GVI in Function Spaces

Gaussian Measures meet Bayesian Deep Learning

Veit D. Wild* , Robert Hu* and Dino Sejdinovic

Department of Statistics

25th of May 2022

UNIVERSITY OF OXFORD

Outline

1. Background
   - Bayesian Deep Learning
   - Variational Inference in Function Spaces
   - Generalised Variational Inference
   - Gaussian Measures on Hilbert Spaces
2. Gaussian Wasserstein Inference
   - Model description
   - Parameterisation of GWI
3. Experiments

Contents

UNIVERSITY OF OXFORD

Bayesian Deep Learning

Bayesian Deep Learning

Let $\mathcal{D} := \{(x_n, y_n) \mid n = 1, \ldots, N\} \subset \mathcal{X} \times \mathcal{Y}$ be data.

Bayesian Deep Learning

Let $\mathcal{D} := \{(x_n, y_n) \mid n = 1, \ldots, N\} \subset \mathcal{X} \times \mathcal{Y}$ be data.

- Supervised Deep Learning:

$$Y = f(x) + \mathcal{N}(0, \sigma^2), \tag{1}$$

where f is a neural network $f(x) = f(x; w)$ with parameters w.

Bayesian Deep Learning

Let $\mathcal{D} := \{(x_n, y_n) \mid n = 1, \ldots, N\} \subset \mathcal{X} \times \mathcal{Y}$ be data.

- Supervised Deep Learning:

$$Y = f(x) + \mathcal{N}(0, \sigma^2), \tag{1}$$

  where f is a neural network $f(x) = f(x; w)$ with parameters w.

- Bayesian Neural Network:
  Sample $W \sim p(w)$ and obtain random function $F(x; W)$ as prior.

UNIVERSITY OF OXFORD

Bayesian Deep Learning

Let $\mathcal{D} := \{(x_n, y_n) \mid n = 1, \dots, N\} \subset \mathcal{X} \times \mathcal{Y}$ be data.

- Supervised Deep Learning:

$$Y = f(x) + \mathcal{N}(0, \sigma^2), \tag{1}$$

  where f is a neural network $f(x) = f(x; w)$ with parameters w.
- Bayesian Neural Network:
  Sample $W \sim p(w)$ and obtain random function $F(x; W)$ as prior.
- Predictions for arbitrary $x^* \in \mathcal{X}$ follow from Bayes rule:

$$p(y^*|\mathcal{D}) = \int p\big(y^*|w\big)p(w|\mathcal{D})\,dw \tag{2}$$

$$= \int p\big(y^*|f(x^*; w)\big)p(w|\mathcal{D})\,dw \tag{3}$$

UNIVERSITY OF OXFORD

Bayesian Deep Learning

Bayesian Deep Learning

Why Bayesian deep learning instead of standard deep learning?

Bayesian Deep Learning

Why Bayesian deep learning instead of standard deep learning?

- Bayesian model averaging may improve predictive performance:

$$p(y^*|\mathcal{D}) = \int p\big(y^*|f(x^*;w)\big)p(w|\mathcal{D})\,dw \tag{4}$$

Bayesian Deep Learning

Why Bayesian deep learning instead of standard deep learning?

- Bayesian model averaging may improve predictive performance:

$$p(y^*|\mathcal{D}) = \int p\big(y^*|f(x^*; w)\big)p(w|\mathcal{D})\, dw \tag{4}$$

- Bayesian posterior can be used for uncertainty quantification

Bayesian Deep Learning

Why Bayesian deep learning instead of standard deep learning?

- Bayesian model averaging may improve predictive performance:

$$p(y^*|\mathcal{D}) = \int p\big(y^*|f(x^*; w)\big) p(w|\mathcal{D}) \, dw \tag{4}$$

- Bayesian posterior can be used for uncertainty quantification

Problem: $p(w|\mathcal{D})$ is intractable! Approximations required.

Bayesian Deep Learning

Why Bayesian deep learning instead of standard deep learning?

- Bayesian model averaging may improve predictive performance:

$$p(y^*|\mathcal{D}) = \int p\big(y^*|f(x^*; w)\big)p(w|\mathcal{D})\,dw \tag{4}$$

- Bayesian posterior can be used for uncertainty quantification

Problem: $p(w|\mathcal{D})$ is intractable! Approximations required.

Sampling based approaches:

Bayesian Deep Learning

Why Bayesian deep learning instead of standard deep learning?

- Bayesian model averaging may improve predictive performance:

$$p(y^*|\mathcal{D}) = \int p\big(y^*|f(x^*;w)\big)p(w|\mathcal{D})\,dw \tag{4}$$

- Bayesian posterior can be used for uncertainty quantification

Problem: $p(w|\mathcal{D})$ is intractable! Approximations required.

Sampling based approaches:

- Hamiltonian Monte Carlo [Neal, 2012, Chen et al., 2014]

Bayesian Deep Learning

Why Bayesian deep learning instead of standard deep learning?

- Bayesian model averaging may improve predictive performance:

$$p(y^*|\mathcal{D}) = \int p\big(y^*|f(x^*; w)\big)p(w|\mathcal{D})\, dw \tag{4}$$

- Bayesian posterior can be used for uncertainty quantification

Problem: $p(w|\mathcal{D})$ is intractable! Approximations required.

Sampling based approaches:

- Hamiltonian Monte Carlo [Neal, 2012, Chen et al., 2014]
- Langevin Dynamics [Welling and Teh, 2011]

BDL: Weight-Space inference

BDL: Weight-Space inference

Variational approach:

BDL: Weight-Space inference

Variational approach:
Let $q(w) = q(w; \nu)$ be a distribution with unknown parameters ν.

BDL: Weight-Space inference

Variational approach:
Let $q(w) = q(w; \nu)$ be a distribution with unknown parameters ν. Learn ν by maximising

$$\mathcal{L}(\nu) := \mathbb{E}_{q(w)}\big[\log p(y|w)\big] - \mathbb{D}_{KL}\big(q(w), p(w)\big), \tag{5}$$

which is (often) tractable. Use $q(w; \nu) \approx p(w|\mathcal{D})$.

UNIVERSITY OF OXFORD

BDL: Weight-Space inference

Variational approach:
Let $\mathrm{q}(\mathrm{w}) = \mathrm{q}(\mathrm{w}; \nu)$ be a distribution with unknown parameters ν. Learn ν by maximising

$$\mathcal{L}(\nu) := \mathbb{E}_{\mathrm{q}(\mathrm{w})}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{w})\big] - \mathbb{D}_{\mathrm{KL}}\big(\mathrm{q}(\mathrm{w}), \mathrm{p}(\mathrm{w})\big), \tag{5}$$

which is (often) tractable. Use $\mathrm{q}(\mathrm{w}; \nu) \approx \mathrm{p}(\mathrm{w}|\mathcal{D})$.
Problems:

UNIVERSITY OF OXFORD

BDL: Weight-Space inference

Variational approach:
Let $\mathrm{q}(\mathrm{w}) = \mathrm{q}(\mathrm{w};\nu)$ be a distribution with unknown parameters ν. Learn ν by maximising

$$\mathcal{L}(\nu) := \mathbb{E}_{\mathrm{q}(\mathrm{w})}\left[\log \mathrm{p}(\mathrm{y}|\mathrm{w})\right] - \mathbb{D}_{\mathrm{KL}}\big(\mathrm{q}(\mathrm{w}), \mathrm{p}(\mathrm{w})\big), \tag{5}$$

which is (often) tractable. Use $\mathrm{q}(\mathrm{w};\nu) \approx \mathrm{p}(\mathrm{w}|\mathcal{D})$.
Problems:

- The parameter space for w is large and the posterior multimodal.

UNIVERSITY OF OXFORD

BDL: Weight-Space inference

Variational approach:
Let $\mathrm{q}(\mathrm{w}) = \mathrm{q}(\mathrm{w};\nu)$ be a distribution with unknown parameters ν. Learn ν by maximising

$$\mathcal{L}(\nu) := \mathbb{E}_{\mathrm{q}(\mathrm{w})}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{w})\big] - \mathbb{D}_{\mathrm{KL}}\big(\mathrm{q}(\mathrm{w}), \mathrm{p}(\mathrm{w})\big), \tag{5}$$

which is (often) tractable. Use $\mathrm{q}(\mathrm{w};\nu) \approx \mathrm{p}(\mathrm{w}|\mathcal{D})$.
Problems:

- The parameter space for w is large and the posterior multimodal.
$\longrightarrow$ challenging for sampling based approaches

UNIVERSITY OF OXFORD

BDL: Weight-Space inference

Variational approach:
Let $q(w) = q(w;\nu)$ be a distribution with unknown parameters ν. Learn ν by maximising

$$\mathcal{L}(\nu) := \mathbb{E}_{q(w)}\left[\log p(y|w)\right] - \mathbb{D}_{KL}\left(q(w), p(w)\right), \tag{5}$$

which is (often) tractable. Use $q(w;\nu) \approx p(w|\mathcal{D})$.
Problems:

- The parameter space for w is large and the posterior multimodal. $\longrightarrow$ challenging for sampling based approaches
- Variational approaches often introduce strong assumptions for tractability.

UNIVERSITY OF OXFORD

BDL: Weight-Space inference

Variational approach:
Let $q(w) = q(w; \nu)$ be a distribution with unknown parameters ν. Learn ν by maximising

$$\mathcal{L}(\nu) := \mathbb{E}_{q(w)}\big[\log p(y|w)\big] - \mathbb{D}_{KL}\big(q(w), p(w)\big), \tag{5}$$

which is (often) tractable. Use $q(w; \nu) \approx p(w|\mathcal{D})$.
Problems:

- The parameter space for w is large and the posterior multimodal.
  $\longrightarrow$ challenging for sampling based approaches
- Variational approaches often introduce strong assumptions for tractability.
  $\longrightarrow$ do we still capture enough of the true posterior? [Foong et al., 2020]

UNIVERSITY OF OXFORD

BDL: Weight-Space inference

Variational approach:
Let $\mathrm{q}(\mathrm{w}) = \mathrm{q}(\mathrm{w}; \nu)$ be a distribution with unknown parameters ν. Learn ν by maximising

$$\mathcal{L}(\nu) := \mathbb{E}_{\mathrm{q}(\mathrm{w})}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{w})\big] - \mathbb{D}_{\mathrm{KL}}\big(\mathrm{q}(\mathrm{w}), \mathrm{p}(\mathrm{w})\big), \tag{5}$$

which is (often) tractable. Use $\mathrm{q}(\mathrm{w}; \nu) \approx \mathrm{p}(\mathrm{w}|\mathcal{D})$.
Problems:

- The parameter space for w is large and the posterior multimodal.
  $\longrightarrow$ challenging for sampling based approaches
- Variational approaches often introduce strong assumptions for tractability.
  $\longrightarrow$ do we still capture enough of the true posterior? [Foong et al., 2020]
- What priors on the function space are induced by $\mathrm{p}(\mathrm{w})$?

UNIVERSITY OF OXFORD

Variational Inference in Function Spaces

Variational Inference in Function Spaces

Idea: perform inference in function space [Ma et al., 2019, Sun et al., 2019, Rudner et al., 2020, Ma and Hernández-Lobato, 2021]

Variational Inference in Function Spaces

Idea: perform inference in function space [Ma et al., 2019, Sun et al., 2019, Rudner et al., 2020, Ma and Hernández-Lobato, 2021]

$$\mathcal{L} = \mathbb{E}_{\mathbb{Q}}\left[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\right] - \mathbb{D}_{\mathrm{KL}}\left(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\right), \tag{6}$$

where $\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}} \in \mathcal{P}(\mathrm{E})$ with:

UNIVERSITY OF OXFORD

Variational Inference in Function Spaces

Idea: perform inference in function space [Ma et al., 2019, Sun et al., 2019, Rudner et al., 2020, Ma and Hernández-Lobato, 2021]

$$\mathcal{L} = \mathbb{E}_{\mathbb{Q}}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\big] - \mathbb{D}_{\mathrm{KL}}\big(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\big), \tag{6}$$

where $\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}} \in \mathcal{P}(\mathrm{E})$ with:

- E an infinite dimensional (Polish) function space

UNIVERSITY OF OXFORD

Variational Inference in Function Spaces

Idea: perform inference in function space [Ma et al., 2019, Sun et al., 2019, Rudner et al., 2020, Ma and Hernández-Lobato, 2021]

$$\mathcal{L} = \mathbb{E}_{\mathbb{Q}}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\big] - \mathbb{D}_{\mathrm{KL}}\big(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\big), \tag{6}$$

where $\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}} \in \mathcal{P}(\mathrm{E})$ with:

- E an infinite dimensional (Polish) function space
- $\mathcal{P}(\mathrm{E})$ the space of Borel probability measures on E

UNIVERSITY OF OXFORD

Variational Inference in Function Spaces

Idea: perform inference in function space [Ma et al., 2019, Sun et al., 2019, Rudner et al., 2020, Ma and Hernández-Lobato, 2021]

$$\mathcal{L} = \mathbb{E}_{\mathbb{Q}}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\big] - \mathbb{D}_{\mathrm{KL}}\big(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\big), \tag{6}$$

where $\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}} \in \mathcal{P}(\mathrm{E})$ with:

- E an infinite dimensional (Polish) function space
- $\mathcal{P}(\mathrm{E})$ the space of Borel probability measures on E

Challenges:

UNIVERSITY OF OXFORD

Variational Inference in Function Spaces

Idea: perform inference in function space [Ma et al., 2019, Sun et al., 2019, Rudner et al., 2020, Ma and Hernández-Lobato, 2021]

$$\mathcal{L} = \mathbb{E}_{\mathbb{Q}}\left[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\right] - \mathbb{D}_{\mathrm{KL}}\left(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\right), \tag{6}$$

where $\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}} \in \mathcal{P}(\mathrm{E})$ with:

- E an infinite dimensional (Polish) function space
- $\mathcal{P}(\mathrm{E})$ the space of Borel probability measures on E

Challenges:

- How to specify priors on infinite dimensional function spaces?

Variational Inference in Function Spaces

Idea: perform inference in function space [Ma et al., 2019, Sun et al., 2019, Rudner et al., 2020, Ma and Hernández-Lobato, 2021]

$$\mathcal{L} = \mathbb{E}_{\mathbb{Q}}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\big] - \mathbb{D}_{\mathrm{KL}}\big(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\big), \tag{6}$$

where $\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}} \in \mathcal{P}(\mathrm{E})$ with:

- E an infinite dimensional (Polish) function space
- $\mathcal{P}(\mathrm{E})$ the space of Borel probability measures on E

Challenges:

- How to specify priors on infinite dimensional function spaces?
  → Gaussian measures on Hilbert spaces

UNIVERSITY OF OXFORD

Variational Inference in Function Spaces

Idea: perform inference in function space [Ma et al., 2019, Sun et al., 2019, Rudner et al., 2020, Ma and Hernández-Lobato, 2021]

$$\mathcal{L} = \mathbb{E}_{\mathbb{Q}}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\big] - \mathbb{D}_{\mathrm{KL}}\big(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\big), \tag{6}$$

where $\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}} \in \mathcal{P}(\mathrm{E})$ with:

- E an infinite dimensional (Polish) function space
- $\mathcal{P}(\mathrm{E})$ the space of Borel probability measures on E

Challenges:

- How to specify priors on infinite dimensional function spaces?
  $\rightarrow$ Gaussian measures on Hilbert spaces
- The KL-divergence is (in general) intractable in infinite dimensions and may even be infinite [Burt et al., 2020].

Variational Inference in Function Spaces

Idea: perform inference in function space [Ma et al., 2019, Sun et al., 2019, Rudner et al., 2020, Ma and Hernández-Lobato, 2021]

$$\mathcal{L} = \mathbb{E}_{\mathbb{Q}}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\big] - \mathbb{D}_{\mathrm{KL}}\big(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\big), \tag{6}$$

where $\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}} \in \mathcal{P}(\mathrm{E})$ with:

- E an infinite dimensional (Polish) function space
- $\mathcal{P}(\mathrm{E})$ the space of Borel probability measures on E

Challenges:

- How to specify priors on infinite dimensional function spaces?
  → Gaussian measures on Hilbert spaces
- The KL-divergence is (in general) intractable in infinite dimensions and may even be infinite [Burt et al., 2020].
  → use generalised variational inference in infinite dimensions!

The Rule of Three [Knoblauch et al., 2019]

The Rule of Three [Knoblauch et al., 2019]

Find posterior as

UNIVERSITY OF OXFORD

The Rule of Three [Knoblauch et al., 2019]

Find posterior as

$$q^*(w) := \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \left\{ \mathbb{E}_{q(w)}\left[\sum_{n=1}^{N} \ell(y_n, w) \right] + D\big(q(w), p(w)\big) \right\}, \tag{7}$$

The Rule of Three [Knoblauch et al., 2019]

Find posterior as

$$q^*(w) := \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \left\{ \mathbb{E}_{q(w)}\left[\sum_{n=1}^{N} \ell(y_n, w) \right] + D\big(q(w), p(w)\big) \right\}, \tag{7}$$

where:

UNIVERSITY OF OXFORD

The Rule of Three [Knoblauch et al., 2019]

Find posterior as

$$q^*(w) := \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \left\{ \mathbb{E}_{q(w)}\left[\sum_{n=1}^{N} \ell(y_n, w) \right] + D\big(q(w), p(w)\big) \right\}, \tag{7}$$

where:

- $\mathcal{Q}$ is a set of tractable pdfs

UNIVERSITY OF OXFORD

The Rule of Three [Knoblauch et al., 2019]

Find posterior as

$$q^*(w) := \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \left\{ \mathbb{E}_{q(w)}\left[\sum_{n=1}^{N} \ell(y_n, w) \right] + D\big(q(w), p(w)\big) \right\}, \tag{7}$$

where:

- $\mathcal{Q}$ is a set of tractable pdfs
- ℓ is a loss function

UNIVERSITY OF OXFORD

The Rule of Three [Knoblauch et al., 2019]

Find posterior as

$$q^*(w) := \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \left\{ \mathbb{E}_{q(w)} \left[\sum_{n=1}^{N} \ell(y_n, w) \right] + D\big(q(w), p(w)\big) \right\}, \qquad (7)$$

where:

- $\mathcal{Q}$ is a set of tractable pdfs
- ℓ is a loss function
- D an arbitrary divergence

UNIVERSITY OF OXFORD

GVI in Function Spaces

GVI in Function Spaces

- Idea: Use rule of three in infinite dimensional function spaces

GVI in Function Spaces

- Idea: Use rule of three in infinite dimensional function spaces
- Theorem 1 in Knoblauch et al. [2019] holds for infinite dimensional parameter spaces

UNIVERSITY OF OXFORD

GVI in Function Spaces

- Idea: Use rule of three in infinite dimensional function spaces
- Theorem 1 in Knoblauch et al. [2019] holds for infinite dimensional parameter spaces
- We can target

$$\mathcal{L} := -\mathbb{E}_{\mathbb{Q}}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\big] + \mathbb{D}\big(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\big), \tag{8}$$

for inference where $\mathbb{D}$ is an arbitrary divergence.

GVI in Function Spaces

- Idea: Use rule of three in infinite dimensional function spaces
- Theorem 1 in Knoblauch et al. [2019] holds for infinite dimensional parameter spaces
- We can target

$$\mathcal{L} := -\mathbb{E}_{\mathbb{Q}}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\big] + \mathbb{D}\big(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\big), \tag{8}$$

  for inference where $\mathbb{D}$ is an arbitrary divergence.
- How to define priors and variational measures $\mathbb{P}^{\mathrm{F}}$ and $\mathbb{Q}^{\mathrm{F}}$ in infinite dimensions?

UNIVERSITY OF OXFORD

Gaussian Measures on Hilbert spaces

UNIVERSITY OF OXFORD

Gaussian Measures on Hilbert spaces

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(\mathrm{H}, \langle \cdot, \cdot \rangle)$ be a Hilbert space.

UNIVERSITY OF OXFORD

Gaussian Measures on Hilbert spaces

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(\mathrm{H}, \langle\cdot,\cdot\rangle)$ be a Hilbert space.

Definition (Gaussian Random Element)

A random mapping $\mathrm{F} : \Omega \to \mathrm{H}$ is called Gaussian random element (GRE) if and only if

$$\langle \mathrm{F}, \mathrm{h} \rangle : \Omega \to \mathbb{R} \tag{9}$$

is a scalar Gaussian variable for every $\mathrm{h} \in \mathrm{H}$.

UNIVERSITY OF OXFORD

Gaussian Measures on Hilbert spaces

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(\mathrm{H}, \langle \cdot, \cdot \rangle)$ be a Hilbert space.

Definition (Gaussian Random Element)

A random mapping $\mathrm{F} : \Omega \to \mathrm{H}$ is called Gaussian random element (GRE) if and only if

$$\langle \mathrm{F}, \mathrm{h} \rangle : \Omega \to \mathbb{R} \tag{9}$$

is a scalar Gaussian variable for every $\mathrm{h} \in \mathrm{H}$.

The mean element of F is defined as

$$\mathrm{m} := \mathbb{E}[\mathrm{F}] := \int \mathrm{F}(\omega)\, \mathrm{d}\mathbb{P}(\omega) \in \mathrm{H}$$

UNIVERSITY OF OXFORD

Gaussian Measures on Hilbert spaces

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(\mathrm{H}, \langle\cdot,\cdot\rangle)$ be a Hilbert space.

Definition (Gaussian Random Element)

A random mapping $\mathrm{F} : \Omega \to \mathrm{H}$ is called Gaussian random element (GRE) if and only if

$$\langle \mathrm{F}, \mathrm{h}\rangle : \Omega \to \mathbb{R} \tag{9}$$

is a scalar Gaussian variable for every $\mathrm{h} \in \mathrm{H}$.

The mean element of F is defined as

$$\mathrm{m} := \mathbb{E}[\mathrm{F}] := \int \mathrm{F}(\omega)\, \mathrm{d}\mathbb{P}(\omega) \in \mathrm{H} \tag{10}$$

and the (linear) covariance operator $\mathrm{C} : \mathrm{H} \to \mathrm{H}$ of F is defined as

$$\mathrm{C}(\mathrm{h}) := \int \langle \mathrm{F}(\omega), \mathrm{h}\rangle \mathrm{F}(\omega)\, \mathrm{d}\mathbb{P}(\omega) - \langle \mathrm{m}, \mathrm{h}\rangle \mathrm{m},\ \mathrm{h} \in \mathrm{H}.$$

UNIVERSITY OF OXFORD

Gaussian Measures on Hilbert spaces

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(\mathrm{H}, \langle\cdot,\cdot\rangle)$ be a Hilbert space.

Definition (Gaussian Random Element)

A random mapping $\mathrm{F} : \Omega \to \mathrm{H}$ is called Gaussian random element (GRE) if and only if

$$\langle \mathrm{F}, \mathrm{h} \rangle : \Omega \to \mathbb{R} \tag{9}$$

is a scalar Gaussian variable for every $\mathrm{h} \in \mathrm{H}$.

The mean element of F is defined as

$$\mathrm{m} := \mathbb{E}[\mathrm{F}] := \int \mathrm{F}(\omega)\, \mathrm{d}\mathbb{P}(\omega) \in \mathrm{H} \tag{10}$$

and the (linear) covariance operator $\mathrm{C} : \mathrm{H} \to \mathrm{H}$ of F is defined as

$$\mathrm{C}(\mathrm{h}) := \int \langle \mathrm{F}(\omega), \mathrm{h} \rangle \mathrm{F}(\omega)\, \mathrm{d}\mathbb{P}(\omega) - \langle \mathrm{m}, \mathrm{h} \rangle \mathrm{m},\ \mathrm{h} \in \mathrm{H}. \tag{11}$$

UNIVERSITY OF OXFORD

Gaussian Measures on Hilbert spaces

Gaussian Measures on Hilbert spaces

By properties of the Bochner integral:

$$\langle \mathrm{F}, \mathrm{h} \rangle \sim \mathcal{N}\big(\langle \mathrm{m}, \mathrm{h} \rangle, \langle \mathrm{Ch}, \mathrm{h} \rangle\big), \tag{12}$$

for any $\mathrm{h} \in \mathrm{H}$.

Gaussian Measures on Hilbert spaces

By properties of the Bochner integral:

$$\langle \mathrm{F}, \mathrm{h} \rangle \sim \mathcal{N}\big(\langle \mathrm{m}, \mathrm{h} \rangle, \langle \mathrm{Ch}, \mathrm{h} \rangle\big), \tag{12}$$

for any $\mathrm{h} \in \mathrm{H}$. Write $\mathrm{F} \sim \mathcal{N}(\mathrm{m}, \mathrm{C})$ for a GRE with mean element $\mathrm{m} \in \mathrm{H}$ and covariance operator C.

UNIVERSITY OF OXFORD

Gaussian Measures on Hilbert spaces

By properties of the Bochner integral:

$$\langle \mathrm{F}, \mathrm{h} \rangle \sim \mathcal{N}\big(\langle \mathrm{m}, \mathrm{h} \rangle, \langle \mathrm{Ch}, \mathrm{h} \rangle\big), \tag{12}$$

for any $\mathrm{h} \in \mathrm{H}$. Write $\mathrm{F} \sim \mathcal{N}(\mathrm{m}, \mathrm{C})$ for a GRE with mean element $\mathrm{m} \in \mathrm{H}$ and covariance operator C.

- $\mathrm{C} : \mathrm{H} \to \mathrm{H}$ of a GRE is a positive self-adjoint trace-class operator.

Gaussian Measures on Hilbert spaces

By properties of the Bochner integral:

$$\langle F, h\rangle \sim \mathcal{N}\big(\langle m, h\rangle, \langle Ch, h\rangle\big), \tag{12}$$

for any $h \in H$. Write $F \sim \mathcal{N}(m, C)$ for a GRE with mean element $m \in H$ and covariance operator C.

- $C : H \to H$ of a GRE is a positive self-adjoint trace-class operator.
- For arbitrary $m \in H$ and arbitrary C positive, self-adjoint and trace-class there exists a GRE such that $F \sim \mathcal{N}(m, C)$.

Gaussian Measures on Hilbert spaces

By properties of the Bochner integral:

$$\langle \mathrm{F}, \mathrm{h} \rangle \sim \mathcal{N}\big(\langle \mathrm{m}, \mathrm{h} \rangle, \langle \mathrm{Ch}, \mathrm{h} \rangle\big), \tag{12}$$

for any $\mathrm{h} \in \mathrm{H}$. Write $\mathrm{F} \sim \mathcal{N}(\mathrm{m}, \mathrm{C})$ for a GRE with mean element $\mathrm{m} \in \mathrm{H}$ and covariance operator C.

- $\mathrm{C} : \mathrm{H} \to \mathrm{H}$ of a GRE is a positive self-adjoint trace-class operator.
- For arbitrary $\mathrm{m} \in \mathrm{H}$ and arbitrary C positive, self-adjoint and trace-class there exists a GRE such that $\mathrm{F} \sim \mathcal{N}(\mathrm{m}, \mathrm{C})$.

Definition (Gaussian Measure)

Let $\mathrm{F} \sim \mathcal{N}(\mathrm{m}, \mathrm{C})$ be a GRE. Then P defined as

$$\mathrm{P}(\mathrm{A}) := \mathbb{P}^{\mathrm{F}}(\mathrm{A}) := \mathbb{P}(\mathrm{F} \in \mathrm{A}) \tag{13}$$

for any (measurable) $\mathrm{A} \subset \mathrm{H}$ is called a Gaussian measure.

Contents

Model description

Model description

Recall the generalised loss:

$$\mathcal{L} := -\mathbb{E}_{\mathbb{Q}}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\big] + \mathbb{D}\big(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\big)$$

Model description

Recall the generalised loss:

$$\mathcal{L} := -\mathbb{E}_{\mathbb{Q}}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\big] + \mathbb{D}\big(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\big) \tag{14}$$

Gaussian Wasserstein Inference:

Model description

Recall the generalised loss:

$$\mathcal{L} := -\mathbb{E}_{\mathbb{Q}}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\big] + \mathbb{D}\big(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\big) \tag{14}$$

Gaussian Wasserstein Inference:

- $\mathrm{E} = \mathrm{L}^2(\mathcal{X}, \rho, \mathbb{R}) := \big\{\mathrm{f} : \mathcal{X} \to \mathbb{R} \mid \int |\mathrm{f}(\mathrm{x})|^2 \,\mathrm{d}\rho(\mathrm{x}) < \infty\big\}$ with ρ input distribution on $\mathcal{X}$

UNIVERSITY OF OXFORD

Model description

Recall the generalised loss:

$$\mathcal{L} := -\mathbb{E}_{\mathbb{Q}}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\big] + \mathbb{D}\big(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\big) \tag{14}$$

Gaussian Wasserstein Inference:

- $\mathrm{E} = \mathrm{L}^2(\mathcal{X}, \rho, \mathbb{R}) := \big\{\mathrm{f} : \mathcal{X} \to \mathbb{R} \mid \int |\mathrm{f}(\mathrm{x})|^2 \,\mathrm{d}\rho(\mathrm{x}) < \infty\big\}$ with ρ input distribution on $\mathcal{X}$
- $\mathrm{P} := \mathbb{P}^{\mathrm{F}} \sim \mathcal{N}\big(\mathrm{m}_{\mathrm{P}}, \mathrm{C}_{\mathrm{P}}\big)$

UNIVERSITY OF OXFORD

Model description

Recall the generalised loss:

$$\mathcal{L} := -\mathbb{E}_{\mathbb{Q}}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\big] + \mathbb{D}\big(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\big) \tag{14}$$

Gaussian Wasserstein Inference:

- $\mathrm{E} = \mathrm{L}^2(\mathcal{X}, \rho, \mathbb{R}) := \big\{\mathrm{f} : \mathcal{X} \to \mathbb{R} \mid \int |\mathrm{f}(\mathrm{x})|^2 \, \mathrm{d}\rho(\mathrm{x}) < \infty\big\}$ with ρ input distribution on $\mathcal{X}$
- $\mathrm{P} := \mathbb{P}^{\mathrm{F}} \sim \mathcal{N}\big(\mathrm{m}_{\mathrm{P}}, \mathrm{C}_{\mathrm{P}}\big)$
- $\mathrm{Q} := \mathbb{Q}^{\mathrm{F}} \sim \mathcal{N}\big(\mathrm{m}_{\mathrm{Q}}, \mathrm{C}_{\mathrm{Q}}\big)$

UNIVERSITY OF OXFORD

Model description

Recall the generalised loss:

$$\mathcal{L} := -\mathbb{E}_{\mathbb{Q}}\left[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\right] + \mathbb{D}\left(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\right) \tag{14}$$

Gaussian Wasserstein Inference:

- $\mathrm{E} = \mathrm{L}^2(\mathcal{X}, \rho, \mathbb{R}) := \left\{\mathrm{f} : \mathcal{X} \to \mathbb{R} \,|\, \int |\mathrm{f}(\mathrm{x})|^2 \,\mathrm{d}\rho(\mathrm{x}) < \infty\right\}$ with ρ input distribution on $\mathcal{X}$
- $\mathrm{P} := \mathbb{P}^{\mathrm{F}} \sim \mathcal{N}\left(\mathrm{m}_{\mathrm{P}}, \mathrm{C}_{\mathrm{P}}\right)$
- $\mathrm{Q} := \mathbb{Q}^{\mathrm{F}} \sim \mathcal{N}\left(\mathrm{m}_{\mathrm{Q}}, \mathrm{C}_{\mathrm{Q}}\right)$
- $\mathbb{D}(\cdot,\cdot) = \mathrm{W}_2(\cdot,\cdot)$ with W_2 given as Wasserstein-distance

Model description

Recall the generalised loss:

$$\mathcal{L} := -\mathbb{E}_{\mathbb{Q}}\big[\log \mathrm{p}(\mathrm{y}|\mathrm{F})\big] + \mathbb{D}\big(\mathbb{Q}^{\mathrm{F}}, \mathbb{P}^{\mathrm{F}}\big) \tag{14}$$

Gaussian Wasserstein Inference:

- $\mathrm{E} = \mathrm{L}^2(\mathcal{X}, \rho, \mathbb{R}) := \big\{\mathrm{f} : \mathcal{X} \to \mathbb{R} \mid \int |\mathrm{f}(\mathrm{x})|^2 \,\mathrm{d}\rho(\mathrm{x}) < \infty\big\}$ with ρ input distribution on $\mathcal{X}$
- $\mathrm{P} := \mathbb{P}^{\mathrm{F}} \sim \mathcal{N}\big(\mathrm{m}_{\mathrm{P}}, \mathrm{C}_{\mathrm{P}}\big)$
- $\mathrm{Q} := \mathbb{Q}^{\mathrm{F}} \sim \mathcal{N}\big(\mathrm{m}_{\mathrm{Q}}, \mathrm{C}_{\mathrm{Q}}\big)$
- $\mathbb{D}(\cdot,\cdot) = \mathrm{W}_2(\cdot,\cdot)$ with W_2 given as Wasserstein-distance

with:

$$\mathrm{C}_{\mathrm{P}}\mathrm{g} := \int \mathrm{k}(\cdot, \mathrm{x}')\mathrm{g}(\mathrm{x}')\,\mathrm{d}\rho(\mathrm{x}'), \qquad \mathrm{C}_{\mathrm{Q}}\mathrm{g} := \int \mathrm{r}(\cdot, \mathrm{x}')\mathrm{g}(\mathrm{x}')\,\mathrm{d}\rho(\mathrm{x}') \tag{15}$$

for all $\mathrm{g} \in \mathrm{L}^2(\mathcal{X}, \rho, \mathbb{R})$ where k and r are trace-class kernels.

UNIVERSITY OF OXFORD

Regression

Regression

For regression:

$$p(y|F) := \prod_{n=1}^{N} p(y_n|F) := \prod_{n=1}^{N} \mathcal{N}(y_n \,|\, F(x_n), \sigma^2), \tag{16}$$

where $\sigma^2 > 0$.

UNIVERSITY OF OXFORD

Regression

For regression:

$$p(y|F) := \prod_{n=1}^{N} p(y_n|F) := \prod_{n=1}^{N} \mathcal{N}(y_n \,|\, F(x_n), \sigma^2), \tag{16}$$

where $\sigma^2 > 0$.
The Wasserstein distance is tractable [Gelbrich, 1990]:

$$W_2^2(P, Q) = \|m_P - m_Q\|_2^2 + \mathrm{tr}(C_P) + \mathrm{tr}(C_Q) - 2 \cdot \mathrm{tr}\left[\left(C_P^{1/2} C_Q C_P^{1/2}\right)^{1/2}\right], \tag{17}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of an operator and $C_P^{1/2}$ is the square root of the positive, self-adjoint operator C_P.

UNIVERSITY OF OXFORD

Approximation of Wasserstein distance

Approximation of Wasserstein distance

Let $\widehat{\rho} := \frac{1}{N}\sum_{n=1}^{N} \delta_{x_n}$

Approximation of Wasserstein distance

Let $\widehat{\rho} := \frac{1}{N}\sum_{n=1}^{N} \delta_{x_n}$ and notice that

$$\|m_P - m_Q\|_2^2 = \int \left(m_P(x) - m_Q(x)\right)^2 d\rho(x) \quad (18)$$

$$\approx \frac{1}{N}\sum_{n=1}^{N} \left(m_P(x_n) - m_Q(x_n)\right)^2 \quad (19)$$

UNIVERSITY OF OXFORD

Approximation of Wasserstein distance

Let $\widehat{\rho} := \frac{1}{N}\sum_{n=1}^{N} \delta_{x_n}$ and notice that

$$\|m_P - m_Q\|_2^2 = \int \left(m_P(x) - m_Q(x)\right)^2 d\rho(x) \quad (18)$$

$$\approx \frac{1}{N}\sum_{n=1}^{N} \left(m_P(x_n) - m_Q(x_n)\right)^2 \quad (19)$$

Further:

Approximation of Wasserstein distance

Let $\widehat{\rho} := \frac{1}{N}\sum_{n=1}^{N}\delta_{x_n}$ and notice that

$$\|m_P - m_Q\|_2^2 = \int \big(m_P(x) - m_Q(x)\big)^2 \, d\rho(x) \tag{18}$$

$$\approx \frac{1}{N}\sum_{n=1}^{N}\big(m_P(x_n) - m_Q(x_n)\big)^2 \tag{19}$$

Further:

$$\mathrm{tr}(C_P) = \int k(x,x)\, d\rho(x) \approx \frac{1}{N}\sum_{n=1}^{N} k(x_n, x_n) \tag{20}$$

$$\mathrm{tr}(C_Q) = \int r(x,x)\, d\rho(x) \approx \frac{1}{N}\sum_{n=1}^{N} r(x_n, x_n) \tag{21}$$

Approximation of Wasserstein distance

Approximation of Wasserstein distance

The last term can be approximated as

$$\mathrm{tr}\left[\left(C_P^{1/2} C_Q C_P^{1/2}\right)^{1/2}\right] \approx \frac{1}{\sqrt{N N_S}} \sum_{s=1}^{N_S} \sqrt{\lambda_s\big(r(X_S, X) k(X, X_S)\big)}, \qquad (22)$$

Approximation of Wasserstein distance

The last term can be approximated as

$$\mathrm{tr}\left[\left(C_P^{1/2} C_Q C_P^{1/2}\right)^{1/2}\right] \approx \frac{1}{\sqrt{N N_S}} \sum_{s=1}^{N_S} \sqrt{\lambda_s\big(r(X_S, X) k(X, X_S)\big)}, \tag{22}$$

where $X_S := (x_{S,1}, \ldots, x_{S,N_S})$, $N_S \in \mathbb{N}$ with:

$$X_{S,1}, \ldots, X_{S,N_S} \overset{\text{ind.}}{\sim} \hat{\rho} \tag{23}$$

$$r(X_S, X) := \big(r(x_{S,s}, x_n)\big)_{s,n} \tag{24}$$

$$k(X, X_S) := \big(k(x_n, x_{S,s})\big)_{n,s} \tag{25}$$

and $\lambda_s\big(r(X_S, X)k(X, X_S)\big)$ denotes the s-th eigenvalue of the matrix $r(X_S, X)k(X, X_S) \in \mathbb{R}^{N_S \times N_S}$.

UNIVERSITY OF OXFORD

Final Loss: Regression

Final Loss: Regression

The final loss:

$$\mathcal{L} = \mathrm{L} + \widehat{\mathrm{W}}^2 \tag{26}$$

Final Loss: Regression

The final loss:

$$\mathcal{L} = \mathrm{L} + \widehat{\mathrm{W}}^2 \tag{26}$$

with:

$$\mathrm{L} := \frac{\mathrm{N}}{2}\log(2\pi\sigma^2) + \sum_{\mathrm{n}=1}^{\mathrm{N}} \frac{\big(\mathrm{y_n} - \mathrm{m_Q}(\mathrm{x_n})\big)^2 + \mathrm{r}(\mathrm{x_n}, \mathrm{x_n})}{2\sigma^2} \tag{27}$$

$$\hat{\mathrm{W}}^2 := \frac{1}{\mathrm{N}}\sum_{\mathrm{n}=1}^{\mathrm{N}} \big(\mathrm{m_P}(\mathrm{x_n}) - \mathrm{m_Q}(\mathrm{x_n})\big)^2 + \frac{1}{\mathrm{N}}\sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{k}(\mathrm{x_n}, \mathrm{x_n}) \tag{28}$$

$$+ \frac{1}{\mathrm{N}}\sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{r}(\mathrm{x_n}, \mathrm{x_n}) - \frac{2}{\sqrt{\mathrm{N}\mathrm{N_S}}}\sum_{\mathrm{s}=1}^{\mathrm{N_S}} \sqrt{\lambda_\mathrm{s}\big(\mathrm{r}(\mathrm{X_S}, \mathrm{X})\mathrm{k}(\mathrm{X}, \mathrm{X_S})\big)}, \tag{29}$$

UNIVERSITY OF OXFORD

Final Loss: Regression

The final loss:

$$\mathcal{L} = \mathrm{L} + \widehat{\mathrm{W}}^2 \tag{26}$$

with:

$$\mathrm{L} := \frac{\mathrm{N}}{2}\log(2\pi\sigma^2) + \sum_{\mathrm{n}=1}^{\mathrm{N}} \frac{\big(\mathrm{y_n} - \mathrm{m_Q}(\mathrm{x_n})\big)^2 + \mathrm{r}(\mathrm{x_n}, \mathrm{x_n})}{2\sigma^2} \tag{27}$$

$$\hat{\mathrm{W}}^2 := \frac{1}{\mathrm{N}}\sum_{\mathrm{n}=1}^{\mathrm{N}} \big(\mathrm{m_P}(\mathrm{x_n}) - \mathrm{m_Q}(\mathrm{x_n})\big)^2 + \frac{1}{\mathrm{N}}\sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{k}(\mathrm{x_n}, \mathrm{x_n}) \tag{28}$$

$$+ \frac{1}{\mathrm{N}}\sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{r}(\mathrm{x_n}, \mathrm{x_n}) - \frac{2}{\sqrt{\mathrm{N}\mathrm{N_S}}}\sum_{\mathrm{s}=1}^{\mathrm{N_S}} \sqrt{\lambda_{\mathrm{s}}\big(\mathrm{r}(\mathrm{X_S}, \mathrm{X})\mathrm{k}(\mathrm{X}, \mathrm{X_S})\big)}, \tag{29}$$

UNIVERSITY OF OXFORD

Final Loss: Properties

Final Loss: Properties

- $\mathcal{L}$ is tractable for any m_P, m_Q, k and r

Final Loss: Properties

- $\mathcal{L}$ is tractable for any $\mathrm{m_P}, \mathrm{m_Q}, \mathrm{k}$ and r
- One evaluation of $\mathcal{L}$ requires:

UNIVERSITY OF OXFORD

Final Loss: Properties

- $\mathcal{L}$ is tractable for any $\mathrm{m_P}, \mathrm{m_Q}, \mathrm{k}$ and r
- One evaluation of $\mathcal{L}$ requires:
  - N evaluations of $\mathrm{m_Q}$ and $\mathrm{m_P}$

UNIVERSITY OF OXFORD

Final Loss: Properties

- $\mathcal{L}$ is tractable for any m_P, m_Q, k and r
- One evaluation of $\mathcal{L}$ requires:
  - N evaluations of m_Q and m_P
  - $N_S \cdot N$ evaluations of r and k

UNIVERSITY OF OXFORD

Final Loss: Properties

- $\mathcal{L}$ is tractable for any m_P, m_Q, k and r
- One evaluation of $\mathcal{L}$ requires:
  - N evaluations of m_Q and m_P
  - $N_S \cdot N$ evaluations of r and k
  - $\mathcal{O}(N + N_S^2 N + N_S^3)$operations for the eigenvalue problem

UNIVERSITY OF OXFORD

Final Loss: Properties

- $\mathcal{L}$ is tractable for any m_P, m_Q, k and r
- One evaluation of $\mathcal{L}$ requires:
  - N evaluations of m_Q and m_P
  - $N_S \cdot N$ evaluations of r and k
  - $\mathcal{O}(N + N_S^2 N + N_S^3)$operations for the eigenvalue problem
- One evaluation of $\mathcal{L}$ in batch-mode requires:

Final Loss: Properties

- $\mathcal{L}$ is tractable for any m_P, m_Q, k and r
- One evaluation of $\mathcal{L}$ requires:
  - N evaluations of m_Q and m_P
  - $N_S \cdot N$ evaluations of r and k
  - $\mathcal{O}(N + N_S^2 N + N_S^3)$operations for the eigenvalue problem
- One evaluation of $\mathcal{L}$ in batch-mode requires:
  - N_B evaluations of m_Q and m_P

UNIVERSITY OF OXFORD

Final Loss: Properties

- $\mathcal{L}$ is tractable for any m_P, m_Q, k and r
- One evaluation of $\mathcal{L}$ requires:
  - N evaluations of m_Q and m_P
  - $N_S \cdot N$ evaluations of r and k
  - $\mathcal{O}(N + N_S^2 N + N_S^3)$operations for the eigenvalue problem
- One evaluation of $\mathcal{L}$ in batch-mode requires:
  - N_B evaluations of m_Q and m_P
  - $N_S \cdot N_B$ evaluations of r and k

Final Loss: Properties

- $\mathcal{L}$ is tractable for any m_P, m_Q, k and r
- One evaluation of $\mathcal{L}$ requires:
  - N evaluations of m_Q and m_P
  - $N_S \cdot N$ evaluations of r and k
  - $\mathcal{O}(N + N_S^2 N + N_S^3)$operations for the eigenvalue problem
- One evaluation of $\mathcal{L}$ in batch-mode requires:
  - N_B evaluations of m_Q and m_P
  - $N_S \cdot N_B$ evaluations of r and k
  - $\mathcal{O}(N_B + N_S^2 N_B + N_S^3)$ operations for the eigenvalue problem

Final Loss: Properties

- $\mathcal{L}$ is tractable for any m_P, m_Q, k and r
- One evaluation of $\mathcal{L}$ requires:
  - N evaluations of m_Q and m_P
  - $N_S \cdot N$ evaluations of r and k
  - $\mathcal{O}(N + N_S^2 N + N_S^3)$operations for the eigenvalue problem
- One evaluation of $\mathcal{L}$ in batch-mode requires:
  - N_B evaluations of m_Q and m_P
  - $N_S \cdot N_B$ evaluations of r and k
  - $\mathcal{O}(N_B + N_S^2 N_B + N_S^3)$ operations for the eigenvalue problem

  $\longrightarrow$ very scalable for typical $N_S, N_B << N$, e.g. $N_S = N_B = 100$

UNIVERSITY OF OXFORD

Recovering Other Methods

Recovering Other Methods

- Stochastic Variational Gaussian processes (SVGP) [Titsias, 2009]:

$$m_Q(x) := m_P(x) + \sum_{m=1}^{M} \beta_m k_m(x) \tag{30}$$

$$r(x, x') := k(x, x') - k_Z(x)^T k(Z, Z)^{-1} k_Z(x) + k_Z(x)^T \Sigma k_Z(x), \tag{31}$$

where $\beta = (\beta_1, \ldots, \beta_M) \in \mathbb{R}^M$ and $\Sigma \in \mathbb{R}^{M \times M}$ are variational parameters. Further $Z = (Z_1, \ldots, Z_M)$ with $\{Z_m\}_{m=1}^M \overset{iid}{\sim} \widehat{\rho}$.

UNIVERSITY OF OXFORD

Recovering Other Methods

- Stochastic Variational Gaussian processes (SVGP) [Titsias, 2009]:

$$\mathrm{m_Q(x)} := \mathrm{m_P(x)} + \sum_{\mathrm{m}=1}^{\mathrm{M}} \beta_{\mathrm{m}} \mathrm{k_m(x)} \tag{30}$$

$$\mathrm{r(x, x')} := \mathrm{k(x, x')} - \mathrm{k_Z(x)^T k(Z, Z)^{-1} k_Z(x)} + \mathrm{k_Z(x)^T} \Sigma \mathrm{k_Z(x)}, \tag{31}$$

  where $\beta = (\beta_1, \ldots, \beta_{\mathrm{M}}) \in \mathbb{R}^{\mathrm{M}}$ and $\Sigma \in \mathbb{R}^{\mathrm{M} \times \mathrm{M}}$ are variational parameters. Further $\mathrm{Z} = (\mathrm{Z}_1, \ldots, \mathrm{Z_M})$ with $\{\mathrm{Z_m}\}_{\mathrm{m}=1}^{\mathrm{M}} \overset{\mathrm{iid}}{\sim} \widehat{\rho}$.
- Decoupled SVGPs [Cheng and Boots, 2017]: Same kernel r as in SVGP but mean

$$\mathrm{m_Q(x)} := \mathrm{m_P(x)} + \sum_{\mathrm{n}=1}^{\tilde{\mathrm{N}}} \beta_{\mathrm{n}} \mathrm{k_n(x)}, \tag{32}$$

  where $\widetilde{\mathrm{N}} > \mathrm{M}$.

UNIVERSITY OF OXFORD

GWI-net

GWI-net

Use neural net for posterior mean

GWI-net

Use neural net for posterior mean

- Let $L \in \mathbb{N}$ be the number of hidden layers.
- Let D_ℓ, $\ell = 0, \ldots, L+1$ be the width of layer ℓ with $D_0 := D$.
- Define $g^1(x) := W^1 x + b^1$ and further

$$h^\ell(x) := \phi\big(g^\ell(x)\big), \tag{33}$$

$$g^{\ell+1}(x) := W^{\ell+1} h^\ell(x) + b^{\ell+1} \tag{34}$$

for $x \in \mathcal{X}$ where ϕ is an activation function.

GWI-net

Use neural net for posterior mean

- Let $\mathrm{L} \in \mathbb{N}$ be the number of hidden layers.
- Let D_ℓ, $\ell = 0, \dots, \mathrm{L}+1$ be the width of layer ℓ with $\mathrm{D}_0 := \mathrm{D}$.
- Define $\mathrm{g}^1(\mathrm{x}) := \mathrm{W}^1\mathrm{x} + \mathrm{b}^1$ and further

$$\mathrm{h}^\ell(\mathrm{x}) := \phi\big(\mathrm{g}^\ell(\mathrm{x})\big), \tag{33}$$

$$\mathrm{g}^{\ell+1}(\mathrm{x}) := \mathrm{W}^{\ell+1}\mathrm{h}^\ell(\mathrm{x}) + \mathrm{b}^{\ell+1} \tag{34}$$

  for $\mathrm{x} \in \mathcal{X}$ where ϕ is an activation function.
- Define

$$\mathrm{m}_\mathrm{Q}(\mathrm{x}) := \mathrm{g}^{\mathrm{L}+1}(\mathrm{x}) \tag{35}$$

  for $\mathrm{x} \in \mathcal{X}$.

UNIVERSITY OF OXFORD

GWI-net

Use neural net for posterior mean

- Let $L \in \mathbb{N}$ be the number of hidden layers.
- Let D_ℓ, $\ell = 0, \dots, L+1$ be the width of layer ℓ with $D_0 := D$.
- Define $g^1(x) := W^1 x + b^1$ and further

$$h^\ell(x) := \phi\big(g^\ell(x)\big), \tag{33}$$

$$g^{\ell+1}(x) := W^{\ell+1} h^\ell(x) + b^{\ell+1} \tag{34}$$

  for $x \in \mathcal{X}$ where ϕ is an activation function.
- Define

$$m_Q(x) := g^{L+1}(x) \tag{35}$$

  for $x \in \mathcal{X}$.

and the SVGP kernel r in (31) for the posterior covariance.

UNIVERSITY OF OXFORD

Contents

UNIVERSITY OF OXFORD

Toy Examples: GWI-net on 1-D data

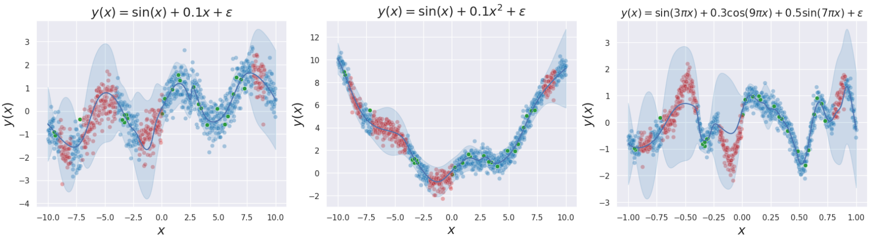
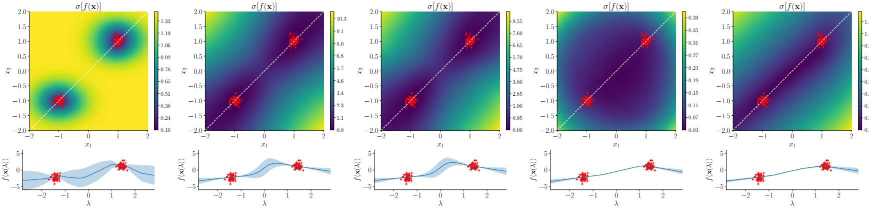


Figure 1: ■ : Training data ■ : Unseen data ■ : Inducing points

We use N = 1000 equidistant points and add white noise with $\epsilon \sim \mathcal{N}(0, 0.5^2)$.
The plot shows $\mathrm{m_Q(x)} \pm 1.96\sqrt{\mathbb{V}[\mathrm{Y^*(x)|Y}]}$ where $\mathbb{V}[\mathrm{Y^*(x)|Y}]$ is the posterior predictive variance given as $\mathrm{r(x,x)} + \sigma^2$.

Toy Examples: GWI-net and "in-between" uncertainty



(a) GWI (b) Inf-width limit GP (c) HMC (d) MFVI (e) MCDO

Figure 2: Regression on a 2D synthetic dataset (red crosses). The colour plots show the standard deviation of the output, $\sigma[\mathrm{f}(\mathbf{x})]$, in 2D input space. The plots beneath show the mean with 2-standard deviation bars along the dashed white line (parameterised by λ). MFVI and MCDO are overconfident for $\lambda \in [-1, 1]$.

UCI Regression

UCI Regression

Dataset	N	D	GWI SVGP	GWI DNN-SVGP	FVI	VIP-BNN	VIP-NP	BBB	VDO	$\alpha = 0.5$	FBNN	**EXACT GP**
BOSTON	506	13	2.8±0.31	**2.27±0.06**	2.33±0.04	2.45±0.04	2.45±0.03	2.76±0.04	2.63±0.10	2.45±0.02	2.30±0.10	2.46±0.04
CONCRETE	1030	8	3.24±0.09	**2.64±0.06**	2.88±0.06	3.02±0.02	3.13±0.02	3.28±0.01	3.23±0.01	3.06±0.03	3.09±0.01	3.05±0.02
ENERGY	768	8	1.81±0.19	0.91±0.12	0.58±0.05	**0.56±0.04**	0.60±0.03	2.17±0.02	1.13±0.02	0.95±0.09	0.68±0.02	0.54±0.02
KIN8NM	8192	8	-0.86±0.38	**-1.2±0.03**	-1.15±0.01	-1.12±0.01	-1.05±0.00	-0.81±0.01	-0.83±0.01	-0.92±0.02	N/A±0.00	N/A±0.00
POWER	9568	4	3.35±0.22	2.74±0.02	**2.69±0.00**	2.92±0.00	2.90±0.00	2.83±0.01	2.88±0.00	2.81±0.00	N/A±0.00	N/A±0.00
PROTEIN	45730	9	**2.84±0.04**	2.87±0.0	2.85±0.00	2.87±0.00	2.96±0.02	3.00±0.00	2.99±0.00	2.90±0.00	N/A±0.00	N/A±0.00
RED WINE	1588	11	0.97±0.02	**0.76±0.08**	0.97±0.06	0.97±0.02	1.20±0.04	1.01±0.02	0.97±0.02	1.01±0.02	1.04±0.01	0.26±0.03
YACHT	308	6	2.37±0.55	0.29±0.1	0.59±0.11	**-0.02±0.07**	0.59±0.13	1.11±0.04	1.22±0.18	0.79±0.11	1.03±0.03	0.10±0.05
NAVAL	11934	16	**-7.25±0.08**	-6.76±0.1	-7.21±0.06	-5.62±0.04	-4.11±0.00	-2.80±0.00	-2.80±0.00	-2.97±0.14	-7.13±0.02	N/A±0.00
Mean Rank			5.5	**2.06**	2.22	3.33	4.94	7	6.11	4.83		

Table 1: The table shows the average test NLL on several UCI regression datasets. We train on random 90% of the data and predict on 10%. This is repeated 10 times and we report mean and standard deviation. The results for our competitors are taken from Ma and Hernández-Lobato [2021].

UNIVERSITY OF OXFORD

Classification

Classification

Model	FMNIST			CIFAR 10		
	Accuracy	NLL	OOD-AUC	Accuracy	NLL	OOD-AUC
GWI-net	**93.25 ±0.09**	**0.250 ±0.00**	**0.959 ±0.01**	**83.82 ±0.00**	**0.553 ±0.00**	0.618 ±0.00
FVI	91.60±0.14	0.254±0.05	0.956±0.06	77.69 ±0.64	0.675±0.03	0.883±0.04
MFVI	91.20±0.10	0.343±0.01	0.782±0.02	76.40±0.52	1.372±0.02	0.589±0.01
MAP	91.39±0.11	0.258±0.00	0.864±0.00	77.41±0.06	0.690±0.00	0.809±0.01
KFAC-LAPLACE	84.42±0.12	0.942±0.01	0.945±0.00	72.49±0.20	1.274±0.01	0.548±0.01
RITTER et al.	91.20±0.07	0.265±0.00	0.947±0.00	77.38±0.06	0.661±0.00	0.796±0.00

Table 2: We report average accuracy, NLL and OOD-AUC on test data for 10 different train/test splits.

References I

Radford M Neal. Bayesian learning for neural networks, volume 118. Springer Science & Business Media, 2012.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In International conference on machine learning, pages 1683–1691. PMLR, 2014.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In Proceedings of the 28th international conference on machine learning (ICML-11), pages 681–688. Citeseer, 2011.

Andrew Foong, David Burt, Yingzhen Li, and Richard Turner. On the expressiveness of approximate inference in bayesian neural networks. Advances in Neural Information Processing Systems, 33:15897–15908, 2020.

References II

Chao Ma, Yingzhen Li, and José Miguel Hernández-Lobato. Variational implicit processes. In International Conference on Machine Learning, pages 4222–4233. PMLR, 2019.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational bayesian neural networks. arXiv preprint arXiv:1903.05779, 2019.

Tim GJ Rudner, Zonghao Chen, and Yarin Gal. Rethinking function-space variational inference in bayesian neural networks. In Third Symposium on Advances in Approximate Bayesian Inference, 2020.

Chao Ma and José Miguel Hernández-Lobato. Functional variational inference based on stochastic process generators. Advances in Neural Information Processing Systems, 34, 2021.

References III

David R Burt, Sebastian W Ober, Adrià Garriga-Alonso, and Mark van der Wilk. Understanding variational inference in function-space. arXiv preprint arXiv:2011.09421, 2020.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized variational inference: Three arguments for deriving new posteriors. arXiv preprint arXiv:1904.02063, 2019.

Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. Mathematische Nachrichten, 147(1):185–203, 1990.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In Artificial intelligence and statistics, pages 567–574. PMLR, 2009.

Ching-An Cheng and Byron Boots. Variational inference for gaussian process models with linear complexity. Advances in Neural Information Processing Systems, 30, 2017.