

Yes, but Did It Work?: Evaluating Variational Inference (Yao, et al., 2018)

Satvik Saha (ss7228)

October 7, 2025

This paper examines two methods of evaluating the performance of a variational approximation $q(\theta)$ of $p(\theta | x)$. These are diagnostics which can detect some (but not all!) issues with q , from overall distributional differences to pointwise biases. It has been pointed out that simply relying on the ELBO is not enough: this is a non-convex objective function which may flatten out even at poor approximations, and we do not know at what value we should stop!

The first method uses an idea from importance sampling: if $q(\theta)$ is a good proposal distribution for sampling from $p(\theta | x)$, then $\sum_{s=1}^S h(\theta_s) r_s / \sum_{s=1}^S r_s$ is a good Monte-Carlo approximation for $\mathbb{E}_{\theta|x}[h(\theta)]$, where $\theta_s \stackrel{iid}{\sim} q$ and $r_s = p(\theta_s, x) / q(\theta_s)$. Note that this weighted average is dominated by samples $h(\theta_s)$ corresponding to the largest values of r_s . The quality of q can be inferred from $\{r_s\}$ by fitting the largest values against a generalized Pareto distribution and examining a particular parameter k . It is recommended that an estimate $\hat{k} > 0.7$ indicates an unreliable VI estimate. This is justified by drawing a connection between k and a Rényi divergence between $q(\theta)$ and $p(\theta | x)$. Similar adjustments on the aforementioned weighted average (smoothing the largest weights) gives a *Pareto Smoothed Importance Sampling* (PSIS) estimate.

The next method uses the median of q as a pointwise estimate to detect systematic bias, when $p(\theta | x)$ is symmetric. Here, multiple datasets are generated as follows: sample $\theta_j^{(0)} \sim p(\theta)$ from the prior, then $x_j \sim p(x | \theta_j^{(0)})$. With this, we can compute variational approximations $q_j(\theta)$ for each $p(\theta | x_j)$. Finally, compute the p -values $p_j = \mathbb{P}(\theta_j^{(0)} < \theta^* | \theta^* \sim q_j)$. If our variational approximations are good, the distribution of these p -values (hence a histogram of $\{p_j\}$) ought to be symmetric; asymmetry (detected by tests such as Kolmogorov-Smirnov) also indicates the direction of bias in q . This is called a *Variational Simulation-Based Calibration* (VSBC) diagnostic.

References

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work?: Evaluating variational inference. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5581–5590. PMLR, 10–15 Jul 2018.