SoftCVI: Contrastive variational inference with self-generated soft labels

Daniel Ward[1], Mark Beaumont[2], Matteo Fasiolo[1]

[1]School of Mathematics, Bristol University, UK
[2]School of Biological Sciences, Bristol University, UK

March 13, 2025

Overview

- Variational inference: use optimization to approximate a posterior $q_\phi(\boldsymbol{\theta}) \approx p(\boldsymbol{\theta}|\boldsymbol{x}_{\text{obs}})$.
- SoftCVI reframes variational inference as a contrastive learning problem.
- Leads to stable to train, mass-covering variational objectives.
- More reliable uncertainty quantification!

Setting up the classification problem:

To perform classification, we need three things:

1. Samples
2. Classification labels
3. A classifier

Setting up the classification problem:

To perform classification, we need three things:

1. Samples
2. Classification labels
3. A classifier

SoftCVI

Setup:

- Define proposal $\pi(\boldsymbol{\theta})$
- Define $p^{-}(\boldsymbol{\theta})$
- $p(\boldsymbol{\theta}|\boldsymbol{x}_{\text{obs}})$ is the unknown positive distribution

Setting up the classification problem:

To perform classification, we need three things:

1. Samples
2. Classification labels
3. A classifier

SoftCVI

Setup:

- Define proposal $\pi(\boldsymbol{\theta})$
- Define $p^{-}(\boldsymbol{\theta})$
- $p(\boldsymbol{\theta}|\boldsymbol{x}_{\text{obs}})$ is the unknown positive distribution

Optimization:

1. Sample $\{\boldsymbol{\theta}_k\}_{k=1}^{K} \sim \pi(\boldsymbol{\theta})$.
2. Analytically assign soft classification labels
3. Optimize a classifier parameterized using $q_{\phi}(\boldsymbol{\theta})$

Labels and predictions

Given a set of K samples, $\{\boldsymbol{\theta}_k\}_{k=1}^K \sim \pi(\boldsymbol{\theta})$

Computing labels

$$y_k = \frac{p(\boldsymbol{\theta}_k, \boldsymbol{x}_{\text{obs}})/p^-(\boldsymbol{\theta}_k)}{\sum_{k'=1}^K p(\boldsymbol{\theta}_{k'}, \boldsymbol{x}_{\text{obs}})/p^-(\boldsymbol{\theta}_{k'})},$$

Computing predictions

$$\hat{y}_k = \frac{q_\phi(\boldsymbol{\theta}_k)/p^-(\boldsymbol{\theta}_k)}{\sum_{k'=1}^K q_\phi(\boldsymbol{\theta}_{k'})/p^-(\boldsymbol{\theta}_{k'})},$$

Labels and predictions

Given a set of K samples, $\{\boldsymbol{\theta}_k\}_{k=1}^K \sim \pi(\boldsymbol{\theta})$

Computing labels

$$y_k = \frac{p(\boldsymbol{\theta}_k, \boldsymbol{x}_{\text{obs}})/p^-(\boldsymbol{\theta}_k)}{\sum_{k'=1}^K p(\boldsymbol{\theta}_{k'}, \boldsymbol{x}_{\text{obs}})/p^-(\boldsymbol{\theta}_{k'})},$$

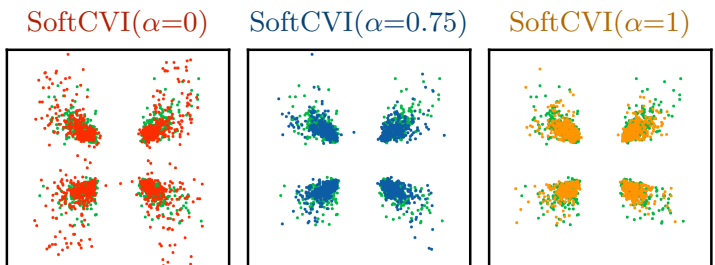
Computing predictions

$$\hat{y}_k = \frac{q_\phi(\boldsymbol{\theta}_k)/p^-(\boldsymbol{\theta}_k)}{\sum_{k'=1}^K q_\phi(\boldsymbol{\theta}_{k'})/p^-(\boldsymbol{\theta}_{k'})},$$

- Use the softmax cross-entropy loss.

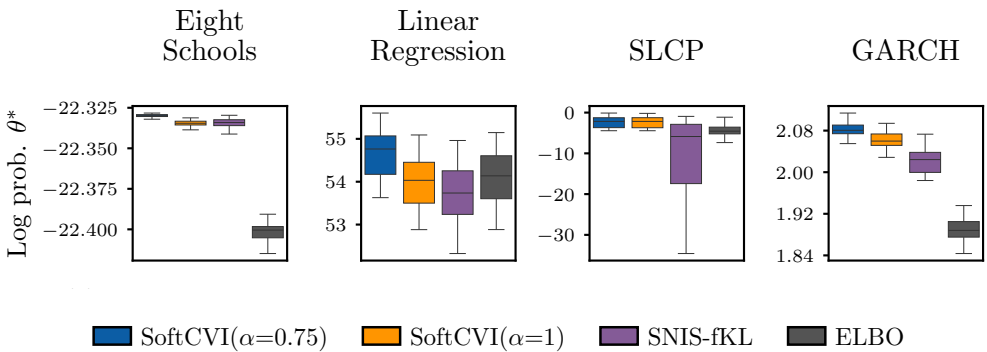
Choices:

- Proposal distribution: $\pi(\boldsymbol{\theta}) = q_{\phi}(\boldsymbol{\theta})$
- Negative distribution: $p^{-}(\boldsymbol{\theta}) = q_{\phi}(\boldsymbol{\theta})^{\alpha}$, where $\alpha \in [0, 1]$
- The negative distribution choice influences the objective properties



Emprical results

Places more mass on ground truth posterior samples:



In the paper

- Gradient variance tends to zero as $q_\phi(\boldsymbol{\theta})$ approaches $p(\boldsymbol{\theta}|\boldsymbol{x}_{\text{obs}})$.
- Has a lower variance gradient approximator for SNIS-fKL as a special case [Jerfel et al., 2021].
- Deep learning examples.

Jerfel, G., Wang, S., Wong-Fannjiang, C., Heller, K. A., Ma, Y., and Jordan, M. I. (2021).
Variational refinement for importance sampling using the forward kullback-leibler divergence.
In *Uncertainty in Artificial Intelligence*, pages 1819–1829. PMLR.