Variational Inference via χ Upper Bound Minimization

Adji B. Dieng[†], Dustin Tran[†], Rajesh Ranganath[*], John Paisley[†], and David M. Blei[†]

[†]Columbia University, [*]New York University

TL;DR

- Different divergence measures have been proposed for variational inference [1, 2, 3].
- These divergences lead to lower bounds of the log marginal likelihood and underdispersed posterior approximations.
- We propose the χ **divergence** for variational inference.
- It leads to an **upper bound** of the model evidence termed the **CUBO** that can be used alongside the ELBO to sandwich estimate the model evidence.
- the χ divergence favors **overdispersed posterior approximations**.
- We propose **CHIVI** – a black box algorithm for minimizing the CUBO.
- CHIVI uses **unbiased gradients** of the exponentiated CUBO.
- CHIVI is a black box alternative to Expectation Propagation (EP) [4].
- When compared to EP and classical VI, CHIVI produces better error rates and more accurate estimates of posterior variance.

Variational Inference

- Probabilistic generative models posit a joint distribution of data $\boldsymbol{x}$ and latent variable $\boldsymbol{z}$:

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z}) \cdot p(\boldsymbol{x}|\boldsymbol{z}).$$

- The quantity of interest for answering questions about the data is the posterior distribution $p(\boldsymbol{z}|\boldsymbol{x})$.
- For many applications, this posterior distribution is intractable and we must resort to approximate posterior inference.
- Variational inference (VI) [5] casts the approximate posterior inference problem into the optimization of some divergence measure $D(p||q)$ between the target posterior p and a chosen tractable parametric family of distribution q.
- Traditional VI minimizes the Kullback-Leibler divergence:

$$KL(q(\boldsymbol{z};\phi)||p(\boldsymbol{z}|\boldsymbol{x})) = E_q(\log q(\boldsymbol{z};\phi) - \log p(\boldsymbol{z}|\boldsymbol{x})).$$

- This KL divergence is intractable but minimizing it is equivalent to maximizing the ELBO – a tractable lower bound to the log marginal likelihood of the data $\log p(\boldsymbol{x})$:

$$\text{ELBO}(\phi) = E_q(\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q(\boldsymbol{z};\phi)).$$

- However maximizing the ELBO often leads to underestimation of posterior uncertainty.

χ Divergence, CUBO, and Friends

- The posterior uncertainty underestimation problem pertains to all divergence measures $D(q||p)$ from q to p in the f-divergence family.
- We propose the χ divergence

$$D_{\chi^n}(p||q) = E_{q(\boldsymbol{z};\phi)}\left[\left(\frac{p(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z};\phi)}\right)^n - 1\right];$$

where n is the order of the divergence.
- The χ divergence has the same **zero-avoiding** property of the objective of EP and therefore does not suffer from posterior uncertainty underestimation.
- This divergence is intractable but minimizing it is equivalent to minimizing the CUBO

$$\text{CUBO}_n(\phi) = \frac{1}{n}\log \mathbb{E}_{q(\boldsymbol{z};\phi)}\left[\left(\frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z};\phi)}\right)^n\right]. \quad (1)$$

- Naively minimizing Equation (1) yields biased gradients and does not guarantee that the original upper bound is preserved.
- We propose to instead minimize the exponentiated CUBO,

$$\mathcal{L}(\phi) = \mathbb{E}_{q(\boldsymbol{z};\phi)}\left[\left(\frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z};\phi)}\right)^n\right].$$

- This objective can be approximated by Monte Carlo:

$$\hat{\mathcal{L}}(\phi) = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z};\phi)}\right)^n \quad (2)$$

CHIVI: χ Divergence Variational Inference

- Contrary to Equation (1), Equation (2) preserves the upper bound during optimization using unbiased noisy gradients,

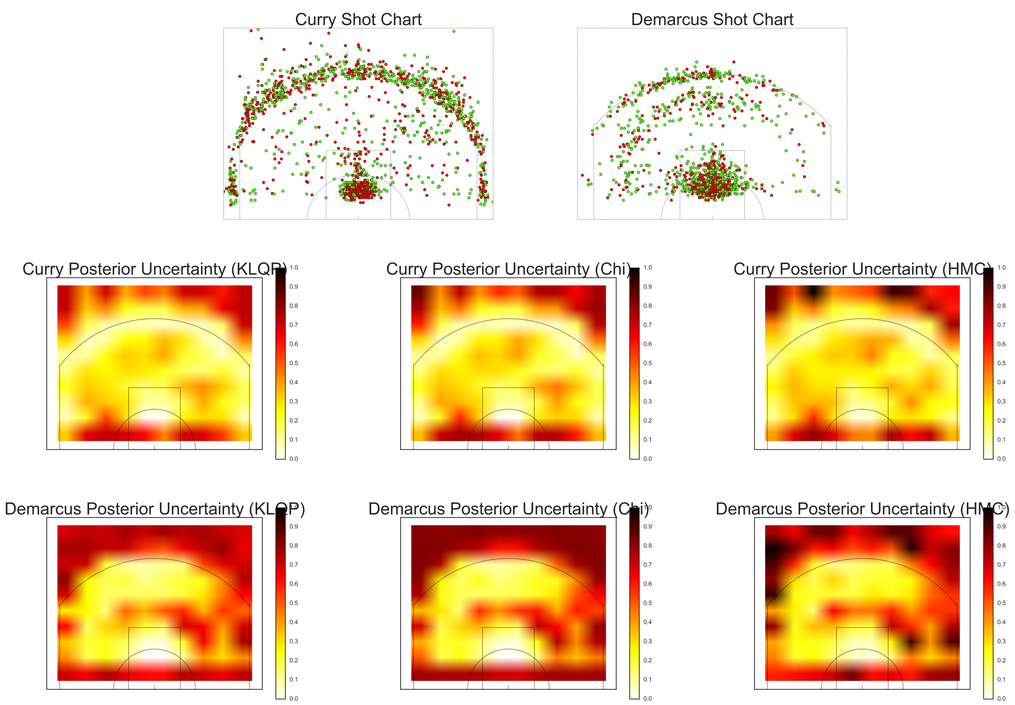
$$\nabla_\phi \hat{\mathcal{L}}(\phi) = \frac{n}{B}\sum_{b=1}^{B}\left(\frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z};\phi)}\right)^n \nabla_\phi \log\left(\frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z};\phi)}\right).$$

- However the exponentiation introduces high variance in the gradients. We use reparameterization gradients to reduce variance by using $\boldsymbol{z} = g(\phi, \epsilon)$ where $\epsilon \sim p(\epsilon)$.
- The resulting black box algorithm CHIVI is detailed below

Algorithm 1: χ-divergence variational inference (CHIVI)

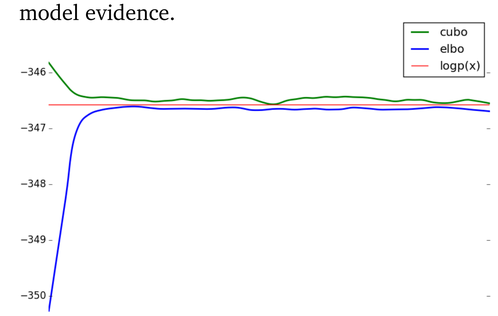
Input: Data $\mathbf{x}$, Model $p(\mathbf{x}, \mathbf{z})$, Variational family $q(\mathbf{z}; \boldsymbol{\lambda})$.
Output: Variational parameters $\boldsymbol{\lambda}$.
Initialize $\boldsymbol{\lambda}$ randomly.
while not converged **do**
  Draw S samples $\mathbf{z}^{(1)}, ..., \mathbf{z}^{(S)}$ from $q(\mathbf{z}; \boldsymbol{\lambda})$ and a data subsample $\{x_{i_1}, ..., x_{i_M}\}$.
  Set ρ_t according to a learning rate schedule.
  Set $\log \mathbf{w}^{(s)} = \log p(\mathbf{z}^{(s)}) + \frac{N}{M}\sum_{j=1}^{M} p(\mathbf{x}_{i_j}|\mathbf{z}) - \log q(\mathbf{z}^{(s)}; \boldsymbol{\lambda}_t)$, $s \in \{1, ..., S\}$.
  Set $\mathbf{w}^{(s)} = \exp(\log \mathbf{w}^{(s)} - \max_s \log \mathbf{w}^{(s)})$, $s \in \{1, ..., S\}$.
  Update $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \frac{(1-n)\cdot\rho_t}{S}\sum_{s=1}^{S}\left[\left(\mathbf{w}^{(s)}\right)^n \nabla_{\boldsymbol{\lambda}} \log q(\mathbf{z}^{(s)}; \boldsymbol{\lambda}_t)\right]$.
end

Better Uncertainty Estimates when Modeling Basketball Plays with Cox Processes



Sandwiching the Log Marginal Likelihood

- The following relationships between ELBO and CUBO hold
  1. $\forall n \geq 1$ $ELBO \leq \log p(\boldsymbol{x}) \leq CUBO_n$.
  2. $\lim_{n\to 0} CUBO_n = \text{ELBO}$.
- This enables black box sandwich estimation of the model evidence.



- This is useful given existing approaches for sandwich estimating the model evidence rely on MCMC [6]

Classification with Gaussian Processes

Settings: We tested CHIVI on GP classification – a model class for which EP has been the method of choice. We chose a factorized Gaussian for the variational approximation. The data are different datasets from the UCI repository. The results are highlighted in Table 1. CHIVI yields lower test error rates when compared to Laplace and EP on most datasets.

Table 1: Test error rate for Gaussian process classification.

Dataset	Laplace	EP	CHIVI
Crabs	**0.02**	**0.02**	0.03 ± 0.03
Sonar	0.154	0.139	**0.055** ± **0.035**
Ionos	0.084	0.08 ± 0.04	**0.069** ± **0.034**

References

[1] R. Ranganath, S. Gerrish, and D. M. Blei. Black box variational inference. In *AISTATS*, 2014.

[2] Y. Li and R. E. Turner. Variational inference with Rényi divergence. In *NIPS*, 2016.

[3] Rajesh Ranganath, Jaan Altosaar, Dustin Tran, and David M. Blei. Operator variational inference. In *NIPS*, 2016.

[4] T. Minka. Expectation propagation for approximate bayesian inference. In *UAI*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

[5] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[6] R. B. Grosse, Z. Ghahramani, and R. P. Adams. Sandwiching the marginal likelihood using bidirectional monte carlo. *arXiv preprint arXiv:1511.02543*, 2015.