A Contrastive Divergence for Combining Variational Inference and MCMC

Francisco J. R. Ruiz [1 2] **Michalis K. Titsias** [3]

Abstract

We develop a method to combine Markov chain Monte Carlo (MCMC) and variational inference (VI), leveraging the advantages of both inference approaches. Specifically, we improve the variational distribution by running a few MCMC steps. To make inference tractable, we introduce the variational contrastive divergence (VCD), a new divergence that replaces the standard Kullback-Leibler (KL) divergence used in VI. The VCD captures a notion of discrepancy between the initial variational distribution and its improved version (obtained after running the MCMC steps), and it converges asymptotically to the symmetrized KL divergence between the variational distribution and the posterior of interest. The VCD objective can be optimized efficiently with respect to the variational parameters via stochastic optimization. We show experimentally that optimizing the VCD leads to better predictive performance on two latent variable models: logistic matrix factorization and variational autoencoders (VAEs).

1. Introduction

Variational inference (VI) and Markov chain Monte Carlo (MCMC) are two of the main approximate Bayesian inference methods (Bishop, 2006; Murphy, 2012). While MCMC is asymptotically exact, VI enjoys other advantages: VI is typically faster, makes it easier to assess convergence, and enables amortized inference—a way to quickly approximate the posterior over the local latent variables.

A natural question is whether it is possible to combine MCMC and VI to leverage the advantages of each inference method. Such topic has attracted a lot of attention in the recent literature (see, e.g., Salimans et al., 2015; Maddison et al., 2017; Naesseth et al., 2018; Le et al., 2018; Hoffman, 2017; Li et al., 2017).

[1]University of Cambridge, Cambridge, UK [2]Columbia University, New York, USA [3]DeepMind, London, UK. Correspondence to: Francisco J. R. Ruiz <f.ruiz@columbia.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

We develop a method for combining VI and MCMC that improves an explicit variational distribution (i.e., with analytic density) by applying MCMC sampling. The method runs a few iterations of an MCMC chain initialized with a sample from the explicit distribution, so that each MCMC step successively improves the initial distribution.

Combining VI and MCMC in this way is challenging. Specifically, fitting the parameters of the explicit variational distribution is intractable under the standard VI framework. This is because the improved distribution, obtained by MCMC sampling, is defined implicitly, i.e., its density cannot be evaluated. Thus, we cannot minimize the Kullback-Leibler (KL) divergence between the improved variational distribution and the true posterior of interest.

To address this challenge, we develop a divergence that allows combining VI and MCMC in a principled manner. We refer to it as *variational contrastive divergence* (VCD). The VCD replaces the standard KL objective of VI, enabling tractable optimization. The key property of the VCD is that it is possible to obtain unbiased estimates of its gradient to perform stochastic optimization (Robbins & Monro, 1951). The VCD differs from the standard KL in that it captures a notion of discrepancy between the improved and the initial variational distributions. The properties of the VCD make it a valid objective function for Bayesian inference: it is non-negative and it becomes zero only when the variational distribution matches the posterior.

Additionally, as the number of MCMC steps increases, the VCD converges asymptotically to the symmetrized KL between the initial explicit variational distribution and the posterior. This implies that the variational distributions fitted by minimizing the VCD will exhibit larger variance than the distributions fitted with the standard KL divergence, even for moderate values of the number of MCMC iterations.

We fit the variational parameters of the initial distribution by following stochastic gradients of the VCD. In contrast to the method of Hoffman (2017) (which optimizes a different objective), the stochastic gradients of the VCD depend on the improved MCMC samples; therefore, the MCMC samples provide feedback to the optimization of the variational parameters. Unlike the method of Li et al. (2017), the VCD leads to stable optimization regardless of the number of MCMC steps, since it is a well-defined divergence.

We demonstrate the VCD on latent variable models, where there is a local latent variable corresponding to each observation, and the goal is to fit some global model parameters via approximate maximum likelihood (ML). In this setting, amortized inference allows us to obtain a quick approximation of the posterior over each latent variable. We refine the amortized variational distribution using MCMC, and we use the VCD objective to fit the parameters of the amortized distribution. We show experimentally that the models fitted with the VCD objective have better predictive performance than when we use alternative approaches, including standard VI and the method of Hoffman (2017).

2. Method Description

Here we formally describe the method for refining the variational distribution with Markov chain Monte Carlo (MCMC) sampling, as well as the variational contrastive divergence (VCD) that enables stochastic optimization.

Section 2.1 provides a brief background on variational inference (VI) and introduces the notation. Section 2.2 describes the improved distribution that uses MCMC sampling. Section 2.3 introduces the VCD, and Section 2.4 describes how to form unbiased estimators of its gradient. Finally, Section 2.5 summarizes the full algorithm.

2.1. Background: Variational inference

Consider the joint distribution $p(x, z)$ over the data x and latent variables z. We are interested in approximating the posterior $p(z \mid x)$. Consider a variational approximation $q_\theta(z)$, a family of distributions with parameter θ. The distribution $q_\theta(z)$ may also depend on x as in amortized VI, where $q_\theta(z) = q_\theta(z \mid x)$. To avoid clutter, we simply write $q_\theta(z)$ throughout this section.

The variational parameter θ is fitted by minimizing the Kullback-Leibler (KL) divergence between the variational distribution $q_\theta(z)$ and the posterior $p(z \mid x)$, $\mathrm{KL}(q_\theta(z) \;||\; p(z \mid x))$. This is equivalent to maximizing the evidence lower bound (ELBO),

$$\mathcal{L}_{\text{standard}}(\theta) = \mathbb{E}_{q_\theta(z)}\left[f_\theta(z)\right], \tag{1}$$

where we use the shorthand notation $f_\theta(z)$ for the argument of the expectation, which we call the "instantaneous ELBO,"

$$f_\theta(z) \triangleq \log p(x, z) - \log q_\theta(z). \tag{2}$$

2.2. Refining the variational approximation

We improve the variational distribution $q_\theta(z)$ by running an MCMC method initialized at $q_\theta(z)$ and whose stationary distribution is the posterior $p(z \mid x)$. Suppose that we apply such a Markov chain for a fixed number of iterations $t \in \mathbb{N}$. This results in a marginal distribution over the latent variables that we denote $q_\theta^{(t)}(z)$,

$$q_\theta^{(t)}(z) = \int Q^{(t)}(z \mid z_0) q_\theta(z_0) dz_0, \tag{3}$$

where $Q^{(t)}(z \mid z_0)$ denotes the overall transition kernel that takes an initial sample from $q_\theta(z_0)$ and after t iterations it produces a sample from $q_\theta^{(t)}(z)$.

The distribution $q_\theta^{(t)}(z)$ is an *improvement* of the variational distribution $q_\theta(z)$ for any t because it is closer to the posterior $p(z \mid x)$ in terms of KL divergence (Cover & Thomas, 2006, p. 81), i.e.,

$$\mathrm{KL}(q_\theta(z) \;||\; p(z \mid x)) \geq \mathrm{KL}(q_\theta^{(t)}(z) \;||\; p(z \mid x)). \tag{4}$$

If $q_\theta(z) \neq p(z \mid x)$, then $q_\theta^{(t)}(z)$ is strictly closer to the posterior and the above becomes a strict inequality. In the case where $q_\theta(z) = p(z \mid x)$, then the KL divergence is zero and cannot be reduced, therefore the KL divergence between $q_\theta^{(t)}(z)$ and the posterior is also zero,

$$\mathrm{KL}(q_\theta(z) \;||\; p(z \mid x)) = \mathrm{KL}(q_\theta^{(t)}(z) \;||\; p(z \mid x)) = 0. \tag{5}$$

We now develop a triangle inequality that will play an important role in the development of the divergence in Section 2.3. Specifically, we add the non-negative term $\mathrm{KL}(q_\theta^{(t)}(z) \;||\; q_\theta(z))$ to the left-hand side of Eq. 4,

$$\begin{aligned} &\mathrm{KL}(q_\theta(z) \;||\; p(z \mid x)) + \mathrm{KL}(q_\theta^{(t)}(z) \;||\; q_\theta(z)) \\ &\geq \mathrm{KL}(q_\theta^{(t)}(z) \;||\; p(z \mid x)). \end{aligned} \tag{6}$$

Triangle inequalities do not hold in general for KL divergences (since the KL is not a norm), but Eq. 6 holds because $q_\theta^{(t)}(z)$ is an improvement of $q_\theta(z)$ with the respect to the target $p(z \mid x)$.

While the original variational distribution $q_\theta(z)$ is typically tractable and allows for direct maximization of the ELBO in Eq. 1, we cannot directly use the improved distribution $q_\theta^{(t)}(z)$ as the variational approximation. The reason is that the resulting ELBO, given by

$$\mathcal{L}_{\text{improved}}(\theta) = \mathbb{E}_{q_\theta^{(t)}(z)}\left[\log p(x, z) - \log q_\theta^{(t)}(z)\right], \tag{7}$$

presents several challenges. First, the log-density of the improved distribution $\log q_\theta^{(t)}(z)$ cannot be evaluated because it is implicitly defined through Eq. 3. This makes standard stochastic optimization techniques for Monte Carlo-based maximization of Eq. 7 intractable to apply. Second, the optimization of Eq. 7 can be difficult because $q_\theta^{(t)}(z)$ depends more weakly on the variational parameters θ than the initial $q_\theta(z)$, since the former is closer to the posterior

$p(z \mid x)$, which is independent of θ. In fact, in the limit when the number of MCMC steps is very large ($t \to \infty$), the objective $\mathcal{L}_{\text{improved}}(\theta)$ becomes independent of θ. Thus, it is challenging to use the improved distribution as the variational distribution directly.

We next introduce a divergence that can be tractably optimized while avoiding these two challenges.

2.3. The variational contrastive divergence

Our goal is to find an alternative divergence that can be tractably and efficiently optimized over θ. The divergence we wish to find needs to satisfy two conditions: (i) it must be non-negative for any value of the variational parameters θ, and (ii) it must become zero only when the variational distribution $q_\theta(z)$ matches the posterior $p(z \mid x)$.

A first idea to form such a divergence is to start from the inequality in Eq. 4, bring everything to the left-hand side, and express the discrepancy between $q_\theta(z)$ and its improvement $q_\theta^{(t)}(z)$ in terms of the difference

$$\mathcal{L}_{\text{diff}}(\theta) = \text{KL}(q_\theta(z) \,||\, p(z \mid x)) - \text{KL}(q_\theta^{(t)}(z) \,||\, p(z \mid x)). \tag{8}$$

This is a proper divergence since it satisfies the two criteria outlined above. It takes non-negative values because $q_\theta^{(t)}(z)$ reduces the KL divergence, and it becomes zero only when $q_\theta(z) = p(z \mid x)$, as discussed in Section 2.2.

However, the objective $\mathcal{L}_{\text{diff}}(\theta)$ is still intractable to minimize due to the term $\log q_\theta^{(t)}(z)$ that appears in a similar way as in Eq. 7. To address that, we make use of the triangle inequality in Eq. 6 and modify the divergence above by adding the regularization term $\text{KL}(q_\theta^{(t)}(z) \,||\, q_\theta(z))$, which adds an extra force for reducing the discrepancy between the initial distribution and its improvement. This leads to the VCD divergence,

$$\mathcal{L}_{\text{VCD}}(\theta) \triangleq \mathcal{L}_{\text{diff}}(\theta) + \text{KL}(q_\theta^{(t)}(z) \,||\, q_\theta(z)). \tag{9}$$

The VCD is also a proper divergence. It is non-negative due to the triangle inequality (see Eq. 6) and it becomes zero only when $q_\theta(z)$ matches the posterior $p(z \mid x)$.

The VCD in Eq. 9 is tractable, in the sense that we can obtain unbiased stochastic gradients with respect to θ without evaluating the density of the improved distribution $q_\theta^{(t)}(z)$. In fact, unlike most common divergences, we can also obtain unbiased stochastic estimates of the actual value of the VCD. The reason is that the problematic term $\log q_\theta^{(t)}(z)$ now cancels out. To see that, we rearrange the terms in Eq. 9 (see Appendix 1) and express the divergence as

$$\mathcal{L}_{\text{VCD}}(\theta) = -\mathbb{E}_{q_\theta(z)}\left[f_\theta(z)\right] + \mathbb{E}_{q_\theta^{(t)}(z)}\left[f_\theta(z)\right]. \tag{10}$$

The VCD contains two terms. The first term is the negative standard ELBO (Eq. 1), which involves only the tractable distribution $q_\theta(z)$. The second term is also an expectation of the instantaneous ELBO (Eq. 2), taken with respect to the improved distribution $q_\theta^{(t)}(z)$ instead. Even though the expectation is taken with respect to $q_\theta^{(t)}(z)$, the argument of the expectation no longer involves the improved distribution. This allows us to form Monte Carlo estimates of the gradient of $\mathcal{L}_{\text{VCD}}(\theta)$ with respect to the variational parameters, as detailed in Section 2.4.

Asymptotic properties of the VCD. In the limit when $t \to \infty$, the improved distribution $q_\theta^{(t)}(z)$ converges to the posterior $p(z \mid x)$, and the divergence $\mathcal{L}_{\text{VCD}}(\theta)$ converges to the symmetrized KL divergence between the variational distribution and the posterior,[1] $\text{KL}_{\text{sym}}(q_\theta(z) \,||\, p(z \mid x)) = \text{KL}(q_\theta(z) \,||\, p(z \mid x)) + \text{KL}(p(z \mid x) \,||\, q_\theta(z))$. This ensures that $\mathcal{L}_{\text{VCD}}(\theta)$ depends on the variational parameters even when the number of MCMC steps is large, unlike the objective in Eq. 7. Moreover, the VCD favors variational distributions $q_\theta(z)$ with larger variance than the standard ELBO, as it converges to the symmetrized KL divergence.

We can form a generalization of the VCD that interpolates between the standard and the symmetrized KL according to a parameter α (see Appendix 2). The experimentation of this generalization is left for future work.

2.4. Taking gradients of the VCD

We now show how to estimate the gradient of the VCD with respect to θ. The first term in Eq. 10 is the negative standard ELBO, for which we can obtain unbiased gradients with respect to θ by using either the score function estimator (or REINFORCE) (Carbonetto et al., 2009; Paisley et al., 2012; Ranganath et al., 2014) or the reparameterization gradient (Rezende et al., 2014; Titsias & Lázaro-Gredilla, 2014; Kingma & Welling, 2014). Assuming that $q_\theta(z)$ is reparameterizable as $\varepsilon \sim q(\varepsilon)$, $z = h_\theta(\varepsilon)$, then the reparameterization gradient of the (negative) first term is

$$\nabla_\theta \mathbb{E}_{q_\theta(z)}\left[f_\theta(z)\right] = \mathbb{E}_{q(\varepsilon)}\left[\nabla_z f_\theta(z)\big|_{z=h_\theta(\varepsilon)} \times \nabla_\theta h_\theta(\varepsilon)\right], \tag{11}$$

which can be estimated with samples $\varepsilon \sim q(\varepsilon)$.

We now focus on the non-standard second term. We express its gradient as an expectation with respect to the improved distribution $q_\theta^{(t)}(z)$, from which we can sample, thus enabling stochastic optimization. To achieve that, we write the gradient of the second term in Eq. 10 as the sum of two expectations with respect to $q_\theta^{(t)}(z)$,

$$\begin{aligned}\nabla_\theta \mathbb{E}_{q_\theta^{(t)}(z)}\left[f_\theta(z)\right] &= -\mathbb{E}_{q_\theta^{(t)}(z)}\left[\nabla_\theta \log q_\theta(z)\right] \\ &+ \mathbb{E}_{q_\theta(z_0)}\left[\mathbb{E}_{Q^{(t)}(z \mid z_0)}[f_\theta(z)]\, \nabla_\theta \log q_\theta(z_0)\right].\end{aligned} \tag{12}$$

This expression allows us to form Monte Carlo estimates of

[1] This can be seen by substituting $q_\theta^{(t)}(z) = p(z \mid x)$ in Eq. 9.

the gradient using samples from $q_\theta^{(t)}(z)$ and does not require to evaluate the intractable log-density $\log q_\theta^{(t)}(z)$. Samples from $q_\theta^{(t)}(z)$ can be obtained by first sampling $z_0 \sim q_\theta(z)$ and then running t MCMC steps, $z \sim Q^{(t)}(z \mid z_0)$.

Proof of Eq. 12. We now show how to derive Eq. 12. We first apply the product rule for derivatives,

$$\nabla_\theta \mathbb{E}_{q_\theta^{(t)}(z)}\left[f_\theta(z)\right] = \int q_\theta^{(t)}(z) \times \nabla_\theta f_\theta(z) dz + \int \nabla_\theta q_\theta^{(t)}(z) \times f_\theta(z) dz. \tag{13}$$

The first integral is straightforward to unbiasedly approximate by drawing samples from $q_\theta^{(t)}(z)$, and hence it is directly one of the terms in Eq. 12. Note that, since the model $p(x, z)$ does not depend on θ, the gradient of the instantaneous ELBO is $\nabla_\theta f_\theta(z) = -\nabla_\theta \log q_\theta(z)$.

For the term $\nabla_\theta q_\theta^{(t)}(z)$ in the second integral, we substitute the definition of the improved distribution in Eq. 3 and apply the log-derivative trick, yielding

$$\begin{aligned}\nabla_\theta q_\theta^{(t)}(z) &= \nabla_\theta \int Q^{(t)}(z \mid z_0) q_\theta(z_0) dz_0 \\ &= \int Q^{(t)}(z \mid z_0) q_\theta(z_0) \nabla_\theta \log q_\theta(z_0) dz_0.\end{aligned} \tag{14}$$

Here we have made use of the fact that the t-step MCMC kernel $Q^{(t)}(z \mid z_0)$ is independent of θ. Finally, we obtain Eq. 12 by substituting Eq. 14 into Eq. 13. □

Controlling the variance. Since the estimator based on (the last line of) Eq. 12 is a score function estimator, it may suffer from high variance. We form a simple control variate to reduce the variance, obtained as an exponentially decaying average of the previous stochastic values.

More in detail, consider the term in the last line of Eq. 12, $\mathbb{E}_{q_\theta(z_0)}\left[w_\theta(z_0) \times \nabla_\theta \log q_\theta(z_0)\right]$, where we have defined $w_\theta(z_0) \triangleq \mathbb{E}_{Q^{(t)}(z \mid z_0)}\left[f_\theta(z)\right]$. This expectation can be equivalently written using a control variate C as $\mathbb{E}_{q_\theta(z_0)}\left[(w_\theta(z_0) - C) \times \nabla_\theta \log q_\theta(z_0)\right]$, as long as C does not depend on z_0. We set C as an exponentially decaying average of the *previous* stochastic values of $w_\theta(z_0)$, i.e., the values from previous iterations of gradient descent. At a given iteration of gradient descent, the one-sample stochastic estimate for $w_\theta(z_0)$ is simply $f_\theta(z)$.

Finally, note that when the number of MCMC iterations t is very large, the resulting estimator based on Eq. 12 has much lower variance, and control variates might not be needed. The reason is that the distribution $q_\theta^{(t)}(z)$ becomes less dependent on θ, and is gradient (Eq. 14) becomes zero. In that case, Eq. 12 simplifies as $\nabla_\theta \mathbb{E}_{q_\theta^{(t)}(z)}\left[f_\theta(z)\right] \approx -\mathbb{E}_{q_\theta^{(t)}(z)}\left[\nabla_\theta \log q_\theta(z)\right]$, and the resulting estimator based on this term only has much lower variance than the estimator based on the two terms.

2.5. Full algorithm

We now summarize the full algorithm to minimize the VCD from Section 2.3. The algorithm forms a one-sample estimator of the gradient of the VCD. For that, it first samples $z_0 \sim q_\theta(z_0)$ from the initial distribution (this can be done using reparameterization) and then runs t MCMC steps to obtain the improved sample $z \sim Q^{(t)}(z \mid z_0)$.

The algorithm uses the initial sample z_0 to form an unbiased estimator of the gradient of the standard term, $\nabla_\theta \mathbb{E}_{q_\theta(z_0)}\left[f_\theta(z_0)\right]$, using Eq. 11 (alternatively, the score function estimator can be used instead). With both samples z_0 and z, the algorithm makes use of Eq. 12 to estimate the non-standard gradient $\nabla_\theta \mathbb{E}_{q_\theta^{(t)}(z)}\left[f_\theta(z)\right]$. The control variate C is used in this step to reduce the variance of the estimator. Finally, the two terms are combined following Eq. 10 to obtain the gradient estimator of the divergence.

The full procedure is given in Algorithm 1.[2] It has two additional parameters. The first one is the decay parameter γ for the updates of the control variate. We set $\gamma = 0.9$. The second parameter is the stepsize ρ. We set the stepsize using RMSProp (Tieleman & Hinton, 2012); at each iteration ℓ we set $\rho^{(\ell)} = \eta/(1 + \sqrt{G^{(\ell)}})$, where η is the learning rate, and the updates of $G^{(\ell)}$ depend on the gradient estimate $\widehat{\nabla}_\theta \mathcal{L}_{\text{VCD}}^{(\ell)}$ as $G^{(\ell)} = 0.9 G^{(\ell-1)} + 0.1(\widehat{\nabla}_\theta \mathcal{L}_{\text{VCD}}^{(\ell)})^2$.

3. Related Work

There are several related works in the literature. For example, Salimans et al. (2015) combine Markov chain Monte Carlo (MCMC) and variational inference (VI) by introducing auxiliary variables (associated with the MCMC iterations) that need to be inferred together with the rest of the variables. Other works use rejection sampling within the variational framework (Naesseth et al., 2017; Grover et al., 2018) or meld sequential Monte Carlo and VI (Maddison et al., 2017; Naesseth et al., 2018; Le et al., 2018).

More related to ours is the work of Hoffman (2017); Li et al. (2017); Zhang et al. (2018); Titsias (2017). Specifically, the method of Hoffman (2017) performs approximate maximum likelihood (ML) estimation in non-linear latent variable models based on the variational autoencoder (VAE). The E-step of the approximate ML procedure minimizes the standard Kullback-Leibler (KL) divergence $\text{KL}(q_\theta(z) \mid\mid p(z \mid x))$, where $q_\theta(z)$ is an explicit amortized distribution; and the M-step updates the model parameters using an improved MCMC distribution $q_\theta^{(t)}(z)$. Since it uses the standard KL

[2] Code is available online at https://github.com/franrruiz/vcd_divergence.

Algorithm 1 Minimization of the VCD

Input: data x, variational family $q_\theta(z)$, number of MCMC iterations t
Output: variational parameters θ
Initialize θ randomly, initialize $C = 0$
while not converged **do**
  `# Sample from q:`
  Sample $z_0 \sim q_\theta(z_0)$
  Sample $z \sim Q^{(t)}(z \mid z_0)$ (run t MCMC steps)
  `# Estimate the gradient:`
  Estimate $\widehat{\nabla}_\theta \mathbb{E}_{q_\theta(z_0)}[f_\theta(z_0)]$ (Eq. 11)
  Estimate $\widehat{\nabla}_\theta \mathbb{E}_{q_\theta^{(t)}(z)}[f_\theta(z)]$ (Eq. 12 with control var.)
  Obtain $\widehat{\nabla}_\theta \mathcal{L}_{\text{VCD}} = \widehat{\nabla}_\theta \mathbb{E}_{q_\theta^{(t)}(z)}[f_\theta(z)] - \widehat{\nabla}_\theta \mathbb{E}_{q_\theta(z)}[f_\theta(z)]$
  `# Update the control variate:`
  Set $C \leftarrow \gamma C + (1-\gamma) f_\theta(z)$
  `# Take gradient step:`
  Set $\theta \leftarrow \theta - \rho \cdot \widehat{\nabla}_\theta \mathcal{L}_{\text{VCD}}$
end while

divergence of VI, the MCMC procedure does not provide feedback when learning the variational parameters θ.

The amortized MCMC technique of Li et al. (2017) incorporates feedback from MCMC back to the parameters of the explicit distribution. Their method aims at learning θ by minimizing the KL divergence between the improved distribution and its initialization, i.e., $\mathrm{KL}(q_\theta^{(t)}(z) \,||\, q_\theta(z))$. This divergence is intractable, and Li et al. (2017) approximate its gradient $\nabla_\theta \mathrm{KL}(q_\theta^{(t)}(z) \,||\, q_\theta(z)) \approx \mathbb{E}_{q_\theta^{(t)}(z)}[\nabla_\theta \log q_\theta(z)]$, which ignores the dependence of $q_\theta^{(t)}(z)$ on θ. Subsequently, this can lead to unstable optimization over θ, especially for small values of t, when $q_\theta^{(t)}(z)$ strongly depends on θ. The procedure becomes stable for large t, when the MCMC chain converges and $q_\theta^{(t)}(z) = p(z \mid x)$. Notice that the expectation $\mathbb{E}_{q_\theta^{(t)}(z)}[\nabla_\theta \log q_\theta(z)]$ appears also in our approach (see Eq. 12), but it is just one part of the overall gradient for optimizing $\mathcal{L}_{\text{VCD}}(\theta)$, a divergence that leads to stable optimization.

Zhang et al. (2018) described a method to combine MCMC with VI that tries to directly minimize $\mathrm{KL}(q_\theta^{(t)}(z) \,||\, p(z \mid x))$, which as discussed in Section 2.2 is intractable. Zhang et al. (2018) drop the intractable entropy term from the evidence lower bound (ELBO); again this can result in unstable and inaccurate optimization for a small number of MCMC steps t.

Finally, Titsias (2017) proposed to firstly apply a model reparameterization so that the exact marginal likelihood is preserved, i.e., $\int p(x,z)dz = \int p(x, g(\epsilon\,;\,\theta))J(\epsilon\,;\,\theta)d\epsilon$, where $g(\epsilon\,;\,\theta)$ is a parametrized invertible transformation and $J(\epsilon\,;\,\theta)$ the determinant of its Jacobian. The method then learns the variational parameters θ by maximizing an ELBO under an MCMC distribution using an EM-like procedure. While such an approach can more accurately minimize a divergence of the form $\mathrm{KL}(q_\theta^{(t)}(z) \,||\, p(z \mid x))$, it can suffer from the weak gradient problem (see Section 2.2), and it is only applicable to differentiable models.

The variational contrastive divergence (VCD) developed in this paper shares also similarities with contrastive divergence procedures for performing ML estimation of model parameters in undirected graphical models such as restricted Boltzmann machines (Hinton, 2002). There, the loss function takes a similar discrepancy form between KL divergences that involve the actual data distribution $p_{\text{data}}(x)$ and an MCMC-improved distribution $p^{(t)}(x)$ that converges to the model distribution $p_{\text{model}}(x)$. The fundamental difference with our method is that these approaches are designed for model parameter estimation while ours is suitable for approximate Bayesian inference.

4. Experiments

Here we demonstrate the algorithm described in Section 2.5, which minimizes the variational contrastive divergence (VCD) with respect to the variational parameters θ.

In Section 4.1, we showcase the procedure on a set of toy experiments involving a two-dimensional target distribution. We show that the variational distribution $q_\theta(z)$ fitted by minimizing the divergence $\mathcal{L}_{\text{VCD}}(\theta)$ has higher variance than the variational distribution fitted by minimizing the standard Kullback-Leibler (KL) divergence. In Section 4.2, we run experiments on two latent variable models, namely, a matrix factorization model and a variational autoencoder (VAE), using amortized variational distributions. We show that the resulting models fitted with the VCD achieve better predictive performance on held-out data.

4.1. Toy experiments

To showcase the VCD, we approximate a set of synthetic distributions defined on a two-dimensional space: a Gaussian, a mixture of two Gaussians, and a banana distribution. Their densities are given in Table 1.

Our goal is to experimentally check that the VCD favors higher variance distributions $q_\theta(z)$ compared to the standard KL divergence used in variational inference (VI). This is because the divergence $\mathcal{L}_{\text{VCD}}(\theta)$ converges asymptotically to the symmetrized KL divergence between $q_\theta(z)$ and the target distribution (see Section 2.3).

We use two different variational families $q_\theta(z)$: a Gaussian distribution and a mixture of two Gaussians. In both cases, the Gaussian components have diagonal covariances. (See Appendix 3 for the mathematical details on minimizing the

Table 1. Synthetic distributions used in the toy experiments.

name	$p(z)$
Gaussian	$\mathcal{N}\left(z \,\middle|\, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.95 \\ 0.95 & 1 \end{bmatrix}\right)$
mixture	$0.3\mathcal{N}\left(z \middle| \begin{bmatrix} 0.8 \\ 0.8 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right) + 0.7\mathcal{N}\left(z \middle| \begin{bmatrix} -2 \\ -2 \end{bmatrix}, \begin{bmatrix} 1 & -0.6 \\ -0.6 & 1 \end{bmatrix}\right)$
banana	$\mathcal{N}\left(\begin{bmatrix} z_1 \\ z_2 + z_1^2 + 1 \end{bmatrix} \middle| \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$

VCD for these specific variational distributions.)

Experimental settings. Our algorithm of choice to improve the variational distribution is Hamiltonian Monte Carlo (HMC) (Neal, 2011). We set the number of HMC iterations $t = 3$. We use 5 leapfrog steps.

We run 20,000 iterations of Algorithm 1 (50,000 iterations instead when $q_\theta(z)$ is a mixture). We set the learning rate $\eta = 0.1$ for the mean parameters, $\eta = 0.005$ for the standard deviation, and $\eta = 0.001$ for the mixture weights. We additionally decrease the learning rate by a factor of 0.9 every 2,000 iterations.

Results. Figure 1 shows the contour plots of the synthetic target distributions (green), together with the contour plots of the fitted variational distribution $q_\theta(z)$. For comparisons, we show the distribution $q_\theta(z)$ obtained when optimizing the standard KL divergence in Eq. 1 (blue), together with the distribution $q_\theta(z)$ obtained when optimizing the VCD in Eq. 10 (red). Both variational methods were initialized to the same values. In Figures 1(a) to 1(c), the variational distribution is a factorized Gaussian; in Figure 1(d) it is a two-component Gaussian mixture.

In all cases, the resulting variational distribution $q_\theta(z)$ has higher variance when the objective is the VCD. As discussed above, this is due to the asymptotic properties of the divergence; specifically, the VCD eventually converges to the symmetrized KL divergence between $q_\theta(z)$ and the target. This effect is apparent despite the small number of HMC iterations ($t = 3$) because the target is a simple two-dimensional distribution.

4.2. Experiments on latent variable models

We now consider latent variable models, where there is a local latent variable z_n for each observation x_n in the dataset, and the joint distribution factorizes as $p_\phi(x, z) = \prod_n p(z_n) p_\phi(x_n \mid z_n)$. Here, ϕ is a model parameter that we wish to learn via maximum likelihood (ML).

We use amortized VI for inference over the latent variables, i.e., the distribution $q_\theta(z_n) \equiv q_\theta(z_n \mid x_n)$. The improved variational distribution, after running t HMC steps, is $q_\theta^{(t)}(z_n \mid x_n) = \int Q_n^{(t)}(z_n \mid z_n^{(0)}) q_\theta(z_n^{(0)} \mid x_n) dz_n^{(0)}$, where the t-step HMC kernel $Q_n^{(t)}(z_n \mid z_n^{(0)})$ has the posterior $p_\phi(z_n \mid x_n)$ as its stationary distribution.

Our goal is to show empirically that the fitted models lead to better predictive performance when we optimize the variational parameters θ by minimizing the VCD, compared to the standard VI setting.

Models. We consider two models. The first one is Bayesian logistic matrix factorization, a model of binary data. The likelihood $p_\phi(x_n \mid z_n) = \prod_d p_\phi(x_{nd} \mid z_n)$ is a product of Bernoulli distributions, each with parameter $\text{sigmoid}(z_n^\top \phi_d + \phi_d^{(0)})$, where $\text{sigmoid}(x) = 1/(1 + e^{-x})$. The model parameters are the weights ϕ_d and intercepts $\phi_d^{(0)}$ for each dimension d.

The second model is a VAE (Kingma & Welling, 2014), which uses a density network (MacKay, 1995) to define the likelihood $p_\phi(x_n \mid z_n)$. That is, the likelihood is parameterized by a neural network with parameters ϕ. We choose a fully connected neural network with two hidden layers of 200 hidden units each and ReLu activation functions. We use a Bernoulli likelihood $p_\phi(x_n \mid z_n)$, similarly to Bayesian logistic matrix factorization, and thus the output layer of the neural network performs a sigmoid transformation.

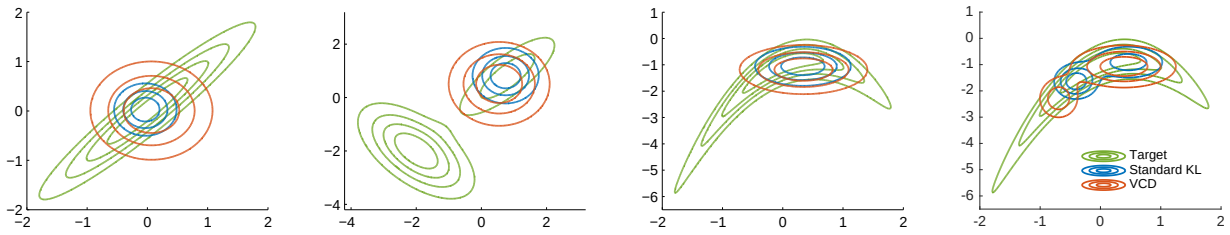
For both models, we consider a standard multivariate Gaussian prior $p(z_n) = \mathcal{N}(z_n \mid 0, I)$.

Methods. We compare three different objectives to perform the optimization: standard KL, the method of Hoffman (2017), and the divergence $\mathcal{L}_{\text{VCD}}(\theta)$. The standard KL objective involves an explicit variational distribution. Both the objective of Hoffman (2017) and the VCD divergence involve an improved HMC distribution $q_\theta^{(t)}(z_n \mid x_n)$.

We fit the model parameters ϕ via ML by maximizing the objective $\sum_n \mathbb{E}_{q_\theta^{(t)}(z_n \mid x_n)} [\log p_\phi(x_n \mid z_n)]$. The variational distribution $q_\theta^{(t)}(z_n \mid x_n) = q_\theta(z_n \mid x_n)$ for the standard KL method, as there is no improved distribution. For the other two methods, $q_\theta^{(t)}(z_n \mid x_n)$ is the distribution improved with HMC, and maximizing the objective with respect to ϕ corresponds essentially to a Monte Carlo expectation maximization algorithm (Wei & Tanner, 1990).

We fit the variational parameters θ by optimizing the corresponding divergence, according to the method. Both the standard KL and the method of Hoffman (2017) optimize the evidence lower bound (ELBO) in Eq. 1. Thus, the method of Hoffman (2017) does not incorporate feedback from the HMC chain into learning θ, as discussed in Section 3. Instead, the VCD method optimizes $\mathcal{L}_{\text{VCD}}(\theta)$ in Eq. 9.

Datasets. We use two datasets. The first one is the binarized MNIST data (Salakhutdinov & Murray, 2008), which contains 50,000 training images and 10,000 test images of hand-written digits. The second dataset is Fashion-MNIST



(a) (b) (c) (d)

Figure 1. Examples of fitting variational distributions to several targets (green contours) by applying the standard KL divergence (blue contours) and the VCD from Section 2.3 (red contours). In the first three panels (a)-(c) the variational distribution is a factorized Gaussian, while in the last panel (d) the variational distribution is a two-component Gaussian mixture. In all cases the VCD leads to distributions with higher variance, since $\mathcal{L}_{\text{VCD}}(\theta)$ converges asymptotically to the symmetrized KL divergence.

Table 2. Marginal log-likelihood on the test set. Fitting an improved distribution by minimizing the VCD leads to the best predictive performance.

method	average test log-likelihood MNIST	Fashion-MNIST
Standard KL	-111.20	-127.43
Hoffman (2017)	-103.61	-121.86
VCD (this paper)	$\mathbf{-101.26}$	$\mathbf{-121.11}$

(a) Bayesian logistic matrix factorization.

method	average test log-likelihood MNIST	Fashion-MNIST
Standard KL	-98.46	-124.63
Hoffman (2017)	-96.23	-117.74
VCD (this paper)	$\mathbf{-95.86}$	$\mathbf{-117.65}$

(b) Variational autoencoder.

(Xiao et al., 2017), which contains 60,000 training images and 10,000 test images of clothing items. We binarize the Fashion-MNIST images with a threshold at 0.5. Images in both datasets are of size 28×28 pixels.

Variational family. The variational family is a Gaussian, $q_\theta(z_n \mid x_n) = \mathcal{N}(z_n \mid \mu_\theta(x_n), \Sigma_\theta(x_n))$, whose mean and covariance are parameterized using two separate fully connected neural networks with two hidden layers of 200 units each. The neural networks have ReLu units, and the covariance $\Sigma_\theta(x_n)$ is set to be diagonal. The neural network for the covariance has non-linear activations in the output layer to ensure positive outputs. In particular, for each entry corresponding to the standard deviation, the neural network activation function is a modified softplus, $\text{softplus}(x) = \log(\exp\{10^{-4}\} + \exp\{x\})$. The modified softplus avoids numerical issues; it ensures that the variances are above a small threshold at 10^{-8}.

Experimental settings. We set the number of HMC iterations $t = 8$, using 5 leapfrog steps. We set the learning rate $\eta = 5 \times 10^{-4}$ for the variational parameters corresponding to the mean, $\eta = 2.5 \times 10^{-4}$ for the variational parameters corresponding to the covariance, and $\eta = 5 \times 10^{-4}$ for the model parameters ϕ. We additionally decrease the learning rate by a factor of 0.9 every 15,000 iterations. We set the dimensionality of z_n to 50 for Bayesian logistic matrix factorization, and to 10 for the VAE.

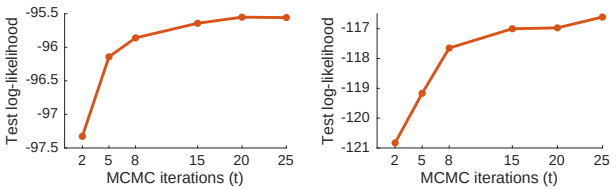
We run 400,000 iterations of each optimization algorithm. We perform stochastic VI by subsampling a minibatch of observations at each iteration (Hoffman et al., 2013); we set the minibatch size to 100.

For the VCD, the control variates are local, i.e., there is a C_n for each datapoint. However, in the earlier iterations of the optimization procedure we set the control variates to the same global value, $C_n = C$. The reason is that in these earlier iterations the model is highly non-stationary, as the model parameters ϕ change more significantly. We found that introducing local control variates at the beginning may lead to unstabilities; therefore we only introduce the local control variates C_n after 3,000 iterations. Before that, we update the global control variate C taking the mean of the stochastic estimates in the minibatch. After iteration 3,000, we let each C_n be updated independently.

Evaluation. We compute the average marginal test log-likelihood. For each test datapoint $x_n^\star$, we estimate the marginal log-likelihood using importance sampling,

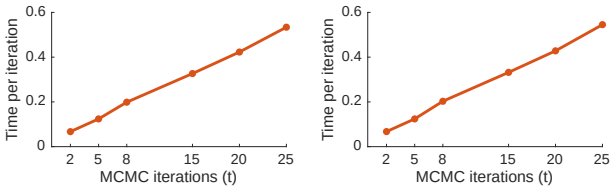
$$\log p_\phi(x_n^\star) \approx \log \frac{1}{S} \sum_{s=1}^{S} \frac{p_\phi(x_n^\star \mid z_n^{(s)}) p(z_n^{(s)})}{r_\theta(z_n^{(s)} \mid x_n^\star)}, \qquad (15)$$

where $z_n^{(s)} \sim r_\theta(z_n \mid x_n^\star)$. For each test instance $x_n^\star$, we use three different proposals $r_\theta(z_n \mid x_n^\star)$ and keep the highest resulting value (note that the approximation in Eq. 15 gives a lower bound of the marginal log-likelihood). The first proposal is an overdispersed version of the amortized variational distribution $q_\theta(z_n \mid x_n^\star)$. It is a Gaussian distribution whose mean is equal to the mean of $q_\theta(z_n \mid x_n^\star)$ but its standard deviation is 1.2 times larger. The second proposal is a Gaussian distribution whose mean is set to the mean of the samples resulting after running an HMC chain initialized at $q_\theta(z_n \mid x_n^\star)$ (we run the chain for 600 iterations and obtain



(a) MNIST. (b) Fashion-MNIST.

Figure 2. Estimates of the marginal log-likelihood on the test set for the VAE as a function of the number of MCMC steps t. Increasing the number of MCMC steps improves the performance.



(a) MNIST. (b) Fashion-MNIST.

Figure 3. Average time (in seconds) per iteration of the inference algorithm for the VAE as a function of the number of MCMC steps.

the mean of the last 300 samples). We set the standard deviation of the proposal 1.2 times larger than the standard deviation of $q_\theta(z_n \mid x_n^\star)$. The third proposal is similar to the second one, but we set its standard deviation 1.2 times larger than the standard deviation of the last 300 HMC samples. In all cases, we set $S = 20{,}000$ samples.

Results. Table 2 displays the average test log-likelihood. Optimizing the VCD leads to the best results for all methods and datasets. We can conclude that optimizing the VCD is in general the best approach, since it is at least as good as the state of the art. It outperforms standard amortized VI because it refines the variational distribution with HMC, and it outperforms the method of Hoffman (2017) because it uses feedback from the HMC samples when fitting the variational parameters θ.

Table 2 was obtained with $t = 8$ HMC steps. We now study the impact of t on the results. Figure 2 shows the test log-likelihood for the VAE as a function of the number of HMC steps, ranging from $t = 2$ to $t = 25$. As expected, increasing the number of steps improves the performance. Even for a small value $t = 2$, the test log-likelihood is better than for the standard KL method (see Table 2b).

The improvement comes at the cost of computational complexity. Figure 3 shows that the average time per iteration for fitting the VAE by minimizing the VCD increases linearly with the number of HMC steps. (No parallelism or GPU acceleration was used.) We also found that the optimization of the VCD is slightly faster than the method of Hoffman (2017) for all models and datasets. This is counter-intuitive, as the VCD requires a few additional computations, although their computational complexity is negligible compared to the HMC steps. Therefore, both methods should run roughly equally as fast. We believe that the differences we observed are implementation-specific.

To sum up, more computation leads to better results, but even a few HMC steps are advantageous compared to the minimization of the standard KL divergence.

5. Conclusion

We have proposed a method to improve the approximating distribution in variational inference (VI) by running a Markov chain Monte Carlo (MCMC) algorithm that targets the posterior of a probabilistic model. This leads to an implicit approximating distribution with an intractable density, which makes it challenging to minimize the Kullback-Leibler (KL) divergence between the approximation and the posterior. To address that, we have developed a divergence, called variational contrastive divergence (VCD), that can be tractably optimized with respect to the variational parameters; in particular, we can form unbiased Monte Carlo estimators of its gradient. The VCD differs from the standard KL in that it measures the discrepancy between the improved and the initial variational distributions. We have shown empirically that minimizing the VCD leads to better predictive performance in latent variable models.

One line for future research is to use the VCD to design tests for assessing the quality (or exactness) of a given variational distribution, similarly to the goal of Yao et al. (2018). For that, we can use a property of the VCD—that it allows us to obtain unbiased estimates of the actual value of the divergence (based on Eq. 10). Such tests can rely on the VCD being a proper divergence, as it takes non-negative values and it becomes zero only when the variational approximation matches the exact posterior.

Acknowledgements

Francisco J. R. Ruiz is supported by the EU Horizon 2020 programme (Marie Skłodowska-Curie Individual Fellowship, grant agreement 706760).

References

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

Carbonetto, P., King, M., and Hamze, F. A stochastic approximation method for inference in probabilistic graphical models. In *Advances in Neural Information Processing Systems*, 2009.

Cover, T. and Thomas, J. A. *Elements of Information Theory*. Wiley-Interscience, 2006.

Grover, A., Gummadi, R., Lázaro-Gredilla, M., Schuurmans, D., and Ermon, S. Variational rejection sampling. In *Artificial Intelligence and Statistics*, 2018.

Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, aug 2002.

Hoffman, M. D. Learning deep latent Gaussian models with Markov chain Monte Carlo. In *International Conference on Machine Learning*, 2017.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, May 2013.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Le, T. A., Igl, M., Rainforth, T., Jin, T., and Wood, F. Auto-encoding sequential Monte-Carlo. In *International Conference on Learning Representations*, 2018.

Li, Y., Turner, R. E., and Liu, Q. Approximate inference with amortised MCMC. In *arXiv:1702.08343*, 2017.

MacKay, D. J. C. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research, A*, 354(1):73–80, 1995.

Maddison, C. J., Lawson, D., Tucker, G., Heess, N., Norouzi, M., Mnih, A., Doucet, A., and Teh, Y. W. Filtering variational objectives. In *Advances in Neural Information Processing Systems*, 2017.

Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

Naesseth, C., Ruiz, F. J. R., Linderman, S., and Blei, D. M. Reparameterization gradients through acceptance-rejection methods. In *Artificial Intelligence and Statistics*, 2017.

Naesseth, C., Linderman, S. W., Ranganath, R., and Blei, D. M. Variational sequential Monte Carlo. In *Artificial Intelligence and Statistics*, 2018.

Neal, R. M. MCMC using Hamiltonian dynamics. In Brooks, S., Gelman, A., Jones, G. L., and Meng, X.-L. (eds.), *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.

Paisley, J. W., Blei, D. M., and Jordan, M. I. Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning*, 2012.

Ranganath, R., Gerrish, S., and Blei, D. M. Black box variational inference. In *Artificial Intelligence and Statistics*, 2014.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400–407, 1951.

Salakhutdinov, R. and Murray, I. On the quantitative analysis of deep belief networks. In *International Conference on Machine Learning*, 2008.

Salimans, T., Kingma, D. P., and Welling, M. Markov chain Monte Carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, 2015.

Tieleman, T. and Hinton, G. Lecture 6.5-RMSPROP: Divide the gradient by a running average of its recent magnitude. Coursera: Neural Networks for Machine Learning, 4, 2012.

Titsias, M. K. Learning model reparametrizations: Implicit variational inference by fitting MCMC distributions. In *arXiv:1708.01529*, 2017.

Titsias, M. K. and Lázaro-Gredilla, M. Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*, 2014.

Wei, G. C. G. and Tanner, M. A. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704, 1990.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. In *arXiv:1708.07747*, 2017.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. Yes, but did it work?: Evaluating variational inference. In *International Conference on Machine Learning*, 2018.

Zhang, Y., Hernández-Lobato, J. M., and Ghahramani, Z. Ergodic measure preserving flows. In *arXiv:1805.10377*, 2018.