Advances in Variational Inference

Cheng Zhang *Member, IEEE,* Judith Bütepage *Member, IEEE,*
Hedvig Kjellström *Member, IEEE,* Stephan Mandt *Member, IEEE,*

Abstract—Many modern unsupervised or semi-supervised machine learning algorithms rely on Bayesian probabilistic models. These models are usually intractable and thus require approximate inference. Variational inference (VI) lets us approximate a high-dimensional Bayesian posterior with a simpler variational distribution by solving an optimization problem. This approach has been successfully used in various models and large-scale applications. In this review, we give an overview of recent trends in variational inference. We first introduce standard mean field variational inference, then review recent advances focusing on the following aspects: (a) *scalable* VI, which includes stochastic approximations, (b) *generic* VI, which extends the applicability of VI to a large class of otherwise intractable models, such as non-conjugate models, (c) *accurate* VI, which includes variational models beyond the mean field approximation or with atypical divergences, and (d) *amortized* VI, which implements the inference over local latent variables with inference networks. Finally, we provide a summary of promising future research directions.

Index Terms—Variational Inference, Approximate Bayesian Inference, Reparameterization Gradients, Structured Variational Approximations, Scalable Inference, Inference Networks.

1 Introduction

Bayesian inference has become a crucial component of machine learning. It allows us to systematically reason about parameter uncertainty. The central object of interest in Bayesian inference is the posterior distribution of model parameters given observations. Most modern applications require complex models, and the corresponding posterior is beyond reach. Practitioners, therefore, resort to approximations. This review focuses on *variational inference* (VI): a collection of approximation tools that make Bayesian inference computationally efficient and scalable to large data sets.

Many Bayesian machine learning methods rely on probabilistic latent variable models. These include Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Latent Dirichlet Allocation (LDA), Stochastic Block Models, and Bayesian deep learning architectures. Exact inference is typically intractable in these models; consequently approximate inference methods are needed. In VI, one approximates the model posterior by a simpler distribution. To this end, one minimizes the Kullback-Leibler divergence between the posterior and the approximating distribution. This approach circumvents computing intractable normalization constants. It only requires knowledge of the joint distribution of the observations and the latent variables. This methodology along with its recent refinements will be reviewed in this paper.

Within the field of approximate Bayesian inference, VI falls into the class of optimization-based approaches [13], [54]. This class also contains methods such as loopy belief propagation [116] and expectation propagation (EP) [113]. On the contrary, Markov Chain Monte Carlo (MCMC) approaches, rely on sampling [19], [53], [134]. Monte Carlo methods are often asymptotically unbiased, but can be slow to converge. Optimization-based methods, on the other hand, are often faster but may suffer from oversimplified posterior approximations [13], [181]. In recent years, there has been considerable progress in both fields [7], [14], and in particular on bridging the gap between these methods [1], [83], [101], [135], [150]. In fact, recent progress in scalable VI partly relies on fusing optimization-based and sampling-based methods. While this review focuses on VI, readers interested in EP and MCMC are referred to, e.g., [155] and [7].

The origins of VI date back to the 1980s. Mean field methods, for instance, have their origins in statistical physics, where they played a prominent role in the statistical mechanics of spin glasses, see [107] and [130]. Early applications of variational methods also include the study of neural networks, see [127] and [132]. The latter work inspired the computer science

- *Cheng Zhang and Stephan Mandt are with Disney Research, Pittsburgh.*
  E-mail: {cheng.zhang, stephan.mandt}@disneyresearch.com
- *Judith Bütepage is with KTH Royal Institute of Technology. Her contribution in this work is mainly done during her internship in Disney Research, Pittsburgh.*
  E-mail: butepage@kth.se
- *Hedvig Kjellström is with KTH Royal Institute of Technology.*
  E-mail: hedvig@kth.se

community of the 1990s to adopt variational methods in the context of probabilistic graphical models [65], [153], see also [71], [126] for early introductions on this topic.

In recent years, several factors have driven a renewed interest in variational methods. The modern versions of VI differ significantly from earlier formulations. Firstly, the availability of large datasets triggered the interest in *scalable* approaches, e.g., based on stochastic gradient descent [17], [59]. Secondly, classical VI is limited to conditionally conjugate exponential family models, a restricted class of models described in [59], [181]. In contrast, black box VI algorithms [66], [71], [135] and probabilistic programs facilitate *generic* VI, making it applicable to a range of complicated models. Thirdly, this generalization has spurred research on more *accurate* variational approximations, such as alternative divergence measures [94], [114], [194] and structured variational families [137]. Finally, *amortized* inference employs complex functions such as neural networks to predict variational distributions conditioned on data points, rendering VI an important component of modern Bayesian deep learning architectures such as variational autoencoders. In this work, we discuss important papers concerned with each of these four aspects.

While several excellent reviews of VI exist, we believe that our focus on recent developments in *scalable, generic, accurate* and *amortized* VI goes beyond those efforts. Both [71] and [126] date back to the early 2000s and do not cover the developments of recent years. Similarly, [181] is an excellent resource, especially regarding structured approximations and the information geometrical aspects of VI. However, it was published prior to the widespread use of stochastic methods in VI. Among recent introductions, [14] contains many examples, empirical comparisons, and explicit model calculations but focuses less on black box and amortized inference; [7] focuses mainly on scalable MCMC. Our review concentrates on the advances of the last 10 years prior to the publication of this paper. Complementing previous reviews, we skip example calculations to focus on a more exhaustive survey of the recent literature.

Here, we present developments in VI in a self-contained manner. We begin by covering the basics of VI in Section 2. In the following sections, we concentrate on recent advances. We identify four main research directions: scalable VI (Section 3), non-conjugate VI (Section 4), non-standard VI (Section 5), and amortized VI (Section 6). We finalize the review with a discussion (Section 7) and concluding remarks (Section 8).

2 Variational Inference

We begin this review with a brief tutorial on variational inference, presenting the mathematical foundations of this procedure and explaining the basic mean-field approximation.

Notation: The generative process is specified by observations $\boldsymbol{x}$, as well as latent variables $\boldsymbol{z}$ and a joint distribution $p(\boldsymbol{x},\boldsymbol{z})$. We use bold font to explicitly indicate sets of variables, i.e. $\boldsymbol{z}=\{z_1,z_2,\cdots,z_N\}$, where N is the total number of latent variables and $\boldsymbol{x}=\{x_1,x_2,\cdots,x_M\}$, where M is the total number of observations in the dataset. The variational distribution $q(\boldsymbol{z};\boldsymbol{\lambda})$ is defined over the latent variables $\boldsymbol{z}$ and has variational parameters $\boldsymbol{\lambda}=\{\lambda_1,\lambda_2,\cdots\lambda_N\}$.

2.1 Inference as Optimization

The central object of interest in Bayesian statistics is the posterior distribution of latent variables given observations:

$$p(\boldsymbol{z}|\boldsymbol{x})=\frac{p(\boldsymbol{x},\boldsymbol{z})}{\int p(\boldsymbol{x},\boldsymbol{z})d\boldsymbol{z}}. \tag{1}$$

For most models, computing the normalization term is impossible.

In order to approximate the posterior distribution $p(\boldsymbol{z}|\boldsymbol{x})$, VI introduces a simpler distribution $q(\boldsymbol{z};\boldsymbol{\lambda})$, parameterized by variational parameters $\boldsymbol{\lambda}$, and a distance measure between the the proxy distribution and the true posterior is considered. We then minimize this distance with respect to the parameters $\boldsymbol{\lambda}$. Finally, the optimized variational distribution is taken as a proxy for the posterior. In this way, VI turns Bayesian inference into an optimization problem.

Distance measures between two distributions $p(y)$ and $q(y)$ are also called *divergences*. As we show below, the divergence between the variational distribution and the posterior cannot be minimized directly in VI, because this involves knowledge of the posterior normalization constant. Instead, a related quantity can be maximized, namely a lower bound on the log marginal probability $p(\boldsymbol{x})$ (this will be explained below).

While various divergence measures exist [4], [6], [114], [161], the most commonly used divergence is the Kullback-Leibler (KL) divergence [13], [87], which is also referred to as relative entropy or information gain:

$$\mathrm{D}_{\mathrm{KL}}(q(y)||p(y))=-\int q(y)\log\frac{p(y)}{q(y)}dy. \tag{2}$$

As seen in Eq. 2, the KL divergence is asymmetric; $\mathrm{D}_{\mathrm{KL}}(q(y)||p(y))\neq\mathrm{D}_{\mathrm{KL}}(p(y)||q(y))$. Depending on the ordering, we obtain two different approximate inference methods. As we show below, VI employs $\mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x}))=-\mathbb{E}_{q(\boldsymbol{z})}\left[\log\frac{p(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z})}\right]$. On the other hand, expectation propagation (EP) [113] optimizes

$D_{KL}(p(\boldsymbol{z}|\boldsymbol{x})||q(\boldsymbol{z}))$ for local moment matching, which is not reviewed in this paper[1]. In Section 5 we discuss alternative divergence measures that can improve the performance of VI.

2.2 Variational Objective

VI aims at determining a variational distribution $q(\boldsymbol{z})$ that is as close as possible to the posterior $p(\boldsymbol{z}|\boldsymbol{x})$, measured in terms of the KL divergence. As for all divergences, $D_{KL}(q||p)$ is only zero if $q = p$. Thus, in an ideal case, the variational distribution takes the form $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x})$. In practice, this is rarely possible; the variational distribution is usually under-parameterized and thus not sufficiently flexible to capture the full complexity of the true posterior.

As discussed below, minimizing the KL divergence is equivalent to maximizing the *evidence lower bound (ELBO)* $\mathscr{L}$. The ELBO is a lower bound on the log marginal probability, $\log p(\boldsymbol{x})$. Since the ELBO is a conservative estimate of this marginal, which can be used for model selection, the ELBO is sometimes taken as an estimate of how well the model fits the data.

The ELBO $\mathscr{L}$ can be derived from $\log p(\boldsymbol{x})$ using Jensen's inequality as follows:

$$\begin{aligned}\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x},\boldsymbol{z}) d\boldsymbol{z} = \log \int \frac{p(\boldsymbol{x},\boldsymbol{z}) q(\boldsymbol{z};\boldsymbol{\lambda})}{q(\boldsymbol{z};\boldsymbol{\lambda})} d\boldsymbol{z} \\ &= \log \mathbb{E}_{q(\boldsymbol{z};\boldsymbol{\lambda})}\left[\frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z};\boldsymbol{\lambda})}\right] \\ &\geq \mathbb{E}_{q(\boldsymbol{z};\boldsymbol{\lambda})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z};\boldsymbol{\lambda})}\right] \equiv \mathscr{L}(\boldsymbol{\lambda}).\end{aligned} \tag{3}$$

It can be shown (see Appendix A.1) that the difference between the true log marginal probability of the data and the ELBO is the KL divergence between the variational distribution and the posterior distribution:

$$\log p(\boldsymbol{x}) = \mathscr{L}(\boldsymbol{\lambda}) + D_{KL}(q||p) \tag{4}$$

Thus, maximizing the ELBO is equivalent to minimizing the KL divergence between q and p. In traditional VI, computing the ELBO amounts to analytically solving the expectations over q, where the model is commonly restricted to the so-called conditionally conjugate exponential family (see Appendix A.2 and [181]). As part of this article, we will present more modern approaches where this is no longer the case. For an exemplary derivation of VI updates for a Gaussian Mixture Model, see [14].

The KL divergence between q and p tends to force q to be close to zero wherever p is zero (zero forcing) [114]. This property leads to automatic symmetry breaking, however, it results in underestimating the variance. In detail, when the posterior has several equivalent modes caused by model symmetry, KL based VI may focus on a single mode due to its zero forcing property. For the same reason, since the variational distribution is commonly not flexible enough, the zero forcing property leads to underestimates of the variance. In contrast, MCMC or EP may show an undesirable mode averaging behavior, while estimating the variance more closely.

It is important to choose $q(\boldsymbol{z})$ to be simple enough to be tractable, but flexible enough to provide a good approximation of the posterior [13]. A common choice is a fully factorized distribution, a mean field distribution, which is introduced in Section 2.3. A mean field distribution assumes that all latent variables are independent, which simplifies derivations. However, this independence assumption also leads less accurate results especially when the true posterior variables are highly dependent. This is e.g. the case in time series or complex hierarchical models. Richer families of distributions may lead to better approximations, but may complicate the mathematics and computation. This motivates extensions such as structured VI which is reviewed in Section 5.

2.3 Mean Field Variational Inference

Mean Field Variational Inference (MFVI) has its origins in the mean field theory of physics [126]. MFVI assumes that the variational distribution factorizes over the latent variables:

$$q(\boldsymbol{z};\boldsymbol{\lambda}) = \prod_{i=1}^{N} q(z_i;\lambda_i). \tag{5}$$

This independence assumption leads to an approximate posterior that is less expressive than when preserving dependencies, but simplifies algorithms and mathematical derivations. For notational simplicity, we omit the variational parameters $\boldsymbol{\lambda}$ for the remainder of this section. We now review how to maximize the ELBO $\mathscr{L}$, defined in Eq. 3, under a mean field assumption.

A fully factorized variational distribution allows one to optimize $\mathscr{L}$ via simple iterative updates. To see this, we focus on updating the variational parameter λ_j associated with latent variable z_j. Inserting the mean field distribution into Eq. 3 allows us to express the ELBO as follows:

$$\begin{aligned}\mathscr{L} = &\int q(z_j)\, \mathbb{E}_{q(\boldsymbol{z}_{\neg j})}\left[\log p(z_j, \boldsymbol{x}|\boldsymbol{z}_{\neg j})\right] dz_j \\ &- \int q(z_j) \log q(z_j) dz_j + c_j.\end{aligned} \tag{6}$$

Above, $\boldsymbol{z}_{\neg j}$ indicates the set $\boldsymbol{z}$ excluding z_j. The constant c_j contains all terms that are constant with respect to z_j, such as the entropy term associated with $\boldsymbol{z}_{\neg j}$. We have thus separated the full expectation into an inner expectation over $\boldsymbol{z}_{\neg j}$, and an outer expectation over z_j.

1. We refer the readers to the EP roadmap for more information about advancement of EP. https://tminka.github.io/papers/ep/roadmap.html

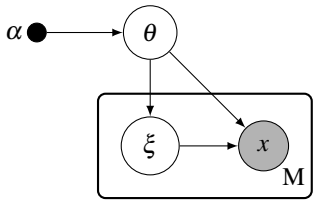


Fig. 1. A graphical model of the observations $\boldsymbol{x}$ that depend on underlying local hidden factors $\boldsymbol{\xi}$ and global parameters θ. We use $\boldsymbol{z} = \{\theta, \boldsymbol{\xi}\}$ to represent all latent variables. M is the number of the data points. N is the number of the latent variables.

Inspecting Eq. 6, it becomes apparent that this formula is a negative KL divergence, which is maximized for variable j by:

$$\log\ q^*(z_j) = \mathbb{E}_{q(\boldsymbol{z}_{\neg j})}\left[\log p(z_j|\boldsymbol{z}_{\neg j}, \boldsymbol{x})\right] + const. \quad (7)$$

Exponentiating and normalizing this result yields:

$$\begin{aligned} q^*(z_j) &\propto \exp(\mathbb{E}_{q(\boldsymbol{z}_{\neg j})}\left[\log p(z_j|\boldsymbol{z}_{\neg j}, \boldsymbol{x})\right]) \\ &\propto \exp(\mathbb{E}_{q(\boldsymbol{z}_{\neg j})}\left[\log p(\boldsymbol{z}, \boldsymbol{x})\right]) \end{aligned} \quad (8)$$

Using Eq. 8, the variational distribution can now be updated iteratively for each latent variable until convergence. Similar updates also form the basis for the variational message passing algorithm [190] (See Appendix A.3).

For more details on the mean field approximation and its geometrical interpretation we refer the reader to [181] and [13]. Having covered the basics of VI, we devote the rest of this paper to reviewing advanced techniques.

3 Scalable Variational Inference

In this section, we survey scalable VI. Big datasets raise new challenges for the computational feasibility of Bayesian algorithms, making scalable inference techniques essential. We begin by reviewing stochastic variational inference (SVI) in Section 3.1, which uses stochastic gradient descent to scale VI to large datasets. Section 3.2 discusses practical aspects of SVI, such as adaptive learning rates or variance reduction. Further approaches to improve on the scalability of VI are discussed in Section 3.3; these include sparse inference, collapsed inference and distributed inference.

Notation: This section follows the general model structure of global and local hidden variables, assumed in [59]. Fig. 1 depicts the corresponding graphical model where the latent variable $\boldsymbol{z} = \{\theta, \boldsymbol{\xi}\}$ includes local (per data point) variables $\boldsymbol{\xi}$ and global variable θ. Similarly, the variational parameters are given by $\boldsymbol{\lambda} = \{\gamma, \boldsymbol{\phi}\}$, where the variational parameter γ corresponds to the global latent variable, and $\boldsymbol{\phi}$ denotes the set of local variational parameters. Furthermore, the model depends on hyperparameters α. The mini-batch size is denoted by S.

3.1 Stochastic Variational Inference

VI frames Bayesian inference as an optimization problem. For many models of interest, the variational objective has a special structure, namely, it is the sum over contributions from all M individual data points. Problems of this type can be solved efficiently using stochastic optimization [17], [143]. Stochastic Variational Inference (SVI) amounts to applying stochastic optimization to the objective function encountered in VI [57], [59], [63], [185], thereby scaling VI to very large datasets. Using stochastic optimization in the context of VI was proposed in [59], [63], [152]. We follow the conventions of [59] which presents SVI for models of the conditionally conjugate exponential family class.

The ELBO of the general graphical model shown in Fig. 1 has the following form:

$$\begin{aligned} \mathscr{L} = \mathbb{E}_q[\log p(\theta|\alpha) - \log q(\theta|\gamma)] + \\ \sum_{i=1}^{M} \mathbb{E}_q\left[\log p(\xi_i|\theta) + \log p(x_i|\xi_i, \theta) - \log q(\xi_i|\phi_i)\right]. \end{aligned} \quad (9)$$

We assume that the variational distribution is given by $q(\boldsymbol{\xi}, \theta) = q(\theta|\gamma)\prod_{i=1}^{M} q(\xi_i|\phi_i)$. For conditionally conjugate exponential families [181], the expectations involved in Eq. 9 can be computed analytically, yielding a closed form objective that can be optimized with coordinate updates. In this case, every iteration or gradient step scales with M, and is therefore expensive for large data.

SVI solves this problem by randomly selecting mini-batches of size S to obtain a stochastic estimate of the ELBO $\hat{\mathscr{L}}$:

$$\begin{aligned} \hat{\mathscr{L}} = \mathbb{E}_q[\log p(\theta|\alpha) - \log q(\theta|\gamma)] + \\ \frac{M}{S}\sum_{s=1}^{S} \mathbb{E}_q\left[\log p(\xi_{i_s}|\theta) + \log p(x_{i_s}|\xi_{i_s}, \theta) - \log q(\xi_{i_s}|\phi_{i_s})\right], \end{aligned} \quad (10)$$

where i_s is the variable index from the mini-batch. This objective is optimized using stochastic gradient ascent. The mini-batch size is chosen as $S \ll M$ with $S > 1$ in order to reduce the variance of the gradient. The choice of S emerges from a trade-off between the computational overhead associated with processing a mini-batch, such as performing inference over global parameters (favoring larger mini-batches which have lower gradient noise and allow larger learning rate), and the cost of iterating over local parameters in the mini-batch (favoring small mini-batches). Additionally, this tradeoff is also affected by memory structures in modern hardware such as GPUs.

As in stochastic optimization, SVI requires the use of a learning rate ρ_t that decreases over iterations t. The

Robbins-Monro conditions $\sum_{t=1}^{\infty} \rho_t = \infty$ and $\sum_{t=1}^{\infty} \rho_t^2 < \infty$ guarantee that every point in parameter space can be reached, while the gradient noise decreases quickly enough to ensure convergence [143]. We deepen the discussion of these topics in Section 3.2.

An important result of [59] is that the SVI procedure automatically produces *natural* stochastic gradients, and that these natural gradients have a simpler form than regular gradients for models in the conditionally conjugate exponential family. Natural gradients are studied in [5] and were first introduced to VI in [61], [62]. They are pre-multiplied with the inverse Fisher information and therefore take the information geometry into account. For details, we refer interested readers to Appendix A.4 and [59].

Sometimes SVI is referred to as online VI [57], [185]. These methods are equivalent under the assumptions that the volume of the data M is known. In streaming applications, the mini-batches arrive sequentially from a data source, but the SVI updates look the same. However, when M is unknown, it is unclear how to set the scale parameter M/S in Eq. 10. To this end, [104] introduces the concept of the population posterior which depends on the unknown size of the dataset. This concept allows one to apply online VI with respect to the expected ELBO over the population.

Stochastic gradient methods have been adapted to various settings, such gamma processes [82] and variational autoencoders [79]. In recent years, most advancements in VI have been developed relying on a SVI scheme. In the following, we detail how to further adapt SVI to accelerate convergence.

3.2 Tricks of the Trade for SVI

The efficiency of stochastic gradient descent (SGD) methods, which are the basis of SVI, depend on the variance of the gradient estimates. Smaller gradient noise allows for larger learning rates and leads to faster convergence. This section covers tricks of the trade in the context of SVI, such as adaptive learning rates and variance reduction. Some of these approaches are generally applicable in SGD setups.

Adaptive Learning Rate and Mini-batch Size: The speed of convergence is influenced by the choice of the learning rate and the mini-batch size [9], [40]. Due to the law of large numbers, increasing the mini-batch size reduces the stochastic gradient noise [40], allowing larger learning rates. To accelerate the learning procedure, one can either optimally adapt the mini-batch size for a given learning rate, or optimally adjust the learning rate to a fixed mini-batch size.

We begin by discussing learning rate adaptation. In each iteration, the empirical gradient variance can guide the adaptation of the learning rate. Popular optimization methods that make use of this idea include RMSProp [170], AdaGrad [36], AdaDelta [191] and Adam [80]. These methods are not specific to SVI but are frequently used in this context; for more details we refer interested reader to [47]. An adaptive learning rate specifically designed for SVI was suggested in [138]. Recall that γ is the global variational parameter. The learning rate for γ results from minimizing the error between the stochastic mini-batch update and the full batch update, and is given by:

$$\rho_t^* = \frac{(\gamma_t^* - \gamma_t)^T (\gamma_t^* - \gamma_t)}{(\gamma_t^* - \gamma_t)^T (\gamma_t^* - \gamma_t) + tr(\Sigma)}, \tag{11}$$

where γ_t^* indicates the batch update (using all data), γ_t is the current variational parameter and Σ is the covariance matrix of the variational parameter in this mini-batch. Eq. 11 indicates that the learning rate should increase when the trace term, i.e. the mini-batch variance, is small. Further estimation of ρ_t is presented in [138]. Although developed for SVI, this method can be adapted to other stochastic optimization methods and resembles the aforementioned adaptive learning rate schemes.

Instead of adapting the learning rate, the mini-batch size can be adapted while keeping the learning rate fixed to achieve similar effects [9], [22], [31]. For example [9] optimizes the mini-batch size to decrease the SGD variance proportionally to the value of the objective function relative to the optimum. In practice, the estimated gradient noise covariance and the magnitude of the gradient are used to estimate the optimal mini-batch size.

Variance Reduction: In addition to controlling the optimization path through the learning rate and mini-batch size, we can reduce the variance, thereby enabling larger gradient steps. Variance reduction is often employed in SVI to achieve faster convergence. In the following, we summarize how to accomplish this with *control variates*, *non-uniform sampling*, and *other approaches*.

Control Variates. A control variate is a stochastic term that can be added to the stochastic gradient such that its expectation remains the same, but its variance is reduced. It is constructed as a vector that should be highly correlated with the stochastic gradient and easy to compute. Using control variates for variance reduction is common in Monte Carlo simulation and stochastic optimization [146], [184]. Several authors have suggested the use of control variates in the context of SVI [70], [129], [135], [184]. [184] provides two examples of model specific control variate construction, focusing on logistic regression and LDA.

The stochastic variance reduced gradient (SVRG) method [70] amounts to constructing a control variate which takes advantage of previous gradients from all

data points. It exploits that gradients along the optimization path are correlated. The standard stochastic gradient update $\gamma_{t+1} = \gamma_t - \rho_t(\nabla\hat{\mathscr{L}}(\gamma_t))$ is replaced by:

$$\gamma_{t+1} = \gamma_t - \rho_t(\nabla\hat{\mathscr{L}}(\gamma_t) - \nabla\hat{\mathscr{L}}(\tilde{\gamma}) + \tilde{\mu}). \tag{12}$$

$\hat{\mathscr{L}}$ indicates the estimated objective (here the negative ELBO) based on the current set of mini-batch indices, $\tilde{\gamma}$ is a snapshot of γ after every m iterations, and $\tilde{\mu}$ is the batch gradient computed over all the data points, $\tilde{\mu} = \nabla\mathscr{L}(\tilde{\gamma})$. Since SVRG requires a full pass through the dataset every mth iteration, it is not feasible for very large datasets. Yet, in contrast to traditional SGD, SVRG enables convergence with constant learning rates. We will encounter different types of control variates in the context of black box variational inference (BBVI) (see Section 4.2 for a detailed discussion).

Non-uniform Sampling. Instead of subsampling data points with equal probability, non-uniform sampling can be used to select mini-batches with a lower gradient variance. Several authors suggested variants of importance sampling in the context of mini-batch selection [27], [131], [198]. Although effective, these methods are not always practical, as the computational complexity of the sampling probability relates to the dimensionality of model parameters [41]. Alternative methods aim at de-correlating similar points and sampling diversified mini-batches. These methods include stratified sampling [197], where one samples data from pre-defined subgroups based on meta-data or labels, clustering-based sampling [41], which amounts to clustering the data using k-means and then sampling data from every cluster with adjusted probabilities, and diversified mini-batch sampling [196] using determinantal point processes (see Appendix A.5) to suppress the probability of data points with similar features in the same mini-batch. All of these methods have been shown to reduce variance and can also be used for learning on imbalanced data.

Other Methods. A number of alternative methods have been developed that contribute to variance reduction for SVI. A popular approach relies on Rao-Blackwellization, which is used in [135]. The Rao-Blackwellization theorem (see Appendix A.6) generally states that a conditional estimation has lower variance if a valid statistic to be conditioned on exists. Inspired by Rao-Blackwellization, the local expectation gradients method [173], [176] has been proposed. It explores the conditional structure of a model to reduce the variance. A related approach has been developed for SVI, which averages expected sufficient statistics over a sliding window of mini-batches to obtain a natural gradient with smaller mean squared error [100]. Further variance reduction methods [77], [135], [144] are designed for black box variational inference (BBVI). These methods make use of the reparametrization trick and are discussed together with BBVI in Section 4. They are different in nature because the sampling space is continuous, while SVI samples from a discrete population of data points.

3.3 Collapsed, Sparse, and Distributed VI

In contrast to using stochastic optimization for faster convergence, this section presents methods that leverage the structure of certain models to achieve the same goal. In particular, we focus on *collapsed*, *sparse*, *parallel*, and *distributed* inference.

Collapsed Inference: Collapsed variational inference (CVI) relies on the idea of analytically integrating out certain model parameters [56], [75], [88], [90], [162], [167], [175]. Due to the reduced number of parameters to be estimated, inference is typically faster. One can either marginalize out these latent variables before the ELBO is derived, or eliminate them afterwards [56], [75].

Several authors have proposed CVI for topic models [88], [167] where one can either collapse the topic proportions [167] or the topic assignment [56]. In addition to these model specific derivations, [56] unifies existing model-specific CVI approaches and presents a general collapsed inference method for models in the conjugate exponential family class.

The computational benefit of CVI depends strongly on the statistics of the collapsed variables. Additionally, collapsing latent random variables in a model can cause other inference techniques to become tractable. For models such as topic models, we can collapse the discrete variables and only infer the continuous ones, allowing e.g. the application of inference networks (Section 6) [108], [160].

Sparse Inference: Sparse inference aims to exploit either sparsely distributed parameters for parametric models, or datasets which can be summarized by a small number of representative data points. In these regimes, low rank approximations enable scalable inference [55], [157], [174]. Sparse inference can be either interpreted as a modeling choice or as an inference scheme [20].

Sparse inference methods are often encountered in the Gaussian Process (GPs) literature. The computational cost of learning GPs is $\mathscr{O}(n^3)$, where n is the number of data points. This cost is caused by the inversion of the kernel matrix K_{nn} of size $n \times n$, which hinders the application of GPs to big data sets. The idea of sparse inference in GPs [157] is to introduce *m inducing points*. These are auxiliary variables which, when integrated out, yield the original GP. Instead of being marginalized out however, they are treated as

latent variables whose distribution is estimated using VI. Inducing points can be interpreted as pseudo-inputs that reflect the original data, but yield a more sparse representation since $m \ll n$. With inducing points, only a $m \times m$ sized matrix needs to be inverted, and consequently the computational complexity of this method is $\mathcal{O}(nm^2)$. [174] collapses the distribution of inducing points, and [55] further extends this work to a stochastic version [59] with a computational complexity of $\mathcal{O}(m^3)$. Additionally, sparse inducing points make inference in Deep Gaussian Processes tractable [30].

Parallel and Distributed Inference: Several computational acceleration techniques, such as distributed computing, can be applied to VI [43], [120], [123], [192]. Distributed inference schemes are often required in large scale scenarios, where data and computations are distributed across several machines. Independent latent variable models are trivially parallelizable. However, model specific designs such as reparametrizations might be required to enable efficient distributed inference [43]. Current computing resources make VI applicable to web-scale data analysis [192].

4 Non-Conjugate Inference

In this section, we review techniques which make VI generic. One aspect of this research is to make VI more broadly applicable to a large class of models, in particular non-conjugate ones. A second aspect is to make VI more automatic, thus eliminating the need for model-specific calculations and approximations, which allows VI to be more accessible in general.

Variational inference was originally limited to conditionally conjugate models, for which the ELBO could be computed analytically before it was optimized [59], [193]. In this section, we introduce methods that relax this requirement and simplify inference. Central to this section are stochastic gradient estimators of the ELBO that can be computed for a broader class of models.

We start with the Laplace approximation in Section 4.1 and illustrate its limitations. We will then introduce black box variational inference (BBVI) in Section 4.2 which employs the REINFORCE or score function gradient. Section 4.3 discusses a different form of BBVI, which uses reparameterization gradients. Other approaches for non-conjugate VI are discussed in Section 4.4.

4.1 Laplace's Method and Its Limitations

The Laplace (or Gaussian) approximation, estimates the posterior by a Gaussian distribution [89]. To this end, one seeks the maximum of the posterior and computes the inverse of its Hessian. These two entities are taken as the mean and covariance of the Gaussian posterior approximation. To make this approach feasible, the log posterior needs to be twice-differentiable. According to the Bernstein von Mises theorem (a.k.a. Bayesian central limit theorem) [23], the posterior approaches a Gaussian asymptotically in the limit of large data, and the Laplace approximation becomes exact (provided that the model is under-parameterized). The approach can be applied to approximate the maximum a posteriori (MAP) mean and covariance, predictive densities, and marginal posterior densities [171]. The Laplace method has also been extended to more complex models such as belief networks with continuous variables [8].

This approximation suffers mainly from being purely local and depending only on the curvature of the posterior around the optimum; KL minimization typically approximates the posterior shape more accurately. Additionally, the Laplace approximation is limited to the Gaussian variational family and does not apply to discrete variables [183]. Computationally, the method requires the inversion of a potentially large Hessian, which can be costly in high dimensions. This makes this approach intractable in setups with a large number of parameters.

4.2 Black Box Variational Inference

In classical variational inference, the ELBO is first derived analytically, and then optimized. This procedure is usually restricted to models in the conditionally conjugate exponential family [59]. For many models, including Bayesian deep learning architectures or complex hierarchical models, the ELBO often contains intractable expectations with no known or simple analytical solution. Even if an analytic solution is available, the analytical derivation of the ELBO often requires time and mathematical expertise. In contrast, black box variational inference proposes a generic inference algorithm for which only the generative process of the data has to be specified. The main idea is to represent the gradient as an expectation and to use Monte Carlo techniques to estimate this expectation.

As discussed in Section 2, variational inference aims at maximizing the ELBO, which is equivalent to minimizing the KL divergence between the variational posterior and target distribution. To maximize the ELBO, one needs to follow the gradient or stochastic gradient of the variational parameters. The key insight of BBVI is that one can obtain an unbiased gradient estimator by sampling from the variational distribution without having to compute the ELBO analytically [129], [135].

For a generic class of models, the gradient of the ELBO can be expressed as an expectation with respect

to the variational distribution:

$$\nabla_{\boldsymbol{\lambda}} \mathscr{L} = \mathbb{E}_q[\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{z}|\boldsymbol{\lambda})(\log p(\boldsymbol{x},\boldsymbol{z}) - \log q(z|\boldsymbol{\lambda}))]. \tag{13}$$

The full gradient $\nabla_{\boldsymbol{\lambda}} \mathscr{L}$, involving the expectation over q, can now be approximated by a stochastic gradient estimator $\nabla_{\boldsymbol{\lambda}} \hat{\mathscr{L}}_s$ by sampling from the variational distribution:

$$\nabla_{\boldsymbol{\lambda}} \hat{\mathscr{L}}_s = \frac{1}{K} \sum_{k=1}^{K} \nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{z}_k|\boldsymbol{\lambda})(\log p(\boldsymbol{x},\boldsymbol{z}_k) - \log q(\boldsymbol{z}_k|\boldsymbol{\lambda})), \tag{14}$$

where $\boldsymbol{z}_k \sim q(\boldsymbol{z}|\boldsymbol{\lambda})$. Thus, BBVI provides black box gradient estimators for VI. Moreover, it only requires a practitioner to provide the joint distribution of observations and latent variables without the need to derive the gradient of the ELBO explicitly. The quantity $\nabla_\lambda \log q(\boldsymbol{z}_k|\boldsymbol{\lambda})$ is also known as the score function and is part of the REINFORCE algorithm [189].

A direct implementation of stochastic gradient ascent based on Eq. 14 suffers from high variances of the estimated gradients. Much of the success of BBVI can be attributed to variance reduction through Rao-Blackwellization and control variates [135]. As one of the most important advancements of modern approximate inference, BBVI as been extended and made amortized inference feasible, see Section 6.1.

Variance Reduction for BBVI: BBVI requires a different set of techniques than those reviewed for SVI in Section 3.2. The gradient noise in SVI resulted from subsampling a finite set, making techniques such as SVRG applicable. In contrast, the BBVI noise originates from continuous random variables. This requires a different approach.

The arguably most important control variate in BBVI is the score function control variate [135], where one subtracts a Monte Carlo expectation of the score function from the gradient estimator:

$$\nabla_{\boldsymbol{\lambda}} \hat{\mathscr{L}}_{control} = \nabla_{\boldsymbol{\lambda}} \hat{\mathscr{L}} - \frac{w}{K} \sum_{k=1}^{K} \nabla_{\boldsymbol{\lambda}} \log q(z_k|\boldsymbol{\lambda}) \tag{15}$$

As required, the score function variate has expectation zero under the variational distribution. The weight w is selected such that it minimizes the variance of the gradient.

While the original BBVI paper introduces both Rao-Blackwellization and control variates, [173] points out that good choices for control variates might be model-dependent. They further elaborate on local expectation gradients, which take only the Markov blanket of each variable into account. A different approach is presented by [148], which introduces overdispersed importance sampling. By sampling from a proposal distribution that belongs to an overdispersed exponential family and that places high mass in the tails of the variational distribution, the variance of the gradient is reduced.

4.3 Reparameterization Gradient VI

An alternative to the reinforce gradients introduced in Section 4.2 are so-called reparameterization gradients. These gradients are obtained by representing the variational distribution as a deterministic parametric transformation of a uniform noise distribution. Empirically, reparameterization gradients are often found to have a lower variance than REINFORCE gradients. Another advantage is that do not depend on the KL divergence, but apply more broadly (see also Section 5).

Reparameterization Gradients: The reparameterization trick simplifies the Monte Carlo computation of the gradient (see Eq. 14) by representing random variables as deterministic functions of noise distributions. This makes backpropagation through random variables possible. In more detail, the trick states that a random variable z with a distribution $q_{\boldsymbol{\lambda}}(z)$ can be expressed as a transformation of a random variable $\boldsymbol{\varepsilon}$ that comes from a noise distribution, such as uniform or Gaussian. For example, if $z \sim \mathscr{N}(z;\mu,\sigma^2)$, then $z = \mu + \sigma\varepsilon$ where $\varepsilon \sim \mathscr{N}(z;0,1)$, see [78], [141].

More generally, the random variable z is given by a parameterized, deterministic function of random noise, $z = g(\varepsilon,\boldsymbol{\lambda}), \varepsilon \sim p(\varepsilon)$. Importantly, the noise distribution $p(\varepsilon)$ is considered independent of the parameters of $q_{\boldsymbol{\lambda}}(z)$, and therefore $q_{\boldsymbol{\lambda}}(z)$ and $g(\varepsilon,\boldsymbol{\lambda})$ share the same parameters $\boldsymbol{\lambda}$. This allows us to compute any expectation over z as an expectation over ε. (This is also called the law of the unconscious statistician [142].)

We can now build a stochastic gradient estimator of the ELBO by pulling the gradient into the expectation, and approximating it by samples from the noise distribution:

$$\nabla_{\boldsymbol{\lambda}} \hat{\mathscr{L}}_r = \frac{1}{S} \sum_{s=1}^{S} \nabla_{\boldsymbol{\lambda}} (\log p(x_i, g(\varepsilon_s,\boldsymbol{\lambda})) - \log q(g(\varepsilon_s,\boldsymbol{\lambda})|\boldsymbol{\lambda})), \quad \varepsilon_s \sim p(\varepsilon). \tag{16}$$

Often, the entropy term can be computed analytically, which can lead to a lower gradient variance [78].

Note that the gradient of the log joint distribution enters the expectation. This is in contrast to the REINFORCE gradient, where the gradient of the variational distribution is taken (Eq. 14). The advantage of taking the gradient of the log joint is that this term is more informed about the direction of the maximum posterior mode. The lower variance of the reparameterization gradient may be attributed to this property.

While the variance of this estimator (Eq. 16) is often lower than the variance of the score function gradient (Eq. 14), a theoretical analysis shows that this

is not guaranteed, see Chapter 3 in [42]. [144] showes that the reparameterization gradient can be divided into a path derivative and the score function. Omitting the score function in the vicinity of the optimum can result in an unbiased gradient estimator with lower variance. Reparameterization gradients are also the key to variational autoencoders [78], [141] which we discuss in detail in Section 6.2.

The reparameterization trick does not trivially extend to many distributions, in particular to discrete ones. Even if a reparameterization function exists, it may not be differentiable. In order for the reparameterization trick to apply to discrete distributions, the variational distributions require further approximations. Several groups have addressed this problem. In [67], [99], the categorical distribution is approximated with the help of the Gumbel-Max trick and by replacing the argmax operation with a softmax operator. Varying a temperature parameter controls the degree to which the softmax can approximate the categorical distribution. The closer it resembles a categorical distribution, the higher the variance of the gradient. The authors propose annealing strategies to improve convergence. Similarly, a stick-breaking process is used in [119] to approximate the Beta distribution with the Kumaraswamy distribution.

As many of these approaches rely on approximations of individual distributions, there is growing interest in more general methods that are applicable without specialized approximations. The generalized reparameterization gradient [147] achieves this by finding an invertible transformation between the noise and the latent variable of interest. The authors derive the gradient of the ELBO which decomposes the expected likelihood into the standard reparameterization gradient and a correction term. The correction term is only needed when the transformation weakly depends on the variational parameters. A similar division is derived by [117] which proposes an accept-reject sampling algorithm for reparameterization gradients that allow one to sample from expressive posteriors. While reparameterization gradients often demonstrate lower variance than the score function, the use of Monte Carlo estimates still suffers from the injected noise. The variance can be further reduced with control variates [109], [144].

4.4 Other Generalizations

Finally, we survey a number of approaches that consider VI in non-conjugate models but do not follow the BBVI principle. Since the ELBO for non-conjugate models contains intractable integrals, these integrals have to be approximated somehow, either using some form of Taylor approximations (including Laplace approximations), lower-bounding the ELBO further such that the resulting integrals are tractable, or using some form of Monte Carlo estimators. Approximation methods which involve inner optimization routines [15], [182], [195] often become prohibitively slow for practical inference tasks. In contrast, approaches based on additional lower bounds with closed-form updates [74], [81], [183] can be computationally more efficient. Examples include extensions of the variational message passing algorithm [190] to non-conjugate models [81], or [183], which adapted ideas from the Laplace approximation (Section 4.1). Furthermore, [151] proposed a variational inference technique based on stochastic linear regression to estimate the parameters of a fixed variational distribution based on Monte Carlo approximations of certain sufficient statistics. Recently, [74] proposed a hybrid approach, where inference is split into a conjugate and a non-conjugate part.

5 Non-standard VI: Beyond KL Divergence and Mean Field

In this section, we present various methods that aim at improving the accuracy of standard VI. Previous sections dealt with making VI scalable and applicable to non-conjugate exponential family models. Most of the work in those areas, however, still addresses the standard setup of MFVI and employs the KL divergence as a measure of distance between distributions. Here we review recent developments that go beyond this setup, with the goal of avoiding poor local optima and increasing the accuracy of VI. Inference networks, normalizing flows, and related methods may also be considered as non-standard VI, but are discussed in Section 6.

We start by reviewing the origins of MFVI in statistical physics and describe its limitations (Section 5.1). We then discuss alternative divergence measures in Section 5.2. Structured variational approximations beyond mean field are discussed in Section 5.3, followed by alternative methods that do not fall into the previous two classes (Section 5.4).

5.1 Origins and Limitations of Mean Field VI

Variational methods have a long tradition in statistical physics. The mean field method was originally applied to spin glass models [126]. A simple example for such a spin glass model is the Ising model, a model of binary variables on a lattice with pairwise couplings. To estimate the resulting statistical distribution of spin states, a simpler, factorized distribution is used as a proxy. This is done in such a way that the marginal probabilities of the spins showing up or down are preserved. The many interactions of a given spin with its neighbors are replaced by a single interaction between a spin and the effective magnetic field (a.k.a. *mean*

field) of all other spins. This explains the name origin. Physicists typically denote the negative log posterior as an energy or Hamiltonian function. This language has been adopted by the machine learning community for approximate inference in both directed and undirected models, summarized in Appendix A.7 for the reader's reference.

Mean field methods were first adopted in neural networks by Anderson and Peterson in 1987 [132], and later gained popularity in the machine learning community [71], [126], [153]. The main limitation of mean field approximations is that they explicitly ignore correlations between different variables e.g., between the spins in the Ising model. Furthermore, [181] showed that the more possible dependencies are broken by the variational distribution, the more non-convex the optimization problem becomes. Conversely, if the variational distribution contains more structure, certain local optima do not exist. A number of initiatives to improve mean field VI have been proposed by the physics community and further developed by the machine learning community [126], [133], [169].

An early example of going beyond the mean field theory in a spin glass system is the Thouless-Anderson-Palmer (TAP) equation approach [169], which introduces perturbative corrections to the variational free energy. A related idea relies on power expansions [133], which has been extended and applied to machine learning models by various authors [72], [125], [128], [139], [164]. Additionally, information geometry provides an insight into the relation between MFVI and TAP equations [165], [166]. [194] further connects TAP equations with divergence measures. We refer the readers to [126] for more information. Next, we review the recent advances beyond MFVI based on divergence measures other than the KL divergence.

5.2 VI with Alternative Divergences

The KL divergence often provides a computationally convenient method to measure the distance between two distributions. It leads to analytically tractable expectations for certain model classes. However, traditional Kullback-Leibler variational inference (KLVI) suffers from problems such as underestimating posterior variances [114]. Furthermore, it is unable to break symmetry when multiple modes are close [124] and is a comparably loose bound [194]. Due to these shortcomings, a number of other divergence measures have been proposed which we survey here.

The idea of using alternative divergence measures for variational methods may leads to various methods such as expectation propagation (EP). Extensions of EP [92], [111], [180], [199] can be viewed as generalizing EP to other divergence measures [114]. While these methods are sophisticated, a practitioner will find them difficult to use due to complex derivations and limited scalability. Recent developments of VI focus mainly on a unified framework in a black box fashion to allow for scalability and accessibility. BBVI rendered the application of other divergence measures, such as the χ divergence [33], possible while maintaining the efficiency and simplicity of the method.

In this section, we introduce relevant divergence measures and show how to use them in the context of VI. The KL divergence, as discussed in Section 2.1, is a special form of the α-divergence, while the α-divergence is a special form of the f-divergence. All above divergences can be written in the form of the Stein discrepancy.

α-divergence: The α-divergence is a family of divergence measures with interesting properties from an information geometrical and computational perspective [4], [6]. Both the KL divergence and the Hellinger distance are special cases of α-divergence.

Different formulations of the α-divergence exist [6], [200], and various VI methods use different definitions [95], [114]. We focus on Renyi's formulation, which is defined as:

$$D_\alpha^R(p||q) = \frac{1}{\alpha-1}\log\int p(x)^\alpha q(x)^{1-\alpha}dx, \qquad (17)$$

where $\alpha \geq 0$. With this definition of α-divergences, a smaller α leads to mass-covering effects, while a larger α results in zero-forcing effects. For $\alpha = 1$ we recover standard VI (involving the KL divergence).

α-divergences have been used in variational inference [94], [95]. Similar to the bound in Eq. 4, using Renyi's definition we can derive another bound with the α-divergence:

$$\begin{aligned}\mathscr{L}_\alpha &= \log p(\boldsymbol{x}) - D_\alpha^R(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})) \\ &= \frac{1}{\alpha-1}\log\mathbb{E}_q\left[\left(\frac{p(\boldsymbol{z},\boldsymbol{x})}{q(\boldsymbol{z})}\right)^{1-\alpha}\right]. \qquad (18)\end{aligned}$$

For $\alpha \geq 0, \alpha \neq 1$, $\mathscr{L}_\alpha$ is a lower bound. Among various possible definitions of the α-divergence, only Renyi's formulation leads to a bound (Eq. 18) in which the marginal likelihood $p(x)$ cancels out.

In the context of big data, there are different ways to derive stochastic updates from the general form of the bound (Eq. 18). Different derivations can recover BBVI for α-divergences [95] and stochastic EP [92].

f-Divergence and Generalized VI: α-divergences are a subset of the more general family of f-divergences [3], [28], which take the form:

$$D_f(p||q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right)dx.$$

f is a convex function with $f(1) = 0$. For example, the KL divergence $KL(p||q)$ is represented by the f-divergence with $f(r) = r\log(r)$, and the Pearson χ^2 distance is an f-divergence with $f(r) = (r-1)^2$.

In general, it is not possible to obtain a useful variational bound for all f-divergences directly (such as in Eq. 18). The bound may non-trivially depend on the marginal likelihood, unless f is chosen in a certain way (such as for the KL and Renyi's α-divergence). In this section, we will review an alternative bound which is derived through Jensen's inequality.

There exists a family of variational bounds, which further generalize Renyi's α bound. [194] lower-bounds the marginal likelihood, using a general function $\tilde{f}$ as follows:

$$p(\boldsymbol{x}) \geq \tilde{f}(p(\boldsymbol{x})) \geq \mathbb{E}_{q(\boldsymbol{z})}\left[\tilde{f}\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right)\right] \equiv \mathscr{L}_{\tilde{f}}. \quad (19)$$

In contrast to the f-divergence, the function $\tilde{f}$ has to be concave with $\tilde{f}(x) < x$. The choice of $\tilde{f}$ allows us to construct variational bounds with different properties. For example, when $\tilde{f}$ is the identity function, the bound is tight and we recover the true marginal likelihood; when $\tilde{f}$ is the logarithm, we obtain the standard ELBO; and when $\tilde{f}(x) \propto x^{(1-\alpha)}$, the bound is equivalent to the α-divergence up to a constant. For $V \equiv \log p(\boldsymbol{x},\boldsymbol{z}) - \log q(\boldsymbol{z})$, the authors propose the following function:

$$\tilde{f}^{(V_0)}(e^{-V}) = e^{-V_0}\Big(1 + (V_0 - V) + \frac{1}{2}(V_0 - V)^2 + \frac{1}{6}(V_0 - V)^3\Big).$$

Above, V_0 is a free parameter that can be optimized, and which absorbs the bound's dependence on the marginal likelihood. The authors show that the terms up to linear order in V correspond to the KL divergence, whereas higher order polynomials are correction terms which make the bound tighter. This connects to earlier work on TAP equations [133], [169] (see Section 5.1), which generally did not result in a bound.

Stein Discrepancy and VI: Stein's method [161] was first proposed as an error bound to measure how well an approximate distribution fits a distribution of interest. [52], [96], [97], [98] introduce the Stein discrepancy to modern VI. Here, we introduce the Stein discrepancy and two VI methods that use it: Stein Variational Gradient Descent (SVDG) [97] and operator VI [136]. These two method share the same objective but are optimized in different manners.

A large class of divergences can be represented in the form of the Stein discrepancy [106], including the general f-divergence. In particular, [97], [136] used the Stein discrepancy as a divergence measure:

$$D_{\text{stein}}(p,q) = sup_{f\in\mathscr{F}} |\mathbb{E}_{q(\boldsymbol{z})}[f(\boldsymbol{z})] - \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[f(\boldsymbol{z})]|^2. \quad (20)$$

$\mathscr{F}$ indicates a set of smooth, real-valued functions. When $q(\boldsymbol{z})$ and $p(\boldsymbol{z}|\boldsymbol{x})$ are identical, the divergence is zero. More generally, the more similar p and q are, the smaller is the discrepancy.

The second term in Eq. 20 involves an expectation under the intractable posterior. Therefore, the Stein discrepancy can only be used in VI for classes of functions $\mathscr{F}$ for which the second term is equal to zero. We can find a suitable class with this property as follows. We define f by applying a differential operator $\mathscr{A}$ on another function ϕ, where ϕ is only restricted to be smooth:

$$f(\boldsymbol{z}) = \mathscr{A}_p\phi(\boldsymbol{z}).$$

The operator $\mathscr{A}$ is constructed in such a way that the second expectation in Eq. 20 is zero for arbitrary ϕ; all operators with this property are valid operators [136]. A popular operator that fulfills this requirement is the Stein operator:

$$\mathscr{A}_p\phi(\boldsymbol{z}) = \phi(\boldsymbol{z})\nabla_{\boldsymbol{z}}\log p(\boldsymbol{z},\boldsymbol{x}) + \nabla_{\boldsymbol{z}}\phi(\boldsymbol{z}).$$

Both operator VI [136] and SVGD [97] use the Stein discrepancy with the Stein operator to construct the variational objective. The main difference between these two methods lies in the optimization of the variational objective using the Stein discrepancy. Operator VI [136] uses a minimax (GAN-style) formulation and BBVI to optimize the variational objective directly; while Stein Variational Gradient Descent (SVGD) [97] uses a kernelized Stein discrepancy. With a particular choice of this kernel and q, it can be shown that SVGD determines the optimal perturbation in the direction of the steepest gradient of the KL divergence [97]. SVGD leads to a scheme where samples in the latent space are sequentially transformed to approximate the posterior. As such, the method is reminiscent of, though formally distinct from, a normalizing flow approach [140].

5.3 Structured Variational Inference

MFVI assumes a fully-factorized variational distribution; as such, it is unable to capture posterior correlations. Fully factorized variational models have limited accuracy, especially when the latent variables are highly dependent such as in models with hierarchical structure. This section examines variational distributions which are not fully factorized, but contain dependencies between the latent variables. These structured variational distributions are more expressive, but often come at higher computational costs.

The choice of which dependencies between latent variables in the variational distribution to maintain may lead to different degrees of performance improvement. To achieve the best performance, a customized choice of structure given a model is needed. For example, structured variational inference with LDA [58] shows

that maintaining global structure is vital, while structured variational inference with a Beta Bernoulli Process [156] shows that maintaining local structure is more important for good performance.

In the following, we detail recent advances in structured inference for hierarchical models and mixed membership models and discuss VI for time series. Certain other structured approximations emerge in the context of inference networks and are covered in Section 6.3.

Hierarchical VI: Maintaining a rich structure of dependencies between latent variables in the variational distribution drastically enhances the expressiveness of the variational approximation for many models. [137] proposes hierarchical variational models (HVM): a general framework that can be applied to different VI models such as inference networks, see Section 6. HVM suggests a general way to include correlations into the variational distribution. To this end, one treats the original mean field parameters as latent variables, places a prior $q(\boldsymbol{\lambda};\boldsymbol{\theta})$ over them, and marginalizes them out:

$$q(\boldsymbol{z};\boldsymbol{\theta}) = \int \left(\prod_i q(z_i;\lambda_i) \right) q(\boldsymbol{\lambda};\boldsymbol{\theta}) d\boldsymbol{\lambda}. \qquad (21)$$

The term $q(\boldsymbol{z};\boldsymbol{\theta})$ thus captures dependencies through the marginalization procedure. The resulting ELBO can be made tractable by further lower-bounding the resulting entropy and sampling from the hierarchical model. Notably, this approach is used in the development of the variational Gaussian Process (VGP) [179], a particular HVM. The VGP applies a Gaussian Process to generate variational estimates, thus forming a Bayesian non-parametric prior. Since GPs can model a rich class of functions, the VGP is able to confidently approximate diverse posterior distributions [179].

Boosting-inspired VI: To model dependencies, mixture models can be employed as variational distributions. Mixed membership models or mixture models are a special type of hierarchical model, and have been used in VI since the 1990s [45], [66], [71]. Mixture models can be fit using the aforementioned auxiliary bounds [137], or using boosting-inspired methods, explained as follows.

Boosting VI and variational boosting [51], [110] have been proposed independently. These algorithms refine the approximate posterior iteratively by adding one component at a time. Here we detail the ideas presented in [110]. Assume that the variational distribution is a mixture model with C components $q_c(z;\lambda_c)$ and corresponding component weights ρ_c, where $\sum_{c=1}^{C} \rho_c = 1$. Initially, one mixture component with $\rho_c = 1$ is fit to the posterior, resulting in a variational parameter λ_1. The second component is added with an initial weight ρ_2 to learn the variational parameter λ_2 and to estimate the weights ρ_1 and ρ_2 with weighted expectation maximization (EM). To guarantee that the weights sum up to 1, when adding a component c with weight ρ_c, the weights of the previously learned components are multiplied by $(1-\rho_c)$. The procedure constructs a multi-modal approximate posterior.

VI for Time Series: One of the most important model classes for structured variational approximations are time series models. Significant examples include Hidden Markov Models (HMM) [38] and Dynamic Topic Models (DTM) [16]. These models have strong dependencies between time steps, leading traditional fully factorized MFVI to produce unsatisfying results. When using VI for time series, one typically employs a structured variational distribution that explicitly captures dependencies between time points, while remaining fully-factorized in the remaining variables [11], [16], [39], [68]. This commonly requires model specific approximations. [39], [68] derive SVI for popular time series models including HMMs, hidden semi-Markov models (HSMM), and hierarchical Dirichlet process-HMMs. Moreover, [68] derived an accelerated SVI for HSMMs. [10], [11] derive a structured BBVI algorithm for non-conjugate latent diffusion models.

5.4 Other Non-Standard VI Methods

In this section, we cover a number of miscellaneous approaches which fall under the broad umbrella of improving VI accuracy but would not be categorized as alternative divergence measures or structured models.

VI by Stochastic Gradient Descent: Stochastic gradient descent on the negative log posterior of a probabilistic model can, under certain circumstances, be seen as an implicit VI algorithm. Here we consider SGD with constant learning rates (constant SGD) [101], [102], and early stopping [37].

Constant SGD can be viewed as a Markov chain that converges to a stationary distribution; as such, it resembles Langevin dynamics [188]. The variance of the stationary distribution is controlled by the learning rate. [101] shows that the learning rate can be tuned to minimize the KL divergence between the resulting stationary distribution and the Bayesian posterior. Additionally, [101] derived formulas for this optimal learning rate which resemble AdaGrad [36] and its relatives. A generalization of SGD that includes momentum and iterative averaging is presented in [102]. In contrast, [37] interprets SGD as a non-parametric VI scheme. The paper proposes a way to track entropy changes in the implicit variational objective based on estimates of the Hessian. As such, the authors consider sampling from distributions that are not stationary.

Robustness to Outliers and Local Optima: VI can benefit from advanced optimization methods, which aim to robustly escape to local optima. Variational tempering [103] adapts an efficient annealing approach [121], [145] to VI. It anneals the likelihood term with an adaptive tempering rate which can be applied either globally or locally to individual data points. Data points with associated small likelihoods under the model (such as outliers) are automatically assigned a high temperature. This reduces their influence on the global variational parameters, making the inference algorithm more robust to local optima. The same method can also be interpreted as data re-weighting [187], the weight being the inverse temperature. In this context, lower weights are assigned to outliers. Other stabilization techniques of note include the trust-region method [168], which uses the KL divergence to regulate the change between updating steps, and population VI [85], which uses bootstrapped populations to increase predictive performance.

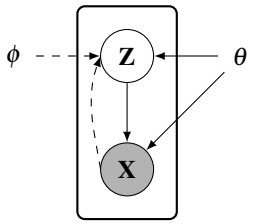
6 Amortized Variational Inference and Deep Learning

Finally, we review amoritzed variational inference. Traditional VI with local latent variables makes it necessary to optimize a separate variational parameter for each data point; this is computationally expensive. Amortized VI circumvents this problem by learning a deterministic function from the data to distributions over latent random variables. This replaces the local latent variables by the globally shared parameters of the function. In this section, we detail the main ideas behind this approach in Section 6.1 and how it is applied in form of variational autoencoders in Sections 6.2 and 6.3.

6.1 Amortized Variational Inference

The term amortized inference refers to utilizing inferences from past computations to support future computations [44]. For VI, amortized inference usually refers to inference over local variables. Instead of approximating separate variables for each data point, amortized VI assumes that these latent variables can be predicted by a parameterized function of the data. Thus, once this function is estimated, the latent variables can be acquired by passing new data points through the function. Deep neural networks used in this context are also called *inference networks*. Amortised VI with inference networks thus combines probabilistic modeling with the representational power of deep learning.

As an example, amortized inference has been applied to Deep Gaussian Processes (DGPs) [30]. Inference in these models is intractable, which is why the authors apply MFVI with inducing points (see Section 3.3) [30]. Instead of estimating these latent variables separately, however, [29] proposes to estimate these latent variables as functions of inference networks, allowing DGPs to scale to bigger datasets and speeding up convergence.



(a) VAE

Fig. 2. The graphical representation of a variational autoencoder; encoding (dashed lines) and decoding (solid lines).

6.2 Variational Autoencoders

Amortised VI has become a popular tool for inference in deep latent Gaussian models (DLGM). This leads to the concept of variational autoencoders (VAEs), which have been proposed independently by two groups [78], [141], and which are discussed in detail below. VAEs apply more generally than to DLGMs, but for simplicity we will focus this discussion on this popular class of models.

The Generative Model: In this paragraph we introduce the class of deep latent Gaussian models. The corresponding graphical model is depicted in Figure 2. The model employs a multivariate normal prior from which we draw a latent variable $\boldsymbol{z}$,

$$p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \mathbb{I}).$$

More generally, this could be a complex prior $p_\theta(\boldsymbol{z})$ that depends on additional parameters θ. The likelihood of the model is:

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \prod_{i=1}^{N} \mathcal{N}(x_i; \mu(z_i), \sigma(z_i)\mathbb{I}).$$

Most importantly, the likelihood depends on $\boldsymbol{z}$ through two non-linear functions $\mu(\cdot)$ and $\sigma(\cdot)$. These are typically neural networks, which take the latent variable as an input and transform it in a non-linear way. The data are then drawn from a normal distribution centered around the transformed latent variables $\mu(z_i)$. This provides a highly flexible density estimator. The parameter θ entails the parameters of the networks $\mu(\cdot)$ and $\sigma(\cdot)$. There exist many modified versions of this model specific to other types of data. For example, for binary data, the Gaussian likelihood can be replaced by a Bernoulli likelihood. Next, we review how amortized inference is applied to this model class.

Variational Autoencoders: Most commonly, VAEs refer to deep latent variable models which are trained using inference networks.

VAEs employ two deep sets of neural networks: a top-down generative model as described above, mapping from the latent variables $\boldsymbol{z}$ to the data $\boldsymbol{x}$, and a bottom-up inference model which approximates the posterior of the $p(\boldsymbol{z}|\boldsymbol{x})$. Commonly, these are referred to as the *generative network* and the *recognition network*.

In order to approximate the posterior, VAEs employ an amortized variational distribution as follows:

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \prod_{i=1}^{N} q_\phi(z_i|x_i).$$

As usual in amortized inference, this distribution does not depend on local variational parameters, but is instead conditioned on the data x_i and is parametrized by global parameters ϕ. This amortized variational distribution is typically chosen as:

$$q_\phi(z_i|x_i) = \mathcal{N}(z_i|\mu(x_i), \sigma^2(x_i)\mathbb{I}). \tag{22}$$

Similar to the generative model, the variational distribution employs non-linear mappings $\mu(x_i)$ and $\sigma(x_i)$ of the data in order to predict the posterior distribution of x_i. The parameter ϕ summarizes the corresponding neural network parameters.

The main contribution of [78], [141] was to derive a scalable and efficient training scheme for deep latent variable models. During optimization, both the inference network and the generative network are trained jointly to optimize the ELBO. The key to training these models is the reparameterization trick (see Section 4.3).

Stochastic gradients for the model's ELBO can be obtained as follows. One draws L local variable samples $\varepsilon_{(l)} \sim p(\varepsilon)$ with $l = 1:L$, to build the ELBO's Monte-Carlo approximation:

$$\begin{aligned}\hat{\mathscr{L}}(\theta, \phi, x_i) = &-D_{KL}(q_\phi(z|x_i)||p_\theta(z)) \\ &+ \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(x_i|g(\varepsilon_{(l)}, \mu(x_i), \sigma^2(x_i))).\end{aligned} \tag{23}$$

One uses the reparameterization trick, Eq. 23 to obtain stochastic gradients with respect to θ and ϕ.

The reparameterization trick also implies that the gradient variance is bounded by a constant [141]. The drawback of this approach however is that we require the posterior to be reparameterizable.

A Probabilistic Encoder-decoder Perspective: The term *variational autoencoder* arises from the fact that the joint training of the generative and recognition network resembles the structure of autoencoders, a class of unsupervised, deterministic models. Autoencoders are deep neural networks that are optimized to reconstruct their input $\boldsymbol{x}$ as closely as possible. Importantly, autoencoders have a hourglass structure which forces the information of the input to be compressed and filtered through a small number of units in the inner layers. These layers are thought to learn a low-dimensional manifold of the data.

In comparison, VAEs assume the low-dimensional representation of the data, the hidden variables $\boldsymbol{z}$, to be random variables and assign prior distributions to them. Thus, instead of a deterministic autoencoder, VAEs learn a probabilistic encoder and decoder which map between the data density and the latent variables.

A technical difference between variational and deterministic autoencoders is whether or not we inject noise into the stochastic layer during training. This can be thought of as a kind of regularization that avoids over-fitting.

6.3 Advancements in VAEs

Since the proposal of VAEs, an ever-growing number of extensions have been proposed.

While exhaustive coverage of the topic would require a review article in its own right, we summarize a few important extensions, including more expressive models and improved posterior inference. We also address specific problems of VAEs, such as the dying units problem.

More Expressive Likelihoods: One drawback of the standard VAE is the assumption that the likelihood factorizes over dimensions. This can be a poor approximation for images. In order to achieve a more expressive decoder, the Deep Recurrent Attentive Writer [49] relies on a recurrent structure that gradually constructs the observations while automatically focusing on regions of interest. In comparison, PixelVAE [50] tackles this problem by modeling dependencies between pixels within an image, using $p_\theta(x_i|z_i) = \prod_j p_\theta(x_i^j|x_i^1, ...x_i^{j-1}, z_i)$, where x_i^j denotes the jth dimension of observation i. The dimensions are generated in a sequential fashion, which accounts for local dependencies.

More Expressive Posteriors: Just as in other models, the mean field approach to VAEs suffers from a lack of expressiveness to model a complex posterior. This can be overcome by loosening the standard modeling assumptions of the inference network, such as the mean field assumption.

In order to tighten the variational bound, importance weighted variational autoencoders (IWAE) have been proposed [21]. IWAEs require L samples from the approximate posterior which are weighted by the ratio:

$$\hat{w}_l = \frac{w_l}{\sum_{l=1}^{L} w_l}, \text{where } w_l = \frac{p_\theta(x_i, z_{(i,l)})}{q_\phi(z_{(i,l)}|x_i)}. \tag{24}$$

A reinterpretation of IWAEs suggests that they are identical to VAEs but sample from a more expressive,

implicit distribution which converges to the true posterior as $L \to \infty$ [26].

The expressiveness of the posterior is also addressed in a series of papers on normalizing flows [25], [34], [35], [140] and [76]. The main idea behind normalizing flows is to transform a simple (e.g. mean field) approximate posterior $q(z)$ into a more expressive one by a series of successive invertible transformations. To this end, we first draw a random variable $z \sim q(z)$, and transform it using an invertible, smooth function f. Let $z' = f(z)$. Then the new distribution $q(z')$ is

$$q(z') = q(z)|\frac{\partial f^{-1}}{\partial z'}| = q(z)|\frac{\partial f}{\partial z'}|^{-1}. \quad (25)$$

It is important that we can compute the determinant since the variational approach requires us to also estimate the entropy of the transformed distribution. By choosing the function f such that $|\frac{\partial f}{\partial z'}|$ is easily computable, this normalizing flow constitutes an efficient method to generate multimodal distributions from a simple distribution. To this end, linear time-transformations, Langevin and Hamilton flow [140], as well as inverse autoregressive flow [76] and autoregressive flow [25] have been proposed.

The Dying Units Problem: While advances in posterior modeling are promising, VAEs can suffer from the optimization challenges that these models impose. The expressiveness of the decoder can, in some cases, be so strong, that the optimization ignores the latent code in the $\boldsymbol{z}$ variables. On one hand, this might be partly caused by the fact that the approximating posterior does not carry relevant information in the early stages of the optimization [18]. On the other hand, the decoder might be strong enough to model $p_\theta(\mathbf{x}|\mathbf{z})$ independent of $\mathbf{z}$ in which case the true posterior is the prior [25]. In these cases, the posterior is set to match the prior in order to satisfy the KL divergence in Eq. 23. Lossy variational autoencoders [25] circumvent this problem by condititioning the decoding distribution for each output dimension on partial input information. This forces the distribution to encode global information in the latent variables. A different approach is to apply an annealing scheme and slowly increase the influence of the prior, i.e. the KL divergence term, during training [159]. Furthermore, the generative distribution can be corrected by recursively adjusting it with a data dependent approximate likelihood term [158].

Inference Networks and Graphical Models: Instead of only relying on a neural network structure, [69] proposes a method to utilize a structured prior for VAEs. In this way, one combines the advantages of traditional graphical models and inference networks.

These hybrid models overcome the intractability of non-conjugate likelihood distributions by learning variational parameters of conjugate distributions with a recognition model. This allows one to approximate the posterior conditioned on the observations while maintaining conjugacy. As the encoder outputs an estimate of natural parameters, message passing, which relies on conjugacy, is applied to carry out the remaining inference.

Implicit Distributions: Traditional VI, including VAEs, relies on parametric models. This facilitates derivations and computations, but also limits our ability to model complex data. One way to enhance expressiveness is to employ *implicit distributions*. These are distributions generated by a deterministic or stochastic transformation of another, possibly simpler distribution. Implicit distributions do not have a parametric likelihood function (preventing us from having access to their entropy), but we can still sample from them, which is often enough to estimate the desired gradients. One popular example is to use Generative Adversarial Networks (GAN) [48] which learn implicit distributions to represent unknown densities.

Several authors have proposed implicit distributions for amortized inference [64], [73], [93], [105], [115]. When employing an implicit distribution as a variational distribution in VAEs, the standard training procedure does not apply because the entropy is intractable. Instead, a GAN-style discriminator can be trained to distinguish between the target distribution and the variational distribution [105]. As part of VI, we aim to approximate an expectation of the log density ratio $\log p(\boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{x})$ under q. employ a GAN-style discriminator T that discriminates the prior from the variational distribution, $T(\boldsymbol{x}, \boldsymbol{z}) = \log q_\phi(\boldsymbol{z}|\boldsymbol{x}) - \log p(\boldsymbol{z})$ [105].

7 Discussion

We have summarized recent advancements in variational inference. Here we outline some selected active research directions and open questions, including, but not limited to: theory of VI, VI and policy gradients, VI for deep learning (DL), and automatic VI.

Theory of VI: Despite progress in modeling and inference, few authors address theoretical aspects of VI. One important direction is quantifying the approximation errors involved when replacing a true posterior with a simplified variational distribution [118]. A related problem is the predictive error, e.g., when approximating the marginalization involved in a Bayesian predictive distribution using VI.

We also conjecture that VI theory could profit from a deeper connection with information theory. This was already exemplified in [165], [166]. Information theory also inspires the development of new models

and inference schemes [2], [12], [172]. For example, the information bottleneck [172] has recently led to the deep variational information bottleneck [2]. We expect more interesting results to come from fusing these two lines of research.

VI and Deep Learning: Despite its recent successes in various areas, deep learning still suffers from a lack of principled uncertainty estimation, a lack in interpretability of its feature representations, and difficulties in including prior knowledge. Bayesian approaches, such as Bayesian neural networks [122] and variational autoencoders (as reviewed in Section 6), are improving all these aspects. Recent work aims at using interpretable probabilistic models as priors for VAEs [32], [69], [84], [149]. In such models, VI is an essential component. Making VI computationally efficient and easy to implement in Bayesian deep architectures is becoming an important research direction [42].

VI and Policy Gradients: Policy gradient estimation is important for reinforcement learning (RL) [163] and stochastic control. The technical challenges in these applications are similar to VI [91], [98], [154], [186] (See Appendix A.8). As an example, SVGD has been applied in the RL setting as the Stein policy gradient [98]. The application of VI in reinforcement learning is currently an active area of research.

Automatic VI: Probabilistic programming allows practitioners to quickly implement and revise models without having to worry about inference. The user is only required to specify the model, and the inference engine will automatically perform the inference. Popular probabilistic programming tools include but are not limited to: *Stan* [24], which covers a large range of advanced VI and MCMC methods, *Infer.Net* [112], which is based on variational message passing and EP, *Automatic Statistician* [46] and *Anglican* [177], which mainly rely on sampling methods, and *Edward* [178], which supports BBVI as well as Monte Carlo sampling. The longstanding goal of these tools is to change the research methodology in probabilistic modeling, allowing users to quickly revise and improve models and to make them accessible to a broader audience.

Despite current efforts to make VI more accessible to practitioners, its usage is still not straightforward for non-experts. For example, manually identifying posterior symmetries and breaking these symmetries is necessary to work with Infer.Net. Furthermore, variance reduction methods such as control variates can drastically accelerate convergence, but a model specific design of control variates is needed to obtain the best performance. At the time of writing, these problems are not yet addressed in current probabilistic programming toolboxes. We believe these and other directions are important to advance the impact of probabilistic modeling in science and technology.

8 Conclusions

In this paper, we review the major advances in variational inference in recent years from four perspectives: scalability, generality, accuracy, and amortized inference. The advancement of variational inference theory and the adoption of approximate inference in new machine learning models are developing rapidly. Although this field has grown in recent years, it remains an open question how to make VI more efficient, more accurate, and easier to use for non-experts. Further development, as discussed in the previous section, is needed.

Acknowledgments

The authors would like to thank Yingzhen Li, Sebastian Nowozin, Tianfan Fu, Robert Bamler, and especially Andrew Hartnett for comments and discussions that greatly improved the manuscript.

References

[1] S Ahn, A Korattikara, and M Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. In *ICML*, 2012.

[2] A. Alemi, I. Fischer, J. Dillon, and K. Murphy. Deep variational information bottleneck. In *ICLR*, 2017.

[3] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1966.

[4] S.I. Amari. *Differential-geometrical methods in statistics*. Springer, 1985.

[5] S.I. Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2), 1998.

[6] S.I. Amari. α-divergence is unique, belonging to both f-divergence and bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11), 2009.

[7] E. Angelino, M. J. Johnson, and R. P. Adams. Patterns of scalable bayesian inference. *Foundations and Trends® in Machine Learning*, 9(2-3), 2016.

[8] A. Azevedo-Filho and R.D. Shachter. Laplace's method approximations for probabilistic inferencein belief networks with continuous variables. In *UAI*, 1994.

[9] L. Balles, J. Romero, and P. Hennig. Coupling adaptive batch sizes with learning rates. In *UAI*, 2017.

[10] R. Bamler and S. Mandt. Dynamic word embeddings. In *ICML*, 2017.

[11] R. Bamler and S. Mandt. Structured black box variational inference for latent time series models. In *ICML WS*, 2017.

[12] D. Barber and F. Agakov. The IM algorithm: a variational approach to information maximization. In *NIPS*, 2003.

[13] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

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

[15] D. M. Blei and J. D. Lafferty. Correlated topic models. In *NIPS*, volume 18, 2006.

[16] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*, 2006.

[17] L. Bottou. Large-scale machine learning with stochastic gradient descent. Springer, 2010.

[18] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Józefowicz, and S. Bengio. Generating sentences from a continuous space. In *CoNLL*, pages 10–21, 2016.

[19] S. Brooks, A. Gelman, G. Jones, and X.L Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

[20] T. D. Bui, J. Yan, and R. E. Turner. A unifying framework for sparse gaussian process approximation using power expectation propagation. *arXiv preprint arXiv:1605.07066*, 2016.

[21] Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[22] R.H. Byrd, G.M. Chin, J. Nocedal, and Y.C. Wu. Soamploample size selection in optimization methods for machine learning. *Mathematical programming*, 134(1), 2012.

[23] L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

[24] B. Carpenter, D. Lee, M. A. Brubaker, A. Riddell, A. Gelman, B. Goodrich, J. Guo, M. Hoffman, M. Betancourt, and P Li. Stan: A probabilistic programming language. *Journal of Statistical Software*, 2016.

[25] X. Chen, D. P. Kingma, T. Salimans, Y. Dua, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel. Variational lossy autoencoder. In *ICLR*, 2017.

[26] C. Cremer, Q. Morris, and D. Duvenaud. Reinterpreting importance-weighted autoencoders. In *ICLR WS*, 2017.

[27] D. Csiba and P. Richtárik. Importance sampling for minibatches. *arXiv preprint arXiv:1602.02283*, 2016.

[28] I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, 8, 1964.

[29] Z.W. Dai, A. Damianou, J. González, and N. Lawrence. Variational auto-encoded deep Gaussian processes. In *ICLR*, 2016.

[30] A. Damianou and N. Lawrence. Deep gaussian processes. In *AISTATS*, 2013.

[31] S. De, A. Yadav, D. Jacobs, and T. Goldstein. Automated inference using adaptive batch sizes. In *AISTATS*, 2017.

[32] Z.W. Deng, R. Navarathna, P. Carr, S. Mandt, Y.S. Yue, I. Matthews, and G. Mori. Factorized variational autoencoders for modeling audience reactions to movies. In *CVPR*, 2017.

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

[34] L. Dinh, D. Krueger, and Y. Bengio. NICE: nonlinear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[35] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. In *ICLR*, 2017.

[36] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12, 2011.

[37] D. Duvenaud, D. Maclaurin, and R. Adams. Early stopping as nonparametric variational inference. In *AISTATS*, 2016.

[38] S. R. Eddy. Hidden Markov Models. *Current opinion in structural biology*, 6(3), 1996.

[39] N. Foti, J. Xu, D. Laird, and E. Fox. Stochastic variational inference for hidden Markov models. In *NIPS*, 2014.

[40] M. P. Friedlander and M. Schmidt. Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing*, 34(3), 2012.

[41] T.F Fu and Z.H. Zhang. CPSG-MCMC: Clustering-based preprocessing method for stochastic gradient mcmc. In *AISTATS*, 2017.

[42] Y. Gal. *Uncertainty in deep learning*. PhD thesis, PhD thesis, University of Cambridge, 2016.

[43] Y. Gal, M. van der Wilk, and C. Rasmussen. Distributed variational inference in sparse gaussian process regression and latent variable models. In *NIPS*, 2014.

[44] S. Gershman and N. Goodman. Amortized inference in probabilistic reasoning. In *CogSci*, volume 36, 2014.

[45] S. J. Gershman, M. D. Hoffman, and D. M. Blei. Nonparametric variational inference. In *ICML*, 2012.

[46] Z. Ghahramani. Probabilistic Machine Learning and Artificial Intelligence. *Nature*, 521(7553), 2015.

[47] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

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

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

[50] I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. V. Visin, D. Vazquez, and A. Courville. PixelVAE: A latent variable model for natural images. 2017.

[51] F.J. Guo, X.Y. Wang, K. Fan, T. Broderick, and D. B. Dunson. Boosting variational inference. *arXiv preprint arXiv:1611.05559*, 2016.

[52] J. Han and Q. Liu. Stein variational adaptive importance sampling. In *UAI*, 2017.

[53] G. Heinrich. Parameter estimation for text analysis, 2008.

[54] P. Hennig. *Approximate inference in graphical models*. PhD thesis, University of Cambridge, 2011.

[55] J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *UAI*, 2013.

[56] J Hensman, M Rattray, and N. D Lawrence. Fast variational inference in the conjugate exponential family. In *NIPS*, 2012.

[57] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for Latent Dirichlet Allocation. In *NIPS*, 2010.

[58] M. D. Hoffman and D. M. Blei. Structured stochastic variational inference. In *AISTATS*, 2015.

[59] M D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *JMLR*, 14(1), 2013.

[60] R. V. Hogg and A. T. Craig. *Introduction to mathematical statistics.(5th edition)*. Upper Saddle River, New Jersey: Prentice Hall, 1995.

[61] A. Honkela, T. Raiko, M. Kuusela, M. Tornio, and J. Karhunen. Approximate riemannian conjugate gradient learning for fixed-form variational bayes. *JMLR*, 11(Nov), 2010.

[62] A. Honkela, M. Tornio, T. Raiko, and J. Karhunen. Natural conjugate gradient in variational inference. In *ICONIP*, pages 305–314, 2007.

[63] A. Honkela and H. Valpola. Online variational bayesian learning. In *International Symposium on Independent Component Analysis and Blind Signal Separation*, 2003.

[64] F. Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

[65] T. S. Jaakkol, L. K. Saul, and M. I. Jordan. Fast learning by bounding likelihoods in sigmoid type belief networks. *NIPS*, 1996.

[66] T. S. Jaakkola and M. I. Jordan. Improving the mean field approximation via the use of mixture distributions. *NATO ASI series D behavioural and social sciences*, 89, 1998.

[67] E. Jang, S.X. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017.

[68] M. Johnson and A. Willsky. Stochastic variational inference for bayesian time series models. In *ICML*, 2014.

[69] M. J. Johnson, D. Duvenaud, A. B. Wiltschko, S. R. Datta, and R. P. Adams. Structured VAEs: Composing probabilistic graphical models and variational autoencoders. In *NIPS*, 2016.

[70] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, 2013.

[71] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2), 1999.

[72] H. J. Kappen and W. Wiegerinck. Second order approximations for probability models. In *NIPS*, 2001.

[73] T. Karaletsos. Adversarial message passing for graphical models. In *NIPS WS*. 2016.

[74] M. E. Khan, P. Baqué, F. Fleuret, and P. Fua. Kullback-Leibler Proximal Variational Inference. In *NIPS*, 2015.

[75] N. King and N. Lawrence. Fast variational inference for gaussian process models through kl-correction. In *ECML*. Springer, 2006.

[76] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improving variational autoencoders with inverse autoregressive flow. In *NIPS*, 2016.

[77] D. P. Kingma, T. Salimans, and M. Welling. Variational Dropout and the Local Reparameterization Trick. In *NIPS*, 2015.

[78] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[79] D. P. Kingma and M. Welling. Stochastic gradient vb and the variational auto-encoder. In *ICLR*, 2014.

[80] D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[81] D. A. Knowles and T. P. Minka. Non-conjugate variational message passing for multinomial and binary regression. In *NIPS*, 2011.

[82] D.A. Knowles. Stochastic gradient variational Bayes for gamma approximating distributions. *arXiv*, page 1509.01631, 2015.

[83] A. Korattikara, V. Rathod, K. Murphy, and M. Welling. Bayesian Dark Knowledge. *arXiv preprint arXiv:1506.04416*, 2015.

[84] R. G. Krishnan, U. Shalit, and D. Sontag. Deep kalman filters. In *NIPS WS*, 2015.

[85] A. Kucukelbir and D. M. Blei. Population empirical bayes. In *UAI*, 2015.

[86] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

[87] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1), 1951.

[88] K. Kurihara, M. Welling, and Y. W. Teh. Collapsed variational dirichlet process mixture models. In *IJCAI*, 2007.

[89] P. S. Laplace. Memoir on the probability of the causes of events. *Statistical Science*, 1(3), 1986.

[90] M. Lázaro-Gredilla, S. Van Vaerenbergh, and N. D. Lawrence. Overlapping mixtures of gaussian processes for the data association problem. *Pattern Recognition*, 45(4), 2012.

[91] S. Levine and V. Koltun. Variational policy search via trajectory optimization. In *NIPS*, 2013.

[92] Y.Z. Li, J.M. Hernández-Lobato, and R.E. Turner. Stochastic expectation propagation. In *NIPS*, 2015.

[93] Y.Z. Li and Q. Liu. Wild variational approximations. In *NIPS WS*, 2016.

[94] Y.Z Li and r. e. Turner. Rényi divergence variational inference. In *NIPS*, 2016.

[95] Y.Z Li, M. Rowland, T. Bui, D. Hernandez-Lobato, and R. Turner. Black-Box Alpha Divergence Minimization. In *ICML*, 2016.

[96] Q. Liu, J. Lee, and M. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *ICML*, 2016.

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

[98] Y. Liu, P. Ramachandran, Q. Liu, and J. Peng. Stein variational policy gradient. In *UAI*, 2017.

[99] C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables, 2017.

[100] S. Mandt and D. Blei. Smoothed gradients for stochastic variational inference. In *NIPS*, 2014.

[101] S. Mandt, M. D. Hoffman, and D. M. Blei. A Variational Analysis of Stochastic Gradient Algorithms. In *ICML*, 2016.

[102] S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic gradient descent as approximate bayesian inference. *JMLR*, 2017.

[103] S. Mandt, J. McInerney, F. Abrol, R. Ranganath, and Blei. Variational Tempering. In *AISTATS*, 2016.

[104] J. McInerney, R. Ranganath, and D. Blei. The population posterior and bayesian modeling on streams. In *NIPS*, 2015.

[105] L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *ICML*, 2017.

[106] L. Mescheder, S. Nowozin, and A. Geiger. The numerics of gans. In *NIPS*, 2017.

[107] M. Mézard, G. Parisi, and M. A. Virasoro. Spin glass theory and beyond. 1990.

[108] Y. Miao, L. Yu, and P. Blunsom. Neural variational inference for text processing. In *ICML*, 2016.

[109] A. C. Miller, N. J. Foti, A. D'Amour, and R. P. Adams. Reducing reparameterization gradient variance. In *NIPS*, 2017.

[110] A.C. Miller, N. Foti, and R. P. Adams. Variational boosting: Iteratively refining posterior approximations. In *ICML*, 2017.

[111] T. Minka. Power EP. Technical report, Technical report, Microsoft Research, Cambridge, 2004.

[112] T. Minka, J. M. Winn, J. P. Guiver, S. Webster, Y. Zaykov, B. Yangel, A. Spengler, and J. Bronskill. Infer.NET 2.6. http://research.microsoft.com/infernet, 2014. Microsoft Research Cambridge.

[113] T. P. Minka. Expectation propagation for approximate bayesian inference. In *UAI*, 2001.

[114] T. P. Minka. Divergence measures and message passing. In *Microsoft Research Technical Report*, 2005.

[115] S. Mohamed and B. Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[116] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *UAI*, 1999.

[117] C. Naesseth, F. Ruiz, S. Linderman, and D. M. Blei. Reparameterization gradients through acceptance-rejection sampling algorithms. In *AISTATS*, 2017.

[118] S. Nakajima and S. Watanabe. Variational bayes solution of linear neural networks and its generalization performance. *Neural Computation*, 19(4), 2007.

[119] Eric Nalisnick and Padhraic Smyth. Stick-breaking variational autoencoders. In *ICLR*, 2017.

[120] R. Nallapati, W. Cohen, and J. Lafferty. Parallelized variational em for latent dirichlet allocation: An experimental evaluation of speed and scalability. In *ICDM WS*, 2007.

[121] R. M. Neal. Probabilistic inference using markov chain monte carlo methods. In *Technical Report*, 1993.

[122] R. M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[123] W. Neiswanger, C. Wang, and E. Xing. Embarrassingly parallel variational inference in nonconjugate models. *arXiv preprint arXiv:1510.04163*, 2015.

[124] M. Opper, M. Fraccaro, U. Paquet, A. Susemihl, and O. Winther. Perturbation theory for variational inference. NIPS WS, 2015.

[125] M. Opper, U. Paquet, and O. Winther. Perturbative corrections for approximate inference in gaussian latent variable models. *JMLR*, 14(1), 2013.

[126] M. Opper and D. Saad. *Advanced mean field methods: Theory and practice*. MIT press, 2001.

[127] M. Opper and O. Winther. A mean field algorithm for bayes learning in large feed-forward neural networks. In *NIPS*, 1996.

[128] M. Opper and O. Winther. Tractable approximations for probabilistic models: The adaptive thouless-anderson-palmer mean field approach. *Physical Review Letters*, 86(17), 2001.

[129] J. Paisley, D. M. Blei, and M. I. Jordan. Variational bayesian inference with stochastic search. In *ICML*, 2012.

[130] G. Parisi. *Statistical field theory*. Addison-Wesley, 1988.

[131] D Perekrestenko, V Cevher, and M Jaggi. Faster coordinate descent via adaptive importance sampling. In *AISTATS*, 2017.

[132] C. Peterson and J. R. Anderson. A mean field theory learning algorithm for neural networks. *Complex systems*, 1, 1987.

[133] T Plefka. Convergence condition of the TAP equation for the infinite-ranged ising spin glass model. *Journal of Physics A: Mathematical and general*, 15(6), 1982.

[134] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. 2008.

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

[136] R. Ranganath, D. Tran, J. Altosaar, and D. M. Blei. Operator variational inference. In *NIPS*. 2016.

[137] R. Ranganath, D. Tran, and D. M. Blei. Hierarchical variational models. In *ICML*, 2016.

[138] R. Ranganath, C. Wang, D. Blei, and E. Xing. An adaptive learning rate for Stochastic Variational Inference. In *ICML*, 2013.

[139] J. Raymond, A. Manoel, and M. Opper. Expectation propagation. *arXiv preprint arXiv:1409.6179*, 2014.

[140] D. Rezende and S. Mohamed. Variational Inference with Normalizing Flows. In *ICML*, 2015.

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

[142] B. Ringner. Law of the unconscious statistician. In *Unpublished Notes, Lund University*.

[143] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, 1951.

[144] G. Roeder, Y. Wu, and D. Duvenaud. Sticking the landing: An asymptotically zero-variance gradient estimator for variational inference. In *NIPS*, 2017.

[145] K. Rose, E. Gurewitz, and G. Fox. A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11(9), 1990.

[146] S. M. Ross. *Simulation*. Elsevier, 2006.

[147] F. J.R. Ruiz, M. K. Titsias, and D. M. Blei. The generalized reparameterization gradient. In *NIPS*. 2016.

[148] F. J.R. Ruiz, M. K. Titsias, and D. M. Blei. Overdispersed black-box variational inference. In *UAI*, 2016.

[149] A. Saeedi, M. D. Hoffman, S. J. DiVerdi, A. Ghandeharioun, M. J. Johnson, and R. P. Adams. Multimodal prediction and personalization of photo edits with deep generative models. *arXiv preprint arXiv:1704.04997*, 2017.

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

[151] T. Salimans and D. A. Knowles. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4), 2013.

[152] M. Sato. Online model selection based on the variational Bayes. *Neural Computation*, 13(7):1649–1681, 2001.

[153] L. K. Saul, T. Jaakkola, and M. I. Jordan. Mean field theory for sigmoid belief networks. *Journal of artificial intelligence research*, 4(1), 1996.

[154] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *ICML*, 2015.

[155] M. Seeger. Expectation propagation for exponential families. Technical report, 2005.

[156] A. Shah, D. A. Knowles, and Z. Ghahramani. An Empirical Study of Stochastic Variational Algorithms for the Beta Bernoulli Process. In *ICML*, 2015.

[157] E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. *NIPS*, 18, 2006.

[158] C. K. Sønderby, R. Raiko, L. Maaløe, S. Sønderby, and O. Winther. Ladder variational autoencoders. In *NIPS*, 2016.

[159] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther. How to train deep variational autoencoders and probabilistic ladder networks. In *ICML*, 2016.

[160] A. Srivastava and C. Sutton. Autoencoding variational inference for topic models. In *ICLR*, 2017.

[161] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Berkeley Symposium on Mathematical Statistics and Probability*, 1972.

[162] J. Sung, Z. Ghahramani, and S. Y. Bang. Latent-space variational bayes. *TPAMI*, 30(12), 2008.

[163] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[164] T. Tanaka. Estimation of third-order correlations within mean field approximation. In *NIPS*, 1998.

[165] T. Tanaka. A theory of mean field approximation. In *NIPS*, 1999.

[166] T. Tanaka. Information geometry of mean-field approximation. *Neural Computation*, 12(8), 2000.

[167] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *NIPS*, 2006.

[168] L. Theis and M. Hoffman. A Trust-region Method for Stochastic Variational Inference with Applications to Streaming Data. In *ICML*, 2015.

[169] D.J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of 'solvable model of a spin glass'. *Philosophical Magazine*, 35(3), 1977.

[170] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012.

[171] L. Tierney and J.B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393), 1986.

[172] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[173] M. Titsias and M. Lázaro-Gredilla. Local Expectation Gradients for Black Box Variational Inference. In *NIPS*, 2015.

[174] M. K. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *AISTATS*, 2009.

[175] M.K. Titsias and M. Lázaro-Gredilla. Variational heteroscedastic gaussian process regression. In *ICML*, 2011.

[176] S. Tokui and I. Sato. Evaluating the variance of likelihood-ratio gradient estimators. In *ICML*, 2017.

[177] D. Tolpin, J. W. van de Meent, H. Yang, and F. Wood. Design and implementation of probabilistic programming language anglican. *arXiv preprint arXiv:1608.05263*, 2016.

[178] D. Tran, A. Kucukelbir, A. B. Dieng, M. Rudolph, D. Liang, and D. M. Blei. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*, 2016.

[179] D. Tran, R. Ranganath, and D. M. Blei. The Variational Gaussian Process. *stat*, 1050, 2016.

[180] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 51(7), 2005.

[181] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2), 2008.

[182] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

[183] C. Wang and D. M. Blei. Variational inference in non-conjugate models. *JMLR*, 14(Apr), 2013.

[184] C. Wang, X. Chen, A. J. Smola, and E. P. Xing. Variance reduction for stochastic gradient optimization. In *NIPS*, 2013.

[185] C. Wang, J. Paisley, and D. M. Blei. Online variational inference for the hierarchical dirichlet process. In *AISTATS*, 2011.

[186] Y. Wang, G. Williams, E. Theodorou, and L. Song. Variational policy for guiding point processes. In *ICML*, 2017.

[187] Y. X. Wang, A. Kucukelbir, and D. M. Blei. Robust Probabilistic Modeling with Bayesian Data Reweighting. In *ICML*, 2017.

[188] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *ICML*, 2011.

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

[190] J. Winn and C. M. Bishop. Variational message passing. *JMLR*, 6, 2005.

[191] M.D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[192] K. Zhai, J. Boyd-Graber, N. Asadi, and M. L. Alkhouja. Mr. lda: A flexible large scale topic modeling package using variational inference in mapreduce. In *WWW*, 2012.

[193] C. Zhang. *Structured Representation Using Latent Variable Models*. PhD thesis, KTH Royal Institute of Technology, 2016.

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

[195] C. Zhang, C. H. Ek, X. Gratal, F. T. Pokorny, and H. Kjellström. Supervised hierarchical Dirichlet processes with variational inference. In *ICCV WS*, 2013.

[196] C. Zhang, H. Kjellstrom, and S. Mandt. Determinantal point processes for mini-batch diversification. In *UAI*, 2017.

[197] P.L. Zhao and T. Zhang. Accelerating minibatch stochastic gradient descent using stratified sampling. *arXiv preprint arXiv:1405.3080*, 2014.

[198] P.L. Zhao and T. Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *ICML*, 2015.

[199] S.D. Zhe, K.C. Lee, K. Zhang, and J. Neville. Online spike-and-slab inference with stochastic expectation propagation. In *NIPS WS*, 2016.

[200] H. Zhu and R. Rohwer. Information geometric measurements of generalisation. Technical report, 1995.

Cheng Zhang obtained her master's degree in System Control and Robotics from KTH Royal Institute of Technology in fall 2011. She received her Ph.D. degree at the Department of Robotics, Perception, and Learning (RPL), KTH, Sep 2016. She is currently a postdoctoral research associate at Disney Research, Pittsburgh. Her research interest lies in probabilistic models and approximate Bayesian inference, as well as their applications in computer vision and health care.

Judith Bütepage holds a Bsc degree in Cognitive Science from the University of Osnabrück. She is a doctoral student at the Department of Robotics, Perception, and Learning (RPL) at KTH in Stockholm, Sweden. Her research interests lie in the development of Bayesian latent variable models and their application to problems in Robotics and Computer Vision.

Hedvig Kjellström is a Professor and the head of the Department of Robotics, Perception, and Learning (RPL) at KTH in Stockholm, Sweden. Her present research focuses on the modeling of perception and production of human non-verbal communicative behavior and activity, with applications in Social Robotics, Performing Arts, and Healthcare. In 2010, she was awarded the Koenderink Prize for fundamental contributions in Computer Vision. She has written around 85 papers in the fields of Robotics, Computer Vision, Information Fusion, Machine Learning, Cognitive Science, Speech, and Human-Computer Interaction, and is an Associate Editor for IEEE TPAMI and IEEE RA-L.

Stephan Mandt is a Research Scientist at Disney Research Pittsburgh, where he leads the statistical machine learning research group. Previously, he was a postdoctoral researcher with David Blei at Columbia University, and a PCCM postdoctoral fellow at Princeton University. Stephan Mandt holds a Ph.D. in theoretical physics from the University of Cologne, supported by the German National Merit Foundation. His interests include scalable probabilistic modeling, stochastic processes, representation learning, variational inference, stochastic optimization, topic modeling, and machine learning applications in the sciences and technology.

APPENDIX A

A.1 ELBO and KL

We show that the difference between the marginal likelihood $\log p(\boldsymbol{x})$ and the ELBO $\mathscr{L}$ is the KL divergence between the variational distribution $q(\boldsymbol{z};\boldsymbol{\lambda})$ and the target distribution $p(\boldsymbol{x},\boldsymbol{z})$:

$$\begin{aligned}
&\log p(\boldsymbol{x}) - \mathscr{L} = \log p(\boldsymbol{x}) - \mathbb{E}_{q(\boldsymbol{z};\boldsymbol{\lambda})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z};\boldsymbol{\lambda})}\right] \\
&= \mathbb{E}_{q(\boldsymbol{z};\boldsymbol{\lambda})}\left[\log p(\boldsymbol{x}) - \log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z};\boldsymbol{\lambda})}\right] \\
&= -\mathbb{E}_{q(\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z})}\right] = D_{\mathrm{KL}}(q||p).
\end{aligned} \tag{26}$$

With this equivalence, the ELBO $\mathscr{L}$ can be derived using either Jensen's inequality as in Eq. 3, or using the KL divergence as $\mathscr{L} = \log p(x) - D_{\mathrm{KL}}(q||p)$.

A.2 Conjugate Exponential family

Many probabilistic models involve exponential family distributions. A random variable x is distributed according to a member of the exponential family if its density takes the form

$$p(x|\theta) = h(x)exp(\eta(\theta)t(x) - a(\eta(\theta))),$$

where θ is a vector of parameters, $\eta(\cdot)$ is the natural parameter, and $t(\cdot)$ are the sufficient statistics. Furthermore, $h(\cdot)$ is the base measure and $a(\cdot)$ is the log-normalizer. Many distributions fall into this class.

In the context of Bayesian statistics, certain exponential family distributions are conjugate pairs. A likelihood and prior distribution are a conjugate pair if the corresponding posterior distribution is in the same family as the prior. Examples for conjugate pairs include a Gaussian distribution with a Gaussian prior, a Poisson distribution with a gamma prior, or a multinomial distribution with a Dirichlet prior.

A.3 Variational Message Passing

Winn *et. al.* formulate MFVI in a message passing manner [190]. MFVI provides a method to update the latent variables of the variational distribution sequentially, as shown in Equation 8. In a Bayesian network, the update for each node only requires information from the nodes in its Markov blanket, which includes this node's parents, children, and co-parents of its children,

$$\begin{aligned}
q^*(z_j) \propto \exp(&\mathbb{E}_{q(\boldsymbol{z}_{\neg j})}\left[\log p(z_j|\boldsymbol{pa}_j)\right] \\
&+ \sum_{c_k \in \boldsymbol{ch}_j} \mathbb{E}_{q(\boldsymbol{z}_{\neg j})}\left[\log p(c_k|\boldsymbol{pa}_k)\right]),
\end{aligned} \tag{27}$$

where $\boldsymbol{pa}_j$ indicates the set of parent nodes of z_j, $\boldsymbol{ch}_j$ includes the set of the child nodes of z_j, and c_k indicates the kth child node. $\boldsymbol{pa}_k$ indicates the set of parent nodes of c_k. Hence, the update of one latent variable only depends on its parents, children, and its children's co-parents.

If we further assume that the model is conjugate-exponential, see Section A.2, a latent variable can be updated by receiving all messages from its parents and children. Here, each child node has already received messages from its co-parents. Thus, to update each node, only nodes in this node's Markov blanket are involved. Finally, z_j is updated with the following three steps: a) receive messages from all parents and children $m_{\boldsymbol{pa}_j \to z_j} = \langle t_{\boldsymbol{pa}_j} \rangle$, $m_{c_k \to z_j} = \tilde{\eta}_{c_k z_j}(\langle t_{c_k} \rangle, \{m_{i \to c_k}\}_{i \in \boldsymbol{pa}_k})$; b) update z_j's natural parameter η_{z_j}; c) update the expectation of z_j's sufficient statistic $\langle t(z_j) \rangle$.

Variational message passing provides a general message passing formulation for the MFVI. It enjoys all the properties of MFVI, but can be used in large scale Bayesian networks and can be automated easily. Together with EP, it forms the basis for the popular probabilistic programming tool Infer.Net [112].

A.4 Natural Gradients and SVI

Following [59], we show that using *natural* gradients instead of standard gradients in SVI simplifies the variational updates. We use the model example as shown in Figure 1 and assume that the true posterior of the global variable is the in exponential family:

$$\begin{aligned}
&p(\theta|x,\xi,\alpha) \\
&= h(\theta)\exp\Big(\eta_g(x,\xi,\alpha)^T t(\theta) - a_g\big(\eta_g(x,\xi,\alpha)^T t(\theta)\big)\Big).
\end{aligned}$$

We also assume that the variational distribution is in the same family:

$$q(\theta|\gamma) = h(\theta)\exp\Big(\gamma^T t(\theta) - a_g(\gamma)\Big).$$

Recall that γ is the variational parameter estimating the global variable θ. The subscript g in η_g and a_g denotes that these are the natural parameter and log-normalizer of the global variable. The natural gradient of a function $f(\gamma)$ is given by $\hat{\nabla}_\gamma f(\gamma) = G(\gamma)^{-1}\nabla_\gamma f(\gamma)$, where $G(\gamma)$ is the Fisher information matrix.

[59] showed that the ELBO has a closed-form solution in terms of its variational parameters γ:

$$\begin{aligned}
&\hat{\mathscr{L}}(\gamma) = \\
&\quad \mathbb{E}_q\left[\eta_g(\boldsymbol{x},\boldsymbol{z},\alpha)\right]\nabla_\gamma a_g(\gamma) - \gamma^T \nabla_\gamma a_g(\gamma) + a_g(\gamma) + c.
\end{aligned} \tag{28}$$

The constant c contains all those terms that are independent of γ. The gradient of Equation 28 is given by

$$\nabla_\gamma \hat{\mathscr{L}}(\gamma) = \nabla^2_\gamma a_g(\gamma)(\mathbb{E}_q\left[\eta_g(\boldsymbol{x},\boldsymbol{z},\alpha)\right] - \gamma).$$

Importantly, when $q(\theta|\gamma)$ is in the exponential family, then it holds that $G(\gamma) = \nabla^2_\gamma a_g(\gamma)$. Thus, the natural gradient simplifies to

$$\hat{\nabla}_\gamma \hat{\mathscr{L}}(\gamma) = \mathbb{E}_q\left[\eta_g(\boldsymbol{x},\boldsymbol{z},\alpha)\right] - \gamma.$$

Hence, the natural gradient has a simpler form than the regular gradient.

Following the natural gradient has the advantage that we do not optimize in the Euclidean space, which is often not able to represent distances between distributions, but in Riemann space, where distance is defined by the KL divergence, i.e. distance between distributions. More information about the advantages of using natural gradients can be found in [5].

A.5 Determinantal Point Processes

Point processes model the probability of a subset of points being sampled from a set of P points $\{1,2,...P\}$. Let L be a similarity matrix in which each entry $L^{i,j}$, describes the pair-wise similarity between two points i and j. The Determinantal Point Processes (DPP) states that the probability to sample a subset of points Y is proportional to the determinant of the sub-matrix $L_Y = [L^{i,j}]_{i,j\in Y}$

$$\mathscr{P}(Y) = \frac{det(L_Y)}{det(L+I)} \propto det(L_Y). \tag{29}$$

This results in a 'repulsive' effect, where similar points are less likely to be sampled together. More information about DPPs in machine learning can be found in [86].

A.6 Rao-Blackwell Theorem

Rao-Blackwellization is used in multiple VI methods for variance reduction such as in BBVI [135]. In general, the Rao-Blackwell Therorem [60] states the following: Let $\hat{\theta}$ be an estimator of parameter θ with $\mathbb{E}(\hat{\theta}^2) < \infty$ for all θ. Suppose that t is a sufficient statistic for θ, and let $\theta^* = \mathbb{E}(\hat{\theta}|t)$. Then for all θ,

$$E(\theta^* - \theta)^2 \leq E(\hat{\theta} - \theta)^2.$$

The inequality is strict unless $\hat{\theta}$ is a function of t. This implies that the conditional estimator $\theta^* = \mathbb{E}(\hat{\theta}|t)$, conditioned on the sufficient statistics, is a better estimator than any other estimator $\hat{\theta}$.

A.7 Physics Notations

In order to facilitate the comprehension of the older literature on VI, we introduce some notation commonly used by the physics community [126]. Distributions are commonly denoted by capital letters P and Q. We can write the KL divergence as:

$$KL(Q||P) = \log Z + \mathbb{E}_Q[\log P] - \mathbb{H}[Q],$$

which corresponds to Equation 26. Here, $\mathbb{H}$ denotes the entropy of a distribution. In the physics community, $-\log Z$ is called free energy. Z is the commonly the marginal likelihood in machine learning, and often called the partition function in physics. $E_Q[\log P]$ is called the variational energy and $F[Q] = E[\log P] - \mathbb{H}[Q]$ is the variational free energy which correspond to the negative ELBO, $F[Q] = -\mathscr{L}$.

A.8 Policy Gradient Estimation as VI

Reinforcement learning (RL) with policy gradients can be formulated as a VI problem [98]. In RL, the objective is to maximize the expected return

$$J(\theta) = J\big(\pi(a|s;\theta)\big) = \mathbb{E}_{\boldsymbol{s},\boldsymbol{a}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right], \tag{30}$$

where $\pi(a|s;\theta)$ indicates the policy parameterized by θ, r is a scalar reward for being in state s_t and performing action a_t at time t, and γ is the discount factor. The policy optimization can be formulated as a VI problem by using $q(\theta)$ – a variational distribution on θ – to maximize $\mathbb{E}_{q(\theta)}[J(\theta)]$. Using a max-entropy regularization, the optimization objective is

$$\mathscr{L} = \mathbb{E}_q(\theta)[J(\theta)] + \alpha H(q(\theta)). \tag{31}$$

This objective is the identical to the ELBO for VI.