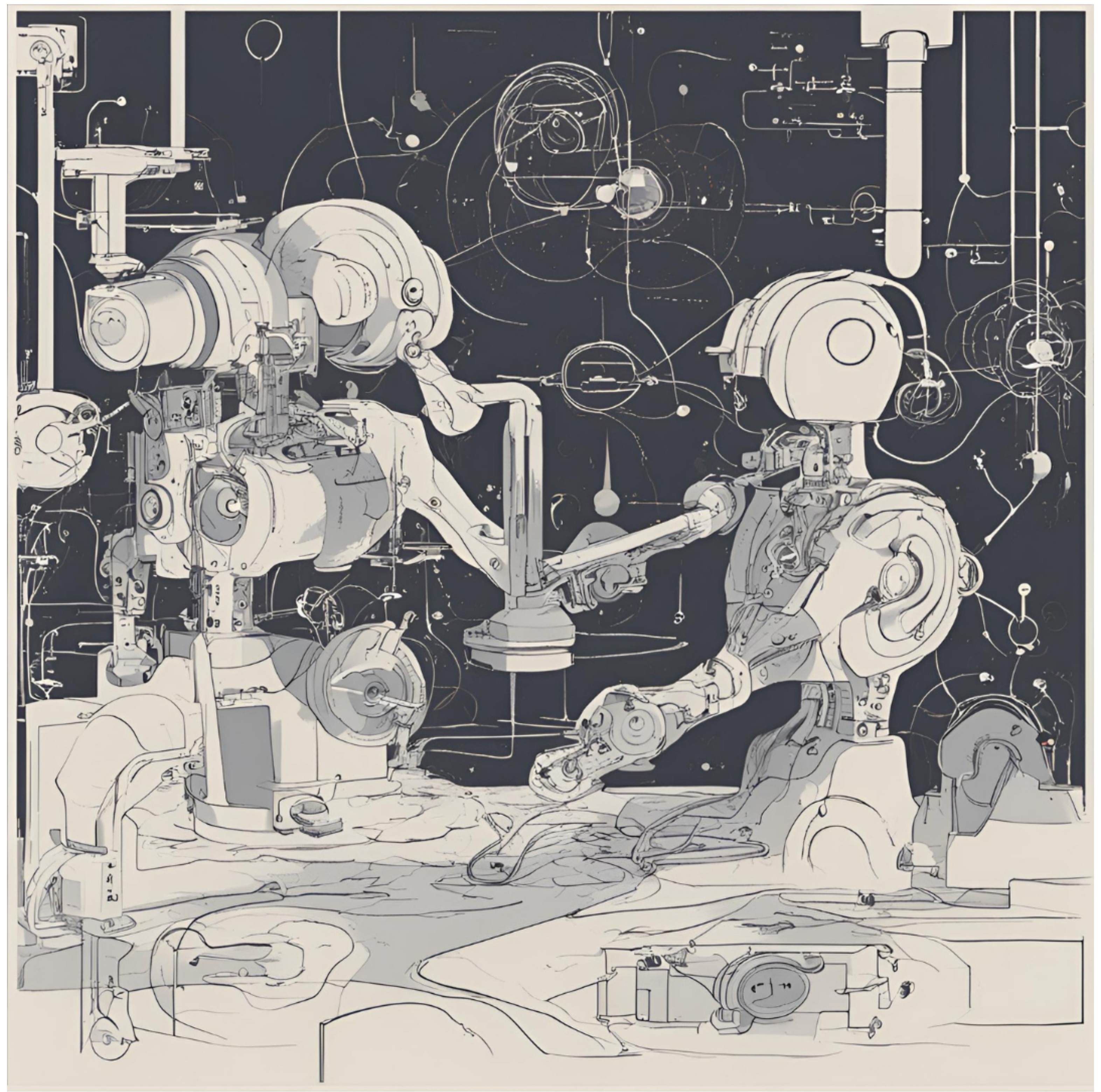
Post-Bayesian Machine Learning



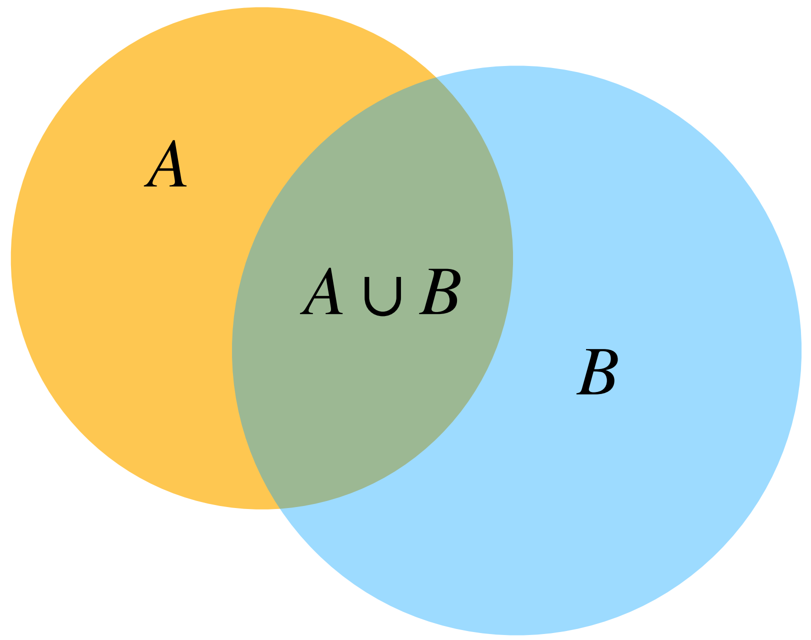
Jeremias Knoblauch; Lecturer & EPSRC Fellow @ UCL Stats

25/07/24

Preamble: Bayesian ML

Bayes' Theorem: Inversion of conditionals

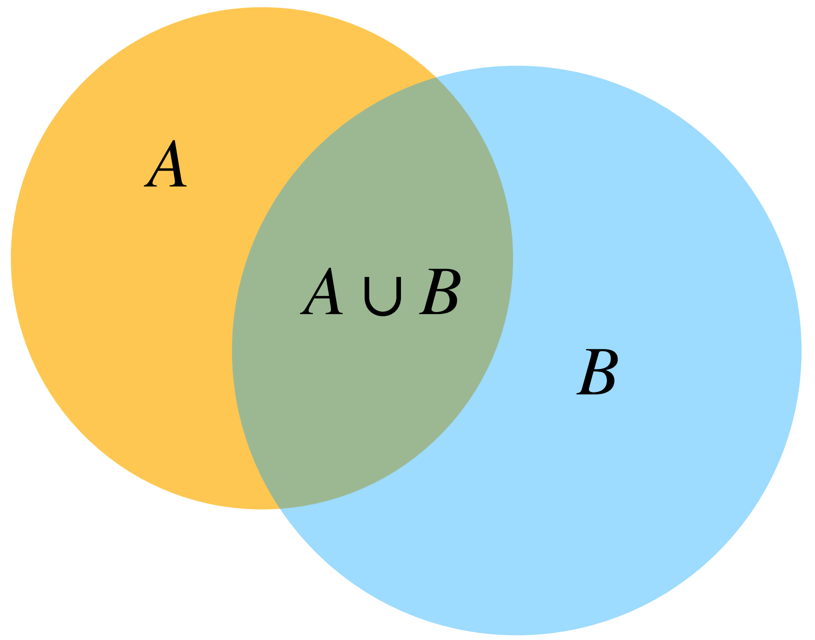
$$\boxed{P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}}$$



Preamble: Bayesian ML

Bayes' Theorem: Inversion of conditionals

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$



Data model: $p(x_{1:n} \mid \theta)$

$x_{1:n} \in \mathcal{X}^n$

Prior probability: $\pi(\theta)$

$\theta \in \Theta$

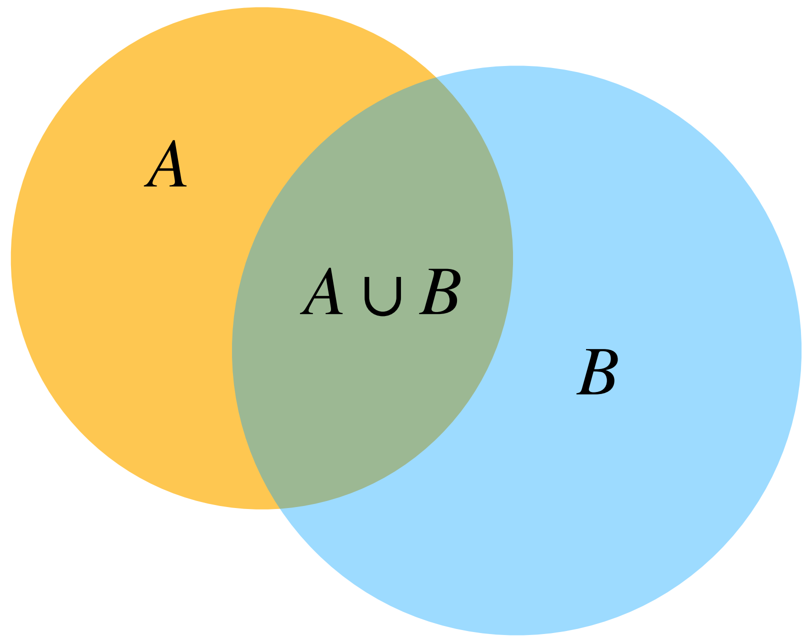
$$\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}$$

(Bayes) Posterior

Preamble: Bayesian ML

Bayes' Theorem: Inversion of conditionals

$$\boxed{P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}}$$



Data model: $p(x_{1:n} \mid \theta)$

$x_{1:n} \in \mathscr{X}^n$

Prior probability: $\pi(\theta)$

$\theta \in \Theta$

$$\boxed{\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}}$$

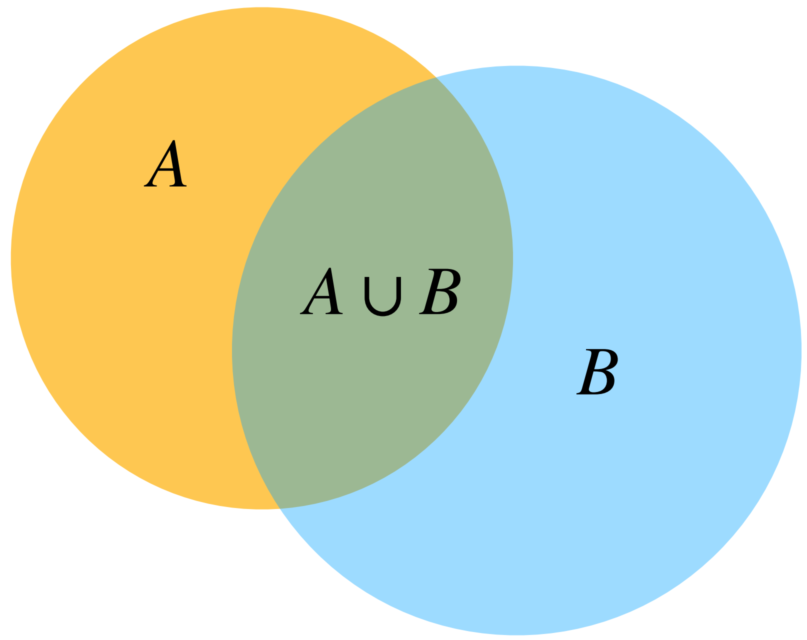
(Bayes) Posterior

- ⊕ Averages models (instead of picking only one)
- ⊕ Quantifies uncertainty about θ via $\pi_n(\theta \mid x_{1:n})$
- ⊕ Inclusion of domain expertise via prior π

Preamble: Bayesian ML

Bayes' Theorem: Inversion of conditionals

$$\boxed{P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}}$$



Data model: $p(x_{1:n} \mid \theta)$

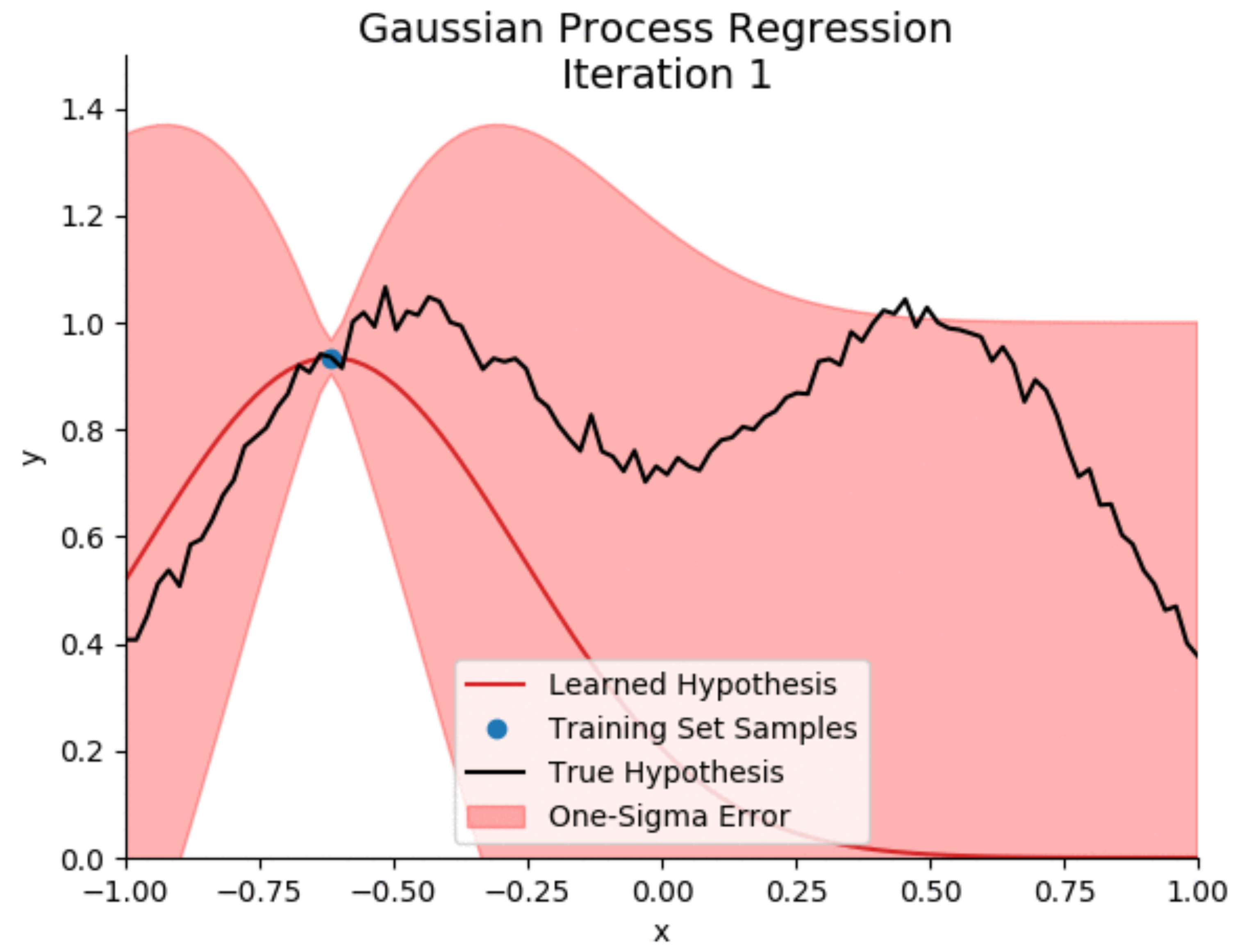
$x_{1:n} \in \mathcal{X}^n$

Prior probability: $\pi(\theta)$

$\theta \in \Theta$

$$\boxed{\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}}$$

(Bayes) Posterior



- (+) Averages models (instead of picking only one)
- (+) Quantifies uncertainty about θ via $\pi_n(\theta \mid x_{1:n})$
- (+) Inclusion of domain expertise via prior π

Preamble: Bayesian ML

Mathematical Foundations

1764—1786: **Bayes' Theorem**

1930: **DeFinetti's Representation Theorem**

1950s/60s: **Savage Axioms, Birnbaum's Likelihood Principle, …**

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Preamble: Bayesian ML

Mathematical Foundations

1764—1786: **Bayes' Theorem**

1930: **DeFinetti's Representation Theorem**

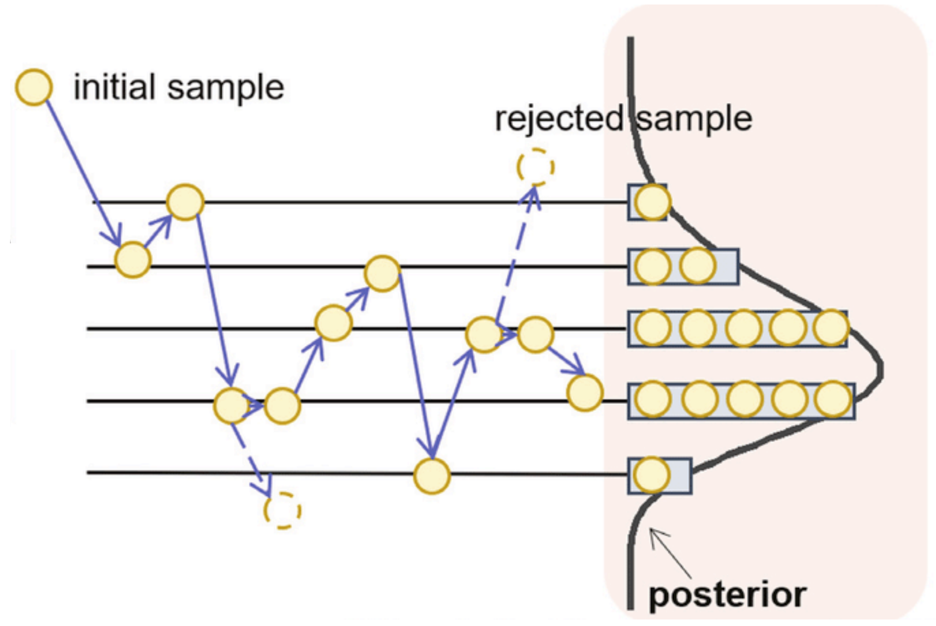
1950s/60s: **Savage Axioms, Birnbaum's Likelihood Principle, …**

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Computation

1960s/70s: **Computational Principles** for Bayesian computation

1980/90s onwards: computation becomes **feasible**



Preamble: Bayesian ML

Mathematical Foundations

1764—1786: **Bayes' Theorem**

1930: **DeFinetti's Representation Theorem**

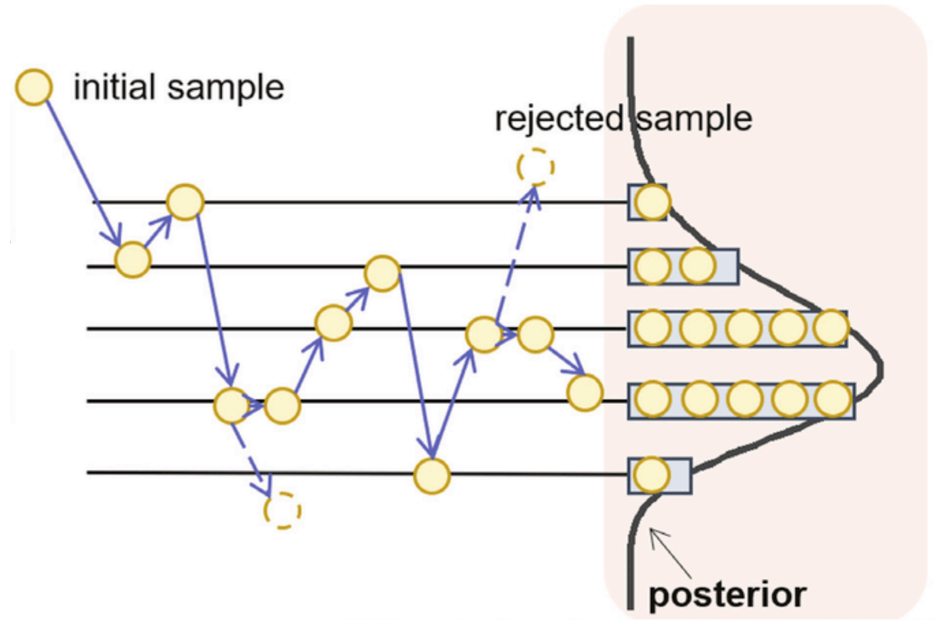
1950s/60s: **Savage Axioms, Birnbaum's Likelihood Principle, …**

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Computation

1960s/70s: **Computational Principles** for Bayesian computation

1980/90s onwards: computation becomes **feasible**

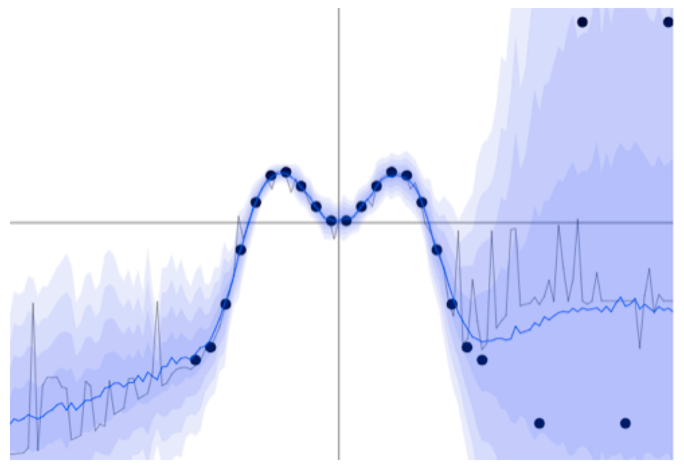


Bayesian ML

1990s: **ML** becomes **data-driven**

Mid 1990s—mid 2000s: **Bayesian ML** emerges

Classic perspective on Bayesian ML



Preamble: Bayesian ML

Mathematical Foundations

1764—1786: **Bayes' Theorem**

1930: **DeFinetti's Representation Theorem**

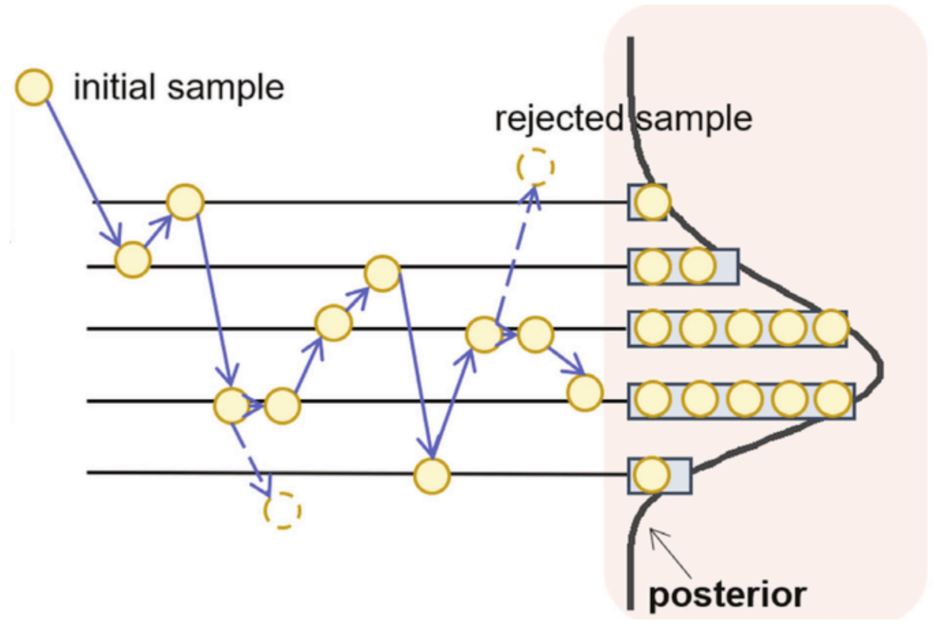
1950s/60s: **Savage Axioms, Birnbaum's Likelihood Principle, …**

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Computation

1960s/70s: **Computational Principles** for Bayesian computation

1980/90s onwards: computation becomes **feasible**



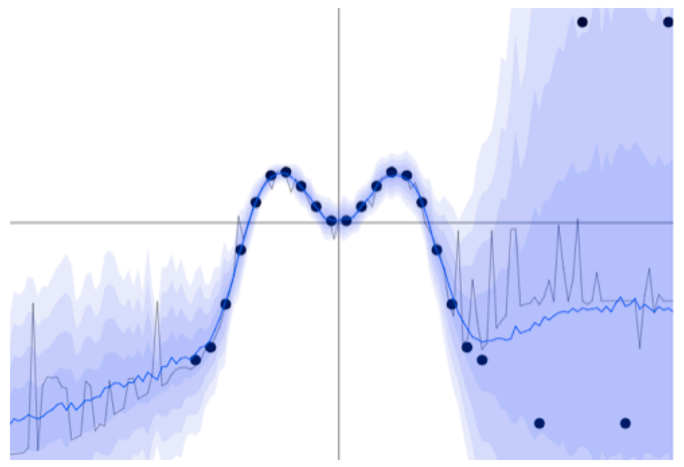
Bayesian ML

1990s: **ML** becomes **data-driven**

Mid 1990s—mid 2000s: **Bayesian ML** emerges

Classic perspective on Bayesian ML

2000s—present: **Adaptations of Bayesian principles** for ML/DL



Post-Bayesian ML and Today's Talk

Today's Talk in a Nutshell

Pitch for ML via Bayes' Rule:

⊕ Uncertainty Quantification

⊕ Inclusion of Prior Knowledge

⊕ Model averaging

Today's Talk in a Nutshell

Pitch for ML via Bayes' Rule:
- ⊕ Uncertainty Quantification
- ⊕ Inclusion of Prior Knowledge
- ⊕ Model averaging

Problem:
- ⊖ ML violates underlying assumptions

$\Longrightarrow$ Unreliable & not robust

Today’s Talk in a Nutshell

Pitch for ML via Bayes’ Rule:

$\oplus$ Uncertainty Quantification

$\oplus$ Inclusion of Prior Knowledge

$\oplus$ Model averaging

Fix: Post-Bayesian ML

Algorithms with ‘Bayesian characteristics’ that are **not using Bayes’ Rule**

Problem:

$\ominus$ ML violates underlying assumptions

$\Longrightarrow$ Unreliable & not robust

Foundations of Bayesian ML

Mathematical Foundations

Reviewer 2: *"The Bayes posterior is principled, and grounded in foundational work of DeFinetti, Savage, Birnbaum, and others. This work gives up these guarantees [...] without clear gain in return."*

1764—1786: **Bayes' Theorem**

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

1930: **DeFinetti's Representation Theorem**

1950s/60s: **Savage Axioms, Birnbaum's Likelihood Principle, ...**

Computation

1960s/70s: **Computational Principles** (Markov chain Monte Carlo)

1980s onwards: computation becomes **feasible**

Bayesian ML

1990s: **ML** becomes **data-driven**

Classic perspective on Bayesian ML

Mid 1990s—mid 2000s: **Bayesian ML** emerges

2000s—present: **Adaptations of Bayesian principles** for ML/DL

Post-Bayesian ML and Today's Talk

Assumptions needed for Foundations

(A1) $\boxed{x_{1:n} \sim p(x_{1:n} \mid \theta^*) \text{ for some } \theta^* \in \Theta}$

$\Theta \;=\;$ Only relevant State of the world

Assumptions needed for Foundations

(A1) $\boxed{x_{1:n} \sim p(x_{1:n} \mid \theta^*) \text{ for some } \theta^* \in \Theta}$

$\Theta \;=\;$ Only relevant State of the world

(A2) $\boxed{\pi(\theta) = \text{ uncertainty about the true State of the world}}$

How rational decision-makers choose the prior

Assumptions needed for Foundations

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

(A1) $\boxed{x_{1:n} \sim p(x_{1:n} \mid \theta^*) \text{ for some } \theta^* \in \Theta}$

Θ = Only relevant State of the world

(A2) $\boxed{\pi(\theta) = \text{ uncertainty about the true State of the world}}$

How rational decision-makers choose the prior

(A3) $\boxed{\pi_n(\theta \mid x_{1:n}) \text{ computable in practice}}$

Guarantees real-world relevance

Assumptions needed for Foundations

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

(A1) $\boxed{x_{1:n} \sim p(x_{1:n} \mid \theta^*) \text{ for some } \theta^* \in \Theta}$

Θ = Only relevant State of the world

(A2) $\boxed{\pi(\theta) = \text{uncertainty about the true State of the world}}$

How rational decision-makers choose the prior

FRAGILE

(A3) $\boxed{\pi_n(\theta \mid x_{1:n}) \text{ computable in practice}}$

Guarantees real-world relevance

Case Study: Regression with Boston Housing Data

Traditional Bayesian analysis in science

(A1) model well-specified

(A2) prior well-specified

(A3) computationally feasible

Case Study: Regression with Boston Housing Data

Traditional Bayesian analysis in science

Expert with research question $\longrightarrow$ Statistical modelling & expert knowledge

$x_{1:n}$ $\longrightarrow$ $p(x_{1:n} \mid \theta), \pi(\theta)$

(A1) model well-specified

(A2) prior well-specified

(A3) computationally feasible

Case Study: Regression with Boston Housing Data

Traditional Bayesian analysis in science

Expert with research question	Statistical modelling & expert knowledge	Inference
$x_{1:n}$	$p(x_{1:n} \mid \theta), \pi(\theta)$	$\pi_n(\theta \mid x_{1:n})$

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

Case Study: Regression with Boston Housing Data

Traditional Bayesian analysis in science

Expert with research question		Statistical modelling & expert knowledge		Inference
$x_{1:n}$	$\longrightarrow$	$p(x_{1:n} \mid \theta), \pi(\theta)$	$\longrightarrow$	$\pi_n(\theta \mid x_{1:n})$

Harrison & Rubinfeld (1978)

Research Question: influence of air pollution on house prices?

(A1) model well-specified

(A2) prior well-specified

(A3) computationally feasible

Case Study: Regression with Boston Housing Data

Traditional Bayesian analysis in science

Expert with research question	Statistical modelling & expert knowledge	Inference
$x_{1:n}$	$p(x_{1:n} \mid \theta), \pi(\theta)$	$\pi_n(\theta \mid x_{1:n})$

Harrison & Rubinfeld (1978)

Research Question: influence of air pollution on house prices?

(A1)

parameters of interest

incidental parameters

$$\log y_i = \sum_{j=1}^{J_1} p_j \log(x_{j,i}) + c_0 + \sum_{j=J_1}^{J_2} c_j \log(x_{j,i}) + \varepsilon_i$$

willingness to pay

pollutants

rooms, sqm, ...

measurement error

$$\theta = (c_0, c_2, \ldots, c_{J_1}, p_1, p_2 \ldots p_{J_2})^\top$$

(A1) model well-specified

(A2) prior well-specified

(A3) computationally feasible

Case Study: Regression with Boston Housing Data

Traditional Bayesian analysis in science

Expert with research question	Statistical modelling & expert knowledge	Inference
$x_{1:n}$	$p(x_{1:n} \mid \theta), \pi(\theta)$	$\pi_n(\theta \mid x_{1:n})$

Harrison & Rubinfeld (1978)

Research Question: influence of air pollution on house prices?

(A1) ✓

parameters of interest

incidental parameters

$$\log y_i = \sum_{j=1}^{J_1} p_j \log(x_{j,i}) + c_0 + \sum_{j=J_1}^{J_2} c_j \log(x_{j,i}) + \varepsilon_i$$

willingness to pay

pollutants

rooms, sqm, ...

measurement error

$\theta = (c_0, c_2, \ldots, c_{J_1}, p_1, p_2 \ldots p_{J_2})^\top$

$\pi(\theta) \sim$ hand-crafted by experts

(A2) ✓

(A1) model well-specified

(A2) prior well-specified

(A3) computationally feasible

Case Study: Regression with Boston Housing Data

Traditional Bayesian analysis in science

Expert with research question	Statistical modelling & expert knowledge	Inference
$x_{1:n}$	$p(x_{1:n} \mid \theta), \pi(\theta)$	$\pi_n(\theta \mid x_{1:n})$

Harrison & Rubinfeld (1978)

Research Question: influence of air pollution on house prices?

(A1)

parameters of interest

incidental parameters

$$\log y_i = \sum_{j=1}^{J_1} p_j \log(x_{j,i}) + c_0 + \sum_{j=J_1}^{J_2} c_j \log(x_{j,i}) + \varepsilon_i$$

willingness to pay

pollutants

rooms, sqm, ...

measurement error

$$\theta = (c_0, c_2, \ldots, c_{J_1}, p_1, p_2 \ldots p_{J_2})^\top$$

$\pi(\theta) \sim$ hand-crafted by experts (A2)

$\pi_n(\theta \mid x_{1:n}) \longrightarrow$ computed exactly (A3)

(A1) model well-specified

(A2) prior well-specified

(A3) computationally feasible

Case Study: Regression with Boston Housing Data

Traditional Bayesian analysis in science

Expert with research question	Statistical modelling & expert knowledge	Inference
$x_{1:n}$	$p(x_{1:n} \mid \theta), \pi(\theta)$	$\pi_n(\theta \mid x_{1:n})$

Modern Bayesian ML

Flexible model

$p(x_{1:n} \mid \theta), \pi(\theta)$

Harrison & Rubinfeld (1978)

Research Question: influence of air pollution on house prices?

(A1)

parameters of interest

incidental parameters

$$\log y_i = \sum_{j=1}^{J_1} p_j \log(x_{j,i}) + c_0 + \sum_{j=J_1}^{J_2} c_j \log(x_{j,i}) + \varepsilon_i$$

willingness to pay

pollutants

rooms, sqm, ...

measurement error

$\theta = (c_0, c_2, \ldots, c_{J_1}, p_1, p_2 \ldots p_{J_2})^\top$

$\pi(\theta) \sim$ hand-crafted by experts (A2)

$\pi_n(\theta \mid x_{1:n}) \longrightarrow$ computed exactly (A3)

(A1) model well-specified

(A2) prior well-specified

(A3) computationally feasible

Case Study: Regression with Boston Housing Data

Traditional Bayesian analysis in science

Expert with research question → Statistical modelling & expert knowledge → Inference

$x_{1:n}$ → $p(x_{1:n} \mid \theta),\ \pi(\theta)$ → $\pi_n(\theta \mid x_{1:n})$

Modern Bayesian ML

Flexible model → Different data/ problems

$p(x_{1:n} \mid \theta),\ \pi(\theta)$ → $x_{1:n}$

Harrison & Rubinfeld (1978)

Research Question: influence of air pollution on house prices?

(A1) ✓

parameters of interest

incidental parameters

$$\log y_i = \sum_{j=1}^{J_1} p_j \log(x_{j,i}) + c_0 + \sum_{j=J_1}^{J_2} c_j \log(x_{j,i}) + \varepsilon_i$$

willingness to pay

pollutants

rooms, sqm, ...

measurement error

$\theta = (c_0, c_2, \ldots, c_{J_1}, p_1, p_2 \ldots p_{J_2})^\top$

$\pi(\theta) \sim$ hand-crafted by experts (A2) ✓

$\pi_n(\theta \mid x_{1:n}) \longrightarrow$ computed exactly (A3) ✓

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

Case Study: Regression with Boston Housing Data

Traditional Bayesian analysis in science

Expert with research question → Statistical modelling & expert knowledge → Inference

$x_{1:n}$ → $p(x_{1:n} \mid \theta), \pi(\theta)$ → $\pi_n(\theta \mid x_{1:n})$

Modern Bayesian ML

Flexible model → Different data/ problems → Inference

$p(x_{1:n} \mid \theta), \pi(\theta)$ → $x_{1:n}$ → $\pi_n(\theta \mid x_{1:n})$

Harrison & Rubinfeld (1978)

Research Question: influence of air pollution on house prices?

(A1) ✓

parameters of interest

incidental parameters

$$\log y_i = \sum_{j=1}^{J_1} p_j \log(x_{j,i}) + c_0 + \sum_{j=J_1}^{J_2} c_j \log(x_{j,i}) + \varepsilon_i$$

willingness to pay

pollutants

rooms, sqm, ...

measurement error

$\theta = (c_0, c_2, \ldots, c_{J_1}, p_1, p_2 \ldots p_{J_2})^\top$

$\pi(\theta) \sim$ hand-crafted by experts (A2) ✓

$\pi_n(\theta \mid x_{1:n}) \longrightarrow$ computed exactly (A3) ✓

(A1) model well-specified

(A2) prior well-specified

(A3) computationally feasible

Case Study: Regression with Boston Housing Data

Traditional Bayesian analysis in science

Expert with research question	Statistical modelling & expert knowledge	Inference
$x_{1:n}$	$p(x_{1:n} \mid \theta), \pi(\theta)$	$\pi_n(\theta \mid x_{1:n})$

Modern Bayesian ML

Flexible model	Different data/ problems	Inference
$p(x_{1:n} \mid \theta), \pi(\theta)$	$x_{1:n}$	$\pi_n(\theta \mid x_{1:n})$

Harrison & Rubinfeld (1978)

Research Question: influence of air pollution on house prices?

(A1) ✓

parameters of interest

incidental parameters

$$\log y_i = \sum_{j=1}^{J_1} p_j \log(x_{j,i}) + c_0 + \sum_{j=J_1}^{J_2} c_j \log(x_{j,i}) + \varepsilon_i$$

willingness to pay

pollutants

rooms, sqm, ...

measurement error

$\theta = (c_0, c_2, \ldots, c_{J_1}, p_1, p_2 \ldots p_{J_2})^\top$

$\pi(\theta) \sim$ hand-crafted by experts (A2) ✓

$\pi_n(\theta \mid x_{1:n}) \longrightarrow$ computed exactly (A3) ✓

Pearce et al. (2020) [AISTATS]

Research Question: Does my algorithm improve prediction on regression tasks like Boston UCI data?

(A1) model well-specified

(A2) prior well-specified

(A3) computationally feasible

Case Study: Regression with Boston Housing Data

Traditional Bayesian analysis in science

Expert with research question → Statistical modelling & expert knowledge → Inference

$x_{1:n}$ → $p(x_{1:n} \mid \theta), \pi(\theta)$ → $\pi_n(\theta \mid x_{1:n})$

Modern Bayesian ML

Flexible model → Different data/ problems → Inference

$p(x_{1:n} \mid \theta), \pi(\theta)$ → $x_{1:n}$ → $\pi_n(\theta \mid x_{1:n})$

Harrison & Rubinfeld (1978)

Research Question: influence of air pollution on house prices?

(A1) ✓

parameters of interest

incidental parameters

$$\log y_i = \sum_{j=1}^{J_1} p_j \log(x_{j,i}) + c_0 + \sum_{j=J_1}^{J_2} c_j \log(x_{j,i}) + \varepsilon_i$$

willingness to pay

pollutants

rooms, sqm, ...

measurement error

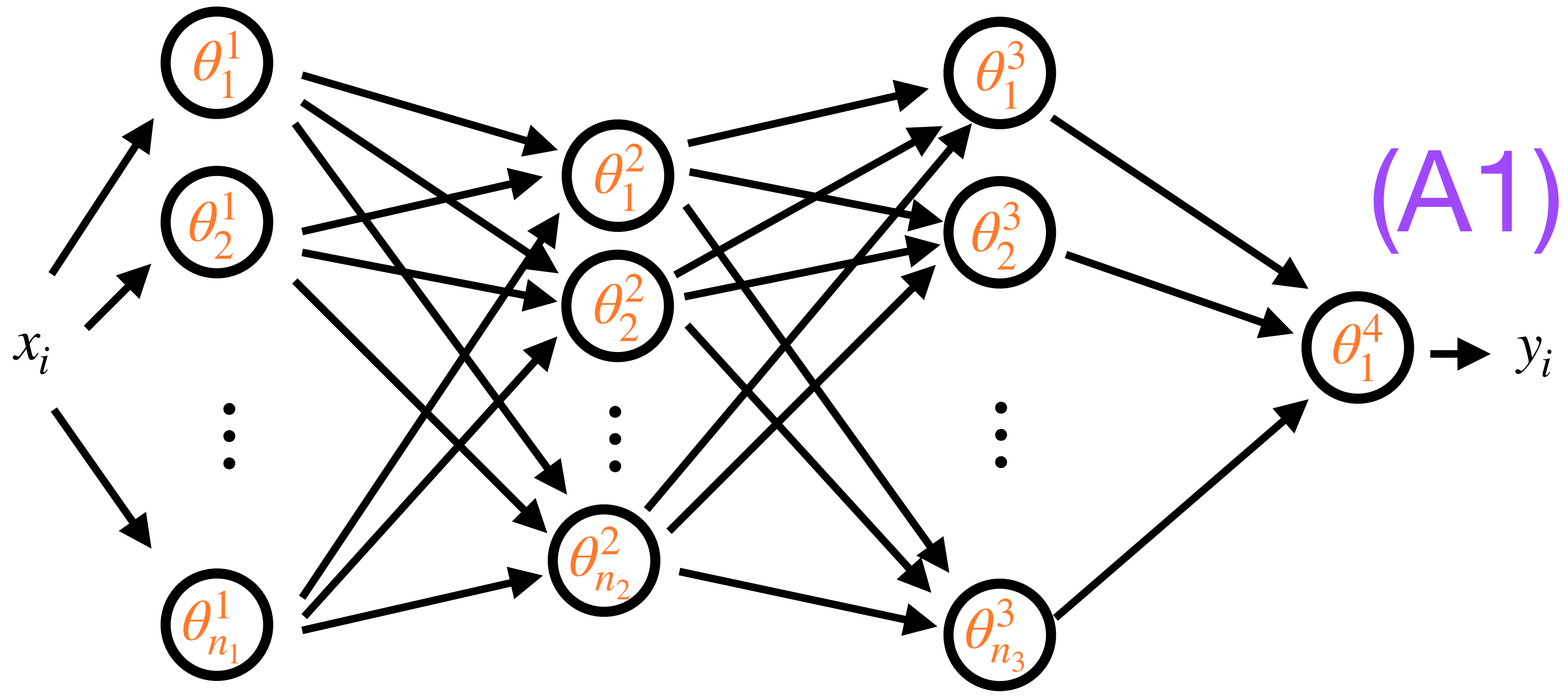
$\theta = (c_0, c_2, \ldots, c_{J_1}, p_1, p_2 \ldots p_{J_2})^\top$

$\pi(\theta) \sim$ hand-crafted by experts (A2) ✓

$\pi_n(\theta \mid x_{1:n}) \longrightarrow$ computed exactly (A3) ✓

Pearce et al. (2020) [AISTATS]

Research Question: Does my algorithm improve prediction on regression tasks like Boston UCI data?



(A1) ✓

F_θ : Input → Output

Case Study: Regression with Boston Housing Data

Traditional Bayesian analysis in science

Expert with research question	Statistical modelling & expert knowledge	Inference
$x_{1:n}$ →	$p(x_{1:n} \mid \theta), \pi(\theta)$ →	$\pi_n(\theta \mid x_{1:n})$

Modern Bayesian ML

Flexible model	Different data/ problems	Inference
$p(x_{1:n} \mid \theta), \pi(\theta)$ →	$x_{1:n}$ →	$\pi_n(\theta \mid x_{1:n})$

Harrison & Rubinfeld (1978)

Research Question: influence of air pollution on house prices?

(A1) ✓

$$\underbrace{\log y_i}_{\text{willingness to pay}} = \sum_{j=1}^{J_1} \overbrace{p_j}^{\text{parameters of interest}} \underbrace{\log(x_{j,i})}_{\text{pollutants}} + c_0 + \sum_{j=J_1}^{J_2} \overbrace{c_j}^{\text{incidental parameters}} \underbrace{\log(x_{j,i})}_{\text{rooms, sqm, ...}} + \underbrace{\varepsilon_i}_{\text{measurement error}}$$

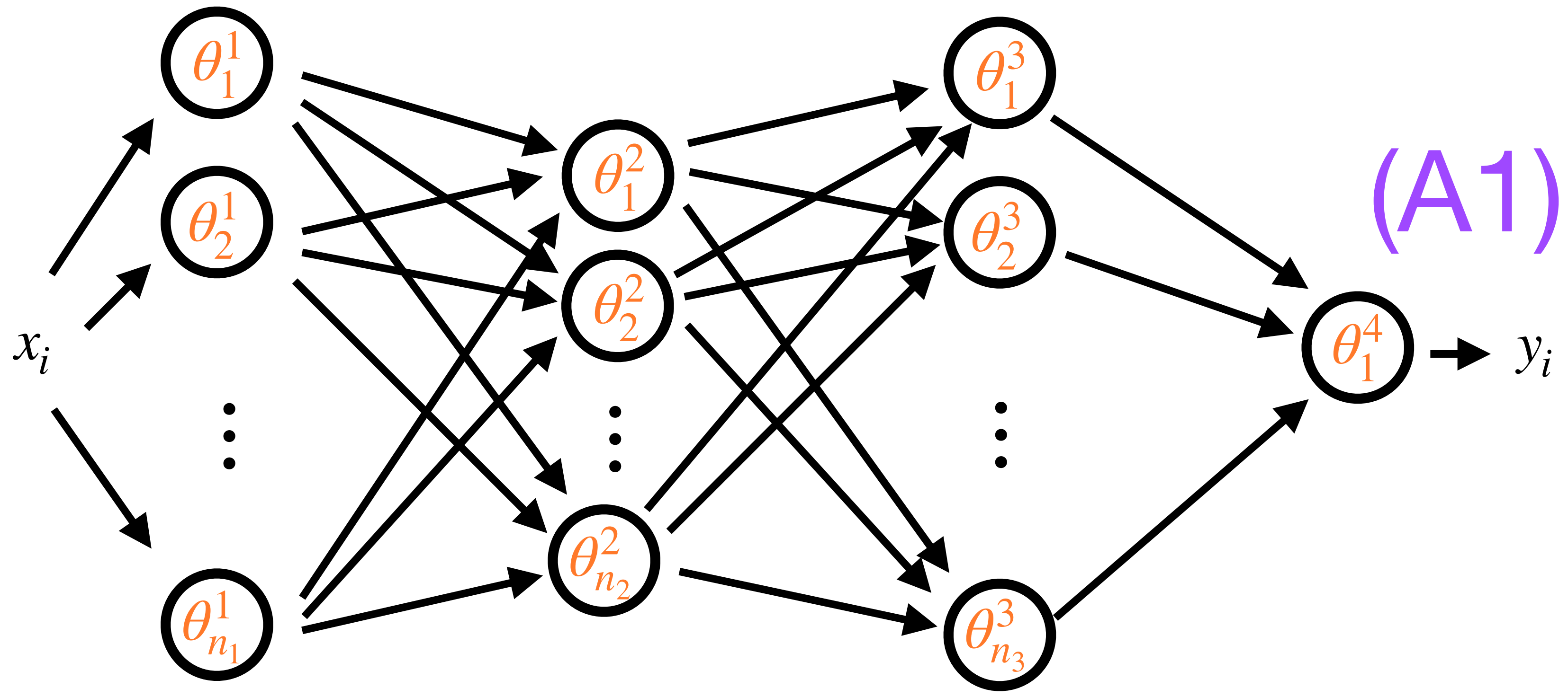
$\theta = (c_0, c_2, \ldots, c_{J_1}, p_1, p_2 \ldots p_{J_2})^\top$

$\pi(\theta) \sim$ hand-crafted by experts (A2) ✓

$\pi_n(\theta \mid x_{1:n}) \longrightarrow$ computed exactly (A3) ✓

Pearce et al. (2020) [AISTATS]

Research Question: Does my algorithm improve prediction on regression tasks like Boston UCI data?



(A1) ✓

F_θ : Input → Output

$\pi(\theta) \sim$ Normal ~~(A2)~~

Case Study: Regression with Boston Housing Data

Traditional Bayesian analysis in science

Expert with research question → Statistical modelling & expert knowledge → Inference

$x_{1:n}$ → $p(x_{1:n} \mid \theta), \pi(\theta)$ → $\pi_n(\theta \mid x_{1:n})$

Modern Bayesian ML

Flexible model → Different data/ problems → Inference

$p(x_{1:n} \mid \theta), \pi(\theta)$ → $x_{1:n}$ → $\pi_n(\theta \mid x_{1:n})$

Harrison & Rubinfeld (1978)

Research Question: influence of air pollution on house prices?

(A1) ✓

parameters of interest

incidental parameters

$$\log y_i = \sum_{j=1}^{J_1} p_j \log(x_{j,i}) + c_0 + \sum_{j=J_1}^{J_2} c_j \log(x_{j,i}) + \varepsilon_i$$

willingness to pay

pollutants

rooms, sqm, ...

measurement error

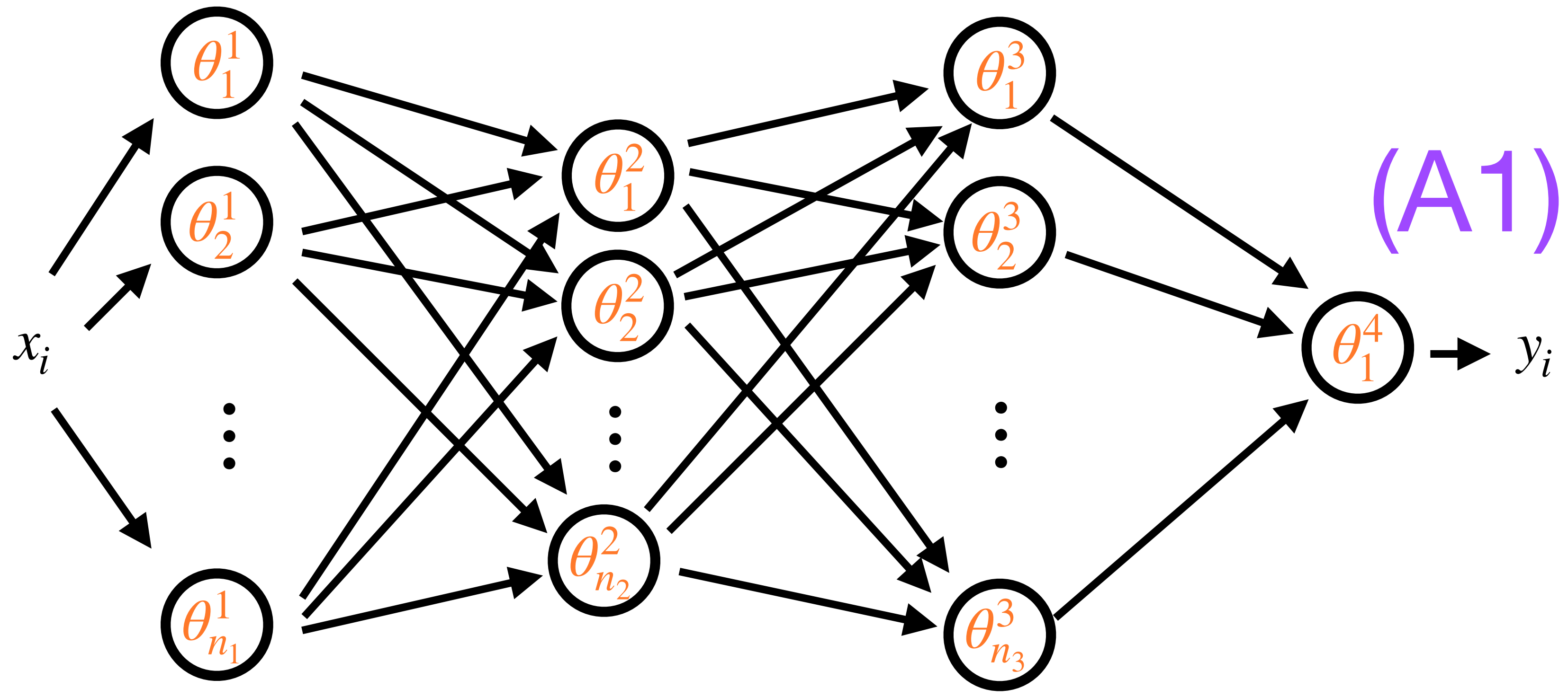
$\theta = (c_0, c_2, \ldots, c_{J_1}, p_1, p_2 \ldots p_{J_2})^\top$

$\pi(\theta) \sim$ hand-crafted by experts (A2) ✓

$\pi_n(\theta \mid x_{1:n}) \longrightarrow$ computed exactly (A3) ✓

Pearce et al. (2020) [AISTATS]

Research Question: Does my algorithm improve prediction on regression tasks like Boston UCI data?



(A1) ✓

F_θ : Input → Output

$\pi(\theta) \sim$ Normal ~~(A2)~~

$\pi_n(\theta \mid x_{1:n}) \longrightarrow$ coarse approximation ~~(A3)~~

Assumptions & Foundations

Traditional Bayesian analysis in science

Expert with research question		Statistical modelling & expert knowledge		Inference
$x_{1:n}$	$\longrightarrow$	$p(x_{1:n} \mid \theta),\ \pi(\theta)$	$\longrightarrow$	$\pi_n(\theta \mid x_{1:n})$

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

Modern Bayesian ML

Flexible model		Different data/ problems		Inference
$p(x_{1:n} \mid \theta),\ \pi(\theta)$	$\longrightarrow$	$x_{1:n}$	$\longrightarrow$	$\pi_n(\theta \mid x_{1:n})$

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

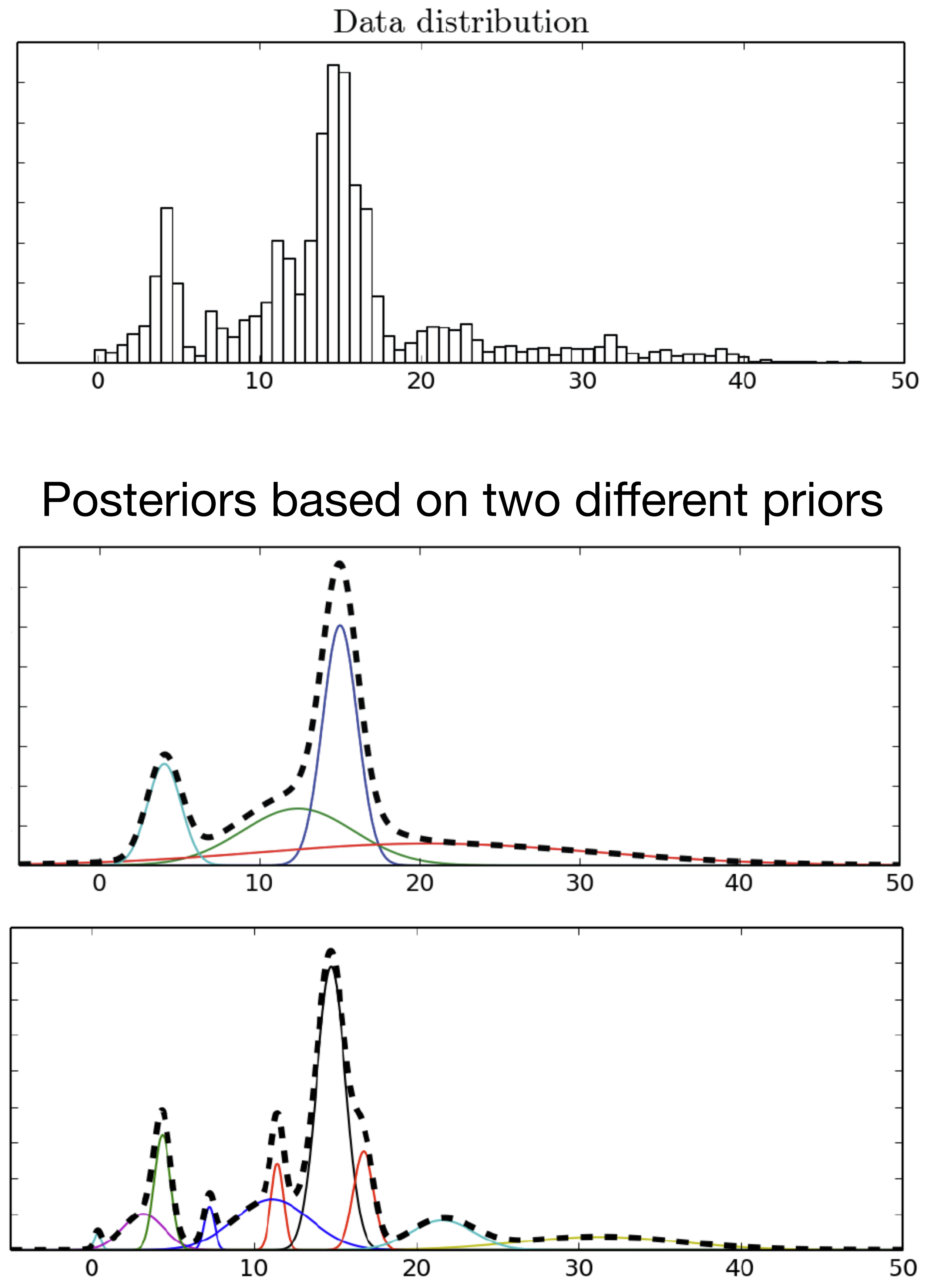
FRAGILE

Ramifications

On the Brittleness of Bayesian Inference
(Ohwadi et al., 2015)

Bayesian inference is generically ill-posed: […]

> *(1) with two […] arbitrarily close models and […] data [one] may reach diametrically opposite conclusions; and*
>
> *(2) any given prior and model can be slightly perturbed to achieve any desired posterior*



Ramifications

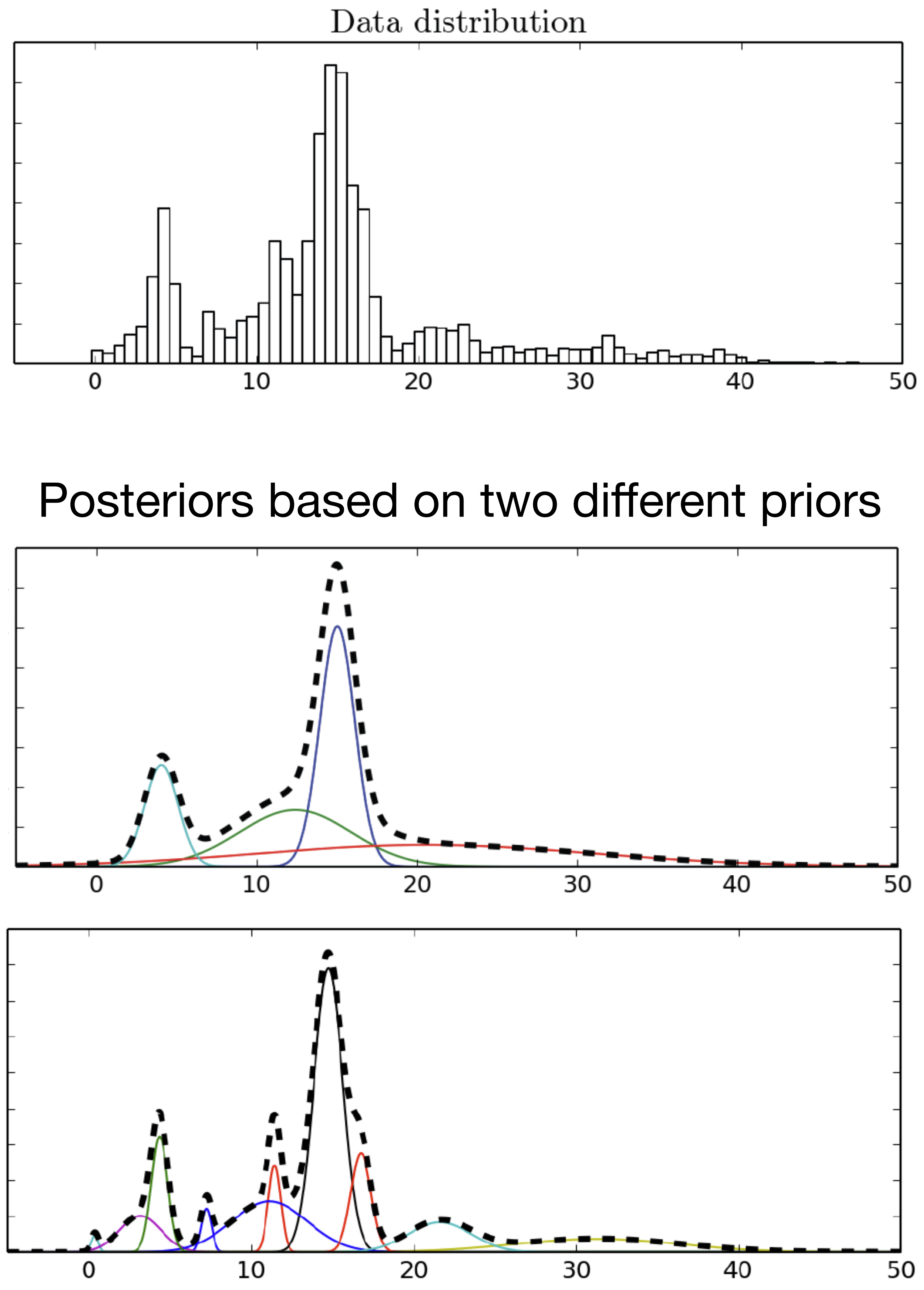
On the Brittleness of Bayesian Inference
(Ohwadi et al., 2015)

Bayesian inference is generically ill-posed: […]

(1) with two […] arbitrarily close models and […] data [one] may reach diametrically opposite conclusions; and

(2) any given prior and model can be slightly perturbed to achieve any desired posterior

Plain English: Bayesian ML is often unreliable!



Ramifications: Post-Bayesian ML

Mathematical Foundations

1764—1786: **Bayes' Theorem**

1930: **DeFinetti's Representation Theorem**

1950s/60s: **Savage Axioms, Birnbaum's Likelihood Principle, …**

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Computation

1960s/70s: **Computational Principles** (Markov chain Monte Carlo)

1980s onwards: computation becomes **feasible**

Bayesian ML

1990s: **ML** becomes **data-driven**

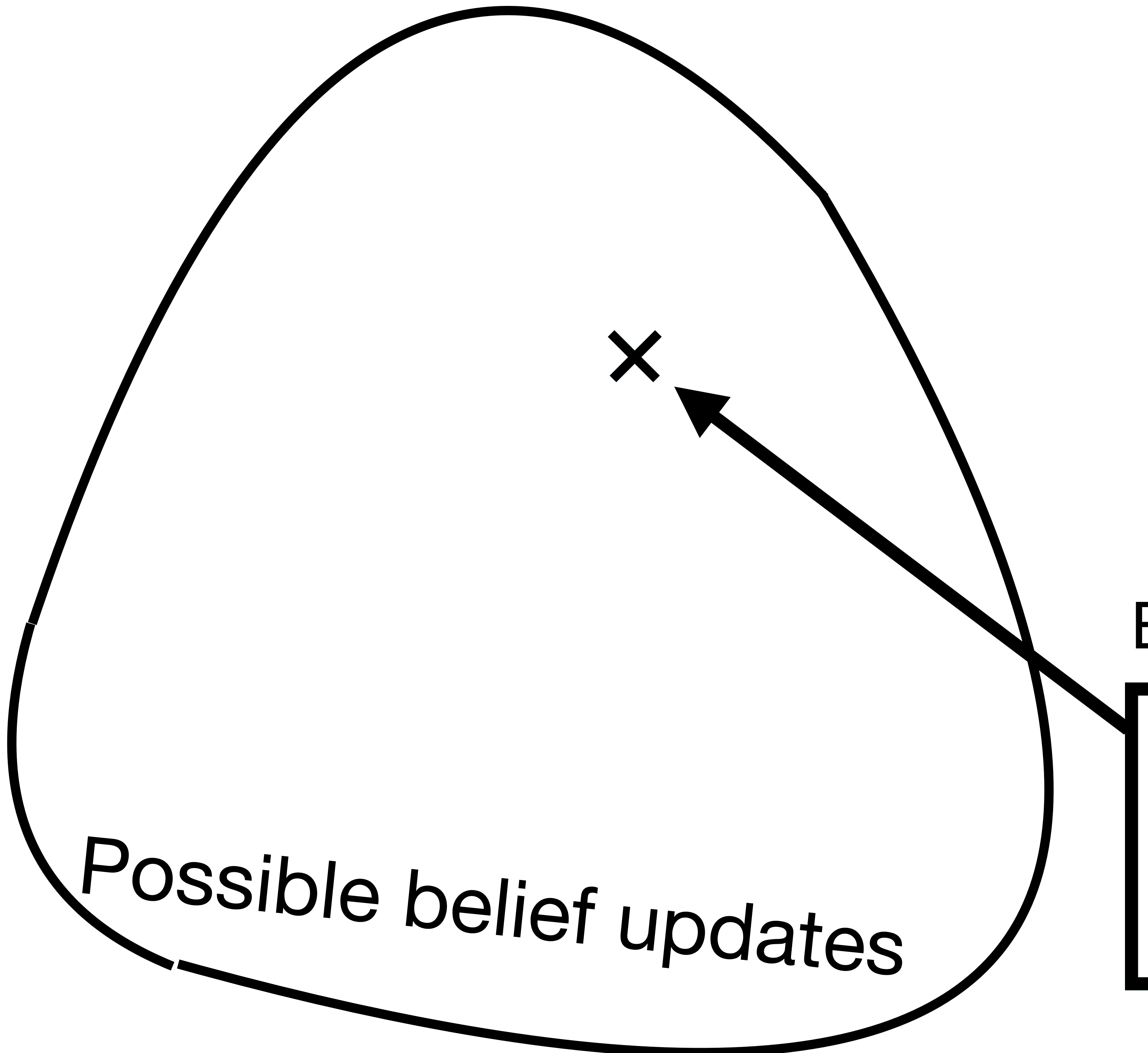
Mid 1990s—mid 2000s: **Bayesian ML** emerges

Classic perspective on Bayesian ML

2000s—present: **Adaptations of Bayesian principles** for ML/DL

Collectively: Post-Bayesian ML

Post-Bayesian ML



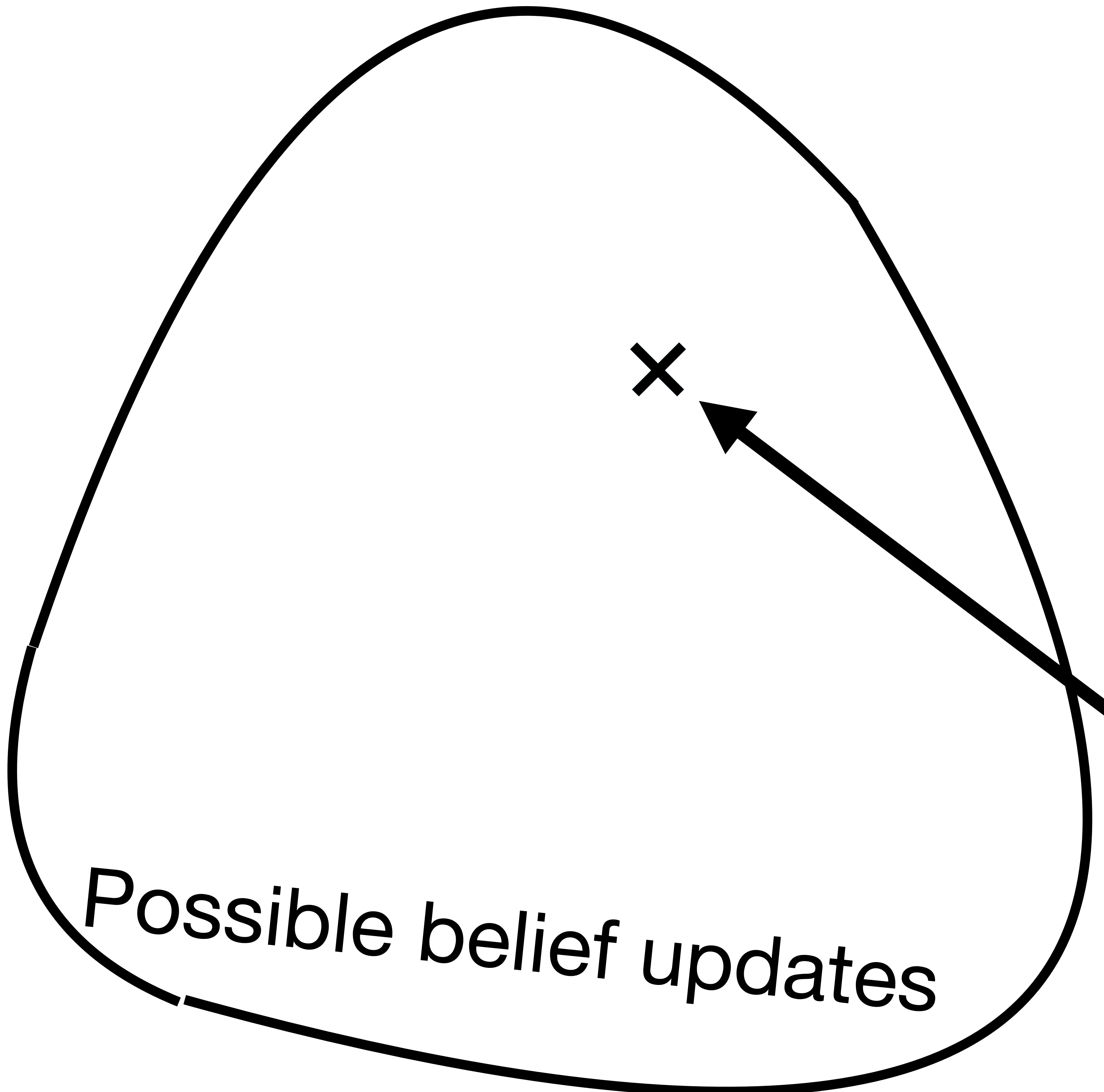
Bayes' Posterior (A1), (A2), (A3)

$$\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}$$

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

Post-Bayesian ML

$$\text{Belief updates} = \Big\{ \mathscr{B} : \overbrace{\mathscr{P}(\Theta)}^{\text{space of priors}} \times \underbrace{\mathscr{X}^n}_{\text{data space}} \times \overbrace{\mathscr{H}}^{\text{space of hyperparameters}} \longrightarrow \underbrace{\mathscr{P}(\Theta)}_{\text{space of posteriors}} \Big\}$$



Bayes' Posterior (A1), (A2), (A3)

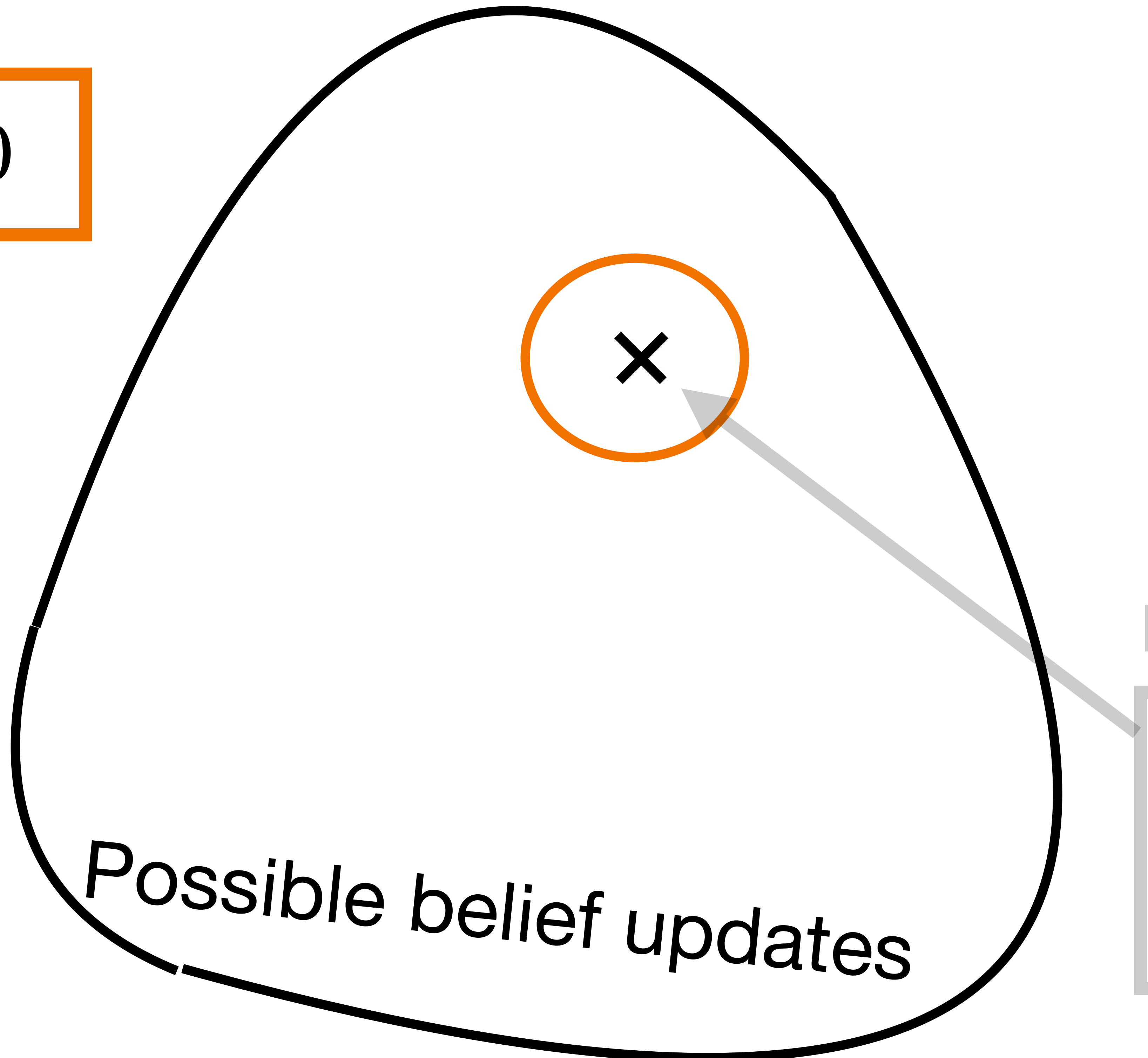
$$\boxed{\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}}$$

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

Post-Bayesian ML

$$\text{Belief updates} = \left\{\mathscr{B} : \overbrace{\mathscr{P}(\Theta)}^{\text{space of priors}} \times \underbrace{\mathscr{X}^n}_{\text{data space}} \times \overbrace{\mathscr{H}}^{\text{space of hyperparameters}} \longrightarrow \underbrace{\mathscr{P}(\Theta)}_{\text{space of posteriors}}\right.$$

$$\boxed{p(x_{1:n} \mid \theta) \longrightarrow p(x_{1:n} \mid \theta)^{\lambda},\ \lambda > 0}$$



Bayes' Posterior (A1), (A2), (A3)

$$\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}$$

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

Post-Bayesian ML

space of priors

space of hyperparameters

$$\text{Belief updates} = \left\{ \mathscr{B} : \mathscr{P}(\Theta) \times \mathscr{X}^n \times \mathscr{H} \longrightarrow \mathscr{P}(\Theta) \right.$$

data space

space of posteriors

Grünwald (2011); COLT
Miller & Dunson (2015); JRSS-B
Bhattacharya, Pati, & Yang (2019); Annals of Statistics
Adlam et al. (2020); preprint
Wenzel et al. (2020); ICML
Aitchison (2021); ICLR
...
McLatchie, Fong, Frazier, & **Knoblauch.** (2024); forthcoming

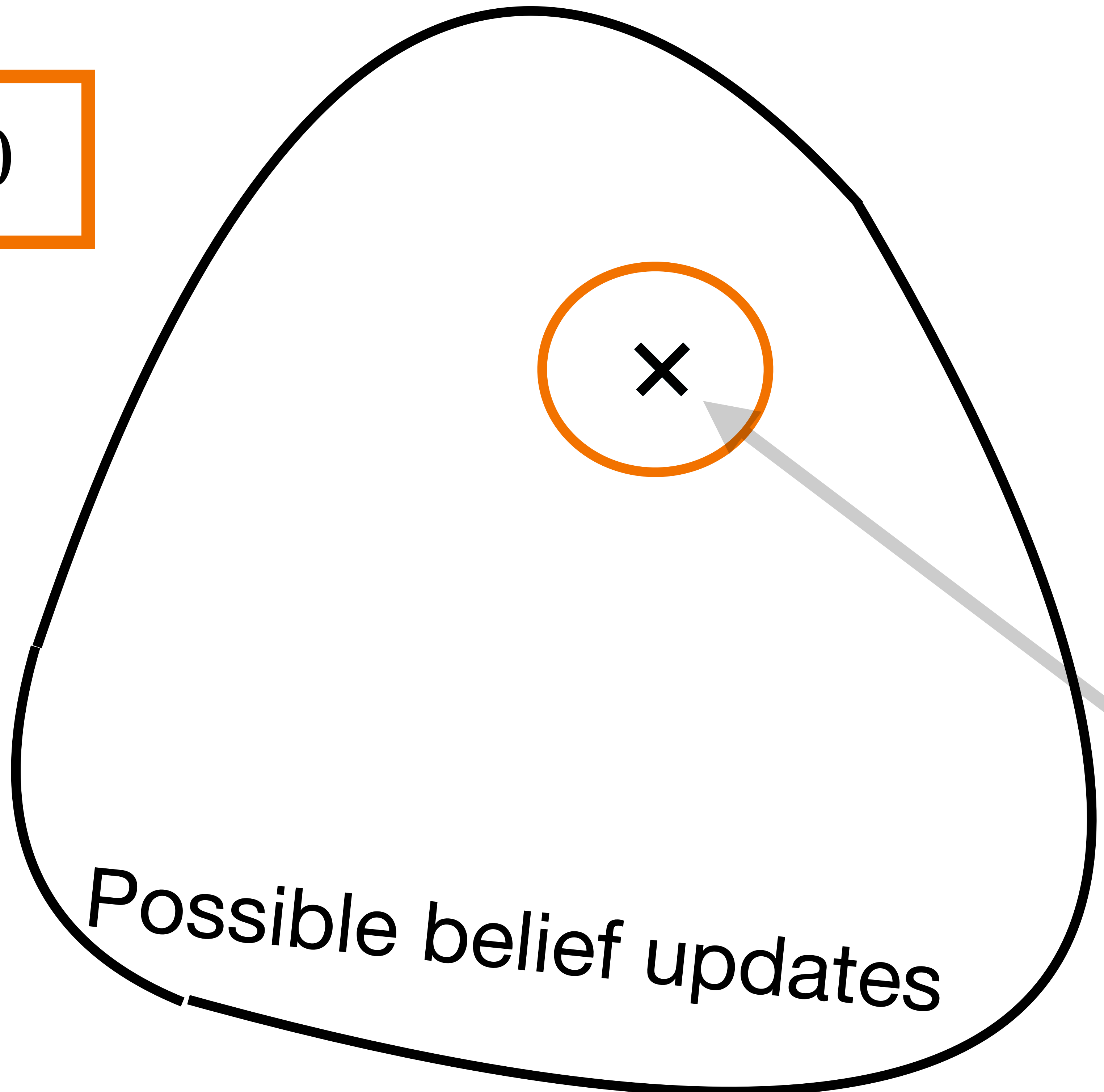
[Generally credited to Grünwald (2011)]

$$\boxed{p(x_{1:n} \mid \theta) \longrightarrow p(x_{1:n} \mid \theta)^{\lambda}, \ \lambda > 0}$$

Power/Fractional/
Cold Posterior

~~(A1)~~, (A2), (A3)

$$\boxed{\pi_n^{(\lambda)}(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta) d\theta}}$$



Bayes' Posterior

(A1), (A2), (A3)

$$\boxed{\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}}$$

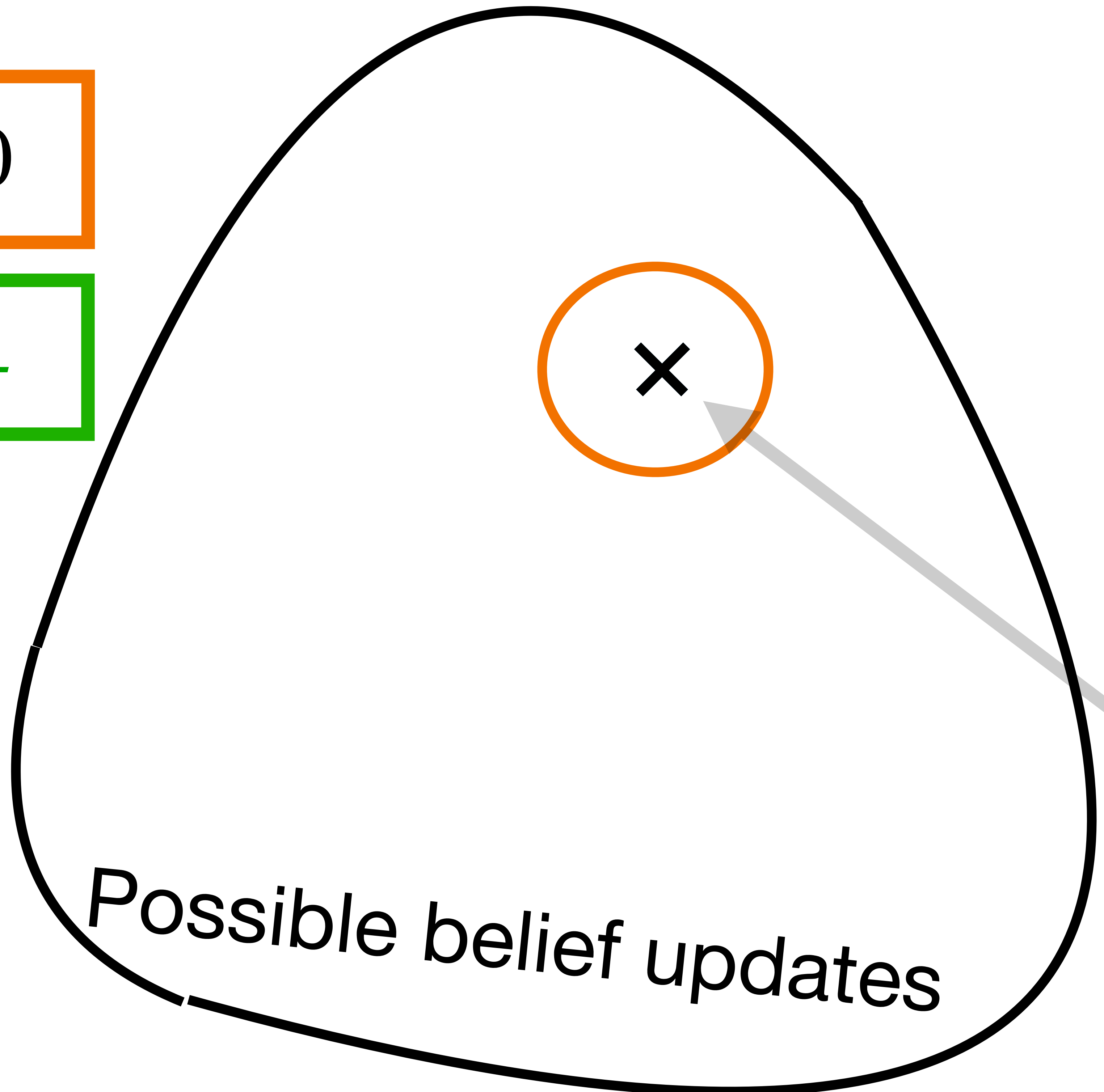
(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

Post-Bayesian ML

$$p(x_{1:n} \mid \theta) \longrightarrow p(x_{1:n} \mid \theta)^{\lambda}, \ \lambda > 0$$

$$p(x_{1:n} \mid \theta) \longrightarrow \exp\{-\mathsf{L}(x_{1:n}, \theta)\}, \text{ loss } \mathsf{L}$$

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible



Power/Fractional/
Cold Posterior

~~(A1)~~, (A2), (A3)

$$\pi_n^{(\lambda)}(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta) d\theta}$$

Bayes' Posterior

(A1), (A2), (A3)

$$\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}$$

Post-Bayesian ML

[Generally credited to Bissiri, Holmes & Walker (2016)]

Langford & Shawe-Taylor (2002); NeurIPS
Seeger (2002); ICML
Bissiri et al. (2016); JRSS-B
...
Knoblauch & Damoulas. (2018); ICML
Knoblauch, Jewson, & Damoulas et al. (2018); NeurIPS
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML spotlight

Gibbs/Generalised/
Pseudo Posterior ~~(A1)~~, (A2), ~~(A3)~~

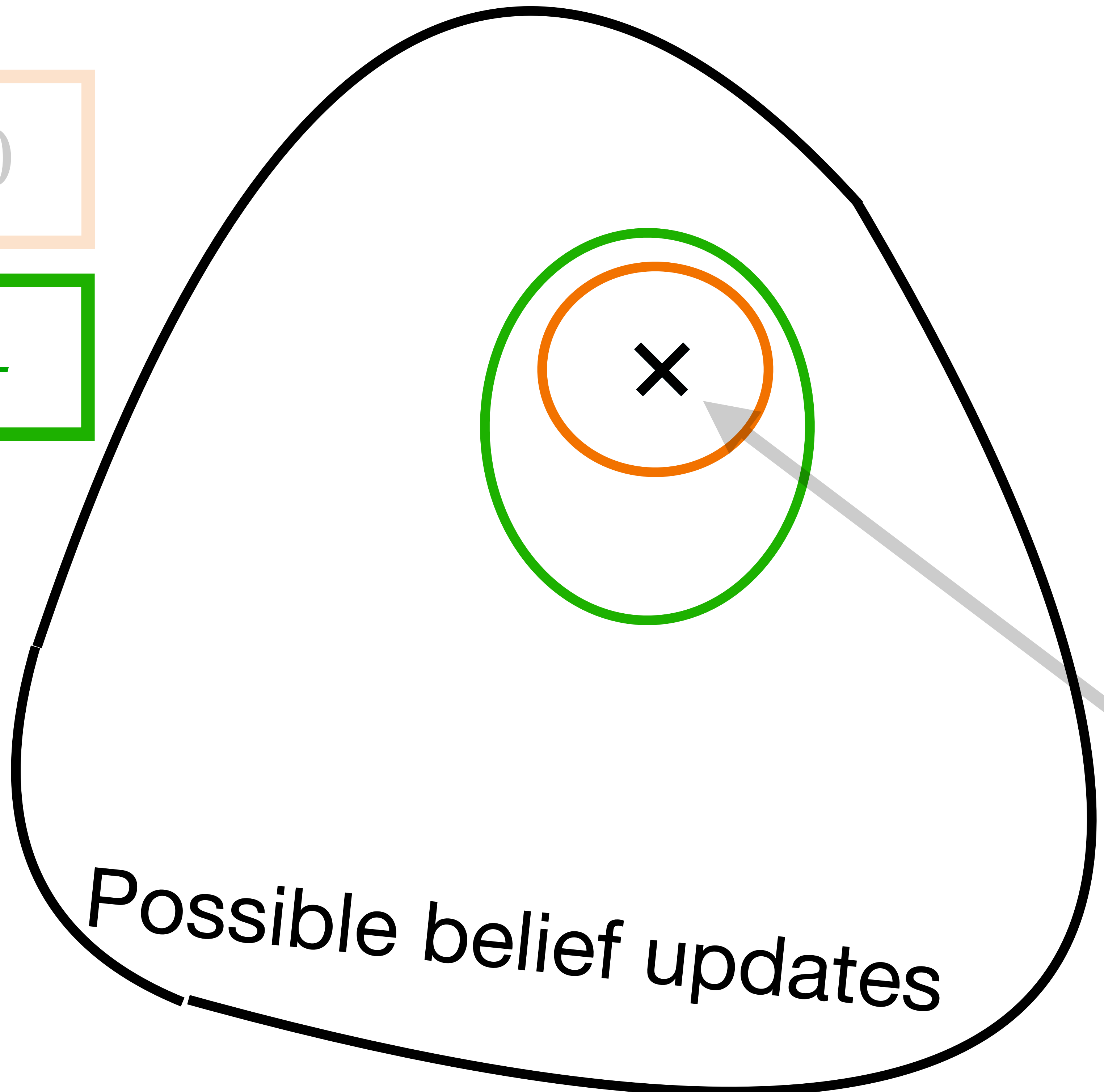
$$\pi_n^{\mathrm{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathrm{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathrm{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}$$

$$p(x_{1:n} \mid \theta) \longrightarrow p(x_{1:n} \mid \theta)^{\lambda}, \; \lambda > 0$$

$$p(x_{1:n} \mid \theta) \longrightarrow \exp\{-\mathrm{L}(x_{1:n}, \theta)\}, \; \text{loss } \mathrm{L}$$

Power/Fractional/
Cold Posterior ~~(A1)~~, (A2), (A3)

$$\pi_n^{(\lambda)}(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta) d\theta}$$



Bayes' Posterior (A1), (A2), (A3)

$$\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}$$

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

Post-Bayesian ML

Optimisation-centric posteriors /
Generalised Variational Inference $\cancel{\text{(A1)}}, \cancel{\text{(A2)}}, \cancel{\text{(A3)}}$

$$q_n^*(\theta) = \arg\min_{q \in \mathcal{Q}} \{\underbrace{\mathcal{L}_{\mathrm{L},\, D}(q)}\}; \quad \mathcal{Q} \subseteq \mathcal{P}(\Theta)$$

$$= \int \mathrm{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta \, + \, \mathrm{D}(q, \pi)$$

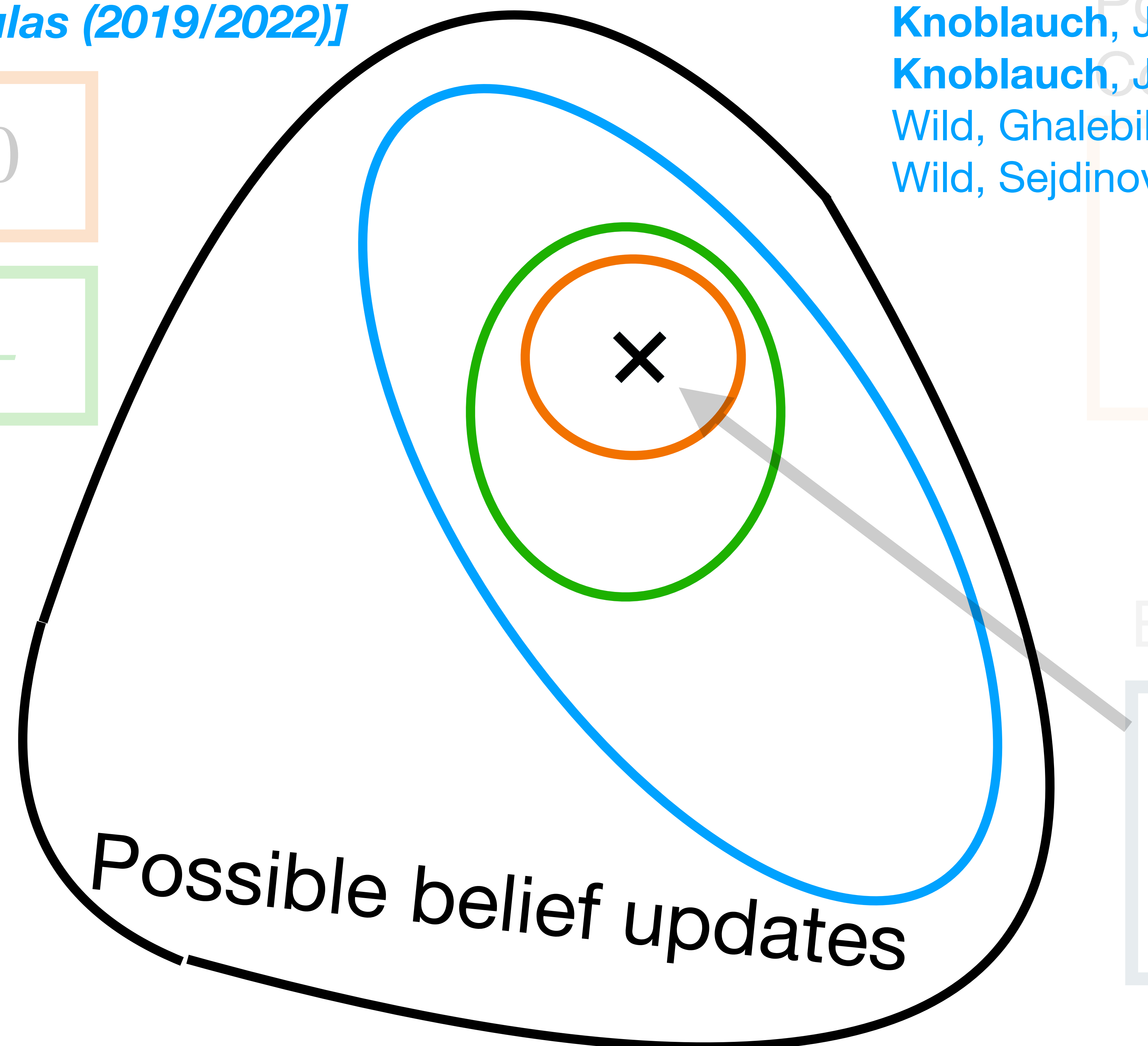
[Generally credited to Knoblauch, Jewson, & Damoulas (2019/2022)]

$$p(x_{1:n} \mid \theta) \longrightarrow p(x_{1:n} \mid \theta)^{\lambda}, \ \lambda > 0$$

$$p(x_{1:n} \mid \theta) \longrightarrow \exp\{-\mathrm{L}(x_{1:n}, \theta)\}, \ \text{loss L}$$

$$\begin{array}{ccc} \mathrm{KL} & \longrightarrow & \mathrm{D} \\ \mathcal{P}(\Theta) & \longrightarrow & \mathcal{Q} \end{array}$$

(A1)	model well-specified
(A2)	prior well-specified
(A3)	computationally feasible



Gibbs/Generalised/
Pseudo Posterior $\cancel{\text{(A1)}}$, (A2), $\cancel{\text{(A3)}}$

$$\pi_n^{\mathrm{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathrm{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathrm{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}$$

Xuan, Wu, Liu, & Lu (2024); UAI
Chi, Zhang, Yang, Ouyang, & Pei (2024); AAAI
…
Knoblauch, Jewson, & Damoulas (2019); NeurIPS
Knoblauch, Jewson, & Damoulas (2022); JMLR
Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2023); NeurIPS (oral)
Wild, Sejdinovic, & **Knoblauch** (2024); forthcoming

Power/Fractional/
Cold Posterior $\cancel{\text{(A1)}}$, (A2), (A3)

$$\pi_n^{(\lambda)}(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta) d\theta}$$

Bayes' Posterior (A1), (A2), (A3)

$$\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}$$

Post-Bayesian ML

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}} = \arg\min_{q \in \mathscr{P}(\Theta)} \left\{ \int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta + \mathrm{KL}\big(q, \pi\big) \right\}$$

Husain & **Knoblauch** (2022); ALT
Knoblauch, Jewson, & Damoulas (2022); JMLR
Wild, Sejdinovic, & **Knoblauch** (2024); forthcoming
Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2024); NeurIPS (oral)

Post-Bayesian ML

~~(A1)~~ model well-specified
(A2) prior well-specified
(A3) computationally feasible

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}} = \arg \min_{q \in \mathscr{P}(\Theta)} \left\{ \int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta + \mathrm{KL}(q, \pi) \right\}$$

~~(A1)~~ by using robust loss L

$$\boxed{q_n^*(\theta) = \arg\min_{q \in \mathscr{Q}} \left\{ \underbrace{\int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta + \mathrm{D}(q, \pi)}_{=: \mathscr{L}_{\mathsf{L},\, \mathrm{D}}(q)} \right\}}$$

Optimisation-centric posteriors / Generalised Variational Inference (GVI)

Husain & **Knoblauch** (2022); ALT
Knoblauch, Jewson, & Damoulas (2022); JMLR
Wild, Sejdinovic, & **Knoblauch** (2024); forthcoming
Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2024); NeurIPS (oral)

Post-Bayesian ML

~~(A1)~~ model well-specified
~~(A2)~~ prior well-specified
(A3) computationally feasible

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n},\theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n},\theta)\} \cdot \pi(\theta) d\theta}} = \arg\min_{q \in \mathscr{P}(\Theta)} \left\{ \int \mathsf{L}(x_{1:n},\theta)\, q(\theta)\, d\theta + \mathrm{KL}\big(q,\pi\big) \right\}$$

~~(A1)~~ by using robust loss L

~~(A2)~~ by using robust regulariser D

$$\boxed{q_n^*(\theta) = \arg\min_{q \in \mathscr{Q}} \underbrace{\left\{ \int \mathsf{L}(x_{1:n},\theta)\, q(\theta)\, d\theta + \mathrm{D}\big(q,\pi\big) \right\}}_{=:\, \mathscr{L}_{\mathsf{L},\,\mathrm{D}}(q)}}$$

Optimisation-centric posteriors / Generalised Variational Inference (GVI)

Husain & **Knoblauch** (2022); ALT
Knoblauch, Jewson, & Damoulas (2022); JMLR
Wild, Sejdinovic, & **Knoblauch** (2024); forthcoming
Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2024); NeurIPS (oral)

Post-Bayesian ML

~~(A1)~~ model well-specified
~~(A2)~~ prior well-specified
~~(A3)~~ computationally feasible

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n},\theta)\}\cdot\pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n},\theta)\}\cdot\pi(\theta)d\theta}} = \arg\min_{q\in\mathscr{P}(\Theta)} \left\{ \int \mathsf{L}(x_{1:n},\theta)\, q(\theta)\, d\theta + \mathrm{KL}(q,\pi) \right\}$$

~~(A3)~~ by optimising over a set $\mathcal{Q} \subseteq \mathscr{P}(\Theta)$

~~(A1)~~ by using robust loss L

~~(A2)~~ by using robust regulariser D

$$\boxed{q_n^*(\theta) = \arg\min_{q\in\mathcal{Q}} \left\{ \underbrace{\int \mathsf{L}(x_{1:n},\theta)\, q(\theta)\, d\theta + \mathrm{D}(q,\pi)}_{=:\ \mathscr{L}_{\mathsf{L},\mathrm{D}}(q)} \right\}}$$

Optimisation-centric posteriors / Generalised Variational Inference (GVI)

Husain & **Knoblauch** (2022); ALT
Knoblauch, Jewson, & Damoulas (2022); JMLR
Wild, Sejdinovic, & **Knoblauch** (2024); forthcoming
Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2024); NeurIPS (oral)

Post-Bayesian ML

Optimisation-centric posteriors / Generalised Variational Inference ~~(A1)~~, ~~(A2)~~, ~~(A3)~~

$$q_n^*(\theta) = \arg\min_{q \in \mathcal{Q}} \{\underbrace{\mathcal{L}_{\mathsf{L},\,D}(q)}\}; \quad \mathcal{Q} \subseteq \mathcal{P}(\Theta)$$

$$= \int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta \; + \; \mathrm{D}\big(q, \pi\big)$$

Gibbs/Generalised/ Pseudo Posterior ~~(A1)~~, (A2), ~~(A3)~~

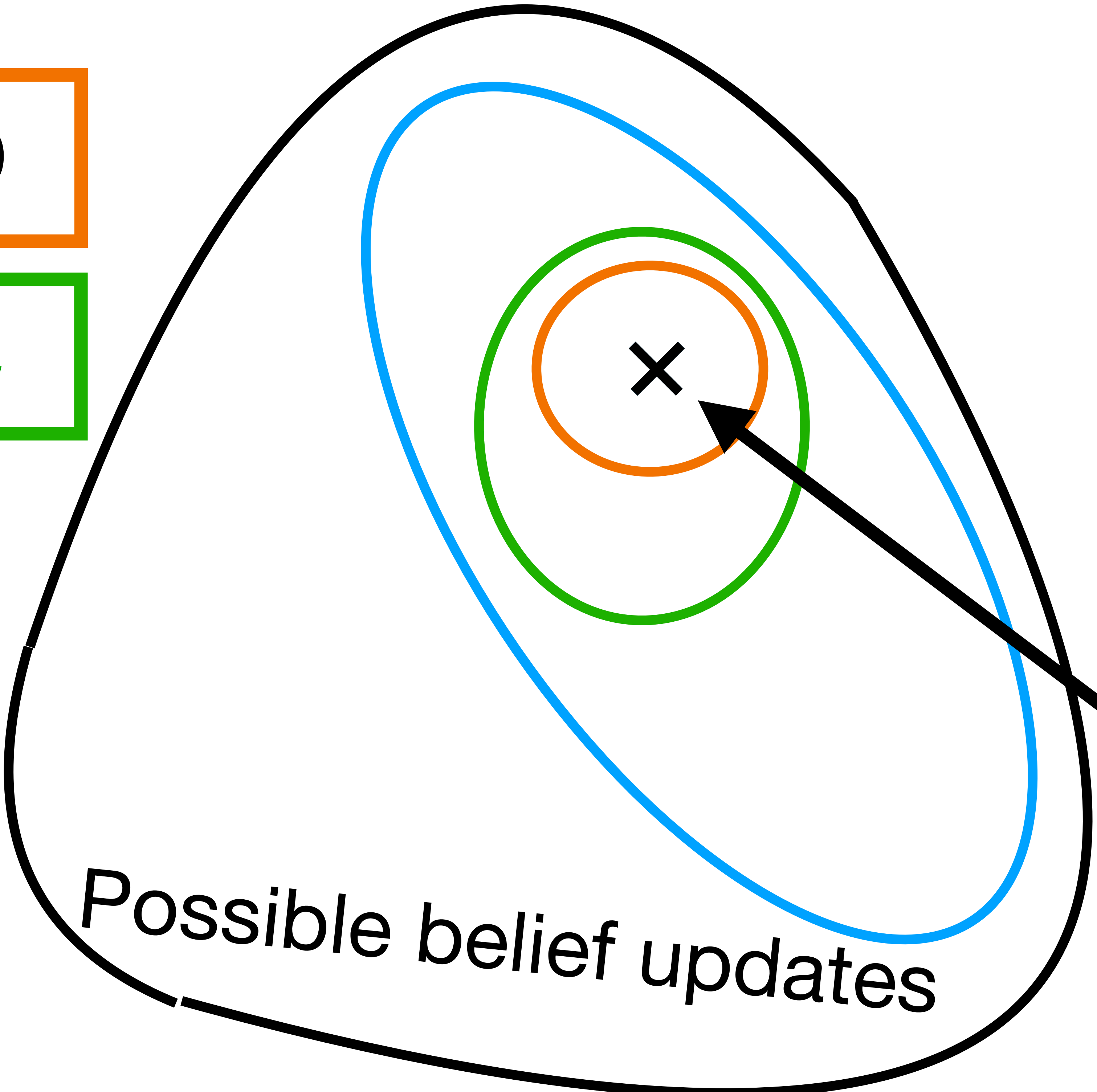
$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}$$

$$p(x_{1:n} \mid \theta) \longrightarrow p(x_{1:n} \mid \theta)^{\lambda}, \; \lambda > 0$$

$$p(x_{1:n} \mid \theta) \longrightarrow \exp\{-\mathsf{L}(x_{1:n}, \theta)\}, \; \text{loss } \mathsf{L}$$

$$\begin{array}{ccc} \mathrm{KL} & \longrightarrow & \mathrm{D} \\ \mathcal{P}(\Theta) & \longrightarrow & \mathcal{Q} \end{array}$$

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible



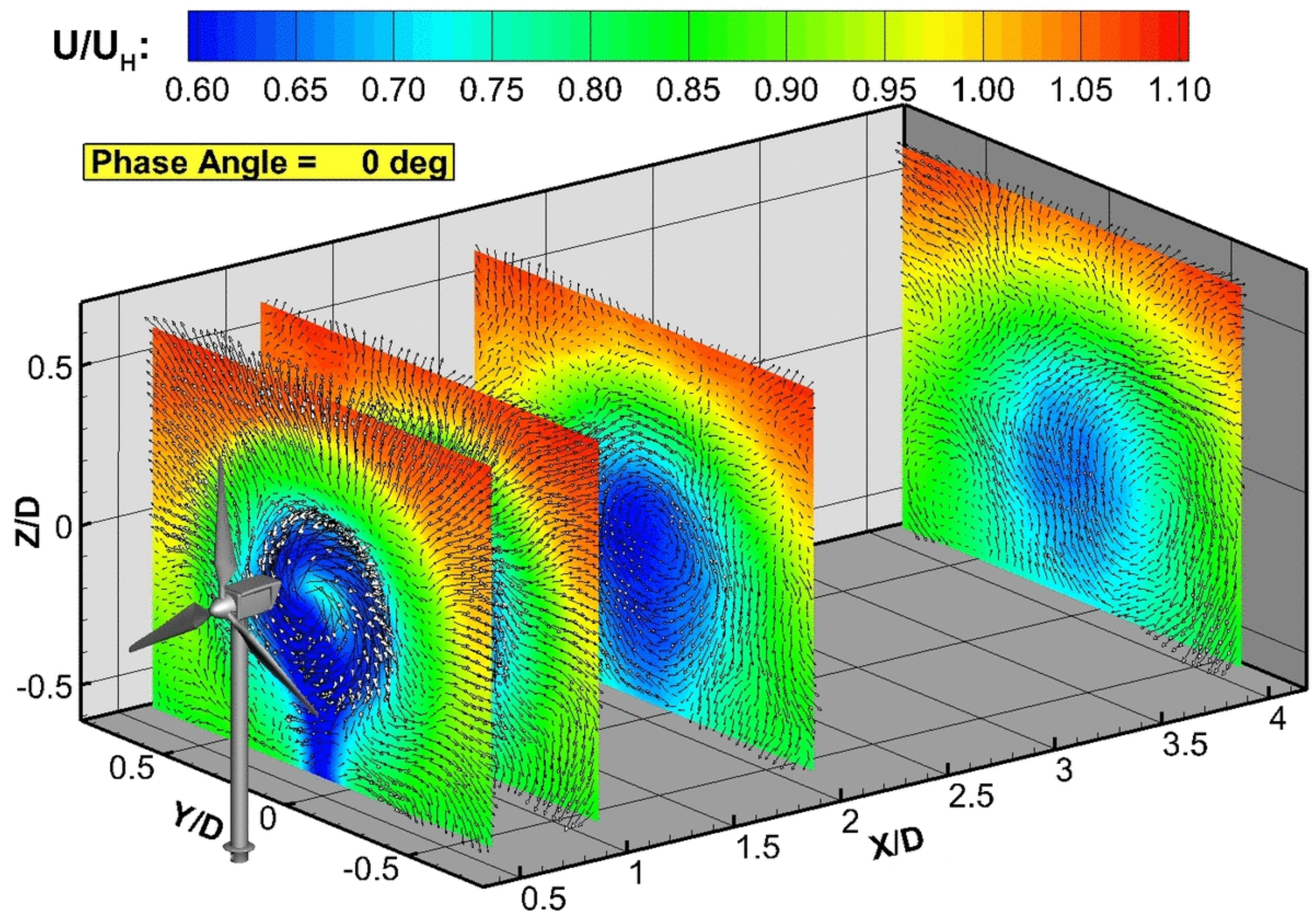
Power/Fractional/ Cold Posterior ~~(A1)~~, (A2), (A3)

$$\pi_n^{(\lambda)}(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta) d\theta}$$

Bayes' Posterior (A1), (A2), (A3)

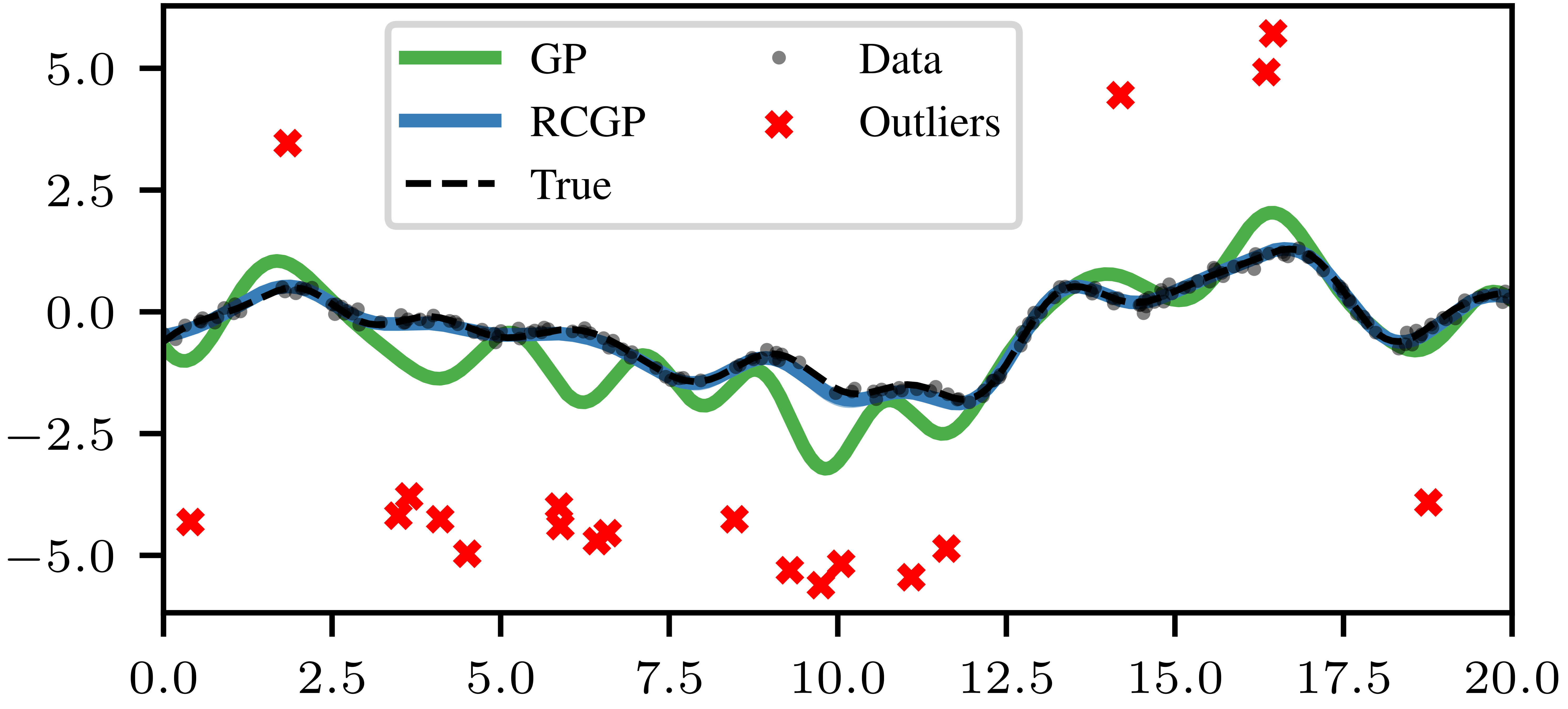
$$\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}$$

Post-Bayesian ML: Recent Success Stories

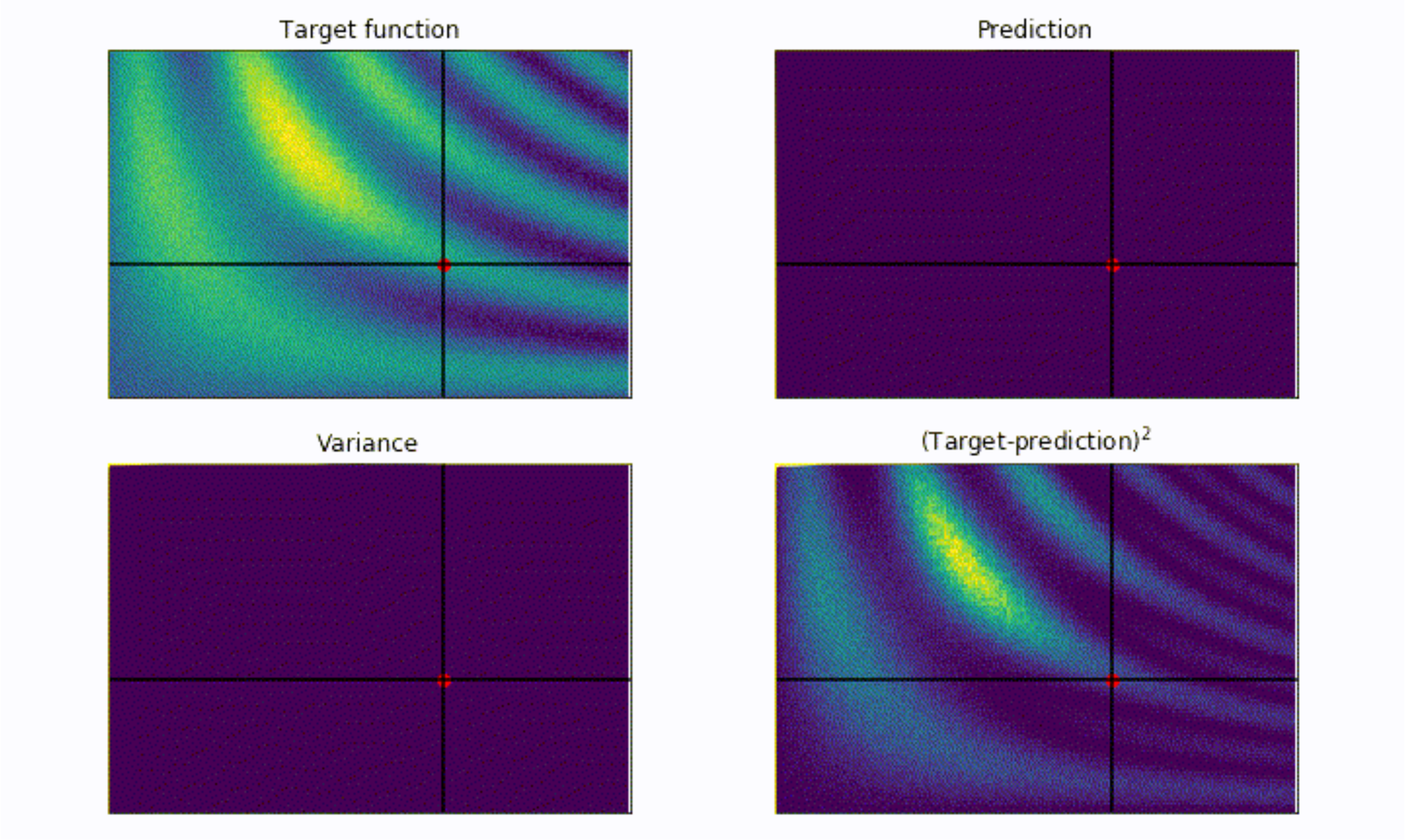


Based on work of Dellaportas, **Knoblauch**, Damoulas, & Briol (2022); AISTATS (best paper award)

Post-Bayesian ML: Recent Success Stories



MATTERHORN STUDIO



Based on work of Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)

Post-Bayesian ML Research

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

1 Foundations

$\boxed{\pi_n^{(\lambda)}(\theta \mid x_{1:n})}$

$\boxed{\pi_n^{\text{L}}(\theta \mid x_{1:n})}$

Post-Bayesian ML Research

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

1 Foundations

$\boxed{\pi_n^{(\lambda)}(\theta \mid x_{1:n})}$

$\boxed{\pi_n^{\text{L}}(\theta \mid x_{1:n})}$

model misspecification

~~(A1)~~ (A2) (A3)

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, **Knoblauch***, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

Post-Bayesian ML Research

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

① **Foundations**

$\boxed{\pi_n^{(\lambda)}(\theta \mid x_{1:n})}$

$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})}$

model misspecification

~~(A1)~~ (A2) (A3)

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, **Knoblauch***, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

② **State of the Art**

$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})}$

Post-Bayesian ML Research

(A1)	model well-specified
(A2)	prior well-specified
(A3)	computationally feasible

1 Foundations

$\boxed{\pi_n^{(\lambda)}(\theta \mid x_{1:n})}$

$\boxed{\pi_n^{\text{L}}(\theta \mid x_{1:n})}$

model misspecification

~~(A1)~~ (A2) (A3)

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, **Knoblauch***, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

2 State of the Art

$\boxed{\pi_n^{\text{L}}(\theta \mid x_{1:n})}$

model misspecification + computation

~~(A1)~~ (A2) ~~(A3)~~

Schmon, Cannon, & **Knoblauch** (2020); AABI
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Dellaporta, **Knoblauch**, Damoulas, & Briol (2022); AISTATS (best paper award)
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)
Duran-Martin, Altamirano, Shestopaloff, Sanchez-Betancourt, **Knoblauch**, Briol, & Murphy (2024); ICML

Post-Bayesian ML Research

(A1)	model well-specified
(A2)	prior well-specified
(A3)	computationally feasible

1 Foundations

$\boxed{\pi_n^{(\lambda)}(\theta \mid x_{1:n})}$

$\boxed{\pi_n^{\mathrm{L}}(\theta \mid x_{1:n})}$

model misspecification

~~(A1)~~ (A2) (A3)

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, **Knoblauch***, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

2 State of the Art

$\boxed{\pi_n^{\mathrm{L}}(\theta \mid x_{1:n})}$

model misspecification + computation

~~(A1)~~ (A2) ~~(A3)~~

Schmon, Cannon, & **Knoblauch** (2020); AABI
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Dellaporta, **Knoblauch**, Damoulas, & Briol (2022); AISTATS (best paper award)
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)
Duran-Martin, Altamirano, Shestopaloff, Sanchez-Betancourt, **Knoblauch**, Briol, & Murphy (2024); ICML

3 The Future

$\boxed{q_n^*(\theta)}$

Post-Bayesian ML Research

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

1 Foundations

$\pi_n^{(\lambda)}(\theta \mid x_{1:n})$

$\pi_n^{L}(\theta \mid x_{1:n})$

model misspecification

~~(A1)~~ (A2) (A3)

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, **Knoblauch***, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

2 State of the Art

$\pi_n^{L}(\theta \mid x_{1:n})$

model misspecification + computation

~~(A1)~~ (A2) ~~(A3)~~

Schmon, Cannon, & **Knoblauch** (2020); AABI
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Dellaporta, **Knoblauch**, Damoulas, & Briol (2022); AISTATS (best paper award)
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)
Duran-Martin, Altamirano, Shestopaloff, Sanchez-Betancourt, **Knoblauch**, Briol, & Murphy (2024); ICML

3 The Future

$q_n^*(\theta)$

model misspecification + prior misspecification + computation

~~(A1)~~ ~~(A2)~~ ~~(A3)~~

Husain & **Knoblauch** (2022); ALT
Knoblauch, Jewson, & Damoulas (2022); JMLR
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Wild, Sejdinovic, & **Knoblauch** (2024); forthcoming
Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2024); NeurIPS (oral)


Matias Altamirano (UCL)

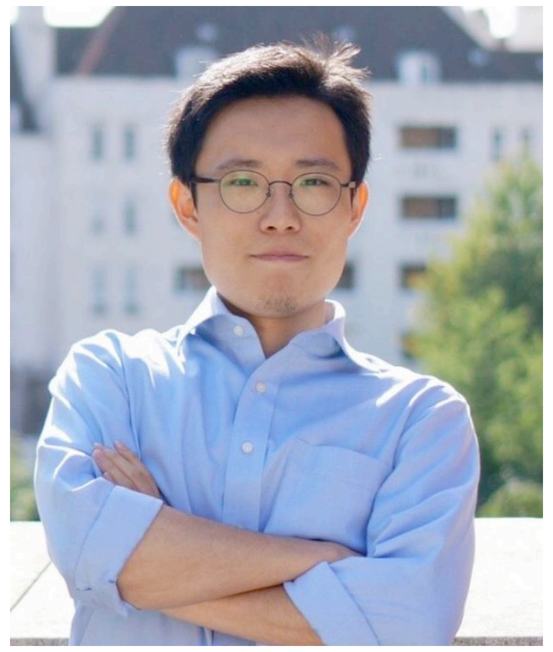

Yann McLatchie (UCL)


Veit Wild (Oxford)


Dino Sejdinovic (Oxford/Adelaide)


Kevin Murphy (DeepMind)


Chris Drovandi (QUT)


Takuo Matsubara (Edinburgh)

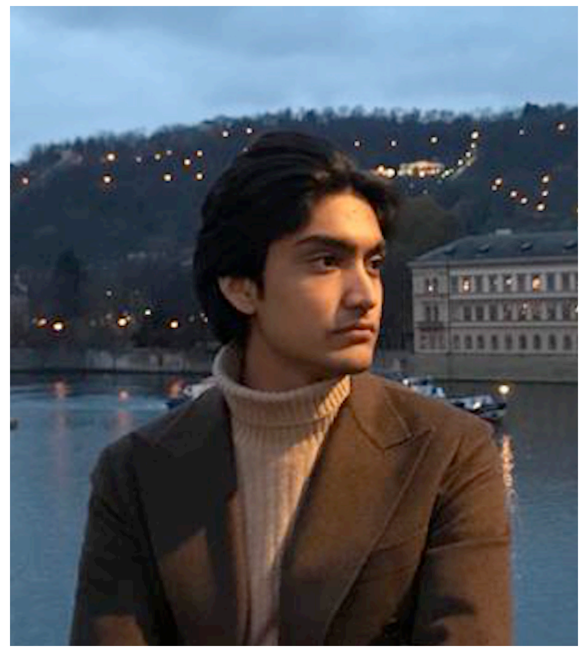

Gerardo Duran-Martin (QMU/Oxford)


Sahra Ghalebikesabi (Oxford/DeepMind)


Edwin Fong (Hong Kong)


David Frazier (Monash)


Miheer Dewaskar (Duke)


Hisham Husain (Amazon)


Francois-Xavier Briol (UCL)


Chris Oates (Newcastle)


Jack Jewson (UPF Barcelona/Monash)


Theo Damoulas (Warwick)


Chris Tosh (Memorial Sloan Kettering Institute)


Harita Dellaporta (Warwick)


David Dunson (Duke)

~~(A1)~~, (A2), (A3) **Tackled Assumptions**

Foundations of Post-Bayesian ML Research

1 Foundations

$\pi_n^{(\lambda)}(\theta \mid x_{1:n})$

$\pi_n^{\text{L}}(\theta \mid x_{1:n})$

model misspecification

~~(A1)~~ (A2) (A3)

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, Knoblauch*, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & Knoblauch (2024); forthcoming

2 State of the Art

$\pi_n^{\text{L}}(\theta \mid x_{1:n})$

model misspecification + computation

~~(A1)~~ (A2) ~~(A3)~~

Schmon, Cannon, & Knoblauch (2020); AABI
Matsubara, Knoblauch, Briol, & Oates (2022); JRSS-B
Dellaporta, Knoblauch, Damoulas, & Briol (2022); AISTATS (best paper award)
Altamirano, Briol, & Knoblauch (2023); ICML
Altamirano, Briol, & Knoblauch (2024); ICML (spotlight)
Duran-Martin, Altamirano, Shestopaloff, Sanchez-Betancourt, Knoblauch, Briol, & Murphy (2024); ICML

3 The Future

$q^*(\theta)$

model misspecification + prior misspecification + computation

~~(A1)~~ ~~(A2)~~ ~~(A3)~~

Husain & Knoblauch (2022); ALT
Knoblauch, Jewson, & Damoulas (2022); JMLR
Matsubara, Knoblauch, Briol, & Oates (2023); JASA
Wild, Sejdinovic, & Knoblauch (2024); forthcoming
Wild, Ghalebikesabi, Sejdinovic, & Knoblauch (2024); NeurIPS (oral)

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

~~(A1)~~, (A2), (A3) **Tackled Assumptions**

Part of the Talk 1

Foundations of Post-Bayesian ML Research

1 Foundations

$\pi_n^{(\lambda)}(\theta \mid x_{1:n})$

$\pi_n^{\text{L}}(\theta \mid x_{1:n})$

model misspecification

~~(A1)~~ (A2) (A3)

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, **Knoblauch***, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

2 State of the Art

$\pi_n^{\text{L}}(\theta \mid x_{1:n})$

model misspecification + computation

~~(A1)~~ (A2) ~~(A3)~~

Schmon, Cannon, & **Knoblauch** (2020); AABI
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Dellaporta, **Knoblauch**, Damoulas, & Briol (2022); AISTATS (best paper award)
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)
Duran-Martin, Altamirano, Shestopaloff, Sanchez-Betancourt, **Knoblauch**, Briol, & Murphy (2024); ICML

3 The Future

$q^*(\theta)$

model misspecification + prior misspecification + computation

~~(A1)~~ ~~(A2)~~ ~~(A3)~~

Husain & **Knoblauch** (2022); ALT
Knoblauch, Jewson, & Damoulas (2022); JMLR
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Wild, Sejdinovic, & **Knoblauch** (2024); forthcoming
Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2024); NeurIPS (oral)

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

~~(A1)~~, (A2), (A3)

1

Foundations of Post-Bayesian ML Research

1 Foundations

$\boxed{\pi_n^{(\lambda)}(\theta \mid x_{1:n})}$

$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})}$

model misspecification

~~(A1)~~ (A2) (A3)

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, **Knoblauch***, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

Q1: Can tuning λ improve robustness?
Q2: What L leads to robust posteriors π_n^{L}?
Q3: How should we design/choose L?

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

~~(A1)~~, (A2), (A3)

1

Foundations of Post-Bayesian ML Research

1 Foundations

$\boxed{\pi_n^{(\lambda)}(\theta \mid x_{1:n})}$

$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})}$

model misspecification

~~(A1)~~ (A2) (A3)

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, **Knoblauch***, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

Q1: Can tuning λ improve robustness?
Q2: What L leads to robust posteriors π_n^{L}?
Q3: How should we design/choose L?

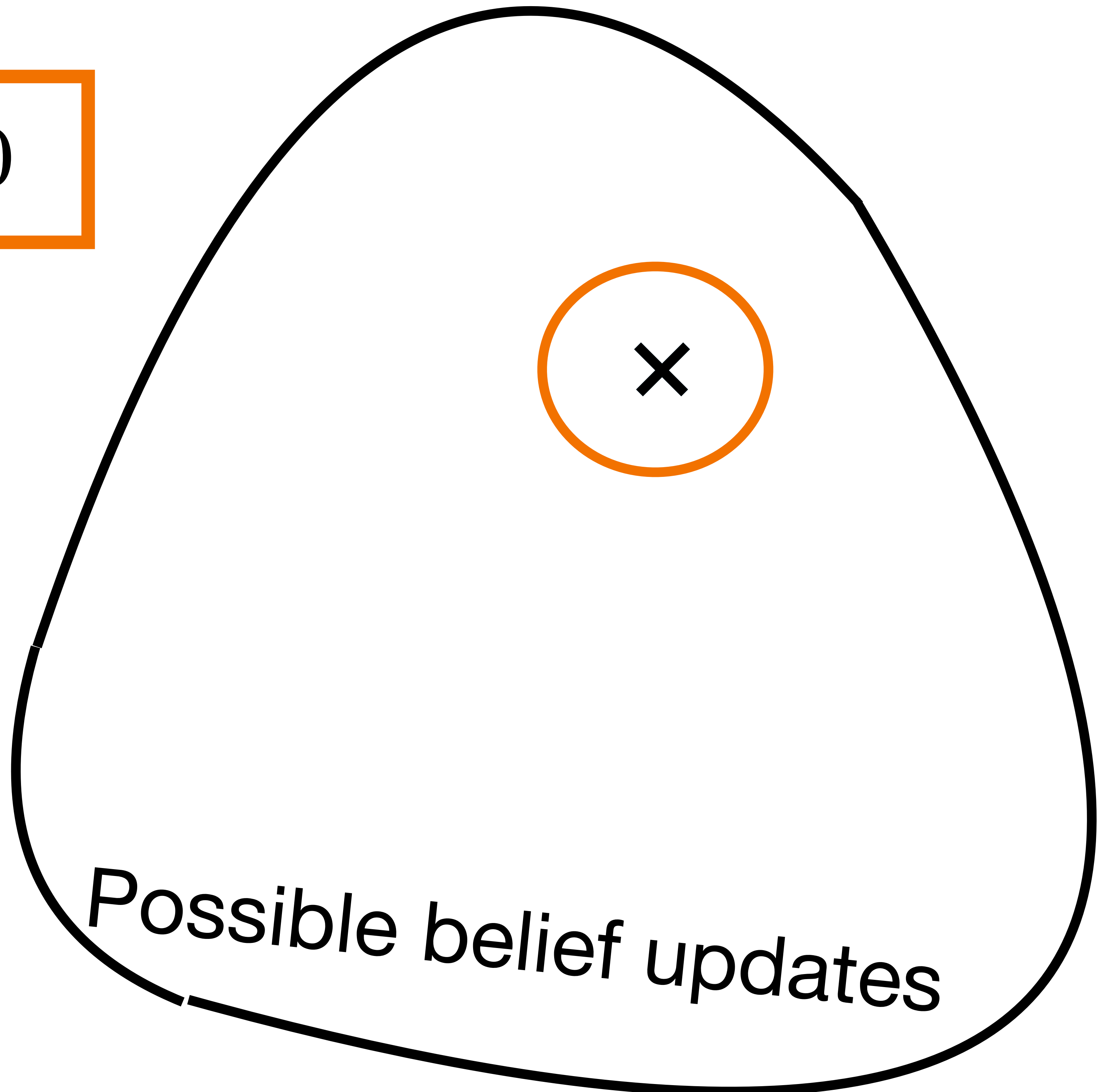
(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

~~(A1)~~, (A2), (A3)

1

Post-Bayesian ML

$$p(x_{1:n} \mid \theta) \longrightarrow p(x_{1:n} \mid \theta)^{\lambda}, \ \lambda > 0$$



Power/Fractional/
Cold Posterior

~~(A1)~~, (A2), (A3)

$$\pi_n^{(\lambda)}(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta) d\theta}$$

Bayes' Posterior

(A1), (A2), (A3)

$$\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}$$

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

~~(A1)~~, (A2), (A3)

1

Q1: Can tuning λ improve Robustness?

(classical statistics)

Grünwald (2012); ALT
Holmes & Walker (2017); Biometrika
Miller & Dunson (2018); JRSS-B
Bhattacharya, Pati, & Yang (2019); Ann. Statist.
…

→ **Frequent claim: λ can deliver robust predictions** ←

(core ML)

Wenzel et al. (2020); ICML
Adlam et al. (2020); preprint
Noci et al. (2021); NeurIPS
Aitchison (2021); ICLR
…

~~(A1)~~, (A2), (A3)

1

Q1: Can tuning λ improve Robustness?

(classical statistics)

Grünwald (2012); ALT
Holmes & Walker (2017); Biometrika
Miller & Dunson (2018); JRSS-B
Bhattacharya, Pati, & Yang (2019); Ann. Statist.
…

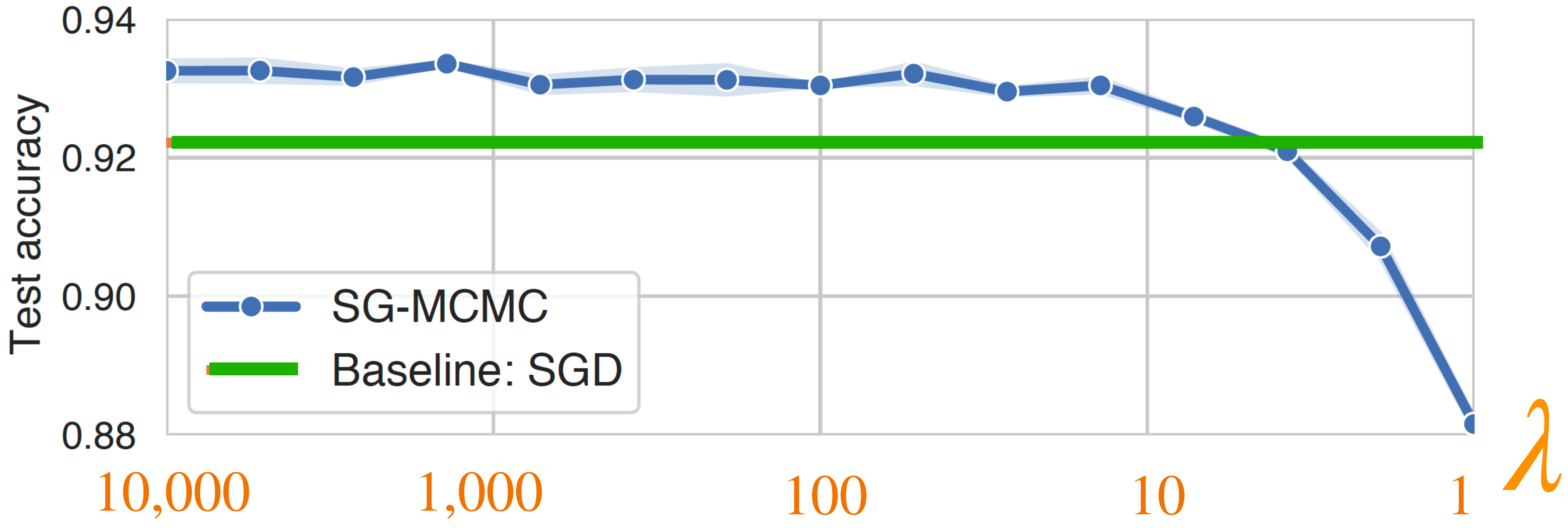
Frequent claim: λ can deliver robust predictions

(core ML)

Wenzel et al. (2020); ICML
Adlam et al. (2020); preprint
Noci et al. (2021); NeurIPS
Aitchison (2021); ICLR
…

How Good is the Bayes Posterior in Deep Neural Networks Really?

Florian Wenzel [*1] **Kevin Roth** [*+2] **Bastiaan S. Veeling** [*+3 1] **Jakub Świątkowski** [4+] **Linh Tran** [5+]
Stephan Mandt [6+] **Jasper Snoek** [1] **Tim Salimans** [1] **Rodolphe Jenatton** [1] **Sebastian Nowozin** [7+]



~~(A1)~~, (A2), (A3)

1

Q1: Can tuning λ improve Robustness?

(classical statistics)

Grünwald (2012); ALT
Holmes & Walker (2017); Biometrika
Miller & Dunson (2018); JRSS-B
Bhattacharya, Pati, & Yang (2019); Ann. Statist.
…

Frequent claim: λ can deliver robust predictions

(core ML)

Wenzel et al. (2020); ICML
Adlam et al. (2020); preprint
Noci et al. (2021); NeurIPS
Aitchison (2021); ICLR
…

Unclear: Why should this be true?

Only regulates trade-off data $\longleftrightarrow$ prior

$$\boxed{\pi_n^{(\lambda)}(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta) d\theta}}$$

~~(A1)~~, (A2), (A3)

①

Q1: Can tuning λ improve Robustness?

Question: What is the predictively optimal λ?

Posterior predictive $= p_n^{\lambda}(z) = \int p(z \mid \theta)\, \pi_n^{(\lambda)}(\theta \mid x_{1:n})\, d\theta$

Predictively optimal λ: $\lambda^* = \mathrm{argmin}_{\lambda>0}\, \mathrm{D}_{\mathrm{TV}}\left(q, p_n^{\lambda}\right)$

Data-generating density: $x_{1:n} \sim q(x_{1:n})$

McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

~~(A1)~~, (A2), (A3)

1

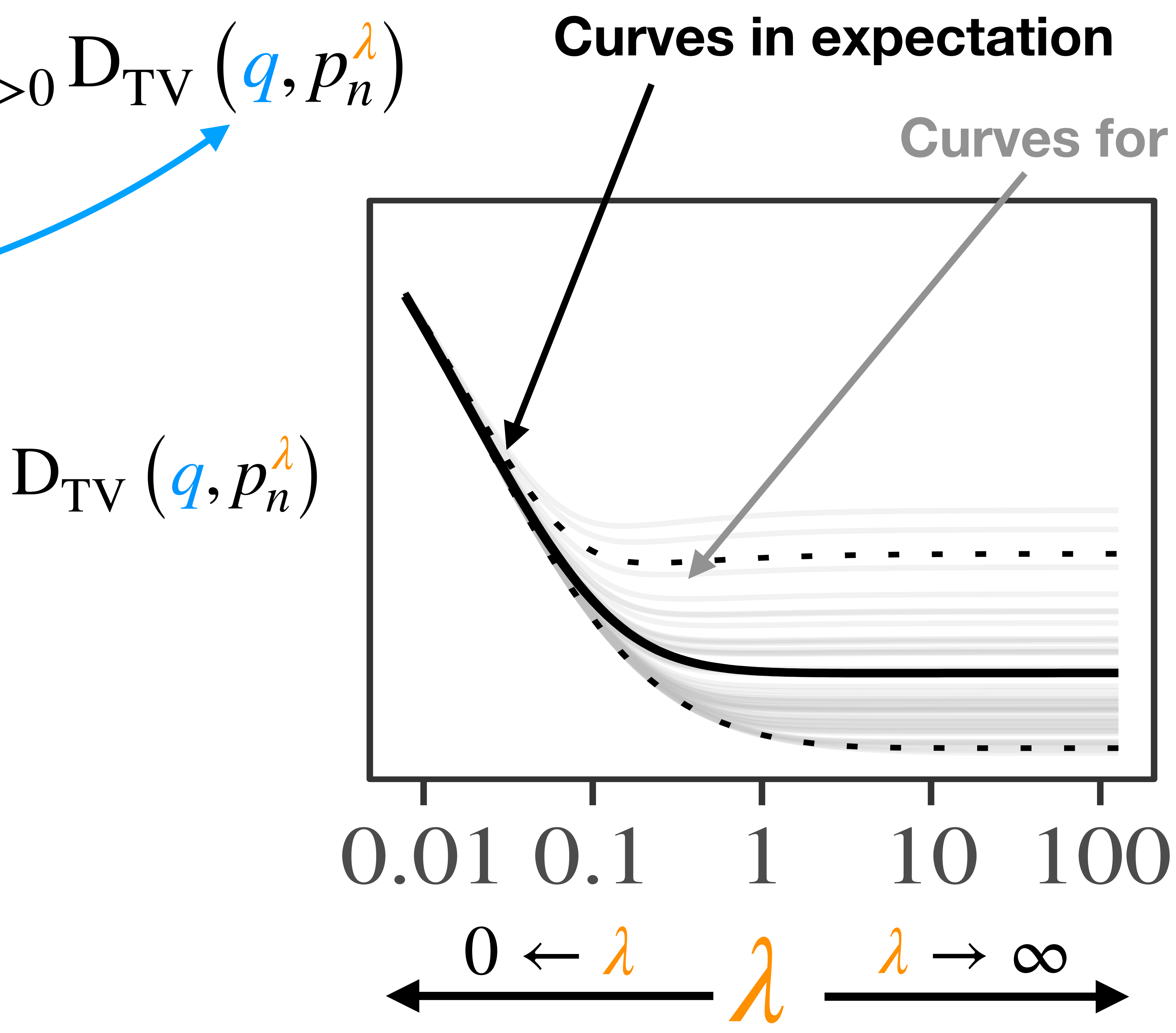
Q1: Can tuning λ improve Robustness?

Question: What is the predictively optimal λ?

Posterior predictive $= p_n^{\lambda}(z) = \int p(z \mid \theta)\, \pi_n^{(\lambda)}(\theta \mid x_{1:n})\, d\theta$

Predictively optimal λ: $\lambda^* = \text{argmin}_{\lambda>0} \, \mathrm{D_{TV}}\left(q, p_n^{\lambda}\right)$

Data-generating density: $x_{1:n} \sim q(x_{1:n})$



McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

~~(A1)~~, (A2), (A3)

①

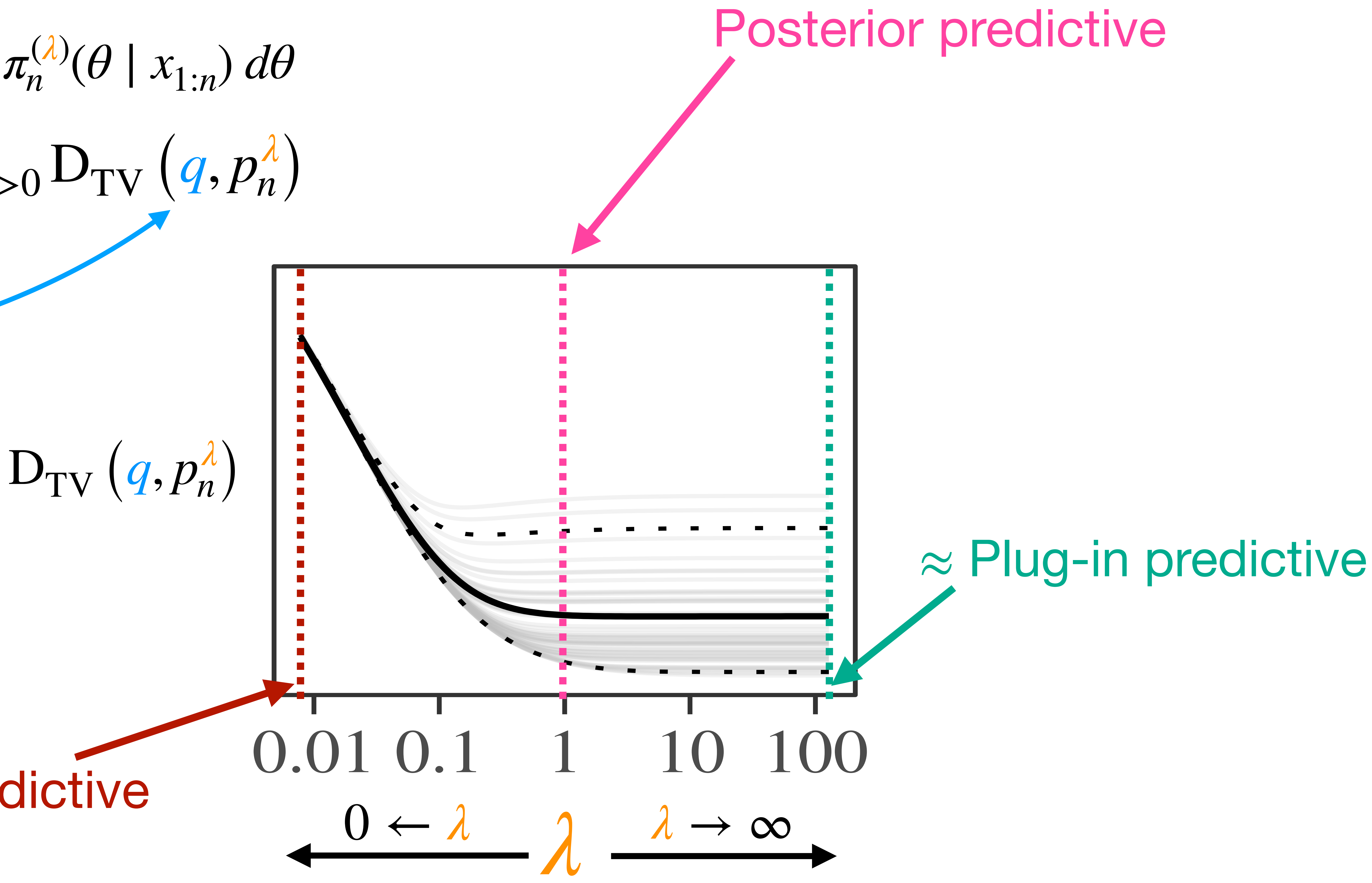
Q1: Can tuning λ improve Robustness?

Question: What is the predictively optimal λ?

Posterior predictive $= p_n^{\lambda}(z) = \int p(z \mid \theta)\, \pi_n^{(\lambda)}(\theta \mid x_{1:n})\, d\theta$

Predictively optimal λ: $\lambda^* = \text{argmin}_{\lambda>0}\, \mathrm{D}_{\mathrm{TV}}\left(q, p_n^{\lambda}\right)$

Data-generating density: $x_{1:n} \sim q(x_{1:n})$



McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

~~(A1)~~, (A2), (A3)

①

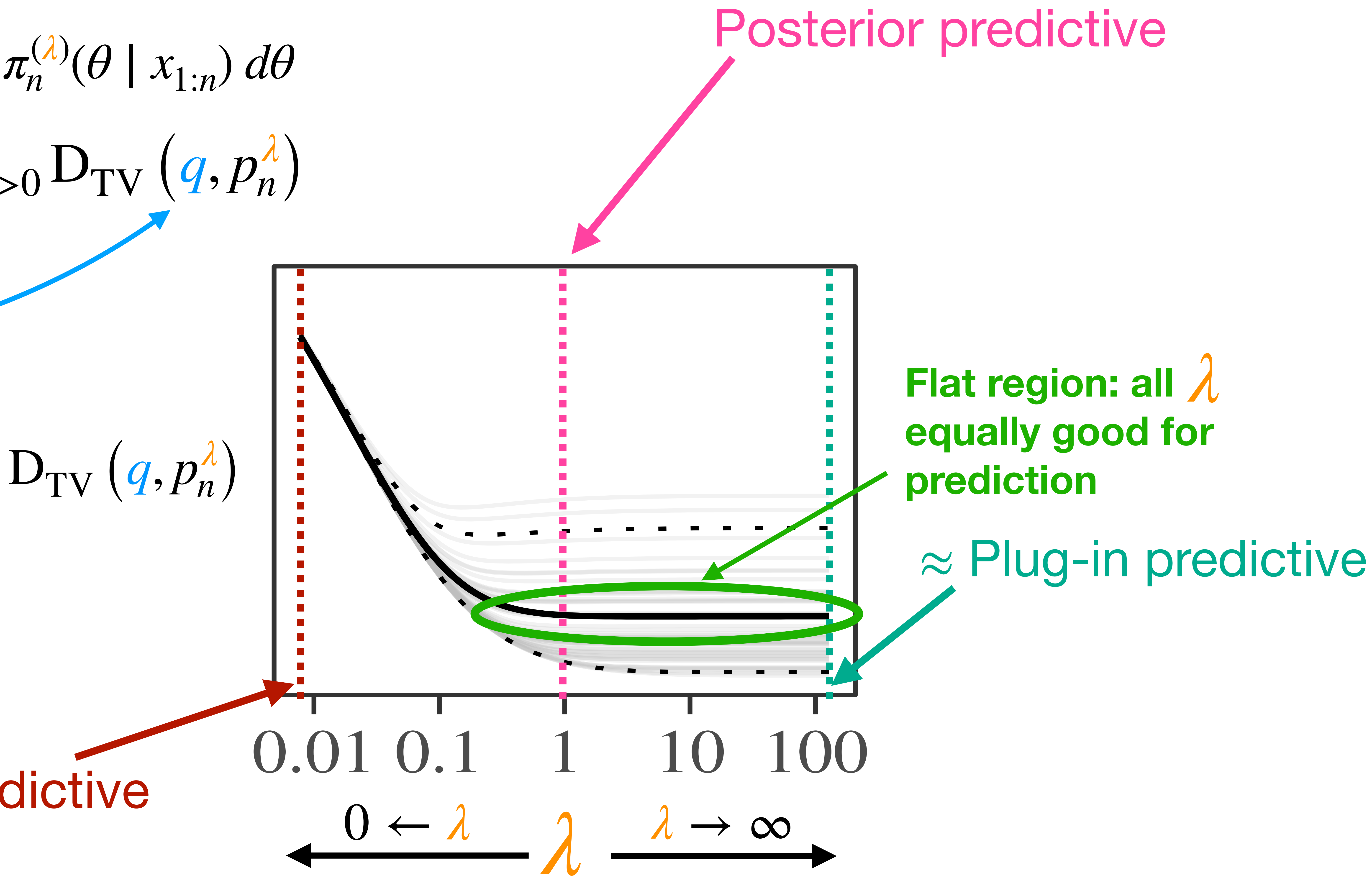
Q1: Can tuning λ improve Robustness?

Question: What is the predictively optimal λ?

Posterior predictive $= p_n^{\lambda}(z) = \int p(z \mid \theta)\, \pi_n^{(\lambda)}(\theta \mid x_{1:n})\, d\theta$

Predictively optimal λ: $\lambda^* = \mathrm{argmin}_{\lambda>0} \mathrm{D_{TV}}\left(q, p_n^{\lambda}\right)$

Data-generating density: $x_{1:n} \sim q(x_{1:n})$



McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

~~(A1)~~, (A2), (A3)

①

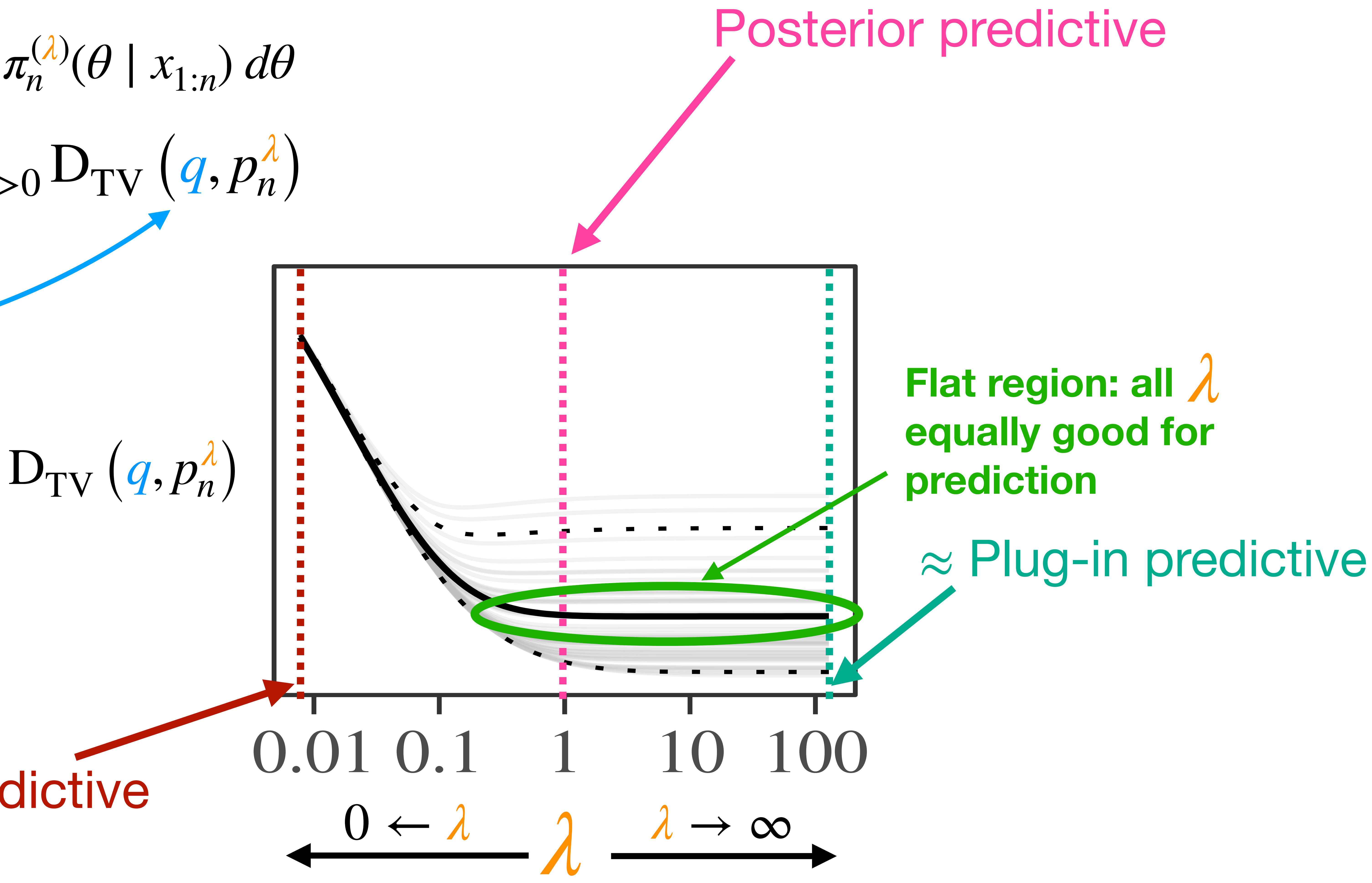
Q1: Can tuning λ improve Robustness?

Question: What is the predictively optimal λ?

Posterior predictive $= p_n^{\lambda}(z) = \int p(z \mid \theta)\, \pi_n^{(\lambda)}(\theta \mid x_{1:n})\, d\theta$

Predictively optimal λ: $\lambda^* = \mathrm{argmin}_{\lambda > 0} \mathrm{D_{TV}}\left(q, p_n^{\lambda}\right)$

Data-generating density: $x_{1:n} \sim q(x_{1:n})$



Theorem: these curves will always look that way.

McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

~~(A1)~~, (A2), (A3)

1

Q1: Can tuning λ improve Robustness?

Findings:

(1) Ill-defined problem: minimiser λ^* doesn't exist
(2) Flat region: infinitely λ yield (nearly) same predictive
(3) Predictively no advantage over MLE / point estimators

McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

~~(A1)~~, (A2), (A3)

1

Q1: Can tuning λ improve Robustness?

Findings:

(1) Ill-defined problem: minimiser λ^* doesn't exist
(2) Flat region: infinitely λ yield (nearly) same predictive
(3) Predictively no advantage over MLE / point estimators

Conclusion: Normally, λ barely has an effect on robustness of predictions.

McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

~~(A1)~~, (A2), (A3)

1

Q1: Can tuning λ improve Robustness?

Findings:	(1) Ill-defined problem: minimiser λ^* doesn't exist (2) Flat region: infinitely λ yield (nearly) same predictive (3) Predictively no advantage over MLE / point estimators

Conclusion: Normally, λ barely has an effect on robustness of predictions.

Reason: As n grows, you almost predict from the plug-in predictive: $p_n^{\infty} \approx p_n^{\lambda}$

McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

~~(A1)~~, (A2), (A3)

1

Q1: Can tuning λ improve Robustness?

Findings:

(1) Ill-defined problem: minimiser λ^* doesn't exist
(2) Flat region: infinitely λ yield (nearly) same predictive
(3) Predictively no advantage over MLE / point estimators

Conclusion: Normally, λ barely has an effect on robustness of predictions.

Reason: As n grows, you almost predict from the plug-in predictive: $p_n^{\infty} \approx p_n^{\lambda}$

Cold Posterior Effect: not only deep learning; more general phenomenon

McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

~~(A1)~~, (A2), (A3)

1

Q1: Can tuning λ improve Robustness?

Findings:
(1) Ill-defined problem: minimiser λ^* doesn't exist
(2) Flat region: infinitely λ yield (nearly) same predictive
(3) Predictively no advantage over MLE / point estimators

Conclusion: Normally, λ barely has an effect on robustness of predictions.

Reason: As n grows, you almost predict from the plug-in predictive: $p_n^{\infty} \approx p_n^{\lambda}$

Cold Posterior Effect: not only deep learning; more general phenomenon

Possible Solution:

$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}$$

McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

~~(A1)~~, (A2), (A3)

1

Foundations of Post-Bayesian ML Research

1 Foundations

$\boxed{\pi_n^{(\lambda)}(\theta \mid x_{1:n})}$

$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})}$

model misspecification

~~(A1)~~ (A2) (A3)

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, **Knoblauch***, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

Q1: Can tuning λ improve robustness?

Q2: What L leads to robust posteriors π_n^{L}?

Q3: How should we design/choose L?

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

~~(A1)~~, (A2), (A3)

Post-Bayesian ML

1

Gibbs/Generalised/
Pseudo Posterior

~~(A1)~~, (A2), (A3)

$$\boxed{\pi_n^{\mathrm{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathrm{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathrm{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}}$$

Power/Fractional/
Cold Posterior

~~(A1)~~, (A2), (A3)

$$\boxed{\pi_n^{(\lambda)}(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta) d\theta}}$$

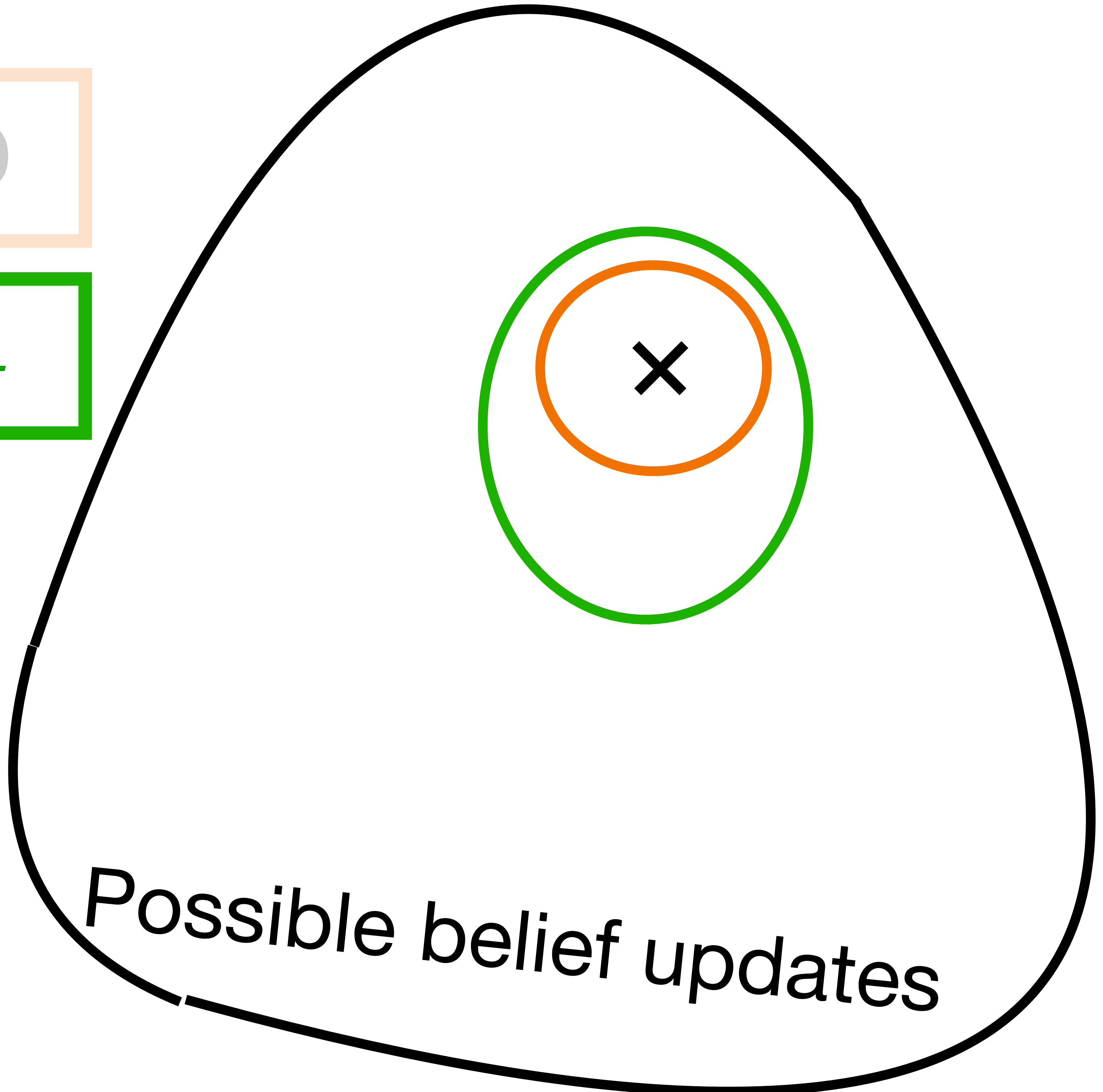
Bayes' Posterior

(A1), (A2), (A3)

$$\boxed{\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}}$$

$$\boxed{p(x_{1:n} \mid \theta) \longrightarrow p(x_{1:n} \mid \theta)^{\lambda}, \ \lambda > 0}$$

$$\boxed{p(x_{1:n} \mid \theta) \longrightarrow \exp\{-\mathrm{L}(x_{1:n}, \theta)\}, \ \text{loss } \mathrm{L}}$$



(A1)	model well-specified
(A2)	prior well-specified
(A3)	computationally feasible

~~(A1)~~, (A2), (A3)

1

Q2: What L leads to robust posteriors?

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}}$$

Setting: for some small $\varepsilon \geq 0$,

Data-generating probability distribution

ε-contamination distribution

$$q_\varepsilon = (1-\varepsilon) \cdot q_0 + \varepsilon \cdot c$$

Part of distribution our model captures

(A1)	model well-specified
(A2)	prior well-specified
(A3)	computationally feasible

~~(A1)~~, (A2), (A3)

①

Q2: What L leads to robust posteriors?

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}}$$

Setting: for some small $\varepsilon \geq 0$,

Data-generating probability distribution

ε-contamination distribution

$$q_\varepsilon = (1-\varepsilon) \cdot q_0 + \varepsilon \cdot c$$

Part of distribution our model captures

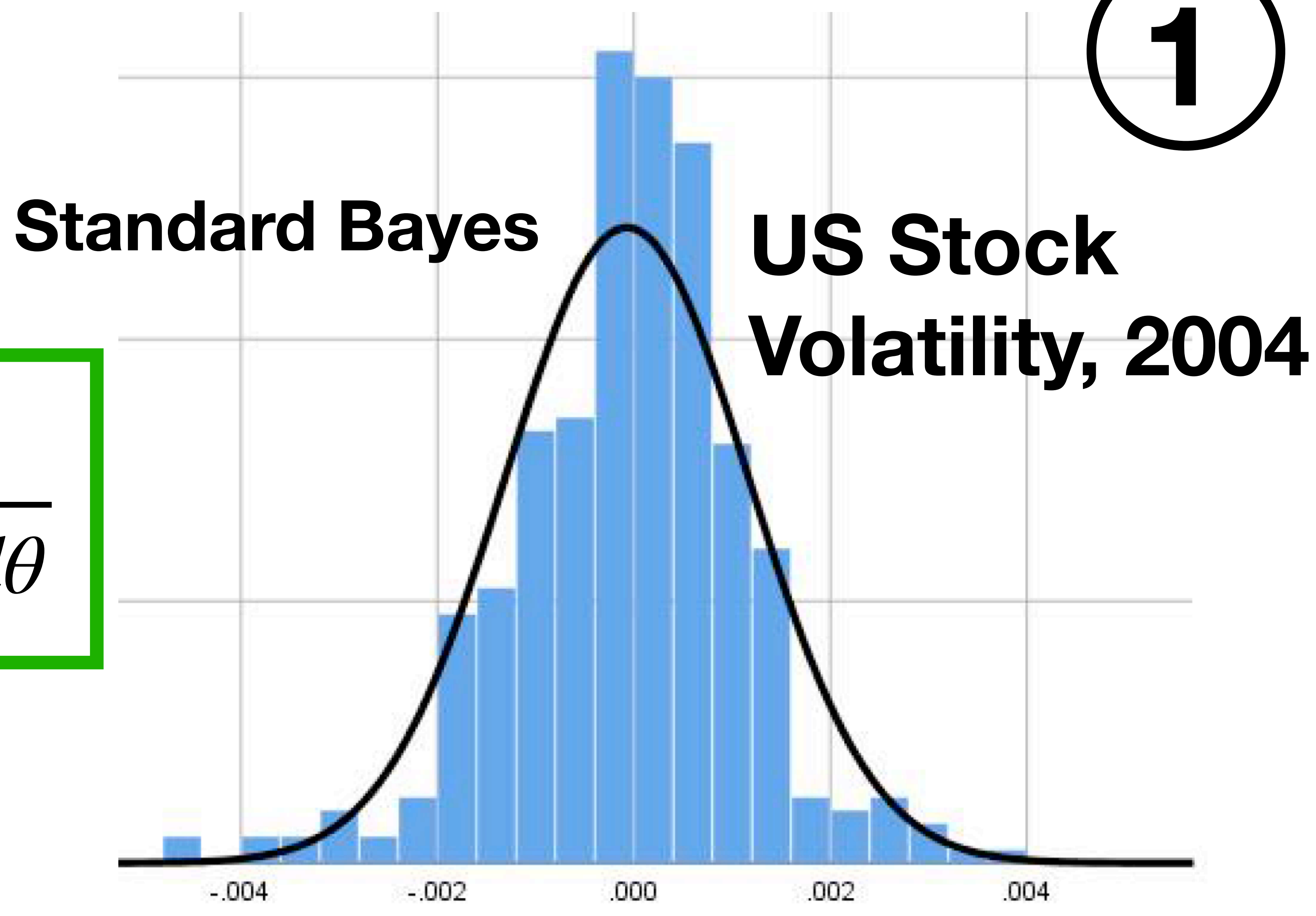
$$\boxed{\textbf{What we want: } \begin{cases} x_{1:n} \sim q_\varepsilon \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) \\ \qquad\qquad\qquad\qquad \approx \\ z_{1:n} \sim q_0 \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid z_{1:n}) \end{cases}}$$

~~(A1)~~, (A2), (A3)

①

Q2: What L leads to robust posteriors?

$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n},\theta)\}\cdot\pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n},\theta)\}\cdot\pi(\theta)d\theta}$$



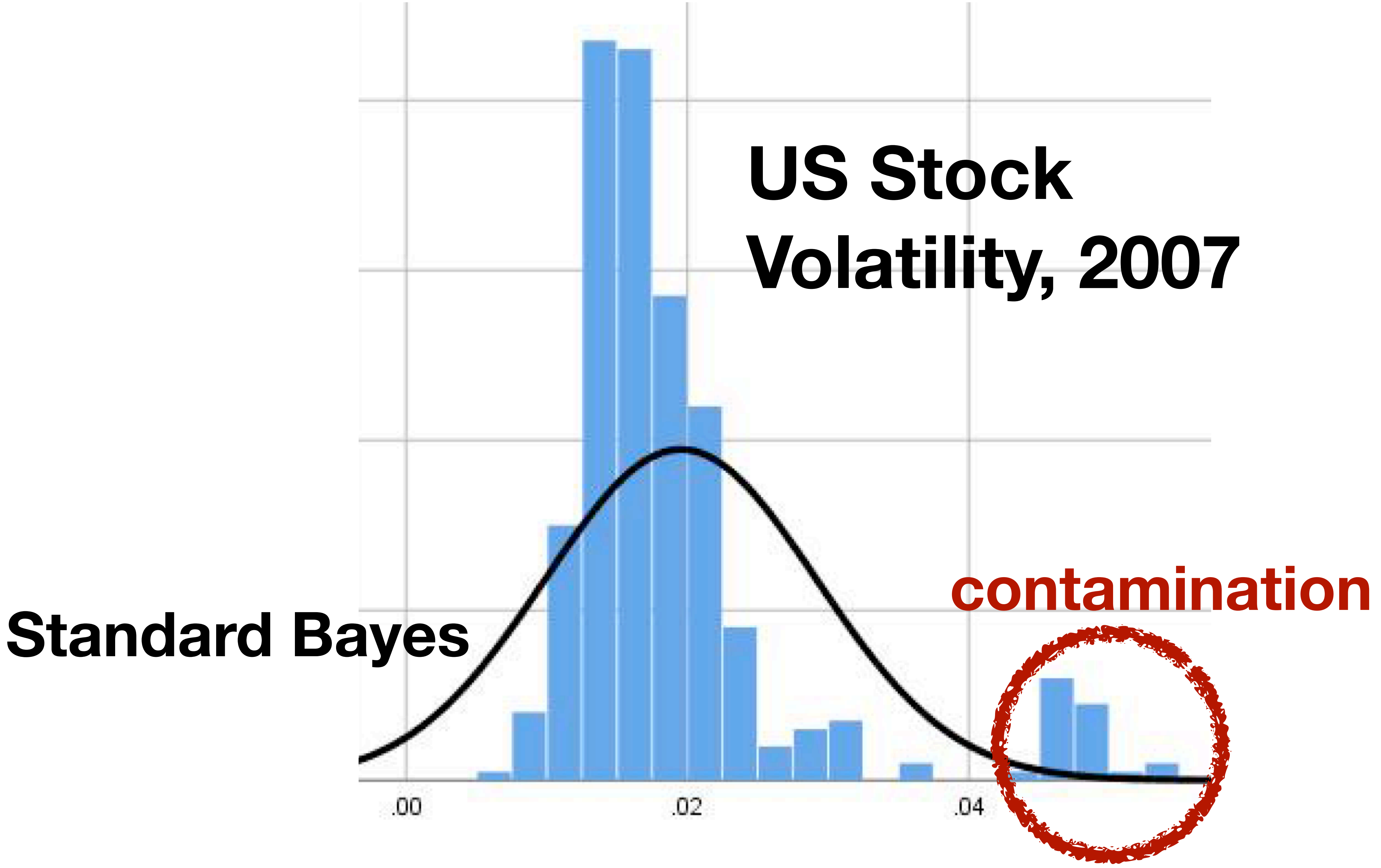
Setting: for some small $\varepsilon \geq 0$,

Data-generating probability distribution

ε-contamination distribution

$$q_\varepsilon = (1-\varepsilon)\cdot q_0 + \varepsilon \cdot c$$

Part of distribution our model captures



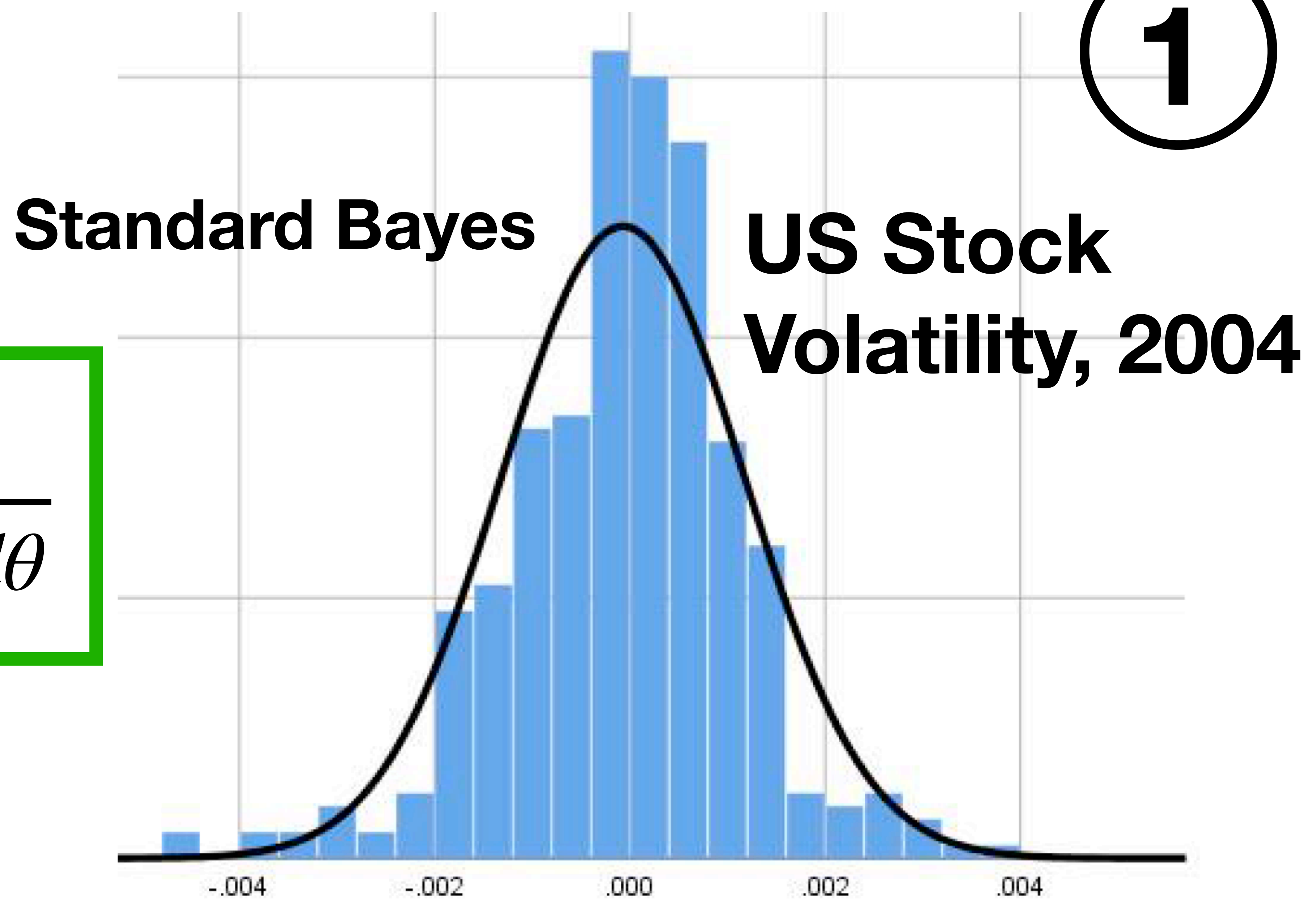
What we want:
$$\begin{cases} x_{1:n} \sim q_\varepsilon \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) \\ \qquad\qquad\qquad\quad \approx \\ z_{1:n} \sim q_0 \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid z_{1:n}) \end{cases}$$

~~(A1)~~, (A2), (A3)

1

Q2: What L leads to robust posteriors?

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}}$$



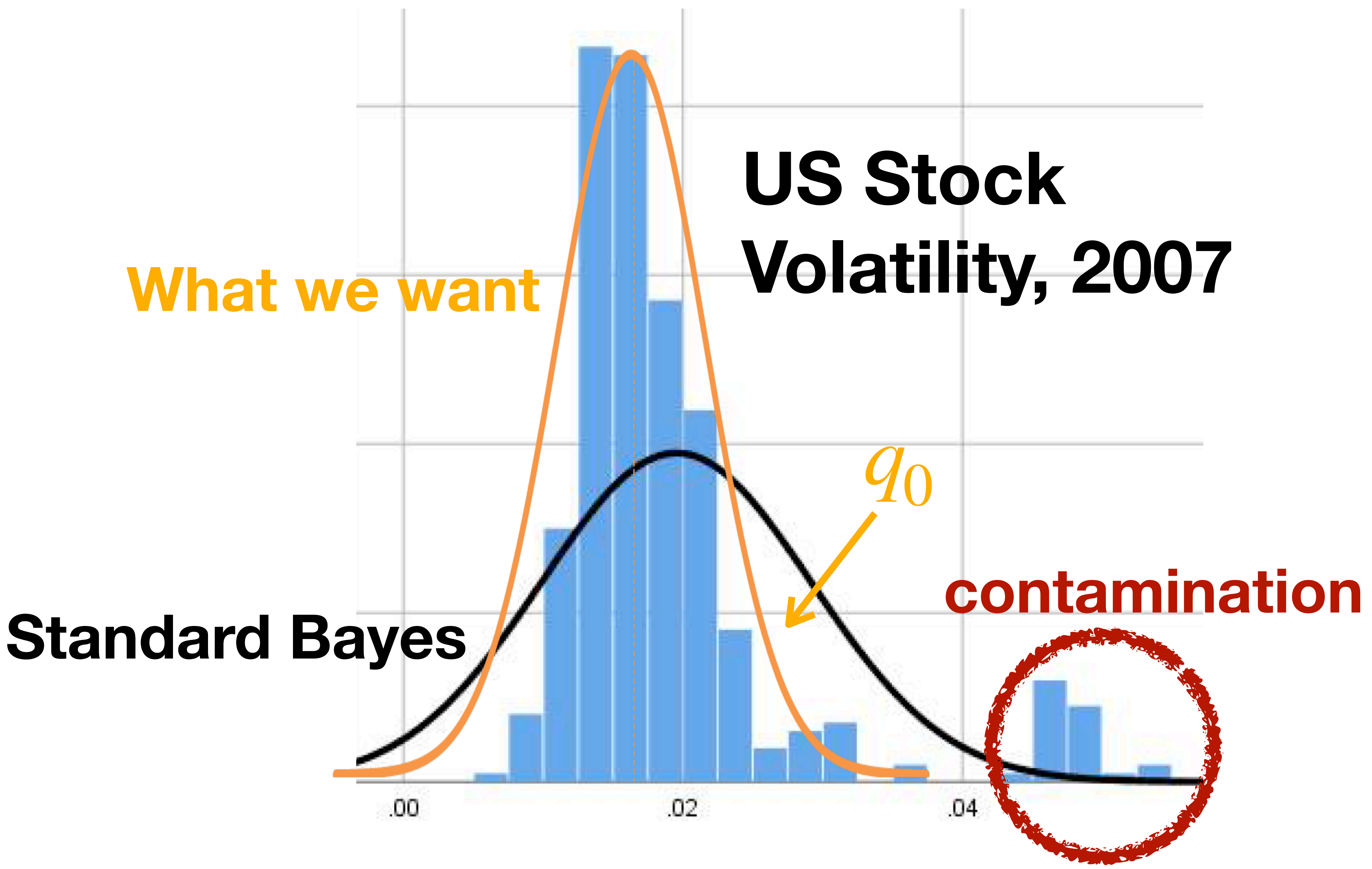
Setting: for some small $\varepsilon \geq 0$,

Data-generating probability distribution

ε-contamination distribution

$$q_\varepsilon = (1 - \varepsilon) \cdot q_0 + \varepsilon \cdot c$$

Part of distribution our model captures



$$\textbf{What we want:} \begin{cases} x_{1:n} \sim q_\varepsilon \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) \\ \qquad\qquad\qquad\qquad \approx \\ z_{1:n} \sim q_0 \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid z_{1:n}) \end{cases}$$

~~(A1)~~, (A2), (A3)

1

Q2: What L leads to robust posteriors?

Setting: $q_\varepsilon = (1-\varepsilon)\cdot q_0 + \varepsilon \cdot c$

$$x_{1:n} \sim q_\varepsilon \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$$

$$z_{1:n} \sim q_0 \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid z_{1:n})$$

Robustness: $\text{distance}\left\{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}), \pi_n^{\mathsf{L}}(\theta \mid z_{1:n})\right\} \leq \text{constant}(c)\cdot\varepsilon$

Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML

~~(A1)~~, (A2), (A3)

1

Q2: What L leads to robust posteriors?

Setting: $q_\varepsilon = (1-\varepsilon)\cdot q_0 + \varepsilon \cdot c$

$x_{1:n} \sim q_\varepsilon \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$

$z_{1:n} \sim q_0 \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid z_{1:n})$

Robustness: $\sup_{c\in\mathcal{S}} \left\{ \text{distance}\left\{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}), \pi_n^{\mathsf{L}}(\theta \mid z_{1:n})\right\}\right\} \leq \text{constant}(\mathcal{S})\cdot\varepsilon$

Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML

~~(A1)~~, (A2), (A3)

1

Q2: What L leads to robust posteriors?

$x_{1:n} \sim q_\varepsilon \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$

Setting: $q_\varepsilon = (1-\varepsilon)\cdot q_0 \, + \, \varepsilon \cdot c$

$z_{1:n} \sim q_0 \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid z_{1:n})$

Robustness: $\sup_{c \in \mathcal{S}} \left\{ \text{distance}\left\{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}), \pi_n^{\mathsf{L}}(\theta \mid z_{1:n})\right\} \right\} \leq \text{constant}(\mathcal{S}) \cdot \varepsilon$

$$= \sup_{\theta \in \Theta} \left| \pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) - \pi_n^{\mathsf{L}}(\theta \mid z_{1:n}) \right|$$

Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML

~~(A1)~~, (A2), (A3)

1

Q2: What L leads to robust posteriors?

Setting: $q_\varepsilon = (1-\varepsilon)\cdot q_0 + \varepsilon \cdot c$

$x_{1:n} \sim q_\varepsilon \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$

$z_{1:n} \sim q_0 \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid z_{1:n})$

Robustness: $\sup_{c\in\mathcal{S}} \left\{ \text{distance}\left\{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}), \pi_n^{\mathsf{L}}(\theta \mid z_{1:n})\right\} \right\} \leq \text{constant}(\mathcal{S}) \cdot \varepsilon$

$= \sup_{\theta\in\Theta} \left| \pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) - \pi_n^{\mathsf{L}}(\theta \mid z_{1:n}) \right|$

Key quantity:
Loss robustness to contamination: $\left.\frac{\partial}{\partial\varepsilon}\mathsf{L}(\theta, x_{1:n})\right|_{\varepsilon=0}$

Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML

~~(A1)~~, (A2), (A3)

1

Q2: What L leads to robust posteriors?

Setting: $q_\varepsilon = (1-\varepsilon)\cdot q_0 + \varepsilon \cdot c$

$x_{1:n} \sim q_\varepsilon \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$

$z_{1:n} \sim q_0 \longrightarrow \pi_n^{\mathsf{L}}(\theta \mid z_{1:n})$

$\boxed{\textbf{Theorem: } \pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) \text{ is robust over all } c \in \mathcal{S} \text{ if } \mathsf{L} \text{ is.}}$

Robustness: $\sup_{c\in\mathcal{S}} \left\{ \text{distance}\left\{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}), \pi_n^{\mathsf{L}}(\theta \mid z_{1:n})\right\}\right\} \le \text{constant}(\mathcal{S})\cdot\varepsilon$

$$= \sup_{\theta\in\Theta} \left| \pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) - \pi_n^{\mathsf{L}}(\theta \mid z_{1:n}) \right|$$

Key quantity:
Loss robustness to contamination: $\left.\frac{\partial}{\partial\varepsilon}\mathsf{L}(\theta, x_{1:n})\right|_{\varepsilon=0}$

Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML

~~(A1)~~, (A2), (A3)

1

Foundations of Post-Bayesian ML Research

1 Foundations

$\pi_n^{(\lambda)}(\theta \mid x_{1:n})$

$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$

model misspecification

~~(A1)~~ (A2) (A3)

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, **Knoblauch***, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

Q1: Can tuning λ improve robustness?
Q2: What L leads to robust posteriors π_n^{L}?
Q3: How should we design/choose L?

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

~~(A1)~~, (A2), (A3)

1

Q3: How should we choose L?

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}}$$

Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Dewaskar, Tosh, **Knoblauch**, & Dunson (2023); preprint
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Knoblauch*, Frazier*, & Drovandi (2024); preprint

Standard Bayes

↓

$$\mathsf{L}(x_{1:n}, \theta) = \sum_{i=1}^{n} -\log p(x_i \mid \theta)$$

~~(A1)~~, (A2), (A3)

1

Q3: How should we choose L?

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n},\theta)\}\cdot\pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n},\theta)\}\cdot\pi(\theta)d\theta}}$$

Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Dewaskar, Tosh, **Knoblauch**, & Dunson (2023); preprint
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Knoblauch*, Frazier*, & Drovandi (2024); preprint

Standard Bayes

$$n \cdot \mathrm{KL}(q_\varepsilon, p(\cdot \mid \theta)) \underset{}{\overset{x_i \sim q_\varepsilon}{\approx}} \mathsf{L}(x_{1:n},\theta) = \sum_{i=1}^{n} -\log p(x_i \mid \theta)$$

~~(A1)~~, (A2), (A3)

1

Q3: How should we choose L?

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}}$$

Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Dewaskar, Tosh, **Knoblauch**, & Dunson (2023); preprint
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Knoblauch*, Frazier*, & Drovandi (2024); preprint

NOT robust to model misspecification ~~(A1)~~

↓

$$n \cdot \mathrm{KL}(q_\varepsilon, p(\cdot \mid \theta))$$

$$\overset{x_i \sim q_\varepsilon}{\approx}$$

Standard Bayes

↓

$$\mathsf{L}(x_{1:n}, \theta) = \sum_{i=1}^{n} -\log p(x_i \mid \theta)$$

~~(A1)~~, (A2), (A3)

1

Q3: How should we choose L?

$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}$$

Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Dewaskar, Tosh, **Knoblauch**, & Dunson (2023); preprint
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Knoblauch*, Frazier*, & Drovandi (2024); preprint

NOT robust to model misspecification ~~(A1)~~

$$n \cdot \mathrm{KL}(q_\varepsilon, p(\cdot \mid \theta))$$

$$x_i \sim q_\varepsilon$$
$$\approx$$

Standard Bayes

$$\mathsf{L}(x_{1:n}, \theta) = \sum_{i=1}^{n} - \log p(x_i \mid \theta)$$

$$n \cdot \mathrm{D}(q_\varepsilon, p(\cdot \mid \theta))$$

Robust discrepancy

$$\boxed{\mathrm{D}(q_\varepsilon, p(\cdot \mid \theta)) \quad \approx \quad \mathrm{D}(q_0, p(\cdot \mid \theta))}$$

~~(A1)~~, (A2), (A3)

1

Q3: How should we choose L?

$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}$$

Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Dewaskar, Tosh, **Knoblauch**, & Dunson (2023); preprint
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Knoblauch*, Frazier*, & Drovandi (2024); preprint

NOT robust to model misspecification ~~(A1)~~

Standard Bayes

$$n \cdot \mathrm{KL}(q_\varepsilon, p(\cdot \mid \theta)) \overset{x_i \sim q_\varepsilon}{\approx} \mathsf{L}(x_{1:n}, \theta) = \sum_{i=1}^{n} -\log p(x_i \mid \theta)$$

$$n \cdot \mathrm{D}(q_\varepsilon, p(\cdot \mid \theta)) \overset{x_i \sim q_\varepsilon}{\approx} \mathsf{L}(x_{1:n}, \theta)$$

Robust discrepancy

$$\boxed{\mathrm{D}(q_\varepsilon, p(\cdot \mid \theta)) \approx \mathrm{D}(q_0, p(\cdot \mid \theta))}$$

~~(A1)~~, (A2), (A3)

1

Q3: How should we choose L?

$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}$$

Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Dewaskar, Tosh, **Knoblauch**, & Dunson (2023); preprint
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Knoblauch*, Frazier*, & Drovandi (2024); preprint

NOT robust to model misspecification ~~(A1)~~

$$n \cdot \mathrm{KL}(q_\varepsilon, p(\cdot \mid \theta)) \overset{x_i \sim q_\varepsilon}{\approx} \mathsf{L}(x_{1:n}, \theta) = \sum_{i=1}^{n} -\log p(x_i \mid \theta)$$

Standard Bayes

$$n \cdot \mathrm{D}(q_\varepsilon, p(\cdot \mid \theta)) \overset{x_i \sim q_\varepsilon}{\approx} \mathsf{L}(x_{1:n}, \theta)$$

Robust discrepancy

Robust loss

$$\boxed{\mathrm{D}(q_\varepsilon, p(\cdot \mid \theta)) \approx \mathrm{D}(q_0, p(\cdot \mid \theta))}$$

L is robust over all $c \in \mathcal{S}$

~~(A1)~~, (A2), (A3)

1

Q3: How should we choose L?

$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}$$

Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Knoblauch, Jewson, & Damoulas (2022); JMLR
Dewaskar, Tosh, **Knoblauch**, & Dunson (2023); preprint
Knoblauch*, Frazier*, & Drovandi (2024); preprint

Examples of this principle:

β-Divergence $\quad \overset{x_i \sim q_\varepsilon}{\approx} \quad \mathsf{L}^{\beta}(x_{1:n}, \theta) \xrightarrow{\beta \downarrow 0} \sum_{i=1}^{n} -\log p(x_i \mid \theta)$

Robust discrepancy **Robust loss**

$\boxed{\mathrm{D}(q_\varepsilon, p(\cdot \mid \theta)) \quad \approx \quad \mathrm{D}(q_0, p(\cdot \mid \theta))}$ ········▶ L is robust over all $c \in \mathcal{S}$

~~(A1)~~, (A2), (A3)

1

Q3: How should we choose L?

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n},\theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n},\theta)\} \cdot \pi(\theta)d\theta}}$$

Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Knoblauch, Jewson, & Damoulas (2022); JMLR
Dewaskar, Tosh, **Knoblauch**, & Dunson (2023); preprint
Knoblauch*, Frazier*, & Drovandi (2024); preprint

Examples of this principle:

γ-Divergence $\quad \overset{x_i \sim q_\varepsilon}{\approx} \quad \mathsf{L}^{\gamma}(x_{1:n},\theta) \xrightarrow{\gamma \downarrow 0} \sum_{i=1}^{n} -\log p(x_i \mid \theta)$

β-Divergence $\quad \overset{x_i \sim q_\varepsilon}{\approx} \quad \mathsf{L}^{\beta}(x_{1:n},\theta) \xrightarrow{\beta \downarrow 0} \sum_{i=1}^{n} -\log p(x_i \mid \theta)$

Robust discrepancy **Robust loss**

$\boxed{\mathrm{D}(q_\varepsilon, p(\cdot \mid \theta)) \approx \mathrm{D}(q_0, p(\cdot \mid \theta))}$ $\dashrightarrow$ L is robust over all $c \in \mathcal{S}$

~~(A1)~~, (A2), (A3)

1

Q3: How should we choose L?

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n},\theta)\}\cdot\pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n},\theta)\}\cdot\pi(\theta)d\theta}}$$

Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Knoblauch, Jewson, & Damoulas (2022); JMLR
Dewaskar, Tosh, **Knoblauch**, & Dunson (2023); preprint
Knoblauch*, Frazier*, & Drovandi (2024); preprint

For small γ / β,

$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) \approx$ Bayes w/o outliers

Examples of this principle:

γ-Divergence $\overset{x_i \sim q_\varepsilon}{\approx}$ $\mathsf{L}^\gamma(x_{1:n},\theta) \xrightarrow{\gamma \downarrow 0} \sum_{i=1}^n -\log p(x_i \mid \theta)$

β-Divergence $\overset{x_i \sim q_\varepsilon}{\approx}$ $\mathsf{L}^\beta(x_{1:n},\theta) \xrightarrow{\beta \downarrow 0} \sum_{i=1}^n -\log p(x_i \mid \theta)$

Robust discrepancy **Robust loss**

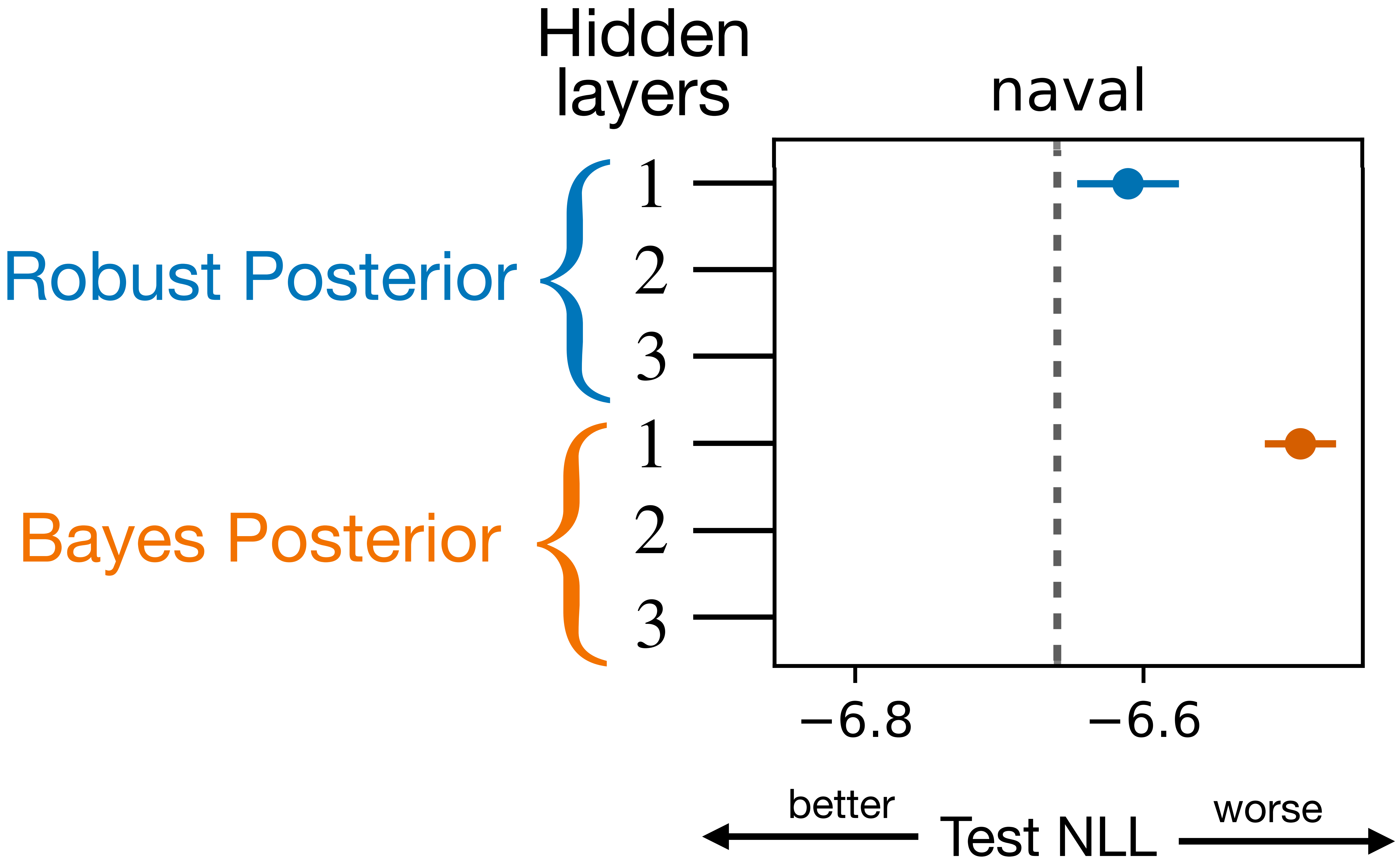
$$\boxed{\mathrm{D}(q_\varepsilon, p(\cdot \mid \theta)) \approx \mathrm{D}(q_0, p(\cdot \mid \theta))} \longrightarrow \mathsf{L} \text{ is robust over all } c \in \mathcal{S}$$

~~(A1)~~, (A2), (A3)

1

Q3: How should we choose L?

Example 1: Deep Gaussian Processes (γ-Divergence)



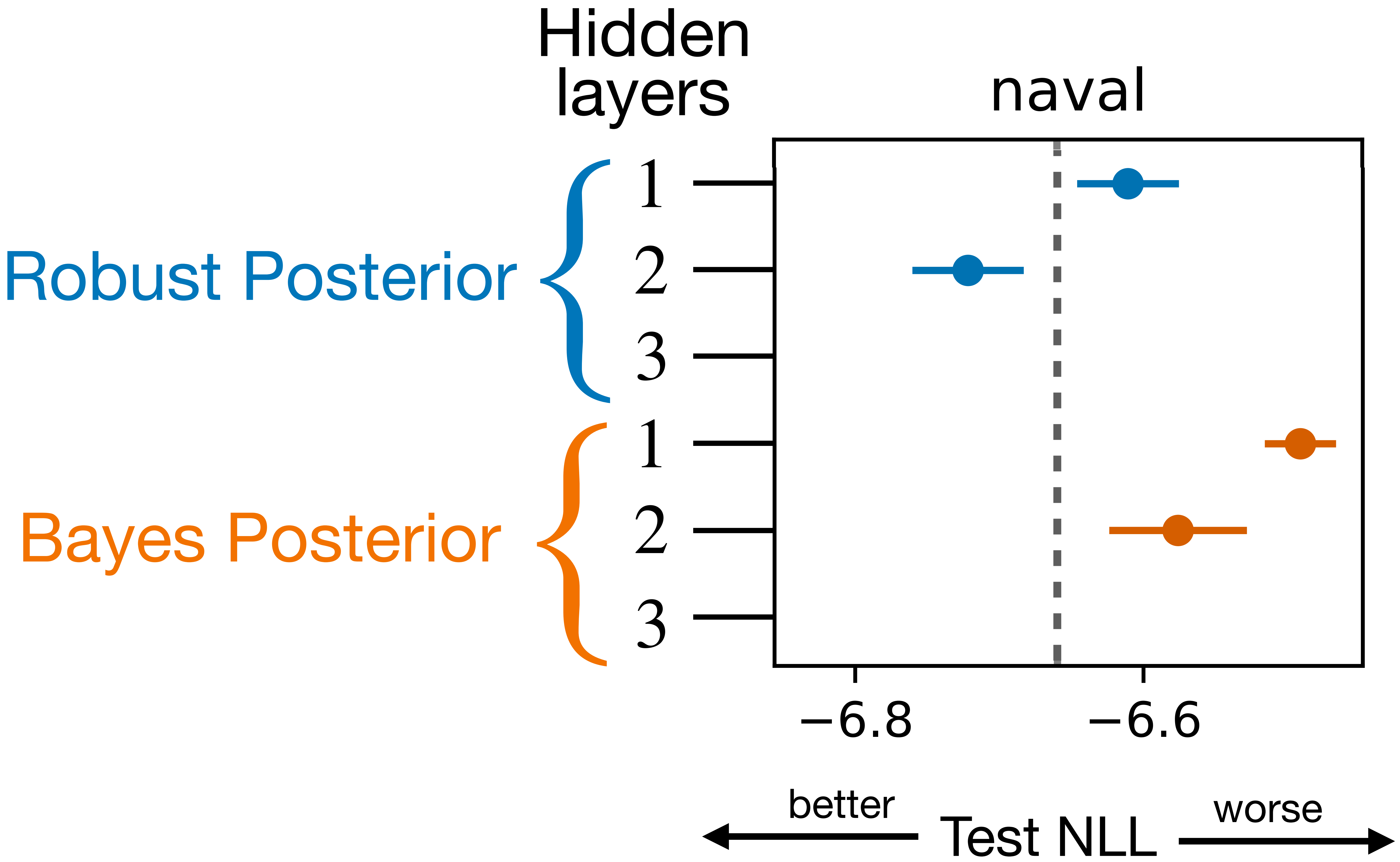
Knoblauch, Jewson, & Damoulas (2022); JMLR

~~(A1)~~, (A2), (A3)

1

Q3: How should we choose L?

Example 1: Deep Gaussian Processes (γ-Divergence)



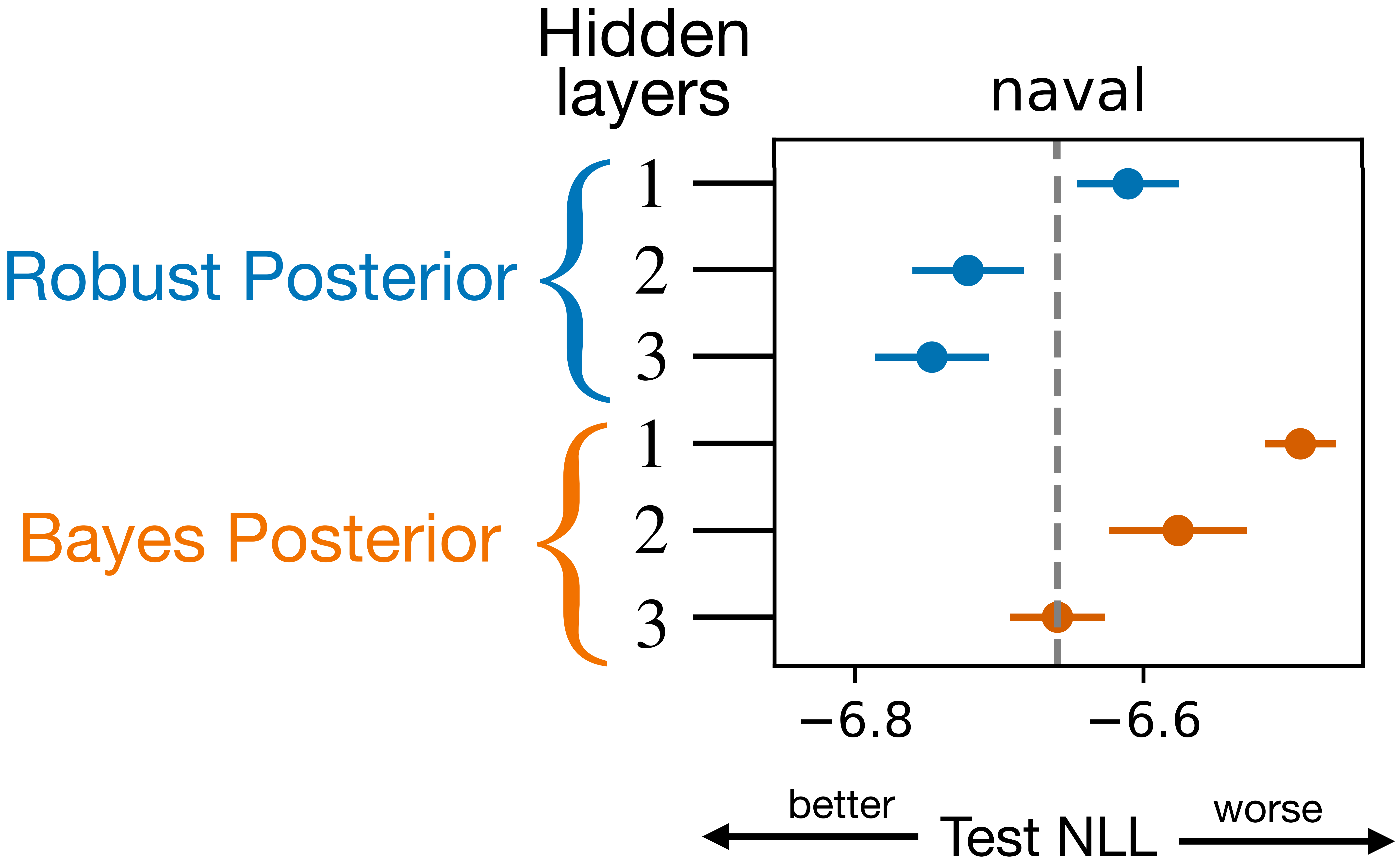
Knoblauch, Jewson, & Damoulas (2022); JMLR

~~(A1)~~, (A2), (A3)

1

Q3: How should we choose L?

Example 1: Deep Gaussian Processes (γ-Divergence)



Knoblauch, Jewson, & Damoulas (2022); JMLR

~~(A1)~~, (A2), (A3)

1

Q3: How should we choose L?

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n},\theta)\}\cdot\pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n},\theta)\}\cdot\pi(\theta)d\theta}}$$

Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Knoblauch, Jewson, & Damoulas (2022); JMLR
Dewaskar, Tosh, **Knoblauch**, & Dunson (2023); preprint
Knoblauch*, Frazier*, & Drovandi (2024); preprint

Generally intractable integrals

analytic form for most exponential families

otherwise: $\approx \frac{1}{S}\sum_{j=1}^{S} p(u_j \mid \theta)^{\beta}, \quad u_j \sim p(u_j \mid \theta)$

Problems with these losses:

γ-Divergence $\quad \overset{x_i \sim q_\varepsilon}{\approx} \quad \mathsf{L}^{\gamma}(x_{1:n},\theta) = n \cdot \int p(u \mid \theta)^{1+\beta}\, du + \ldots$

β-Divergence $\quad \overset{x_i \sim q_\varepsilon}{\approx} \quad \mathsf{L}^{\beta}(x_{1:n},\theta) = n \cdot \log \int p(u \mid \theta)^{1+\gamma}\, du + \ldots$

Robust discrepancy **Robust loss**

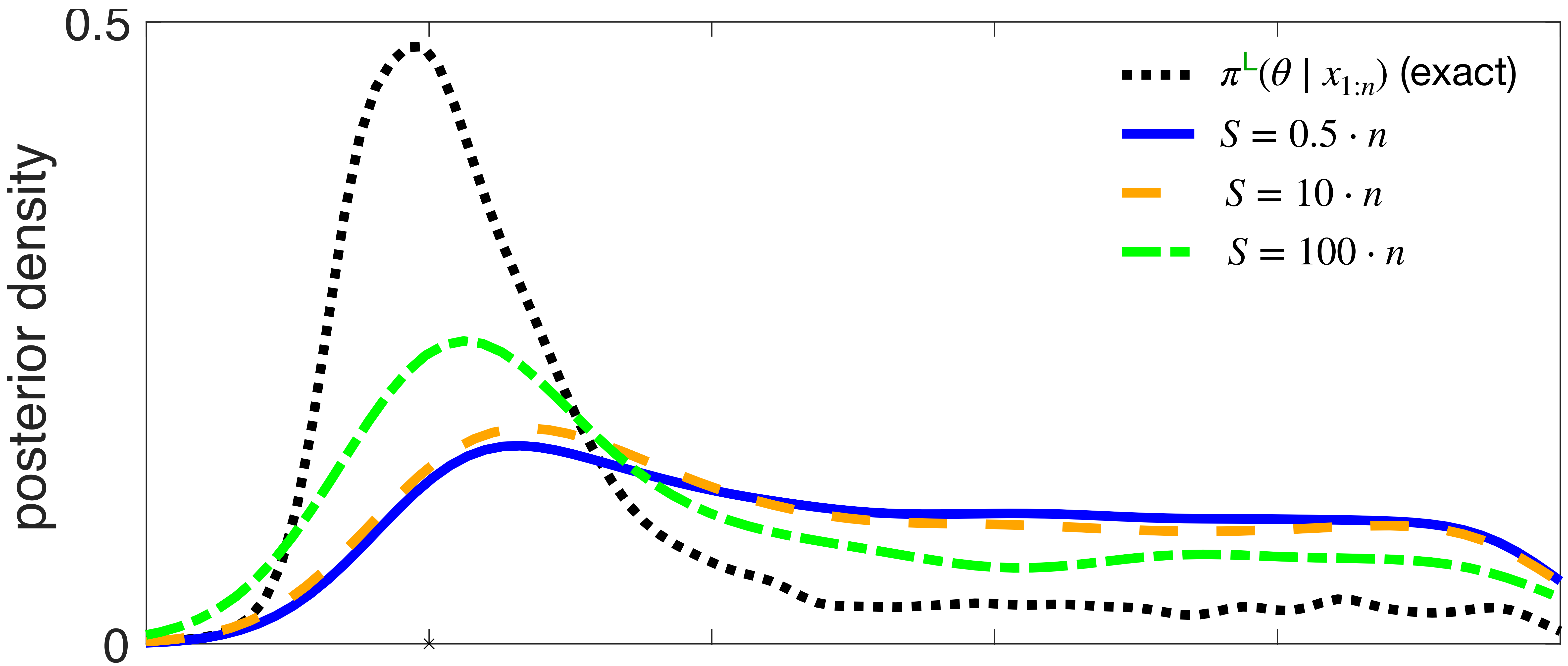
$$\boxed{\big(\mathrm{D}(q_\varepsilon, p(\cdot \mid \theta)) \approx \mathrm{D}(q_0, p(\cdot \mid \theta))\big)} \dashrightarrow \mathsf{L} \text{ is robust over all } c \in \mathcal{S}$$

~~(A1)~~, (A2), (A3)

1

Q3: How should we choose L?

Question: How many simulations for good approximation quality of $\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$?



Terrible approximations, even for large S!

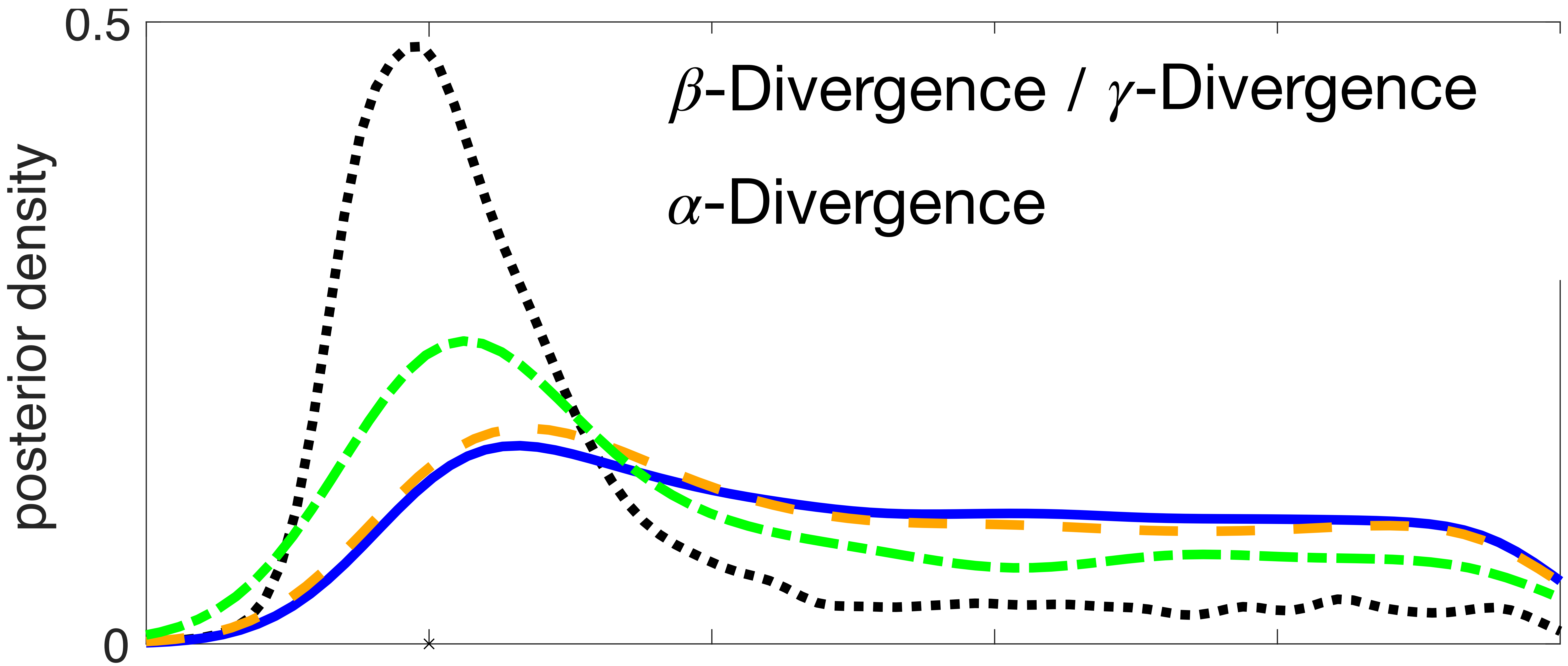
Knoblauch*, Frazier*, & Drovandi (2024); preprint

~~(A1)~~, (A2), (A3)

1

Q3: How should we choose L?

Question: How many simulations for good approximation quality of $\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$?



$\boxed{S = O(n^{2.5})}$ ✗

$\boxed{S = O(n^{2.25})}$ ✗

Terrible approximations, even for large S! ✗

Knoblauch*, Frazier*, & Drovandi (2024); preprint

~~(A1)~~, (A2), (A3)

1

Summary: Foundations of Post-Bayesian ML Research

1 Foundations

$\boxed{\pi_n^{(\lambda)}(\theta \mid x_{1:n})}$

$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})}$

model misspecification

~~(A1)~~ (A2) (A3)

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, **Knoblauch***, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

Q1: Can tuning λ improve robustness? — **A:** No. Work with **L**

Q2: What L leads to robust posteriors π_n^{L}? — **A:** robust **L** $\Longrightarrow$ robust $\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$

Q3: How should we design/choose L? — **A:** **L** = (robust) divergence

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

~~(A1)~~, (A2), (A3)

1

Summary: Foundations of Post-Bayesian ML Research

1 Foundations

$\boxed{\pi_n^{(\lambda)}(\theta \mid x_{1:n})}$

$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})}$

model misspecification

~~(A1)~~ (A2) (A3)

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, **Knoblauch***, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

Q1: Can tuning λ improve robustness? **A:** No. Work with **L**

Q2: What L leads to robust posteriors π_n^{L}? **A:** robust **L** $\Longrightarrow$ robust $\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$

Q3: How should we design/choose L? **A: L** = (robust) divergence

Robustness $\Longleftrightarrow$ Tractability

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

~~(A1)~~, (A2), ~~(A3)~~ Tackled Assumptions

Part of the Talk 2

Post-Bayesian ML Research: State of the Art

(A1)	model well-specified
(A2)	prior well-specified
(A3)	computationally feasible

1 **Foundations**

$\pi_n^{(\lambda)}(\theta \mid x_{1:n})$

$\pi_n^{\text{L}}(\theta \mid x_{1:n})$

model misspecification

~~(A1)~~ (A2) (A3)

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, **Knoblauch***, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

2 **State of the Art**

$\pi_n^{\text{L}}(\theta \mid x_{1:n})$

model misspecification + computation

~~(A1)~~ (A2) ~~(A3)~~

Schmon, Cannon, & **Knoblauch** (2020); AABI
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Dellaporta, **Knoblauch**, Damoulas, & Briol (2022); AISTATS (best paper award)
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)
Duran-Martin, Altamirano, Shestopaloff, Sanchez-Betancourt, **Knoblauch**, Briol, & Murphy (2024); ICML

3 **The Future**

$q_n^*(\theta)$

model misspecification + prior misspecification + computation

~~(A1)~~ ~~(A2)~~ ~~(A3)~~

Husain & **Knoblauch** (2022); ALT
Knoblauch, Jewson, & Damoulas (2022); JMLR
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Wild, Sejdinovic, & **Knoblauch** (2024); forthcoming
Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2024); NeurIPS (oral)

~~(A1)~~, (A2), ~~(A3)~~

2

Post-Bayesian ML Research: State of the Art

(A1)	model well-specified
(A2)	prior well-specified
(A3)	computationally feasible

2 **State of the Art**

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})}$$

model misspecification + computation

~~(A1)~~ (A2) ~~(A3)~~

Schmon, Cannon, & **Knoblauch** (2020); AABI
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Dellaporta, **Knoblauch**, Damoulas, & Briol (2022); AISTATS (best paper award)
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)
Duran-Martin, Altamirano, Shestopaloff, Sanchez-Betancourt, **Knoblauch**, Briol, & Murphy (2024); ICML

~~(A1)~~, (A2), ~~(A3)~~

2

Post-Bayesian ML Research: State of the Art

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

2 **State of the Art**

$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})}$

model misspecification + computation

~~(A1)~~ (A2) ~~(A3)~~

Schmon, Cannon, & **Knoblauch** (2020); AABI
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Dellaporta, **Knoblauch**, Damoulas, & Briol (2022); AISTATS (best paper award)
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)
Duran-Martin, Altamirano, Shestopaloff, Sanchez-Betancourt, **Knoblauch**, Briol, & Murphy (2024); ICML

Problem identified in 1 : **Robustness $\Longleftrightarrow$ Tractability**

Q: Can we design losses L that are **both robust and tractable** ?

~~(A1)~~, (A2), ~~(A3)~~

Post-Bayesian ML

2

Gibbs/Generalised/
Pseudo Posterior

~~(A1)~~, (A2), ~~(A3)~~

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}}$$

Power/Fractional/
Cold Posterior

~~(A1)~~, (A2), (A3)

$$\boxed{\pi_n^{(\lambda)}(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta) d\theta}}$$

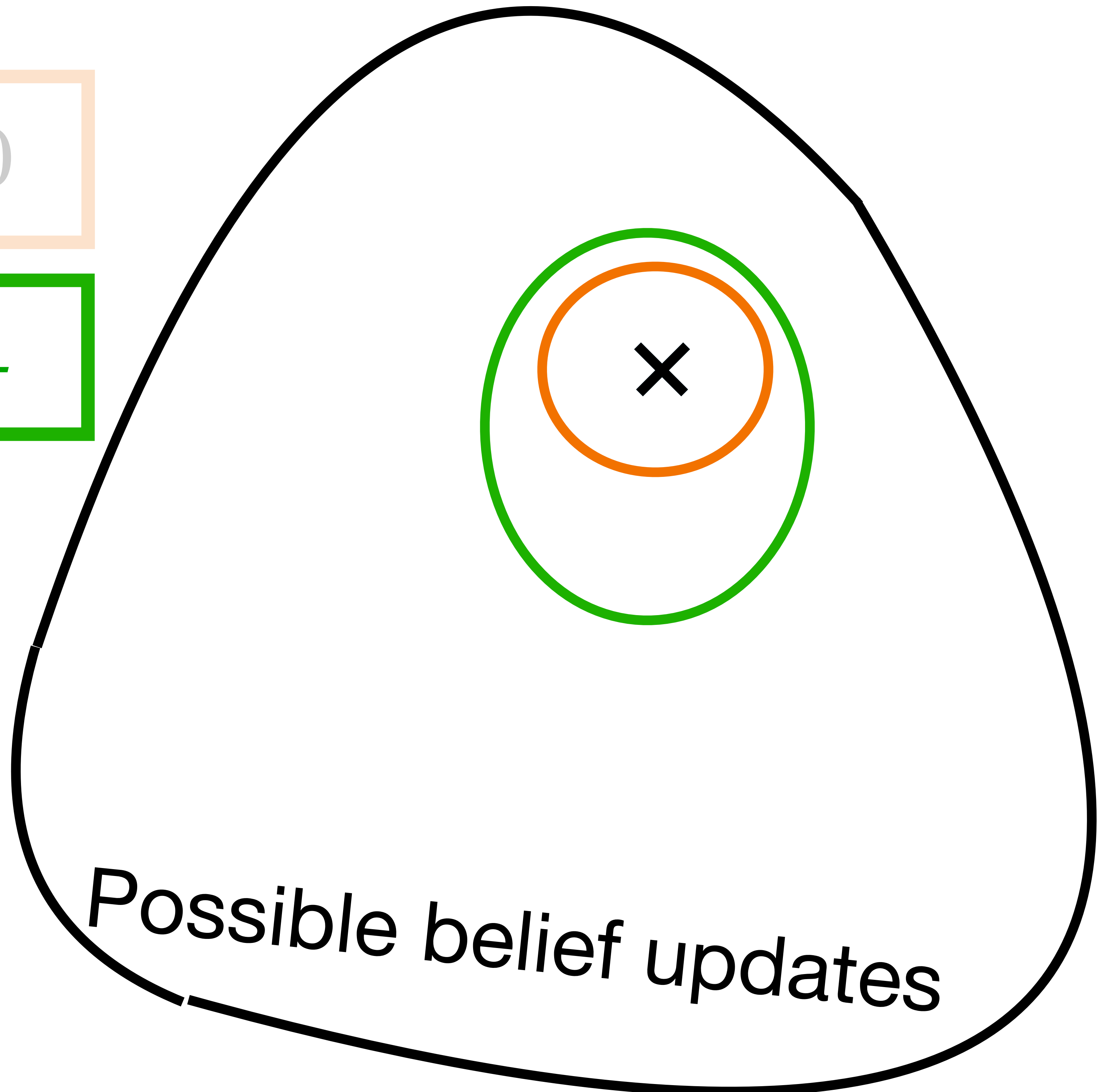
Bayes' Posterior

(A1), (A2), (A3)

$$\boxed{\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}}$$

$$\boxed{p(x_{1:n} \mid \theta) \longrightarrow p(x_{1:n} \mid \theta)^{\lambda}, \ \lambda > 0}$$

$$\boxed{p(x_{1:n} \mid \theta) \longrightarrow \exp\{-\mathsf{L}(x_{1:n}, \theta)\}, \ \text{loss } \mathsf{L}}$$



(A1)	model well-specified
(A2)	prior well-specified
(A3)	computationally feasible

~~(A1)~~, (A2), ~~(A3)~~

②

L for Robustness + Tractability

$$\boxed{\pi_n^{L}(\theta \mid x_{1:n}) = \frac{\exp\{-L(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-L(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}}$$

Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)

Inspiration: dealing with unnormalised likelihoods

Setting: $p(\cdot \mid \theta) = \underbrace{v(\cdot \mid \theta)}_{\text{can be evaluated}} \; / \; \underbrace{Z_\theta}_{= \int v(u \mid \theta)\, du}$ (intractable integral)

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

$$\pi_n^{L}(\theta \mid x_{1:n}) = \frac{\exp\{-L(x_{1:n},\theta)\} \cdot \pi(\theta)}{\int \exp\{-L(x_{1:n},\theta)\} \cdot \pi(\theta) d\theta}$$

Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)

Inspiration: dealing with unnormalised likelihoods

Setting: $p(\cdot \mid \theta) = \underbrace{v(\cdot \mid \theta)}_{\text{can be evaluated}} \,/\, \underbrace{Z_\theta}_{= \int v(u \mid \theta)\, du}$ (intractable integral)

Brute Force: $u_j \sim v(u_j \mid \theta)$

$$\int v(u \mid \theta)\, du \approx \frac{1}{S} \sum_{j=1}^{S} v(u_j \mid \theta)$$

Tractability ✗

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}$$

Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)

Inspiration: dealing with unnormalised likelihoods

Setting: $p(\cdot \mid \theta) = \underbrace{v(\cdot \mid \theta)}_{\text{can be evaluated}} \; / \; Z_\theta$

$Z_\theta = \int v(u \mid \theta)\, du$ (intractable integral)

Brute Force: $u_j \sim v(u_j \mid \theta)$

$$\int v(u \mid \theta)\, du \approx \frac{1}{S} \sum_{j=1}^{S} v(u_j \mid \theta)$$

Tractability ✗

Next!

Smart strategy:

$\Longrightarrow \mathsf{L}(x_{1:n}, \theta) =$ Score Matching / Stein Discrepancies

~~(A1)~~, (A2), ~~(A3)~~

②

L for Robustness + Tractability

Suppose: $p(x_{1:n} \mid \theta) \;=\; v(x_{1:n} \mid \theta) \,/\, Z_\theta$

$$\underbrace{\nabla_x \log v(x_{1:n} \mid \theta)}_{\text{can be evaluated}}$$

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

Suppose: $p(x_{1:n} \mid \theta) \;=\; v(x_{1:n} \mid \theta) \,/\, Z_\theta$

Stein / Hyvärinen score

$$\underbrace{\nabla_x \log v(x_{1:n} \mid \theta)}_{\text{can be evaluated}} = \frac{\nabla_x v(x_{1:n} \mid \theta)}{v(x_{1:n} \mid \theta)} = \frac{\nabla_x v(x_{1:n} \mid \theta) \,/\, Z_\theta}{v(x_{1:n} \mid \theta) \,/\, Z_\theta} = \underbrace{\nabla_x \log p(x_{1:n} \mid \theta)}_{\text{can be evaluated}}$$

Hyvärinen (2005); JMLR

~~(A1)~~, (A2), ~~(A3)~~

②

L for Robustness + Tractability

Suppose: $p(x_{1:n} \mid \theta) \;=\; v(x_{1:n} \mid \theta) \,/\, Z_\theta$

Stein / Hyvärinen score

$$\underbrace{\nabla_x \log v(x_{1:n} \mid \theta)}_{\text{can be evaluated}} = \frac{\nabla_x v(x_{1:n} \mid \theta)}{v(x_{1:n} \mid \theta)} = \frac{\nabla_x v(x_{1:n} \mid \theta) \,/\, Z_\theta}{v(x_{1:n} \mid \theta) \,/\, Z_\theta} = \underbrace{\nabla_x \log p(x_{1:n} \mid \theta)}_{\text{can be evaluated}}$$

true data-generating density q_ε

$$\underbrace{\| \nabla_x \log p(x_{1:n} \mid \theta) - \nabla_x \log q_\varepsilon(x_{1:n}) \|_2^2}_{\textbf{Score Matching}}$$

Hyvärinen (2005); JMLR

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

Suppose: $p(x_{1:n} \mid \theta) \;=\; v(x_{1:n} \mid \theta) \,/\, Z_\theta$

Stein / Hyvärinen score

$$\underbrace{\nabla_x \log v(x_{1:n} \mid \theta)}_{\text{can be evaluated}} = \frac{\nabla_x v(x_{1:n} \mid \theta)}{v(x_{1:n} \mid \theta)} = \frac{\nabla_x v(x_{1:n} \mid \theta) \,/\, Z_\theta}{v(x_{1:n} \mid \theta) \,/\, Z_\theta} = \underbrace{\nabla_x \log p(x_{1:n} \mid \theta)}_{\text{can be evaluated}}$$

true data-generating density q_ε

$$\underbrace{\| \nabla_x \log p(x_{1:n} \mid \theta) - \nabla_x \log q_\varepsilon(x_{1:n}) \|_2^2}_{\textbf{Score Matching}} \overset{+C}{=} \underbrace{\| \nabla_x \log p(x_{1:n} \mid \theta) \|_2^2 + 2 \nabla \cdot \nabla_x \log p(x_{1:n} \mid \theta)}_{\text{can be evaluated}} \;=\; \mathrm{L}(x_{1:n}, \theta)$$

Hyvärinen (2005); JMLR

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

$$n \cdot \mathrm{D_F}(q_\varepsilon, p(\,\cdot \mid \theta)) \overset{x_i \sim q_\varepsilon}{\approx} \mathsf{L}(x_{1:n}, \theta)$$

Fisher Divergence **Score Matching**

Barp, Briol, Duncan, Girolami, & Mackey (2019); NeurIPS
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

$$x_i \sim q_\varepsilon$$

$$n \cdot \mathrm{D_F}(q_\varepsilon, p(\,\cdot \mid \theta)) \quad \approx \quad \mathrm{L}(x_{1:n}, \theta)$$

Fisher Divergence | **Score Matching**

❌ NOT a robust divergence ⟶ L NOT robust

Barp, Briol, Duncan, Girolami, & Mackey (2019); NeurIPS
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

Can we generalise $\mathrm{D_F}(q_\varepsilon, p(\cdot \mid \theta))$ **to make it robust?**

$$x_i \sim q_\varepsilon$$

$$n \cdot \mathrm{D_F}(q_\varepsilon, p(\cdot \mid \theta)) \approx \mathrm{L}(x_{1:n}, \theta)$$

Fisher Divergence — **Score Matching**

✗ NOT a robust divergence ⇢ L NOT robust

Barp, Briol, Duncan, Girolami, & Mackey (2019); NeurIPS
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

✓ CAN be made robust

Stein Discrepancy

$$n \cdot \mathrm{D}_{\mathrm{SD}}(q_\varepsilon, p(\cdot \mid \theta))$$

Can we generalise $\mathrm{D}_{\mathrm{F}}(q_\varepsilon, p(\cdot \mid \theta))$ **to make it robust?**

$$n \cdot \mathrm{D}_{\mathrm{F}}(q_\varepsilon, p(\cdot \mid \theta)) \overset{x_i \sim q_\varepsilon}{\approx} \mathrm{L}(x_{1:n}, \theta)$$

Fisher Divergence **Score Matching**

✗ NOT a robust divergence $\dashrightarrow$ L NOT robust

Barp, Briol, Duncan, Girolami, & Mackey (2019); NeurIPS
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

✓ CAN be made robust ⇢ L robust

Stein Discrepancy — **Generalised Score Matching**

$$n \cdot \mathrm{D}_{\mathrm{SD}}(q_\varepsilon, p(\,\cdot \mid \theta)) \underset{\approx}{x_i \sim q_\varepsilon} \mathrm{L}(x_{1:n}, \theta)$$

Can we generalise $\mathrm{D}_{\mathrm{F}}(q_\varepsilon, p(\,\cdot \mid \theta))$ **to make it robust?**

$$n \cdot \mathrm{D}_{\mathrm{F}}(q_\varepsilon, p(\,\cdot \mid \theta)) \underset{\approx}{x_i \sim q_\varepsilon} \mathrm{L}(x_{1:n}, \theta)$$

Fisher Divergence — **Score Matching**

✗ NOT a robust divergence ⇢ L NOT robust

Barp, Briol, Duncan, Girolami, & Mackey (2019); NeurIPS
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)

~~(A1)~~, (A2), ~~(A3)~~

②

L for Robustness + Tractability

$$n \cdot \mathrm{D}_{\mathrm{SD}}(p(\,\cdot \mid \theta), q_\varepsilon) \;=\; \sup_{f \in \mathcal{F}_\theta} \left| \mathbb{E}_{X \sim q_\varepsilon}\left[f(X)\right] - \mathbb{E}_{X \sim p(X|\theta)}\left[f(X)\right] \right|$$

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

$$n \cdot \mathrm{D}_{\mathrm{SD}}(p(\cdot \mid \theta), q_\varepsilon) = \sup_{f \in \mathscr{F}_\theta} \left| \mathbb{E}_{X \sim q_\varepsilon}\left[f(X)\right] - \underbrace{\cancel{\mathbb{E}_{X \sim p(X|\theta)}\left[f(X)\right]}}_{=0} \right|$$

Designed so that

$$\boxed{\mathbb{E}_{X \sim p(X|\theta)}\left[f(X)\right] = 0 \text{ for all } f \in \mathscr{F}_\theta}$$

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

$$n \cdot \mathrm{D}_{\mathrm{SD}}(p(\cdot \mid \theta), q_\varepsilon) = \sup_{f \in \mathscr{F}_\theta} \left| \mathbb{E}_{X \sim q_\varepsilon}[f(X)] - \underbrace{\cancel{\mathbb{E}_{X \sim p(X|\theta)}[f(X)]}}_{=0} \right| = \mathbb{E}_{X \sim q_\varepsilon}[f^*(X)]$$

(weighted Langevin-Stein operator
+ Stein set as RKHS / $C^1 \cup L^2$)

Designed so that

$$\boxed{\mathbb{E}_{X \sim p(X|\theta)}[f(X)] = 0 \text{ for all } f \in \mathscr{F}_\theta}$$

$$\boxed{\mathscr{F}_\theta \text{ ensures that supremum has a closed form solution}}$$

~~(A1)~~, (A2), ~~(A3)~~

②

L for Robustness + Tractability

$$n \cdot \mathrm{D}_{\mathrm{SD}}(p(\cdot \mid \theta), q_\varepsilon) = \sup_{f \in \mathcal{F}_\theta} \left| \mathbb{E}_{X\sim q_\varepsilon}\left[f(X)\right] - \underbrace{\cancel{\mathbb{E}_{X\sim p(X|\theta)}\left[f(X)\right]}}_{=0} \right| = \mathbb{E}_{X\sim q_\varepsilon}\left[f^*(X)\right] \overset{x_i \sim q_\varepsilon}{\approx} \mathrm{L}(x_{1:n}, \theta)$$

(weighted Langevin-Stein operator
+ Stein set as RKHS / $C^1 \cup L^2$)

Designed so that

$\boxed{\mathbb{E}_{X\sim p(X|\theta)}\left[f(X)\right] = 0 \text{ for all } f \in \mathcal{F}_\theta}$

$\boxed{\mathcal{F}_\theta \text{ ensures that supremum has a closed form solution}}$

$\boxed{\text{All } f(\cdot) \in \mathcal{F}_\theta \text{ depend on } \theta \text{ only via } w(\cdot) \cdot \nabla_x \log p(\cdot \mid \theta)}$

~~(A1)~~, (A2), ~~(A3)~~

②

L for Robustness + Tractability

$$n \cdot \mathrm{D}_{\mathrm{SD}}(p(\cdot \mid \theta), q_\varepsilon) = \sup_{f \in \mathscr{F}_\theta} \left| \mathbb{E}_{X \sim q_\varepsilon}[f(X)] - \underbrace{\cancel{\mathbb{E}_{X \sim p(X|\theta)}[f(X)]}}_{=0} \right| = \mathbb{E}_{X \sim q_\varepsilon}[f^*(X)] \overset{x_i \sim q_\varepsilon}{\approx} \mathrm{L}(x_{1:n}, \theta)$$

Designed so that

$\boxed{\mathbb{E}_{X \sim p(X|\theta)}[f(X)] = 0 \text{ for all } f \in \mathscr{F}_\theta}$

$\boxed{\mathscr{F}_\theta \text{ ensures that supremum has a closed form solution}}$

(weighted Langevin-Stein operator
+ Stein set as RKHS / $C^1 \cup L^2$)

$\boxed{\text{All } f(\cdot) \in \mathscr{F}_\theta \text{ depend on } \theta \text{ only via } w(\cdot) \cdot \underbrace{\nabla_x \log p(\cdot \mid \theta)}_{\text{can be evaluated}}}$

Robustness

can be evaluated

$\Longrightarrow$ L based on Stein Discrepancies = robust & tractable

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

Comparison: $\pi_n^{\text{L}}(\theta \mid x_{1:n}) \iff \pi_n(\theta \mid x_{1:n})$

All settings	$\pi_n^{\text{L}}(\theta \mid x_{1:n})$ **robust** via $w(\cdot)$	$\pi_n(\theta \mid x_{1:n})$ **is not**	✓

$\cancel{(A1)}$, (A2), $\cancel{(A3)}$

②

L for Robustness + Tractability

Comparison:	$\pi_n^{L}(\theta \mid x_{1:n})$	$\Longleftrightarrow$	$\pi_n(\theta \mid x_{1:n})$	
All settings	$\pi_n^{L}(\theta \mid x_{1:n})$	robust via $w(\cdot)$	$\pi_n(\theta \mid x_{1:n})$ **is not**	✓
$p(\cdot \mid \theta) = \underbrace{v(\cdot \mid \theta)}_{\text{tractable}} / \underbrace{Z_\theta}_{\text{intractable}}$	$\pi_n^{L}(\theta \mid x_{1:n})$	**faster** to compute than	$\pi_n(\theta \mid x_{1:n})$	✓

~~(A1)~~, (A2), ~~(A3)~~

②

L for Robustness + Tractability

Comparison: $\pi_n^{\mathrm{L}}(\theta \mid x_{1:n}) \iff \pi_n(\theta \mid x_{1:n})$

All settings	$\pi_n^{\mathrm{L}}(\theta \mid x_{1:n})$	robust via $w(\cdot)$	$\pi_n(\theta \mid x_{1:n})$ **is not**	✓
$p(\cdot \mid \theta) = \underbrace{v(\cdot \mid \theta)}_{\text{tractable}} / \underbrace{Z_\theta}_{\text{intractable}}$	$\pi_n^{\mathrm{L}}(\theta \mid x_{1:n})$	**faster** to compute than	$\pi_n(\theta \mid x_{1:n})$	✓
$\underbrace{p(\cdot \mid \theta)}_{\text{tractable}}$	$\pi_n^{\mathrm{L}}(\theta \mid x_{1:n})$	**roughly as fast** as	$\pi_n(\theta \mid x_{1:n})$	✓

~~(A1)~~, (A2), ~~(A3)~~ ②

L for Robustness + Tractability

Comparison: $\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) \iff \pi_n(\theta \mid x_{1:n})$

All settings	$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$	robust via $w(\cdot)$	$\pi_n(\theta \mid x_{1:n})$ **is not**	✓
$p(\cdot \mid \theta) = \underbrace{v(\cdot \mid \theta)}_{\text{tractable}} / \underbrace{Z_\theta}_{\text{intractable}}$	$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$	**faster** to compute than	$\pi_n(\theta \mid x_{1:n})$	✓
$\underbrace{p(\cdot \mid \theta)}_{\text{tractable}}$	$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$	**roughly as fast** as	$\pi_n(\theta \mid x_{1:n})$	✓
$p(\cdot \mid \theta)$ exponential family + $\pi(\theta)$ conjugate	$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$	**??? Slower, surely ???**	$\pi_n(\theta \mid x_{1:n})$	**???**

Next!

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

$$n \cdot \mathrm{D}_{\mathrm{SD}}(p(\cdot \mid \theta), q_\varepsilon) \overset{x_i \sim q_\varepsilon}{\approx} \mathrm{L}(x_{1:n}, \theta) = \boxed{\sum_{i=1}^{n} \big(\theta - \mu(x_i)\big)^\top \Lambda(x_i)\big(\theta - \mu(x_i)\big)}$$

exponential family
(possibly with intractable normaliser)

Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

$$\exp\left\{-n \cdot \mathrm{D}_{\mathrm{SD}}(p(\cdot \mid \theta), q_\varepsilon)\right\} \overset{x_i \sim q_\varepsilon}{\approx} \exp\left\{-\mathsf{L}(x_{1:n}, \theta)\right\} = \underbrace{\exp\left\{-\boxed{\sum_{i=1}^{n}\left(\theta-\mu(x_i)\right)^\top \Lambda(x_i)\left(\theta-\mu(x_i)\right)}\right\}}$$

exponential family

(possibly with intractable normaliser)

= unnormalised Squared Exponential / Gaussian in θ

$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}$$

Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)

$\cancel{(A1)}$, (A2), $\cancel{(A3)}$

L for Robustness + Tractability

$$\exp\left\{-n\cdot D_{SD}(p(\cdot\mid\theta), q_\varepsilon)\right\} \overset{x_i\sim q_\varepsilon}{\approx} \exp\left\{-L(x_{1:n},\theta)\right\} = \underbrace{\exp\left\{-\boxed{\sum_{i=1}^{n}(\theta-\mu(x_i))^\top\Lambda(x_i)(\theta-\mu(x_i))}\right\}}$$

exponential family

(possibly with intractable normaliser)

= unnormalised Squared Exponential / Gaussian in θ

conjugate prior!

$$\boxed{\pi_n^L(\theta\mid x_{1:n})} \propto \exp\left\{-\sum_{i=1}^{n}(\theta-\mu(x_i))^\top\Lambda(x_i)(\theta-\mu(x_i))\right\}\underbrace{\exp\left\{(\theta-\mu_0)^\top\Lambda_0(\theta-\mu_0)\right\}}_{\text{squared exponential prior}}$$

$$\boxed{\pi_n^L(\theta\mid x_{1:n}) = \frac{\exp\{-L(x_{1:n},\theta)\}\cdot\pi(\theta)}{\int\exp\{-L(x_{1:n},\theta)\}\cdot\pi(\theta)d\theta}}$$

Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)

~~(A1)~~, (A2), ~~(A3)~~

②

L for Robustness + Tractability

$$\exp\left\{-n\cdot D_{SD}(p(\,\cdot\mid\theta), q_\varepsilon)\right\} \overset{x_i\sim q_\varepsilon}{\approx} \exp\left\{-L(x_{1:n},\theta)\right\} = \underbrace{\exp\left\{-\boxed{\sum_{i=1}^{n}\left(\theta-\mu(x_i)\right)^\top\Lambda(x_i)\left(\theta-\mu(x_i)\right)}\right\}}$$

exponential family
(possibly with intractable normaliser)

= unnormalised Squared Exponential / Gaussian in θ

conjugate prior!

$$\boxed{\pi_n^L(\theta\mid x_{1:n})} \propto \exp\left\{-\sum_{i=1}^{n}\left(\theta-\mu(x_i)\right)^\top\Lambda(x_i)\left(\theta-\mu(x_i)\right)\right\}\underbrace{\exp\left\{\left(\theta-\mu_0\right)^\top\Lambda_0\left(\theta-\mu_0\right)\right\}}_{\text{squared exponential prior}}$$

$$\boxed{=\mathcal{N}\left(\theta;\mu_L(x_{1:n}),\Sigma_L(x_{1:n})\right)}$$

Closed form / conjugate Post-Bayesian posterior

Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

$$\exp\left\{-n\cdot D_{SD}(p(\cdot\mid\theta),q_\varepsilon)\right\}\overset{x_i\sim q_\varepsilon}{\approx}\exp\left\{-L(x_{1:n},\theta)\right\}=\underbrace{\exp\left\{-\boxed{\sum_{i=1}^n(\theta-\mu(x_i))^\top\Lambda(x_i)(\theta-\mu(x_i))}\right\}}$$

exponential family

(possibly with intractable normaliser)

= unnormalised Squared Exponential / Gaussian in θ

conjugate prior!

$$\boxed{\pi_n^L(\theta\mid x_{1:n})}\propto\exp\left\{-\sum_{i=1}^n(\theta-\mu(x_i))^\top\Lambda(x_i)(\theta-\mu(x_i))\right\}\underbrace{\exp\left\{(\theta-\mu_0)^\top\Lambda_0(\theta-\mu_0)\right\}}_{\text{squared exponential prior}}$$

$$\boxed{=\mathcal{N}\left(\theta;\mu_L(x_{1:n}),\Sigma_L(x_{1:n})\right)}$$

Closed form / conjugate Post-Bayesian posterior

even if $p(\cdot\mid\theta)=\underbrace{v(\cdot\mid\theta)}_{\text{tractable}}\;/\;\underbrace{Z_\theta}_{\text{intractable}}$!!!

Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

Comparison: $\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) \iff \pi_n(\theta \mid x_{1:n})$

All settings	$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$	robust via $w(\cdot)$	$\pi_n(\theta \mid x_{1:n})$ **is not**	✓
$p(\cdot \mid \theta) = \underbrace{v(\cdot \mid \theta)}_{\text{tractable}} / \underbrace{Z_\theta}_{\text{intractable}}$	$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$	**faster** to compute than	$\pi_n(\theta \mid x_{1:n})$	✓
$\underbrace{p(\cdot \mid \theta)}_{\text{tractable}}$	$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$	**roughly as fast** as	$\pi_n(\theta \mid x_{1:n})$	✓
$p(\cdot \mid \theta)$ exponential family + $\pi(\theta)$ conjugate	$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$	**as fast as / faster** to compute than	$\pi_n(\theta \mid x_{1:n})$	✓

$\cancel{(A1)}$, (A2), $\cancel{(A3)}$

②

L for Robustness + Tractability

$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta} \boxed{= \mathcal{N}\left(\theta; \mu_{\mathsf{L}}(x_{1:n}), \Sigma_{\mathsf{L}}(x_{1:n})\right)}$$

Like sufficient statistics!
can be updated with simple algebra

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta} \boxed{= \mathcal{N}\left(\theta; \mu_{\mathsf{L}}(x_{1:n}), \Sigma_{\mathsf{L}}(x_{1:n})\right)}$$

Like sufficient statistics!
can be updated with simple algebra

Graphical Modelling

Proposition 2. *Consider $\mathcal{X} = \mathbb{R}^d$ and the Langevin Stein operator $\mathcal{S}_{\mathbb{P}_\theta}$ in (5), where $\mathbb{P}_\theta$ is the exponential family in (10), and a kernel $K \in C_b^{1,1}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^{d\times d})$. Assuming the prior has a p.d.f. π, the KSD-Bayes generalised posterior has a p.d.f.*

$$\pi_n^D(\theta) \propto \pi(\theta) \exp\left(-\beta n\{\eta(\theta) \cdot \Lambda_n \eta(\theta) + \eta(\theta) \cdot \nu_n\}\right),$$

where $\Lambda_n \in \mathbb{R}^{k\times k}$ and $\nu_n \in \mathbb{R}^k$ are defined as

$$\boxed{\begin{aligned} \Lambda_n &:= \frac{1}{n^2}\sum_{i,j=1}^n \nabla t(x_i) \cdot K(x_i, x_j) \nabla t(x_j), \\ \nu_n &:= \frac{1}{n^2}\sum_{i,j=1}^n \nabla t(x_i) \cdot (\nabla_{x_j} \cdot K(x_i,x_j)) + \nabla t(x_j) \cdot (\nabla_{x_i} \cdot K(x_i,x_j)) + 2\nabla t(x_i) \cdot K(x_i,x_j)\nabla b(x_j). \end{aligned}}$$

For a natural exponential family we have $\eta(\theta) = \theta$, and the prior $\pi(\theta) \propto \exp(-\frac{1}{2}(\theta-\mu) \cdot \Sigma^{-1}(\theta-\mu))$ leads to a generalised posterior

$$\pi_n^D(\theta) \propto \exp\left(-\frac{1}{2}(\theta - \mu_n) \cdot \Sigma_n^{-1}(\theta - \mu_n)\right),$$

where $\Sigma_n^{-1} := \Sigma^{-1} + 2\beta n \Lambda_n$ and $\mu_n := \Sigma_n^{-1}(\Sigma^{-1}\mu - \nu_n)$.

Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B

Changepoints

Proposition 3.1. *If p_θ is given by (5), then*

$$\pi_\omega^{\mathcal{D}_m}(\theta | x_{1:T}) \propto \pi(\theta) \exp(-\omega T[\eta(\theta)^\top \Lambda_T \eta(\theta) + \eta(\theta)^\top \nu_T]),$$

for $\Lambda_T = \frac{1}{T}\sum_{t=1}^T \Lambda(x_t)$, $\nu_T = \frac{2}{T}\sum_{t=1}^T \nu(x_t)$, and

$$\boxed{\begin{aligned} \Lambda(x) &= (\nabla r^\top m m^\top \nabla r)(x), \\ \nu(x) &= \left(\nabla r^\top m m^\top \nabla b + \nabla \cdot (m m^\top \nabla r)\right)(x). \end{aligned}}$$

Taking $\eta(\theta) = \theta$ and choosing a squared exponential prior $\pi(\theta) \propto \exp\left(-\frac{1}{2}(\theta-\mu)^\top \Sigma^{-1}(\theta-\mu)\right)$, also makes $\pi_\omega^{\mathcal{D}_m}(\theta|x_{1:T})$ a (truncated) normal of the form

$$\pi_\omega^{\mathcal{D}_m}(\theta|x_{1:T}) \propto \exp\left(-\tfrac{1}{2}(\theta - \mu_T)^\top \Sigma_T^{-1}(\theta - \mu_T)\right),$$

for $\Sigma_T^{-1} = \Sigma^{-1} + 2\omega T \Lambda_T$ and $\mu_T = \Sigma_T\left(\Sigma^{-1}\mu - \omega T \nu_T\right)$.

Altamirano, Briol, & **Knoblauch** (2023); ICML

Gaussian Processes

Proposition 3.1. *Suppose $f \sim \mathcal{GP}(m,k)$ and $\varepsilon \sim \mathcal{N}(0, I_n\sigma^2)$. Then, the RCGP posterior is*

$$p^w(\mathbf{f}|\mathbf{y},\mathbf{x}) = \mathcal{N}(\mathbf{f}; \mu^R, \Sigma^R),$$

$$\boxed{\begin{aligned} \mu^R &= \mathbf{m} + K\left(K + \sigma^2 J_{\mathbf{w}}\right)^{-1}(\mathbf{y} - \mathbf{m}_{\mathbf{w}}), \\ \Sigma^R &= K\left(K + \sigma^2 J_{\mathbf{w}}\right)^{-1}\sigma^2 J_{\mathbf{w}}, \end{aligned}}$$

for $\mathbf{w} = (w(x_1,y_1), \ldots, w(x_n,y_n))^\top$, $\mathbf{m}_{\mathbf{w}} = \mathbf{m} + \sigma^2\nabla_y \log(\mathbf{w}^2)$ and $J_{\mathbf{w}} = \mathrm{diag}(\frac{\sigma^2}{2}\mathbf{w}^{-2})$. The RCGP's posterior predictive over $f_\star = f(x_\star)$ at $x^\star \in \mathcal{X}$ is

$$\begin{aligned} p^w(f_\star|x_\star,\mathbf{x},\mathbf{y}) &= \int_{\mathbb{R}^n} p(f_\star|x_\star,\mathbf{f},\mathbf{x},\mathbf{y}) p^w(\mathbf{f}|\mathbf{y},\mathbf{x}) d\mathbf{f} \\ &= \mathcal{N}(f_\star; \mu_\star^R, \Sigma_\star^R), \end{aligned}$$

$$\begin{aligned} \mu_\star^R &= m_\star + \mathbf{k}_\star^\top\left(K + \sigma^2 J_{\mathbf{w}}\right)^{-1}(\mathbf{y} - \mathbf{m}_{\mathbf{w}}), \\ \Sigma_\star^R &= k_{\star\star} - \mathbf{k}_\star^\top (K + \sigma^2 J_{\mathbf{w}})^{-1}\mathbf{k}_\star. \end{aligned}$$

Altamirano, Briol, & **Knoblauch** (2024); ICML spotlight

~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability

$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta} \boxed{= \mathcal{N}\left(\theta; \mu_{\mathsf{L}}(x_{1:n}), \Sigma_{\mathsf{L}}(x_{1:n})\right)}$$

Next!

Like sufficient statistics!
can be updated with simple algebra

$\Longrightarrow$ Feasible for on-line problems!

Graphical Modelling

Proposition 2. *Consider $\mathcal{X} = \mathbb{R}^d$ and the Langevin Stein operator $\mathcal{S}_{\mathbb{P}_\theta}$ in (5), where $\mathbb{P}_\theta$ is the exponential family in (10), and a kernel $K \in C_b^{1,1}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^{d\times d})$. Assuming the prior has a p.d.f. π, the KSD-Bayes generalised posterior has a p.d.f.*

$$\pi_n^D(\theta) \propto \pi(\theta) \exp\left(-\beta n\{\eta(\theta) \cdot \Lambda_n \eta(\theta) + \eta(\theta) \cdot \nu_n\}\right),$$

where $\Lambda_n \in \mathbb{R}^{k\times k}$ and $\nu_n \in \mathbb{R}^k$ are defined as

$$\boxed{\begin{aligned}\Lambda_n &:= \frac{1}{n^2} \sum_{i,j=1}^n \nabla t(x_i) \cdot K(x_i, x_j) \nabla t(x_j),\\ \nu_n &:= \frac{1}{n^2} \sum_{i,j=1}^n \nabla t(x_i) \cdot (\nabla_{x_j} \cdot K(x_i, x_j)) + \nabla t(x_j) \cdot (\nabla_{x_i} \cdot K(x_i, x_j)) + 2\nabla t(x_i) \cdot K(x_i, x_j) \nabla b(x_j).\end{aligned}}$$

For a natural exponential family we have $\eta(\theta) = \theta$, and the prior $\pi(\theta) \propto \exp(-\frac{1}{2}(\theta - \mu) \cdot \Sigma^{-1}(\theta - \mu))$ leads to a generalised posterior

$$\pi_n^D(\theta) \propto \exp\left(-\frac{1}{2}(\theta - \mu_n) \cdot \Sigma_n^{-1}(\theta - \mu_n)\right),$$

where $\Sigma_n^{-1} := \Sigma^{-1} + 2\beta n \Lambda_n$ and $\mu_n := \Sigma_n^{-1}(\Sigma^{-1}\mu - \nu_n)$.

Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B

Changepoints

Proposition 3.1. *If p_θ is given by (5), then*

$$\pi_\omega^{\mathcal{D}_m}(\theta | x_{1:T}) \propto \pi(\theta) \exp(-\omega T[\eta(\theta)^\top \Lambda_T \eta(\theta) + \eta(\theta)^\top \nu_T]),$$

for $\Lambda_T = \frac{1}{T}\sum_{t=1}^T \Lambda(x_t)$, $\nu_T = \frac{2}{T}\sum_{t=1}^T \nu(x_t)$, and

$$\boxed{\begin{aligned}\Lambda(x) &= (\nabla r^\top m m^\top \nabla r)(x),\\ \nu(x) &= \left(\nabla r^\top m m^\top \nabla b + \nabla \cdot (m m^\top \nabla r)\right)(x).\end{aligned}}$$

Taking $\eta(\theta) = \theta$ and choosing a squared exponential prior $\pi(\theta) \propto \exp\left(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\right)$, also makes $\pi_\omega^{\mathcal{D}_m}(\theta | x_{1:T})$ a (truncated) normal of the form

$$\pi_\omega^{\mathcal{D}_m}(\theta | x_{1:T}) \propto \exp\left(-\tfrac{1}{2}(\theta - \mu_T)^\top \Sigma_T^{-1}(\theta - \mu_T)\right),$$

for $\Sigma_T^{-1} = \Sigma^{-1} + 2\omega T \Lambda_T$ and $\mu_T = \Sigma_T\left(\Sigma^{-1}\mu - \omega T \nu_T\right)$.

Altamirano, Briol, & **Knoblauch** (2023); ICML

Gaussian Processes

Proposition 3.1. *Suppose $f \sim \mathcal{GP}(m, k)$ and $\varepsilon \sim \mathcal{N}(0, I_n\sigma^2)$. Then, the RCGP posterior is*

$$p^w(\mathbf{f}|\mathbf{y}, \mathbf{x}) = \mathcal{N}(\mathbf{f}; \mu^R, \Sigma^R),$$

$$\boxed{\begin{aligned}\mu^R &= \mathbf{m} + K\left(K + \sigma^2 J_{\mathbf{w}}\right)^{-1}(\mathbf{y} - \mathbf{m}_{\mathbf{w}}),\\ \Sigma^R &= K\left(K + \sigma^2 J_{\mathbf{w}}\right)^{-1}\sigma^2 J_{\mathbf{w}},\end{aligned}}$$

for $\mathbf{w} = (w(x_1, y_1), \ldots, w(x_n, y_n))^\top$, $\mathbf{m}_{\mathbf{w}} = \mathbf{m} + \sigma^2 \nabla_y \log(\mathbf{w}^2)$ and $J_{\mathbf{w}} = \mathrm{diag}(\frac{\sigma^2}{2}\mathbf{w}^{-2})$. The RCGP's posterior predictive over $f_\star = f(x_\star)$ at $x^\star \in \mathcal{X}$ is

$$\begin{aligned}p^w(f_\star | x_\star, \mathbf{x}, \mathbf{y}) &= \int_{\mathbb{R}^n} p(f_\star | x_\star, \mathbf{f}, \mathbf{x}, \mathbf{y}) p^w(\mathbf{f}|\mathbf{y}, \mathbf{x}) d\mathbf{f}\\ &= \mathcal{N}(f_\star; \mu_\star^R, \Sigma_\star^R),\end{aligned}$$

$$\begin{aligned}\mu_\star^R &= m_\star + \mathbf{k}_\star^\top\left(K + \sigma^2 J_{\mathbf{w}}\right)^{-1}(\mathbf{y} - \mathbf{m}_{\mathbf{w}}),\\ \Sigma_\star^R &= k_{\star\star} - \mathbf{k}_\star^\top (K + \sigma^2 J_{\mathbf{w}})^{-1}\mathbf{k}_\star.\end{aligned}$$

Altamirano, Briol, & **Knoblauch** (2024); ICML spotlight

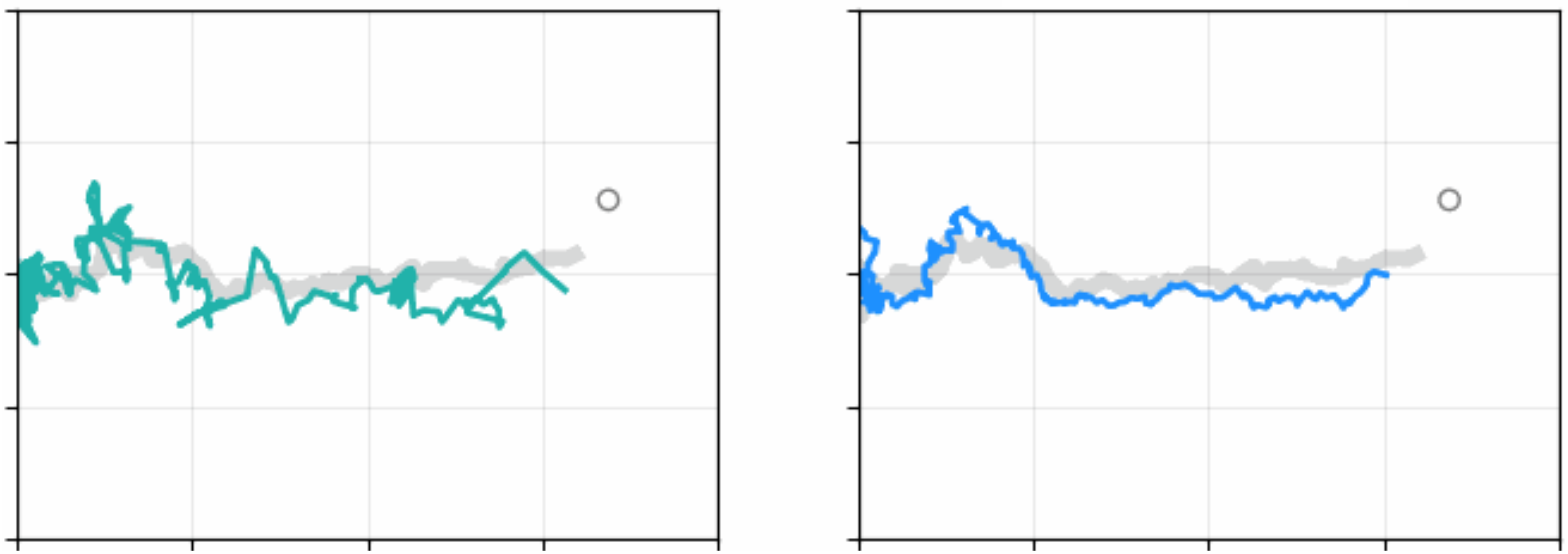
~~(A1)~~, (A2), ~~(A3)~~

2

L for Robustness + Tractability: Kalman Filter

$\Longrightarrow$ Closed form Post-Bayesian posteriors: feasible for on-line problems!

Standard Kalman Filter | **Robust version**



Duran-Martin, Altamirano, Shestopaloff, Sanchez-Betancourt, **Knoblauch**, Briol, & Murphy (2024); ICML

~~(A1)~~, (A2), ~~(A3)~~

Summary: State of the Art in Post-Bayesian ML

2

(A1)	model well-specified
(A2)	prior well-specified
(A3)	computationally feasible

2 **State of the Art**

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})}$$

model misspecification + computation

~~(A1)~~ (A2) ~~(A3)~~

Schmon, Cannon, & **Knoblauch** (2020); AABI
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Dellaporta, **Knoblauch**, Damoulas, & Briol (2022); AISTATS (best paper award)
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)
Duran-Martin, Altamirano, Shestopaloff, Sanchez-Betancourt, **Knoblauch**, Briol, & Murphy (2024); ICML

Q: Can we design losses L that are **both robust and tractable** ?

A: Yes! Stein Discrepancies. (And weighted likelihoods.)

~~(A1)~~, ~~(A2)~~, ~~(A3)~~ Tackled Assumptions

Part of the Talk 3

The Future of Post-Bayesian ML

1 **Foundations**

$\pi_n^{(\lambda)}(\theta \mid x_{1:n})$

$\pi_n^{L}(\theta \mid x_{1:n})$

model misspecification

~~(A1)~~ (A2) (A3)

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, **Knoblauch***, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

2 **State of the Art**

$\pi_n^{L}(\theta \mid x_{1:n})$

model misspecification + computation

~~(A1)~~ (A2) ~~(A3)~~

Schmon, Cannon, & **Knoblauch** (2020); AABI
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Dellaporta, **Knoblauch**, Damoulas, & Briol (2022); AISTATS (best paper award)
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** (2024); ICML (spotlight)
Duran-Martin, Altamirano, Shestopaloff, Sanchez-Betancourt, **Knoblauch**, Briol, & Murphy (2024); ICML

3 **The Future**

$q_n^*(\theta)$

model misspecification + prior misspecification + computation

~~(A1)~~ ~~(A2)~~ ~~(A3)~~

Husain & **Knoblauch** (2022); ALT
Knoblauch, Jewson, & Damoulas (2022); JMLR
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Wild, Sejdinovic, & **Knoblauch** (2024); forthcoming
Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2024); NeurIPS (oral)

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

3

The Future of Post-Bayesian ML

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible

3 **The Future**

$\boxed{q_n^*(\theta)}$

model misspecification +
prior misspecification +
computation

~~(A1)~~ ~~(A2)~~ ~~(A3)~~

Husain & **Knoblauch** (2022); ALT
Knoblauch, Jewson, & Damoulas (2022); JMLR
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Wild, Sejdinovic, & **Knoblauch** (2024); forthcoming
Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2024); NeurIPS (oral)

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

3

The Future of Post-Bayesian ML

Optimisation-centric posteriors / Generalised Variational Inference ~~(A1)~~, ~~(A2)~~, ~~(A3)~~

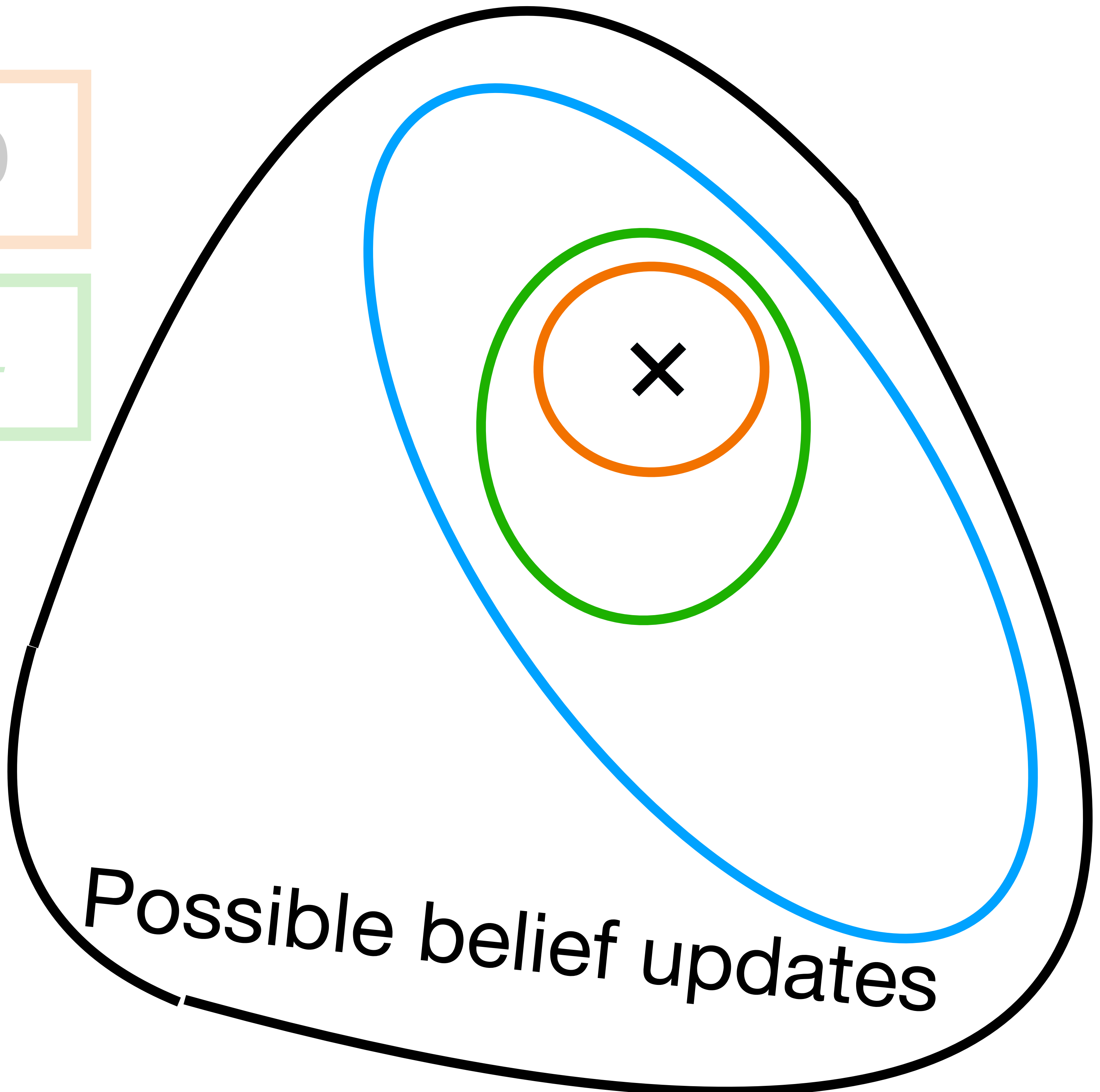
$$q_n^*(\theta) = \arg\min_{q \in \mathcal{Q}} \left\{ \underbrace{\mathcal{L}_{\mathsf{L},\, D}(q)}_{} \right\}; \quad \mathcal{Q} \subseteq \mathcal{P}(\Theta)$$
$$= \int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta \; + \; \mathrm{D}\big(q, \pi\big)$$

$$p(x_{1:n} \mid \theta) \longrightarrow p(x_{1:n} \mid \theta)^{\lambda}, \;\; \lambda > 0$$

$$p(x_{1:n} \mid \theta) \longrightarrow \exp\{-\mathsf{L}(x_{1:n}, \theta)\}, \;\; \text{loss } \mathsf{L}$$

$$\begin{array}{ccc} \mathrm{KL} & \longrightarrow & \mathrm{D} \\ \mathcal{P}(\Theta) & \longrightarrow & \mathcal{Q} \end{array}$$

(A1) model well-specified
(A2) prior well-specified
(A3) computationally feasible



Gibbs/Generalised/ Pseudo Posterior ~~(A1)~~, (A2), ~~(A3)~~

$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}$$

Power/Fractional/ Cold Posterior ~~(A1)~~, (A2), (A3)

$$\pi_n^{(\lambda)}(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta) d\theta}$$

Bayes' Posterior (A1), (A2), (A3)

$$\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}$$

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

3

The Future of Post-Bayesian ML

$$\boxed{\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}} = \arg \min_{q \in \mathcal{P}(\Theta)} \left\{ \int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta + \mathrm{KL}(q, \pi) \right\}$$

~~(A3)~~ by optimising over a set $\mathcal{Q} \subseteq \mathcal{P}(\Theta)$

~~(A1)~~ by using robust loss L

~~(A2)~~ by using robust regulariser D

$$\boxed{q_n^*(\theta) = \arg\min_{q \in \mathcal{Q}} \left\{ \underbrace{\int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta + \mathrm{D}(q, \pi)}_{=: \mathcal{L}_{\mathsf{L},\mathrm{D}}(q)} \right\}}$$

Optimisation-centric posteriors / Generalised Variational Inference (GVI)

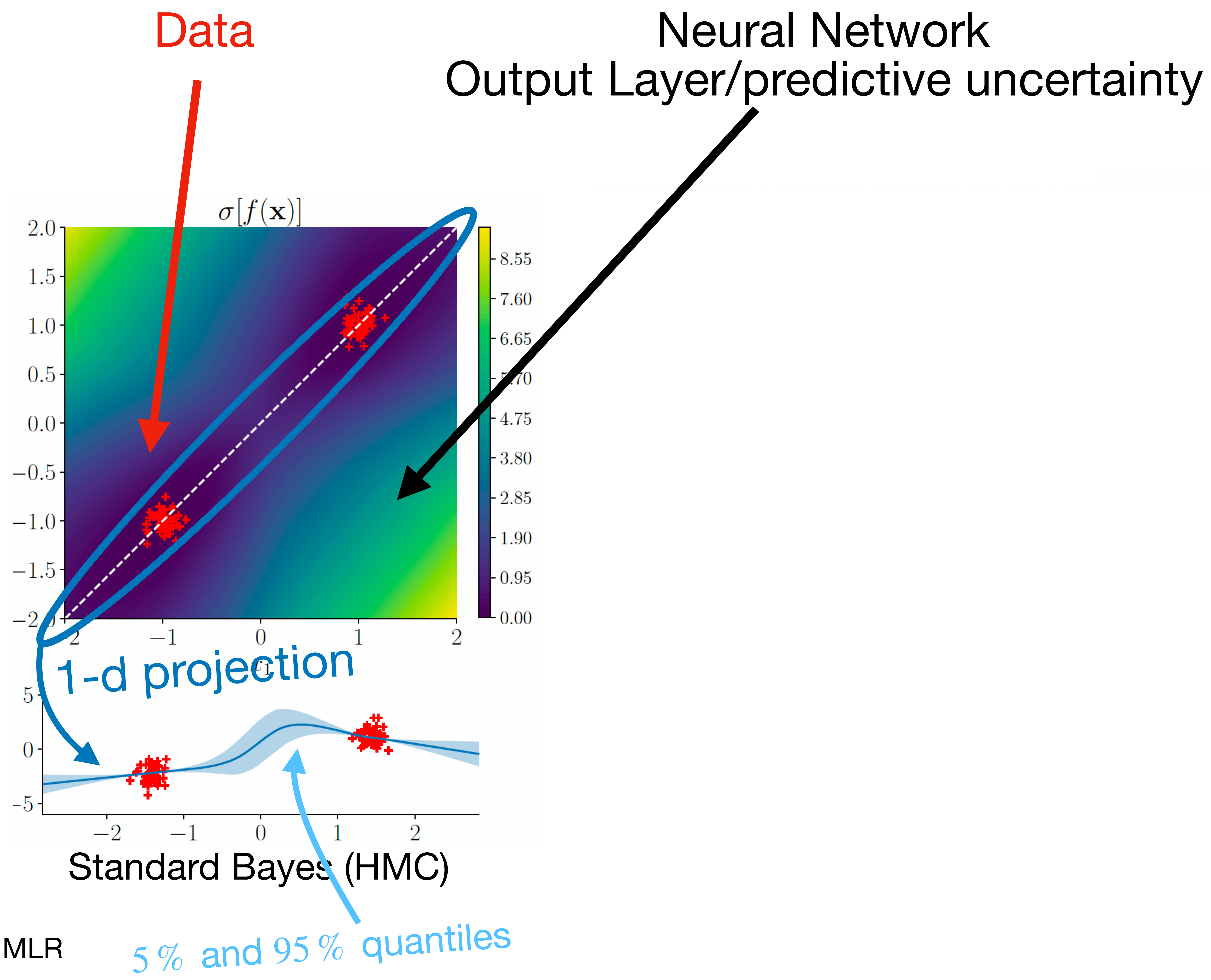
~~(A1)~~	model well-specified
~~(A2)~~	prior well-specified
~~(A3)~~	computationally feasible

Husain & **Knoblauch** (2022); ALT
Knoblauch, Jewson, & Damoulas (2022); JMLR
Wild, Sejdinovic, & **Knoblauch** (2024); forthcoming
Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2024); NeurIPS (oral)

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

③

The Future of Post-Bayesian ML

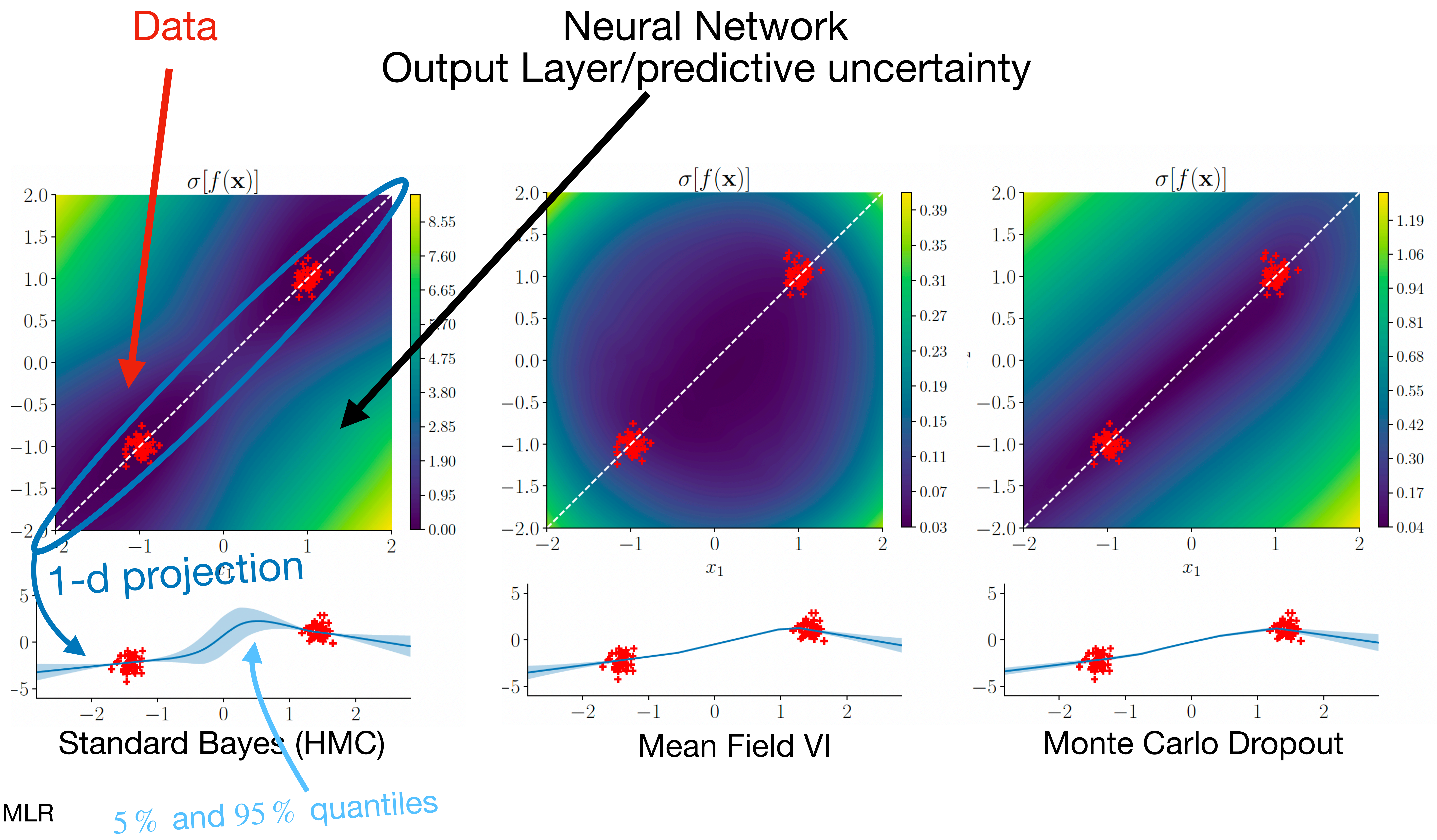


Knoblauch, Jewson, & Damoulas (2022); JMLR
Wild, Hu, & Sejdinovic, NeurIPS (2022)

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

The Future of Post-Bayesian ML

3



Knoblauch, Jewson, & Damoulas (2022); JMLR
Wild, Hu, & Sejdinovic, NeurIPS (2022)

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

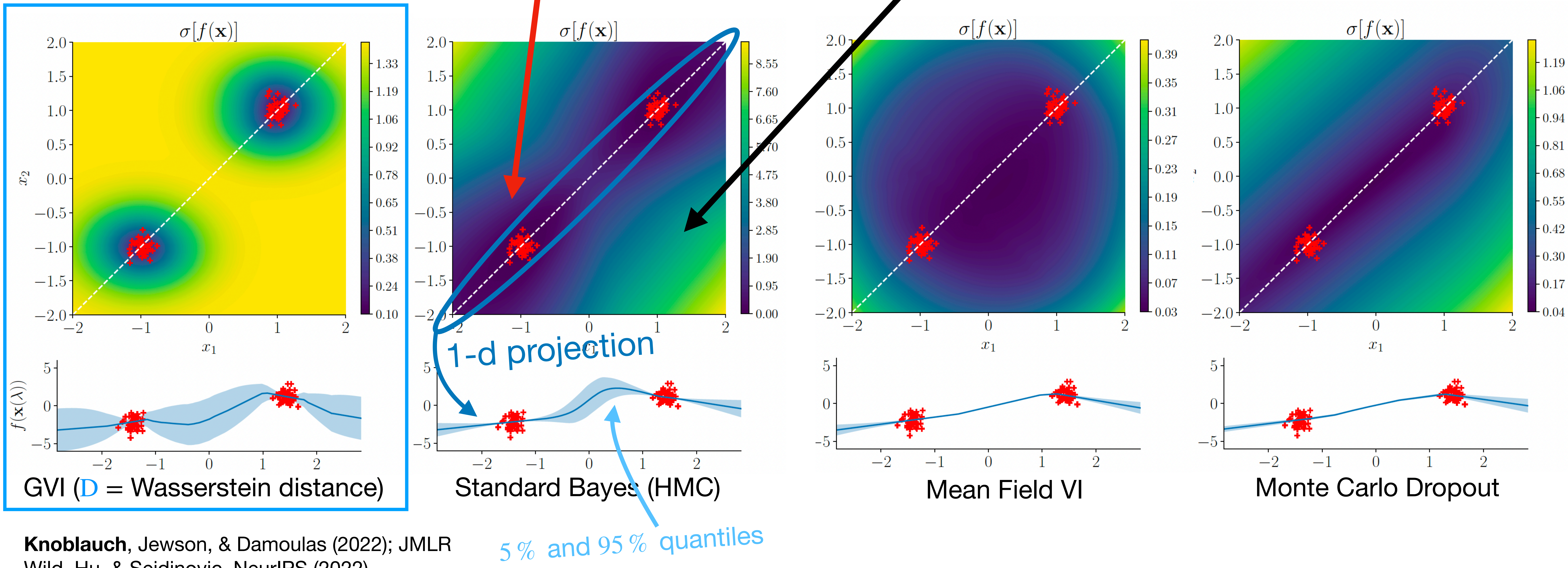
3

The Future of Post-Bayesian ML

D = Wasserstein Distance Improves Uncertainty

Data

Neural Network
Output Layer/predictive uncertainty



Knoblauch, Jewson, & Damoulas (2022); JMLR
Wild, Hu, & Sejdinovic, NeurIPS (2022)

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

The Future of Post-Bayesian ML

$$\boxed{q_n^*(\theta) \quad = \underset{q \in \mathcal{Q}}{\arg\min} \left\{ \quad \int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta \; + \; \mathsf{D}\big(q, \pi\big) \right\}}$$

New Problem: Unclear how to compute $q_n^*(\theta)$ unless $\mathcal{Q}$ is a parameterised set of distributions

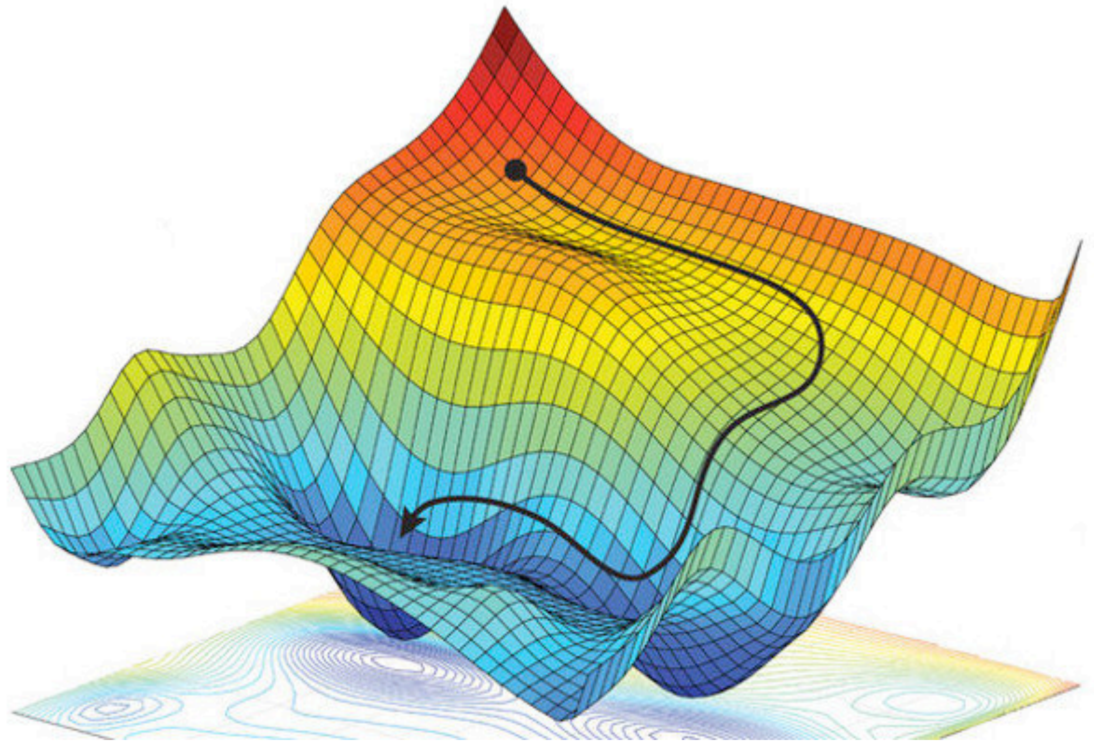
Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2023); NeurIPS Oral

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

③

The Future of Post-Bayesian ML

$$\boxed{q_n^*(\theta) \quad = \arg\min_{q \in \mathcal{Q}} \left\{ \int L(x_{1:n}, \theta)\, q(\theta)\, d\theta + D\big(q, \pi\big) \right\}}$$



New Problem: Unclear how to compute $q_n^*(\theta)$ unless $\mathcal{Q}$ is a parameterised set of distributions

New Algorithmic Solution: Infinite-dimensional Gradient Descent on $\mathcal{P}_2(\Theta)$

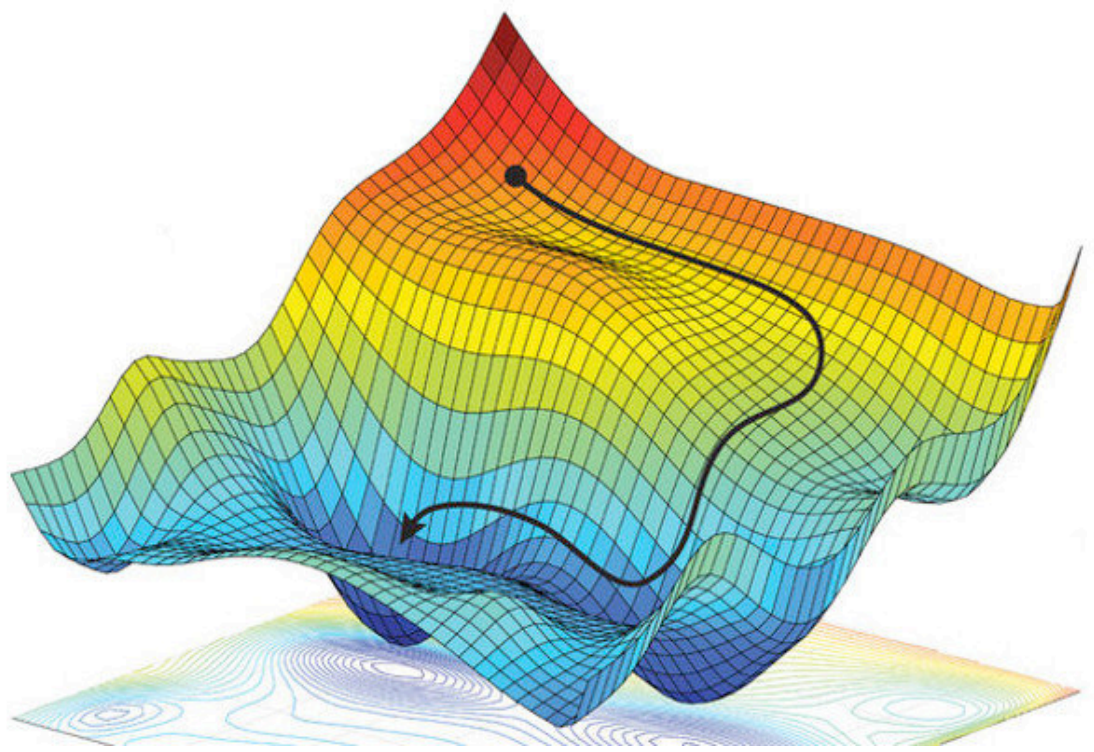
Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2023); NeurIPS Oral

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

3

The Future of Post-Bayesian ML

$$\boxed{q_n^*(\theta) = \arg\min_{q \in \mathcal{Q}} \left\{ \int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta + \mathsf{D}\big(q, \pi\big) \right\}}$$



New Problem: Unclear how to compute $q_n^*(\theta)$ unless $\mathcal{Q}$ is a parameterised set of distributions

New Algorithmic Solution: Infinite-dimensional Gradient Descent on $\mathcal{P}_2(\Theta)$

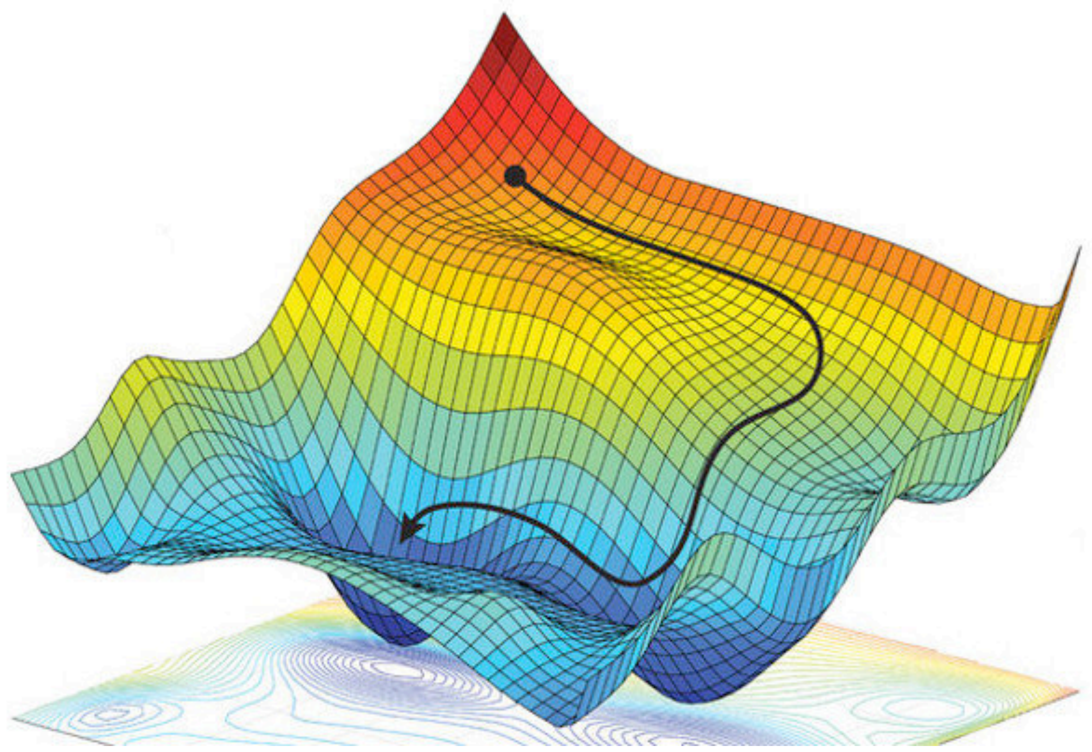
Implementation: Wasserstein Gradient Flow on $q \mapsto \int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta + \mathsf{D}\big(q, \pi\big)$

Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2023); NeurIPS Oral

$\cancel{(A1)}$, $\cancel{(A2)}$, $\cancel{(A3)}$

3

The Future of Post-Bayesian ML



$$\boxed{q_n^*(\theta) \quad = \underset{q \in \mathcal{Q}}{\arg\min} \left\{ \int \mathrm{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta + \mathrm{D}\big(q, \pi\big) \right\}}$$

New Problem: Unclear how to compute $q_n^*(\theta)$ unless $\mathcal{Q}$ is a parameterised set of distributions

New Algorithmic Solution: Infinite-dimensional Gradient Descent on $\mathcal{P}_2(\Theta)$

Implementation: Wasserstein Gradient Flow on $q \mapsto \int \mathrm{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta + \mathrm{D}\big(q, \pi\big)$

Result: Sampling algorithm that doesn't need access to its target $q_n^*(\theta)$

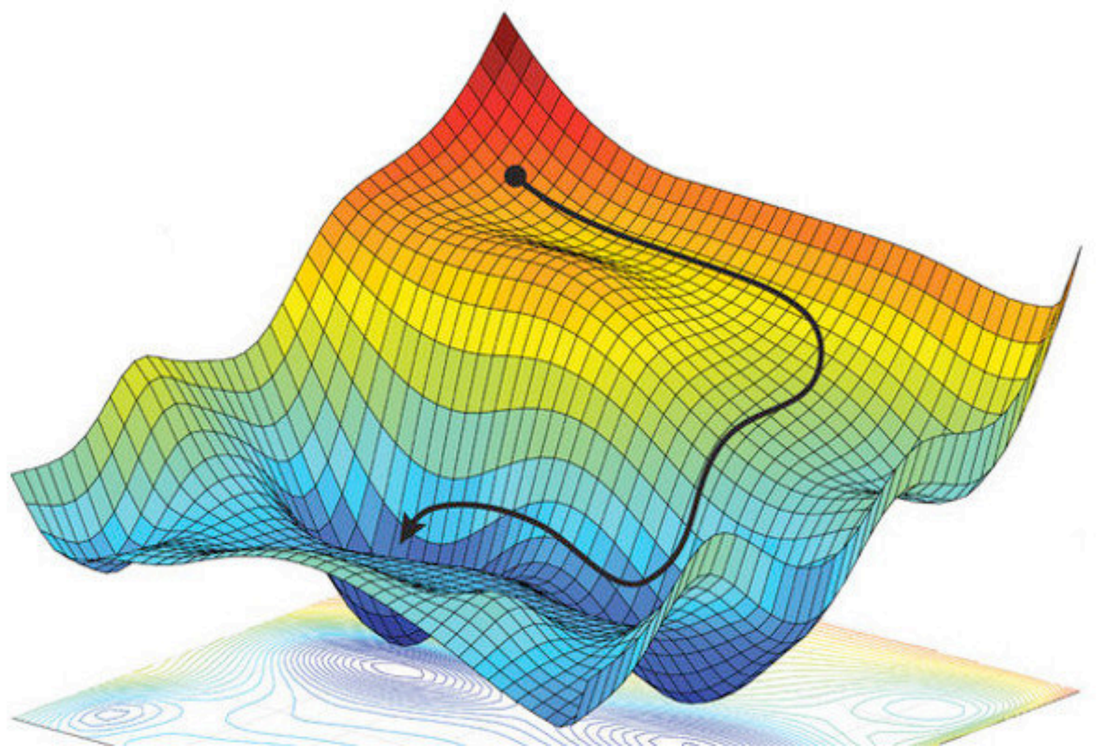
First of its kind.

Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2023); NeurIPS Oral

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

③

The Future of Post-Bayesian ML



$$\boxed{q_n^*(\theta) = \arg\min_{q \in \mathcal{Q}} \left\{ \int \mathrm{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta + \mathrm{D}(q, \pi) \right\}}$$

New Problem: Unclear how to compute $q_n^*(\theta)$ unless $\mathcal{Q}$ is a parameterised set of distributions

New Algorithmic Solution: Infinite-dimensional Gradient Descent on $\mathcal{P}_2(\Theta)$

Implementation: Wasserstein Gradient Flow on $q \mapsto \int \mathrm{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta + \mathrm{D}(q, \pi)$

Result: Sampling algorithm that doesn't need access to its target $q_n^*(\theta)$

Bonus: recovers various deep learning algorithms + emblematic for why Post-Bayesian ML matters

First of its kind.

Will revisit this in a second.

Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2023); NeurIPS Oral

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

3

The Future of Post-Bayesian ML

$$\boxed{q_n^*(\theta) \quad = \arg\min_{q \in \mathcal{Q}} \left\{ \int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta + \mathsf{D}\big(q, \pi\big) \right\}}$$

A Selection of unresolved challenges:

- How should we choose **D**?

Knoblauch, Jewson, & Damoulas (2022); JMLR

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

3

The Future of Post-Bayesian ML

$$\boxed{q_n^*(\theta) \quad = \arg\min_{q \in \mathcal{Q}} \left\{ \int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta \; + \; \mathsf{D}\big(q, \pi\big) \right\}}$$

A Selection of unresolved challenges:

- How should we choose **D**?
- Can we characterise all divergence-based **L** that provide conjugate posteriors?

Knoblauch, Jewson, & Damoulas (2022); JMLR

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

3

The Future of Post-Bayesian ML

$$\boxed{q_n^*(\theta) \quad = \underset{q \in \mathcal{Q}}{\arg\min} \left\{ \quad \int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta \; + \; \mathsf{D}\big(q, \pi\big) \right\}}$$

A Selection of unresolved challenges:

- How should we choose **D**?
- Can we characterise all divergence-based **L** that provide conjugate posteriors?
- How should we **formalise** robustness to prior misspecification?

Knoblauch, Jewson, & Damoulas (2022); JMLR

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

The Future of Post-Bayesian ML

$$\boxed{q_n^*(\theta) \quad = \arg\min_{q \in \mathcal{Q}} \left\{ \quad \int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta \; + \; \mathsf{D}\left(q, \pi\right) \right\}}$$

A Selection of unresolved challenges:

- How should we choose **D**?
- Can we characterise all divergence-based **L** that provide conjugate posteriors?
- How should we **formalise** robustness to prior misspecification?
- How should we **choose between** different Post-Bayesian methods?

Knoblauch, Jewson, & Damoulas (2022); JMLR

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

③

The Future of Post-Bayesian ML

$$\boxed{q_n^*(\theta) \quad = \arg\min_{q \in \mathcal{Q}} \left\{ \quad \int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta \, + \, \mathsf{D}\big(q, \pi\big) \right\}}$$

A Selection of unresolved challenges:

- How should we choose **D**?
- Can we characterise all divergence-based **L** that provide conjugate posteriors?
- How should we **formalise** robustness to prior misspecification?
- How should we **choose between** different Post-Bayesian methods?
- (basic) asymptotics

Knoblauch, Jewson, & Damoulas (2022); JMLR

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

3

The Future of Post-Bayesian ML

$$\boxed{q_n^*(\theta) \quad = \arg\min_{q \in \mathcal{Q}} \left\{ \quad \int \mathrm{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta \,+\, \mathrm{D}\big(q, \pi\big) \right\}}$$

A Selection of unresolved challenges:

- How should we choose **D**?
- Can we characterise all divergence-based **L** that provide conjugate posteriors?
- How should we **formalise** robustness to prior misspecification?
- How should we **choose between** different Post-Bayesian methods?
- (basic) asymptotics
- What is the overlap / difference with Martingale Posteriors?
- …

… But rather than expanding on technical challenges, let me end with a morality tale

Knoblauch, Jewson, & Damoulas (2022); JMLR

$\cancel{(A1)}$, $\cancel{(A2)}$, $\cancel{(A3)}$

3

Morality Tale: Why Post-Bayesian ML matters

$$\boxed{q_n^*(\theta) \quad = \underset{q \in \mathcal{Q}}{\arg\min} \left\{ \quad \lambda \cdot \int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta \; + \; \mathrm{D}\big(q, \pi\big) \right\}}$$

Objective: $q \mapsto \mathbb{E}_{\theta \sim q}\left[-\log p(x_{1:n} \mid \theta)^{\lambda}\right] \; + \; \mathrm{KL}(q, \pi)$

Target: **Cold Posterior** $(\lambda \gg 1)$

/ Bayes Posterior $(\lambda = 1)$

Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2023); NeurIPS Oral

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

3

Morality Tale: Why Post-Bayesian ML matters

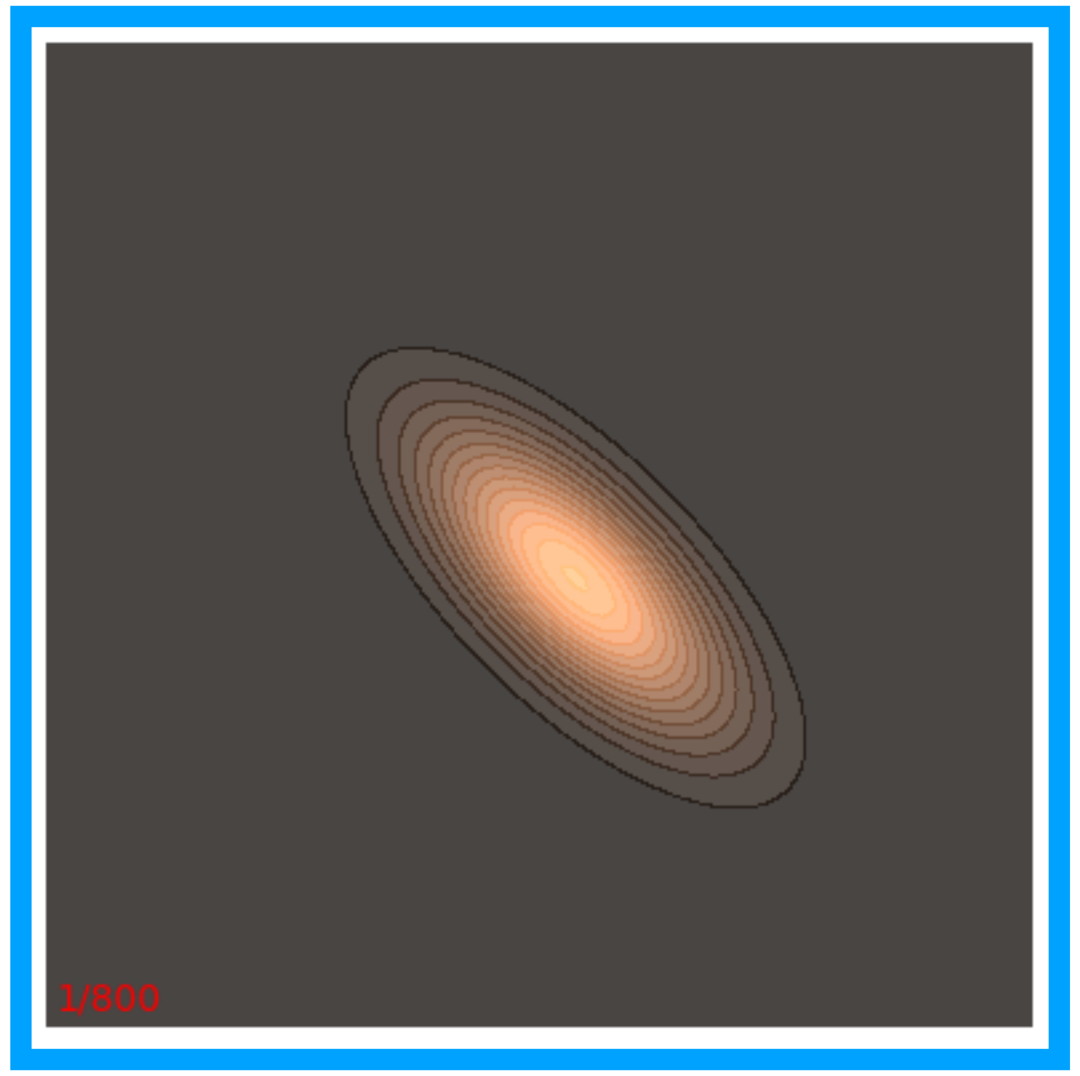
$$\boxed{q_n^*(\theta) \quad = \arg\min_{q \in \mathcal{Q}} \left\{ \quad \lambda \cdot \int \mathrm{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta \; + \; \mathrm{D}\big(q, \pi\big) \right\}}$$

Objective: $\quad q \mapsto \mathbb{E}_{\theta \sim q}\left[-\log p(x_{1:n} \mid \theta)^{\lambda}\right] \; + \; \mathrm{KL}(q, \pi)$

Target: **Cold Posterior** $(\lambda \gg 1)$

/ Bayes Posterior $(\lambda = 1)$

Wasserstein Gradient Flow = Langevin Diffusion



Converges to well-defined density

$$q_n^*(\theta) = \pi_n^{(\lambda)}(\theta \mid x_{1:n})$$

Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2023); NeurIPS Oral

~~(A1)~~, ~~(A2)~~, ~~(A3)~~ 3

Morality Tale: Why Post-Bayesian ML matters

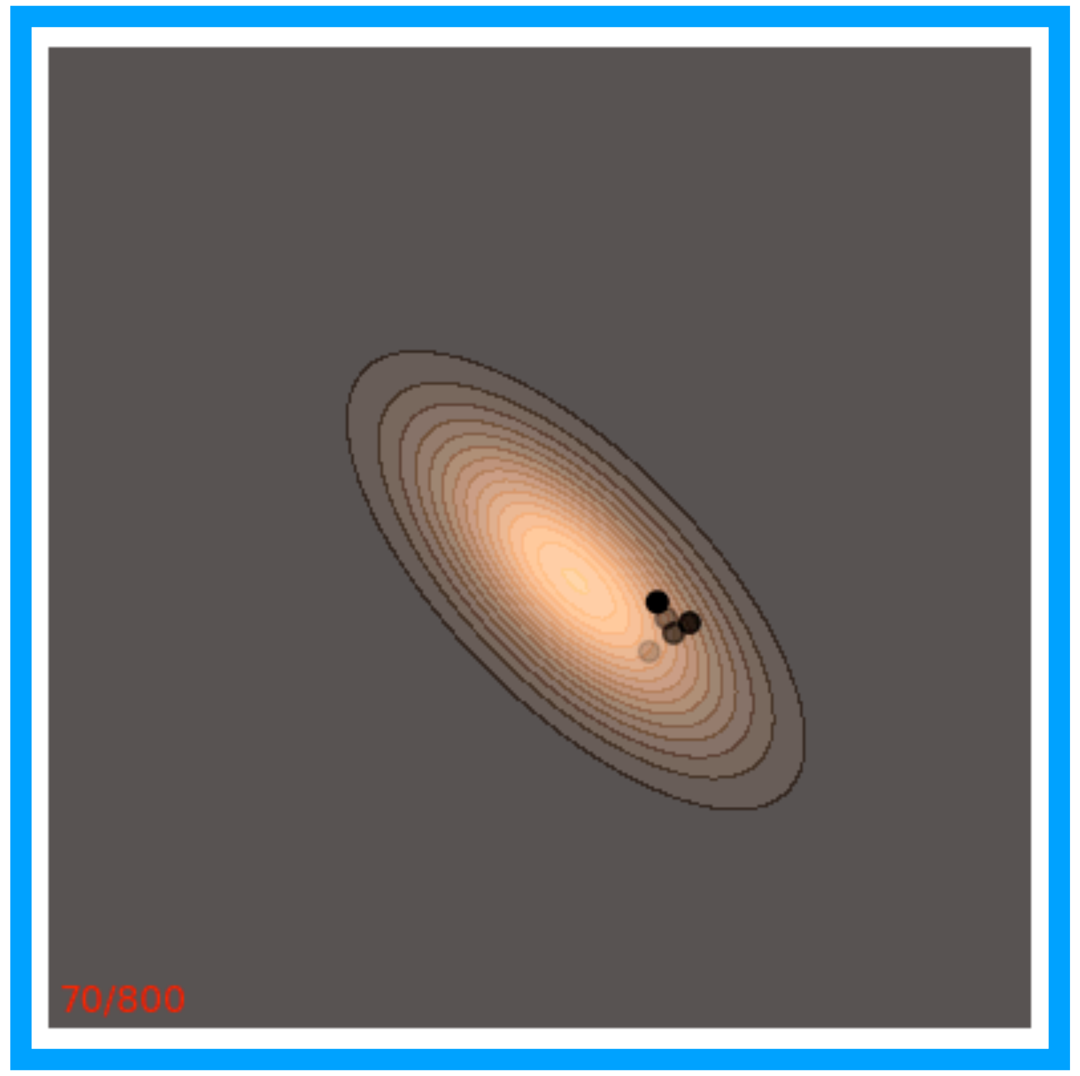
$$\boxed{q_n^*(\theta) \quad = \underset{q \in \mathcal{Q}}{\arg\min} \left\{ \quad \lambda \cdot \int \mathsf{L}(x_{1:n}, \theta)\, q(\theta)\, d\theta \, + \, \mathsf{D}\big(q, \pi\big) \right\}}$$

Objective: $q \mapsto \mathbb{E}_{\theta \sim q}\left[-\log p(x_{1:n} \mid \theta)^{\lambda}\right] + \mathrm{KL}(q, \pi) \xrightarrow{\lambda \to \infty} q \mapsto \mathbb{E}_{\theta \sim q}\left[-\log p(x_{1:n} \mid \theta)\right]$

Target: **Cold Posterior** $(\lambda \gg 1)$
/ Bayes Posterior $(\lambda = 1)$

Deep Ensemble (DE) $(\lambda \to \infty)$

Wasserstein Gradient Flow = Langevin Diffusion

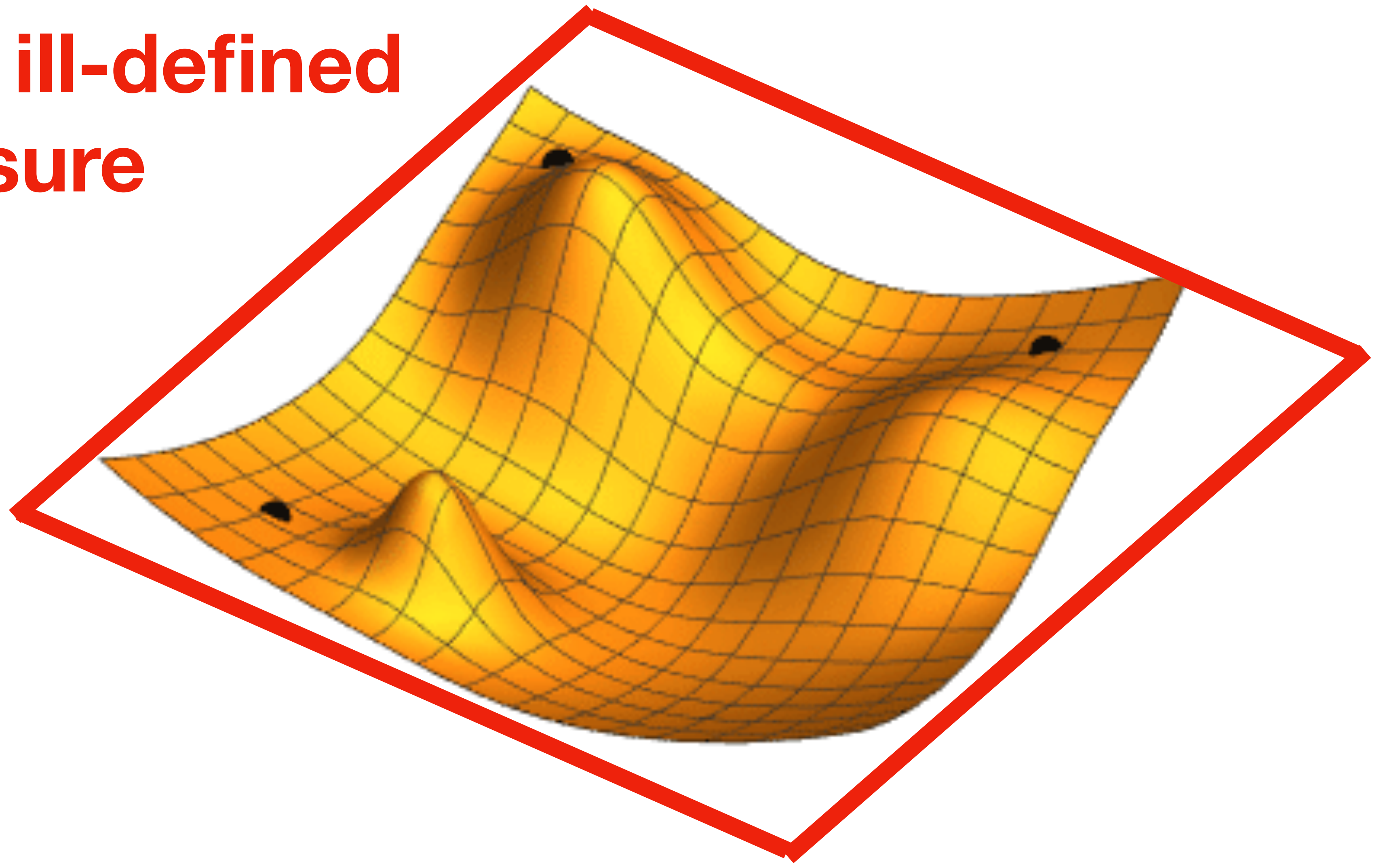


Converges to well-defined density

$$q_n^*(\theta) = \pi_n^{(\lambda)}(\theta \mid x_{1:n})$$

Wasserstein Gradient Flow = DE

Converges to ill-defined discrete measure



Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2023); NeurIPS Oral

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

3

Morality Tale: Why Post-Bayesian ML matters

Claim: ‘Deep Ensembles = Bayesian Inference’

[…] Deep ensembles (Lakshminarayanan et al., 2017) are not a competing approach to Bayesian inference, but […] a compelling mechanism for Bayesian marginalization.

Published by a group specialised in Bayesian ML (2020) @ NeurIPS (cited > 650 times according to Google scholar)

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

Morality Tale: Why Post-Bayesian ML matters

Claim: ‘Deep Ensembles = Bayesian Inference’

[…] Deep ensembles (Lakshminarayanan et al., 2017) are not a competing approach to Bayesian inference, but […] a compelling mechanism for Bayesian marginalization.

Published by a group specialised in Bayesian ML (2020) @ NeurIPS (cited > 650 times according to Google scholar)

Unfortunately, as we just saw this is not correct. ✗

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

3

Conclusion: Why Post-Bayesian ML matters

I. In practice, orthodox Bayesian ML has already been abandoned

(Bayes posterior: prior regulariser, densities ;

Deep Ensembles: no prior regulariser, discrete measures)

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

3

Conclusion: Why Post-Bayesian ML matters

I. In practice, orthodox Bayesian ML has already been abandoned
(Bayes posterior: prior regulariser, densities ;
Deep Ensembles: no prior regulariser, discrete measures)

II. But as a field, we often don't pay enough attention to the ramifications, and lag behind in developing coherent Post-Bayesian formalisms and a mathematical grammar to talk about them.

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

3

Conclusion: Why Post-Bayesian ML matters

I. In practice, orthodox Bayesian ML has already been abandoned
(Bayes posterior: prior regulariser, densities ;
Deep Ensembles: no prior regulariser, discrete measures)

II. But as a field, we often don't pay enough attention to the ramifications, and lag behind in developing coherent Post-Bayesian formalisms and a mathematical grammar to talk about them.

III. This in turns leads to incorrect claims and conclusions.
('Deep Ensembles are Bayesian')

References: Post-Bayesian ML

(A1)	model well-specified
(A2)	prior well-specified
(A3)	computationally feasible

1 Foundations

$\pi_n^{(\lambda)}(\theta \mid x_{1:n})$

$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$

model missp

~~(A1)~~ (A2)

@ ISBA 2024

Knoblauch & Damoulas (2018); ICML
Knoblauch, Jewson, & Damoulas (2018); NeurIPS
Frazier*, **Knoblauch***, & Drovandi (2024); preprint
McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

2 State of the Art

$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$

model misspecification + computation

~~(A1)~~ (A2)

@ ICML 2024

Schmon, Cannon, & **Knoblauch** (2020); AABI
Matsubara, **Knoblauch**, Briol, & Oates (2022); JRSS-B
Dellaporta, **Knoblauch**, Damoulas, & Briol (2022); AISTATS (best paper award)
Altamirano, Briol, & **Knoblauch** (2023); ICML
Altamirano, Briol, & **Knoblauch** **(2024); ICML (spotlight)**
Duran-Martin, **Altamirano**, Shestopaloff, Sanchez-Betancourt, **Knoblauch**, Briol, & Murphy **(2024); ICML**

3 The Future

$q_n^*(\theta)$

model missp
prior misspe
compu

~~(A1)~~ ~~(A2)~~ ~~(A3)~~

Husain & **Knoblauch** (2022); ALT
Knoblauch, Jewson, & Damoulas (2022); JMLR
Matsubara, **Knoblauch**, Briol, & Oates (2023); JASA
Wild, Sejdinovic, & **Knoblauch** (2024); forthcoming
Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2023); NeurIPS (oral)

Foundations

Mathematical Foundations

1930: **DeFinetti's Representation Theorem** [cf. Hewitt & Savage (1955), Diaconis & Freedman (1984, 1987)]

For all $k \leq n$ and all permutations σ, $p(x_1, x_2, \ldots, x_k) \stackrel{\forall \sigma}{=} p(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(k)},)$

$$\Updownarrow$$

There is a parameter space Θ and $\pi(\theta)$ s.t. $p(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(k)},) \stackrel{\forall \sigma}{=} \int \prod_{i=1}^{k} p(x_i \mid \theta)\pi(\theta)d\theta$

$\Longrightarrow$ If data exchangeable, there MUST be

1) model $p(\cdot \mid \theta)$

2) prior $\pi(\theta)$

that represent the data through a Bayesian approach.

Justification for Bayesianism

Foundations

Mathematical Foundations

1930: **DeFinetti's Representation Theorem** [cf. Hewitt & Savage (1955), Diaconis & Freedman (1984, 1987)]

For all $k \leq n$ and all permutations σ, $p(x_1, x_2, \ldots, x_k) \overset{\forall \sigma}{=} p(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(k)},)$

$$\Updownarrow$$

There is a parameter space Θ and $\pi(\theta)$ s.t. $p(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(k)},) \overset{\forall \sigma}{=} \int \prod_{i=1}^{k} p(x_i \mid \theta)\pi(\theta)d\theta$

$\Longrightarrow$ Problem: Does NOT tell you what $\pi(\theta)$ and $p(\cdot \mid \theta)$ are!

1) π generally depends on $p(\cdot \mid \theta)$ and n

2) θ generally ∞-dimensional

[e.g. $p(x_i \mid \theta) = \theta(x_i)$ for θ a probability density on x_i]

Not how we practice Bayesianism

Foundations

Mathematical Foundations

1954: **Savage Axioms** connects Bayes' Theorem to **Decision Theory**

Actions $\mathscr{A} = \{a : \text{States} \longrightarrow \text{Consequences}\}$ $\quad a_2 \lesssim a_1 \iff a_1$ preferred to a_2

$\lesssim$ satisfies Savage Axioms on $\mathscr{A}$

$$\Downarrow$$

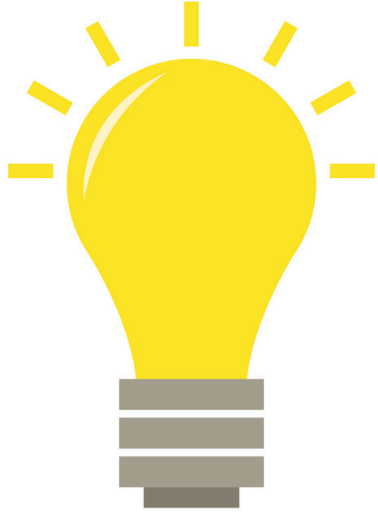
$\exists$ utility function u : Consequences $\rightarrow \mathbb{R}$ and π on States s.t.

$$\forall a_1, a_2 \in \mathscr{A} : a_1 \lesssim a_2 \iff \int u[a_1(s)]\, \pi(s)\, ds \le \int u[a_2(s)]\, \pi(s)\, ds$$

$$\forall a_1, a_2 \in \mathscr{A} : a_1 \lesssim a_2 \text{ given } x_{1:n} \iff \int u[a_1(s)]\, \pi_n(s \mid x_{1:n})\, ds \le \int u[a_2(s)]\, \pi_n(s \mid x_{1:n})\, ds$$

Foundations

Mathematical Foundations



$\Longrightarrow$ Prior $\pi(s)$ = beliefs IMPLIED by rational agent's preferences

$\Longrightarrow$ Bayes' Posterior $\pi_n(s \mid x_{1:n})$ = rational agent's belief update given data

$\lesssim$ satisfies Savage Axioms on $\mathscr{A}$

$$\Downarrow$$

$\exists$ utility function u : Consequences $\rightarrow \mathbb{R}$ and π on States s.t.

$$\forall a_1, a_2 \in \mathscr{A} : a_1 \lesssim a_2 \iff \int u[a_1(s)]\, \pi(s)\, ds \leq \int u[a_2(s)]\, \pi(s)\, ds$$

$$\forall a_1, a_2 \in \mathscr{A} : a_1 \lesssim a_2 \text{ given } x_{1:n} \iff \int u[a_1(s)]\, \pi_n(s \mid x_{1:n})\, ds \leq \int u[a_2(s)]\, \pi_n(s \mid x_{1:n})\, ds$$

Foundations

Mathematical Foundations

Problem: Prior $\pi(s)$ is defined on States (not parameters θ!)

$\implies$ parameter space Θ = relevant State $\iff x_{1:n} \sim p_{\theta^*}(x_{1:n})$ for some $\theta^* \in \Theta$

$\implies$ Does NOT tell you what $\pi(\theta)$ and $p(\cdot \mid \theta)$ are!

$\lesssim$ satisfies Savage Axioms on $\mathscr{A}$

$$\Downarrow$$

$\exists$ utility function u : Consequences $\to \mathbb{R}$ and π on States s.t.

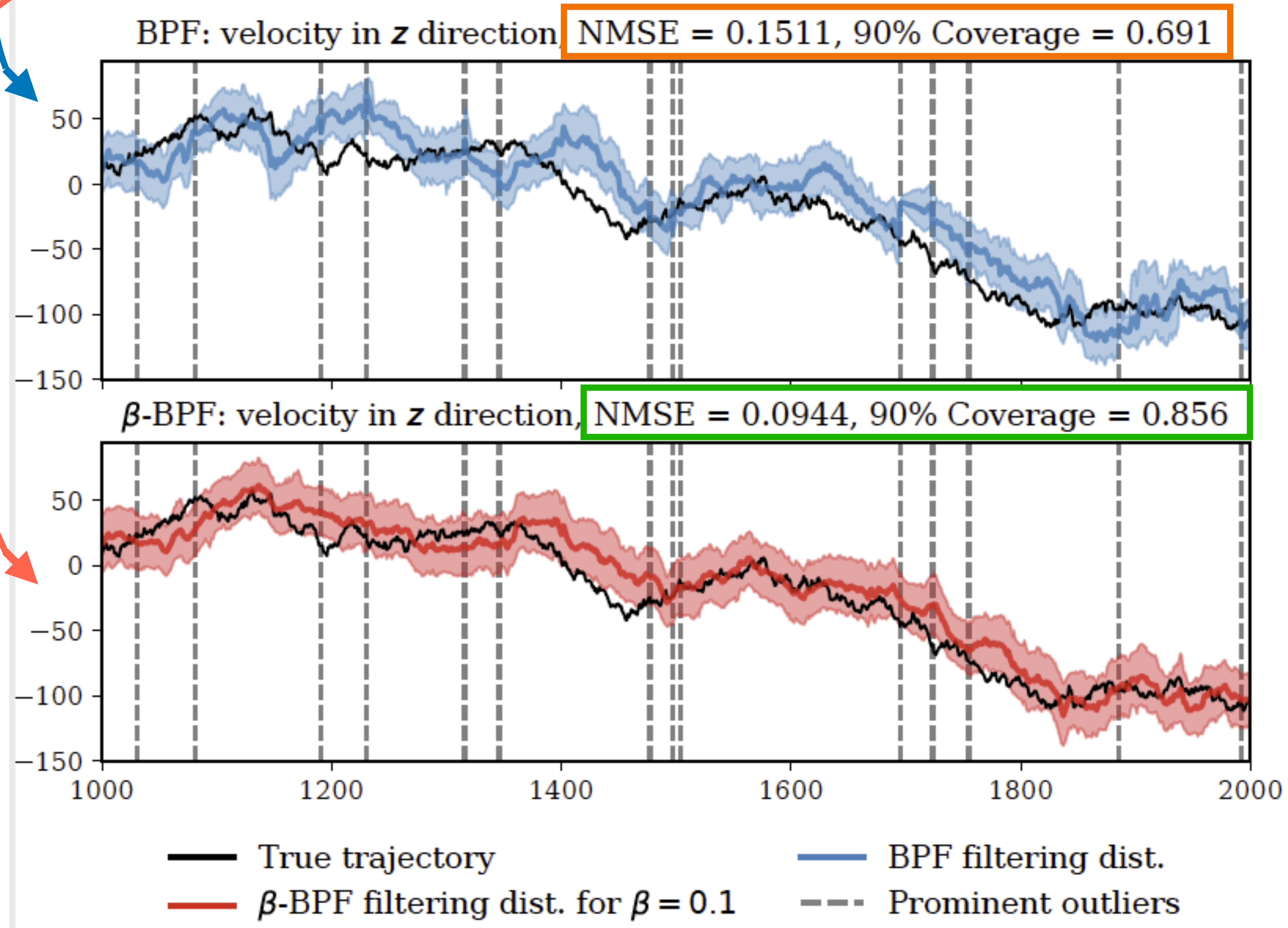
$$\forall a_1, a_2 \in \mathscr{A} : a_1 \lesssim a_2 \iff \int u[a_1(s)]\, \pi(s)\, ds \leq \int u[a_2(s)]\, \pi(s)\, ds$$

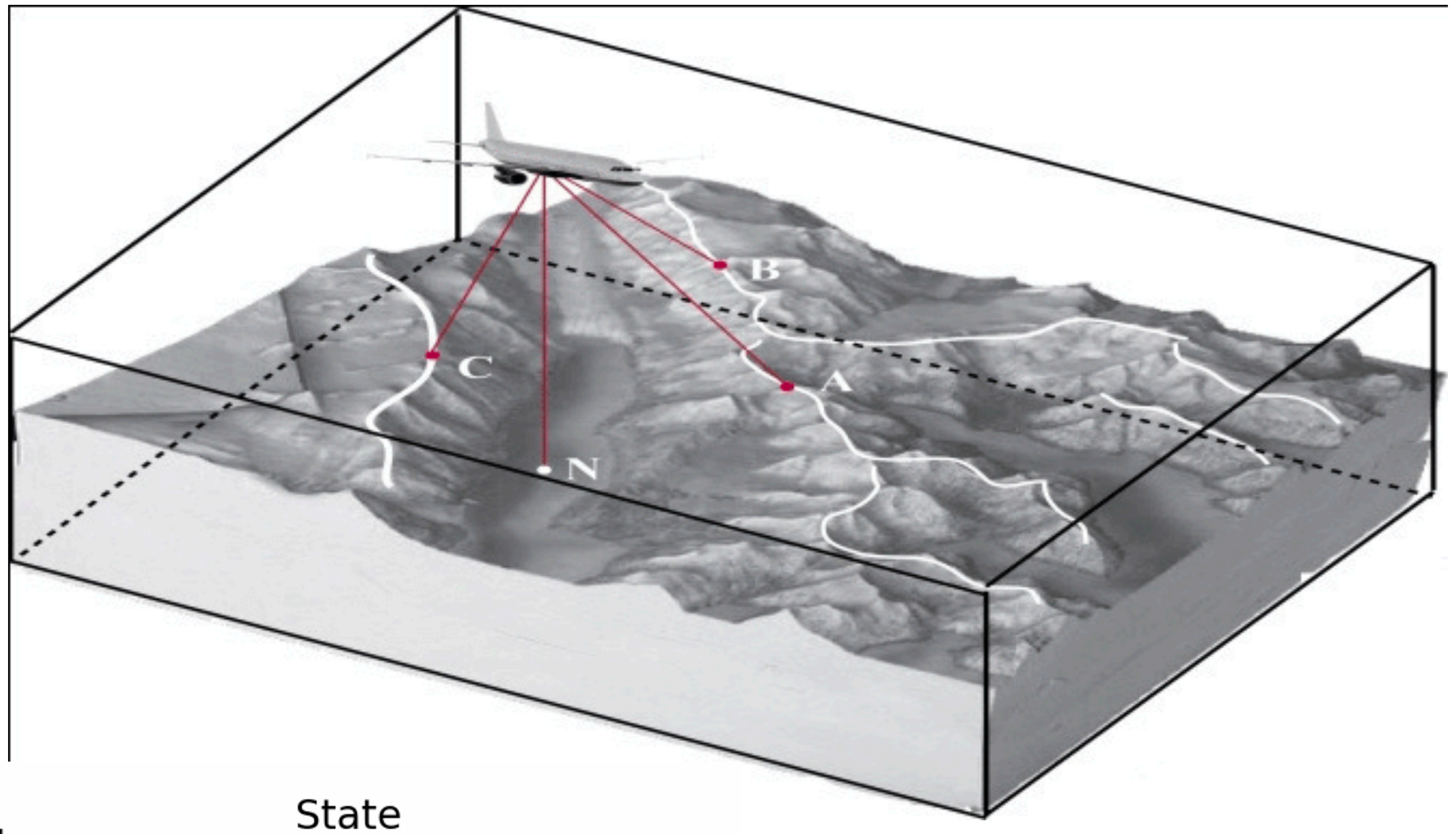
$$\forall a_1, a_2 \in \mathscr{A} : a_1 \lesssim a_2 \text{ given } x_{1:n} \iff \int u[a_1(s)]\, \pi_n(s \mid x_{1:n})\, ds \leq \int u[a_2(s)]\, \pi_n(s \mid x_{1:n})\, ds$$

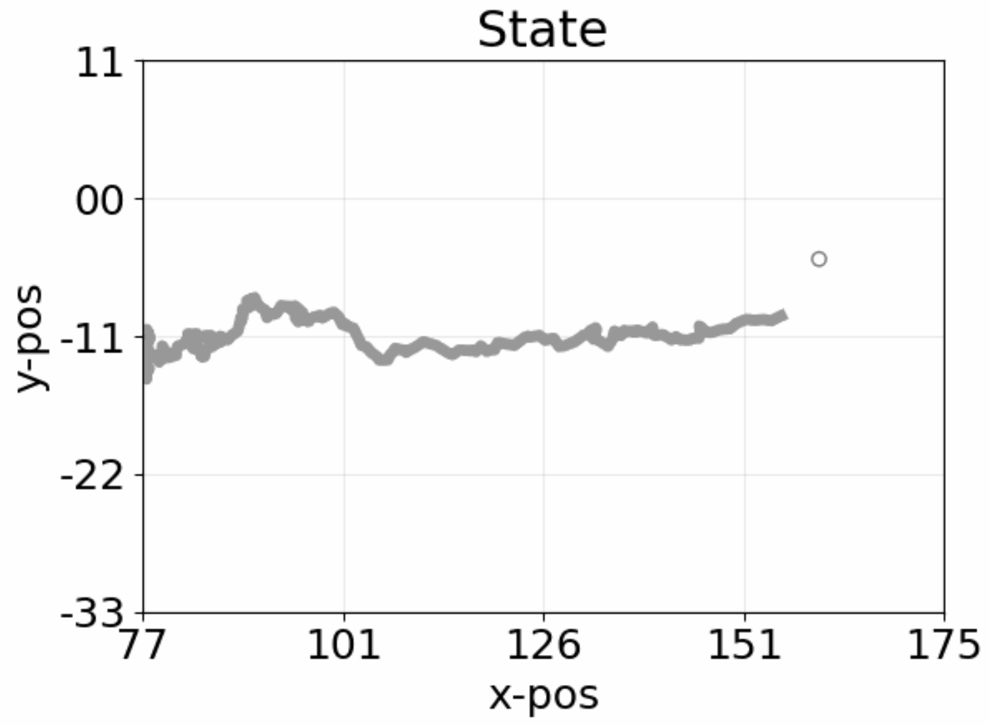
Post-Bayesian ML: Success Stories

Standard Kalman Filter for Terrain Aided Navigation (Drone over Terrain Map)

Robustified version







Based on work of Boustati, Akyildiz, Damoulas, & Johansen, NeurIPS (2020)

~~(A1)~~, (A2), (A3)

1

Q1: Can tuning λ improve Robustness?

Question: What is the predictively optimal λ?

Posterior predictive $= p_n^{\lambda}(z) = \int p(z \mid \theta)\, \pi_n^{(\lambda)}(\theta \mid x_{1:n})\, d\theta$

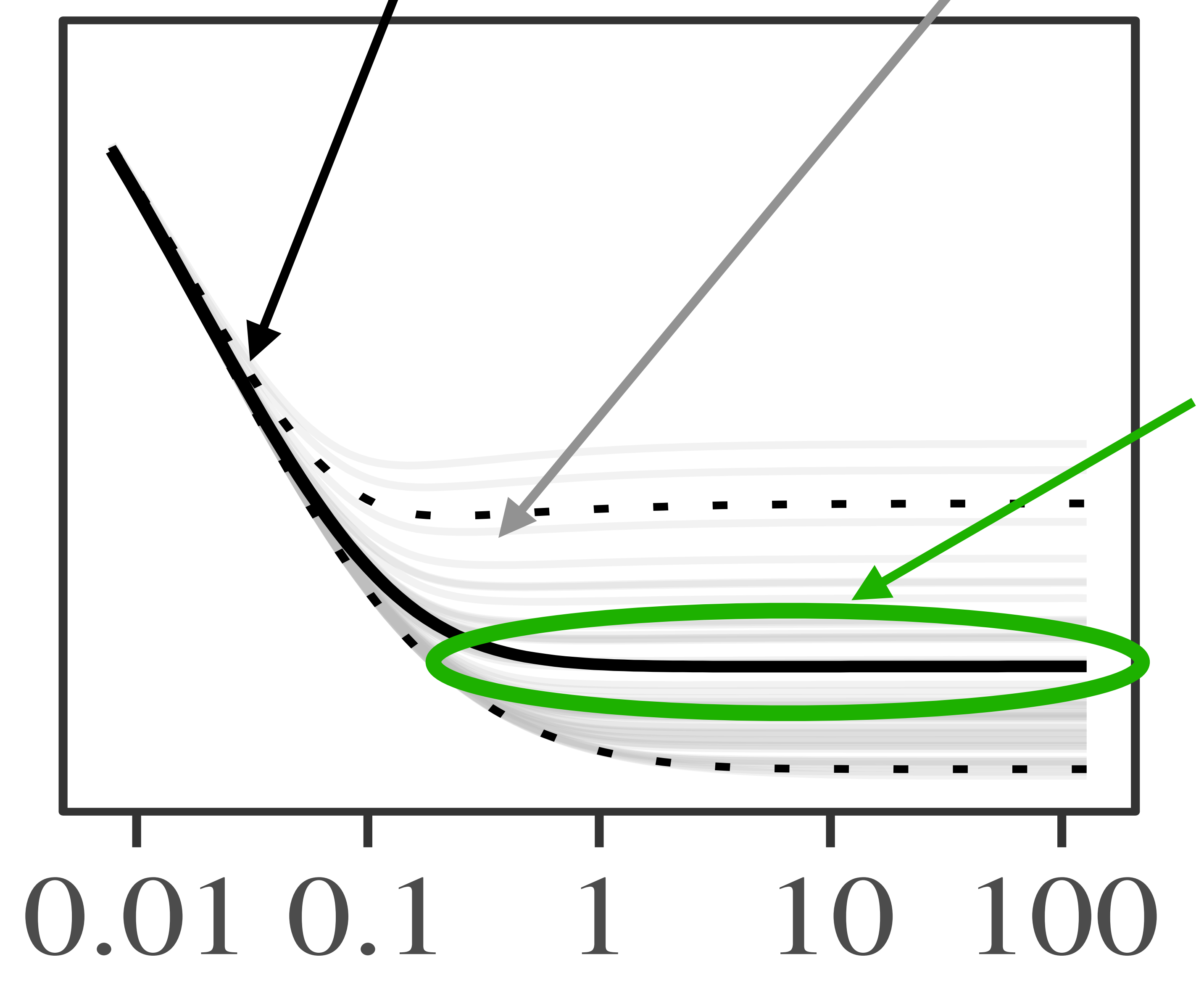
Predictively optimal λ: $\lambda^* = \mathrm{argmin}_{\lambda>0}\, \mathrm{D_{TV}}\left(q, p_n^{\lambda}\right)$

Data-generating density: $x_{1:n} \sim q(x_{1:n})$

Curves in expectation

Curves for individual data set

$\mathrm{D_{TV}}\left(q, p_n^{\lambda}\right)$



Flat region: all λ equally good for prediction

$\pi_n^{(0)} = \pi \xleftarrow{0 \leftarrow \lambda} \lambda \xrightarrow{\lambda \to \infty} \pi_n^{(\infty)} =$ Dirac at MLE

Theorem: these curves will always look that way.

McLatchie, Fong, Frazier, & **Knoblauch** (2024); forthcoming

What λ leads to Robustness?

Findings:
(1) Ill-defined problem: minimiser λ^* doesn't exist
(2) Flat region: infinitely many choices yield (nearly) same predictive
(3) Predictively, there is no advantage over MLE / point estimators

Q: Why does this happen?

p^* = oracle predictive induced by θ^* $\qquad$ $\theta^* = \text{argmin}_{\theta \in \Theta} \lim_{n \to \infty} n^{-1} - \log p(x_{1:n} \mid \theta)$

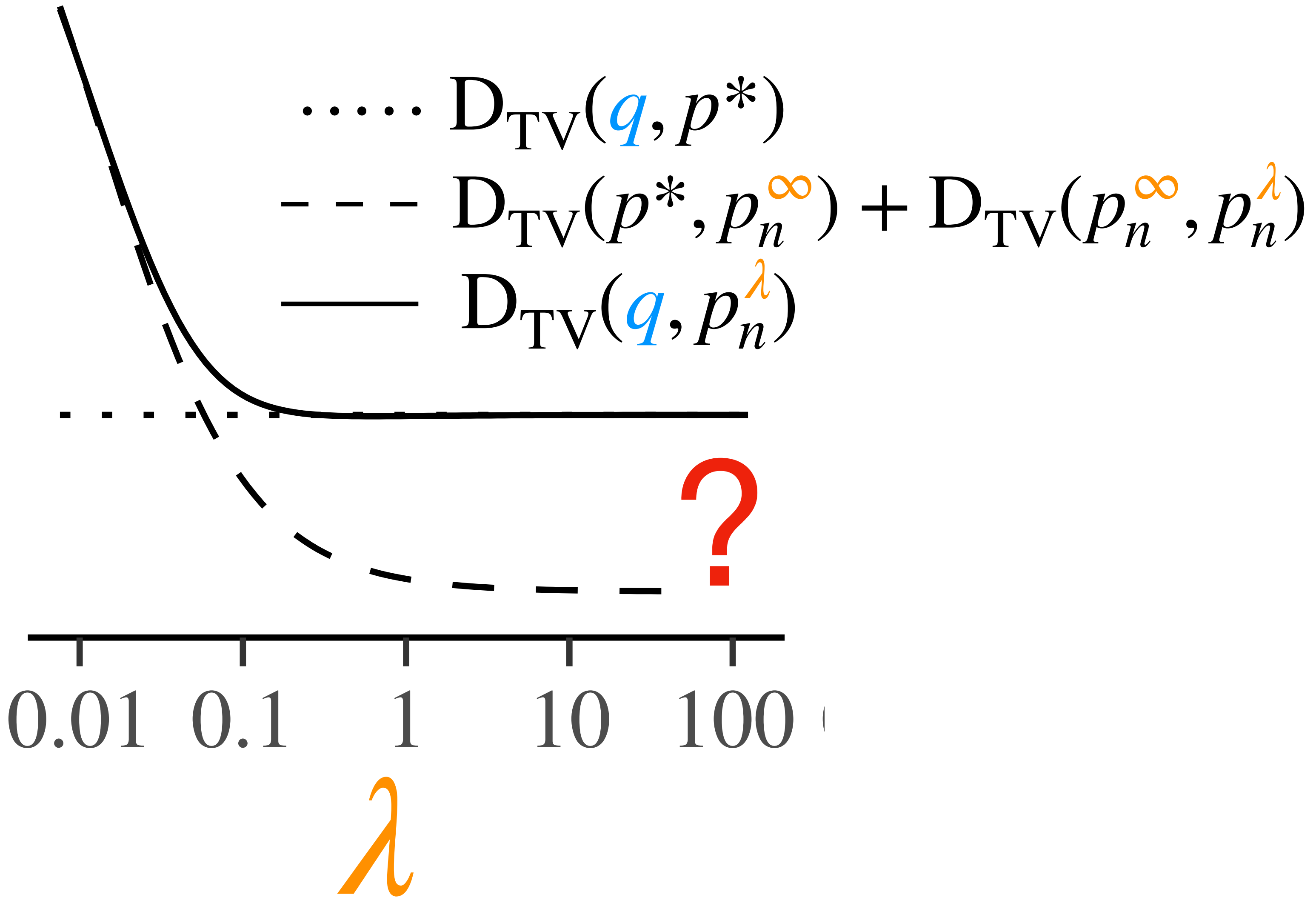
p_n^∞ = MLE predictive induced by $\hat{\theta}_n$ $\qquad$ $\hat{\theta}_n = \text{argmin}_{\theta \in \Theta} - \log p(x_{1:n} \mid \theta)$

$$\mathrm{D}_{\mathrm{TV}}(q, p_n^\lambda) \underset{\text{w.h.p}}{\leq} \underbrace{\mathrm{D}_{\mathrm{TV}}(q, p^*)}_{\text{(irreducible error)} \; = \text{ constant}} + \underbrace{\underbrace{\mathrm{D}_{\mathrm{TV}}(p^*, p_n^\infty)}_{\text{(error of MLE/point estimator)} \; \lesssim \nu_n \text{ if MLE converges at rate } \nu_n} + \underbrace{\mathrm{D}_{\mathrm{TV}}\left(p_n^\infty, p_n^\lambda\right)}_{\text{(difference MLE vs } p_n^\lambda) \; \lesssim \varepsilon_n \text{ if } \pi_n^{(\lambda)} \text{ concentrates at rate } \varepsilon_n}}$$

In practice / experimentally:
- goes to 0 MUCH faster than $\nu_n + \varepsilon_n$
- even works WITHOUT concentration and consistency

McLatchie, Fong, Frazier, & **K.** (2024); forthcoming

What λ leads to Robustness?



Similar for CV and $D = KL$ ✓

Exponentially fast ?

$$\boxed{D_{TV}(q, p_n^\lambda)} \underset{\text{w.h.p}}{\leq} \underbrace{D_{TV}(q, p^*)}_{= \text{ constant}}^{\text{(irreducible error)}} + \underbrace{\overbrace{D_{TV}(p^*, p_n^\infty)}^{\text{(error of MLE/point estimator)}}_{\lesssim \nu_n \text{ if MLE converges at rate } \nu_n} + \overbrace{D_{TV}\left(p_n^\infty, p_n^\lambda\right)}^{\text{(difference MLE vs } p_n^\lambda)}_{\lesssim \varepsilon_n \text{ if } \pi_n^{(\lambda)} \text{ concentrates at rate } \varepsilon_n}}$$

? In practice / experimentally:
- goes to 0 MUCH faster than $\nu_n + \varepsilon_n$
- even works WITHOUT concentration and consistency

McLatchie, Fong, Frazier, & **K.** (2024); forthcoming

What λ leads to Robustness?

$$\boxed{\pi_n^{(\lambda)}(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta)^{\lambda} \cdot \pi(\theta) d\theta} = \operatorname{argmin}_{q \in \mathscr{P}(\Theta)} \left\{ \lambda \cdot \int -\log p(x_{1:n}, \theta)\, q(\theta)\, d\theta + \mathrm{KL}\big(q, \pi\big) \right\}}$$

$\boxed{0 < \lambda \ll 1}$ $\longleftrightarrow$ $\boxed{\lambda \gg 1}$

How to square with our finding that λ is largely irrelevant?

Grünwald (2012); ALT
Holmes & Walker (2017); Biometrika
Miller & Dunson (2018); JRSS-B
Bhattacharya, Pati, & Yang (2019); Ann. Statist.
…

Wenzel et al. (2020); ICML
Adlam et al. (2020); preprint
Noci et al. (2021); NeurIPS
Aitchison (2021); ICLR
…

essential	Parameter uncertainty	incidental
incidental	Predictive uncertainty	essential
n>d	Model complexity	d<<n
good	Prior quality	bad
theory	Evidence	Experimental

Prior quality should be VERY important

McLatchie, Fong, Frazier, & **K.** (2024); forthcoming

What L leads to Robustness?

$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}$$

Ghosh & Basu (2016); AISM
K., Jewson, & Damoulas (2018); NeurIPS
Boustati, Akyildiz, Damoulas, & Johansen (2020); NeurIPS
K., Jewson, & Damoulas (2022); JMLR
Frazier*, **K.***, & Drovandi (2024); preprint

Q2: Which discrepancies $\mathrm{D}(q, p(\cdot \mid \theta))$ should we construct $\mathsf{L}(x_{1:n}, \theta)$ from?

Initial Work:

$$\mathrm{D}^{\beta}(q, p(\cdot \mid \theta)) = \underbrace{\int p(z \mid \theta)^{1+\beta}\, dz}_{\substack{\text{Generally intractable...} \\ \text{(and hard to approximate)} \\ \text{Sometimes: tractable as } A(\theta,\beta)}} - \frac{1+\beta}{\beta} \underbrace{\int p(z \mid \theta)^{\beta} q(z)\, dz}_{\approx \frac{1}{n}\sum_{i=1}^{n} p(x_i \mid \theta)^{\beta}} + \underbrace{\frac{1}{\beta}\int q^{1+\beta}(z)\, dz}_{\text{Independent of } \theta}$$

(β-Divergence)

Issues:

Intractability
Numerical instability
Scale dependence

scale

instability

$$\mathsf{L}(x_{1:n}, \theta) = n \cdot A(\theta, \beta) + \frac{\beta + 1}{\beta} \sum_{i=1}^{n} p(x_i \mid \theta)^{\beta} \xrightarrow{\beta \downarrow 0} \sum_{i=1}^{n} -\log p(x_i \mid \theta)$$

What L leads to Robustness?

$$\boxed{\pi_n^{\mathrm{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathrm{L}(x_{1:n},\theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathrm{L}(x_{1:n},\theta)\} \cdot \pi(\theta) d\theta}}$$

Futami, Sato & Sugiyama (2018); AISTATS
Cherrief-Abdellatif & Alquier (2020); AABI
Pacchiardi, Khoo, & Dutta (2021); preprint
Frazier*, **K.***, & Drovandi (2024); preprint

Q2: Which discrepancies $\mathrm{D}(q, p(\cdot \mid \theta))$ should we construct $\mathrm{L}(x_{1:n}, \theta)$ from?

Other Proposals:

Bias!

$$\mathrm{D}^{\gamma}(q, p(\cdot \mid \theta)) = \log \underbrace{\int p(z \mid \theta)^{1+\gamma}\, dz}_{\text{Generally intractable...}} - \frac{1+\gamma}{\gamma} \log \underbrace{\int p(z \mid \theta)^{\gamma} q(z)\, dz}_{\approx \frac{1}{n}\sum_{i=1}^{n} p(x_i \mid \theta)^{\gamma}} + \frac{1}{\gamma} \log \underbrace{\int q^{1+\gamma}(z)\, dz}_{\text{Independent of } \theta}$$

(γ-Divergence)

Like β-divergence

Include Hellinger Divergence as well!

(and hard to approximate)

$$\mathrm{D}_k^2(q, p(\cdot \mid \theta)) = \underbrace{\mathbb{E}_{x\sim q(x), x'\sim q(x')}\left[k(x,x')\right]}_{\text{Independent of } \theta} - 2\underbrace{\mathbb{E}_{x\sim q(x), x'\sim p(x'|\theta)}\left[k(x,x')\right]}_{\text{Intractable}} + \underbrace{\mathbb{E}_{x\sim p(x|\theta), x'\sim p(x'|\theta)}\left[k(x,x')\right]}_{\text{Intractable}}$$

(MMD^2)

Intractable Integrals

Not defined for conditional models

(But useful for
intractable/simulation models)

(but easy to approximate!)

(but easy to approximate!)

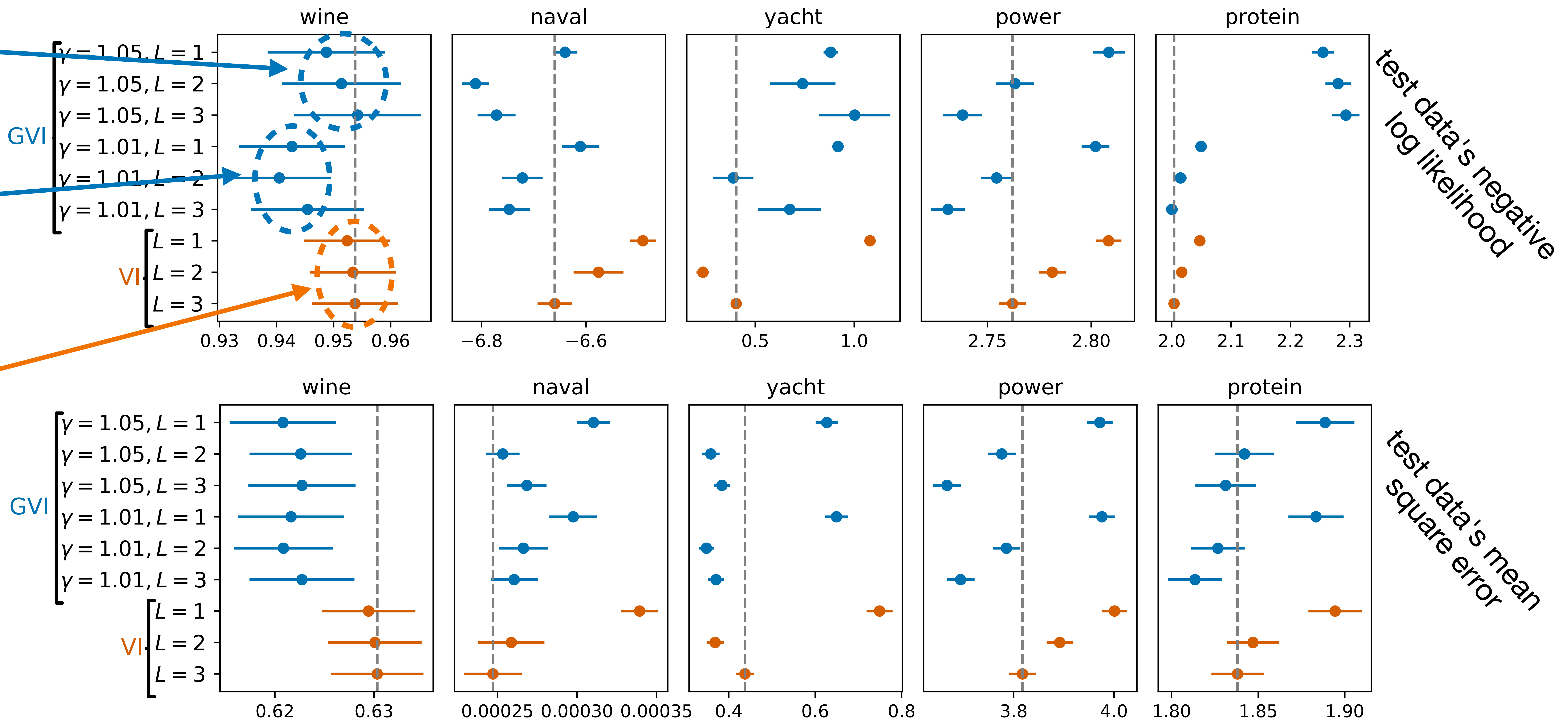
What L leads to Robustness?

Example 2: Deep Gaussian Processes (γ-Divergence)

$L(x_{1:n} \mid \theta) \approx \mathrm{D}_\gamma(p_0 \| p_\theta)$
(γ-divergence [$\gamma = 0.05$])

$L(x_{1:n} \mid \theta) \approx \mathrm{D}_\gamma(p_0 \| p_\theta)$
(γ-divergence [$\gamma = 0.01$])

$-\log p(x_{1:n} \mid \theta) \approx \mathrm{KLD}(p_0 \| p_\theta)$

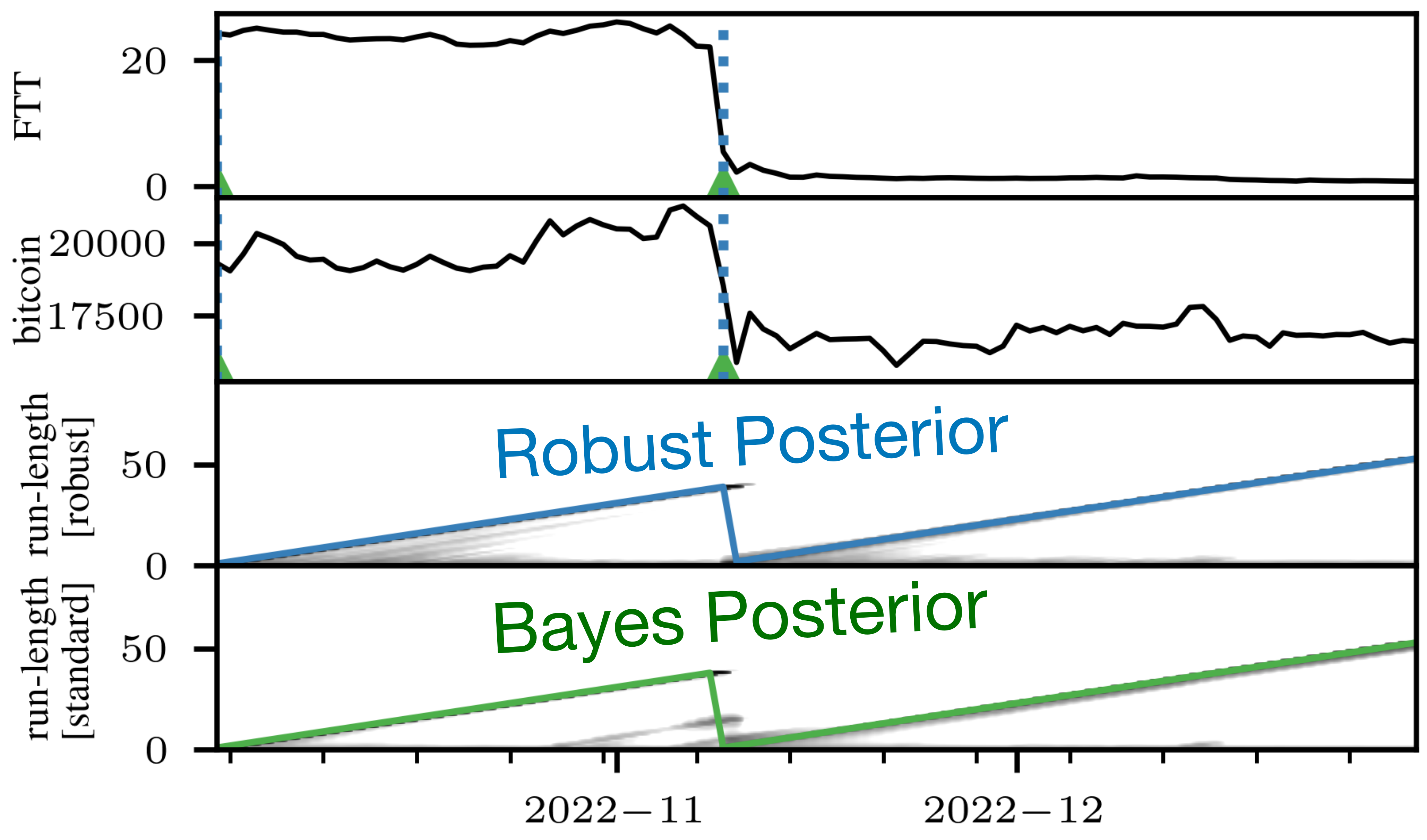


K., Jewson, & Damoulas (2022); JMLR

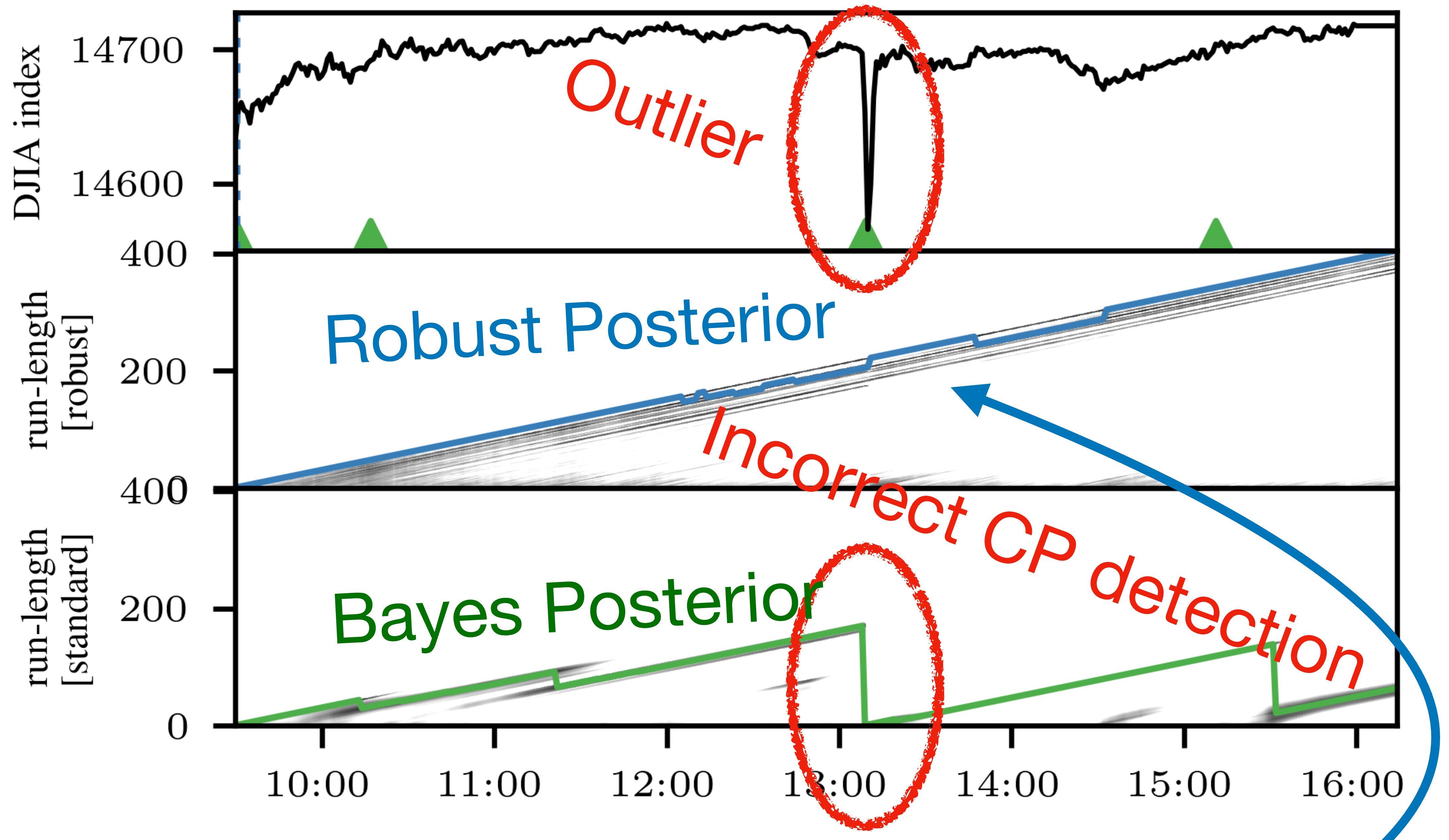
Q2: What L leads to Robustness?

Example 1: Bayesian On-line Changepoint Detection (β-Divergence)

FTX / crytpocurrency crash



Twitter flash crash



Theorem: the robust algorithm cannot declare a change point after a single outlier.

K. & Damoulas (2018); ICML
K., Jewson, & Damoulas (2018); NeurIPS

Approximating L

Q3: How does approximating $\mathsf{L}(x_{1:n}, \theta)$ affect $\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$?

Setting: $\mathsf{L}(x_{1:n}, \theta)$ involves intractable components

(like integrals $I(\theta) = \int f(u)\, p(u \mid \theta)\, du$)

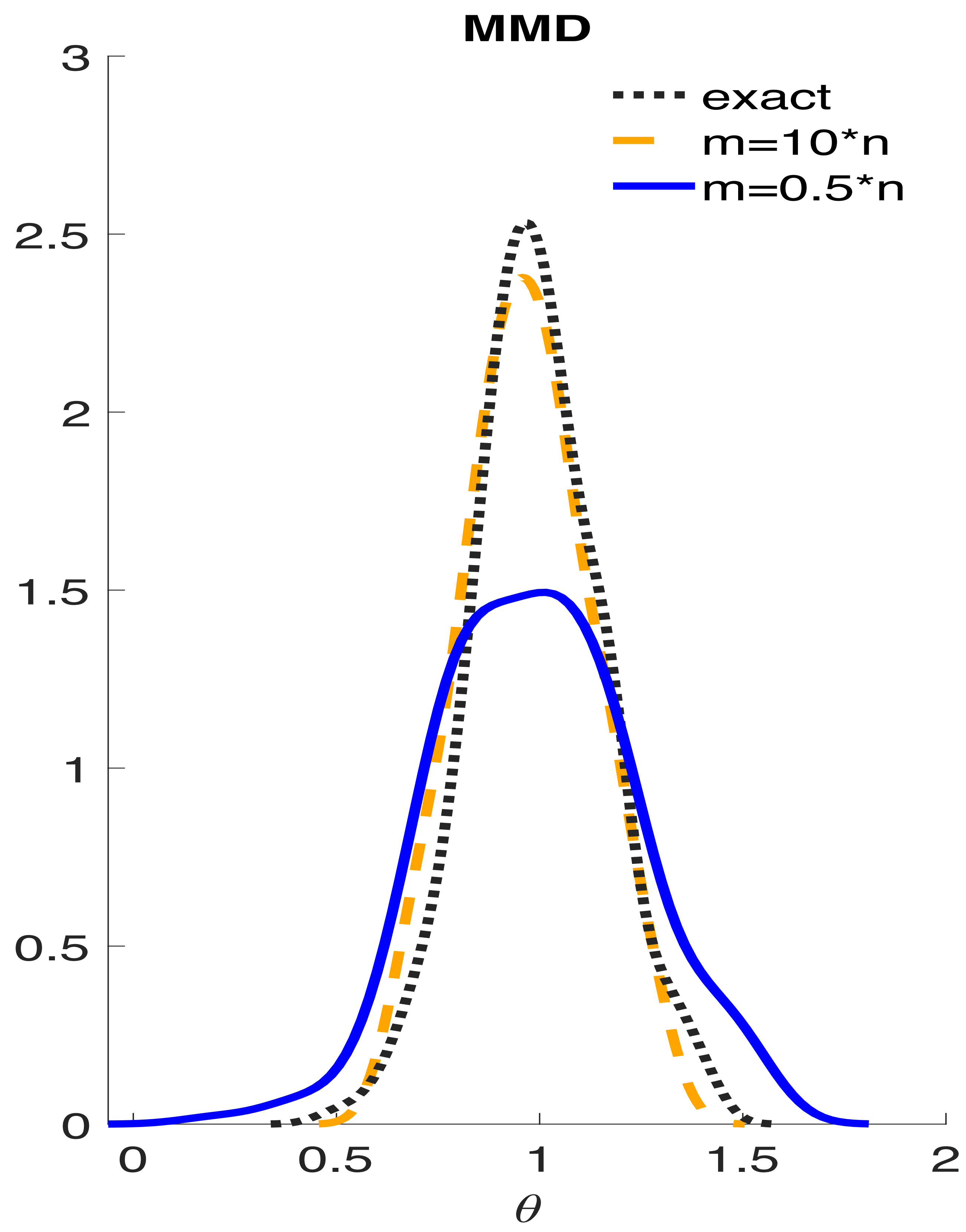
$$\mathsf{L}_m(u_{1:m}, x_{1:n}, \theta) \approx \mathsf{L}(x_{1:n}, \theta)$$

(e.g. $I(\theta) \approx \frac{1}{m} \sum_{j=1}^{m} f(u_j)$ and $u_j \sim p(u_j \mid \theta)$)

Example:

$$\mathsf{L}^k(x_{1:n}, \theta) = n \cdot \iint k(u, u')\, p(u \mid \theta)\, p(u' \mid \theta)\, du du' - 2 \sum_{i=1}^{n} \int k(x_i, u)\, p(u \mid \theta)\, du + C$$

$$\mathsf{L}^k{}_m(u_{1:m}, x_{1:n}, \theta) = \frac{1}{n} \sum_{j=1}^{n} \sum_{l=1}^{n} k(u_j, u_l) - 2 \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} k(x_i, u_j) + C$$



K.*, Frazier*, & Drovandi (2024); preprint

Approximating L

Q3: How does approximating $\mathsf{L}(x_{1:n}, \theta)$ affect $\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$?

Assumption 1: There are $\kappa_1 > 0$, $\kappa_2 > 0$ and $\{\boxed{\sigma_n^2(\theta)} : \Theta \to \mathbb{R}_+\}_{n=1}^{\infty}$ so that

a.s. bounded by RHS for m large enough

$$\overbrace{\mathbb{E}_{u_{1:m} \sim p(u_{1:m}|\theta)}\left[\mathsf{L}_m(u_{1:m}, x_{1:n}, \theta)\right] - \mathsf{L}(x_{1:n}, \theta)}^{= \text{bias}_m(\theta)} \lesssim \boxed{\sigma_n^2(\theta)} \cdot m^{-\kappa_1}$$

Rates of convergence

$$\overbrace{\mathbb{E}_{u_{1:m} \sim p(u_{1:m}|\theta)}\left[\left\{\mathsf{L}_m(u_{1:m}, x_{1:n}, \theta) - \mathbb{E}_{u'_{1:m} \sim p(u_{1:m}|\theta)}\left[\mathsf{L}_m(u'_{1:m}, x_{1:n}, \theta)\right]\right\}^2\right]}^{= \text{variance}_m(\theta)} \lesssim \boxed{\sigma_n^2(\theta)} \cdot m^{-\kappa_2}$$

Assumption 2: $\{\boxed{\sigma_n^2(\theta)} : \Theta \to \mathbb{R}_+\}_{n=1}^{\infty}$ satisfies some prior + posterior integrability conditions

$$\left(\Longrightarrow \int \text{bias}_m(\theta)\, \pi_n^{\mathsf{L}}(\theta \mid x_{1:n})\, d\theta \lesssim m^{-\kappa_1}, \quad \int \text{variance}_m(\theta)\, \pi_n^{\mathsf{L}}(\theta \mid x_{1:n})\, d\theta \lesssim m^{-\kappa_2} \right)$$

K.*, Frazier*, & Drovandi (2024); preprint

Approximating L

Q3: How does approximating $\mathsf{L}(x_{1:n}, \theta)$ **affect** $\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$ **?**

Actual target posterior: $\pi_{n,m}^{\mathsf{L}}(\theta \mid x_{1:n}) \propto \exp\left\{-\mathbb{E}_{u_{1:m}\sim p(u_{1:m}|\theta)}\left[\mathsf{L}_m(u_{1:m}, x_{1:n}, \theta)\right]\right\} \cdot \pi(\theta)$

Theorem 1: Under **Assumptions 1 + 2,** for all $\xi \in [0,2]$ and any fixed $x_{1:n}$,

a.s. bounded by RHS for m large enough

$$\int \|\theta\|^{\xi} \left|\pi_{n,m}^{\mathsf{L}}(\theta \mid x_{1:n}) - \pi_{n,m}^{\mathsf{L}}(\theta \mid x_{1:n})\right| d\theta \quad \lesssim \quad m^{-\min\{\kappa_1, \kappa_2\}}$$

(also implies convergence in TV and TV of all ξth moments)

Rate of convergence:
slower between bias and variance decay
$\Longrightarrow$ trading off bias vs variance improves approximation

K.*, Frazier*, & Drovandi (2024); preprint

Approximating L

Q3: How does approximating $\mathsf{L}(x_{1:n}, \theta)$ affect $\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$?

Assumption 3: Standard (mild) regularity conditions for posterior concentration of $\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$

Also, we have $m = m(n)$ and there are $\eta_1 > 0, \eta_2 > 0$ so that

a.s. of same order for n large enough

$$m(n)^{-\kappa_1} \underbrace{\int \sigma_n^2(\theta)\, \pi_n^{L}(\theta \mid x_{1:n})\, d\theta}_{} \asymp m(n)^{-\eta_1}$$

a.s. of same order for n large enough

$$m(n)^{-\kappa_2} \int \sigma_n^2(\theta)\, \pi_n^{L}(\theta \mid x_{1:n})\, d\theta \asymp m(n)^{-\eta_2}$$

e.g., if $\sigma_n^2(\theta) =$ constant, then $\kappa_1 = \eta_1$

$$\left(\Longrightarrow \int \text{bias}_m(\theta)\, \pi_n^{\mathsf{L}}(\theta \mid x_{1:n})\, d\theta \quad \text{doesn't diverge as } n \to \infty \right)$$

K.*, Frazier*, & Drovandi (2024); preprint

Approximating L

Q3: How does approximating $\mathsf{L}(x_{1:n}, \theta)$ affect $\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$?

Theorem 2: Under **Assumptions 1–3,** for $m = m(n)$, and any $M_n \to \infty$, as $n \to \infty$

$$\lim_{n\to\infty} \mathbb{P}\left(\int \left| \min_{\theta' \in \Theta} \mathsf{L}^{\infty}(\theta') - \mathsf{L}^{\infty}(\theta) \right| \pi_{m,n}^{\mathsf{L}}(\theta \mid x_{1:n})\, d\theta > M_n / \min\left\{ \sqrt{n}, m(n)^{\min\{\eta_1, \eta_2\}} \right\} \right) = 0$$

$$\mathsf{L}^{\infty}(\theta) = \lim_{n\to\infty} \mathbb{E}_{x_{1:n}}\left[\frac{1}{n} \mathsf{L}(x_1 : n, \theta) \right]$$

Standard concentration rate of un-approximated posterior $\pi_n^{\mathsf{L}}(\theta \mid x_{1:n})$

Need to choose $\boxed{m(n) \asymp \left(\sqrt{n}\right)^{1/\min\{\eta_1,\eta_2\}}}$ to avoid slower concentration due to $\mathsf{L}_m(u_{1:m}, x_{1:n}, \theta) \approx \mathsf{L}(x_{1:n}, \theta)$

Tells us how good loss approximation needs to be
Helps evaluate which losses are computationally infeasible

K.*, Frazier*, & Drovandi (2024); preprint

Stein Discrepancies

$$\mathrm{D}_{\text{Stein}}(p(\cdot \mid \theta), q_\varepsilon) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X\sim q_\varepsilon}\left[f(X)\right] - \underbrace{\cancel{\mathbb{E}_{X\sim p(X|\theta)}\left[f(X)\right]}}_{=0} \right| = \mathbb{E}_{X\sim q_\varepsilon}\left[f^*_\theta(X)\right] \approx \frac{1}{n}\sum_{i=1}^{n} f^*_\theta(x_i) = \frac{1}{n}\mathrm{L}(x_{1:n}, \theta)$$

$$\mathcal{F} = \left\{ \mathcal{A}_{p(\cdot|\theta)}(f) : f \in \mathcal{F}_0 \right\}$$

Closed form

depends on $p(\cdot \mid \theta)$ only via $\nabla_x \log p(\cdot \mid \theta)$!

(For all $\mathcal{A}_{p(\cdot|\theta)}$ and $\mathcal{F}_0$ below)

Stein Operator $\mathcal{A}_{p(\cdot\|\theta)}$	Stein Set $\mathcal{F}_0$	$\mathrm{D}_{\text{Stein}}(p(\cdot \mid \theta), q_\varepsilon)$
Langevin-Stein operator $\mathcal{A}_{p(\cdot\|\theta)}(f)(x) = f(x) \cdot \nabla_x \log p(x \mid \theta) + \nabla \cdot f(x)$	$\mathcal{F}_0 = \left\{ f \in C^1(\mathcal{X}, \mathbb{R}) \cap L^2(\mathcal{X}; p(\cdot \mid \theta)) : \\|f\\|_{L^2(\mathcal{X}; p(\cdot\|\theta))} \leq 1 \right\}$	Fisher Divergence (FD)/ Score Matching
Diffusion Stein operator $\mathcal{A}_{p(\cdot\|\theta)}(f)(x) = f(x) \cdot m(x)^T \nabla_x \log p(x \mid \theta) + \nabla \cdot f(x)$	$\mathcal{F}_0 = \left\{ f \in C^1(\mathcal{X}, \mathbb{R}) \cap L^2(\mathcal{X}; p(\cdot \mid \theta)) : \\|f\\|_{L^2(\mathcal{X}; p(\cdot\|\theta))} \leq 1 \right\}$	Weighted Fisher Divergence/ Diffusion Score Matching (DSM) (robust for right choice of m)
	$\mathcal{F}_0 = \left\{ f \in \mathrm{RKHS}(K) : \\|f\\|_K \leq 1 \right\}$	Kernel Stein Discrepancy (KSD) (robust for right choice of m)

$m(x)$: can be tuned for robustness

Computationally Efficient Beliefs: Methodology

$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta} \propto \exp\left\{-\sum_{i=1}^{n} \left(\theta - \mu(x_i)\right)^\top \Lambda(x_i) \left(\theta - \mu(x_i)\right)\right\} \underbrace{\exp\left\{\left(\theta - \mu_0\right)^\top \Lambda_0 \left(\theta - \mu_0\right)\right\}}_{\text{squared exponential prior}}$$

$$\boxed{\stackrel{!}{=} \mathcal{N}\left(\theta; \mu_{\mathsf{L}}(x_{1:n}), \Sigma_{\mathsf{L}}(x_{1:n})\right)}$$

exponential family

$$p(x \mid \theta) = h(x) \cdot \exp\left\{T(x)^\top \theta - A(\theta)\right\}$$

$f^*(x)$ is a 2nd order polynomial

$$\mathsf{L}(p_\theta(x_{1:n})) = \sum_{i=1}^{n} f_\theta^*(x_i) = \sum_{i=1}^{n} f^*(\nabla_x \log p(x_i \mid \theta)) = \sum_{i=1}^{n} a_i + b_i \cdot (\nabla_x \log p(x_i \mid \theta)) + c_i \cdot (\nabla_x \log p(x_i \mid \theta))^2$$

$$= \sum_{i=1}^{n} \left[a_i \tilde{h}_i + b_i \tilde{h}_i^{\,2}\right] + \left[b_i \tilde{T}_i + 2 c_i \tilde{h}_i \tilde{T}_i\right]^\top \theta + \theta^\top \left[c_i \tilde{T}_i \tilde{T}_i^{\top}\right] \theta$$

$$= \nabla_x \log h(x_i) + \nabla_x T(x_i)^\top \theta$$

$$\boxed{\stackrel{+C}{=} \sum_{i=1}^{n} \left(\theta - \mu(x_i)\right)^\top \Lambda(x_i) \left(\theta - \mu(x_i)\right)}$$

$$= \tilde{h}_i \qquad = \tilde{T}_i$$

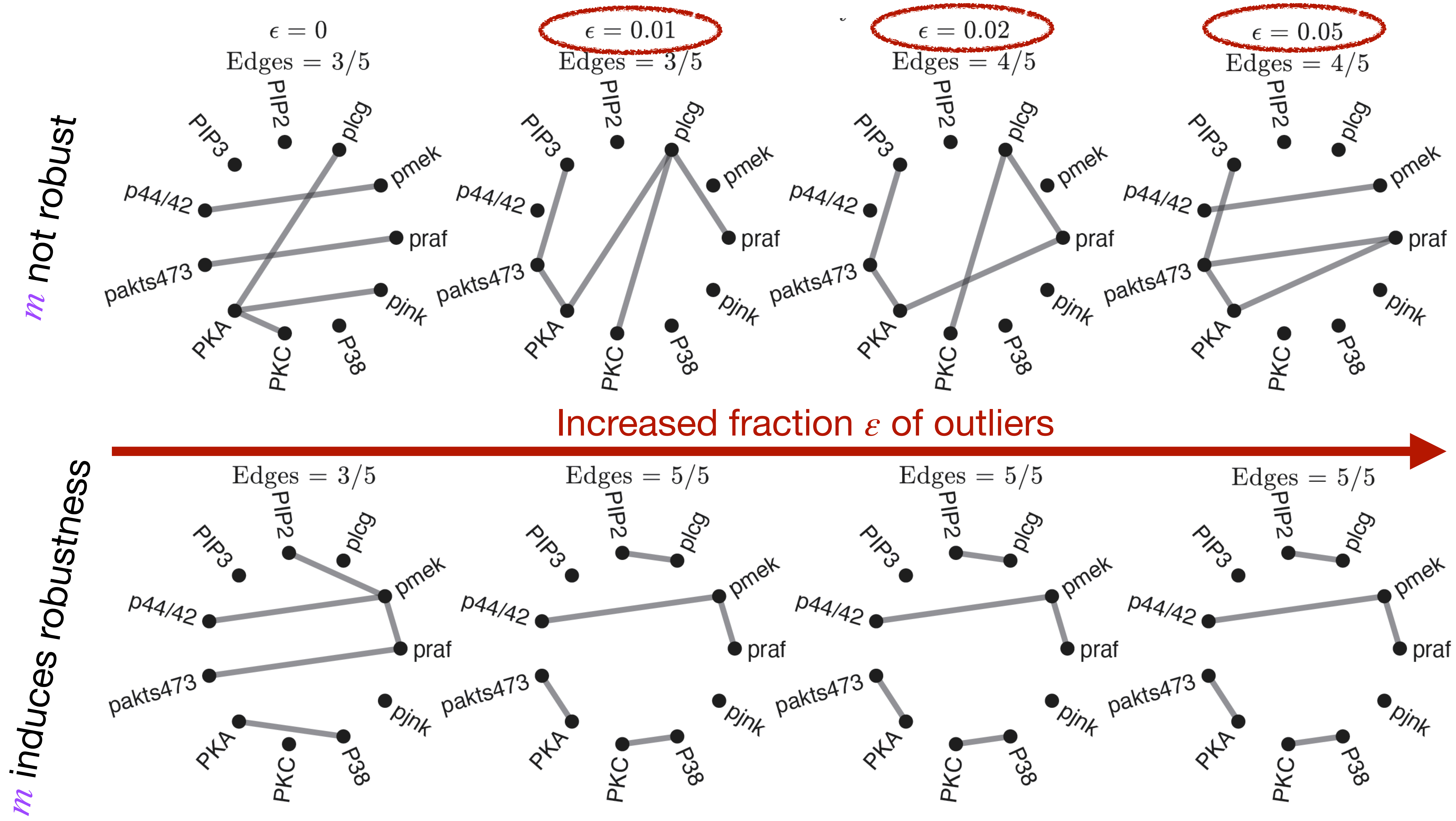
$\exp\{-\ldots\}$ of this looks like a Gaussian in θ!

$\Longrightarrow$ Put squared exponential prior on θ!

L for Robustness & Computation: Graphical Models

$$p(x_i \mid \theta) \propto \exp\left\{ -\sum_{d=1}^{D} \theta_d x_{i,(d)} - \sum_{d<d'}^{D} \theta_{d,d'} x_{i,(d)} x_{i,(d')} \right\}; \qquad \theta_d > 0, \theta_{d,d'} \geq 0.\ D = 11 \quad \text{(proteins)}, \quad n = 7466 \quad \text{(number of cells measured)}$$

L = $\boxed{\text{Kernel Stein Discrepancy}}$ (robust / non-robust loss depending only on $\nabla_x \log p(x_{1:n} \mid \theta)$)

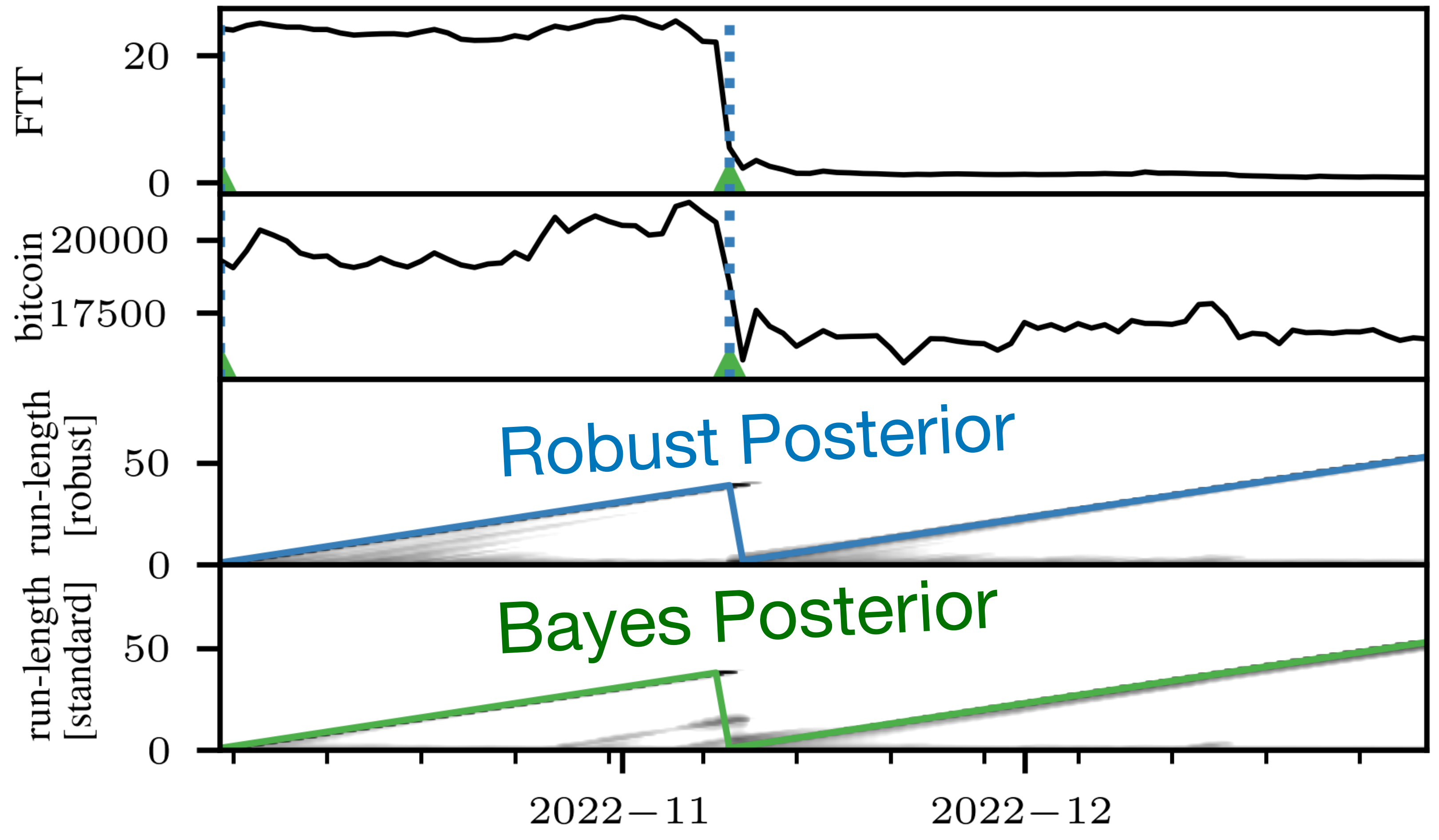


Matsubara, **K.**, Briol, & Oates (2022); JRSS-B

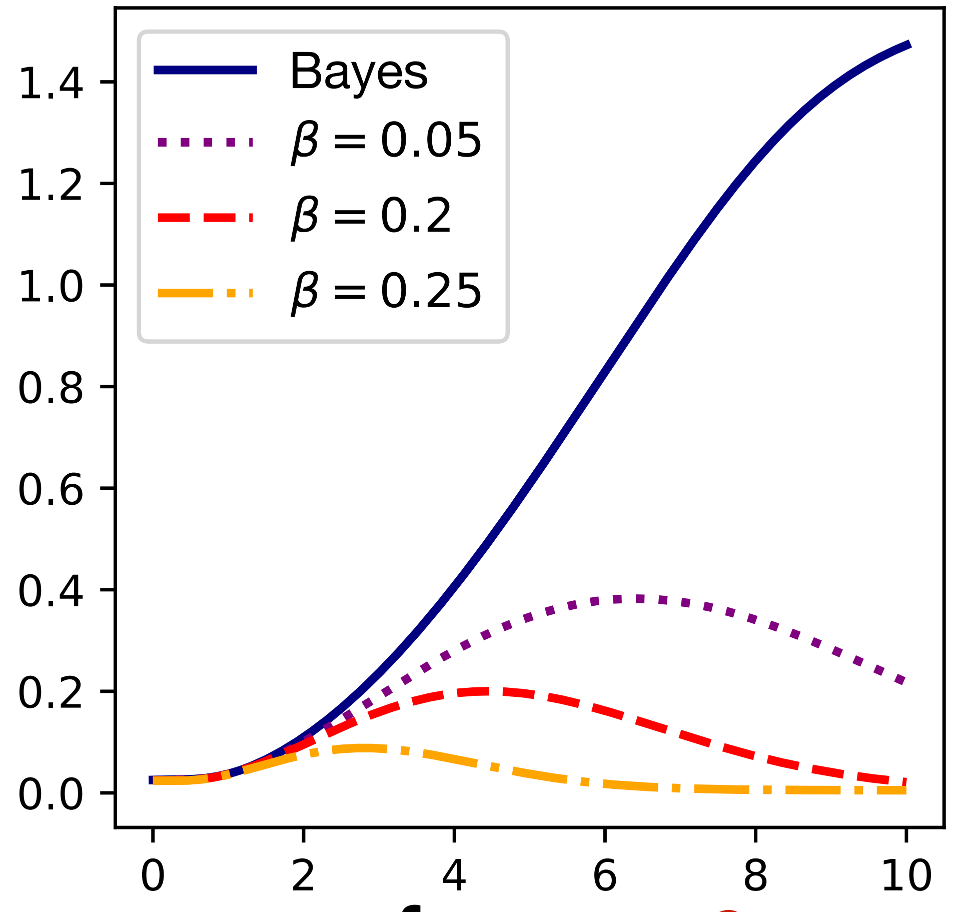
Choices of L

Example 1: Bayesian On-line Changepoint Detection (β-Divergence)

FTX / crytpocurrency crash



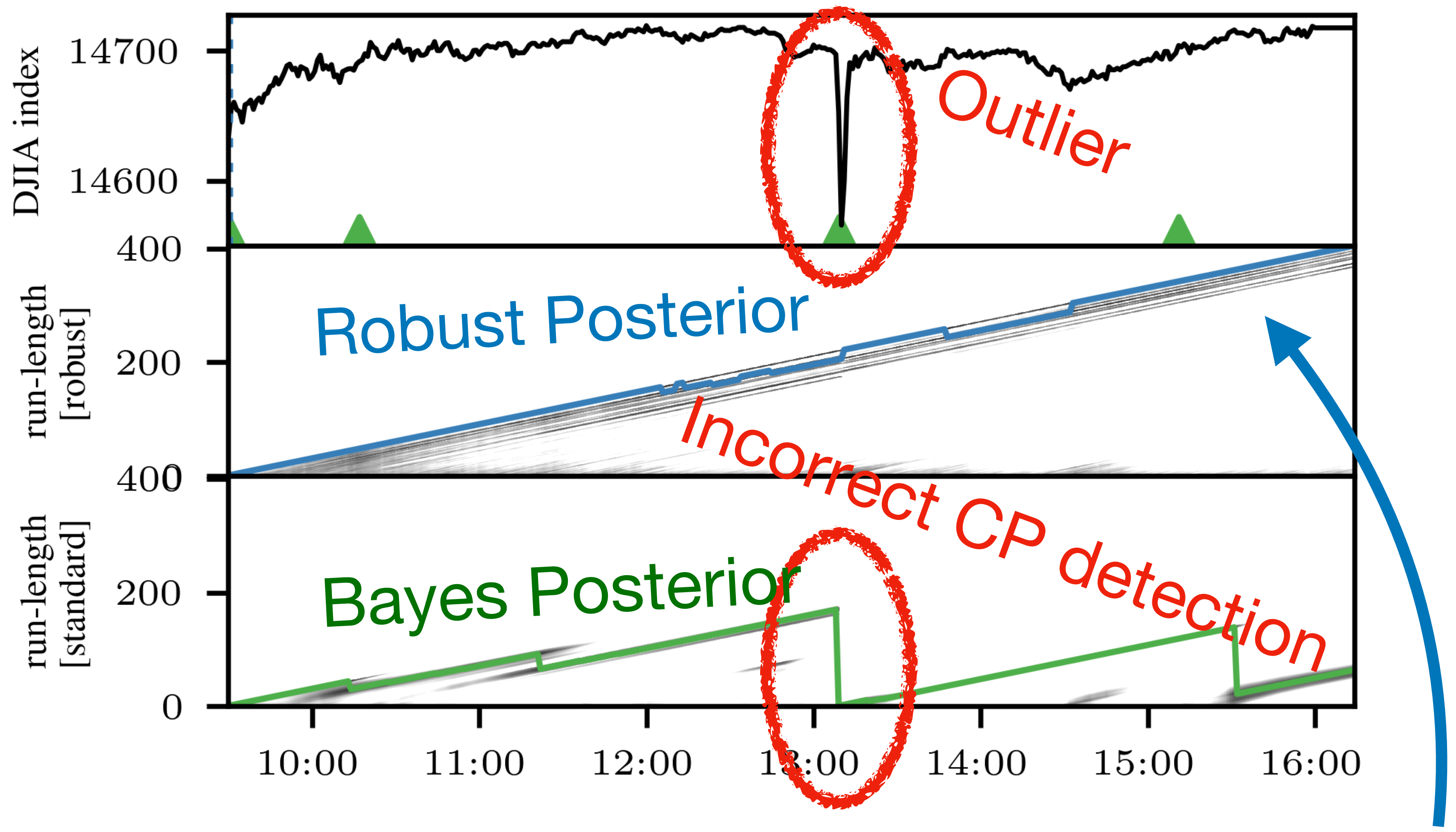
$$\lim_{\varepsilon \to 0} \mathrm{D}_{\mathrm{FR}}\left(\pi_n^{\mathrm{L}}(\theta \mid q_\varepsilon), \pi_n^{\mathrm{L}}(\theta \mid q_0)\right)$$



y for $c = \delta_y$

$$q_\varepsilon = (1 - \varepsilon) \cdot q_0 + \varepsilon \cdot c$$

Twitter flash crash

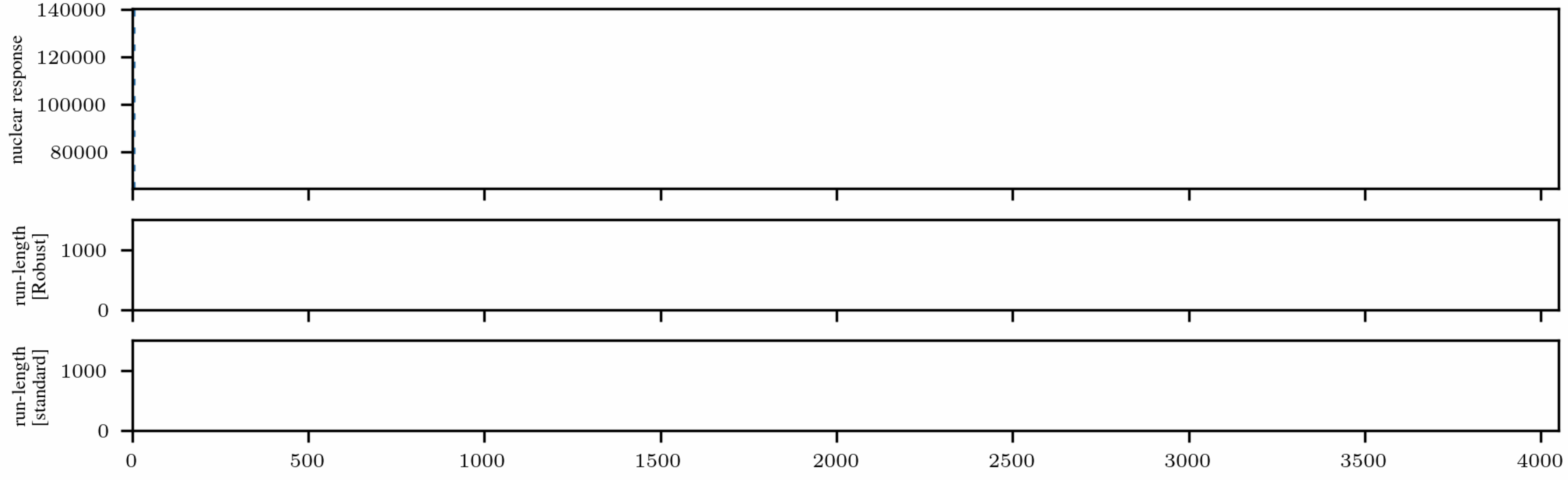


Theorem: the robust algorithm cannot declare a change point after a single outlier.

K. & Damoulas (2018); ICML
K., Jewson, & Damoulas (2018); NeurIPS

L for Robustness & Computation: Changepoints

L = $\boxed{\text{Weighted Fisher Divergence}}$ (robust / non-robust loss depending only on $\nabla_x \log p(x_{1:n} \mid \theta)$)

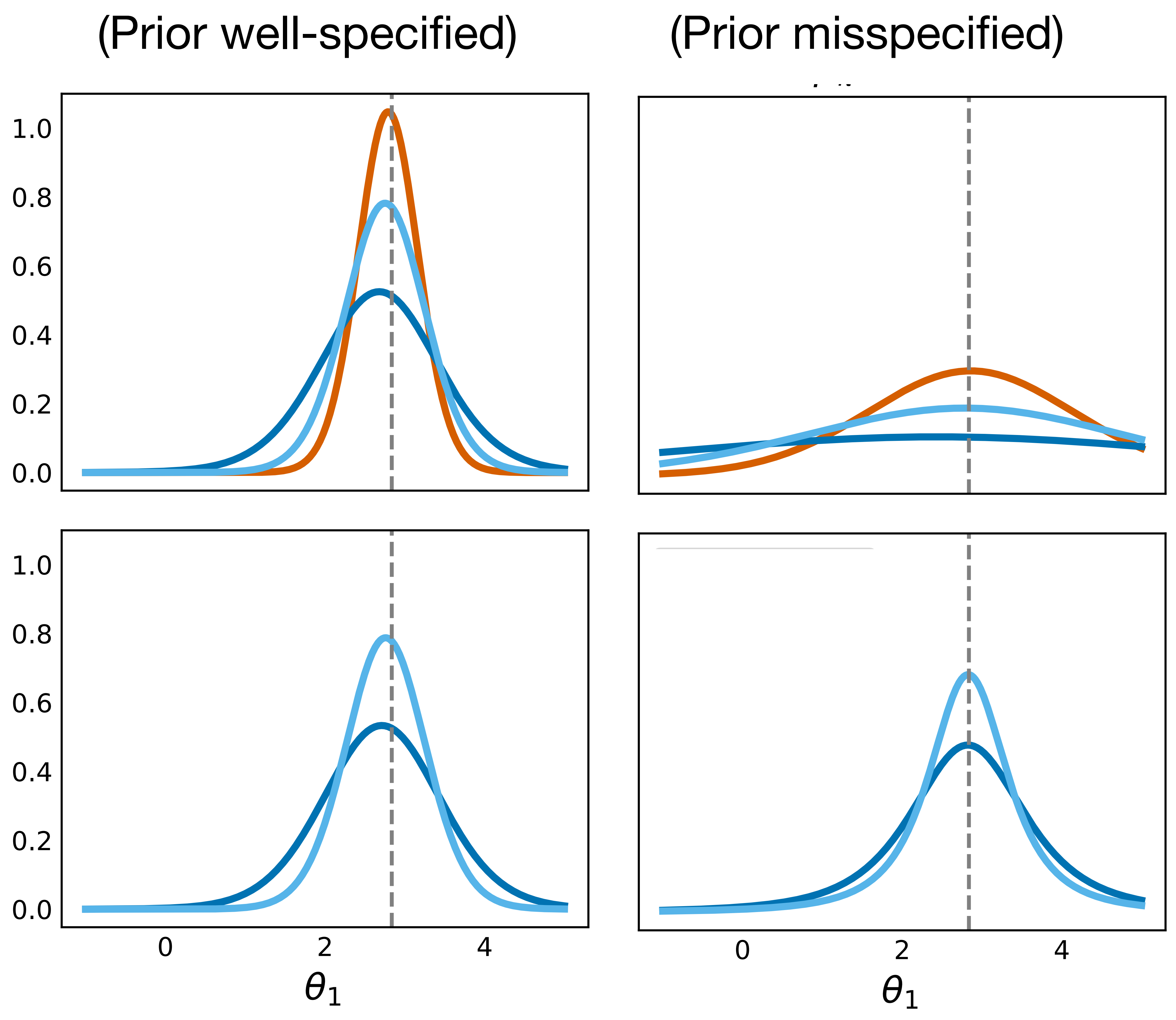


Altamirano, Briol, & **K.** (2023); ICML

Post-Bayesian ML: Optimisation-centric posteriors

Assumptions

(A1)	model well-specified	✓
(A2)	prior well-specified	✗
(A3)	computationally feasible	✗



Mean Field VI for:

Standard Bayes

Power posterior; $\lambda = 0.6$

Power posterior; $\lambda = 0.3$

GVI/Optimisation-centric:

D = Renyi divergence; $\alpha = 0.6$

D = Renyi divergence; $\alpha = 0.3$

$$q_n^*(\theta) = \operatorname{argmin}_{q \in \text{Normals}} \left\{ \mathscr{L}^{(\lambda)}_{\mathrm{L},\, D,\, \pi}(q) \right\}$$

$$= -\int \log p(x_{1:n} \mid \theta)\, q(\theta)\, d\theta + \mathrm{D}(q, \pi)$$

Picture from **K.**, Jewson, & Damoulas (2022); JMLR

~~(A1)~~, ~~(A2)~~, ~~(A3)~~

③

Case Study: Why Post-Bayesian ML matters

Objective: $q \mapsto \mathbb{E}_{\theta \sim q}\left[-\log p(x_{1:n} \mid \theta)\right] + \frac{1}{\lambda} \cdot \mathrm{KL}(q, \pi) \xrightarrow{\lambda \to \infty} q \mapsto \mathbb{E}_{\theta \sim q}\left[-\log p(x_{1:n} \mid \theta)\right]$

Target: **Cold Posterior** $(\lambda \gg 1)$ **/ Bayes Posterior** $(\lambda = 1)$ — **Deep Ensemble** $(\lambda \to \infty)$

Wasserstein Gradient Flow:

Step 1: Sample $\theta_k(0) \sim \pi, k = 1,2,\ldots K$

Step 2: For $t \in [0,T]$, evolve as

Brownian motions $\{B_n(t)\}_{t \geq 0}$

$$d\theta_k(t) = -\left(\lambda \cdot \nabla_\theta \mathrm{L}(x_{1:n}, \theta_k(t)) - \nabla \log \pi(\theta_k(t))\right) dt + \sqrt{2 dB_k(t)}$$

$$\boxed{q_n^*(\theta) = \pi_n^{(\lambda)}(\theta \mid x_{1:n}) \approx \frac{1}{K} \sum_{n=1}^{K} \theta_k(T) \text{ as } T \to \infty, K \to \infty .}$$

Converges to well-defined density

Wasserstein Gradient Flow = Deep Ensemble Algorithm

Step 1: Sample $\theta_k(0) \sim \pi, k = 1,2,\ldots K$

Step 2: For $t \in [0,T]$, evolve as

$$d\theta_k(t) = -\lambda \cdot \nabla_\theta \mathrm{L}(x_{1:n}, \theta_k(t))$$

$$\boxed{\text{Deep Ensemble} = \frac{1}{K} \sum_{n=1}^{K} \theta_k(T)}$$

Converges to ill-defined atomic measure

Wild, Ghalebikesabi, Sejdinovic, & **Knoblauch** (2023); NeurIPS Oral

$\cancel{(A1)}$, $\cancel{(A2)}$, $\cancel{(A3)}$

③

The Future of Post-Bayesian ML: Success Stories

$q \mapsto \mathbb{E}_{\theta\sim q}\left[\mathsf{L}(x_{1:n}, \theta)\right]$ ⟶ **Deep Ensemble**

$q \mapsto \mathbb{E}_{\theta\sim q}\left[\mathsf{L}(x_{1:n}, \theta)\right] + w_2 \cdot \mathrm{KL}(q, \pi)$ ⟶ **Cold Posterior** $(w_2 \ll 1)$ / **Bayes Posterior** $(w_2 = 1)$

$q \mapsto \mathbb{E}_{\theta\sim q}\left[\mathsf{L}(x_{1:n}, \theta)\right] + w_1 \cdot \mathrm{MMD}^2(q, \pi) + w_2 \cdot \mathrm{KL}(q, \pi)$ ⟶ **BDL Ensemble with repulsive particles**

Infinite-dimensional gradient descent / Wasserstein Gradient Flow:

Step 1: Sample $\theta_k(0) \sim \pi, k = 1,2,\ldots K$

Step 2: Evolve via SDE for $t \in [0,T]$ as

independent Brownian motions $\{B_n(t)\}_{t\geq 0}$

$$d\theta_k(t) = -\left(\nabla_\theta \mathsf{L}(x_{1:n}, \theta_k(t)) - w_1 \nabla \mu_\pi(\theta_k(t)) - w_2 \nabla \log \pi(\theta_k(t)) + \frac{w_1}{K}\sum_{j=1}^{K} w_1 \kappa(\theta_k(t), \theta_j(t))\right) dt + \sqrt{2 w_2 dB_k(t)}$$

$$q_n^*(\theta) \approx \frac{1}{K}\sum_{n=1}^{K} \theta_k(T) \text{ as } T \to \infty, K \to \infty\,.$$

Wild, Ghalebikesabi, Sejdinovic, & **K.** (2023); NeurIPS Oral

Adversarial Robustness: How to think about general D?

Theorem: $\mathrm{L}(x_{1:n}, \theta)$ bounded + measurable, $\lambda > 0$, $\mathcal{Q}$ is closed, bounded, convex:

$$\inf_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim q} \left[\lambda \cdot \mathrm{L}(x_{1:n}, \theta) \right] + \mathrm{D}(q, \pi) \right\} = \sup_{h \in F_b(\Theta)} \left\{ \inf_{q \in \mathcal{Q}} \mathbb{E}_{\theta \sim q} \underbrace{\left[\lambda \cdot \mathrm{L}(x_{1:n}, \theta) + \lambda \cdot h(\theta) \right]}_{\text{Perturbed / worsened loss}} - \underbrace{\Delta_{\mathrm{D},\pi}(\lambda \cdot h)}_{\text{Penalty for perturbation}} \right\}$$

inner minimisation over (perturbed) loss

outer maximisation (of adversary) over perturbations
$h \in F_b(\Theta) = \{$bounded + measurable functions $f : \Theta \to \mathbb{R}\}$

Legendre-Fenchel dual of $\mathrm{D}(\cdot, \pi) = \sup_{\rho \in \mathcal{P}(\Theta)} \left\{ \int_\Theta h(\theta)\, \rho(\theta)\, d\theta - \mathrm{D}(\rho, \pi) \right\} = \Delta_{\mathrm{D},\pi}(h)$

(penalty for increasing $h(\theta)$ highest where $\pi(\theta)$ largest)

Husain & **K.** (2022); ALT

Adversarial Robustness: How to think about general D?

Theorem: $\mathsf{L}(x_{1:n}, \theta)$ bounded + measurable, $\lambda > 0$, $\mathcal{Q}$ is closed, bounded, convex:

$$\inf_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim q}\left[\lambda \cdot \mathsf{L}(x_{1:n}, \theta)\right] + \mathrm{D}(q, \pi) \right\} = \sup_{h \in F_b(\Theta)} \left\{ \inf_{q \in \mathcal{Q}} \underbrace{\mathbb{E}_{\theta \sim q}\left[\lambda \cdot \mathsf{L}(x_{1:n}, \theta) + \lambda \cdot h(\theta)\right]}_{\text{Perturbed / worsened loss}} - \underbrace{\Delta_{\mathrm{D},\pi}(\lambda \cdot h)}_{\text{Penalty for perturbation}} \right\}$$

Example 1: $\mathrm{D} = \mathrm{KL}$, $\quad \Delta_{\mathrm{D},\pi}(\lambda \cdot h) = \log \int \exp\{\lambda \cdot h(\theta)\}\ \pi(\theta)\, d\theta$

Example 2: $\mathrm{D} = \chi^2$, $\quad \Delta_{\mathrm{D},\pi}(\lambda \cdot h) = \lambda \cdot \int h(\theta)\, \pi(\theta)\, d\theta + \frac{\lambda^2}{4} \mathrm{Var}_{\theta \sim \pi}\left(h(\theta)\right)$

Example 3: $\mathrm{D} = \mathrm{IPM}_{\mathscr{W}}$, $\quad \Delta_{\mathrm{D},\pi}(\lambda \cdot h) = \begin{cases} \lambda \cdot \int h(\theta)\, \pi(\theta)\, d\theta & \text{if } h \in \mathscr{W} \\ \infty & \text{if } h \notin \mathscr{W} \end{cases}$

π = **preferences**. More density/mass where we want perturbations to be more expensive for adversary.

D = **cost/penalty function**. Determines actual cost of perturbations to adversary

Husain & **K.** (2022); ALT

How to approach computation with general D?

More Elegant Strategy: analytical solutions

Advantage: further insight on effect of D

Disadvantage:

- only applicable for a small selection of D
- computationally intractable

How could we fix this?

$$q_n^*(\theta) = \nabla f^*\big(Z - \lambda \cdot \mathsf{L}(x_{1:n}, \theta)\big)\ \pi(\theta) = \arg\min_{q \in \mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta \sim q}\left[\lambda \cdot \mathsf{L}(x_{1:n}, \theta)\right] + \mathrm{D}_f(q, \pi) \right\}$$

$$\mathrm{D}_f(q, \pi) = \mathbb{E}_{\theta \sim \pi}\left[f\left(\frac{q(\theta)}{\pi(\theta)}\right)\right]$$

$Z =$ normaliser defined via $\int \nabla f^*\big(Z - \lambda \cdot \mathsf{L}(x_{1:n}, \theta)\big)\ \pi(\theta)\, d\theta = 1$

f^* is the Fenchel conjugate of f, $\quad f^*(x) = \sup_{x' \in \mathbb{R}} \left\{ \langle x, x' \rangle - f(x') \right\}$

$f(x) = x^2 - 1$

Example: $\mathrm{D}_f = \chi^2, \quad \mathsf{L} \geq 0$

$$q_n^*(\theta) = \frac{1}{2} \max\left\{0, Z - \lambda \cdot \mathsf{L}(x_{1:n}, \theta)\right\} \pi(\theta)$$

Alquier, P. (2021); ICML

Post-Bayesian ML

Optimisation-centric posteriors / GVI

$$q_n^*(\theta) = \operatorname{argmin}_{q \in \mathcal{Q}} \left\{ \mathscr{L}^{(\lambda)}_{\mathsf{L},\, D,\, \pi}(q) \right\}; \quad \mathcal{Q} \subseteq \mathscr{P}(\Theta)$$

Gibbs/Generalised/Quasi/Pseudo Posterior

$$\pi_n^{(\lambda,\mathsf{L})}(\theta \mid x_{1:n}) = \frac{\exp\{-\lambda \cdot \mathsf{L}(x_{1:n}, \theta) \cdot \pi(\theta)}{\int \exp\{-\lambda \cdot \mathsf{L}(x_{1:n}, \theta) \cdot \pi(\theta) d\theta}$$

Martingale Posterior

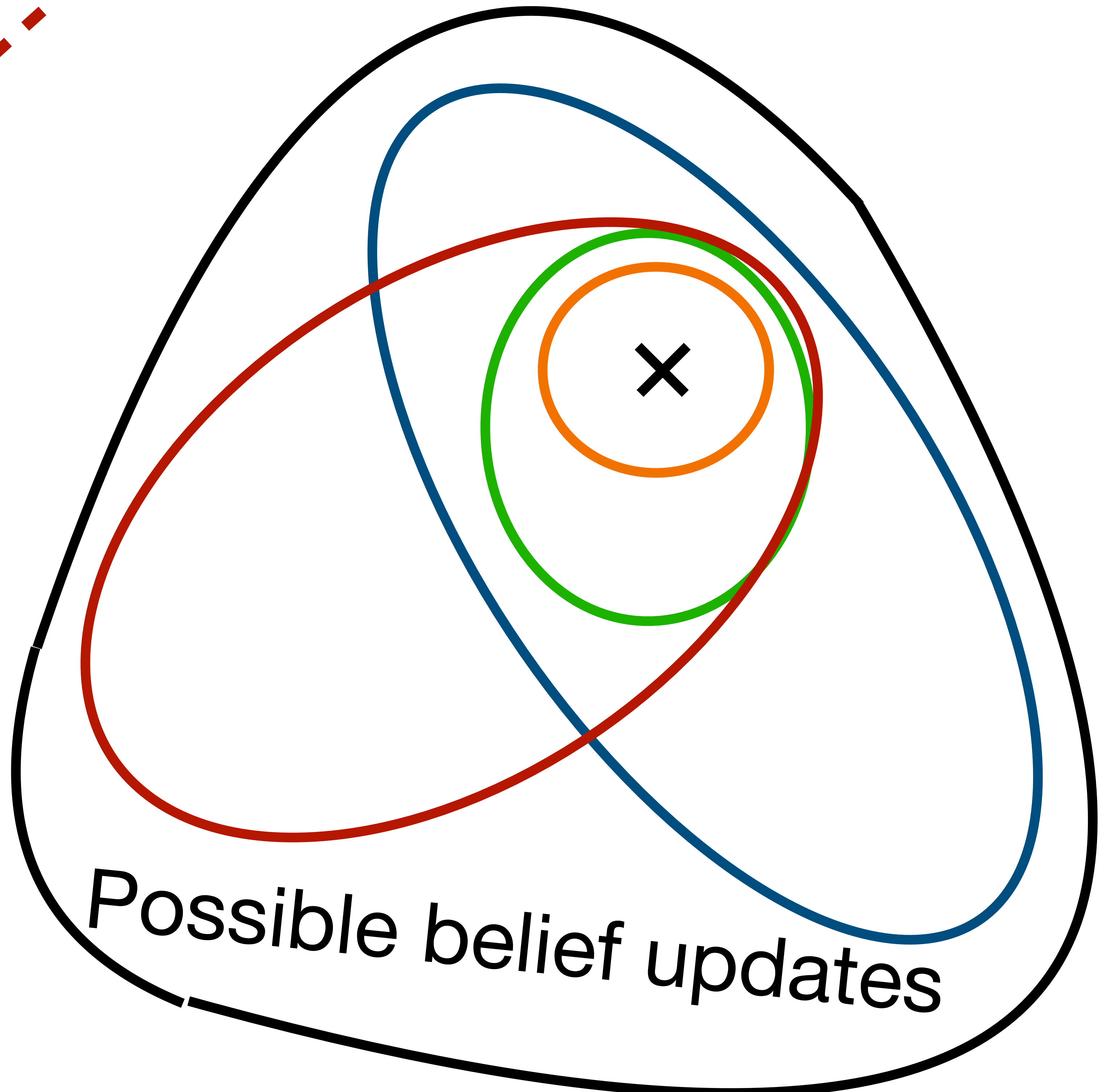
For $i = 1, 2, \ldots$

$$X_{n+i+1} \sim p(X_{n+i} \mid x_{1:n}, X_{n+1:n+i})$$

$$\theta^\infty = \operatorname{argmin}_{\theta \in \Theta} \mathsf{L}\left([x_{1:n}, X_{n+1:\infty}], \theta\right)$$

Power/Fractional/Cold Posterior

$$\pi_n^{(\lambda)}(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)^\lambda \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta)^\lambda \cdot \pi(\theta) d\theta}$$



Bayes' Posterior

$$\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}$$

Post-Bayesian ML

For $i = 1,2,\ldots$

$$X_{n+i+1} \sim p(X_{n+i+1} \mid \{x_{1:n} \cup X_{n+1:n+i}\})$$

$$\theta^\infty = \text{argmin}_{\theta\in\Theta} - \log p(\{x_{1:n} \cup X_{n+1:\infty}\} \mid \theta)$$

Bayes' posterior predictive:

$$p(x_{n+1} \mid x_{1:n}) = \int p(x_{n+1} \mid \theta)\, \pi_n(\theta \mid x_{1:n})\, d\theta$$

$$\theta^\infty \sim \pi_n(\theta \mid x_{1:n})$$

(Doob's Consistency Theorem)

$=$

Bayes' Posterior

$$\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}$$

Post-Bayesian ML

For $i = 1,2,\ldots$

$$X_{n+i+1} \sim p(X_{n+i+1} \mid \{x_{1:n} \cup X_{n+1:n+i}\})$$

$$\theta^\infty = \mathrm{argmin}_{\theta\in\Theta} \mathsf{L}(\{x_{1:n} \cup X_{n+1:\infty}\}, \theta)$$

Generalised Bayes' posterior predictive:

$$p(x_{n+1} \mid x_{1:n}) = \int p(x_{n+1} \mid \theta)\, \pi_n^{(\lambda,\mathsf{L})}(\theta \mid x_{1:n})\, d\theta$$

(If parameter θ indexes a model $p(\cdot \mid \theta)$)

$$\theta^\infty \sim \pi_n^{(\lambda,\mathsf{L})}(\theta \mid x_{1:n})$$

(Doob's Consistency)

$$=$$

Gibbs/Generalised/Quasi/Pseudo Posterior

$$\pi_n^{\mathsf{L}}(\theta \mid x_{1:n}) = \frac{\exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta)}{\int \exp\{-\mathsf{L}(x_{1:n}, \theta)\} \cdot \pi(\theta) d\theta}$$

Post-Bayesian ML

For $i = 1,2,\ldots$

$$X_{n+i+1} \sim p(X_{n+i+1} \mid \{x_{1:n} \cup X_{n+1:n+i}\})$$

$$\theta^\infty = \text{argmin}_{\theta \in \Theta} \mathsf{L}\left(\{x_{1:n} \cup X_{n+1:\infty}\}, \theta\right)$$

Choose predictive & parameter/loss:

$p(x_{n+1} \mid x_{1:n})$ = Your choice

$\mathsf{L}(x_{1:n}, \theta)$ = Loss determining θ

Martingale condition:

$$\mathbb{E}_{X_{n+1} \sim p(X_{n+1} \mid x_{1:n})}\left[p\left(z \mid \{x_{1:n} \cup X_{n+1}\}\right) \middle| x_{1:n}\right] = p(z \mid x_{1:n})$$

Martingale Posterior

$$\theta^\infty \sim \pi_n^{(\mathsf{L},p)}(\theta \mid x_{1:n})$$

Post-Bayesian ML

For $i = 1,2,\ldots$

$$X_{n+i+1} \sim p(X_{n+i+1} \mid \{x_{1:n} \cup X_{n+1:n+i}\})$$

$$\theta^\infty = \mathrm{argmin}_{\theta\in\Theta} \mathsf{L}\left(\{x_{1:n} \cup X_{n+1:\infty}\}, \theta\right)$$

Predictive replaces model & prior specification

Good specification

=

predictive capturing uncertainty about observables

Depends on choices for predictive & loss

Martingale Posterior

$$\theta^\infty \sim \pi_n^{(\mathsf{L},p)}(\theta \mid x_{1:n})$$

Assumptions of Bayesian inference

(A1) model well-specified N/A?

(A2) prior well-specified N/A?

(A3) inversion computationally feasible ✓

Post-Bayesian ML

Optimisation-centric posteriors / GVI

$$q_n^*(\theta) = \text{argmin}_{q\in\mathcal{Q}} \left\{ \mathcal{L}^{(\lambda)}_{\mathsf{L},\, D,\, \pi}(q) \right\}; \quad \mathcal{Q} \subseteq \mathcal{P}(\Theta)$$

Gibbs/Generalised/Quasi/Pseudo Posterior

$$\pi_n^{(\lambda,\mathsf{L})}(\theta \mid x_{1:n}) = \frac{\exp\{-\lambda \cdot \mathsf{L}(x_{1:n},\theta) \cdot \pi(\theta)}{\int \exp\{-\lambda \cdot \mathsf{L}(x_{1:n},\theta) \cdot \pi(\theta) d\theta}$$

Martingale Posterior

For $i = 1,2,\ldots$

$$X_{n+i+1} \sim p(X_{n+i} \mid x_{1:n}, X_{n+1:n+i})$$

$$\theta^\infty = \text{argmin}_{\theta\in\Theta} \mathsf{L}\left([x_{1:n}, X_{n+1:\infty}], \theta\right)$$

Power/Fractional/Cold Posterior

$$\pi_n^{(\lambda)}(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)^\lambda \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta)^\lambda \cdot \pi(\theta) d\theta}$$

Bayes' Posterior

$$\pi_n(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta) \cdot \pi(\theta)}{\int p(x_{1:n} \mid \theta) \cdot \pi(\theta) d\theta}$$