Inference and validation of post-Bayesian methods

Zheyang Shen
Newcastle University

Lancaster University: CSML talk

Papers

- *Prediction-centric uncertainty quantification via MMD*. Z. Shen, J. Knoblauch, S. Power, C. J. Oates.
- *A computable measure of suboptimality for entropy-regularised variational objectives*. C. Chazal, H. Kanagawa, Z. Shen, A. Korba, C. J. Oates.
- *Predictively oriented posteriors*. Y. McLatchie, B.-E. Cherief-Abdellatif, D. T. Frazier, J. Knoblauch.

Optimization-centric probabilistic inference

Bayesian posterior and its generalization are minimizers of an *entropy-regularized objective*

$$P := \operatorname*{arg\,min}_{Q \in \mathcal{P}(\mathbb{R}^d)} \mathcal{J}(Q), \qquad \mathcal{J}(Q) := \mathcal{L}(Q) + \mathrm{KL}(Q || Q_0).$$

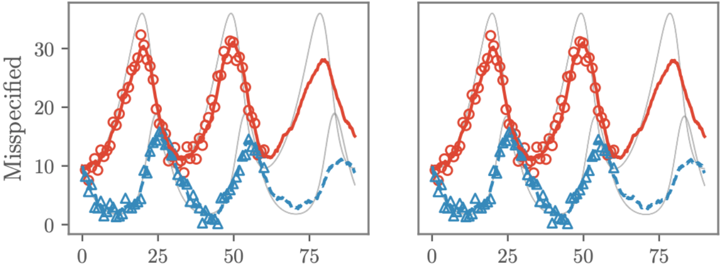
Examples of $\mathcal{L}(Q)$

- Standard Bayes: $-\sum_{i=1}^{n} \int \log p(y_i|\theta) \, \mathrm{d}Q(\theta)$.
- Generalized Bayes[1]: $\sum_{i=1}^{n} \int \ell(y_i, \theta) \, \mathrm{d}Q(\theta)$.

[1] Bissiri et al. [2016]

Why post-Bayes?

- Bayesian posterior in a misspecified model lack in parameter uncertainty;
- Generalized Bayes confronts aspects of model misspecification (e.g., outlier robustness), yet is still overconfident.



Lotka-Volterra example[1]

[1] Shen et al. [2025]

Predictively-oriented (PrO) posteriors[1]

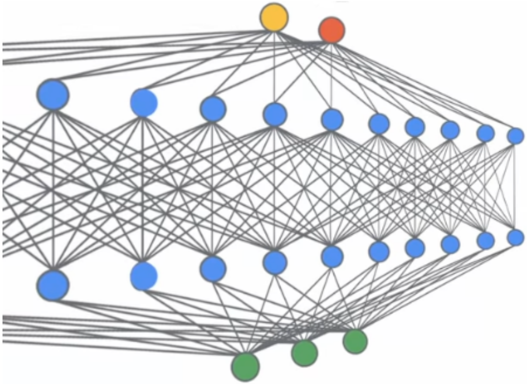
Generalized Bayes	Predictively-oriented
$\underbrace{\sum_{i=1}^{n} \int S(y_i, P_\theta)\, \mathrm{d}Q(\theta)}_{\text{average fit}}$	$\underbrace{\sum_{i=1}^{n} S\left(y_i, \int P_\theta\, \mathrm{d}Q(\theta)\right)}_{\text{predictive fit}}$

What happens when the model is well-specified?

McLatchie et al. [2025] illustrate that the PrO posterior still concentrates around the data-generating distribution (Theorem 1).

[1] McLatchie et al. [2025]

Mean field neural networks (MFNNs)



The output of MFNN is "infinitely wide"

$$f_Q(x) = \int \Phi(x, \theta) \, \mathrm{d}Q(\theta)$$

Training loss is a nonlinear $\mathcal{L}(Q)$:

$$\mathcal{L}(Q) = \sum_{i=1}^{n} \ell\left(y_i, f_Q(x_i)\right).$$

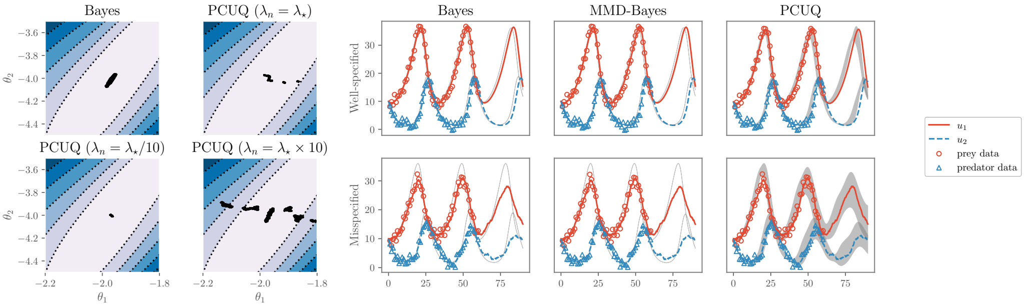
Challenges of a nonlinear $\mathcal{L}(Q)$

$$P := \underset{Q \in \mathcal{P}(\mathbb{R}^d)}{\arg\min}\ \mathcal{L}(Q) + \mathrm{KL}(Q||Q_0).$$

- P is identifiable up to normalization when $\mathcal{L}(Q)$ is linear: applies to (generalized) Bayes;
- The nonlinearity of $\mathcal{L}(Q)$ makes it so that P is only identifiable via optimization: mean-field Langevin dynamics evolve a set of interacting particles $Q_N^t := \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i^t}$:

$$\theta_i^{t+1} = \theta_i^t + \epsilon[\nabla \log q_0(\theta_i^t) - \underbrace{\nabla_{\mathrm{V}}\mathcal{L}(Q_N^t)(\theta_i^t)}_{\text{variational gradient}}] + \sqrt{2\epsilon}Z_i^t.$$

Given a sample-based approximation $Q_N := \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$, *is it any good?*



Bayes
PCUQ ($\lambda_n = \lambda_\star$)
PCUQ ($\lambda_n = \lambda_\star/10$)
PCUQ ($\lambda_n = \lambda_\star \times 10$)
θ_1
θ_2
Bayes
MMD-Bayes
PCUQ
Well-specified
Misspecified
u_1
u_2
prey data
predator data

Revisiting Stein's identity

We would like to generalize Stein's discrepancy to validate how well a set of particles satisfies the minimizer of $\mathcal{J}(Q)$.

$$\forall \phi \in \mathcal{F}, \quad \mathbb{E}_{x\sim Q}\left[\langle \nabla \log p(x), \phi(x)\rangle + \langle \nabla, \phi(x)\rangle\right] = 0 \quad \Leftrightarrow \quad Q = P.$$

Noting that $\mathcal{J}_{\mathrm{Bayes}}(Q) = \mathrm{KL}(Q\|P)$, we have:

$$\begin{aligned}
&\mathbb{E}_{x\sim Q}\left[\langle \nabla \log p(x), \phi(x)\rangle + \langle \nabla, \phi(x)\rangle\right] = \mathbb{E}_{x\sim Q}\left[\langle \nabla \log p(x) - \nabla \log q(x), \phi(x)\rangle\right] \\
=&\mathbb{E}_{x\sim Q}\left[\langle -\nabla_V \mathcal{J}_{\mathrm{Bayes}}(Q)(x), \phi(x)\rangle\right] = -\left.\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{J}_{\mathrm{Bayes}}(Q_t^{\phi})\right|_{t=0}.
\end{aligned}$$

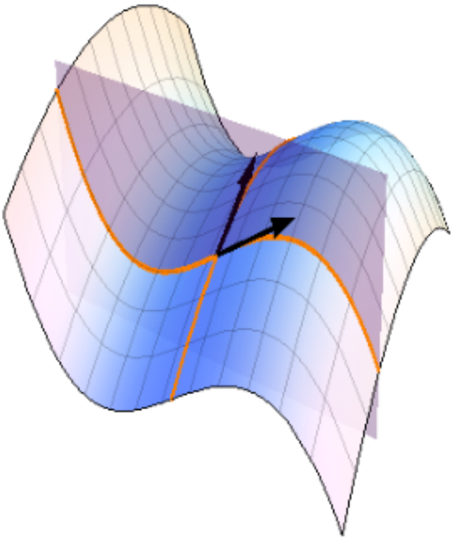
Drawing any geodesic curve $\left(Q_t^{\phi}\right)_{t\in(-\epsilon,\epsilon)}$ around Q, its time derivative w.r.t. $\mathcal{J}_{\mathrm{Bayes}}(Q_t)$ is always zero.

Measuring approximation quality

Idea: See "how well Q minimises $\mathcal{J}$"

Concretely: If Q does **not** minimise $\mathcal{J}$ then there is some "direction" ϕ such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{J}(Q_t^{\phi})\bigg|_{t=0} < 0.$$



Variational Gradients

Variational gradients: From the fundamental theorem of calculus

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{J}(Q_t^{\phi})\Big|_{t=0} = \int \langle \nabla_{\mathrm{V}}\mathcal{J}(Q)(x), \phi(x)\rangle \,\mathrm{d}Q(x).$$

where the *variational gradient* is $\nabla_{\mathrm{V}}\mathcal{J}(Q)(x) := \nabla_x \mathcal{J}'(Q)(x)$ for each $x \in \mathbb{R}^d$.

(the *first variation* $\mathcal{F}'(Q)$ is defined as $\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathcal{F}(Q+\epsilon\chi)|_{\epsilon=0} = \int \mathcal{F}'(Q)\,\mathrm{d}\chi$)

Computing the variational gradient of $\mathcal{J}$: Letting Q_0 have a density $q_0 > 0$,

$$\nabla_{\mathrm{V}}\mathcal{J}(Q)(\theta) = \nabla_{\mathrm{V}}\mathcal{L}(Q)(\theta) - (\nabla \log q_0)(\theta) + \underbrace{(\nabla \log q)(\theta)}_{\text{problematic}}$$

Main idea: Can still evaluate integrals of the variational gradient:

$$\int \underbrace{(\nabla \log q)(x)}_{\text{problematic}} \cdot \phi(x)\,\mathrm{d}Q(x) = \underbrace{-\int (\nabla \cdot \phi)(x)\,\mathrm{d}Q(x).}_{\text{fine}}$$

Gradient Discrepancy

Gradient discrepancy: For a given set $\mathcal{F}$ of differentiable vector fields on $\mathbb{R}^d$, define the *gradient discrepancy* as

$$\mathrm{GD}(Q) := \sup_{\substack{\phi \in \mathcal{F} \text{ s.t.} \\ (\mathcal{T}_Q \phi)_- \in \mathcal{L}^1(Q)}} \left| \int \mathcal{T}_Q \phi(x) \, \mathrm{d}Q(x) \right|$$

where $\mathcal{T}_Q \phi(x) := [(\nabla \log q_0)(x) - \nabla_{\mathrm{V}} \mathcal{L}(Q)(x)] \cdot \phi(x) + (\nabla \cdot \phi)(x)$.

Example: For $\mathcal{L}'(Q)(x) = -\log p(y|x)$, we recover *(Langevin) Stein discrepancy* [Gorham and Mackey, 2015]

$$\mathrm{SD}(Q) := \sup_{\substack{\phi \in \mathcal{F} \text{ s.t.} \\ (\mathcal{S}_P \phi)_- \in \mathcal{L}^1(Q)}} \left| \int \mathcal{S}_P \phi(x) \, \mathrm{d}Q(x) \right|$$

where $\mathcal{S}_P \phi(x) := (\nabla \log p)(x) \cdot \phi(x) + (\nabla \cdot v)(x)$, where $p(x) \propto q_0(x) p(y|x)$ is a density for P.

$\therefore$ **Stein discrepancy is measuring the size of a variational gradient**

Computing gradient discrepancy

Kernel gradient discrepancy: Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel with associated Hilbert space $\mathcal{H}_k$. The Kernel Gradient Discrepancy (KGD) is defined as

$$\mathrm{KGD}_k(Q) := \sup_{\substack{\|\phi\|_{\mathcal{H}_k^d} \leq 1 \text{ s.t.} \\ (\mathcal{T}_Q \phi)_- \in \mathcal{L}^1(Q)}} \left| \int \mathcal{T}_Q \phi(x) \, \mathrm{d}Q(x) \right|.$$

$\therefore$ **KGD generalises kernel Stein discrepancy (KSD) to nonlinear $\mathcal{L}$**

Computable form of KGD: Let $s(Q)(\theta) := (\nabla \log q_0)(\theta) - \nabla_{\mathrm{V}} \mathcal{L}(Q)(\theta)$. Then

$$\mathrm{KGD}_k(Q) = \left(\iint k_Q(x, x') \, \mathrm{d}Q(x) \mathrm{d}Q(x') \right)^{1/2}$$

with the Q-dependent kernel

$$\begin{aligned} k_Q(x, x') := \nabla_1 \cdot \nabla_2 k(x, x') + \nabla_1 k(x, x') \cdot s(Q)(x') \\ + \nabla_2 k(x, x') \cdot s(Q)(x) + k(x, x') s(Q)(x) \cdot s(Q)(x'). \end{aligned}$$

Measuring approximation quality via KGD

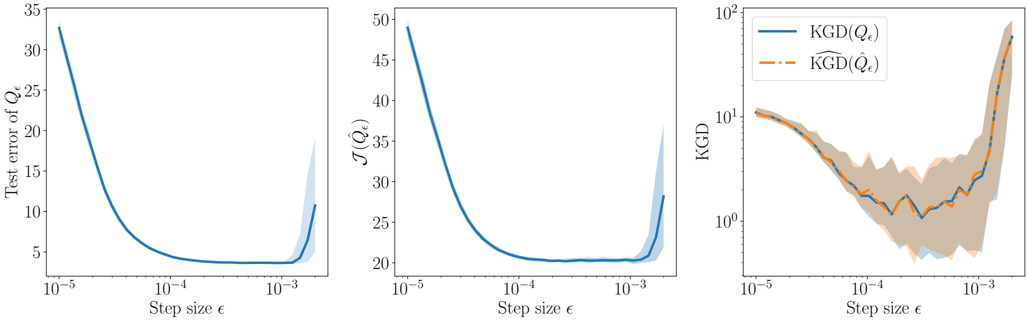


Figure: Selecting the step size ϵ in mean field Langevin dynamics for training a mean field neural network.

Rather than use MFLD, why not directly optimise KGD?

New Algorithms Based on KGD (I)

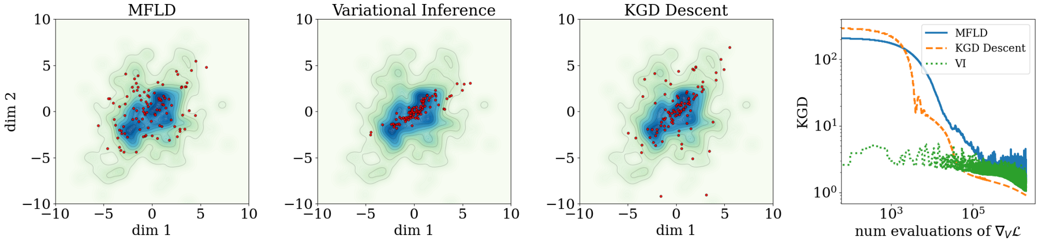


Figure: Comparing MFLD with new algorithms based on KGD, in the setting of training mean field neural networks.

Variational Gradient Descent

Steepest descent: Following Wang and Liu [2019], pick the vector field ϕ_Q from a vector-valued reproducing kernel Hilbert space $\mathcal{H}_k^d$ corresponding to steepest descent:

$$\phi_Q(\cdot) \propto \int \{k(x,\cdot)(\nabla \log q_0 - \nabla_{\mathrm{V}} \mathcal{L}(Q))(x) + \nabla_1 k(x,\cdot)\} \,\mathrm{d}Q(x),$$

Variational gradient descent: Initialise $\{x_i^0\}_{i=1}^N$ as independent samples from μ_0 at time $t = 0$ and then update $\{x_i^t\}_{i=1}^N$ deterministically, via the coupled system of ODEs

$$\frac{\mathrm{d}x_k^t}{\mathrm{d}t} = \frac{1}{N}\sum_{j=1}^N k(x_i^t, x_j^t)(\nabla \log q_0 - \nabla_{\mathrm{V}} \mathcal{L}(Q_N^t))(x_j^t) + \nabla_1 k(x_j^t, x_i^t), \qquad Q_N^t := \frac{1}{N}\sum_{j=1}^N \delta_{x_j^t}$$

up to a time horizon T.

New Algorithms Based on KGD (II)

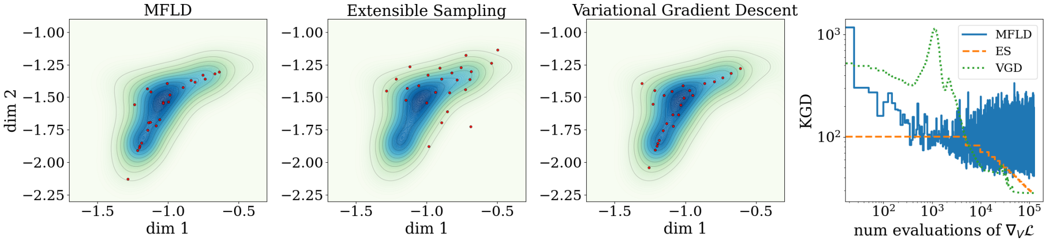


Figure: Comparing MFLD with new algorithms based on KGD, in the setting of prediction-centric uncertainty quantification.

Properties of Kernel Gradient Discrepancy

Informal Theorem (Properties of KGD)

One can pick the kernel k such that the following hold:

- **Identification:** $\mathrm{KGD}_k(Q) = 0$ *iff Q a stationary point of $\mathcal{J}$*
  (roughly: convexity of $\mathcal{L}$ implies a unique stationary point P of $\mathcal{J}$)
- **Continuity:** $Q_n \xrightarrow{\alpha} Q$ *implies* $\mathrm{KGD}_k(Q_n) \to \mathrm{KGD}_k(Q)$
  ($Q_n \xrightarrow{\alpha} Q$ means $\int f \, \mathrm{d}Q_n \to \int f \, \mathrm{d}Q$ for all $f(\theta) \lesssim 1 + \|\theta\|^\alpha$)
- **Convergence control:** $\mathrm{KGD}_k(Q_n) \to 0$ *implies* $Q_n \xrightarrow{\alpha} P$
  (requires a dissipativity condition on P)

$\therefore$ **minimisation of KGD leads to consistent approximation of the target**

Summary

In a nutshell:

- (kernel) gradient discrepancy (KGD) enables approximation quality to be measured...
- ... and unlocks new classes of algorithms for $\arg\min \mathcal{J}$
- can be considered a **nonlinear** generalisation of kernel Stein discrepancy (KSD)
- sheds light on KSD as measuring the size of a variational gradient

Open questions:

- All the usual challenges apply, e.g. high-dimensions, manifold-constrained targets, computational efficiency, mode collapse, etc.
- Stein's identity generalizes to the diffusion Stein operator [Gorham et al., 2019] – what would be the corresponding nonlinear generalization?

Thank you for your attention!

References I

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, 2016.

J. Gorham and L. Mackey. Measuring sample quality with Stein's method. *Advances in Neural Information Processing Systems*, 28, 2015.

J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928, 2019.

Y. McLatchie, B.-E. Cherief-Abdellatif, D. T. Frazier, and J. Knoblauch. Predictively oriented posteriors. *arXiv preprint arXiv:2510.01915*, 2025.

Z. Shen, J. Knoblauch, S. Power, and C. J. Oates. Prediction-centric uncertainty quantification via MMD. In *International Conference on Artificial Intelligence and Statistics*, pages 649–657. PMLR, 2025.

D. Wang and Q. Liu. Nonlinear Stein variational gradient descent for learning diversified mixture models. In *International Conference on Machine Learning*, pages 6576–6585. PMLR, 2019.