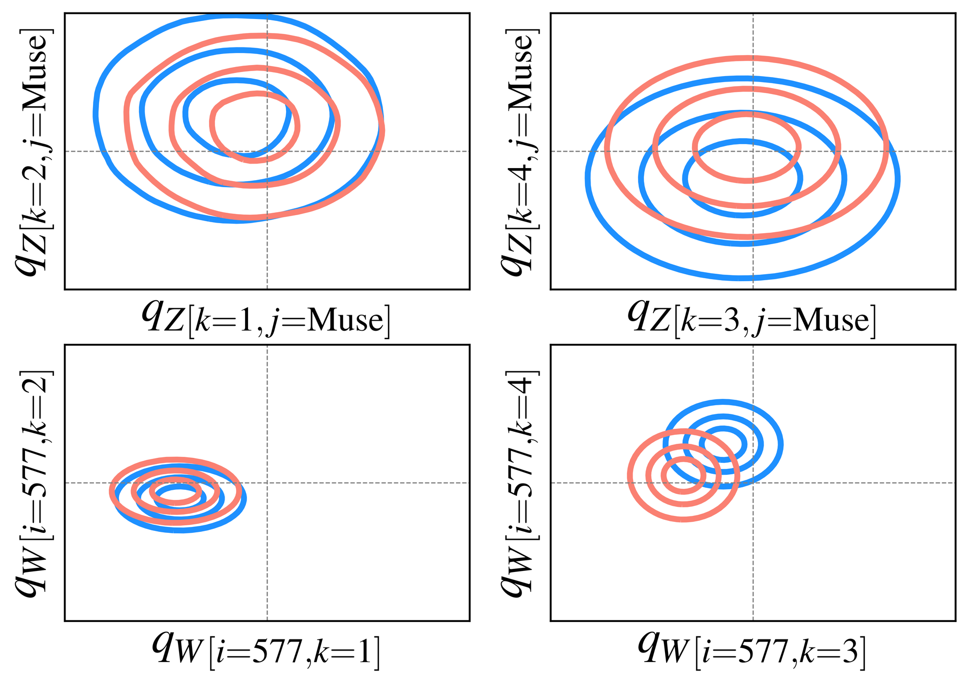
Variational Bayesian Decision-making for Continuous Utilities

Tomasz Kuśmierczyk Joseph Sakaya Arto Klami

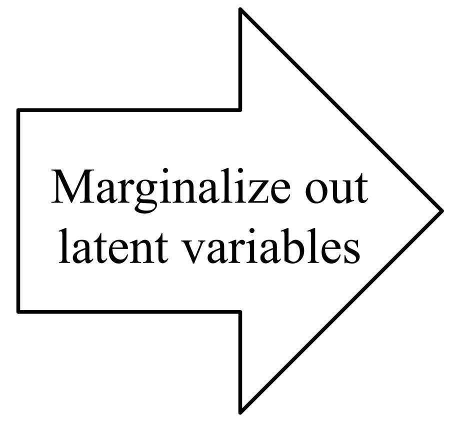
Helsinki Institute for Information Technology HIIT
Department of Computer Science, University of Helsinki

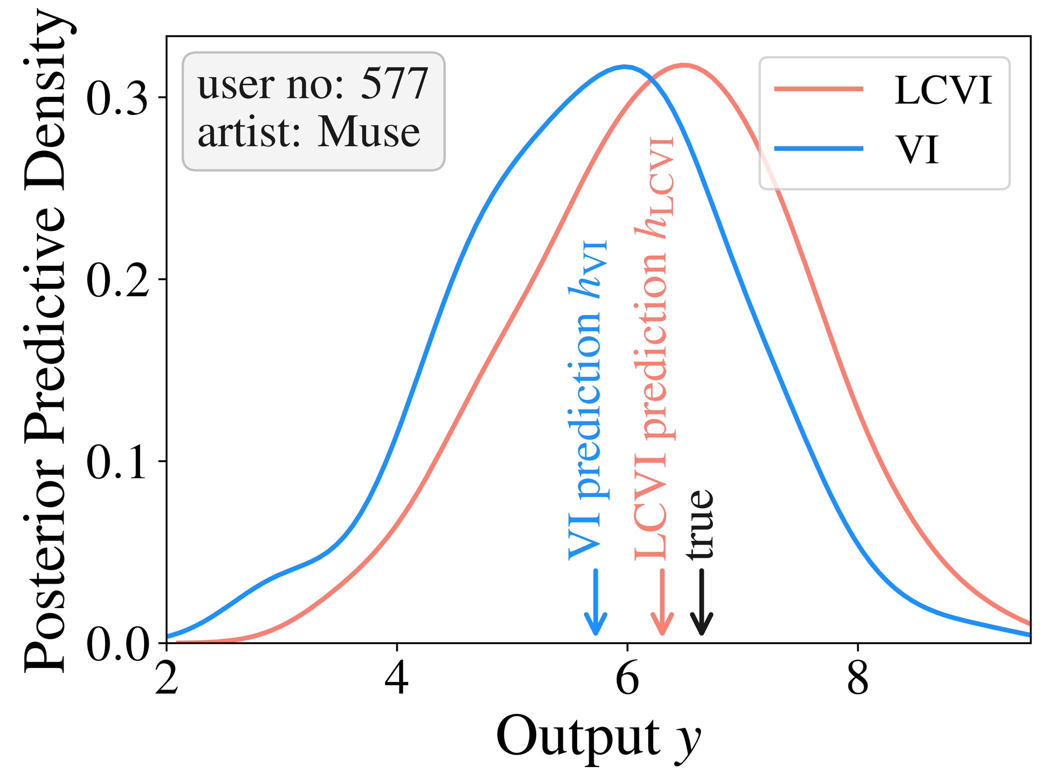
Goal: Improving Variational Approximations for Predictive Tasks

Modify the posterior approximation *to* improve given predictive utility



Approximate Posterior Densities





- **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 $p(\theta|\mathcal{D})$ sufficient for optimal decisions h_p
- Maximize the *gain* [1]

$$\mathcal{G}_u(h) = \int p(\theta|\mathcal{D})\tilde{u}(\theta,h)d\theta,$$

where $\tilde{u}(\theta,h)$ is the *utility*

- For predictive problems $\tilde{u}(\theta,h) = \int p(y|\theta,\mathcal{D})u(y,h)dy$
- Closed-form decisions available for some utilities

Variational Inference

- Approximate the posterior $p(\theta|\mathcal{D})$ with $q_\lambda(\theta)$ parameterized by λ
- Maximize a lower bound $\mathcal{L}_{\mathrm{VI}}(\lambda)$ for the marginal log-likelihood

$$\log p(\mathcal{D}) \geq \int q_\lambda(\theta)\log\frac{p(\mathcal{D},\theta)}{q_\lambda(\theta)}d\theta =: \mathcal{L}_{\mathrm{VI}}(\lambda)$$

- 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]

$$\log\mathcal{G}_u(h) \geq \mathcal{L}_{\mathrm{VI}}(\lambda) + \underbrace{\mathbb{E}_q[\log\int p(y|\theta,\mathcal{D})u(y,h)dy]}_{\mathbb{U}(\lambda,h)} =: \mathcal{L}_{\mathrm{LCVI}}(\lambda,h)$$

- Reparameterization of both the approximation $q_\lambda(\theta)$ [3] and the predictive distribution $p(y|\theta,\mathcal{D})$
- Joint gradient-based optimization of h and λ
- Calibration maximized for utilities with $\inf_{y,h} u(y,h) = 0$

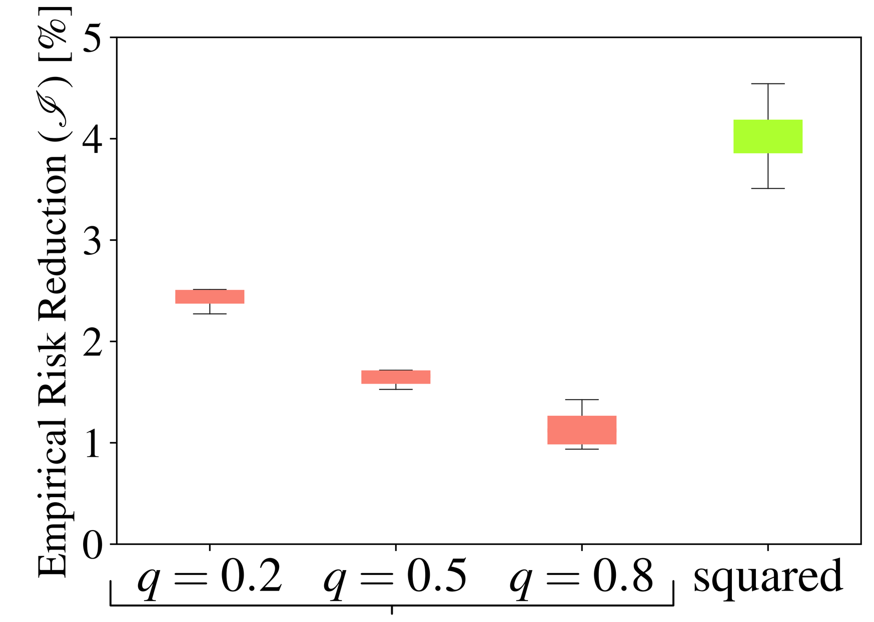
Utilities and Losses

- Losses $\ell(y,h)$ need to be first converted into utilities $u(y,h)$
- **Problem:** $u(y,h) = M - \ell(y,h)$ does not change optimal decisions, but requires $M = \sup_{y,h}\ell(y,h)$. Large M reduces calibration
- **Solutions:**
  1. Linearize $u(y,h)$ and use M_q that is the qth quantile of the loss distribution
  2. Use $\exp\left(-\frac{\ell(y,h)}{M_q}\right)$ to approximately retain the decisions

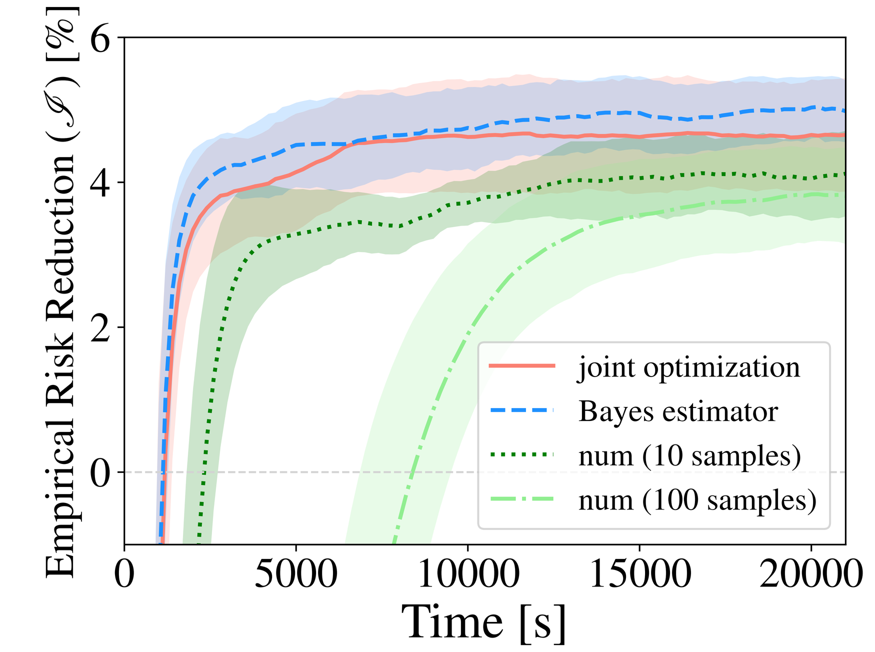
Experiments

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

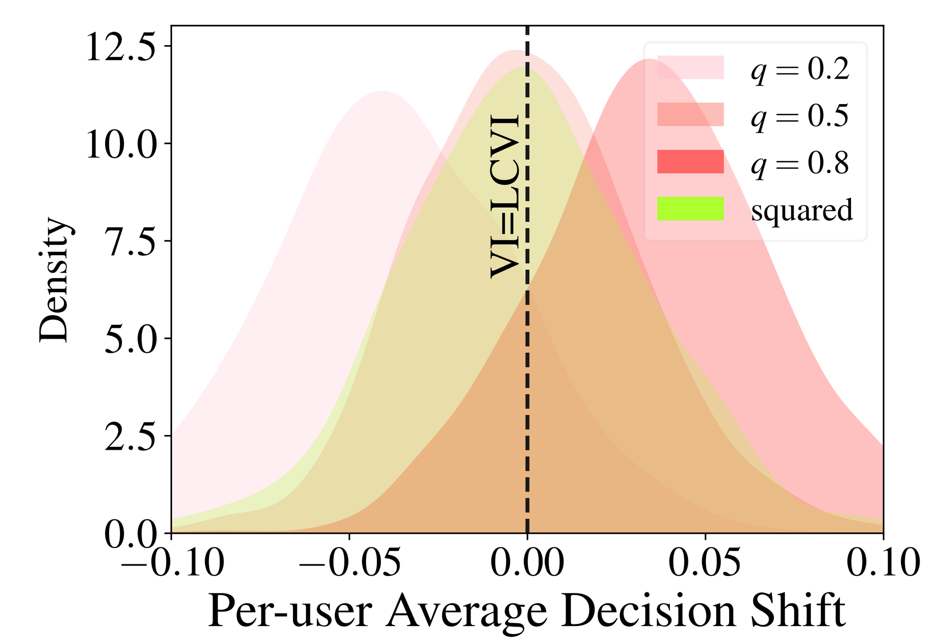
$$\mathcal{I} = \frac{\mathcal{ER}_{\mathrm{VI}} - \mathcal{ER}_{\mathrm{LCVI}}}{\mathcal{ER}_{\mathrm{VI}}}, \quad \mathcal{ER}_{\mathrm{ALG}} = \frac{1}{|\mathcal{D}_{\mathrm{test}}|}\sum_{i\in[\mathcal{D}_{\mathrm{test}}]}\ell(y_i, h_i^{ALG}).$$



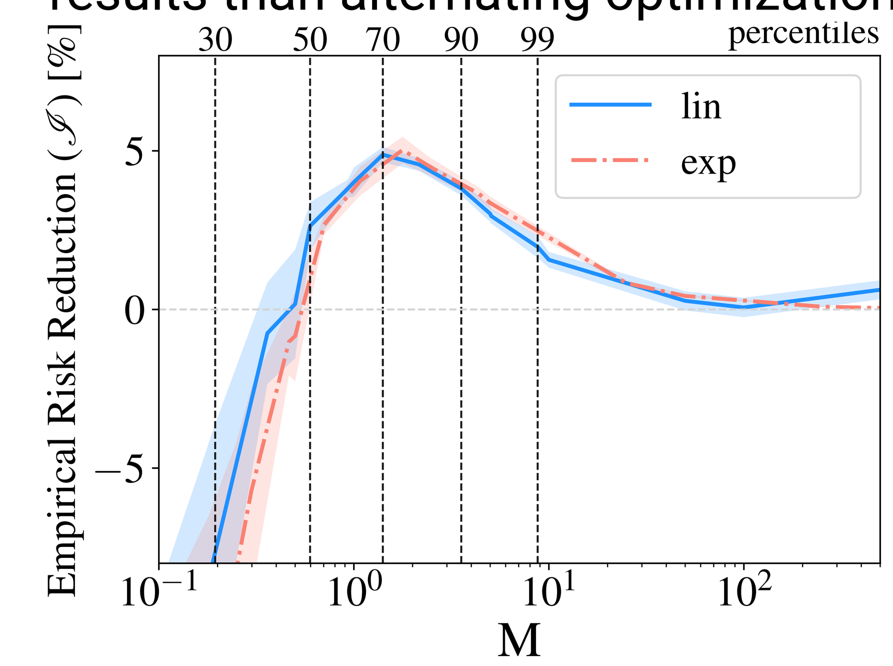
LCVI outperforms VI on different losses



Joint optimization achieves better results than alternating optimization



LCVI changes the decisions in a non-trivial manner



$M_q = 70\%$ quantile achieves optimal calibration

References

[1] James O Berger. Statistical Decision Theory and Bayesian Analysis; 2nd edition. Springer Series in Statistics. Springer, New York, 1985.

[2] Simon Lacoste-Julien, Ferenc Huszár, and Zoubin Ghahramani. Approximate inference for the loss-calibrated Bayesian. In ICML, 2011.

[3] Adam D Cobb, Stephen J Roberts, and Yarin Gal. Loss-calibrated Approximate Inference in Bayesian Neural Networks. In ICML Workshop on Theory of Deep Learning workshop, 2018.