Some recent advances in Variational Inference

François BERTHOLOM[1] François ROUEFF[2]

[1]SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, France

[2]LTCI, Télécom Paris, Institut Polytechnique de Paris, France

Résumé – Les développements récents en inférence variationnelle visent à dépasser les limites des approches traditionnelles en adoptant des techniques telles que l'échantillonnage préférentiel et l'utilisation de nouvelles divergences, en lien avec des stratégies de reparamétrisation. Cet article propose une présentation générale de l'inférence variationnelle illustrée par un exemple numérique dans le cadre d'un modèle classique de traitement du signal, puis fait le point sur certaines innovations récentes telles que l'inférence variationnelle par poids d'importance et l'inférence variationnelle avec les divergences alpha.

Abstract – Recent advances in Variational Inference (VI) have aimed to overcome the shortcomings of traditional methods by adopting refined techniques such as importance weighting and alternative divergence metrics, leveraging reparameterization strategies. In this paper, we present a comprehensive introduction to VI, illustrated through a numerical experiment based on a simple and classical signal processing model. We discuss significant innovations in the field, particularly highlighting Importance Weighted VI and alpha-divergence VI, while outlining both their theoretical foundations and practical benefits.

1 Introduction

Variational Inference (VI) has emerged as an essential method for approximate Bayesian inference [2], enabling scalable probabilistic modeling in scenarios where exact inference is computationally infeasible. Following, e.g., the seminal work of [11], VI has also become instrumental in AI research. At its core, VI approximates complex posterior distributions using a family of simpler, tractable densities. A primary drawback of traditional VI is the limited expressiveness of the variational family, which can lead to inaccurate posterior approximations. Recent advances, including Importance Weighted VI (IWVI) and alpha-divergence VI, have focused on improving the flexibility of VI methods. IWVI tightens the variational bound using multiple importance-weighted samples, while alpha-divergence VI modifies the learning objective to allow better posterior coverage. This paper begins with an accessible introduction to VI, illustrated through an example relevant to the signal processing community. We then examine recent advances, highlighting both theoretical perspectives and practical implications.

2 Variational Inference

General principles. Consider a probabilistic latent variable model in which the data $\mathbf{y} \in \mathsf{Y}$ is generated by a hidden variable $\mathbf{x} \in \mathsf{X}$. We assume a joint density $p_\theta(\mathbf{y}, \mathbf{x}) = p_\theta(\mathbf{y} \mid \mathbf{x}) p_\theta(\mathbf{x})$ parameterized by $\theta \in \mathsf{T}$, where $p_\theta(\mathbf{x})$ is the prior density of the latent variable and $p_\theta(\mathbf{y} \mid \mathbf{x})$ denotes the conditional density function of $\mathbf{y}$ given $\mathbf{x}$. Typically, two central problems arise in this context: computing the likelihood $p_\theta(\mathbf{y})$ and obtaining moments or samples from the posterior $p_\theta(\mathbf{x} \mid \mathbf{y})$. In most interesting models, both tasks are intractable. Variational Inference (VI) sidesteps this computational difficulty by approximating the true posterior $p_\theta(\mathbf{x} \mid \mathbf{y})$ with a density $q^*(\mathbf{x})$ selected from a family of simpler, tractable densities $\mathcal{Q} = \{q_\varphi \ : \ \varphi \in \mathsf{H}\}$. The quality of the resulting approximation is quantified by a chosen criterion, usually the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(q_\varphi(\cdot) \,\|\, p_\theta(\cdot \mid \mathbf{y}))$. In this framework, VI can be framed as the maximization of the Evidence Lower Bound (ELBO), defined as

$$\mathrm{ELBO}(\theta, \varphi, \mathbf{y}) = \int_{\mathsf{X}} \ln\left(\frac{p_\theta(\mathbf{y}, \mathbf{x})}{q_\varphi(\mathbf{x})}\right) q_\varphi(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \quad (1)$$

By writing $p_\theta(\mathbf{y}, \mathbf{x}) = p_\theta(\mathbf{x}|\mathbf{y}) p_\theta(\mathbf{y})$ and using the properties of the logarithm, we get

$$\mathrm{ELBO}(\theta, \varphi, \mathbf{y}) = \ln p_\theta(\mathbf{y}) - D_{\mathrm{KL}}(q_\varphi(\cdot) \,\|\, p_\theta(\cdot \mid \mathbf{y})). \quad (2)$$

In particular, $\mathrm{ELBO}(\theta, \varphi, \mathbf{y}) \leq \ln p_\theta(\mathbf{y})$ and minimizing $D_{\mathrm{KL}}(q_\varphi(\cdot) \,\|\, p_\theta(\cdot \mid \mathbf{y}))$ with respect to φ aligns the ELBO with $\ln p_\theta(\mathbf{y})$. Thus, jointly maximizing the ELBO for θ and φ approximates the Maximum Likelihood Estimator (MLE).

Variational Expectation-Maximization algorithm. Alternatively maximizing the ELBO with respect to φ and θ is known as the Variational Expectation-Maximization (VEM) algorithm. This method initializes the parameter θ at some θ_0 and alternates updates of φ and θ through an ascent procedure:

1. $\varphi_{k+1} = \arg\max_{\varphi \in \mathsf{H}} \mathrm{ELBO}(\theta_k, \varphi, \mathbf{y})$,
2. $\theta_{k+1} = \arg\max_{\theta \in \mathsf{T}} \mathrm{ELBO}(\theta, \varphi_{k+1}, \mathbf{y})$.

Variational EM can be viewed as a generalization of the classical Expectation-Maximization (EM) algorithm. When the variational family $\mathcal{Q}$ is sufficiently expressive to include the posterior $p_\theta(\cdot \mid \mathbf{y})$ for all $\theta \in \Theta$, these two steps simplify to:

1. $q_{\varphi_k}(\cdot) = p_{\theta_k}(\cdot \mid \mathbf{y})$,
2. $\theta_{k+1} = \arg\max_{\theta \in \mathsf{T}} \mathbb{E}_{q_{\varphi_k}}[\ln p_\theta(\mathbf{x}, \mathbf{y}) \mid \mathbf{y}]$.

EM involves the density $p_{\theta_k}(\cdot \mid \mathbf{y})$, which is often intractable. VEM addresses this by using an approximating family. A key distinction is that EM increases the likelihood at each step, whereas VEM optimizes a lower bound (see (2)).

3 A signal processing example

We now provide an illustrative example where both the EM and the variational EM based on the so-called *mean-field* approximation have closed forms.

Setting. Consider the AR(1) model observed with additive noise. Denoting real-valued observations by $\mathbf{y} = Y_{1:T} = (Y_t)_{1\le t\le T}$ and hidden variables by $\mathbf{x} = X_{0:T} = (X_t)_{0\le t\le T}$, the model is defined as

$$\begin{aligned} X_{t+1} &= \psi X_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_h^2), \\ Y_t &= \zeta X_t + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, \sigma_v^2), \end{aligned} \tag{3}$$

where $\theta = (\psi, \zeta, \sigma_h^2, \sigma_v^2)$ denotes the unknown parameters. For simplicity, we assume that X_0 is known (or it can be set to zero). The joint density is thus given by

$$p_\theta(\mathbf{x}, \mathbf{y}) = \prod_{t=1}^{T} p_\theta(Y_t \mid X_t) p_\theta(X_t \mid X_{t-1}),$$

where $p_\theta(Y_t \mid X_t) = \mathcal{N}(Y_t \mid \zeta X_t, \sigma_v^2)$ and $p_\theta(X_t \mid X_{t-1}) = \mathcal{N}(X_t \mid \psi X_{t-1}, \sigma_h^2)$.

Expectation-maximization. The EM algorithm amounts to iteratively maximizing the integral function

$$\theta' \mapsto Q_\theta(\theta') = \int_{\mathbb{R}^T} \ln(p_{\theta'}(\mathbf{x}, \mathbf{y}))\, p_\theta(\mathbf{x} \mid \mathbf{y})\, \mathrm{d}\mathbf{x}.$$

It is tractable in this model, since Q_θ only involves $\mathbb{E}_\theta[X_t \mid Y_{1:T}]$, $\mathbb{E}_\theta[X_t^2 \mid Y_{1:T}]$ and $\mathbb{E}_\theta[X_t X_{t-1} \mid Y_{1:T}]$, all of which can be computed using the Kalman filter and smoother. Optimizing $Q_\theta(\theta')$ with respect to θ' yields the EM iterative update $\theta^\star = \big(\psi^\star, \zeta^\star, (\sigma_h^2)^\star, (\sigma_v^2)^\star\big)$ for this AR(1) model:

$$\begin{aligned} \psi^\star &= \frac{\sum_{t=1}^T \mathbb{E}_\theta[X_t X_{t-1} \mid Y_{1:T}]}{\sum_{t=1}^T \mathbb{E}_\theta[X_{t-1}^2 \mid Y_{1:T}]}, \\ \zeta^\star &= \frac{\sum_{t=1}^T Y_t\, \mathbb{E}_\theta[X_t \mid Y_{1:T}]}{\sum_{t=1}^T \mathbb{E}_\theta[X_t^2 \mid Y_{1:T}]}, \\ (\sigma_h^2)^\star &= \frac{1}{T}\sum_{t=1}^T \mathbb{E}_\theta[(X_t - \psi^\star X_{t-1})^2 \mid Y_{1:T}], \\ (\sigma_v^2)^\star &= \frac{1}{T}\sum_{t=1}^T \mathbb{E}_\theta[(Y_t - \zeta^\star X_t)^2 \mid Y_{1:T}]. \end{aligned}$$

Variational EM. To approximate the posterior density $p_\theta(X_{1:T} \mid Y_{1:T})$ in the traditional VI framework, we choose the mean-field variational family:

$$q_{\boldsymbol{\varphi}}(X_{1:T}) = \prod_{t=1}^{T} q_{\varphi_t}(X_t),$$

where we denote $\boldsymbol{\varphi} = (\varphi_t)_{1\le t\le T}$ with $\varphi_t = (\mu_t, \mathrm{v}_t)$ and $q_{\varphi_t}(X) = \mathcal{N}(X \mid \mu_t, \mathrm{v}_t)$. By convention, we set $\mu_0 = X_0$, $\mathrm{v}_0^2 = 0$. In this setup, $\mathrm{ELBO}(\theta, \boldsymbol{\varphi}, X_{1:T})$ (see (1)) admits a simple closed form and the variational means $(\mu_t^\star)_{1\le t\le T}$ that maximize it satisfy the tridiagonal system of equations

$$\begin{aligned} \mathrm{a}_\theta(\mu_{t-1}^\star + \mu_{t+1}^\star) + \mathrm{b}_\theta \mu_t^\star &= \mathrm{c}_\theta Y_t, \qquad t \in [\![1, T-1]\!], \\ \mathrm{a}_\theta \mu_{T-1}^\star + \mathrm{b}'_\theta \mu_T^\star &= \mathrm{c}_\theta Y_T, \end{aligned}$$

where $\mathrm{a}_\theta = -\sigma_v^2\psi$, $\mathrm{b}_\theta = \sigma_h^2\zeta + \sigma_v^2(1+\psi^2)$, $\mathrm{b}'_\theta = \sigma_h^2\zeta + \sigma_v^2$ and $\mathrm{c}_\theta = \sigma_h^2\zeta$. This system of equations can be solved using the Thomas algorithm, a special case of Gaussian elimination. Regarding the variances, we have

$$\mathrm{v}_T^\star = \frac{\sigma_v^2\sigma_h^2}{\mathrm{b}'_\theta}, \qquad \mathrm{v}_t^\star = \frac{\sigma_v^2\sigma_h^2}{\mathrm{b}_\theta}, \qquad t \in [\![1, T-1]\!].$$

The optimization of $\mathrm{ELBO}(\theta, \boldsymbol{\varphi}, X_{1:T})$ with respect to θ also has an explicit solution $\theta^\star = \big(\psi^\star, \zeta^\star, (\sigma_h^2)^\star, (\sigma_v^2)^\star\big)$:

$$\begin{aligned} \psi^\star &= \frac{\sum_{t=1}^T \mu_t \mu_{t-1}}{\sum_{t=1}^T (\mu_{t-1}^2 + \sigma_{t-1}^2)}, \quad \zeta^\star = \frac{\sum_{t=1}^T Y_t \mu_t}{\sum_{t=1}^T (\mu_t^2 + \sigma_t^2)}, \\ (\sigma_h^2)^\star &= \frac{1}{T}\sum_{t=1}^T \Big[(\mu_t - \psi^\star \mu_{t-1})^2 + (\psi^\star)^2 \sigma_{t-1}^2 + \sigma_t^2\Big], \\ (\sigma_v^2)^\star &= \frac{1}{T}\sum_{t=1}^T \Big[(Y_t - \zeta^\star \mu_t)^2 + (\zeta^\star)^2 \sigma_t^2\Big]. \end{aligned}$$

Experiment. We generate synthetic data of length $T = 500$ from the model (3), with $X_0 = 1.2$, $\psi = 0.88$, $\zeta = 1.23$, $\sigma_h^2 = 1.34$, and $\sigma_v^2 = 0.95$. We use EM (with the Kalman filter) and traditional VI (VEM) to infer the parameters $\theta = (\psi, \zeta, \sigma_h^2, \sigma_v^2)$. Initial values for ψ and ζ are drawn from a standard Gaussian, while σ_h^2 and σ_v^2 are drawn uniformly in $[0.1, 10]$. We give each method a time budget and a fixed number of iterations. Within the allotted time, we repeatedly run each method from random initializations, stopping after the predetermined number of iterations and retaining the last parameter estimates. Among the available repeated runs, the one with the best likelihood is our final estimator. This procedure is repeated 20 times to visualize the distribution of the obtained estimators through boxplots in Figure 1. VEM takes advantage of more available runs up to 8 iterations. EM eventually achieves higher accuracy than VEM, but the gain mainly leads to overfitting. The difference between the ELBO and the true likelihood is called the variational gap, a smaller gap is synonym of a tighter and more reliable approximation. Here, the gap remains significant, meaning that the mean-field approximation fails to converge to the actual posterior. Notice that VEM runs almost three times as fast as EM (both scale as $\mathcal{O}(T)$) with similar memory requirements, making it significantly more scalable while maintaining comparable estimation quality on this example. In more intricate models, the E-step of the EM can be intractable. A proper choice of variational family then allows to perform gradient-based optimization of the ELBO with respect to both the model parameter θ and variational parameter φ using the reparameterization trick [11]. This makes VI a particularly efficient and scalable inference technique, even for the very high-dimensional models used in current AI applications.

4 Importance Weighted VI

Monte-Carlo objectives as variational bounds. Let the sequence $(\widehat{\mathcal{P}}(\theta, \mathbf{y}, N))_{N\ge1}$ be an unbiased estimator of the normalizing constant $p_\theta(\mathbf{y})$, i.e., $\mathbb{E}[\widehat{\mathcal{P}}(\theta, \mathbf{y}, N)] = p_\theta(\mathbf{y})$. Define the associated Monte-Carlo objective (MCO) [15] by

$$\mathcal{L}_{\widehat{\mathcal{P}}}(\theta, \mathbf{y}, N) = \mathbb{E}[\ln \widehat{\mathcal{P}}(\theta, \mathbf{y}, N)].$$

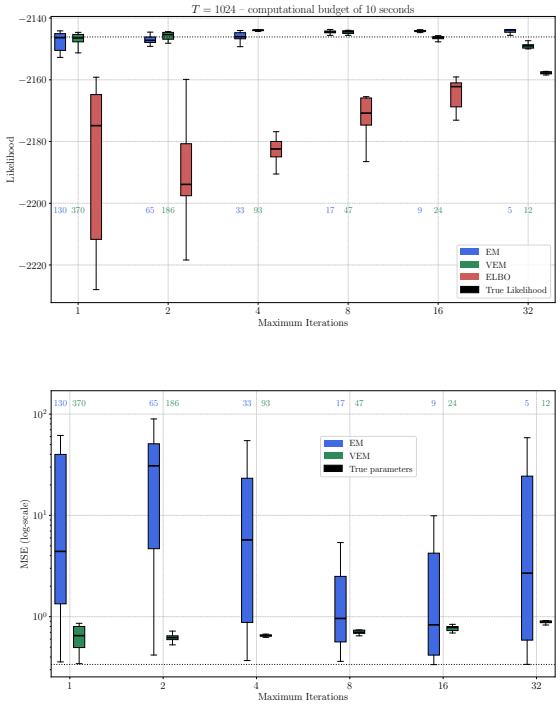

Figure 1 – Synthetic data with $T = 1024$, runtime budget of 10 sec to perform a fixed number of iterations as many times as possible. Boxplots show the best likelihood across 20 experiments. The colored numbers below the boxplots correspond to the average number of runs in each experiment.

By Jensen's inequality, we have $\mathcal{L}_{\widehat{\mathcal{P}}}(\theta, \mathbf{y}, N) \leq \ln p_\theta(\mathbf{y})$. When the estimator $\widehat{\mathcal{P}}(\theta, \mathbf{y}, N)$ is consistent, the MCO $\mathcal{L}_{\widehat{\mathcal{P}}}(\theta, \mathbf{y}, N)$ converges to $\ln p_\theta(\mathbf{y})$ at a rate that can be specified under relatively mild assumptions [13]. By setting $\widehat{\mathcal{P}}^{\mathrm{ELBO}}(\theta, \varphi, \mathbf{y}, N) = p_\theta(\mathbf{y}, \mathbf{x}_N)/q_\varphi(\mathbf{x}_N)$, we recover the ELBO as a special case. The Importance Weighted Auto-Encoder (IWAE) bound [3] is obtained by taking $\widehat{\mathcal{P}}^{\mathrm{IW}}(\theta, \varphi, \mathbf{y}, N) = \frac{1}{N}\sum_{i=1}^N \frac{p_\theta(\mathbf{y}, \mathbf{x}_i)}{q_\varphi(\mathbf{x}_i)}$, and yields the MCO:

$$\mathcal{L}_{\mathrm{IW}}(\theta, \varphi, \mathbf{y}, N) = \mathbb{E}_{q_\varphi^{\otimes n}}\left[\ln\left(\frac{1}{N}\sum_{i=1}^N \frac{p_\theta(\mathbf{y}, \mathbf{x}_i)}{q_\varphi(\mathbf{x}_i)}\right)\right]. \quad (4)$$

Note that we have $\mathrm{ELBO}(\theta, \varphi, \mathbf{y}) = \mathcal{L}_{\mathrm{IW}}(\theta, \varphi, \mathbf{y}, 1)$.

Theoretical insights on the IWAE estimator. While it may not be obvious at first glance, the IWAE estimator indeed minimizes a KL divergence [9] and provides a tighter bound on the evidence than the traditional ELBO [3]. Specifically, for all $N \in \mathbb{N}^*$, we have

$$\mathcal{L}_{\mathrm{IW}}(\theta, \varphi, \mathbf{y}, N) \leq \mathcal{L}_{\mathrm{IW}}(\theta, \varphi, \mathbf{y}, N+1) \leq \ln p_\theta(\mathbf{y}).$$

Remarkably, $\mathcal{L}_{\mathrm{IW}}(\theta, \varphi, \mathbf{y}, N)$ converges to the true log-evidence $\ln p_\theta(\mathbf{y})$ as the number of samples N goes to infinity, independently of the expressiveness of the variational family $\mathcal{Q}$. While approaching the evidence more closely may sound beneficial, it must be noted that the per-iteration computational cost of IWVI grows with N. More importantly, it has been shown in [16] that increasing N can degrade inference performance in certain settings due to a worsening signal-to-noise ratio of the gradient estimator. The asymptotics of Importance-Weighted Variational Inference (IWVI) have since been studied extensively in [4]. The analyses reveal new insights and establish consistency and asymptotic efficiency of parameter estimates under smoothness conditions, as both the sample size N and the size of observed data tend to infinity.

5 Alpha-divergence VI

Shortcomings of the KL divergence. Although minimizing the forward KL divergence often yields accurate and efficient estimation of the posterior mean, it may fail to capture complexity expressed in higher order posterior moments [14, 17]. Recall the expression

$$D_{\mathrm{KL}}(q_\varphi(\cdot) \,\|\, p_\theta(\cdot \,|\, \mathbf{y})) = \int_{\mathsf{X}} \ln\left(\frac{q_\varphi(\mathbf{x})}{p_\theta(\mathbf{x} \,|\, \mathbf{y})}\right) q_\varphi(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

This learning objective heavily penalizes variational approximations that put mass in regions of X where $p_\theta(\mathbf{x} \,|\, \mathbf{y})$ vanishes. As a result, q_φ is forced towards high-density regions of the posterior, effectively leading to a mode-seeking behavior and potentially neglecting lower-density regions. Note that the opposite behavior is expected if we consider the reverse KL divergence, $D_{\mathrm{KL}}(p_\theta(\cdot \,|\, \mathbf{y}) \,\|\, q_\varphi(\cdot))$. In that case, q_φ is discouraged from being zero in regions where the posterior density is non-negligible. This drives the variational approximation to extend over the full support of $p_\theta(\mathbf{x} \,|\, \mathbf{y})$, promoting a mass-covering behavior.

Alpha-divergences. The class of alpha-divergences provides a flexible framework that encompasses both the forward and reverse KL divergences as special cases and allows tuning between mode-seeking, and mass covering tendencies through a scalar hyperparameter $\alpha \in \mathbb{R} \setminus \{0, 1\}$. Following the convention established in [5], we define

$$D_\alpha(q_\varphi \,\|\, p_\theta(\cdot \,|\, \mathbf{y})) = \frac{1}{\alpha(\alpha-1)}\left[\int_{\mathsf{X}} q_\varphi^\alpha(\mathbf{x})\, p_\theta^{1-\alpha}(\mathbf{x} \,|\, \mathbf{y})\, \mathrm{d}\mathbf{x} - 1\right].$$

The definition can be extended by continuity to $\alpha = 1$, in which case we recover the forward KL divergence, and $\alpha = 0$ which corresponds to the reverse KL divergence. Other notable special cases include the Hellinger distance at $\alpha = 0.5$ and chi-square divergence at $\alpha = 2$.

Variational-Rényi bounds. Minimizing $D_\alpha(q_\varphi \,\|\, p_\theta(\cdot \,|\, \mathbf{y}))$ for $\alpha \neq 1$ (or, equivalently, maximizing $\ln p_\theta(\mathbf{y}) - D_\alpha(q_\varphi \,\|\, p_\theta(\cdot|\mathbf{y}))$) can be done by maximizing the Variational-Rényi (VR) bound [12],

$$\mathcal{L}_\alpha(\theta, \varphi, \mathbf{y}) = \frac{1}{1-\alpha}\ln\left(\int_{\mathsf{X}} q_\varphi^\alpha(\mathbf{x}) p_\theta^{1-\alpha}(\mathbf{y}, \mathbf{x})\mathrm{d}\mathbf{x}\right).$$

Similarly to (4), a multi-sample version of the VR bound can be defined. The VR-IWAE bound [6] writes

$$\mathcal{L}_\alpha(\theta, \varphi, \mathbf{y}, N) = \frac{1}{1-\alpha}\mathbb{E}_{q_\varphi^{\otimes n}}\left[\ln\left(\frac{1}{N}\sum_{i=1}^N \left(\frac{p_\theta(\mathbf{y}, \mathbf{x}_i)}{q_\varphi(\mathbf{x}_i)}\right)^{1-\alpha}\right)\right]. \quad (5)$$

Variational parameters optimization. The improved flexibility offered by the class of alpha-divergences comes at the price of greater complexity and additional challenges when optimizing the parameters, notably biased gradient estimators. Alternatively to gradient-based optimization of the VR-bound, an iterative optimization scheme was proposed in [7], with the following update rule:

$$\varphi_{k+1} = \arg\max_{\varphi \in \mathsf{H}} \int_{\mathsf{X}} \zeta_\alpha(\mathbf{x}; \theta, \varphi_k) \ln \left(\frac{q_\varphi(\mathbf{x})}{q_{\varphi_k}(\mathbf{x})} \right) \mathrm{d}\mathbf{x}, \quad (6)$$

where $\zeta_\alpha(\mathbf{x}; \theta, \varphi) \propto q_\varphi^\alpha(\mathbf{x})\, p_\theta^{1-\alpha}(\mathbf{y}, \mathbf{x})$ is a probability density. Assuming that the update (6) can be computed exactly, the algorithm is monotonic in the sense that $D_\alpha(q_{\varphi_{k+1}} \,\|\, p_\theta(\cdot|\mathbf{y})) \leq D_\alpha(q_{\varphi_k} \,\|\, p_\theta(\cdot \mid \mathbf{y}))$. When the variational family $\mathcal{Q}$ is an exponential family with natural parameter φ, sufficient statistic S and log-partition function A, that is, $q_\varphi(\mathbf{x}) = h(y) \exp\left(\langle \varphi, S(\mathbf{x}) \rangle - A(\varphi)\right)$, it can be parameterized by its mean $\mu = \nabla A(\varphi)$. The updates can be expressed through μ:

$$\mu_{k+1} = \int_{\mathsf{X}} S(\mathbf{x})\, \zeta_\alpha(\mathbf{x}; \theta, \varphi_k)\, \mathrm{d}\mathbf{x}. \quad (7)$$

The natural parameters $(\varphi_k)_{k \geq 1}$ can be recovered by computing $\varphi_k = (\nabla A)^{-1}(\mu_k)$. Under relatively mild assumptions, the sequence $(\varphi_k)_{k \geq 1}$ converges to a minimizer of the alpha-divergence at an asymptotically exponential rate [1]. However, the integral in (7) is generally intractable and we must resort to Monte-Carlo estimation to approach it. The resulting estimator is biased, due to the normalizing constant $\int_{\mathsf{X}} q_\varphi^\alpha(\mathbf{x})\, p_\theta^{1-\alpha}(\mathbf{y}, \mathbf{x})\, \mathrm{d}\mathbf{x}$ in $\zeta_\alpha(\cdot\,; \theta, \varphi)$. An alternative update rule involving only unbiased estimators can be considered [1], but this scheme seems to be far less stable and its interest primarily lies in fine-tuning pre-trained models. In the context of Variational Auto-Encoders (VAEs), a link can be drawn with gradient-based optimization of the VR bound [7, 12].

Joint optimization of θ and φ. Recall first that the EM algorithm corresponds precisely to the case where joint optimization is performed alternatively (and exactly), the true posterior lies within the variational family, $\alpha = 0$, and $N = 1$. In the general case, gradient-based updates are generally preferred. For example, promising empirical results were reported in [12], using VAEs. Later theoretical analyses [6, 8] extended the results from [16] on the IWAE bound to the VR-IWAE bound defined in (5), demonstrating clear advantages to choosing $\alpha \in (0, 1)$. However, in high-dimensional settings, gradient estimators collapse unless N grows exponentially with the dimension [8], reinforcing concerns stated in [10]. In short, the theory tends to confirm that there is little benefit in using $N > 1$ in very high-dimensional parameter settings.

References

[1] François Bertholom, Randal Douc, and François Roueff. Asymptotics of alpha-divergence variational inference algorithms with exponential families. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[2] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[3] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *4th International Conference on Learning Representations (ICLR)*, 2016.

[4] Badr-Eddine Cherief-Abdellatif, Randal Douc, Arnaud Doucet, and Hugo Marival. On the asymptotics of importance weighted variational inference, 2025.

[5] Andrzej Cichocki and Shun-ichi Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

[6] Kamélia Daudel, Joe Benton, Yuyang Shi, and Arnaud Doucet. Alpha-divergence variational inference meets importance weighted auto-encoders: Methodology and asymptotics. *Journal of Machine Learning Research*, 24(243):1–83, 2023.

[7] Kamélia Daudel, Randal Douc, and François Roueff. Monotonic alpha-divergence minimisation for variational inference. *Journal of Machine Learning Research*, 24(62):1–76, 2023.

[8] Kamélia Daudel and François Roueff. Learning with importance weighted variational inference: Asymptotics for gradient estimators of the vr-iwae bound, 2024.

[9] Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. *Advances in neural information processing systems*, 31, 2018.

[10] Tomas Geffner and Justin Domke. On the difficulty of unbiased alpha divergence minimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 3650–3659, 2021.

[11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[12] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[13] Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering variational objectives. *Advances in Neural Information Processing Systems*, 30, 2017.

[14] Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, January 2005.

[15] Andriy Mnih and Danilo J Rezende. Variational inference for Monte Carlo objectives. *arXiv:1602.06725*, 2016.

[16] Tom Rainforth, Adam Kosiorek, Tuan Anh Le, Chris Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 4277–4285, 2018.

[17] Yue Yang, Ryan Martin, and Howard Bondell. Variational approximations using fisher divergence. *arXiv preprint arXiv:1905.05284*, 2019.