# Wild Variational Approximations

**Yingzhen Li**
University of Cambridge
Cambridge, CB2 1PZ, UK
yl494@cam.ac.uk

**Qiang Liu**
Dartmouth College
Hanover, NH 03755, USA
qiang.liu@dartmouth.edu

## Abstract

We formalise the research problem of approximate inference in the wild: developing new variants of variational methods that work for arbitrary variational approximation families for which inference (e.g., sampling or calculating expectation) is tractable, but probability density function may be intractable. We provide several examples for this type of approximations, discuss energy/gradient approximation for existing techniques, and further comment on developing other variational objectives and amortising stochastic dynamics. Connections and comparisons to existing approaches are also briefly discussed.

## 1 Introduction

For many machine learning tasks, a probabilistic model is fitted to the underlying distribution of the observed data. In the following we discuss w.l.o.g. latent variable models denoted by $p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})$ with prior $p_0(\boldsymbol{z})$, although the presented approaches extend to the general case. Here $\boldsymbol{z}$ denotes the latent variables that a Bayesian approach would integrate out, e.g. the latent representation of deep generative models and the weight matrices of Bayesian neural networks. The hyper-parameters are loaded in $\boldsymbol{\theta}$ which will be learned by (approximate) maximum likelihood estimation (MLE), which requires the marginal likelihood $p(\boldsymbol{x}|\boldsymbol{\theta})$. Also given an observation $\boldsymbol{x}$, inference requires computing the exact posterior $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{p(\boldsymbol{x}|\boldsymbol{\theta})} p_0(\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})$. Through out this paper we assume the log-likelihood terms are tractable, but even so both quantities are intractable and hence require approximations in most cases, .

Practical approaches for Bayesian inference include sampling-based and optimisation-based methods. Sampling-based methods, e.g. Markov chain Monte Carlo (MCMC) [1, 2, 3, 4] approximate these quantities by drawing samples from the exact posterior $\boldsymbol{z}^k \sim p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})$ and use them later for inference and prediction. However most of these approaches are unbiased only in asymptotes, and in practice they can be computationally challenging for big models. In contrast, optimisation-based methods provide fast yet powerful tools for Bayesian inference of large scale. These methods explicitly define an approximate posterior distribution $q(\boldsymbol{z}|\boldsymbol{x})$, fit it to the exact posterior by optimising some objective function $\mathcal{L}(\boldsymbol{\theta}, q; \boldsymbol{x})$, and replaces the exact posterior with $q(\boldsymbol{z}|\boldsymbol{x})$ in inference/prediction time. An example of optimisation-based methods is variational inference (VI) [5, 6], which maximises the *variational lower-bound* in some distribution family $\mathcal{Q}$:

$$\max_{q \in \mathcal{Q}} \mathcal{L}_{\text{VI}}(\boldsymbol{\theta}, q; \boldsymbol{x}) = \mathbb{E}_q \left[ \log \frac{p_0(\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z}|\boldsymbol{x})} \right]. \tag{1}$$

Maximising this lower-bound is equivalent to minimising the KL-divergence $\text{KL}[q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})]$. Furthermore this lower-bound can also be used as a surrogate loss function for maximum likelihood estimation (MLE), which optimises $\mathcal{L}_{\text{VI}}(\boldsymbol{\theta}, q; \boldsymbol{x})$ w.r.t. $\boldsymbol{\theta}$ given the training data $\boldsymbol{x}$.

The first batch of VI publications utilised $q$ distributions of simple forms, e.g. factorised Gaussians and exponential families, in order to allow a closed-form calculation for inference. Recently Monte Carlo

Submitted to the Approximate Inference workshop, NIPS 2016, Barcelona, Spain.

(MC) approximation [7, 8] has been introduced to the field, allowing a wider class of distributions to be deployed. Critically, the introduction of MC approaches blurs the boundary of sampling-based and optimisation-based methods, since then inference is also computed with the samples from the approximate posterior. These methods are often referred to as *black-box variational inference* [8] because they can be conveniently applied to generic posteriors $p(\boldsymbol{z}|\boldsymbol{x})$ without significant case-by-case consideration. However, except a very recent work [9], most of these algorithms still require tractability and fast evaluation of $\log q(\boldsymbol{z}|\boldsymbol{x})$, that is, the variational approximation is still "white-box" in terms of $q$. This requirement has become the major constraint for designing flexible approximations to exact inference. In fact the state of the art methods, e.g. [10, 11] to name a few, are hand-crafted by domain experts, with carefully designed $q$ distributions and/or auxiliary distributions/objective functions. New approaches that do not require the evaluation of $q$ will significantly simplify the development of variational approximation methods, allowing the practitioners to focus on model design for their specific tasks.

In this paper we formalise the research problem of using optimisation-based method to construct "truly black-box" approximations, or in the language of this paper *wild variational approximations* to distinguish from [8]. We provide several examples of such distributions, and discuss three classes of methods towards fitting them to the exact posterior. A straight-forward idea is *energy approximation*, i.e. approximating the terms involving $q$ in the optimisation objective. Another interesting proposal is *direct gradient approximation*, and potential methods in this line includes model-based approximation [12, 13]. The third proposal considers other optimisation objectives that do not require the tractability of $q$. Finally amortisation of stochastic dynamics is briefly sketched. Hybrid methods combining these four and other directions (e.g. gradient-free optimisation) are left to future work. We present all these proposals in VI context (which is then named *wild variational inference*), although they can be extended to other methods such as Bethe free energy [14, 15] and $\alpha$-divergence methods [16, 17].

## 2 Wild variational approximations

### 2.1 Definition and examples

A wild variational approximation to the exact posterior is defined as follows.

**Definition 1.** *A distribution $q(\boldsymbol{z}|\boldsymbol{x})$ is said to be a wild variational approximation to $p(\boldsymbol{z}|\boldsymbol{x})$ if*
*(i) it is fitted to the $p(\boldsymbol{z}|\boldsymbol{x})$ using an optimisation-based method;*
*(ii) inference with $q(\boldsymbol{z}|\boldsymbol{x})$ is comparatively easier, i.e. for the function $F(\boldsymbol{z})$ in interest, it is easier to compute (or estimate with MC methods) $\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[F(\boldsymbol{z})]$ than $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[F(\boldsymbol{z})]$;*
*(iii) the density $q(\boldsymbol{z}|\boldsymbol{x})$ is not necessary computable in a fast way.*

We provide several examples in the following, with $\boldsymbol{\phi}$ denoting all the trainable parameters for the $q$ distribution. In this paper we consider gradient-based optimisation methods for learning $\boldsymbol{\phi}$.

**Example 1.** *(Generative model)* Sampling $\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{x})$ is defined by a generative model that transforms random noise $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$ with a mapping $\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{\epsilon}, \boldsymbol{x})$. A prevalent example of such mapping is a (deep) neural network which takes $\boldsymbol{x}$ and $\boldsymbol{\epsilon}$ as input and $\boldsymbol{z}$ as the output. It has been introduced to VI as the *reparameterization trick* [18, 19, 20], hence we also use *reparameterizable* proposal for reference. The underlying distribution $q$ is often intractable, or requires further approximations to be computed efficiently.

**Example 2.** *(Truncated Markov chain)* Here the samples $\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{x})$ is defined by $T$-step transitions of a Markov chain. Examples include $T$-step Gibbs sampling process of a restricted Boltzmann machine [21], or $T$-step simulation of an SG-MCMC algorithm such as SGLD [4]. In the latter case the trainable parameters are the step-size and/or the preconditioning matrix and so on. Related work includes [22] which proposed an auxiliary lower-bound as the optimisation objective for this type of variational approximations.

**Example 3.** *(Stochastic gradient descent (SGD) with constant/adaptive learning rates)* It has been shown in [23, 24] that the trajectory of SGD with constant/adaptive learning rates near a local optimum can be viewed as a variational approximation to the exact posterior. It has similar trainable parameters $\boldsymbol{\phi}$ as for truncated SG-MCMC algorithms. These methods are more expensive when producing samples since they require evaluations of $\nabla_{\boldsymbol{z}} \log p(\boldsymbol{x}, \boldsymbol{z})$, but still they can be much cheaper than sampling from the exact posterior.

**Example 4.** *(Stochastic regularisation techniques (SRT))* SRT methods for deep neural network training, e.g. dropout [25] and related variants [26, 27], have been shown as a variational inference method for network weights $\boldsymbol{z} = \{\boldsymbol{W}\}$ [28, 29]. In this case we consider $\boldsymbol{\phi} = \{\boldsymbol{M}\}$ (the weight matrix used without SRT) as the variational parameters, and the network output is computed as $\boldsymbol{h} = \sigma((\boldsymbol{\epsilon} \odot \boldsymbol{x})\boldsymbol{M})$ with $\sigma(\cdot)$ the activation function and $\boldsymbol{\epsilon}$ some random noise. This is equivalent to set $\boldsymbol{W} = \text{diag}(\boldsymbol{\epsilon})\boldsymbol{M}$, making SRT a type of reparameterizable approximations as in example 1. Though in this case $q(\boldsymbol{z}|\boldsymbol{x})$ is tractable, practical evaluation during training can be slow since different noise values $\boldsymbol{\epsilon}$ are sampled for different datapoints in a mini-batch.

The gradient of the trainable parameters $\nabla_{\boldsymbol{\phi}}\mathcal{L}_{\text{VI}}$ contains two parts: terms that involve evaluating $\log q$ and/or $\nabla_{\boldsymbol{\phi}} \log q$, and other terms involving $\log p(\boldsymbol{z}, \boldsymbol{x})$. As an example we consider reparameterizable $q$ distributions, where we might still keep the term $\nabla_{\boldsymbol{\phi}} \log p(\boldsymbol{z}, \boldsymbol{x})$ (with $\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{\epsilon}, \boldsymbol{x})$), and approximate $\nabla_{\boldsymbol{\phi}} \log q(\boldsymbol{z}|\boldsymbol{x})$. In the following we discuss three potential solutions to this request.

### 2.2 Approximations for Optimisation

**Energy Approximation.** Assume $q$ is reparameterizable, then by the chain rule the gradient $\nabla_{\boldsymbol{\phi}}\mathcal{L}$ is computed as $\nabla_{\boldsymbol{f}}\mathcal{L}\nabla_{\boldsymbol{\phi}}\boldsymbol{f}$. This means if we have an approximation $\hat{\mathcal{L}}$ to the objective function, then the gradient can be approximated by $\nabla_{\boldsymbol{\phi}}\mathcal{L} \approx \nabla_{\boldsymbol{f}}\hat{\mathcal{L}}\nabla_{\boldsymbol{\phi}}\boldsymbol{f}$. We name this approach as *energy approximation*, since often the optimising objective can be interpreted as an energy function. For non-reparameterizable distributions $\nabla_{\boldsymbol{\phi}}\boldsymbol{f}$ can be computed with further approximations such as the generalised reparameterization trick [30].

A straight-forward method in this class considers density estimation based on the samples generated by $q$. In this case a density estimator $\hat{q}$, e.g. kernel density estimators (KDE) [31] or those parametrised by a neural network [32], is fitted to the samples $\{\boldsymbol{z}^k = \boldsymbol{f}(\boldsymbol{\epsilon}^k, \boldsymbol{x})\} \sim q$, and the gradient of $\log q$ is approximated as $\nabla_{\boldsymbol{\phi}} \log q(\boldsymbol{z}|\boldsymbol{x}) \approx \nabla_{\boldsymbol{z}} \log \hat{q}(\boldsymbol{z}|\boldsymbol{x})\nabla_{\boldsymbol{\phi}}\boldsymbol{f}$. One might even want to directly estimate $\log q$ if it turns out to be more accurate. However practitioners should be careful for implementations with automatic differentiation tools, since the parameters of the density estimator $\hat{q}$ should not be differentiated through (even though they depend on the samples $\boldsymbol{z}^k$).

The next proposal directly approximate (part of) the energy function, e.g. the entropy term $\mathbb{H}[q]$ or the KL-divergence $\text{KL}[q||p_0]$ in $\mathcal{L}_{\text{VI}}$, and let the MC approach handle the rest. In MC-dropout [28] $\text{KL}[q||p_0]$ is approximated by the Frobenius norm of the network weights $\frac{pl^2}{2}||\boldsymbol{M}||_F^2$ with dropout rate $p$ and length-scale $l$ defined by the prior. Similar approximations are in development process for $\alpha$-divergence approaches, aiming at extending SRTs to those variational methods.

A new direction for energy approximation applies density ratio estimation methods [33, 34, 35]. This is done by introducing an auxiliary distribution $\tilde{q}$ and rewrite the variational lower-bound:

$$\mathcal{L}_{\text{VI}}(\boldsymbol{\theta}, q; \boldsymbol{x}) = \mathbb{E}_q \left[ \log \frac{p_0(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\theta})}{\tilde{q}(\boldsymbol{z}|\boldsymbol{x})} + \log \frac{\tilde{q}(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{x})} \right]. \tag{2}$$

The auxiliary distribution $\tilde{q}$ is required to have tractable density and is easy to sample from. Then one can use sample-based density ratio estimation methods to fit a model $\tilde{R}$ for the ratio between $\tilde{q}$ and $q$. The gradient approximation for general $\tilde{q}$ distributions can be derived similarly as

$$\nabla_{\boldsymbol{\phi}}\mathcal{L}_{\text{VI}} = \mathbb{E}_q \left[ \nabla_{\boldsymbol{\phi}} \log \frac{p_0(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\theta})}{\tilde{q}(\boldsymbol{z}|\boldsymbol{x})} + \nabla_{\boldsymbol{z}}\tilde{R}(\boldsymbol{z})\nabla_{\boldsymbol{\phi}}\boldsymbol{f} \right]. \tag{3}$$

A simple example considers $\tilde{q} = p_0$ and the classification approach for ratio estimation. This means we train a classifier $D(\boldsymbol{z} \text{ sampled from } p_0|\boldsymbol{z}) = (1 + \exp[-\tilde{R}(\boldsymbol{z})])^{-1}$ to distinguish samples from $p_0$ and $q$, A related approach is the adversarial auto-encoder [36] which uses the prior distribution as an auxiliary. However the objective function proposed by [36] replaces the $\text{KL}[q||p_0]$ in the variational lower-bound with Jensen-Shannon divergence. Also the presented method can be extended to a sequence of auxiliary distributions (in similar spirit as the annealed importance sampling [37]), which can also be adapted slowly during training in order to obtain a better approximation.

**Direct Gradient Approximation.** The gradient of an accurate energy approximation might not necessarily be a good estimator for the exact gradient $\nabla_{\boldsymbol{\phi}}\mathcal{L}$. Therefore *direct gradient approximation* to the exact gradient might be preferred, if one cares less about the accuracy of the approximate variational lower-bound. There exists a rich literature on (non-parametric) derivative estimation

[38, 39, 40, 41, 42]; however, many of them require at least a noisy version of $\log q$ at the sampled locations, which is intractable in our case. Instead [12] applied a kernel estimator directly to the first and higher order derivatives, and [13] improved upon this idea by performing Kernel Ridge regression directly on the derivatives. The usage of integration by parts avoided evaluations on the actual gradients in [13] to, making this algorithm applicable in our context.

**New Optimisation Objectives.** In variational inference the KL-divergence $\mathrm{KL}[q||p]$ is minimised to obtain the approximate posterior. In general the KL-divergence minimisation can be replaced by other optimisation-based approximation methods, as long as with the guarantee of recovering the exact posterior if $\mathcal{Q}$ contains it. However simply replacing the objective with say other $f$-divergences will not make the optimisation easier as $q$ has an intractable density. Neither the variational techniques for estimating $f$-divergence [43, 44] as the exact posterior is difficult to sample from.

One promising direction is to replace KL divergence with the Stein discrepancy which has a special form that does not require evaluating $q(x)$ for minimisation. Briefly speaking, Stein discrepancy involves a linear functional operator $\mathcal{T}$, called Stein operator, on a set of test functions $\mathcal{G} = \{g(\boldsymbol{z})\}$ such that $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[(\mathcal{T}g)(\boldsymbol{z})] = 0$ for $\forall g \in \mathcal{G}$. Then the associated *Stein discrepancy* is defined as $\mathcal{S}(q \mid\mid p) = \sup_{g\in\mathcal{G}} \mathbb{E}_q[(\mathcal{T}g)(\boldsymbol{z})]$. For continuous density functions, a generic Stein operator is $\mathcal{T}g = \langle \nabla_{\boldsymbol{z}} \log p(\boldsymbol{z}|\boldsymbol{x}), g(\boldsymbol{z})\rangle + \nabla_{\boldsymbol{z}} \cdot g(\boldsymbol{z})$, for which $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[(\mathcal{T}g)(\boldsymbol{z})] = 0$, called Stein's identity, can be easily verified using integration by parts. Very recently [9] defined $\mathcal{G}$ as parametric functions represented by neural networks, and approximate the minimax optimisation with gradient descent (similar to GAN [45]). In contrast analytic solution for the maximiser $g^*$ exists if $\mathcal{G}$ is defined as the unit ball of a RKHS, in which case we can find the optimal $q$ by standard stochastic optimisation for minimising $\mathcal{S}(q \mid\mid p)$.

**Amortising Stochastic Dynamics.** MCMC and particle-based approximate inference methods [46, 47], though very accurate, become inefficient when inference from multiple different distributions is repeatedly required. As an example consider learning a (deep) generative model, where fast (approximate) marginalisation of latent variables is desirable. Instead we consider amortized inference here, which learns an inference network to mimic a selected stochastic dynamics. More precisely, we sample $\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{x})$, simulate $T$-step stochastic dynamics to obtain the updated particle $\boldsymbol{z}' = \boldsymbol{z} + \epsilon\Delta\boldsymbol{z}$, and update the $q$ distribution by minimising the $l_2$-distance between $\boldsymbol{z}$ and $\boldsymbol{z}'$, i.e. $\mathbb{E}_q[||\boldsymbol{z} - \boldsymbol{z}'||_2^2]$. When the step-size $\epsilon$ is small, the update can be approximated by one-step gradient descent of the $l_2$-norm, resulting in $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \epsilon\mathbb{E}_q[\nabla_{\boldsymbol{\phi}}\boldsymbol{z}\Delta\boldsymbol{z}]$. Very recently [48] applied the amortized SVGD idea to learning energy-based models. Future work in this direction will consider SG-MCMC as the stochastic transition model, and alternative measure to $l_2$-norm (e.g. maximum mean discrepancy) will also be tested.

## 3 Discussion

In this short paper we presented the research problem of constructing wild variational approximations to the exact posterior. The development of wild variational approximation methods aims at simplifying the design and application of approximate inference methods, allowing practitioners to focus more on selecting an appropriate approximate distribution that suits the best with their needs. But still our approach encourages the control of inference procedure (through the design of approximate posterior and optimisation procedure), unlike previous research of automated methods that implemented the inference engine transparently to the users. Future studies will develop better methods, more applications and theoretical understandings for wild variational approximations, and we hope our efforts can potentially motivate new ideas in the field.

**Acknowledgements**: Yingzhen Li thanks the Schlumberger Foundation FFTF fellowship.

### References

[1] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid monte carlo," *Physics Letters B*, vol. 195, no. 2, pp. 216 – 222, 1987.

[2] I. Murray, *Advances in Markov chain Monte Carlo methods*. PhD thesis, Gatsby computational neuroscience unit, University College London, 2007.

[3] R. M. Neal *et al.*, "Mcmc using hamiltonian dynamics," *Handbook of Markov Chain Monte Carlo*, vol. 2, pp. 113–162, 2011.

[4] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 681–688, 2011.

[5] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[6] M. J. Beal, *Variational algorithms for approximate Bayesian inference*. PhD thesis, University College London, 2003.

[7] J. Paisley, D. Blei, and M. Jordan, "Variational Bayesian inference with stochastic search," in *ICML*, 2012.

[8] R. Ranganath, S. Gerrish, and D. M. Blei, "Black box variational inference," in *AISTATS*, 2014.

[9] R. Ranganath, J. Altosaar, D. Tran, and D. M. Blei, "Operator variational inference," in *NIPS*, 2016.

[10] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *ICML*, 2015.

[11] R. Ranganath, D. Tran, and D. M. Blei, "Hierarchical variational models," in *ICML*, 2016.

[12] R. S. Singh, "Improvement on some known nonparametric uniformly consistent estimators of derivatives of a density," *The Annals of Statistics*, pp. 394–399, 1977.

[13] H. Sasaki, Y.-K. Noh, and M. Sugiyama, "Direct density-derivative estimation and its application in kl-divergence approximation.," in *AISTATS*, 2015.

[14] H. A. Bethe, "Statistical theory of superlattices," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 150, no. 871, pp. 552–575, 1935.

[15] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Bethe free energy, kikuchi approximations, and belief propagation algorithms," in *Neural Information Processing Systems*, 2001.

[16] J. M. Hernández-Lobato, Y. Li, M. Rowland, D. Hernández-Lobato, T. Bui, and R. E. Turner, "Black-box $\alpha$-divergence minimization," in *ICML*, 2016.

[17] Y. Li and R. E. Turner, "Rényi divergence variational inference," in *NIPS*, 2016.

[18] T. Salimans, D. A. Knowles, *et al.*, "Fixed-form variational posterior approximation through stochastic linear regression," *Bayesian Analysis*, vol. 8, no. 4, pp. 837–882, 2013.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[20] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *ICML*, 2014.

[21] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[22] T. Salimans, D. P. Kingma, and M. Welling, "Markov chain monte carlo and variational inference: Bridging the gap," in *ICML*, 2015.

[23] D. Maclaurin, D. Duvenaud, and R. P. Adams, "Early stopping is nonparametric variational inference," in *AISTATS*, 2016.

[24] S. Mandt, M. D. Hoffman, and D. M. Blei, "A variational analysis of stochastic gradient algorithms," in *ICML*, 2016.

[25] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[26] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus, "Regularization of neural networks using dropconnect," in *ICML*, 2013.

[27] S. Singh, D. Hoiem, and D. Forsyth, "Swapout: Learning an ensemble of deep architectures," in *NIPS*, 2016.

[28] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, 2016.

[29] Y. Gal, *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

[30] F. J. Ruiz, M. K. Titsias, and D. M. Blei, "The generalized reparameterization gradient," in *NIPS*, 2016.

[31] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Transactions on information theory*, vol. 21, no. 1, pp. 32–40, 1975.

[32] H. Larochelle and I. Murray, "The neural autoregressive distribution estimator.," in *AISTATS*, 2011.

[33] J. Qin, "Inferences for case-control and semiparametric two-sample density ratio models," *Biometrika*, vol. 85, no. 3, pp. 619–630, 1998.

[34] M. Sugiyama, T. Kanamori, T. Suzuki, S. Hido, J. Sese, I. Takeuchi, and L. Wang, "A density-ratio framework for statistical data processing," *Information and Media Technologies*, vol. 4, no. 4, pp. 962–987, 2009.

[35] M. Sugiyama, T. Suzuki, and T. Kanamori, "Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation," *Annals of the Institute of Statistical Mathematics*, vol. 64, no. 5, pp. 1009–1044, 2012.

[36] A. Makhzani, J. Shlens, N. Jaitly, and I. Goodfellow, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[37] R. M. Neal, "Annealed importance sampling," *Statistics and Computing*, vol. 11, no. 2, pp. 125–139, 2001.

[38] C. J. Stone, "Additive regression and other nonparametric models," *The annals of Statistics*, pp. 689–705, 1985.

[39] S. Zhou and D. A. Wolfe, "On derivative estimation in spline regression," *Statistica Sinica*, pp. 93–108, 2000.

[40] D. Ruppert and M. P. Wand, "Multivariate locally weighted least squares regression," *The annals of statistics*, pp. 1346–1370, 1994.

[41] J. Fan and I. Gijbels, *Local polynomial modelling and its applications*. Chapman & Hall, 1996.

[42] K. De Brabanter, J. De Brabanter, B. De Moor, and I. Gijbels, "Derivative estimation with local polynomial fitting," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 281–301, 2013.

[43] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization.," in *NIPS*, 2007.

[44] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.

[45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[46] B. Dai, N. He, H. Dai, and L. Song, "Provable bayesian inference via particle mirror descent," *arXiv preprint arXiv:1506.03101*, 2015.

[47] Q. Liu and D. Wang, "Stein variational gradient descent: A general purpose bayesian inference algorithm," in *Advances In Neural Information Processing Systems*, pp. 2370–2378, 2016.

[48] D. Wang and Q. Liu, "Learning to draw samples: With application to amortized mle for generative adversarial learning," *arXiv preprint arXiv:1611.01722*, 2016.