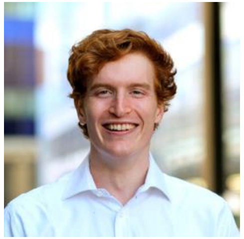
UTN Technische Universität Nürnberg · HELMHOLTZ AI Artificial Intelligence Cooperation Unit · MCML Munich Center for Machine Learning · TUM · UCL · HELMHOLTZ MUNICH

Gaussian Mean Field Variational Inference can Overestimate Predictive Variance

James Odgers Ben Riegler Siddharth Swaroop Vincent Fortuin



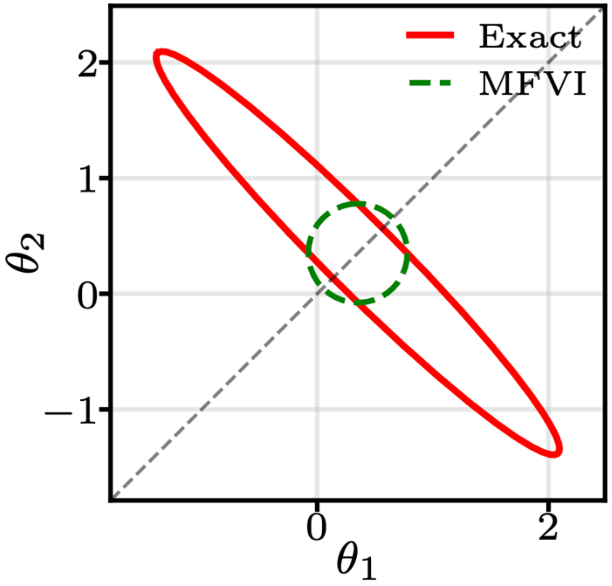

Intro

- Mean Field Variational Inference (MFVI) is a technique to approximate the exact Bayesian Posterior

$$q^*(\theta) = \arg\min_{q \in \mathcal{Q}} D_{KL}(q(\theta) || p(\theta|D))$$

- MFVI is known to **underestimate uncertainty in the parameter space**

- We show that (for conjugate models) **MFVI predictions overestimate predictive uncertainty on in distribution data**



Setup

- We study **conjugate Bayesian linear regression with an isotropic Gaussian prior** and **approximate the posterior with a mean field Gaussian**

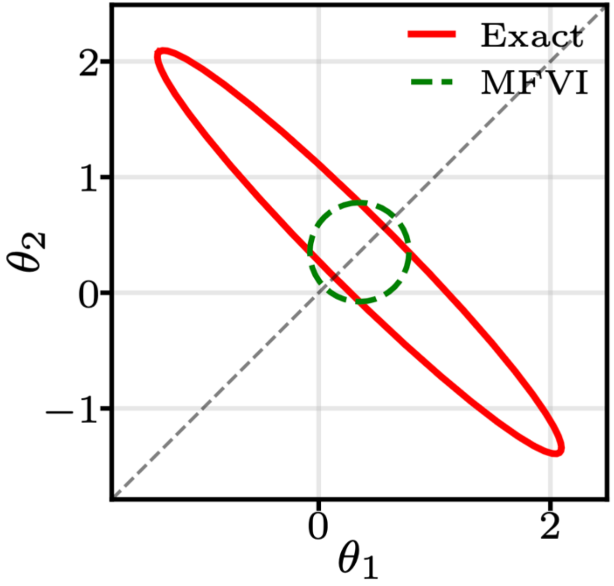
- We are interested in **comparing the predictions** from the exact posterior and MFVI posterior for **different distributions of test points**

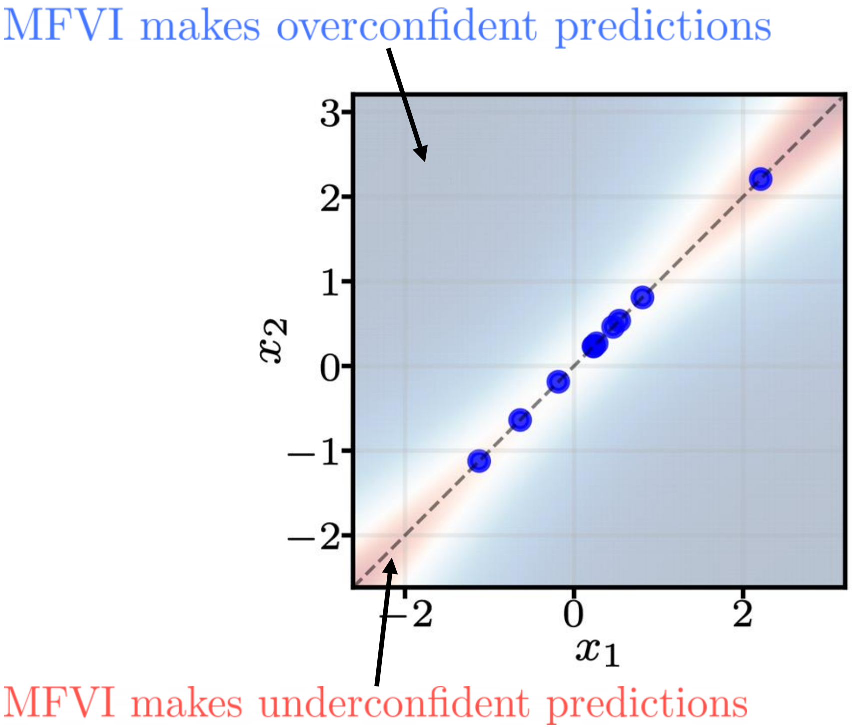
$$p(f|D) = N(\mu^\top x, x^\top \Sigma x)$$
$$q^*(f) = N(m^* x, x^\top S^* x)$$

- The mean of the MFVI and exact posterior are equal, so **we only need to compare the predicted variance**

Example: Low Rank Linear Regression

Consider a 2D regression, where the data is in a 1D subspace





Theory

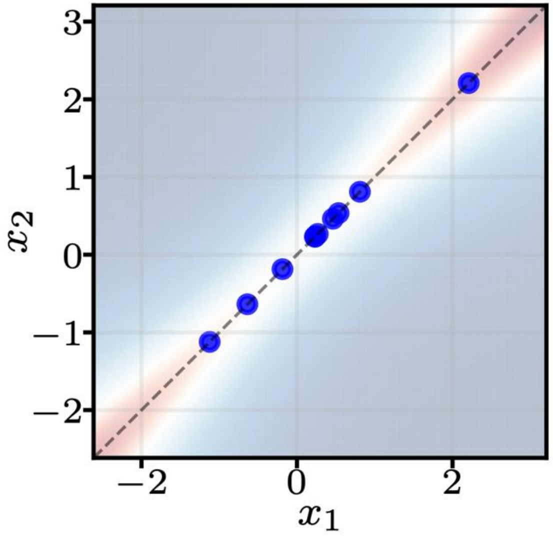
- We show theoretically that **this behavior is true for all input point distributions**

Theorem 3.6. *For test data points distributed according to the empirical distribution* $x \sim \hat{p}(x)$*, the difference in the expected predicted variance of the MFVI and exact posterior is given by*

$$\mathbb{E}_{x \sim \hat{p}(x)} \left[x^\top \Sigma x - x^\top S^* x \right] \leq 0$$

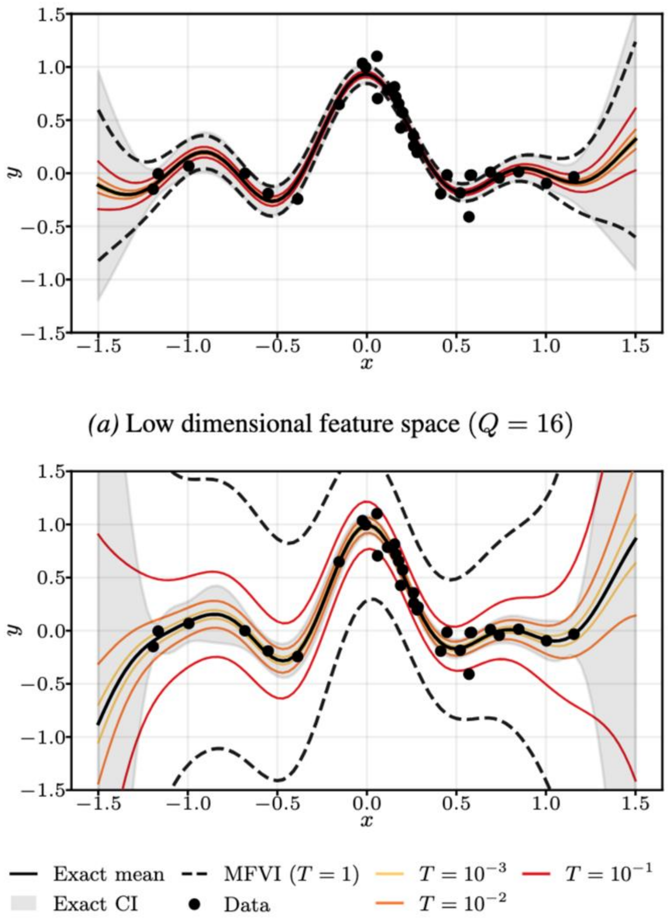
Exact posterior predictive variance

MFVI posterior predictive variance



Cold Posterior Effects in Linear Regression

- In **high dimensions** the variance overestimation can be large
- This can be partially **corrected by applying a Cold Posterior**
- We conjecture that **the deep learning setting behaves similarly to our linear setting**



(a) Low dimensional feature space ($Q = 16$)

(b) High dimensional feature space ($Q = 1024$)

Come to the poster or read the paper

- We provide **more theoretical results** about underconfident predictions in MFVI
- We provide **more experimental validation** for linear regression
- We discuss **conceptual similarities to deep learning**