Empirical Risk Reduction (I) [%]

Goal: Improving Variational Approximations for Predictive Tasks

- Why: The posterior distribution is sufficient for making optimal decisions in down-stream tasks, but approximate posteriors are not.
- What: We calibrate variational approximations to improve decisions, by accounting for the decision task already during inference.
- Outcome: First practical solution for prediction tasks with continuous utilities, with systematic improvement in expected utility.

Preliminaries

Bayesian Decision Theory
- Posterior \( p(\theta | D) \) sufficient for optimal decisions \( h_{\theta} \).
- Maximize the gain \( J \)

\[ J = \int p(\theta | D) \hat{u}(\theta, h) d\theta \]

where \( \hat{u}(\theta, h) \) is the utility.
- For predictive problems \( \hat{u}(\theta, h) = \int p(y|\theta, D) u(y, h) dy \).
- Closed-forms available for some utilities.

Variational Inference
- Approximate the posterior \( p(\theta | D) \) with \( q(\theta | \lambda) \) parameterized by \( \lambda \).
- Maximize a lower bound \( \mathbb{E}_q \mathbb{L}_\lambda(\lambda) \) for the marginal log-likelihood \( \log p(D) \)

\[ \log p(D) \geq \int q(\theta) \log \frac{p(D, \theta)}{q(\theta)} d\theta = \mathbb{L}_\lambda(\lambda) \]
- Gradient-based optimization via reparameterization of the approximation and Monte Carlo integration.

Loss Calibrated Variational Inference – LCVI

General Framework
Bound the logarithmic gain using Jensen's inequality [2]

\[ \mathbb{E}_q [\log p(y|\theta, D)u(y, h)dy] = : \mathbb{L}_{LCVI}(\lambda, h) \]

- Reparameterization of both the approximation \( q(\theta | \lambda) \) and the predictive distribution \( p(y|\theta, D) \).
- Joint gradient-based optimization of \( h \) and \( \lambda \).
- Calibration maximized for utilities with \( \int u_{\lambda}u(y, h) = 0 \).

Utilities and Losses
- Losses \( l(y, h) \) need to be first converted into utilities \( u(y, h) \).
- Problem: \( u(y, h) = M - l(y, h) \) does not change optimal decisions, but requires \( M = \sup_y u(y, h) \). Large \( M \) reduces calibration.

- Solutions:
  1. Linearize \( u(y, h) \) and use \( M_q \) that is the \( q \)th quantile of the loss distribution
  2. Use \( exp(-\frac{M_q}{M_q}) \) to approximately retain the decisions.

Experiments
- Bayesian matrix factorization on the Last.fm dataset.
- We measure empirical risk reduction on test data.

\[ J = \frac{ER_{\text{ALG}} - ER_{\text{LCVI}}}{ER_{\theta}} \]

- LCVI outperforms VI on different losses.
- Joint optimization achieves better results than alternating optimization.

References