate them on synthetic and real-world tasks. These models can be fit by maximizing a tractable lower bound on their log-likelihood.
- We show that EIMs with auxiliary variable variational inference provide a unifying framework for understanding recent tighter variational lower bounds, simplifying their analysis and exposing potentially sub-optimal design choices.

2 Background

In this work, we consider learned probabilistic models of data $p(x)$. Energy-based models [43] define $p(x)$ in terms of an energy function $U(x)$

$$p(x) = \frac{\pi(x)\exp(-U(x))}{Z},$$

where π is a tractable "prior" distribution and $Z = \int \pi(x)\exp(-U(x))\,dx$ is a generally intractable partition function. To fit the model, many approximate methods have been developed (*e.g.*, pseudo log-likelihood [6], contrastive divergence [29, 65], score matching estimator [30], minimum probability flow [64], noise contrastive estimation [27]) to bypass the calculation of the partition function. Empirically, previous work has found that convolutional architectures that score images (*i.e.*, map x to a real number) tend to have strong inductive biases that match natural data [22]. These networks are a natural fit for energy-based models. Because drawing exact samples from these models is intractable, samples are typically approximated by Monte Carlo schemes, for example, Hamiltonian Monte Carlo [54].

Alternatively, latent variables z allow us to construct complex distributions by defining the likelihood $p(x) = \int p(x|z)p(z)\,dz$ in terms of tractable components $p(z)$ and $p(x|z)$. While marginalizing z is generally intractable, we can instead optimize a tractable lower bound on $\log p(x)$ using the identity

$$\log p(x) = \mathbb{E}_{q(z|x)}\left[\log \frac{p(x,z)}{q(z|x)}\right] + D_{\mathrm{KL}}\left(q(z|x)||p(z|x)\right), \tag{1}$$

where $q(z|x)$ is a variational distribution and the positive D_{KL} term can be omitted to form a lower bound commonly referred to as the evidence lower bound (ELBO) [32, 7]. The tightness of the bound

is controlled by how accurately $q(z|x)$ models $p(z|x)$, so limited expressivity in the variational family can negatively impact the learned model.

3 Energy-Inspired Models

Instead of viewing the sampling procedure as drawing approximate samples from the energy-based models, we treat the sampling procedure as the model of interest. We represent the randomness in the sampler as latent variables, and we obtain a tractable lower bound on the marginal likelihood using the ELBO. Explicitly, if $p(\lambda)$ represents the randomness in the sampler and $p(x|\lambda)$ is the generative process, then

$$\log p(x) \geq \mathbb{E}_{q(\lambda|x)}\left[\log \frac{p(\lambda)p(x|\lambda)}{q(\lambda|x)}\right], \tag{2}$$

where $q(\lambda|x)$ is a variational distribution that can be optimized to tighten the bound. In this section, we explore concrete instantiations of models in this paradigm: one based on truncated rejection sampling (TRS), one based on self-normalized importance sampling (SNIS), and another based on Hamiltonian importance sampling (HIS) [53].

Algorithm 1 TRS(π, U, T) generative process

Require: Proposal distribution $\pi(x)$, energy function $U(x)$, and truncation step T.
1: **for** $t = 1, \ldots, T-1$ **do**
2: Sample $x_t \sim \pi(x)$.
3: Sample $b_t \sim$ Bernoulli$(\sigma(-U(x_t)))$.
4: **end for**
5: Sample $x_T \sim \pi(x)$ and set $b_T = 1$.
6: Compute $i = \min t$ s.t. $b_t = 1$.
7: **return** $x = x_i$.

3.1 Truncated Rejection Sampling (TRS)

Consider the truncated rejection sampling process (Algorithm 1) used in [5], where we sequentially draw a sample x_t from $\pi(x)$ and accept it with probability $\sigma(-U(x_t))$. To ensure that the process ends, if we have not accepted a sample after T steps, then we return x_T.

In this case, $\lambda = (x_{1:T}, b_{1:T-1}, i)$, so we need to construct a variational distribution $q(\lambda|x)$. The optimal $q(\lambda|x)$ is $p(\lambda|x)$, which motivates choosing a similarly structured variational distribution. It is straightforward to see that $p(i|x) \propto (1-Z)^{i-1}\sigma(-U(x))^{\delta_{i<T}}$, where $Z = \int \pi(x)\sigma(-U(x))\,dx$ is generally intractable. So, we choose $q(i|x) \propto (1-\hat{Z})^{i-1}\sigma(-U(x))^{\delta_{i<T}}$, where $\hat{Z}$ is a learnable variational parameter. Then, we sample $x_{1:T}$ and $b_{i+1:T-1}$ as in the generative process. This results in a simple variational bound

$$\log p_{TRS}(x) \geq \mathbb{E}_{q(i|x)}\mathbb{E}_{\prod_{t=1}^{i}\pi(x_t)}\left[\log \pi(x)\sigma(-U(x)) + \sum_{t=1}^{i-1}\log\left(1-\sigma(-U(x_t))\right) - \log q(i|x)\right].$$

The TRS generative process is the same process as the Learned Accept/Reject Sampling (LARS) model [5]. The key difference is the training procedure. LARS tries to directly estimate the gradient of the log likelihood. Without truncation, such a process is attractive because unbiased gradients of its log likelihood can easily be computed without knowing the normalizing constant. Unfortunately, after truncating the process, we require estimating a normalizing constant. In practice, Bauer and Mnih [5] estimate the normalizing constant using 1024 samples during training and 10^{10} samples during evaluation. Even so, LARS requires additional implementation tricks (*e.g.*, evaluating the target density, using an exponential moving average to estimate the normalizing constant) to ensure successful training, which complicate the implementation and analysis of the algorithm. On the other hand, we optimize a tractable log likelihood lower bound. As a result, no implementation tricks are necessary.

3.2 Self-Normalized Importance Sampling (SNIS)

Consider the sampling process defined by self-normalized importance sampling. That is, first sampling a set of K candidate x_is from a proposal distribution $\pi(x_i)$, and then sampling x from the empirical distribution composed of atoms located at each x_i and weighted proportionally to $\exp(-U(x_i))$ (Algorithm 2). In this case, the latent variables λ are the locations of the proposal samples $x_1, \ldots, x_K$ (abbreviated $x_{1:K}$) and the index of the selected sample, i.

Explicitly, the model is defined by

$$p(x_{1:K}, i) = \left(\prod_{k=1}^{K} \pi(x_k)\right) \frac{\exp(-U(x_i))}{\sum_k \exp(-U(x_k))}, \qquad p(x|x_{1:K}, i) = \delta_{x_i}(x),$$

with $\lambda = (x_{1:K}, i)$. We denote the density of the process by $p_{SNIS}(x)$. Choosing $q(\lambda|x) = \frac{1}{K}\delta_{x_i}(x)\prod_{j\neq i}\pi(x_j)$ in Eq. (2), yields

$$\log p_{SNIS}(x) \geq \mathbb{E}_{x_{2:K}} \log \left[\frac{\pi(x)\exp(-U(x))}{\frac{1}{K}\left(\sum_{j=2}^{K}\exp(-U(x_j)) + \exp(-U(x))\right)}\right]. \tag{3}$$

To summarize, $p_{SNIS}(x)$ can be sampled from exactly and has a tractable lower bound on its log-likelihood. For the same K, we expect p_{SNIS} to outperform p_{TRS} because it considers all candidate samples simultaneously instead of sequentially.

Algorithm 2 SNIS(π, U) generative process

Require: Proposal distribution $\pi(x)$ and energy function $U(x)$.
1: **for** $k = 1, \ldots, K$ **do**
2: Sample $x_k \sim \pi(x)$.
3: Compute $w(x_k) = \exp(-U(x_k))$.
4: **end for**
5: Compute $\hat{Z} = \sum_{k=1}^{K} w(x_k)$
6: Sample $i \sim \text{Categorical}(w(x_1)/\hat{Z}, \ldots, w(x_K)/\hat{Z})$.
7: **return** $x = x_i$.

As $K \to \infty$, $p_{SNIS}(x)$ becomes proportional to $\pi(x)\exp(-U(x))$. For finite K, $p_{SNIS}(x)$ interpolates between the tractable proposal $\pi(x)$ and the energy model $\pi(x)\exp(-U(x))$. Furthermore, Equation (3) is closely connected with the *ranking* NCE loss [33, 45], a popular objective for training energy-based models. In fact, if we consider $\pi(x)$ as our noise distribution $p_N(x)$ and set $U(x) = \log p_N(x) - s(x)$, then up to a constant (in s), we recover the ranking NCE loss using the notation from [45]. The ranking NCE loss is motivated by the fact that it is a consistent objective for any $K > 1$ when the true data distribution is in our model family. As a result, it is straightforward to adapt the consistency proof from [45] to our setting. Furthermore, our perspective gives a coherent objective for jointly learning the noise distribution and the energy function and shows that the ranking NCE loss can be viewed as a lower bound on the log likelihood of a well-specified model regardless of whether the true data distribution is in our model family. In addition, we can recover the recently proposed InfoNCE [55] bound on mutual information by using SNIS as the variational distribution in the classic variational bound by Barber and Agakov [4] (see Appendix C for details).

To train the SNIS model, we perform stochastic gradient ascent on Eq. (3) with respect to the parameters of the proposal distribution π and the energy function U. When the data x are continuous, reparameterization gradients can be used to estimate the gradients to the proposal distribution [59, 37]. When the data are discrete, score function gradient estimators such as REINFORCE [66] or relaxed gradient estimators such as the Gumbel-Softmax [47, 31] can be used.

3.3 Hamiltonian importance sampling (HIS)

Simple importance sampling scales poorly with dimensionality, so it is natural to consider more complex samplers with better scaling properties. We evaluated models based on Hamiltonian importance sampling (HIS) [53], which evolve an initial sample under deterministic, discretized

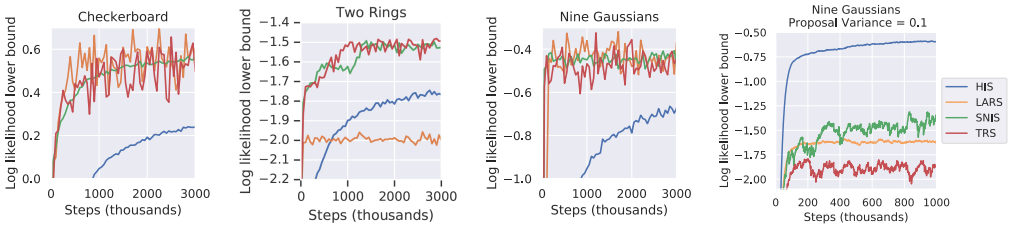

Figure 1: **Performance of LARS, TRS, SNIS, and HIS on synthetic data.** LARS, TRS, and SNIS achieve comparable data log-likelihood lower bounds on the first two synthetic datasets, whereas HIS converges slowly on these low dimensional tasks. The results for LARS on the Nine Gaussians problem match previously-reported results in [5]. We visualize the target and learned densities in Appendix Fig. 2.

Hamiltonian dynamics with a learned energy function. In particular, we sample initial location and momentum variables, and then transition the candidate sample and momentum with leap frog integration steps, changing the temperature at each step (Algorithm 3). While the quality of samples from SNIS are limited by the samples initially produced by the proposal, a model based on HIS updates the positions of the samples directly, potentially allowing for more expressive power. Intuitively, the proposal provides a coarse starting sample which is further refined by gradient optimization on the energy function. When the proposal is already quite strong, drawing additional samples as in SNIS may be advantageous.

In practice, we parameterize the temperature schedule such that $\prod_{t=0}^{T} \alpha_t = 1$. This ensures that the deterministic invertible transform from (x_0, ρ_0) to (x_T, ρ_T) has a Jacobian determinant of 1 (*i.e.*, $p(x_0, \rho_0) = p(x_T, \rho_T)$). Applying Eq. (2) yields a tractable variational objective

$$\log p_{HIS}(x_T) \geq \mathbb{E}_{q(\rho_T|x_T)}\left[\log \frac{p(x_T, \rho_T)}{q(\rho_T|x_T)}\right] = \mathbb{E}_{q(\rho_T|x_T)}\left[\log \frac{p(x_0, \rho_0)}{q(\rho_T|x_T)}\right].$$

We jointly optimize $\pi, U, \epsilon, \alpha_{0:T}$, and the variational parameters with stochastic gradient ascent. Goyal et al. [25] propose a similar approach that generates a multi-step trajectory via a learned transition operator.

Algorithm 3 HIS$(\pi, U, \epsilon, \alpha_{0:T})$ generative process

Require: Proposal distribution $\pi(x)$, energy function $U(x)$, step size ϵ, temperature schedule $\alpha_0, \ldots, \alpha_T$.

Sample $x_0 \sim \pi(x)$ and $\rho_0 \sim \mathcal{N}(0, I)$.
$\rho_0 = \alpha_0 \rho_0$
for $t = 1, \ldots T$ **do**
  $\rho_t = \rho_{t-1} - \frac{\epsilon}{2} \odot \nabla U(x_{t-1})$
  $x_t = x_{t-1} + \epsilon \odot \rho_t$
  $\rho_t = \alpha_t \left(\rho_t - \frac{\epsilon}{2} \odot \nabla U(x_t)\right)$
end for
return x_T

4 Experiments

We evaluated the proposed models on a set of synthetic datasets, binarized MNIST [42] and Fashion MNIST [67], and continuous MINST, Fashion MNIST, and CelebA [44]. See Appendix D for details on the datasets, network architectures, and other implementation details. To provide a competitive baseline, we use the recently developed Learned Accept/Reject Sampling (LARS) model [5].

4.1 Synthetic data

As a preliminary experiment, we evaluated the methods on modeling synthetic densities: a mixture of 9 equally-weighted Gaussian densities, a checkerboard density with uniform mass distributed in 8

Method	Static MNIST	Dynamic MNIST	Fashion MNIST
VAE w/ Gaussian prior	-89.20 ± 0.08	-84.82 ± 0.12	-228.70 ± 0.09
VAE w/ TRS prior	-86.81 ± 0.06	-82.74 ± 0.10	-227.66 ± 0.14
VAE w/ SNIS prior	-86.28 ± 0.14	-82.52 ± 0.03	-227.51 ± 0.09
VAE w/ HIS prior	$\mathbf{-86.00 \pm 0.05}$	$\mathbf{-82.43 \pm 0.05}$	-227.63 ± 0.04
VAE w/ LARS prior	-86.53	-83.03	$\mathbf{-227.45}^{\dagger}$
ConvHVAE w/ Gaussian prior	-82.43 ± 0.07	-81.14 ± 0.04	-226.39 ± 0.12
ConvHVAE w/ TRS prior	-81.62 ± 0.03	-80.31 ± 0.04	-226.04 ± 0.19
ConvHVAE w/ SNIS prior	$\mathbf{-81.51 \pm 0.06}$	$\mathbf{-80.19 \pm 0.07}$	$\mathbf{-225.83 \pm 0.04}$
ConvHVAE w/ HIS prior	-81.89 ± 0.02	-80.51 ± 0.07	-226.12 ± 0.13
ConvHVAE w/LARS prior	-81.70	-80.30	-225.92
SNIS w/ VAE proposal	-87.65 ± 0.07	-83.43 ± 0.07	-227.63 ± 0.06
SNIS w/ ConvHVAE proposal	$\mathbf{-81.65 \pm 0.05}$	$\mathbf{-79.91 \pm 0.05}$	$\mathbf{-225.35 \pm 0.07}$
LARS w/ VAE proposal	—	-83.63	—

Table 1: **Performance on binarized MNIST and Fashion MNIST.** We report 1000 sample IWAE log-likelihood lower bounds (in nats) computed on the test set. LARS results are copied from [5]. †We note that our implementation of the VAE (on which our models are based) underperforms the reported VAE results in [5] on Fashion MNIST.

Method	MNIST	Fashion MNIST	CelebA
Small VAE	-1258.81 ± 0.49	-2467.91 ± 0.68	-60130.94 ± 34.15
LARS w/ small VAE proposal	-1254.27 ± 0.62	-2463.71 ± 0.24	-60116.65 ± 1.14
SNIS w/ small VAE proposal	-1253.67 ± 0.29	-2463.60 ± 0.31	-60115.99 ± 19.75
HIS w/ small VAE proposal	$\mathbf{-1186.06 \pm 6.12}$	$\mathbf{-2419.83 \pm 2.47}$	$\mathbf{-59711.30 \pm 53.08}$
VAE	-991.46 ± 0.39	-2242.50 ± 0.70	-57471.48 ± 11.65
LARS w/ VAE proposal	$\mathbf{-987.62 \pm 0.16}$	$\mathbf{-2236.87 \pm 1.36}$	-57488.21 ± 18.41
SNIS w/ VAE proposal	-988.29 ± 0.20	$\mathbf{-2238.04 \pm 0.43}$	-57470.42 ± 6.54
HIS w/ VAE proposal	-990.68 ± 0.41	-2244.66 ± 1.47	$\mathbf{-56643.64 \pm 8.78}$
MAF	-1027	—	—

Table 2: **Performance on continuous MNIST, Fashion MNIST, and CelebA.** We report 1000 sample IWAE log-likelihood lower bounds (in nats) computed on the test set. As a point of comparison, we include a similar result from a 5 layer Masked Autoregressive Flow distribution [56].

squares, and two concentric rings (Fig. 1 and Appendix Fig. 2 for visualizations). For all methods, we used a unimodal standard Gaussian as the proposal distribution (see Appendix D for further details).

TRS, SNIS, and LARS perform comparably on the Nine Gaussians and Checkerboard datasets. On the Two Rings datasets, despite tuning hyperparameters, we were unable to make LARS learn the density.

On these simple problems, the target density lies in the high probability region of the proposal density, so TRS, SNIS, and LARS only have to reweight the proposal samples appropriately. In high-dimensional problems when the proposal density is mismatched from the target density, however, we expect HIS to outperform TRS, SNIS, and LARS. To test this we ran each algorithm on the Nine Gaussians problem with a Gaussian proposal of mean 0 and variance 0.1 so that there was a significant mismatch in support between the target and proposal densities. The results in the rightmost panel of Fig. 1 show that HIS was almost unaffected by the change in proposal while the other algorithms suffered considerably.

4.2 Binarized MNIST and Fashion MNIST

Next, we evaluated the models on binarized MNIST and Fashion MNIST. MNIST digits can be either statically or dynamically binarized — for the statically binarized dataset we used the binarization

from [60], and for the dynamically binarized dataset we sampled images from Bernoulli distributions with probabilities equal to the continuous values of the images in the original MNIST dataset. We dynamically binarize the Fashion MNIST dataset in a similar manner.

First, we used the models as the prior distribution in a Bernoulli observation likelihood VAE. We summarize log-likelihood lower bounds on the test set in Table 1 (referred to as VAE w/ *method* prior). SNIS outperformed LARS on static MNIST and dynamic MNIST even though it used only 1024 samples for training and evaluation, whereas LARS used 1024 samples during training and 10^{10} samples for evaluation. As expected due to the similarity between methods, TRS performed comparably to LARS. On all datasets, HIS either outperformed or performed comparably to SNIS. We increased K and T for SNIS and HIS, respectively, and find that performance improves at the cost of additional computation (Appendix Fig. 3). We also used the models as the prior distribution of a convolutional heiarachical VAE (ConvHVAE, following the architecture in [5]). In this case, SNIS outperformed all methods.

Then, we used a VAE as the proposal distribution to SNIS. A limitation of the HIS model is that it requires continuous data, so it cannot be used in this way on the binarized datasets. Initially, we thought that an unbiased, low-variance estimator could be constructed similarly to VIMCO [49], however, this estimator still had high variance. Next, we used the Gumbel Straight-Through estimator [31] to estimate gradients through the discrete samples proposed by the VAE, but found that method performed worse than ignoring those gradients altogether. We suspect that this may be due to bias in the gradients. Thus, for the SNIS model with VAE proposal, we report results on training runs which ignore those gradients. Future work will investigate low-variance, unbiased gradient estimators. In this case, SNIS again outperforms LARS, however, the performance is worse than using SNIS as a prior distribution. Finally, we used a ConvHVAE as the proposal for SNIS and saw performance improvements over both the vanilla ConvHVAE and SNIS with a VAE proposal, demonstrating that our modeling improvements are complementary to improving the proposal distribution.

4.3 Continuous MNIST, Fashion MNIST, and CelebA

Finally, we evaluated SNIS and HIS on continuous versions of MNIST, Fashion MNIST, and CelebA (64x64). We use the same preprocessing as in [18]. Briefly, we dequantize pixel values by adding uniform noise, rescale them to $[0, 1]$, and then transform the rescaled pixel values into logit space by $x \to \text{logit}(\lambda + (1 - 2\lambda)x)$, where $\lambda = 10^{-6}$. When we calculate log-likelihoods, we take into account this change of variables.

We speculated that when the proposal is already strong, drawing additional samples as in SNIS may be better than HIS. To test this, we experimented with a smaller VAE as the proposal distribution. As we expected, HIS outperformed SNIS when the proposal was weaker, especially on the more complex datasets, as shown in Table 2.

5 Variational Inference with EIMs

To provide a tractable lower bound on the log-likelihood of EIMs, we used the ELBO (Eq. (1)). More generally, this variational lower bound has been used to optimize deep generative models with latent variables following the influential work by Kingma and Welling [37], Rezende et al. [59], and models optimized with this bound have been successfully used to model data such as natural images [58, 38, 11, 26], speech and music time-series [12, 23, 39], and video [2, 28, 17]. Due to the usefulness of such a bound, there has been an intense effort to provide improved bounds [9, 48, 51, 41, 68, 50, 63]. The tightness of the ELBO is determined by the expressiveness of the variational family [69], so it is natural to consider using flexible EIMs as the variational family. As we explain, EIMs provide a conceptual framework to understand many of the recent improvements in variational lower bounds.

In particular, suppose we use a conditional EIM $q(z|x)$ as the variational family (*i.e.*, $q(z|x) = \int q(z, \lambda|x)\, d\lambda$ is the marginalized sampling process). Then, we can use the ELBO lower bound on $\log p(x)$ (Eq. (1)), however, the density of the EIM $q(z|x)$ is intractable. Agakov and Barber [1], Salimans et al. [61], Ranganath et al. [57], Maaløe et al. [46] develop an auxiliary variable

variational bound

$$\mathbb{E}_{q(z|x)}\left[\log \frac{p(x,z)}{q(z|x)}\right] = \mathbb{E}_{q(z,\lambda|x)}\left[\log \frac{p(x,z)r(\lambda|z,x)}{q(z,\lambda|x)}\right] + \mathbb{E}_{q(z|x)}\left[D_{\mathrm{KL}}\left(q(\lambda|z,x)||r(\lambda|z,x)\right]\right.$$
$$\geq \mathbb{E}_{q(z,\lambda|x)}\left[\log \frac{p(x,z)r(\lambda|z,x)}{q(z,\lambda|x)}\right], \tag{4}$$

where $r(\lambda|z,x)$ is a variational distribution meant to model $q(\lambda|z,x)$, and the identity follows from the fact that $q(z|x) = \frac{q(z,\lambda|x)}{q(\lambda|z,x)}$. Similar to Eq. (1), Eq. (4) shows the gap introduced by using $r(\lambda|z,x)$ to deal with the intractability of $q(z|x)$. We can form a lower bound on the original ELBO and thus a lower bound on the log marginal by omitting the positive D_{KL} term. This provides a tractable lower bound on the log-likelihood using flexible EIMs as the variational family and precisely characterizes the bound gap as the sum of D_{KL} terms in Eq. (1) and Eq. (4). For different choices of EIM, this bound recovers many of the recently proposed variational lower bounds.

Furthermore, the bound in Eq. (4) is closely related to partition function estimation because $\frac{p(x,z)r(\lambda|z,x)}{q(z,\lambda|x)}$ is an unbiased estimator of $p(x)$ when $z, \lambda \sim q(z,\lambda|x)$. To first order, the bound gap is related to the variance of this partition function estimator (e.g., [48]), which motivates sampling algorithms used in lower variance partition function estimators such as SMC [21] and AIS [52].

5.1 Importance Weighted Auto-encoders (IWAE)

To tighten the ELBO without explicitly expanding the variational family, Burda et al. [9] introduced the importance weighted autoencoder (IWAE) bound,

$$\mathbb{E}_{z_{1:K}\sim\prod_i \tilde{q}(z_i|x)}\left[\log\left(\frac{1}{K}\sum_{i=1}^{K}\frac{p(x,z_i)}{\tilde{q}(z_i|x)}\right)\right] \leq \log p(x). \tag{5}$$

The IWAE bound reduces to the ELBO when $K = 1$, is non-decreasing as K increases, and converges to $\log p(x)$ as $K \to \infty$ under mild conditions [9]. Bachman and Precup [3] introduced the idea of viewing IWAE as auxiliary variable variational inference and Naesseth et al. [51], Cremer et al. [13], Domke and Sheldon [20] formalized the notion.

Consider the variational family defined by the EIM based on SNIS (Algorithm 2). We use a learned, tractable distribution $\tilde{q}(z|x)$ as the proposal $\pi(z|x)$ and set $U(z|x) = \log \tilde{q}(z|x) - \log p(x,z)$ motivated by the fact that $p(z|x) \propto \tilde{q}(z|x)\exp(\log p(x,z) - \log \tilde{q}(z|x))$ is the optimal variational distribution. Similar to the variational distribution used in Section 3.2, setting

$$r(z_{1:K}, i|z, x) = \frac{1}{K}\delta_{z_i}(z)\prod_{j\neq i}\tilde{q}(z_j|x) \tag{6}$$

yields the IWAE bound Eq. (5) when plugged into to Eq. (4) (see Appendix A for details).

From Eq. (4), it is clear that IWAE is a lower bound on the standard ELBO for the EIM $q(z|x)$ and the gap is due to $D_{\mathrm{KL}}(q(z_{1:K}, i|z, x)||r(z_{1:K}, i|z, x))$. The choice of $r(z_{1:K}, i|z, x)$ in Eq. (6) was for convenience and is suboptimal. The optimal choice of r is

$$q(z_{1:K}, i|z, x) = q(i|z, x)q(z_{1:K}|i, z, x) = \frac{1}{K}\delta_{z_i}(z)q(z_{-i}|i, z, x).$$

Compared to the optimal choice, Eq. (6) makes the approximation $q(z_{-i}|i, z, x) \approx \prod_{j\neq i}\tilde{q}(z_j|x)$ which ignores the influence of z on z_{-i} and the fact that z_{-i} are not independent given z. A simple extension could be to learn a factored variational distribution conditional on z: $r(z_{1:k}, i|z, x) = \frac{1}{K}\delta_{z_i}(z)\prod_{j\neq i} r(z_j|z, x)$. Learning such an r could improve the tightness of the bound, and we leave exploring this to future work.

5.2 Semi-implicit variational inference

As a way of increasing the flexibility of the variational family, Yin and Zhou [68] introduce the idea of semi-implicit variational families. That is they define an implicit distribution $q(\lambda|x)$ by transforming a random variable $\epsilon \sim q(\epsilon|x)$ with a differentiable deterministic transformation (*i.e.*,

$\lambda = g(\epsilon, x)$). However, Sobolev and Vetrov [63] keenly note that $q(z, \lambda|x) = q(z|\lambda, x)q(\lambda|x)$ can be equivalently written as $q(z|\epsilon, x)q(\epsilon|x)$ with two explicit distributions. As a result, semi-implicit variational inference is simply auxiliary variable variational inference by another name.

Additionally, Yin and Zhou [68] provide a multi-sample lower bound on the log likelihood which is generally applicable to auxiliary variable variational inference.

$$\log p(x) \geq \mathbb{E}_{q(\lambda_{1:K-1}|x)q(z,\lambda|x)}\left[\log \frac{p(x,z)}{\frac{1}{K}\left(q(z|\lambda,x) + \sum_i q(z|\lambda_i, x)\right)}\right] \tag{7}$$

We can interpret this bound as using an EIM for $r(\lambda|z, x)$ in Eq. (4). Generally, if we introduce additional auxiliary random variables γ into $r(\lambda, \gamma|z, x)$, we can tractably bound the objective

$$\mathbb{E}_{q(z,\lambda|x)}\left[\log \frac{p(x,z)r(\lambda|z,x)}{q(z,\lambda|x)}\right] \geq \mathbb{E}_{q(z,\lambda|x)s(\gamma|z,\lambda,x)}\left[\log \frac{p(x,z)r(\lambda,\gamma|z,x)}{q(z,\lambda|x)s(\gamma|z,\lambda,x)}\right], \tag{8}$$

where $s(\gamma|z, \lambda, x)$ is a variational distribution. Analogously to the previous section, we set $r(\lambda|z, x)$ as an EIM based on the self-normalized importance sampling process with proposal $q(\lambda|x)$ and $U(\lambda|x, z) = -\log q(z|\lambda, x)$. If we choose

$$s(\lambda_{1:K}, i|z, \lambda, x) = \frac{1}{K}\delta_{\lambda_i}(\lambda)\prod_{j\neq i} q(\lambda_j|x),$$

with $\gamma = (\lambda_{1:K}, i)$, then Eq. 8 recovers the bound in [68] (see Appendix B for details). In a similar manner, we can continue to recursively augment the variational distribution s (*i.e.*, add auxiliary latent variables to s).

This view reveals that the multi-sample bound from [68] is simply one approach to choosing a flexible variational $r(\lambda|z, x)$. Alternatively, Ranganath et al. [57] use a learned variational $r(\lambda|z, x)$. It is unclear when drawing additional samples is preferable to learning a more complex variational distribution. Furthermore, the two approaches can be combined by using a learned proposal $r(\lambda_i|z, x)$ instead of $q(\lambda_i|x)$, which results in a bound described in [63].

5.3 Additional Bounds

Finally, we can also use the self-normalized importance sampling procedure to extend a proposal family $q(z, \lambda|x)$ to a larger family (instead of solely extending $r(\lambda|z, x)$) [63]. Self-normalized importance sampling is a particular choice of taking a proposal distribution and moving it closer to a target. Hamiltonian Monte Carlo [54] is another choice which can also be embedded in this framework as done by [61, 10]. Similarly, SMC can be used as a sampling procedure in an EIM and when used as the variational family, it succinctly derives variational SMC [48, 51, 41] without any instance specific tricks. In this way, more elaborate variational bounds can be constructed by specific choices of EIMs without additional derivation.

6 Discussion

We proposed a flexible, yet tractable family of distributions by treating the approximate sampling procedure of energy-based models as the model of interest, referring to them as *energy-inspired models*. The proposed EIMs bridge the gap between learning and inference in EBMs. We explore three instantiations of EIMs induced by truncated rejection sampling, self-normalized importance sampling, and Hamiltonian importance sampling and we demonstrate comparably or stronger performance than recently proposed generative models. The results presented in this paper use simple architectures on relatively small datasets. Future work will scale up both the architectures and size of the datasets.

Interestingly, as a by-product, exploiting the EIMs to define the variational family provides a unifying framework for recent improvements in variational bounds, which simplifies existing derivations, reveals potentially suboptimal choices, and suggests ways to form novel bounds.

Acknowledgments

We thank Ben Poole, Abhishek Kumar, and Diederick Kingma for helpful comments. We thank Matthias Bauer for answering implementation questions about LARS.

References

[1] Felix V Agakov and David Barber. An auxiliary variational method. In *International Conference on Neural Information Processing*, pages 561–566. Springer, 2004.

[2] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *International Conference on Learning Representations*, 2017.

[3] Philip Bachman and Doina Precup. Training deep generative models: Variations on a theme. In *NIPS Approximate Inference Workshop*, 2015.

[4] David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pages 201–208. MIT Press, 2003.

[5] Matthias Bauer and Andriy Mnih. Resampled priors for variational autoencoders. *arXiv preprint arXiv:1810.11428*, 2018.

[6] Julian Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195, 1975.

[7] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 2017.

[8] Lawrence D Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. Ims, 1986.

[9] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *nternational Conference on Learning Representations*, 2015.

[10] Anthony L Caterini, Arnaud Doucet, and Dino Sejdinovic. Hamiltonian variational auto-encoder. In *Advances in Neural Information Processing Systems*, pages 8167–8177, 2018.

[11] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *International Conference on Learning Representations*, 2016.

[12] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988, 2015.

[13] Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*, 2017.

[14] Bo Dai, Hanjun Dai, Arthur Gretton, Le Song, Dale Schuurmans, and Niao He. Kernel exponential family estimation via doubly dual embedding. *arXiv preprint arXiv:1811.02228*, 2018.

[15] Zihang Dai, Amjad Almahairi, Philip Bachman, Eduard Hovy, and Aaron Courville. Calibrating energy-based generative adversarial networks. *arXiv preprint arXiv:1702.01691*, 2017.

[16] Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The helmholtz machine. *Neural computation*, 7(5):889–904, 1995.

[17] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. *International Conference on Machine Learning*, 2018.

[18] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[19] Justin Domke and Daniel Sheldon. Divide and couple: Using monte carlo variational objectives for posterior approximation. *arXiv preprint arXiv:1906.10115*, 2019.

[20] Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. In *Advances in Neural Information Processing Systems*, pages 4471–4480, 2018.

[21] Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001.

[22] Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019.

[23] Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. Sequential neural models with stochastic layers. In *Advances in neural information processing systems*, pages 2199–2207, 2016.

[24] Yoav Freund and David Haussler. A fast and exact learning rule for a restricted class of boltzmann machines. *Advances in Neural Information Processing Systems*, 4:912–919, 1992.

[25] Anirudh Goyal Alias Parth Goyal, Nan Rosemary Ke, Surya Ganguli, and Yoshua Bengio. Variational walkback: Learning a transition operator as a stochastic recurrent net. In *Advances in Neural Information Processing Systems*, pages 4392–4402, 2017.

[26] Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. Pixelvae: A latent variable model for natural images. *International Conference on Learning Representations*, 2016.

[27] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[28] David Ha and Jürgen Schmidhuber. World models. *Advances in neural information processing systems*, 2018.

[29] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[30] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

[31] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[32] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[33] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

[34] Taesup Kim and Yoshua Bengio. Deep directed generative models with energy-based probability estimation. *arXiv preprint arXiv:1606.03439*, 2016.

[35] R. Kinderman and S.L. Snell. *Markov random fields and their applications*. American mathematical society, 1980.

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[37] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *nternational Conference on Learning Representations*, 2013.

[38] 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.

[39] Rahul G Krishnan, Uri Shalit, and David Sontag. Deep kalman filters. *arXiv preprint arXiv:1511.05121*, 2015.

[40] John D Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc., 2001.

[41] Tuan Anh Le, Maximilian Igl, Tom Rainforth, Tom Jin, and Frank Wood. Auto-encoding sequential monte carlo. *International Conference on Learning Representations*, 2017.

[42] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[43] Yann LeCun, Sumit Chopra, and Raia Hadsell. A tutorial on energy-based learning. 2006.

[44] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[45] Zhuang Ma and Michael Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. *arXiv preprint arXiv:1809.01812*, 2018.

[46] Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Auxiliary deep generative models. *arXiv preprint arXiv:1602.05473*, 2016.

[47] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[48] Chris J Maddison, Dieterich Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering variational objectives. In *Advances in Neural Information Processing Systems*, pages 6573–6583, 2017.

[49] Andriy Mnih and Danilo J Rezende. Variational inference for monte carlo objectives. *International Conference on Machine Learning*, 2016.

[50] Dmitry Molchanov, Valery Kharitonov, Artem Sobolev, and Dmitry Vetrov. Doubly semi-implicit variational inference. *arXiv preprint arXiv:1810.02789*, 2018.

[51] Christian Naesseth, Scott Linderman, Rajesh Ranganath, and David Blei. Variational sequential monte carlo. In *International Conference on Artificial Intelligence and Statistics*, pages 968–977, 2018.

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

[53] Radford M Neal. Hamiltonian importance sampling. In *In talk presented at the Banff International Research Station (BIRS) workshop on Mathematical Issues in Molecular Dynamics*, 2005.

[54] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

[55] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[56] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.

[57] Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333, 2016.

[58] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538, 2015.

[59] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

[60] Ruslan Salakhutdinov and Iain Murray. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th international conference on Machine learning*, pages 872–879. ACM, 2008.

[61] Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pages 1218–1226, 2015.

[62] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, Colorado Univ at Boulder Dept of Computer Science, 1986.

[63] Artem Sobolev and Dmitry Vetrov. Importance weighted hierarchical variational inference. In *Bayesian Deep Learning Workshop*, 2018.

[64] Jascha Sohl-Dickstein, Peter Battaglino, and Michael R DeWeese. Minimum probability flow learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 905–912. Omnipress, 2011.

[65] Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071. ACM, 2008.

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

[67] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[68] Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. *arXiv preprint arXiv:1805.11183*, 2018.

[69] Arnold Zellner. Optimal information processing and bayes's theorem. *The American Statistician*, 42(4):278–280, 1988.

[70] Song Chun Zhu, Yingnian Wu, and David Mumford. Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998.