Importance Weighted Hierarchical Variational Inference

Artem Sobolev[1]
asobolev@bayesgroup.ru

Dmitry Vetrov[1,2]
dvetrov@hse.ru

[1] Samsung AI Center in Moscow
[2] National Research University Higher School of Economics, Joint Samsung-HSE lab

Abstract

Variational Inference is a power tool in a Bayesian toolkit, however its effectiveness is determined by expressivity of family of variational distributions in their ability to match the true posterior distribution. However, using more expressive variational distributions is limited by the requirement of tractable density. To overcome this roadblock we introduce a new family of variational upper bounds on marginal log density in the case of latent variable models. This allows us to upper bound KL divergence and derive a family of increasingly tight variational lower bounds on the otherwise intractable standard Evidence Lower Bound for Hierarchical Variational Models, enabling the use of more expressive approximate posteriors. We show that previously known methods like HVM, SIVI and DSIVI can be seen as special cases of the proposed approach.

1 Introduction

Bayesian Inference is an important statistical tool. However, exact inference is possible only in a small class of conjugate problems, and for many practically interesting cases one has to resort to Approximate Inference techniques. Variational Inference (Wainwright et al. [2008]) being one of them is an efficient and scalable approach that gained a lot of interest in recent years due to advances in Neural Networks.

However, efficiency and accuracy of Variational Inference heavily depend on how close the approximate posterior is to the true posterior. As a result there is a lot of interest in leveraging Neural Networks' universal approximation abilities to generate powerful posterior approximates, however a major obstacle in this direction is a need for tractable density $q(z)$, which is not available in a closed form for arbitrary generators. Existing approaches to this problem fall into roughly three classes: a) approximate log-density / ratios of densities, b) augment $q(z)$ with some invertible and expressive transformation that keeps the density tractable, c) overcome the need for computation of the marginal density $q(z)$, e.g. by means of bounds. We will discuss these approaches in greater detail in section 2.

In this paper we propose a novel method that can be seen as a generalization of Hierarchical Variational Models, Semi-Implicit Variational Inference and Doubly Semi-Implicit Variational Inference. The method provides a family of tighter bounds on the marginal log-likelihood $\log p(x)$ in the case of Hierarchical Variational Model $q(z \mid x) = \int q(z \mid \psi, x) q(\psi \mid x) d\psi$, which any semi-implicit model could be transformed into. Finally, our method can be combined with a multisample bound of Burda et al. [2015] to further tighten the marginal log-likelihood estimate.

Preprint. Work in progress.

2 Related Work

More expressive variational distributions have been under active investigation for a while. One possible approach is to augment some standard $q(z)$ while keeping the density tractable using copulas (Tran et al. [2015]), mixtures (Guo et al. [2016]), invertible transformations with tractable Jacobians also known as normalizing flows (Rezende and Mohamed [2015], Kingma et al. [2016]). Flow-based models have demonstrated some success (Kingma and Dhariwal [2018]), but the requirement for invertability might lead to an inefficiency in parameters usage. Alternative direction is to embrace implicit distributions (ones we can only sample from), and overcome the need for tractable density by means of bounds or estimates (Huszár [2017]). Methods based on estimates (Mescheder et al. [2017], Shi et al. [2017]) (for example, via Density Ratio Estimation trick) typically estimate the densities indirectly, introducing bias, which prevents one from using such models during evaluation. Moreover such methods tend to hide dependency of the density $q(z)$ on variational parameters, hence biasing the optimization procedure.

Ranganath et al. [2016] proposed Hierarchical Variational Models and introduced a lower bound on (possibly differential) entropy of intractable marginal density $q(z)$ in hierarchical model $q(z) = \int q(z \mid \psi)q(\psi)d\psi$, giving an upper bound for the KL-divergence in hierarchical case. Yin and Zhou [2018] introduced Semi-Implicit Variational Inference for hierarchical models with implicit $q(\psi)$ and suggested a multisample ELBO surrogate to be used in optimization. Molchanov et al. [2018] have shown that the proposed surrogate is actually a lower bound on ELBO, and effectively gives a novel lower bound on entropy of the marginal $q(z)$. Titsias and Ruiz [2018] have shown that in gradient-based ELBO optimization in case of a hierarchical model with tractable $q_\phi(z \mid \psi)$ and $q_\phi(\psi)$ one does not need the marginal log density $\log q_\phi(z)$ per se, only the its gradient which can be estimated using MCMC. However this approach is incompatible with multisample objectives of Burda et al. [2015], Nowozin [2018] since the gradient of these objectives depends on the value of the marginal log-density.

3 Background

Having a Latent Variable Model $p_\theta(x, z) = p_\theta(x \mid z)p_\theta(z)$ for observable objects x, we're interested in two tasks: inference and learning. The problem of (bayesian) inference consists in finding the true posterior $p_\theta(z \mid x)$, which is often intractable, and thus is approximated by some $q_\phi(z \mid x)$. The problem of learning is that of finding parameters θ s.t. the marginal model distribution $p_\theta(x)$ approximates the true data-generating process of x as good as possible.

Variational Inference provides a way to solve both tasks simultaneously by lower-bounding the intractable marginal log-likelihood $\log p_\theta(x)$ using the Evidence Lower Bound (ELBO):

$$\log p_\theta(x) \geq \mathop{\mathbb{E}}_{q_\phi(z|x)} \log \frac{p_\theta(x, z)}{q_\phi(z \mid x)}$$

The bound requires analytically tractable density for both $q_\phi(z \mid x)$ and $p_\theta(z)$. The gap between the marginal log-likelihood and the bound is equal to $D_{KL}(q_\phi(z \mid x) \mid\mid p_\theta(z \mid x))$, which can be reduced by using a better posterior approximation. Burda et al. [2015] proposed a family of tighter bounds, generalizing the ELBO:

$$\log p_\theta(x) \geq \mathop{\mathbb{E}}_{q_\phi(z_{1:M}|x)} \log \frac{1}{M} \sum_{m=1}^{M} \frac{p_\theta(x, z_m)}{q_\phi(z_m \mid x)}$$

Domke and Sheldon [2018] have shown that this bound essentially corresponds to using a more powerful approximate posterior obtained using self-normalized importance sampling. However, the price of this expressivity is higher computation complexity, and thus we might want to come up with a more expressive posterior approximation. Ranganath et al. [2016] proposed to use a hierarchical variational model (HVM) for $q(z) = \int q(z \mid \psi)q(\psi)d\psi$ and to overcome the intractability of the density $q(z)$ via a lower bound

$$\log p_\theta(x) \geq \mathop{\mathbb{E}}_{q_\phi(z,\psi|x)} \log \frac{p_\theta(x,z)}{\frac{q_\phi(z,\psi|x)}{\tau(\psi|x,z)}}$$

Recently Yin and Zhou [2018] also proposed to use a hierarchical model $q(z) = \int q_\phi(z \mid \psi) q_\phi(\psi) d\psi$ with possibly implicit (but reparametrizable) $q(\psi)$, but explicit $q_\phi(z \mid \psi)$, and provided the following objective, which was shown to be a lower bound (SIVI bound) by Molchanov et al. [2018]:

$$\log p_\theta(x) \geq \mathop{\mathbb{E}}_{q_\phi(z,\psi_0|x)} \mathop{\mathbb{E}}_{q_\phi(\psi_{1:K}|x)} \log \frac{p_\theta(x,z)}{\frac{1}{K+1}\sum_{k=0}^{K} q_\phi(z \mid \psi_k, x)}$$

An important observation that we'll make use of is that many hierarhical models with implicit and reparametrizable mixing distribution $q_\phi(\psi \mid x)$ can be equivalently reformulated as a mixture of two explicit distributions: due to reparametrizability of $q_\phi(\psi \mid x)$ we have $\psi = g(\varepsilon \mid \phi, x)$ for some $\varepsilon \sim q_\phi(\varepsilon)$ with tractable density. We can then consider an equivalent hierarchical model $q_\phi(z) = \int q_\phi(z \mid g(\varepsilon \mid \phi, x), x) q_\phi(\varepsilon) d\varepsilon$ that first transforms ε into ψ and then generates samples from $q_\phi(z \mid \psi, x)$.

In the next section we derive a general variational upper bound on the marginal log density $\log q(z)$, and show that both HVM and SIVI can be derived as a special cases of a more general lower bound on marginal log-likelihood $\log p_\theta(x)$.

4 Importance Weighted Hierarchical Variational Inference

Having intractable $\log q_\phi(z)$ as a source of our problems, we seek a tractable and efficient bound, which is provided by the following theorem

Theorem (Marginal log density upper bound)**.** *For any $q(z, \psi)$, $K \in \mathbb{N}_0$ and $\tau_+(\psi \mid z)$ (under some regularity conditions) consider the following*

$$\mathcal{U}_{K_+} = \mathop{\mathbb{E}}_{q(\psi_0|z)} \mathop{\mathbb{E}}_{\tau_+(\psi_{1:K}|z)} \log \left(\frac{1}{K+1} \sum_{k=0}^{K} \frac{q(z, \psi_k)}{\tau_+(\psi_k \mid z)} \right)$$

where we write $\tau_+(\psi_{1:K} \mid z) = \prod_{k=1}^{K} \tau_+(\psi_k \mid z)$ for brevity. Then the following holds:

1. $\mathcal{U}_K \geq \log q(z)$
2. $\mathcal{U}_K \geq \mathcal{U}_{K+1}$
3. $\lim_{M\to\infty} \mathcal{U}_M = \log q(z)$

We can further combine the proposed bound with the standard Importance Weighted lower bound of Burda et al. [2015] on marginal log density to obtain a family of sandwich bounds for KL divergence in case of both $q(z)$ and $p(z)$ being different (maybe structuraly) hierarchical models (we use the same K for both bound on $q(z)$ and $p(z)$ to simplify notation, but keep in mind they might be different):

$$D_{KL}(q(z) \,||\, p(z))) \geq \mathop{\mathbb{E}}_{q(z)} \Big[\mathop{\mathbb{E}}_{\tau_-(\psi_{1:K}|z)} \log \left(\frac{1}{K} \sum_{k=1}^{K} \frac{q(z, \psi_k)}{\tau_-(\psi_k \mid z)} \right) - \mathop{\mathbb{E}}_{p(\zeta_0|z)\rho_+(\zeta_{1:K}|z)} \log \left(\frac{1}{K+1} \sum_{k=1}^{K} \frac{p(z, \zeta_k)}{\rho_+(\zeta_k \mid z)} \right) \Big] \tag{1}$$

$$D_{KL}(q(z) \,||\, p(z))) \leq \mathop{\mathbb{E}}_{q(z,\psi_0)} \Big[\mathop{\mathbb{E}}_{\tau_+(\psi_{1:K}|z)} \log \left(\frac{1}{K+1} \sum_{k=0}^{K} \frac{q(z, \psi_k)}{\tau_+(\psi_k \mid z)} \right) - \mathop{\mathbb{E}}_{\rho_-(\zeta_{1:K}|z)} \log \left(\frac{1}{K} \sum_{k=1}^{K} \frac{p(z, \zeta_k)}{\rho_-(\zeta_k \mid z)} \right) \Big] \tag{2}$$

The proposed bound lets us give a lower bound on ELBO for models with latent variable model $q(z \mid x)$ and $p(z)$, leading to **Importance Weighted Hierarchical Variational Inference** (IWHVI):

$$\begin{aligned}\log p(x) &\geq \mathop{\mathbb{E}}_{q(z|x)} \log \frac{p(x,z)}{q(z \mid x)} = \mathop{\mathbb{E}}_{q(z|x)} \left[\log p(x \mid z) + \log p(z) - \log q(z \mid x)\right] \\ &\geq \mathop{\mathbb{E}}_{q(\psi_0|x)} \mathop{\mathbb{E}}_{q(z|x,\psi_0)} \Bigg[\log p(x \mid z) - \mathop{\mathbb{E}}_{\tau_+(\psi_{1:K}|z,x)} \log \left(\frac{1}{K+1} \sum_{k=0}^{K} \frac{q(z, \psi_k \mid x)}{\tau_+(\psi_k \mid z, x)}\right) \\ &\quad + \mathop{\mathbb{E}}_{\rho_-(\zeta_{1:K}|z)} \log \left(\frac{1}{K} \sum_{k=1}^{K} \frac{p(z, \zeta_k)}{\rho_-(\zeta_k \mid z)}\right)\Bigg] \end{aligned} \tag{3}$$

This bound introduces two additional variational distributions ρ and τ that are learned by maximizing the bound w.r.t. their parameters, tightening the bound. While the optimal distributions are $\tau(\psi \mid z) = q(\psi \mid z)$ and $\rho(\zeta \mid z) = p(\zeta \mid z)$, one can see that for some choices of these distributions the bound leads to previously known methods. In particular:

- For arbitrary K, $\tau_+(\psi \mid z, x) = q(\psi \mid x)$ and $\rho_-(\zeta \mid z) = p(\zeta)$ we recover DSIVI bound (Molchanov et al. [2018])
$$\log p(x) \geq \mathop{\mathbb{E}}_{q(z,\psi_0|x)} \mathop{\mathbb{E}}_{q(\psi_{1:K}|x)} \mathop{\mathbb{E}}_{p(\zeta_{1:K})} \log \frac{p(x \mid z) \frac{1}{K} \sum_{k=1}^{K} p(z \mid \zeta_k)}{\frac{1}{K+1} \sum_{k=0}^{K} q(z \mid \psi_k, x)}$$
- For arbitrary K, $\tau_+(\psi \mid z, x) = q(\psi \mid x)$ and explicit prior $p(z)$ we recover SIVI bound (Yin and Zhou [2018])
$$\log p(x) \geq \mathop{\mathbb{E}}_{q(z,\psi_0|x)} \mathop{\mathbb{E}}_{q(\psi_{1:K}|x)} \log \frac{p(x,z)}{\frac{1}{K+1} \sum_{k=0}^{K} q(z \mid \psi_k, x)}$$
- For $K = 0$, arbitrary $\tau_+(\psi \mid z, x)$ and explicit prior $p(z)$ we recover HVM bound (Ranganath et al. [2016])
$$\log p(x) \geq \mathop{\mathbb{E}}_{q(z,\psi_0|x)} \log \frac{p(x,z)}{\frac{q(z,\psi_0|x)}{\tau_+(\psi_0|z,x)}}$$
- For arbitrary K, factorized inference and prior models $q_\phi(z, \psi \mid x) = q_\phi(z \mid x) q_\phi(\psi \mid x)$, $p_\theta(z, \zeta) = p_\theta(z) p_\theta(\zeta)$, optimal $\tau_+(\psi \mid z, x) = q_\phi(\psi \mid x)$ and $\rho_-(\zeta \mid z) = p_\theta(\zeta)$ we recover the standard ELBO
$$\log p(x) \geq \mathop{\mathbb{E}}_{q(z|x)} \log \frac{p(x,z)}{q_\phi(z \mid x)}$$

Moreover, we tighten the bound (3) further in a way similar to Burda et al. [2015] (Theorem A.4), leading to **Doubly Importance Weighted Hierarchical Variational Inference** (DIWHVI).

$$\log p(x) \geq \mathbb{E}\left[\log \frac{1}{M} \sum_{m=1}^{M} \frac{p(x \mid z_m) \frac{1}{K} \sum_{k=1}^{K} \frac{p(z_m, \zeta_{m,k})}{\rho_-(\zeta_{m,k}|z_m)}}{\frac{1}{K+1} \sum_{k=0}^{K} \frac{q(z_m, \psi_{m,k}|x)}{\tau_+(\psi_{m,k}|z_m, x)}}\right] \tag{4}$$

Where the expectation is taken over the following generative process:

1. Sample $\psi_{m,0} \sim q(\psi \mid x)$ for $1 \leq m \leq M$
2. Sample $z_m \sim q(z \mid x_n, \psi_{m,0})$ for $1 \leq m \leq M$
3. Sample $\psi_{m,k} \sim \tau_+(\psi \mid z_m, x)$ for $1 \leq m \leq M$ and $1 \leq k \leq K$
4. Sample $\zeta_{m,k} \sim \rho_-(\zeta \mid z_m)$ for $1 \leq m \leq M$ and $1 \leq k \leq K$

The downside of this tighter bound is increased sampling complexity: it requires $M(1 + K)$ samples of ψ and MK samples of ζ. However, as has been shown in Yin and Zhou [2018] when $\tau_+(\psi \mid z, x) = \tau_+(\psi \mid x)$ and $\rho_-(\zeta \mid z) = \rho_-(\zeta)$ (independent of z, but not necessarily equal to the prior), it's possible to reuse samples $\psi_{m,1:K}$ and $\zeta_{m,1:K}$ for different m at the expense of a looser bound and possibly higher variance, thus bringing sample complexity down to $M + 2K$.

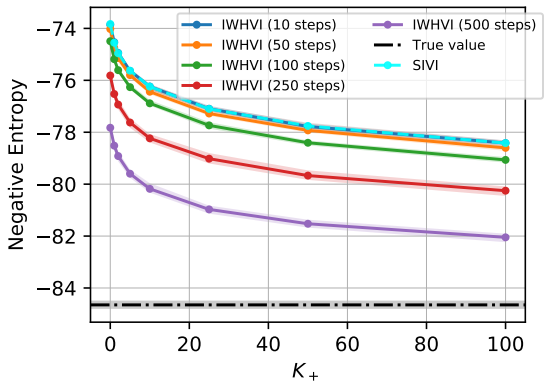

Figure 1: Negative entropy bound for 50-dimensional Laplace distribution. Shaded area denotes 90 % confidence interval

Method	test log-likelihood
DIWHVI	-92.67
IWHVI	-93.41
SIVI-IW	-93.16
SIVI	-93.86
UIVI	-94.09

Table 1: log-likelihood on static binarization on MNIST

5 Experiments

5.1 Toy Experiment

As a toy experiment we consider 50-dimensional factorized standard Laplace distribution $q(z)$ as a hierarchical scale-mixture model:

$$q(z) = \prod_{d=1}^{50} \text{Laplace}(z_d \mid 0, 1) = \int \prod_{d=1}^{50} \mathcal{N}(z_d \mid 0, \psi_d)\text{Exp}(\psi_d \mid \tfrac{1}{2}) d\psi_{1:50}$$

We do not make use of factorized join distribution $p(z, \psi)$ to explore bound's behaviour in high dimensions. We use the proposed bound from Theorem A.1 and compare it to SIVI (Yin and Zhou [2018]) on the task of upper-bounding the negative differential entropy $\mathbb{E}_{q(z)} \log q(z)$. For IWHVI we take $\tau_+(\psi \mid z)$ to be a Gamma distribution whose concentration and rate are generated by a neural network with 3 500-dimensional hidden layers from z. We initialize the network at prior, namely, we also add a sigmoid "gate" output with large initial negative bias and use the gate to combine prior concentration and rate with those generated by the network. This way we're guaranteed to perform no worse than SIVI even at a randomly initialized τ_+. Figure 1 shows the value of the bound for different number of optimization steps over τ parameters, minimizing the bound. The whole process (including random initialization of neural networks) was repeated 50 times to compute empirical 90% confidence intervals.

5.2 Variational Autoencoder

For VAE we follow Titsias and Ruiz [2018] setup: we use single stochastic layer with $p(z) = \mathcal{N}(z \mid 0, I)$ prior, decoder $p_\theta(x \mid z) = \text{Bernoulli}(x \mid \pi_\theta(z))$ where π_θ is a neural network with two hidden 200-neurons layers, and latent variable model encoder $q_\phi(z \mid x) = \int \mathcal{N}(z \mid \mu_\phi(x, \psi), \sigma^2_\phi(x, \psi))\mathcal{N}(\psi \mid 0, 1)d\psi$ where $\mu_\phi(x, \psi)$ and $\sigma^2_\phi(x, \psi)$ are outputs of a neural network with two hidden 200-neurons layers. We take $\tau_\vartheta(\psi \mid z, x) = \mathcal{N}(\psi \mid \nu_\vartheta(x, z), \varsigma^2_\vartheta(x, z))$ where mean and variance are outputs of another neural network with two hidden 200-neurons layers. We take $z, \psi \in \mathbb{R}^{10}$, and evaluate the VAE on the problem of generative modeling in terms of marginal log-likelihood on the MNIST dataset (Salakhutdinov and Murray [2008]).

We used the proposed bound eq. (4) with analytically tractable prior $p(z) = \mathcal{N}(z \mid 0, 1)$ with increasing number K: we used $K = 1$ for first 250 epochs, $K = 5$ for next 250 epochs, and $K = 20$ for the rest 500 epochs. For DIWHVI and SIVI-IW we used the same schedule for the number of IWAE samples M, while for other methods we always used just 1 sample of z per x.

To estimate the marginal log-likelihood in a way comparable to prior work, we use the following lower [1] bound for $M = 1000$, $K = 500$ [2]. Results are shown in table 1.

$$\log p(x_n) \geq \log \frac{1}{M} \sum_{m=1}^{M} \frac{p(x_n, z_{n,m})}{\frac{1}{K+1} \sum_{k=0}^{K} q(z_{n,m} \mid \psi_{n,m,k}^{\tau} \mid x_n)} \tag{5}$$

6 Conclusion

We presented a variational upper bound on marginal log density, which allowed us to upper bound $D_{KL}(q(x) \mid\mid p(x))$ for the case of latent variable model $q(z)$ in addition to prior works that only provided upper bounds for the case of latent variable model $p(z)$. We applied it to lower bound the intractable ELBO with a tractable one for the case of latent variable model approximate posterior $q(z \mid x)$. We combined the resulting bound with IWAE-like lower bound, which led to a tighter bound of the marginal log-likelihood. Proposed variational inference method allows the use of much more expressive approximate posterior, which will be useful for many variational models.

Acknowledgements

Authors would like to thank Aibek Alanov, Dmitry Molchanov and Oleg Ivanov for valuable discussions and feedback.

References

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Dustin Tran, David Blei, and Edo M Airoldi. Copula variational inference. In *Advances in Neural Information Processing Systems*, pages 3564–3572, 2015.

Fangjian Guo, Xiangyu Wang, Kai Fan, Tamara Broderick, and David B Dunson. Boosting variational inference. *arXiv preprint arXiv:1611.05559*, 2016.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4743–4751. Curran Associates, Inc., 2016. URL `http://papers.nips.cc/paper/6581-improved-variational-inference-with-inverse-autoregressive-flow.pdf`.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10235–10244. Curran Associates, Inc., 2018. URL `http://papers.nips.cc/paper/8224-glow-generative-flow-with-invertible-1x1-convolutions.pdf`.

Ferenc Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning (ICML)*, 2017.

[1] Unlike the estiamte used in Titsias and Ruiz [2018] this expression is guaranteed to be a lower bound

[2] Using fewer samples K only makes the estimate worse

Jiaxin Shi, Shengyang Sun, and Jun Zhu. Kernel implicit variational inference. *arXiv preprint arXiv:1705.10119*, 2017.

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

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

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

Michalis K Titsias and Francisco JR Ruiz. Unbiased implicit variational inference. *arXiv preprint arXiv:1808.02078*, 2018.

Sebastian Nowozin. Debiasing evidence approximations: On importance-weighted autoencoders and jackknife variational inference. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=HyZoi-WRb`.

Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4471–4480. Curran Associates, Inc., 2018. URL `http://papers.nips.cc/paper/7699-importance-weighting-and-variational-inference.pdf`.

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.

A Proofs

Theorem A.1 (Marginal log-density upper bound)**.** *For any $p(z, \psi)$, $K_+ \in \mathbb{N}_0$ and $\tau_+(\psi \mid z)$ consider the following holds*

$$\mathcal{U}_{K_+} = \mathop{\mathbb{E}}_{p(\psi_0|z)} \mathop{\mathbb{E}}_{\tau_+(\psi_{1:K_+}|z)} \log \left(\frac{1}{K_+ + 1} \sum_{k=0}^{K_+} \frac{p(z, \psi_k)}{\tau_+(\psi_k \mid z)} \right)$$

where we write $\tau_+(\psi_{1:K_+} \mid z) = \prod_{k=1}^{K_+} \tau_+(\psi_k \mid z)$ for brevity. Then the following holds:

1. $\mathcal{U}_{K_+} \geq \log p(z)$
2. $\mathcal{U}_{K_+} \geq \mathcal{U}_{K_+ + 1}$
3. $\lim_{M \to \infty} \mathcal{U}_M = \log p(z)$ *if* $\mathbb{E}_{\tau_+(\psi|z)} \frac{p(z,\psi)}{\tau_+(\psi|z)} < \infty$

Proof. 1. Consider a gap between the proposed bound at the marginal log density:

$$\begin{aligned}
\text{Gap} &= \mathop{\mathbb{E}}_{p(\psi_0|z)} \mathop{\mathbb{E}}_{\tau_+(\psi_{1:K_+}|z)} \log \left(\frac{1}{K_+ + 1} \sum_{k=0}^{K_+} \frac{p(z, \psi_k)}{\tau_+(\psi_k \mid z)} \right) - \log p(z) \\
&= \mathop{\mathbb{E}}_{p(\psi_0|z)} \mathop{\mathbb{E}}_{\tau_+(\psi_{1:K_+}|z)} \log \left(\frac{1}{K_+ + 1} \sum_{k=0}^{K_+} \frac{p(\psi_k \mid z)}{\tau_+(\psi_k \mid z)} \right) \\
&= \mathop{\mathbb{E}}_{p(\psi_0|z)} \mathop{\mathbb{E}}_{\tau_+(\psi_{1:K_+}|z)} \log \left(\frac{p(\psi_0 \mid z)\tau_+(\psi_{1:K_+} \mid z)}{\omega(\psi_{0:K_+} \mid z)} \right) \\
&= D_{KL}(p(\psi_0 \mid z)\tau_+(\psi_{1:K_+} \mid z) \mid\mid \omega(\psi_{0:K_+} \mid z)) \geq 0
\end{aligned}$$

Where the last line holds due to ω being a normalized density function (see Lemma A.2):

$$\omega(\psi_{0:K_+} \mid z) = \frac{p(\psi_0 \mid z)\tau_+(\psi_{1:K_+} \mid z)}{\frac{1}{K_+ + 1} \sum_{k=0}^{K_+} \frac{p(\psi_k|z)}{\tau_+(\psi_k|z)}}$$

2. Now we will prove the second claim.

$$\begin{aligned}
\mathcal{U}_{K_+} - \mathcal{U}_{K_+ + 1} &= \mathbb{E}_{p(\psi_0|z)} \mathbb{E}_{\tau_+(\psi_{1:K_+ + 1}|z)} \log \frac{\frac{1}{K_+ + 1} \sum_{k=0}^{K_+} \frac{p(z,\psi_k)}{\tau_+(\psi_k|z)}}{\frac{1}{K_+ + 2} \sum_{k=0}^{K_+ + 1} \frac{p(z,\psi_k)}{\tau_+(\psi_k|z)}} \\
&= \mathbb{E}_{p(\psi_0|z)} \mathbb{E}_{\tau_+(\psi_{1:K_+ + 1}|z)} \log \frac{p(\psi_0 \mid z)\tau_+(\psi_{1:K_+ + 1} \mid z)}{\nu(\psi_{0:K_+ + 1} \mid z)} \\
&= D_{KL} \left(p(\psi_0 \mid z)\tau_+(\psi_{1:K_+ + 1} \mid z) \mid\mid \nu(\psi_{0:K_+ + 1} \mid z) \right) \geq 0
\end{aligned}$$

Where we used the fact that $\nu_{\tau_+}(\psi_{0:K_+ + 1} \mid z)$ is normalized density due to Lemma A.3

$$\nu_{\tau_+}(\psi_{0:K_+ + 1} \mid z) = \omega(\psi_{0:K_+} \mid z)\tau_+(\psi_{K_+ + 1} \mid z) \frac{1}{K_+ + 2} \sum_{k=0}^{K_+ + 1} \frac{p(\psi_k \mid z)}{\tau_+(\psi_k \mid z)}$$

3. For the last claim we follow Burda et al. [2015]. Consider

$$M_{K_+} = \frac{1}{K_+ + 1} \sum_{k=0}^{K_+} \frac{p(z, \psi_k)}{\tau_+(\psi_k \mid z)} = \overbrace{\frac{1}{K_+ + 1} \frac{p(z, \psi_0)}{\tau_+(\psi_0 \mid z)}}^{A_{K_+}} + \overbrace{\frac{K_+}{K_+ + 1}}^{B_{K_+}} \overbrace{\frac{1}{K_+} \sum_{k=1}^{K_+} \frac{p(z, \psi_k)}{\tau_+(\psi_k \mid z)}}^{X_{K_+}}$$

We have

$$A_{K_+} \xrightarrow[K_+\to\infty]{a.s.} 0, \qquad X_{K_+} \xrightarrow[K_+\to\infty]{a.s.} \mathop{\mathbb{E}}_{\tau_+(\psi|z)} \frac{p(z,\psi)}{\tau_+(\psi \mid z)} = p(z), \qquad B_{K_+} \xrightarrow[K_+\to\infty]{a.s.} 1$$

Thus

$$M_{K_+} \xrightarrow[K_+\to\infty]{a.s.} p(z), \qquad \mathcal{U}_{K_+} = \mathop{\mathbb{E}}_{\tau_+(\psi_{0:K_+}|z)} \log M_{K_+} \xrightarrow[K_+\to\infty]{} \log p(z)$$

□

Lemma A.2 (ω distribution, following Domke and Sheldon [2018])**.** *Given z, consider a following generative process:*

- *Sample $K+1$ i.i.d. samples from $\hat{\psi}_k \sim \tau(\psi \mid z)$*
- *For each sample compute its weight $w_k = \frac{p(\hat{\psi}_k, z)}{\tau(\hat{\psi}_k|z)}$*
- *Sample $h \sim$ Cat $\left(\frac{w_0}{\sum_{k=0}^K w_k}, \dots, \frac{w_K}{\sum_{k=0}^K w_k}\right)$*
- *Put h-th sample first, and then the rest: $\psi_0 = \hat{\psi}_h$, $\psi_{1:K} = \hat{\psi}_{\setminus h}$*

Then the marginal density of $\psi_{0:K}$

$$\omega_\tau(\psi_{0:K} \mid z) = \frac{p(\psi_0 \mid z)\tau(\psi_{1:K} \mid z)}{\frac{1}{K+1}\sum_{k=0}^K \frac{p(\psi_k|z)}{\tau(\psi_k|z)}}$$

Proof. The joint density for the generative process described above is

$$\omega_\tau(\hat{\psi}_{0:K}, h, \psi_{0:K} \mid z) = \tau(\hat{\psi}_{0:K} \mid z)\frac{w_h}{\sum_{k=0}^K w_k}\delta(\psi_0 - \hat{\psi}_h)\delta(\psi_{1:K} - \hat{\psi}_{\setminus h})$$

One can see that this is indeed a normalized density

$$\begin{aligned}\int \sum_{h=0}^K \left(\int \omega_\tau(\hat{\psi}_{0:K}, h, \psi_{0:K} \mid z) d\psi_{0:K}\right) d\hat{\psi}_{0:K} &= \int \sum_{h=0}^K \tau(\hat{\psi}_{0:K} \mid z)\frac{w_h}{\sum_{k=0}^K w_k} d\hat{\psi}_{0:K} \\ &= \int \tau(\hat{\psi}_{0:K} \mid z) \sum_{h=0}^K \frac{w_h}{\sum_{k=0}^K w_k} d\hat{\psi}_{0:K} = \int \tau(\hat{\psi}_{0:K} \mid z) d\hat{\psi}_{0:K} = 1\end{aligned}$$

The marginal density $\omega_\tau(\psi_{0:K} \mid z)$ then is

$$\begin{aligned}\omega_\tau(\psi_{0:K} \mid z) &= \int \sum_{h=0}^K \tau(\hat{\psi}_{0:K} \mid z)\frac{w_h}{\sum_{k=0}^K w_k}\delta(\psi_0 - \hat{\psi}_h)\delta(\psi_{1:K} - \hat{\psi}_{\setminus h}) d\hat{\psi}_{0:K} \\ &= (K+1)\int \tau(\hat{\psi}_{0:K} \mid z)\frac{w_0}{\sum_{k=0}^K w_k}\delta(\psi_0 - \hat{\psi}_0)\delta(\psi_{1:K} - \hat{\psi}_{1:K}) d\hat{\psi}_{0:K} \\ &= \int \tau(\hat{\psi}_{1:K} \mid z)\frac{p(z,\hat{\psi}_0)}{\frac{1}{K+1}\sum_{k=0}^K w_k}\delta(\psi_0 - \hat{\psi}_0)\delta(\psi_{1:K} - \hat{\psi}_{1:K}) d\hat{\psi}_{0:K} \\ &= \tau(\psi_{1:K} \mid z)\frac{p(z,\psi_0)}{\frac{1}{K+1}\sum_{k=0}^K \frac{p(\psi_k,z)}{\tau(\psi_k|z)}} = \frac{p(\psi_0|z)\tau(\psi_{1:K}|z)}{\frac{1}{K+1}\sum_{k=0}^K \frac{p(\psi_k|z)}{\tau(\psi_k|z)}}\end{aligned}$$

Where on the second line we used the fact that integrand is symmetric under the choice of h.

□

Lemma A.3. *Let*

$$\nu_\tau(\psi_{0:K+1} \mid z) = \omega_\tau(\psi_{0:K} \mid z)\tau(\psi_{K+1} \mid z)\frac{1}{K+2}\sum_{k=0}^{K+1}\frac{p(\psi_k \mid z)}{\tau(\psi_k \mid z)}$$

Then $\nu_\tau(\psi_{0:K+1} \mid z)$ is a normalized density.

Proof. $\nu_\tau(\psi_{0:K+1} \mid z)$ is non-negative due to all the terms being non-negative. Now we'll show it integrates to 1 (colors denote corresponding terms):

$$\begin{aligned}
&\int \omega_\tau(\psi_{0:K} \mid z)\tau(\psi_{K+1} \mid z)\frac{1}{K+2}\sum_{k=0}^{K+1}\frac{p(\psi_k \mid z)}{\tau(\psi_k \mid z)}d\psi_{0:K+1} \\
&= \frac{1}{K+2}\int \omega_\tau(\psi_{0:K} \mid z)\left[\sum_{k=0}^{K}\frac{p(\psi_k \mid z)}{\tau(\psi_k \mid z)} + \int \tau(\psi_{K+1} \mid z)\frac{p(\psi_{K+1} \mid z)}{\tau(\psi_{K+1} \mid z)}d\psi_{K+1}\right]d\psi_{0:K} \\
&= \frac{1}{K+2}\left[\int \frac{p(\psi_0 \mid z)\tau(\psi_{1:K} \mid z)}{\frac{1}{K+1}\sum_{k=0}^{K}\frac{p(\psi_k|z)}{\tau(\psi_k|z)}}\sum_{k=0}^{K}\frac{p(\psi_k \mid z)}{\tau(\psi_k \mid z)}d\psi_{0:K} + 1\right] = \frac{K+1+1}{K+2} = 1
\end{aligned}$$

□

Theorem A.4.

$$\log p(x) \geq \mathbb{E}\left[\log \frac{1}{M}\sum_{m=1}^{M}\frac{p(x \mid z_m)\frac{1}{K_-}\sum_{k=1}^{K_-}\frac{p(z_m,\psi^\rho_{m,k})}{\rho_-(\psi^\rho_{m,k}|z_m)}}{\frac{1}{K_++1}\sum_{k=0}^{K_+}\frac{q(z_m,\psi^\tau_{m,k})}{\tau_+(\psi^\tau_{m,k}|z_m)}}\right] \tag{6}$$

Proof. Consider a random variable

$$X_M = \frac{1}{M}\sum_{m=1}^{M}\frac{p(x \mid z_m)\frac{1}{K_-}\sum_{k=1}^{K_-}\frac{p(z_m,\psi^\rho_{m,k})}{\rho_-(\psi^\rho_{m,k}|z_m)}}{\frac{1}{K_++1}\sum_{k=0}^{K_+}\frac{q(z_m,\psi^\tau_{m,k}|x)}{\tau_+(\psi^\tau_{m,k}|z_m,x)}}$$

We'll show it's unbiased estimate of $p(x)$ (colors denote corresponding terms):

$$\begin{aligned}
\mathbb{E}\, X_M &= \int \left[\left(\prod_{m=1}^{M} q(\psi^{\tau}_{m,0} \mid x) q(z_m \mid \psi^{\tau}_{m,0}, x) \tau_+(\psi^{\tau}_{m,1:K_+} \mid z_m, x) \rho_-(\psi^{\rho}_{m,1:K_-} \mid z_m) \right) \right. \\
&\qquad \left. \frac{1}{M} \sum_{m=1}^{M} \frac{p(x \mid z_m) \frac{1}{K_-} \sum_{k=1}^{K_-} \frac{p(z_m, \psi^{\rho}_{m,k})}{\rho_-(\psi^{\rho}_{m,k} \mid z_m)}}{\frac{1}{K_+ + 1} \sum_{k=0}^{K_+} \frac{q(z_m, \psi^{\tau}_{m,k} \mid x)}{\tau_+(\psi^{\tau}_{m,k} \mid z_m, x)}} \right] d\psi^{\tau}_{1:M,0:K_+} d\psi^{\rho}_{1:M,1:K_-} dz_{1:M} \\
&= \int \left[\left(\prod_{m=1}^{M} q(\psi^{\tau}_{m,0} \mid x) q(z_m \mid \psi^{\tau}_{m,0}, x) \tau_+(\psi^{\tau}_{m,1:K_+} \mid z_m, x) \right) \right. \\
&\qquad \left. \frac{1}{M} \sum_{m=1}^{M} \frac{p(x \mid z_m) \mathbb{E}_{\rho_-(\psi^{\rho}_{m,1:K_-} \mid z_m)} \frac{1}{K_-} \sum_{k=1}^{K_-} \frac{p(z_m, \psi^{\rho}_{m,k})}{\rho_-(\psi^{\rho}_{m,k} \mid z_m)}}{\frac{1}{K_+ + 1} \sum_{k=0}^{K_+} \frac{q(z_m, \psi^{\tau}_{m,k} \mid x)}{\tau_+(\psi^{\tau}_{m,k} \mid z_m, x)}} \right] d\psi^{\tau}_{1:M,0:K_+} dz_{1:M} \\
&= \int \left[\left(\prod_{m=1}^{M} q(\psi^{\tau}_{m,0} \mid x) q(z_m \mid \psi^{\tau}_{m,0}, x) \tau_+(\psi^{\tau}_{m,1:K_+} \mid z_m, x) \right) \right. \\
&\qquad \left. \frac{1}{M} \sum_{m=1}^{M} \frac{p(x \mid z_m) p(z_m)}{\frac{1}{K_+ + 1} \sum_{k=0}^{K_+} \frac{q(z_m, \psi^{\tau}_{m,k} \mid x)}{\tau_+(\psi^{\tau}_{m,k} \mid z_m, x)}} \right] d\psi^{\tau}_{1:M,0:K_+} dz_{1:M} \\
&= \frac{1}{M} \sum_{m=1}^{M} \int p(x, z_m) \frac{q(z_m, \psi^{\tau}_{m,0} \mid x) \tau_+(\psi^{\tau}_{m,1:K_+} \mid z_m)}{\frac{1}{K_+ + 1} \sum_{k=0}^{K_+} \frac{q(z_m, \psi^{\tau}_{m,k} \mid x)}{\tau_+(\psi^{\tau}_{m,k} \mid z_m, x)}} d\psi^{\tau}_{m,0:K_+} dz_m \\
&= \frac{1}{M} \sum_{m=1}^{M} \int p(x, z_m) \omega(\psi^{\tau}_{m,0:K_+} \mid z_m, x) d\psi^{\tau}_{m,0:K_+} dz_m = \frac{1}{M} \sum_{m=1}^{M} \int p(x, z_m) dz_m \\
&= p(x)
\end{aligned}$$

□

Corollary A.4.1. *All statements of Theorem 1 of Burda et al. [2015] apply to this bounds as well.*