Supplement: A Contrastive Divergence for Combining Variational Inference and MCMC

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

1. Simplification of the VCD

Here we show how to express the simplified expression of the variational contrastive divergence (VCD) as the difference of two expectations. We start from the definition in terms of the Kullback-Leibler (KL) divergences,

$$\begin{aligned}&\mathcal{L}_{\text{VCD}}(\theta)\\&= \mathcal{L}_{\text{diff}}(\theta) + \text{KL}(q_\theta^{(t)}(z) \,||\, q_\theta(z))\\&= \text{KL}(q_\theta(z) \,||\, p(z \,|\, x)) - \text{KL}(q_\theta^{(t)}(z) \,||\, p(z \,|\, x))\\&\quad + \text{KL}(q_\theta^{(t)}(z) \,||\, q_\theta(z)).\end{aligned} \tag{1}$$

We now apply the definition of the KL divergence,

$$\begin{aligned}&\mathcal{L}_{\text{VCD}}(\theta)\\&= \mathbb{E}_{q_\theta(z)}\left[\log \frac{q_\theta(z)}{p(z \,|\, x)}\right] - \mathbb{E}_{q_\theta^{(t)}(z)}\left[\log \frac{q_\theta^{(t)}(z)}{p(z \,|\, x)}\right]\\&\quad + \mathbb{E}_{q_\theta^{(t)}(z)}\left[\log \frac{q_\theta^{(t)}(z)}{q_\theta(z)}\right].\end{aligned} \tag{2}$$

Next we expand the logarithms. The expectation of the log-density $\log q_\theta^{(t)}(z)$ cancels out because it appears with different signs in the second and third terms. We also rewrite the posterior $p(z \,|\, x) = p(x, z)/p(x)$,

$$\begin{aligned}&\mathcal{L}_{\text{VCD}}(\theta)\\&= \mathbb{E}_{q_\theta(z)}\left[\log \frac{q_\theta(z)p(x)}{p(x, z)}\right] - \mathbb{E}_{q_\theta^{(t)}(z)}\left[\log \frac{q_\theta(z)p(x)}{p(x, z)}\right].\end{aligned} \tag{3}$$

The marginal log-likelihood $\log p(x)$ does not depend on z and can be taken out of the expectation. Since it appears with different signs, it cancels out. We now recognize that the argument of each expectation—after having canceled out the marginal log-likelihood—is the negative instantaneous evidence lower bound (ELBO), defined as

$$f_\theta(z) \triangleq \log p(x, z) - \log q_\theta(z). \tag{4}$$

Thus, we finally have

$$\mathcal{L}_{\text{VCD}}(\theta) = -\mathbb{E}_{q_\theta(z)}\left[f_\theta(z)\right] + \mathbb{E}_{q_\theta^{(t)}(z)}\left[f_\theta(z)\right]. \tag{5}$$

2. Generalization of the VCD

The VCD divergence can be generalized with a parameter α that downweights the two KL terms involving the improved distribution $q_\theta^{(t)}(z)$. More in detail, we define the α-generalized VCD as

$$\begin{aligned}&\mathcal{L}_{\text{VCD}}^{(\alpha)}(\theta) = \text{KL}(q_\theta(z) \,||\, p(z \,|\, x))\\&+\alpha\left[\text{KL}(q_\theta^{(t)}(z) \,||\, q_\theta(z)) - \text{KL}(q_\theta^{(t)}(z) \,||\, p(z \,|\, x))\right].\end{aligned} \tag{6}$$

For any $0 \leq \alpha \leq 1$, the α-generalized VCD is also a proper divergence because it satisfies the two desired criteria—it is non-negative and it becomes zero only when $q_\theta = p(z \,|\, x)$. Moreover, it leads to tractable optimization because the intractable log-density $\log q_\theta^{(t)}(z)$ also cancels out in this expression.

By varying α, the α-generalized VCD interpolates between the standard KL divergence of variational inference (VI) (for $\alpha = 0$) and the VCD in Eq. 1 (for $\alpha = 1$). When the number of Markov chain Monte Carlo (MCMC) steps is large, this is effectively an interpolation between the standard KL and the symmetrized KL divergence.

The α-generalized VCD is useful when the MCMC method does not mix well. To see this, consider that due to slow mixing, the improved distribution $q_\theta^{(t)}(z) \approx q_\theta(z)$. In this case, the divergence $\mathcal{L}_{\text{VCD}}(\theta) \approx 0$ for any value of the variational parameters, but the α-generalized VCD becomes proportional to the standard KL, $\mathcal{L}_{\text{VCD}}^{(\alpha)}(\theta) \approx (1 - \alpha)\text{KL}(q_\theta(z) \,||\, p(z \,|\, x))$. Therefore, the α-generalized VCD may lead to more robust optimization when the MCMC method does not mix well.

In our experiments, we consider the non-generalized VCD in Eq. 1 because we did not find any mixing issues with our MCMC method. The derivations of the gradients in the main paper can be straightforwardly generalized for the case where the objective is $\mathcal{L}_{\text{VCD}}^{(\alpha)}(\theta)$.

[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.

3. Particularizations of the Gradients

Here we derive the gradients of the VCD for two choices of the variational distribution $q_\theta(z)$, namely, a Gaussian and a mixture of Gaussians.

3.1. Gaussian Variational Distribution

We now show how to obtain the gradient of the VCD in the case where the distribution $q_\theta(z)$ is Gaussian.

Consider $q_\theta(z) = \mathcal{N}(z \,|\, \mu, \Sigma)$, i.e., a Gaussian distribution with mean μ and covariance Σ. That is,

$$\begin{aligned}\log q_\theta(z) = &-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| \\ &-\frac{1}{2}(z-\mu)^\top \Sigma^{-1}(z-\mu)\end{aligned} \tag{7}$$

Here, $\theta = [\mu, \Sigma]$ denotes the variational parameters, and D is the dimensionality of the latent variable, $z \in \mathbb{R}^D$.

The definition of the VCD is in Eq. 5. Since it consists of a difference of two expectations of the same function $f_\theta(z)$, the terms that are constant with respect to z in $f_\theta(z)$ cancel out. Specifically, the term $-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma|$ from Eq. 7, which appears in $f_\theta(z)$ (Eq. 4), is constant with respect to z; therefore it cancels out. This leads to the simplified objective

$$\mathcal{L}_{\text{VCD}}(\theta) = -\mathbb{E}_{q_\theta(z)}\left[g_\theta(z)\right] + \mathbb{E}_{q_\theta^{(t)}(z)}\left[g_\theta(z)\right], \tag{8}$$

where we have introduced the shorthand notation $g_\theta(z)$ for the (simplified) argument of the expectation,

$$g_\theta(z) \triangleq \log p(x,z) + \frac{1}{2}(z-\mu)^\top \Sigma^{-1}(z-\mu). \tag{9}$$

Eq. 8 can be further simplified by computing the exact expectation of the quadratic form that appears in the first expectation,

$$\mathcal{L}_{\text{VCD}}(\theta) = -\mathbb{E}_{q_\theta(z)}\left[\log p(x,z)\right] - \frac{D}{2} + \mathbb{E}_{q_\theta^{(t)}(z)}\left[g_\theta(z)\right]. \tag{10}$$

These expressions are also valid when using amortized inference with a Gaussian variational distribution. That is, when μ and Σ are functions of the data x, $\mu = \mu_\theta(x)$ and $\Sigma = \Sigma_\theta(x)$.

Taking the gradients. We now derive the expressions for the gradient of the VCD with respect to the variational parameters. We parameterize the Gaussian in terms of its mean and the Cholesky decomposition of its covariance. That is, the variational parameters are μ and L, where L is a lower triangular matrix such that $LL^\top = \Sigma$. The reparameterization transformation in terms of a standard Gaussian $q(\varepsilon) = \mathcal{N}(\varepsilon \,|\, 0, I)$ is given as $\varepsilon \sim q(\varepsilon)$, $z = h_\theta(\varepsilon) = \mu + L\varepsilon$.

The gradient of the (negative) first term in Eq. 10 is directly given by the reparameterization gradient,

$$\begin{aligned}&\nabla_\theta \mathbb{E}_{q_\theta(z)}\left[\log p(x,z)\right] \\ &= \mathbb{E}_{q(\varepsilon)}\left[\nabla_z \log p(x,z)\big|_{z=h_\theta(\varepsilon)} \times \nabla_\theta h_\theta(\varepsilon)\right],\end{aligned} \tag{11}$$

where $\theta = [\mu, L]$ denotes the variational parameters.

For the second expectation in Eq. 10, we apply the derivation in the main paper,

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

Note that the gradient $\nabla_\theta g_\theta(z)$ only involves the gradient of the quadratic form, since the model $p(x,z)$ does not depend on θ. The gradient $\nabla_\theta \log q_\theta(z_0)$ is the gradient of the Gaussian log-density,

$$\begin{aligned}&\nabla_\mu \log q_\theta(z_0) = L^{-\top}L^{-1}(z_0 - \mu), \\ &\nabla_L \log q_\theta(z_0) \\ &= -\Omega + \left(L^{-\top}L^{-1}(z_0-\mu)(z_0-\mu)^\top L^{-\top}\right) \odot M,\end{aligned}$$

where Ω is a diagonal matrix whose entries are given by the element-wise inverse of the diagonal entries of L, the symbol $\odot$ denotes the element-wise product, and M is a lower triangular masking matrix of ones (it contains zeros above the main diagonal).

The above expressions are also valid when the variational distribution is a fully factorized Gaussian, in which case L is a diagonal matrix whose entries correspond to the standard deviation of each component. This is the setting that we consider in the paper.

3.2. Mixture of Gaussians

Consider now that the variational distribution $q_\theta(z)$ is a mixture of K components,

$$q_\theta(z) = \sum_{k=1}^{K} w_k q_{\theta_k}(z). \tag{13}$$

where each $q_{\theta_k}(z)$ is a Gaussian,

$$q_{\theta_k}(z) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_k|}} e^{-\frac{1}{2}(z-\mu_k)^\top \Sigma_k^{-1}(z-\mu_k)}. \tag{14}$$

The variational parameters are $\theta_k = [\mu_k, L_k]$ for each component, where L_k is the Cholesky decomposition of Σ_k, i.e., $L_k L_k^\top = \Sigma_k$, as well as the mixture weights w_k.

The VCD objective is given in Eq. 5; however in this case there are no constant terms in $f_\theta(z)$ that cancel out as in the Gaussian case. Thus, we obtain the gradient of

the VCD by computing the gradients $\nabla_\theta \mathbb{E}_{q_\theta(z)}[f_\theta(z)]$ and $\nabla_\theta \mathbb{E}_{q_\theta^{(t)}(z)}[f_\theta(z)]$ separately.

Taking the gradient of the first term. We first rewrite the standard ELBO term as a sum over the mixture components,

$$\mathbb{E}_{q_\theta(z)}[f_\theta(z)] = \sum_{k=1}^{K} w_k \mathbb{E}_{q_{\theta_k}(z)}[f_\theta(z)]. \tag{15}$$

We next reparameterize each component, with $\varepsilon \sim q(\varepsilon) = \mathcal{N}(\varepsilon\,;0,I)$ and $z = h_{\theta_k}(\varepsilon) = \mu_k + L_k\varepsilon$. We rewrite the ELBO in terms of this reparameterization,

$$\mathbb{E}_{q_\theta(z)}[f_\theta(z)] = \sum_{k=1}^{K} w_k \mathbb{E}_{q(\varepsilon)}\left[f_\theta(z)\big|_{z=h_{\theta_k}(\varepsilon)}\right]. \tag{16}$$

We now take the gradient of the ELBO. The gradient with respect to each component θ_k is

$$\begin{aligned}
&\nabla_{\theta_k}\mathbb{E}_{q_\theta(z)}[f_\theta(z)] \\
&= \sum_{k'=1}^{K} w_{k'}\mathbb{E}_{q(\varepsilon)}\left[\nabla_z f_\theta(z)\big|_{z=h_{\theta_{k'}}(\varepsilon)}\nabla_{\theta_k}h_{\theta_{k'}}(\varepsilon)\right] \\
&\quad + \sum_{k'=1}^{K} w_{k'}\mathbb{E}_{q_{\theta_{k'}}(z)}[\nabla_{\theta_k} f_\theta(z)] \\
&= w_k\mathbb{E}_{q(\varepsilon)}\left[\nabla_z f_\theta(z)\big|_{z=h_{\theta_k}(\varepsilon)}\nabla_{\theta_k}h_{\theta_k}(\varepsilon)\right] \\
&\quad + \mathbb{E}_{q_\theta(z)}[-\nabla_{\theta_k}\log q_\theta(z)] \\
&= w_k\mathbb{E}_{q(\varepsilon)}\left[\nabla_z f_\theta(z)\big|_{z=h_{\theta_k}(\varepsilon)}\nabla_{\theta_k}h_{\theta_k}(\varepsilon)\right].
\end{aligned} \tag{17}$$

We now obtain the gradient of the ELBO w.r.t. the mixture weights. We build a score function estimator,

$$\begin{aligned}
&\nabla_{w_k}\mathbb{E}_{q_\theta(z)}[f_\theta(z)] \\
&= \mathbb{E}_{q_\theta(z)}[f_\theta(z)\nabla_{w_k}\log q_\theta(z)] \\
&= \mathbb{E}_{q_\theta(z)}\left[f_\theta(z)\frac{1}{q_\theta(z)}q_{\theta_k}(z)\right] \\
&= \mathbb{E}_{q_{\theta_k}(z)}[f_\theta(z)].
\end{aligned} \tag{18}$$

To sum up, we have obtained a reparameterization gradient for the parameters θ_k (Eq. 17) and a score function gradient for the mixture weights w_k (Eq. 18).

For the reparameterization gradient, one of the quantities that we need to evaluate is the gradient $\nabla_z \log q_\theta(z)$. We compute this gradient in a numerically stable manner using the log-derivative trick,

$$\begin{aligned}
\nabla_z \log q_\theta(z) &= \frac{1}{q_\theta(z)}\sum_{k=1}^{K} w_k \nabla_z q_{\theta_k}(z) \\
&= \frac{1}{q_\theta(z)}\sum_{k=1}^{K} w_k q_{\theta_k}(z)\nabla_z \log q_{\theta_k}(z) \\
&= \sum_{k=1}^{K} q_\theta(k\mid z)\nabla_z \log q_{\theta_k}(z),
\end{aligned} \tag{19}$$

where $q_\theta(k\mid z)$ is the "posterior probability" (under the variational model) of component k given z, i.e.,

$$q_\theta(k\mid z) \propto \exp\{\log w_k + \log q_{\theta_k}(z)\}. \tag{20}$$

Taking the gradient of the second term. The second term in th VCD of Eq. 5 involves an expectation with respect to the improved distribution $q_\theta^{(t)}(z)$. As derived in the main paper, the gradient of the second term can in turn be split into two terms. We first obtain the gradient w.r.t. the parameters θ_k of each Gaussian component,

$$\begin{aligned}
&\nabla_{\theta_k}\mathbb{E}_{q_\theta^{(t)}(z)}[f_\theta(z)] = \mathbb{E}_{q_\theta^{(t)}(z)}[-\nabla_{\theta_k}\log q_\theta(z)] \\
&+ \mathbb{E}_{q_\theta(z_0)}[w_\theta(z_0)\nabla_{\theta_k}\log q_\theta(z_0)],
\end{aligned} \tag{21}$$

where we have defined

$$w_\theta(z_0) \triangleq \mathbb{E}_{Q^{(t)}(z\mid z_0)}[f_\theta(z)]. \tag{22}$$

Now we make use of the definition of $q_\theta(z)$ and substitute Eq. 13 in the two expressions above, yielding

$$\begin{aligned}
&\nabla_{\theta_k}\mathbb{E}_{q_\theta^{(t)}(z)}[f_\theta(z)] \\
&= \sum_{k=1}^{K} w_k \mathbb{E}_{q_{\theta_k}(z_0)}\left[\mathbb{E}_{Q^{(t)}(z\mid z_0)}[-\nabla_{\theta_k}\log q_\theta(z)]\right] \\
&\quad + \sum_{k=1}^{K} w_k \mathbb{E}_{q_{\theta_k}(z_0)}[w_\theta(z_0)\nabla_{\theta_k}\log q_\theta(z_0)].
\end{aligned} \tag{23}$$

We can find an alternative expression for the latter term. Starting with the latter term in Eq. 23, we first use the definition in Eq. 13 and then apply the log-derivative trick, yielding the following expression,

$$\begin{aligned}
&\sum_{k=1}^{K} w_k \mathbb{E}_{q_{\theta_k}(z_0)}[w_\theta(z_0)\nabla_{\theta_k}\log q_\theta(z_0)] \\
&= \mathbb{E}_{q_\theta(z_0)}[w_\theta(z_0)\nabla_{\theta_k}\log q_\theta(z_0)] \\
&= \mathbb{E}_{q_\theta(z_0)}\left[w_\theta(z_0)\frac{1}{q_\theta(z_0)}w_k q_{\theta_k}(z_0)\nabla_{\theta_k}\log q_{\theta_k}(z_0)\right] \\
&= w_k \mathbb{E}_{q_{\theta_k}(z_0)}[w_\theta(z_0)\nabla_{\theta_k}\log q_{\theta_k}(z_0)].
\end{aligned} \tag{24}$$

Finally, we obtain the gradient $\nabla_{w_k}\mathbb{E}_{q_\theta^{(t)}(z)}\left[f_\theta(z)\right]$, taken w.r.t. the mixture weights. We apply the expression derived in the main paper,

$$\begin{aligned}\nabla_{w_k}\mathbb{E}_{q_\theta^{(t)}(z)}\left[f_\theta(z)\right] &= \mathbb{E}_{q_\theta^{(t)}(z)}\left[-\nabla_{w_k}\log q_\theta(z)\right] \\ &+ \mathbb{E}_{q_\theta(z_0)}\left[w_\theta(z_0)\nabla_{w_k}\log q_\theta(z_0)\right].\end{aligned} \tag{25}$$

We rewrite the first term in Eq. 25 by taking the exact expectation with respect to the mixture indicator,

$$\begin{aligned}&\mathbb{E}_{q_\theta^{(t)}(z)}\left[-\nabla_{w_k}\log q_\theta(z)\right] \\ &= \sum_{k'=1}^{K} w_{k'}\mathbb{E}_{q_{\theta_{k'}}(z_0)}\left[\mathbb{E}_{Q^{(t)}(z\,|\,z_0)}\left[-\frac{q_{\theta_k}(z)}{q_\theta(z)}\right]\right].\end{aligned} \tag{26}$$

We rewrite the second term in Eq. 25 using the log-derivative trick,

$$\begin{aligned}&\mathbb{E}_{q_\theta(z_0)}\left[w_\theta(z_0)\nabla_{w_k}\log q_\theta(z_0)\right] \\ &= \mathbb{E}_{q_\theta(z_0)}\left[w_\theta(z_0)\frac{1}{q_\theta(z_0)}q_{\theta_k}(z_0)\nabla_{w_k}\log q_{\theta_k}(z_0)\right] \\ &= \mathbb{E}_{q_{\theta_k}(z_0)}\left[w_\theta(z_0)\nabla_{w_k}\log q_{\theta_k}(z_0)\right].\end{aligned} \tag{27}$$