Trading-off Learning and Inference in Deep Latent Variable Models

Daniel Levy, Stefano Ermon
Department of Computer Science
Stanford University
{danilevy,ermon}@cs.stanford.edu

Abstract

Latent variable models parameterized by deep neural networks are a popular class of probabilistic models for sampling, likelihood estimation and representation learning over large datasets. These models are often accompanied with an inference method that both enables learning as well as allows for approximate recovery of the unobserved latent variables. The quality of the inference scheme is particularly important when aiming for good performance on downstream tasks. However, current training objectives do not enable trading-off between learning –how well we model the observations– and inference –how far is our inference model from the true posterior. In this work, we introduce a new learning objective that leverages the score matching divergence and allows balancing learning and inference. We empirically evaluate this new learning objective on the MNIST dataset and demonstrate that we can achieve arbitrarily good inference (at the cost of likelihood). We also show improvement on semi-supervised learning.

1 INTRODUCTION

Latent variable models parameterized by deep neural networks have achieved great success at modeling high-dimensional, complex distributions [9, 13]. In order to learn these models, directly maximizing the log-likelihood of observed points is computationally intractable; to remedy that, one can form a variational lower bound by introducing an approximate posterior. These models can then be efficiently fitted to large datasets using stochastic variational inference [4], and jointly training both the approximate posterior and the generative model.

When training these models, several choices have to be made. First of all, we must choose both a variational family and a family of generative networks. However, in order to keep training tractable, we are often restricted to simple variational families and using approximations such as mean-field. This restriction regularly results in misspecified problems (i.e. the true posterior is not in the variational family). Recently, there has been plethora of work extending the variational family and showing the impact on the resulting generative model [14, 8, 3, 11].

In this work, we try to adopt a different viewpoint. Instead of trying to increase the expressiveness of the variational posterior, we aim at shifting some of the focus from the generative model to the inference model. Indeed, instead of interpreting maximizing the variational lower bound as a mere surrogate to maximizing the log-likelihood, we can look at the objective as a *penalized* log-likelihood where we actually value both terms: learning and inference. The question then becomes: how can we efficiently trade-off between those two terms? However, in the standard VAE framework, this is currently impossible, as recent work has shown [16].

The long-standing promise of unsupervised learning and generative models has always been the hope of discovering structure in the data; as such, being able to do efficient inference on useful posterior models remains a key challenge.

In this work, we introduce a new learning objective that allows balancing learning and inference by using a surrogate divergence: score matching [5]. This objective depends on an additional hyper-parameter λ that corresponds to the relative preference given to inference over learning. On the MNIST dataset, we empirically show that by varying this new parameter, we can arbitrarily tighten the variational lower bound. We additionally demonstrate that on the same model class, one can achieve better performance on a downstream task –here semi-supervised learning– by choosing the right λ.

2 BACKGROUND

2.1 DEEP LATENT VARIABLE MODELS

Deep latent variable models is an unsupervised modeling technique that posits the existence of latent variables from which the observation is generated. Formally, the model consists of a latent variable $z \sim p(z)$, the prior, and a conditional distribution $p(x|z)$ that generates the observation x. We often choose $p(z) = \mathcal{N}(z; 0, I)$. Given a family of parametric decoders $\{p(x|z;\theta), \theta \in \Theta\}$, and a set of samples $\mathcal{D} = \{x^{(i)}\}_{i \leq m}$, the goal is to find:

$$\theta^* = \arg\max_{\theta \in \Theta} p(\mathcal{D}; \theta). \tag{1}$$

However, this likelihood is intractable as it involves computing a high-dimensional integral. To circumvent this issue, one can introduce a variational posterior $q(z|x;\phi)$ to approximate the true (intractable) posterior $p(z|x;\theta)$. Following [9, 13], we can introduce a lower bound (ELBO) on the log-likelihood of a sample x:

$$\begin{aligned} \log p(x) &\geq \log p(x) - \mathrm{KL}(q_\phi(z|x), p_\theta(z|x)) \\ &= \mathbf{E}_{q_\phi}\left[\log p_\theta(x|z)\right] - \mathrm{KL}(q_\phi(z|x), p(z)) \\ &\triangleq \mathcal{L}(x;\phi,\theta). \end{aligned} \tag{2}$$

This lower-bound (and its gradients) can now be efficiently estimated; we can thus maximize it by jointly optimizing θ and ϕ using stochastic gradient methods.

2.2 SCORE MATCHING

Introduced by [5], the score matching divergence is defined as follows:

$$\mathrm{D_{SM}}(p, q) \triangleq \mathbf{E}_p \|\nabla_x \log p(x) - \nabla_x \log q(x)\|_2^2. \tag{3}$$

It is a *strictly* proper scoring rule as $\forall p \neq q, \mathrm{D_{SM}}(p, q) > 0$ and $\mathrm{D_{SM}}(p, p) = 0$. This divergence is particularly interesting as it can be applied to un-normalized distribution due to the fact that $\nabla_x \log p(x) = \nabla_x p(x)/p(x)$ and thus one needs to know p (and q) only up to a constant. Moreover, convergence in score matching is stronger than convergence in KL, Hellinger and Total Variation distances [6]. We additionally defer to [15] for a thorough treatment of training un-normalized models using such divergences.

3 VAE CANNOT TRADE-OFF LEARNING AND INFERENCE

The approximate posterior used in Equation 2 to maximize the log-likelihood can be interpreted in at least two different ways. One can see q_ϕ as a tool to perform optimization with, that can be discarded after training. Alternatively, we can see q_ϕ as the star of the show; allowing us to perform efficient posterior inference which could prove very useful for downstream tasks. As such, we can now see the ELBO not as a lower bound but as a *penalized* log-likelihood objective where we wish for a model that not only models the data well but where q_ϕ is a good approximation of the true posterior. As such, it is natural to introduce the following objective:

$$\mathcal{L}_\lambda(x;\phi,\theta) \triangleq \log p_\theta(x) - \lambda \mathrm{KL}(q_\phi(z|x), p_\theta(z|x)), \tag{4}$$

where we can vary $\lambda > 0$ to trade-off learning and inference. However, this objective is intractable for all $\lambda \neq 1$, in the sense that one cannot efficiently evaluate or differentiate the objective. We defer to [16] for an enumeration of tractable models.

A note on β-VAE: introduced in [2], the β-VAE model allows to trade-off between the two terms of the *rewritten* ELBO (i.e. reconstruction and regularization); this is different from balancing learning and inference.

4 SCORE MATCHING VARIATIONAL INFERENCE

In the previous section, we acknowledged that 1) we are interested in the inference gap for downstream task and 2) the VAE model cannot trade-off between learning and inference. In this section, we introduce a new objective for learning latent variable models in which we can trade-off between learning and inference.

We place ourselves in the usual setting of [9]. The prior is a unit Gaussian, the approximate posterior is a multivariate Gaussian with diagonal covariance $q(z|x) = \mathcal{N}(z; \mu_\phi(x), \mathrm{diag}(\sigma_\phi(x)))$, where μ_ϕ and σ_ϕ are neural networks. The generative part is either independent Bernoulli random variables or a diagonal Gaussian depending on the nature of the inputs; also parameterized by neural networks.

First of all, we can notice that the score matching divergence between $q_\phi(z|x)$ and $p_\theta(z|x)$ is *directly* computable without the need of additional likelihood term. Indeed, under the mean-field approximation, we can analytically compute $q_\phi(z|x)$ and thus its gradients. Similarly, $\nabla_z p_\theta(z|x)$ can also be computed as:

$$p_\theta(z|x) \propto p(z) p_\theta(x|z), \tag{5}$$

and thus we know $p_\theta(z|x)$ up to a constant. We can thus form an unbiased estimate of $\mathrm{D_{SM}}(q_\phi(z|x), p_\theta(z|x))$.

We can now define the following objective:

$$\widetilde{\mathcal{L}}_\lambda(x;\theta,\phi) \triangleq \mathcal{L}(x;\theta,\phi) - \lambda \mathrm{D}_{\mathrm{SM}}(q_\phi(z|x), p_\theta(z|x)), \tag{6}$$

which can be estimated and optimized for all $\lambda \in \mathbf{R}$. For $\lambda = 0$, we retrieve the original VAE objective, for $\lambda \to \infty$, the objective only amounts to the inference gap. This new objective effectively trades-off between learning and inference. We thus aim to solve:

$$\max_{\phi,\theta} \frac{1}{m} \sum_{i \leq m} \widetilde{\mathcal{L}}_\lambda(x^{(i)};\theta,\phi), \tag{7}$$

which can be done using stochastic gradient methods. It is important to note that our learning objective is meant to trade-off learning and inference for a *fixed* model class; as such, it can easily be combined with recent advances in increasing the expressiveness of the variational family.

We now proceed to empirically evaluate this objective in two different ways: first by measuring the ELBO gap (i.e. $\mathrm{KL}(q_\phi(z|x), p_\theta(z|x))$) and second, by evaluating the performance of semi-supervised learning for different values of λ.

5 EXPERIMENTS

We now evaluate the learning objective we presented in the previous section. We first show that, by varying the value of λ, we can arbitrarily tighten the ELBO gap, at the cost of likelihood. We also show that inference is somewhat important for downstream tasks and that sacrificing some likelihood for better inference can result in better performance.

Our encoder is a neural network with 2 fully-connected layers, with 1024 units each and softplus activation function. It outputs mean and variance for the approximate posterior. We use the same architecture for the encoder although it returns Bernoulli activation probabilities for each pixel. Our model was trained for 300 epochs with [7] and a learning rate $\alpha = 10^{-3}$. All experiments were done on the dynamically binarized MNIST dataset [10].

5.1 ELBO GAP

We start by measuring the KL between the approximate posterior and the true posterior of various latent variable models trained using varying values of λ. We also refer to it as the ELBO gap. To obtain that quantity, we measure the log-likelihood of a data point using annealed importance sampling (AIS; [12]). While AIS does not give an *exact* value for $\log p(x)$, it provides much tighter estimate. We can also verify that we have a very accurate estimate by evaluating it for log-scaled number of anneal steps and making sure the values converge. The ELBO gap can thus be measured as the difference between the AIS estimate and the value of the ELBO. We show in Figure 1, the results on both the train and test set.

We observe that, as expected, increasing the value of λ allows to arbitrarily shrink the ELBO gap, albeit at the cost of likelihood.

5.2 SEMI-SUPERVISED LEARNING

We now proceed to evaluate the effect of better inference on downstream tasks. To that aim, we first train a deep latent variable model on the full MNIST training set. We then sub-samples s examples from $\mathcal{D}$ for which we provide labels (equally distributed amongst classes). We extract features using the trained encoder of the generative model (namely $\mu_\phi(\cdot)$) and train an SVM [1] on the labeled examples. We then report the accuracy on the test set. We show the results in Figure 2 for varying values of λ as well as $s \in \{600, 1000, 3000\}$.

We observe that by increasing the preference on the inference term of the objective, we can improve the performance of the semi-supervised learning downstream task. We also see that for large values of λ, the loss of likelihood is too high and thus the learned features are not useful for the classification task.

6 DISCUSSION

In this work, we presented a new objective to allow trading-off learning and inference in deep latent variable models. Additionally, we showed that empirically, we can tighten the ELBO as well as we wish at the cost of likelihood. We also showed that sacrificing some likelihood for better inference can improve performance on a downstream task like semi-supervised learning.

Improving the inference method in latent variable models has been thoroughly explored in the literature, for example by increasing the expressiveness of the variational family of approximate posteriors. This has been done in parametric [8] and non-parametric ways [3, 11]. However, it is important to not that the aim is very different than our work. Indeed, in this work, we explore balancing the objective for a fixed model class and thus do not consider increasing the variational family; it would however be easy to combine the two.

For future work, there are several ways to use this new objective. For example, one could want to leverage more of the advantages of score matching by considering unnormalized approximate posteriors (e.g. RBM or rejection sampling). Another direction, would be to consider the objective $\widetilde{\mathcal{L}}$ as the Lagrangian of a constrained objec-

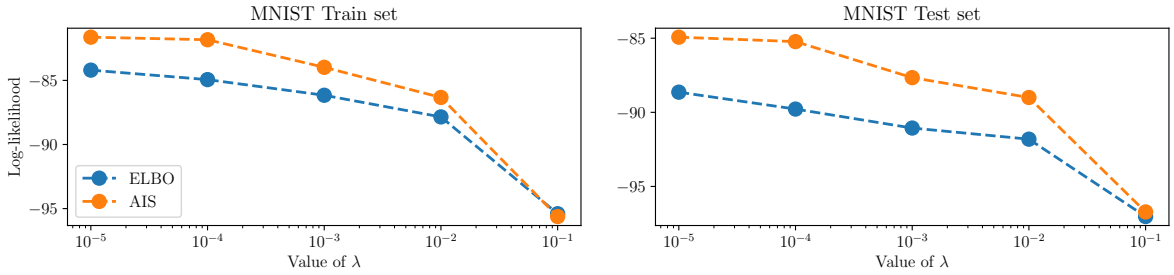

Figure 1: Increasing the value of λ in our new training objective allows us to trade-off learning and inference, i.e. arbitrarily tightening the ELBO at the cost of likelihood.

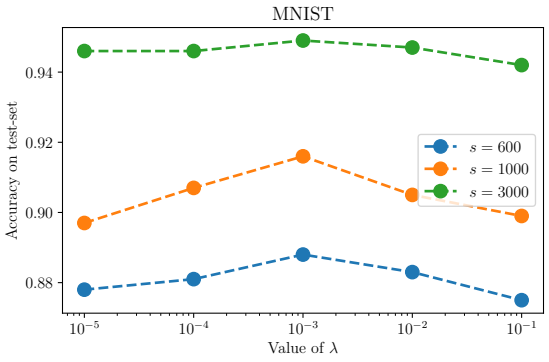

Figure 2: Semi-supervised learning performance for varying values of λ and s (size of the labeled sample). Improving inference at the cost of likelihood can improve performance on downstream tasks.

tive and optimize for λ as well (similar in spirit to [16]).

ACKNOWLEDGMENTS

We thank Matt Hoffman, Aditya Grover and Ben Poole for insightful discussions. This research was supported by Intel Corporation, TRI, a Hellman Faculty Fellowship, ONR, NSF (#1651565, #1522054, #1733686) and FLI (#2017-158687).

References

[1] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[2] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[3] Matthew D Hoffman. Learning deep latent Gaussian models with Markov chain Monte Carlo. In *International Conference on Machine Learning*, pages 1510–1519, 2017.

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

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

[6] Oliver Thomas Johnson. *Information theory and the central limit theorem*. World Scientific, 2004.

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

[8] 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.

[9] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

[11] Daniel Levy, Matthew D Hoffman, and Jascha Sohl-Dickstein. Generalizing hamiltonian monte carlo with neural networks. *arXiv preprint arXiv:1711.09268*, 2017.

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

[13] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

[14] Tim Salimans, Diederik Kingma, and Max Welling. Markov chain Monte Carlo and variational inference: Bridging the gap. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1218–1226, 2015.

[15] Jascha Sohl-Dickstein, Peter B Battaglino, and Michael R DeWeese. New method for parameter estimation in probabilistic models: minimum probability flow. *Physical review letters*, 107(22):220601, 2011.

[16] Shengjia Zhao, Jiaming Song, and Stefano Ermon. The information-autoencoding family: A lagrangian perspective on latent variable generative modeling.