Operator Variational Inference

1. Can we formalize computational **tradeoffs** in inference?
2. Can we leverage **intractable** distributions as approximate posteriors?

Rajesh Ranganath

Dustin Tran

David Blei

Background

Given

- Data set $\mathbf{x}$.
- Generative model $p(\mathbf{x}, \mathbf{z})$ with latent variables $\mathbf{z} \in \mathbb{R}^d$.

Goal

- Infer posterior $p(\mathbf{z} \mid \mathbf{x})$.

Background

Variational inference

- Posit a family of distributions $q \in \mathcal{Q}$.
- Typically minimize $\mathrm{KL}\left(q \parallel p\right)$, which is equivalent to maximizing

$$\mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z})].$$

Operator Objectives

There are three ingredients:

1. An operator $O^{p,q}$ that depends on $p(\mathbf{z} \,|\, \mathbf{x})$ and $q(\mathbf{z})$.
2. A family of test functions $f \in \mathcal{F}$, where each $f(\mathbf{z}) : \mathbb{R}^d \to \mathbb{R}^d$.
3. A distance function $t(a) : \mathbb{R} \to [0, \infty)$.

These three ingredients combine to form an operator objective,

$$\sup_{f \in \mathcal{F}} t\big(\ \mathbb{E}_{q(\mathbf{z})}[(O^{p,q} f)(\mathbf{z})]\ \big).$$

It is the worst-case expected value among all functions $f \in \mathcal{F}$.

Operator Objectives

The goal is to minimize this objective,

$$\inf_{q \in \mathcal{Q}} \sup_{f \in \mathcal{F}} t(\ \mathbb{E}_{q(\mathbf{z})}[(O^{p,q} f)(\mathbf{z})]\ ).$$

In practice, we parameterize the variational family, $\{q(\mathbf{z}; \lambda)\}$. We also parameterize the test functions $\{f(\mathbf{z}; \theta)\}$ with a neural network.

$$\lambda^* = \min_{\lambda} \max_{\theta} t(\ \mathbb{E}_{\lambda}[(O^{p,q} f_\theta)(z)]\ )$$

Operator Objectives

$$\sup_{f \in \mathcal{F}} t\big(\ \mathbb{E}_{q(\mathbf{z})}[(O^{p,q} f)(\mathbf{z})]\ \big).$$

To use these objectives for variational inference, we impose two conditions:

1. *Closeness*. Its minimum is achieved at the posterior,

$$\mathbb{E}_{p(\mathbf{z} \mid \mathbf{x})}[(O^{p,p} f)(\mathbf{z})] = 0 \text{ for all } f \in \mathcal{F}.$$

2. *Tractability*. The operator $O^{p,q}$—originally in terms of $p(\mathbf{z} \mid \mathbf{x})$ and $q(\mathbf{z})$—can be written in terms of $p(\mathbf{x}, \mathbf{z})$ and $q(\mathbf{z})$.

Operator Objectives: Examples

KL variational objective. The operator is

$$(O^{p,q} f)(z) = \log q(\mathbf{z}) - \log p(\mathbf{x}, \mathbf{z}) \quad \forall f \in \mathcal{F}.$$

With distance function $t(a) = a$, the objective is

$$\mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z}) - \log p(\mathbf{x}, \mathbf{z})].$$

Operator Objectives: Examples

KL variational objective. The operator is

$$(O^{p,q} f)(\mathbf{z}) = \log q(\mathbf{z}) - \log p(\mathbf{x}, \mathbf{z}) \quad \forall f \in \mathcal{F}.$$

With distance function $t(a) = a$, the objective is

$$\mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z}) - \log p(\mathbf{x}, \mathbf{z})].$$

Langevin-Stein variational objective. The operator is

$$(O^{p} f)(\mathbf{z}) = \nabla_z \log p(\mathbf{x}, \mathbf{z})^\top f(\mathbf{z}) + \nabla^\top f,$$

where $\nabla^\top f$ is the divergence of f. With distance function $t(a) = a^2$, the objective is

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}_{q(\mathbf{z})}[\nabla_z \log p(\mathbf{x}, \mathbf{z})^\top f(\mathbf{z}) + \nabla^\top f] \right)^2.$$

Operator Variational Inference

$$\min_{\lambda} \max_{\theta} t(\mathbb{E}_{\lambda}[(O^{p,q} f_{\theta})(z)]).$$

Fix $t(a) = a^2$; the case of $t(a) = a$ easily applies.

Operator Variational Inference

$$\min_{\lambda} \max_{\theta} t(\mathbb{E}_{\lambda}[(O^{p,q} f_{\theta})(z)]).$$

Fix $t(a) = a^2$; the case of $t(a) = a$ easily applies.

Gradient with respect to λ. (Variational approximation)

$$\nabla_{\lambda}\mathcal{L}_{\theta} = 2\, \mathbb{E}_{\lambda}[(O^{p,q} f_{\theta})(Z)]\; \nabla_{\lambda}\mathbb{E}_{\lambda}[(O^{p,q} f_{\theta})(Z)].$$

We use the score function gradient (Ranganath et al., 2014) and reparameterization gradient (Kingma & Welling, 2014).

Operator Variational Inference

$$\min_{\lambda} \max_{\theta} t\left(\mathbb{E}_{\lambda}[(O^{p,q} f_{\theta})(z)] \right).$$

Fix $t(a) = a^2$; the case of $t(a) = a$ easily applies.

Gradient with respect to λ. (Variational approximation)

$$\nabla_{\lambda}\mathcal{L}_{\theta} = 2\, \mathbb{E}_{\lambda}[(O^{p,q} f_{\theta})(Z)]\, \nabla_{\lambda}\mathbb{E}_{\lambda}[(O^{p,q} f_{\theta})(Z)].$$

We use the score function gradient (Ranganath et al., 2014) and reparameterization gradient (Kingma & Welling, 2014).

Gradient with respect to θ. (Test function)

$$\nabla_{\theta}\mathcal{L}_{\lambda} = 2\, \mathbb{E}_{\lambda}[(O^{p,q} f_{\theta})(z)]\, \mathbb{E}_{\lambda}[\nabla_{\theta} O^{p,q} f_{\theta}(z)].$$

We construct stochastic gradients with two sets of Monte Carlo estimates.

Characterizing Objectives: Data Subsampling

Stochastic optimization scales variational inference to massive data (Hoffman et al., 2013; Salimans & Knowles, 2013). The idea is to subsample data and scale the log-likelihood,

$$\log p(x_{1:n}, z_{1:n}, \beta) = \log p(\beta) + \sum_{n=1}^{N} \Big[\log p(x_n \mid z_n, \beta) + \log p(z_n \mid \beta) \Big].$$

$$\approx \log p(\beta) + \frac{M}{N} \sum_{m=1}^{M} \Big[\log p(x_m \mid z_m, \beta) + \log p(z_m \mid \beta) \Big].$$

Characterizing Objectives: Data Subsampling

Stochastic optimization scales variational inference to massive data (Hoffman et al., 2013; Salimans & Knowles, 2013). The idea is to subsample data and scale the log-likelihood,

$$\log p(x_{1:n}, z_{1:n}, \beta) = \log p(\beta) + \sum_{n=1}^{N} \Big[\log p(x_n \mid z_n, \beta) + \log p(z_n \mid \beta) \Big].$$
$$\approx \log p(\beta) + \frac{M}{N} \sum_{m=1}^{M} \Big[\log p(x_m \mid z_m, \beta) + \log p(z_m \mid \beta) \Big].$$

One class of operators which admit data subsampling are linear operators with respect to $\log p(\mathbf{x}, \mathbf{z})$.

The LS and KL operators are examples in this class. (An operator for f-divergences is not.)

Characterizing Objectives: Variational Programs

Recent advances in variational inference aim to develop expressive approximations, such as with transformations (Rezende & Mohamed, 2015; Tran et al., 2015; Kingma et al., 2016) and auxiliary variables (Salimans et al., 2015; Tran et al., 2016; Ranganath et al., 2016).

In variational inference, our design of the variational family $q \in \mathcal{Q}$ is limited by a tractable density.

Characterizing Objectives: Variational Programs

We can design operators that do not depend on q, $O^{p,q} = O^p$, such as the LS objective

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}_{q(\mathbf{z})}[\nabla_z \log p(\mathbf{x}, \mathbf{z})^\top f(\mathbf{z}) + \nabla^\top f] \right)^2.$$

The class of approximating families is much larger, which we call *variational programs*.

Consider a generative program of latent variables,

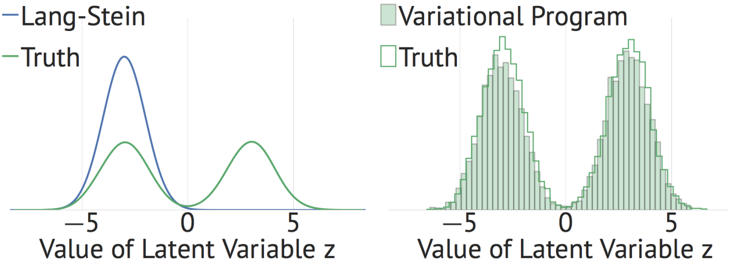
$$\epsilon \sim \text{Normal}(0, 1), \quad \mathbf{z} = G(\epsilon; \lambda),$$

where G is a neural network. The program is differentiable and generates samples for $\mathbf{z}$. Moreover, its density does not have to be tractable.

Experiments

Variational program:

1. Draw $\epsilon, \epsilon' \sim \mathrm{Normal}(0, 1)$.
2. If $\epsilon' > 0$, return $G_1(\epsilon)$; else if $\epsilon' \leq 0$, return $G_2(\epsilon)$.



1-D Mixture of Gaussians. LS with a Gaussian family fits a mode. LS with a variational program approaches the truth.

Experiments

We model binarized MNIST, $\mathbf{x}_n \in \{0, 1\}^{28 \times 28}$, with

$$\mathbf{z}_n \sim \text{Normal}(0, 1),$$
$$\mathbf{x}_n \sim \text{Bernoulli}(\text{logistic}(\mathbf{z}_n^\top \mathbf{W} + \mathbf{b})),$$

where $\mathbf{z}_n$ has latent dimension 10 and with parameters $\{\mathbf{W}, \mathbf{b}\}$.

Inference method	Completed data log-likelihood
Mean-field Gaussian + KL$(q\|\|p)$	-59.3
Mean-field Gaussian + LS	-75.3
Variational Program + LS	-58.9

References

- J. Altosaar, R. Ranganath, and D.M. Blei. Proximity variational inference. *NIPS, Approximate Inference Workshop*, 2016.
- R. Ranganath, J. Altosaar, D. Tran, and D.M. Blei. Operator variational inference. *NIPS*, 2016.