SAND2022-4462C

Sandia National Laboratories

Exceptional service in the national interest

A Study of Bias-Variance Trade-off in Variational Inferencing Using Delta Method

SIAM UQ 2022 : **Variational Inference Bridging Application and Theory**

Niladri Das

Thomas A. Catanach

April 14, 2022

Atlanta, GA, USA

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

CONTENTS

Variational Inference

Variational inference (VI) is a method for approximating probability densities

$$\min_{\phi}[q(\theta \mid \phi) \,||\, p(\theta \mid \mathcal{D})]$$

$$= \min_{\phi} \int q(\theta \mid \phi) \log \frac{q(\theta \mid \phi)}{p(\theta \mid \mathcal{D})} d\theta$$

$$= \boxed{\min_{\phi} \int q(\theta \mid \phi) \log \frac{q(\theta \mid \phi)}{p(\mathcal{D} \mid \theta) p(\theta)} d\theta} + \log p(\mathcal{D})$$

Motivation

- In sequential data assimilation, calculating the evidence term is tractable.
- We optimize for the VI distribution parameter over sample estimates.
- Can we study the bias and variance of those sample estimates ?
- Low variance estimate can enable larger steps in the optimization problem, leading to faster convergence.

BIAS-VARIANCE

Let $\mathbf{X} \in \mathbb{R}^d$ be a random variable with probability distribution F with finite mean μ and variance Σ.

We are interested in $f(\mu) \in \mathbb{R}$ having access to samples from F. What can we say about the confidence in our sample estimate of $f(\mu)$?

Sampling from $F : \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$

Estimating $\mu : \hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{X}_j$

Estimating $f(\mu) : f(\hat{\mu}) = f\left(\frac{1}{n}\sum_{j=1}^{n} \mathbf{X}_j\right)$

What can we say about the bias and variance (confidence) in the estimate $f(\hat{\mu})$?

ESTIMATION BIAS-I

Taylor series expansion of $f(\hat{\mu})$:

$$f(\hat{\mu}) = f(\mu) + (\hat{\mu} - \mu)^T Df(\mu) + \frac{1}{2}(\hat{\mu} - \mu)^T Hf(\mu)(\hat{\mu} - \mu) + \text{H.O.T.}$$

Assumption 1:

Second order approximation of $f(\hat{\mu})$ yields:

$$\mathbb{E}\big(f(\hat{\mu}) - f(\mu)\big) = \frac{1}{2}\mathbb{E}\big((\hat{\mu} - \mu)^T Hf(\mu)(\hat{\mu} - \mu)\big)$$

ESTIMATION BIAS-II

If we assume that X is a 1-D Gaussian random variable, we analyze the exact nature of the bias. The Taylor series expansion:

$$f(\hat{\mu}) = f(\mu) + \sum_{i=1}^{\infty} \frac{1}{i!}(\hat{\mu} - \mu)^i f^i(\mu)$$

$$\mathbb{E}\big(f(\hat{\mu}) - f(\mu)\big) = \mathbb{E}\left(\sum_{i=1}^{\infty} \frac{1}{i!}(\hat{\mu} - \mu)^i f^i(\mu)\right)$$

If $f(x) = \log(x) \rightarrow f^1(x) = \frac{1}{x} \rightarrow f^2(x) = -\frac{1}{x^2} \rightarrow f^3(x) = +2\frac{1}{x^3} \rightarrow \ldots \rightarrow f^i(x) = (-1)^{i-1}\frac{(i-1)!}{x^i}$

ESTIMATION BIAS-IV

Let us see how the co-coefficients of the term $\frac{\Sigma^{i/2}}{n^{i/2}\mu^{i}}$ look like. For $i = \{2,4,6,8,10,12\}$ the values look like

$$\{-\frac{1}{2}, -\frac{3}{4}, -\frac{5}{2}, -\frac{105}{8}, -\frac{189}{2}, -\frac{3465}{4}\}$$

For $i = 2$ we have,

$$\mathbb{E}\big(f(\hat{\mu}) - f(\mu)\big) = -\frac{\Sigma}{2n\mu^2}$$

ESTIMATION VARIANCE-I

Assumption 3:

We approximate the estimate $f(\hat{\mu})$ using the first order term only and analyze the confidence based upon this assumption.

Since $\mathbb{E}(\hat{\mu}) = \mu$, the expected value of the estimate $f(\hat{\mu})$ is $f(\mu)$ and the variance is:

$$\mathbb{E}\big((\hat{\mu} - \mu)^T f'(\mu) f'(\mu)^T (\hat{\mu} - \mu)\big) = f'(\mu)^T \frac{1}{n} \Sigma f'(\mu)$$

We can only calculate it if we know the actual μ and the variance Σ of F. Since $\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} \quad X_j$, using C.L.T. we have:

$$\sqrt{n}(\hat{\mu} - \mu) \to \mathcal{N}(0, \Sigma)$$

Using assumption 3 and C.L.T. we have:

$$\sqrt{n}\big(f(\hat{\mu}) - f(\mu)\big) \to \mathcal{N}\big(0, f'(\mu)^T \Sigma f'(\mu)\big)$$

Delta Method

ESTIMATION VARIANCE-II

Assumption 4:

If we do not have access to actual μ and the variance Σ of F we estimate the $f'(\mu)^T \Sigma f'(\mu)$ by $\hat{\mu}$ and $\hat{\Sigma}$ by:

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu})\,(X_i - \hat{\mu})^T$$

Combining assumption 1 and 2 we can state different levels of confidence in our estimate of $f(\mu)$ and say that $f(\mu)$ lies with in:

$$f(\hat{\mu}) \pm \frac{2.58}{\sqrt{n}} (f'^{(\hat{\mu})^T \hat{\Sigma}} f'^{(\hat{\mu})})^{0.5}$$

with 99% confidence.

DELTA METHOD ON VI

The KL divergence between the variational distribution and the posterior is:

$$J = \int q(\theta \mid \phi) \log \frac{q(\theta \mid \phi)}{p(\theta)} d\theta - \int q(\theta \mid \phi) \log p(\mathcal{D} \mid \theta) d\theta + \log \int \frac{p(\mathcal{D} \mid \theta) p(\theta)}{q(\theta \mid \phi)} q(\theta \mid \phi) d\theta$$

$$= f(\phi) - \int \boxed{\frac{q(\theta(\zeta) \mid \phi)}{r(\theta(\zeta) \mid \varphi)}} \log p\big(\mathcal{D} \mid \theta(\zeta)\big) p(\zeta) d\zeta + \log \int \boxed{\frac{p\big(\mathcal{D} \mid \theta(\zeta)\big) p\big(\theta(\zeta)\big)}{r(\theta(\zeta) \mid \varphi)}} p(\zeta) d\zeta$$

where $r(\theta \mid \varphi)$ is an arbitrary importance distribution

- ζ is the re-parameterized random variable.
- We assume that the first term $f(\phi)$ can be exactly calculated and do not need to be a MC estimate.
- The remaining two terms are approximated as a MC estimate. Do we need re-parameterization or should we sample directly from the importance distribution?

$$-\int \boxed{\frac{q(\theta(\zeta)|\phi)}{r(\theta(\zeta)|\varphi)}} \log p\left(\mathcal{D} \mid \theta(\zeta)\right) p(\zeta) d\zeta + \log \int \boxed{\frac{p\left(\mathcal{D}|\theta(\zeta)\right)p\left(\theta(\zeta)\right)}{r(\theta(\zeta)|\varphi)}} p(\zeta) d\zeta$$

$$\approx -\frac{1}{N}\sum_{i=1}^{N} \boxed{\frac{q(\theta(\zeta_i)|\phi)}{r(\theta(\zeta_i)|\varphi)}} \log p\left(\mathcal{D} \mid \theta(\zeta_i)\right) + \log\left(\frac{1}{M}\sum_{j=1}^{M} \boxed{\frac{p\left(\mathcal{D}|\theta(\zeta_j)\right)p\left(\theta(\zeta_j)\right)}{r(\theta(\zeta_j)|\varphi)}}\right)$$

We assume that the ϕ (Variational parameters) and φ (Importance parameters) are coupled by:

$$\varphi = \phi + \eta$$

First calculate the variance and bias of the estimates:

$$\hat{\mathcal{X}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{q(\theta(\zeta_i) \mid \phi)}{r(\theta(\zeta_i) \mid \phi + \eta)}\log p\big(\mathcal{D} \mid \theta(\zeta_i)\big) \quad and \quad \hat{\mathcal{Y}} = \frac{1}{M}\sum_{j=1}^{M}\frac{p\left(\mathcal{D} \mid \theta(\zeta_j)\right)p\left(\theta(\zeta_j)\right)}{r\big(\theta(\zeta_j) \mid \phi + \eta\big)}$$

Now,

$$\mathcal{V}ar(\hat{\mathcal{X}}) \quad = \frac{1}{N}\mathcal{V}ar\left(-\frac{q(\theta(\zeta) \mid \phi)}{r(\theta(\zeta) \mid \phi + \eta)}\log p\big(\mathcal{D} \mid \theta(\zeta)\big)\right) = \Sigma_{\mathcal{X}}/N$$

We assume that we know the Σ_X (How to calculate this?), then

$$\sqrt{N}(\widehat{X} - X)XX \rightarrow \mathcal{N}(0, \Sigma_X)$$

Now,

$$\mathcal{V}ar(\hat{y}) \quad = \frac{1}{M}\mathcal{V}ar\left(\frac{p(\mathcal{D} \mid \theta(\zeta))p(\theta(\zeta))}{r(\theta(\zeta_j) \mid \phi + \eta)}\right) = \Sigma_y/M$$

We assume that we know the Σ_y (How to calculate this?), then

$$\sqrt{M}(\hat{y} - y)XX \rightarrow d\mathcal{N}(0, \Sigma_y)$$

Use Delta-method to calculate the variance and bias of

$$\log\left(\frac{1}{M}\sum_{j=1}^{M}\frac{p\left(\mathcal{D} \mid \theta(\zeta_j)\right)p\left(\theta(\zeta_j)\right)}{r(\theta(\zeta_j) \mid \phi + \eta)}\right)$$

We have shown in the earlier step,

$$\sqrt{M}(\hat{y} - y)XX \rightarrow d\mathcal{N}(0, \Sigma_y)$$

Using Delta method,

$$\sqrt{M}\left(\log(\hat{y}) - \log(y)\right)XX \rightarrow d\mathcal{N}\left(0, \frac{\Sigma_y}{y^2}.\right)$$

Then calculate the variance and bias of

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{q(\theta(\zeta_i) \mid \phi)}{r(\theta(\zeta_i) \mid \phi + \eta)}\log p\left(\mathcal{D} \mid \theta(\zeta_i)\right) \quad and \quad \log\left(\frac{1}{M}\sum_{j=1}^{M}\frac{p\left(\mathcal{D} \mid \theta(\zeta_j)\right)p\left(\theta(\zeta_j)\right)}{r\left(\theta(\zeta_j) \mid \phi + \eta\right)}\right)$$

using delta method for multiple variables.

- If M and N samples are drawn independently then, (Shouldn't this be called MSE rather than Var?)

$$var\left(\hat{X} + \log(\hat{Y})\right) = \frac{\Sigma_X}{N} + \frac{\Sigma_Y}{MY^2}$$

- If If M=N samples are the same samples then, using multivariate Delta method,

$$var\left(\hat{X} + \log(\hat{Y})\right) = \frac{\Sigma_X}{N} + \frac{\Sigma_Y}{NY^2} + 2\frac{\Sigma_{XY}}{NY}$$

- Hence,

$$\frac{\Sigma_X}{N} + \frac{\Sigma_Y}{NY^2} - 2\frac{\sqrt{\Sigma_X\Sigma_Y}}{NY} < var\left(\hat{X} + \log(\hat{Y})\right) < \frac{\Sigma_X}{N} + \frac{\Sigma_Y}{NY^2} + 2\frac{\sqrt{\Sigma_X\Sigma_Y}}{NY}$$