Variational Inference: "Does it work?"

Anna, Chris and Sahra

December 2, 2019

Bayesian inference

- Data $\mathbf{x}$
- Latent parameter θ
- Given prior $p(\theta)$ and likelihood $p(\mathbf{x}|\theta)$ posterior $p(\theta|\mathbf{x})$ is given by

$$\pi(\theta) = p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

- In practice $\pi(\theta)$ is intractable $\longrightarrow$ an approximation: $\hat{\pi}$ is needed.

Variational Inference

- Approximation of posterior by variational distribution $\pi^*(\theta)$ in variational family $\mathcal{Q}$, such that:

$$\pi^*(\theta) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \; D_{KL}(q(\theta)||\pi(\theta))$$

- Diagnostics presented:
  - Yao et al. (2018)
  - Huggins et al. (2019)

- Particle approximation to the posterior distribution is given by sampling from:

  - $\hat{\pi}_{IS}(\theta) = \sum_{i=1}^{S} \tilde{w}_i \delta(\theta^i)$, with $\tilde{w}_i = \frac{w_i}{\sum_{i=1}^{S} w_i}$
  - $\hat{\pi}_{PSIS}(\theta) = \sum_{i=1}^{S} \tilde{r}_i \delta(\theta^i)$, with $\tilde{r}_i = \frac{r_i}{\sum_{i=1}^{S} r_i}$

Pareto-Smoothed Importance Sampling (PSIS)

Given variational approximation to the posterior $\hat{\pi}(\theta)$, what is the most appropriate way to estimate the integral $\mathrm{E}_\pi[h(\theta)]$ sampling from $\theta^i \overset{i.i.d}{\sim} \hat{\pi}$?

- **Monte Carlo**: $T_{MC} = \frac{1}{S}\sum_{i=1}^{S} h(\theta^i)$ - biased, inconsistent, low variance
- **Importance Sampling**: $w_i = \frac{p(\theta^i, \mathbf{x})}{\hat{\pi}(\theta^i)}$ $T_{IS} = \frac{\sum_{i=1}^{S} w_i h(\theta^i)}{\sum_{i=1}^{S} w_i}$ - asymptotically unbiased, consistent, high variance?
- **Pareto Smoothed Importance Sampling**: $T_{PSIS} = \frac{\sum_{i=1}^{S} r_i h(\theta^i)}{\sum_{i=1}^{S} r_i}$ - asymptotically unbiased, lower variance

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman (2018). *"Yes, but Did It Work?: Evaluating Variational Inference"*

Pareto-Smoothed Importance Sampling (PSIS) ctd...

The weights r_i are derived by fitting the generalised Pareto distribution to the highest $\min(\frac{S}{5}, 3\sqrt{S})$ importance weights w_i, replacing the largest importance weights w_i with the expected value of the fitted distribution.

$$p(y|k,\mu,\sigma) = \frac{1}{\sigma}(1 + k(\frac{y-\mu}{\sigma}))^{-\frac{1}{k}-1} \quad (1)$$

- Extreme value distributions of random variables: $X|X > u$ converge to generalised Pareto distributions.
- Parameter k links to moments of random variables: $k = \inf\{c \in \mathbb{R} : \mathrm{E}[X^{\frac{1}{c}}] < \infty\}$
- Hence, $\hat{k}$ is estimator of $k = \inf\{c \in \mathbb{R} : \mathrm{E}_{\pi^*}[\frac{p(\theta,\mathbf{x})}{\pi^*(\theta)}^{\frac{1}{c}}] < \infty\}$
- Low $\hat{k}$ - fast convergence of (Pareto smoothed) importance sampling, high $\hat{k}$, slow convergence

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman (2018). *"Yes, but Did It Work?: Evaluating Variational Inference"*

Pareto-Smoothed Importance Sampling (PSIS) ctd...

- Finiteness of the moments of $\mathrm{E}_{\pi^*}[\frac{p(\theta,\mathbf{x})}{\pi^*(\theta)}^{\frac{1}{k}}]$ corresponds to the finiteness of the Rényi divergence $D_{\frac{1}{k}}(\pi||\pi^*) = \frac{k}{1-k}\log(\int_\Theta \pi(\theta)^{\frac{1}{k}}\pi^*(\theta)^{1-\frac{1}{k}})$.
- Thus, also serves as a measure of accuracy of samples from the posterior.
- $\hat{k} \geq 1$ implies infinite KL divergence...!

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman (2018). *"Yes, but Did It Work?: Evaluating Variational Inference"*

Variation Simulation Based Calibration (VSBC)

Given a defined Bayesian model, with D dimensional parameter $\theta = (\theta_1, \ldots, \theta_D)^T$ the following steps are conducted:

1) Generate parameter θ^i from the prior $p(\theta)$
2) Generate dataset $\mathbf{x}^i$ from the likelihood $p(\mathbf{x}|\theta^i)$ - the resulting values $(\theta^i, \mathbf{x}^i)$ are a sample from the joint distribution $p(\mathbf{x}^i, \theta)$, and therefore θ^i is a sample from $p(\theta|\mathbf{x}^i)$
3) Use a VI approach to approximate $p(\theta|\mathbf{x}^i)$ with $\hat{\pi}_i(\theta)$

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman (2018). *"Yes, but Did It Work?: Evaluating Variational Inference"*

Variation Simulation Based Calibration (VSBC) ctd...

4) Generate sufficiently large S samples of $\theta^{ij} \sim q_i(\theta)$
5) For each parameter component, record $U^i_d = \hat{F}^i_d(\theta^i_d)$, where $\hat{F}^i_d(c) = \frac{1}{S}\sum_{j=1}^{S} \mathbb{I}[\theta^{ij}_d \leq c]$

This method assesses the performance of VI under a Bayesian model for **any** set of responses from the prior predictive $p(\mathbf{x})$.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman (2018). *"Yes, but Did It Work?: Evaluating Variational Inference"*

Combining PSIS with VSBC

We can use VSBC to assess the impact of adjusting variational posteriors with PSIS:

- Particle approximation to the posterior distribution is given by sampling from:
  - $\hat{\pi}_{IS}(\theta) = \sum_{i=1}^{S} \tilde{w}_i \delta(\theta^i)$, with $\tilde{w}_i = \frac{w_i}{\sum_{i=1}^{S} w_i}$
  - $\hat{\pi}_{PSIS}(\theta) = \sum_{i=1}^{S} \tilde{r}_i \delta(\theta^i)$, with $\tilde{r}_i = \frac{r_i}{\sum_{i=1}^{S} r_i}$

Sample from these distributions, instead of from $\hat{\pi}$.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman (2018). *"Yes, but Did It Work?: Evaluating Variational Inference"*

Posterior Error Bounds From Variational Objectives

- **goal:** post-hoc accuracy measure
- **method:** bounds on the error of posterior mean & uncertainty estimates
- **requirement:** approximating & exact posterior have polynomial momements
  - + computational efficiency
  - + weak tail restrictions
  - + evaluation of relevant targets

Jonathan H. Huggins, Mikolaj Kasprzak, Trevor Campbell, and Tamara Broderick (2019). *"Practical Posterior Error Bounds from Variational Objectives"*

Workflow for Variational Inference (Part 1)

1) Select variational family $\mathcal{Q}$ with sufficiently heavy tails.
2) Minimize discrepancy measure to find variational approximation $\hat{\pi}$.
   $\rightarrow$ **KLVI**: maximizing ELBO $\hat{=}$ minimizing KL-divergence
   $\rightarrow$ **CHIVI**: minimizing CUBO $\hat{=}$ minimizing α-Rényi divergence
3) Compute $\hat{k}$,
   $\rightarrow$ **if** there is no guarantee that $k \leq 0$
   $\rightarrow$ **if** $\hat{k} > 0$,
   then refine $\mathcal{Q}$ or reparameterize the model.

Jonathan H. Huggins, Mikolaj Kasprzak, Trevor Campbell, and Tamara Broderick (2019). *"Practical Posterior Error Bounds from Variational Objectives"*

Workflow for Variational Inference (Part 2)

4) Compute $\mathrm{ELBO}(\hat{\pi})$ and $\mathrm{CUBO}_2(\hat{\pi})$.
5) **(optional)** Further optimize the $\mathrm{ELBO}(\xi)$.
6) Compute bound on α-divergence $\bar{\delta} \geq D_2(\pi|\hat{\pi})$.
7) Compute bound on p-Wasserstein distance $\bar{w}_2 \geq \mathcal{W}_2(\pi, \hat{\pi})$.

Jonathan H. Huggins, Mikolaj Kasprzak, Trevor Campbell, and Tamara Broderick (2019). *"Practical Posterior Error Bounds from Variational Objectives"*

Workflow for Variational Inference (Part 3)

8) **If** $\bar{\delta}_2 \uparrow$ and $\bar{w}_2 \uparrow$, **then** refine $\mathcal{Q}$ or reparameterize the model.

9) **If** $\bar{\delta}_2 \downarrow$ and $\bar{w}_2 \uparrow$, **then** use IS or PSIS to refine the posterior expectations produced by $\hat{\pi}$.

10) **If** $\bar{\delta}_2 \downarrow$ and $\bar{w}_2 \downarrow$, **then** use $\hat{\pi}$ to approximate π.

Jonathan H. Huggins, Mikolaj Kasprzak, Trevor Campbell, and Tamara Broderick (2019).*"Practical Posterior Error Bounds from Variational Objectives"*

Diabetes Dataset

input: age, sex, bmi value, blood pressure
output: measure of disease progression

	age	sex	bmi	bp	y
125	-2.437433	1.0	-3.695818	-0.984661	5.081404
333	12.012805	1.0	-2.743032	12.711079	5.099866
32	15.223969	1.0	55.376907	12.711079	5.831882
160	-4.043015	0.0	-27.515464	-32.941387	3.970292
104	-12.070925	0.0	28.698902	-0.984661	4.553877

Bayesian Linear Regression Model Setup

Prior distributions

$$p(\alpha) = \mathcal{N}(\alpha;\ 0,\ 10)$$
$$p(\beta) = \mathcal{N}(\beta;\ 0,\ 1)$$
$$p(\sigma) = Gamma(\sigma;\ 1,\ 1)$$

Likelihood

$$\mu_i = \alpha + \beta_{age} \cdot \mathrm{age}_i + \beta_{sex} \cdot \mathrm{sex}_i + \beta_{bmi} \cdot \mathrm{bmi}_i + \beta_{bp} \cdot \mathrm{bp}_i$$
$$p(\mathbf{y}|\alpha, \beta) = \mathcal{N}(y; \mu, \sigma^2)$$

Variational Distribution Setup

Variational Inference Method: Stochastic Mean-Field VI
Variational Family: Gaussians

$$q(\alpha) = \mathcal{N}(\alpha;\ \mu_\alpha,\ \sigma_\alpha)$$
$$q(\beta) = \mathcal{N}(\beta;\ \mu_\beta,\ \sigma_\beta)$$
$$q(\sigma) = \mathcal{N}(\sigma;\ \mu_\sigma,\ \sigma_\sigma)$$

Initialization:

$$q(\alpha) = \mathcal{N}(\alpha;\ 0,\ 1)$$
$$q(\beta) = \mathcal{N}(\beta;\ 0,\ 1)$$
$$q(\sigma) = \mathcal{N}(\sigma;\ 1, 0.05)$$

Gaussian Process Regression Model Setup

Prior distributions and Likelihood:

$$y_i = f(x_i) + \epsilon_i$$

where

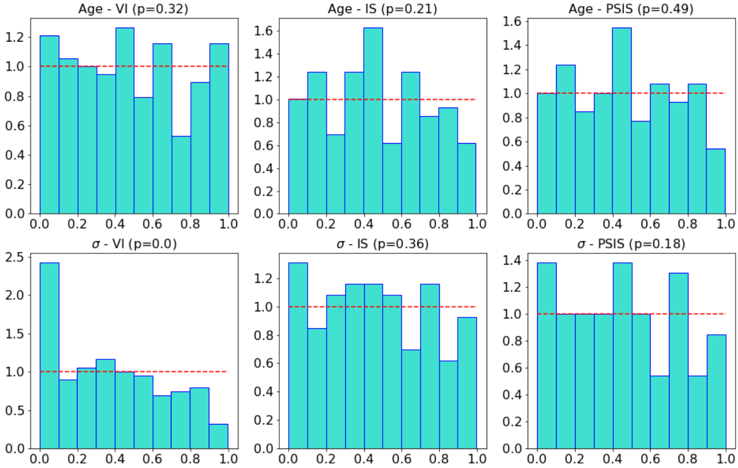
$$f \sim \mathcal{GP}(0, K),$$
K is the squared exponential kernel,
$$\epsilon_i \sim \text{Student-t(df)}$$

Variational Inference Method: variational Gaussian approximation (Opper and Archambeau, 2009)

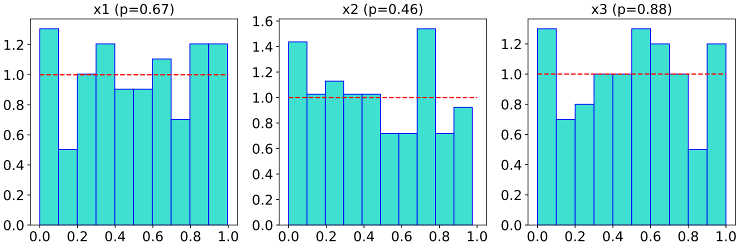
Variational Family: Gaussians

$$q(f) = N(\mu_f, K_f)$$

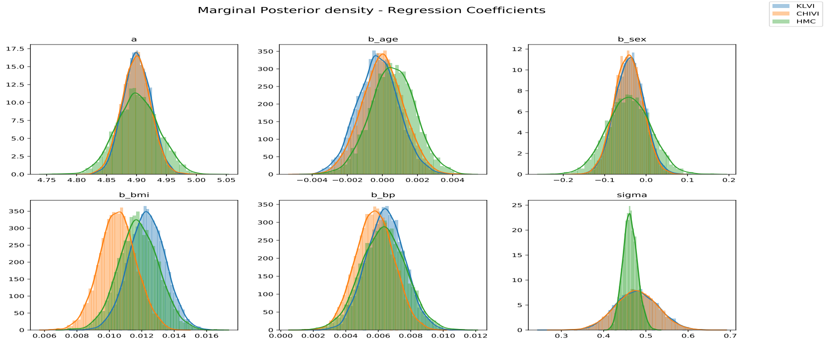
Investigating VI for the model: PSIS with VSBC



Investigating VI for the GP model: PSIS with VSBC



KLVI vs. CHIVI vs. HMC



Error Bounds on Posterior Quantities

KLVI: SVI with KL-divergence as variational objective
CHIVI: SVI with 2-Rényi divergence as variational objective
GP KLVI: Variational Gaussian Approximation with KL-divergence as variational objective and Gaussian process prior

	KLVI	CHIVI	GP KLVI
D_2 bound	9.84	11.50	0.07
$\mathcal{W}_2$ bound	1.64	2.48	4.25
mean error	1.64	2.48	4.25
std error	3.16	4.79	8.22

Conclusion

PSIS	+ simple rules - arbitrary decision bounds
VSBC	+ focuses on measure +/- does not assess for a particular dataset +/- looks at marginals - computationally expensive
divergence/Wasserstein bounds	+ computationally efficient + weak tail restrictions + possible stopping criteria

References I

Jonathan H. Huggins, Mikolaj Kasprzak, Trevor Campbell, and Tamara Broderick. Practical posterior error bounds from variational objectives. 2019.

Manfred Opper and Cédric Archambeau. The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work?: Evaluating variational inference. *arXiv preprint arXiv:1802.02538*, 2018.