Generalized Variational Inference (GVI)

Posterior beliefs with the rule of three

Jeremias Knoblauch[1,3], Jack Jewson[1,3],
Theodoros Damoulas[1,2,3]

September 29, 2020

[1]University of Warwick, Department of Statistics
[2]University of Warwick, Department of Computer Science
[3]The Alan Turing Institute for Data Science and AI

Structure of the talk

1. The form of the Bayesian problem
   1.1 The traditional perspective
   1.2 The optimization perspective
   1.3 The loss-minimization perspective
   1.4 The new perspective
2. The form of the Generalized Bayesian problem
   2.1 Provable modularity
   2.2 Axiomatic derivation
   2.3 Relationship to existing methods
3. Reinterpreting standard **VI**
   3.1 Optimality & reinterpretation of standard **VI**
   3.2 Why does **DVI** produce better posteriors?
   3.3 Towards **GVI**
4. **GVI**: What does it do?
   4.1 M-open vs M-closed
   4.2 The losses
   4.3 Uncertainty Quantification
   4.4 Three **GVI** use cases
5. **GVI**: Inference & Experiments
   5.1 Black box **GVI**
   5.2 Robust Bayesian On-line Changepoint Detection
   5.3 Bayesian Neural Networks
   5.4 Deep Gaussian Processes

1 The form of the Bayesian problem

Purpose of part 1: Motivate the rule of three

(1) Bayesian inference minimizes **losses**

(2) Bayesian inference **regularizes** with the prior

(3) Bayesian inference = **optimization** over **(sub)spaces of probability measures**

1.1 The Bayesian problem: Traditional perspective

Ingredients (for the simplest case) are:

- $n = n_1 + n_2$ observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_{n_1+n_2})^T$,
- **prior** $\pi(\boldsymbol{\theta})$,
- **likelihoods** $\{p(x_i|\boldsymbol{\theta})\}_{i=1}^{n_1+n_2}$

Output = **posterior belief**:

$$q^*(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^{n_1+n_2} p(x_i|\boldsymbol{\theta}) = \widetilde{\pi}(\boldsymbol{\theta}) \prod_{i=n_1+1}^{n_2} p(x_i|\boldsymbol{\theta}), \text{ for } \widetilde{\pi}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) \prod_{i=1}^{n_1} p(x_i|\boldsymbol{\theta})$$

Inference interpretation = belief updates:

- likelihoods $\{p(x_i|\boldsymbol{\theta})\}_{i=1}^{n_1+n_2}$ update prior about $\boldsymbol{\theta}$
- Old posterior $\widetilde{\pi}(\boldsymbol{\theta})$ = new prior (**coherence**/**Bayesian additivity**)

1.2 The Bayesian problem: The optimization perspective

Zellner (1988) shows that the Bayes posterior $q^*(\boldsymbol{\theta})$ solves

$$q^*(\boldsymbol{\theta}) = \underset{q \in \mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \underbrace{\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} -\log(p(x_i|\boldsymbol{\theta}))\right]}_{\text{minimized by } q(\boldsymbol{\theta}) = \delta_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{\theta}),\ \hat{\boldsymbol{\theta}}_n = \text{MLE}} + \underbrace{\mathrm{KLD}\left(q||\pi\right)}_{\text{minimized by } q = \pi} \right\}, \quad (1)$$

Notation:

- $\mathcal{P}(\boldsymbol{\Theta})$ = all probability distributions on $\boldsymbol{\Theta}$
- KLD = Kullback-Leibler divergence = $\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log q(\boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta})\right]$

Inference interpretation = regularized loss-minimization:

- $-\log(p(x_i|\boldsymbol{\theta}))$ = **loss** of $\boldsymbol{\theta}$ for x_i
- Inference = regularizing MLE $\hat{\boldsymbol{\theta}}_n$ with $\mathrm{KLD}(q||\pi)$

1.3 The Bayesian problem: The loss-minimization perspective

Bissiri et al. (2016): Bayes posteriors $q^*(\boldsymbol{\theta})$ for general loss $\ell(\boldsymbol{\theta}, x_i)$:

$$q^*(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \exp\left\{-\sum_{i=1}^{n_1+n_2} \ell(\boldsymbol{\theta}, x_i)\right\} = \widetilde{\pi}(\boldsymbol{\theta}) \exp\left\{-\sum_{i=n_1+1}^{n_2} \ell(\boldsymbol{\theta}, x_i)\right\}$$

$$\text{for } \; \widetilde{\pi}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) \exp\left\{-\sum_{i=1}^{n_1} \ell(\boldsymbol{\theta}, x_i)\right\}$$

Inference interpretation = belief updates:

- Again: losses $\{\ell(\boldsymbol{\theta}, x_i)\}_{i=1}^{n_1+n_2}$ update prior about $\boldsymbol{\theta}$
- Again: Old posterior $\widetilde{\pi}(\boldsymbol{\theta})$ = new prior (**coherence**)
- Difference: $\boldsymbol{\theta}$ arbitrary, e.g. $\ell(\boldsymbol{\theta}, x_i) = |x_i - \boldsymbol{\theta}|$ admissible

1.4 The Bayesian problem: The new perspective I/II

Easy to show: Zellner's representation valid for any $\ell(\boldsymbol{\theta}, x_i)$:

$$q^*(\boldsymbol{\theta}) = \underset{q \in \mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \underbrace{\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\right]}_{\text{minimized by } \delta_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{\theta})} + \underbrace{\mathrm{KLD}\left(q || \pi\right)}_{\text{minimized by } q = \pi} \right\}$$

Bissiri et al. (2016)'s generalization (preserves coherence):

- Replacing $-\log(p(x_i|\boldsymbol{\theta}))$ with other losses $\ell(\boldsymbol{\theta}, x_i)$

Two more generalizations (break coherence):

- Replacing $\mathcal{P}(\boldsymbol{\Theta})$ with $\mathcal{Q} \subset \mathcal{P}(\boldsymbol{\Theta})$ ($=$ **VI**)
- Replacing KLD with inference-problem specific regularizers

1.4 The Bayesian problem: The new perspective II/II

Our **generalized representation of Bayesian inference**:

$$q^*(\boldsymbol{\theta}) = \underset{q \in \Pi}{\arg\min} \left\{ \underbrace{\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\right]}_{\text{minimized by } \delta_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{\theta})} + \underbrace{D(q\|\pi)}_{\text{minimized by } q = \pi} \right\}$$

Notation:

- if $\boldsymbol{\Pi}$ = variational family, write $\mathcal{Q}$.
- $\ell_n(\boldsymbol{\theta}, \boldsymbol{x}) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)$

Inference interpretation = regularized & constrained minimization:

- $\ell_n(\boldsymbol{\theta}, \boldsymbol{x})$ = **loss** of $\boldsymbol{\theta}$ to minimize
- **D** = **divergence**, acting as uncertainty quantifier/regularizer
- $\boldsymbol{\Pi}$ = set of **admissible posterior** beliefs
- **Inference** = constrained, regularized optimization

  $\Rightarrow$ **Shorthand Notation:** $P(\ell_n, D, \boldsymbol{\Pi})$

2 The form of the Generalized Bayesian problem

Purpose of part 2: Investigate $P(\ell_n, D, \Pi)$

(1) Interpretations & modularity of ℓ_n, D and Π?

(2) Is there an axiomatic justification?

(3) Which existing methods does this (not) encompass?

2.1 Generalized Bayesian problem: provable modularity

$$q^*(\boldsymbol{\theta}) = \underset{q \in \Pi}{\arg\min} \left\{ \underbrace{\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\right]}_{\text{minimized by } \delta_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{\theta})} + \underbrace{D\left(q \| \pi\right)}_{\text{minimized by } q = \pi} \right\}$$

Roles of ℓ_n, D, $\boldsymbol{\Pi}$::

- ℓ_n: which parameter $\boldsymbol{\theta}$ do we care about?
- **D**: How is uncertainty quantified/what does q^* look like?
- $\boldsymbol{\Pi}$: Which beliefs are allowed?

  $\Rightarrow$ (provable) modularity of $P(\ell_n, D, \boldsymbol{\Pi})$!

Theorem 1 (GVI modularity)

For Bayesian inference with $P(\ell_n, D, \boldsymbol{\Pi})$, making it robust to model misspecification amounts to changing ℓ_n. Conversely, adapting uncertainty quantification (fixing $\boldsymbol{\Pi}$, π, $\boldsymbol{\theta}^*$, $\hat{\boldsymbol{\theta}}_n$) amounts to changing D.

2.2 Generalized Bayesian problem: Axiomatic derivation I/II

Axiom 1 (Representation)

The posterior $q^* \in \mathcal{P}(\boldsymbol{\Theta})$ is constructed by solving an optimization problem over some space $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$. The optimization seeks to minimize exactly two criteria that do not interact:

(i) The in-sample loss $\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)$ to be expected under $q^*(\boldsymbol{\theta})$.

(ii) The deviation from the prior $\pi(\boldsymbol{\theta})$ as measured by the magnitude of some statistical divergence D.

Axiom 2 (Information Equivalence)

Take any two constants $C, M \in \mathbb{R}_{>0}$ and let the posteriors $q_1^*, q_2^* \in \mathcal{P}(\boldsymbol{\Theta})$ be computed based on the same optimization problem, but with two different loss functions $\ell^{(1)}$ and $\ell^{(2)}$ as well as two different divergences $D^{(1)}, D^{(2)}$ so that $D^{(1)} = D^{(2)} + M$ and so that $\ell^{(1)} = \ell^{(2)} + C$. Then, $q_1^*(\boldsymbol{\theta}) = q_2^*(\boldsymbol{\theta})$ (almost everywhere).

2.2 Generalized Bayesian problem: Axiomatic derivation I/II

Axiom 3 (Generalized Likelihood Principle)

Suppose the posteriors $q_n^*, q_{n+m}^* \in \Pi$ are computed based on an optimization problem satisfying Axiom 1. Assume that the same optimization problem is used for both posteriors, except for a difference in the samples $x_{1:n}$ and $x_{1:n+m}$ and potentially different loss functions $\ell^{(1)}$ and $\ell^{(2)}$ that are used, respectively.

(i) Provided that there is an information difference between $x_{1:n}$ and $x_{1:n+m}$ for $m > 0$, the posteriors are different. In other words, $\sum_{i=1}^{n} \ell^{(1)}(\boldsymbol{\theta}, x_i) \neq \sum_{i=1}^{n+m} \ell^{(2)}(\boldsymbol{\theta}, x_i)$ implies that $q_n^* \neq q_{n+m}^*$, even if $\ell^{(1)} = \ell^{(2)}$.

(ii) Provided that there is a difference in the measure of information between $\ell^{(1)}$ and $\ell^{(2)}$, the posteriors are different. In other words, $\ell^{(1)} \neq \ell^{(2)}$ implies that $q_n^* \neq q_{n+m}^*$, even if $m = 0$ so that $x_{1:n} = x_{1:n+m}$.

2.2 Generalized Bayesian problem: Axiomatic derivation II/II

Theorem 2 (Form 1)

If Axiom 1 holds, $P(\ell_n, D, \Pi)$ has form $\arg\min_{q\in\Pi}\{L(q|\boldsymbol{x}, \ell_n, D)\}$ for $L(q|\boldsymbol{x}, \ell_n, D) = f\left(\mathbb{E}_{q(\boldsymbol{\theta})}[\boldsymbol{\ell}_n(\boldsymbol{\theta}, \boldsymbol{x})], D(q||\pi)\right)$, for some $f : \mathbb{R}^2 \to \mathbb{R}$.

Theorem 3 (Form 2)

For $P(\ell_n, D, \Pi)$ being $\arg\min_{q\in\Pi}\{L(q|\boldsymbol{x}, \ell_n, D)\}$ and $\circ$ an elementary operation on $\mathbb{R}$, $L(q|\boldsymbol{x}, \ell_n, D) = \mathbb{E}_{q(\boldsymbol{\theta})}[\ell_n(\boldsymbol{\theta}, \boldsymbol{x})] \circ D(q||\pi)$ satisfies Axioms 1, 2 and 3 only if $\circ = +$.

Implications/relevance:

- Bayesian inference = constrained, regularized optimization
- Objective only depends on $\mathbb{E}_{q(\boldsymbol{\theta})}[\boldsymbol{\ell}_n(\boldsymbol{\theta}, \boldsymbol{x})]$ and $D(q||\pi)$
- For elementary $f\left(\mathbb{E}_{q(\boldsymbol{\theta})}[\boldsymbol{\ell}_n(\boldsymbol{\theta}, \boldsymbol{x})], D(q||\pi)\right)$, f must be addition. (**Note:** Axiom 2 excludes most non-elementary f)

2.3 Generalized Bayesian problem & existing methods I/III

$$q^*(\boldsymbol{\theta}) = \underset{q \in \boldsymbol{\Pi}}{\arg\min} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[\ell_n(\boldsymbol{\theta}, \boldsymbol{x}) \right] + D\left(q || \pi\right) \right\}$$

$P(\boldsymbol{\ell_n}, D, \boldsymbol{\Pi})$ covers & gives **insight** into existing methods, e.g.

- **Power Bayes**: $P(w\ell_n, D, \Pi) = P(\ell_n, \frac{1}{w}D, \Pi)$.
  ($\Longrightarrow$ w-power likelihood $= \frac{1}{w}\times$ **more trust in your prior**.)
- **Regularized Bayes**: Adding $\Phi(q(\boldsymbol{\theta}, \boldsymbol{x})) = \mathbb{E}_{q(\boldsymbol{\theta}, \boldsymbol{x})}\left[\phi(\boldsymbol{\theta}, \boldsymbol{x})\right]$ into the objective corresponds to $P\left(\ell_n + \phi, D, \Pi\right)$.
  ($\Longrightarrow$ RegBayes $=$ a form of **GVI** that changes ℓ_n)
- **PAC Bayes**: Operationalizes a variational inference procedure based on PAC-bounds whose complexity penalty is some $D \neq \mathrm{KLD}$ (e.g. Alquier and Guedj, 2018)

2.3 Generalized Bayesian problem & existing methods II/III

Method	$\ell(\boldsymbol{\theta}, x_i)$	D	Π
Standard Bayes	$-\log(p(\boldsymbol{\theta}\|x_i))$	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Generalized Bayes[1]	any ℓ	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Power Bayes[2]	$-\log(p(\boldsymbol{\theta}\|x_i))$	$\frac{1}{w}\text{KLD}$, $w > 1$	$\mathcal{P}(\boldsymbol{\Theta})$
Divergence Bayes[3]	divergence-based ℓ	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Standard VI	$\boldsymbol{-\log(p(\theta\|x_i))}$	**KLD**	$\mathcal{Q}$
Power VI[4]	$-\log(p(\boldsymbol{\theta}\|x_i))$	$\frac{1}{w}\text{KLD}$, $w > 1$	$\mathcal{Q}$
Regularized Bayes[5]	$-\log(p(\boldsymbol{\theta}\|x_i)) + \phi(\boldsymbol{\theta}, x_i)$	KLD	$\mathcal{Q}$
Gibbs VI[6]	any ℓ	KLD	$\mathcal{Q}$
Generalized VI	**any** $\boldsymbol{\ell}$	**any** $\boldsymbol{D}$	$\mathcal{Q}$

Table 1 – $P(\ell_n, D, Q)$ & existing methods. [1](Bissiri et al., 2016), [2](e.g. Holmes and Walker, 2017; Grünwald et al., 2017; Miller and Dunson, 2018), [3](e.g. Hooker and Vidyashankar, 2014; Ghosh and Basu, 2016; Futami et al., 2017; Jewson et al., 2018), [4](e.g. Yang et al., 2017; Huang et al., 2018) [5](Ganchev et al., 2010; Zhu et al., 2014), [6](Alquier et al., 2016; Futami et al., 2017)

2.3 Generalized Bayesian problem & existing methods II/III

$$
\begin{tikzcd}[column sep=huge, row sep=large]
\begin{matrix}\text{Rule of Three} \\ P(\ell, D, \Pi)\end{matrix} \arrow[r, "\begin{matrix} D = \mathrm{KLD} \\ \Pi = \mathcal{P}(\Theta) \end{matrix}"] \arrow[d, "\Pi = \mathcal{Q} \subset \mathcal{P}(\Theta)"] & \begin{matrix}\text{PAC/Gibbs Bayes} \\ P(\ell, \mathrm{KLD}, \mathcal{P}(\Theta))\end{matrix} \arrow[r, "\ell = -\log p(\cdot|\theta)"] \arrow[d, "\Pi = \mathcal{Q} \subset \mathcal{P}(\Theta)"] & \begin{matrix}\text{Standard Bayes} \\ P(-\log p(\cdot|\theta), \mathrm{KLD}, \mathcal{P}(\Theta))\end{matrix} \arrow[d, "\Pi = \mathcal{Q} \subset \mathcal{P}(\Theta)"] \\
\begin{matrix}\textbf{Generalized VI} \\ P(\ell, D, \mathcal{Q})\end{matrix} \arrow[r, "D = \mathrm{KLD}"'] & \begin{matrix}\text{Gibbs VI} \\ P(\ell, \mathrm{KLD}, \mathcal{Q})\end{matrix} \arrow[r, "\ell = -\log p(\cdot|\theta)"'] & \begin{matrix}\textbf{Standard VI} \\ P(-\log p(\cdot|\theta), \mathrm{KLD}, \mathcal{Q})\end{matrix}
\end{tikzcd}
$$

Figure 1 – A taxonomy of some important belief distributions as special cases.

2.3 Generalized Bayesian problem & existing methods III/III

Not everything fits $P(\ell_n, D, \mathbf{\Pi})$:

(1) **Laplace approximations** (e.g., INLA)

(2) **Divergence-Variational inference (DVI)**: VI based on discrepancy $D \neq \text{KLD}$ (locally) solving $q^* = \arg\min_{q \in \mathcal{Q}} D(q \| \widetilde{q})$ for $\widetilde{q} =$ standard Bayesian posterior, e.g.

$D =$ Rényi's α-divergence (Li and Turner, 2016; Saha et al., 2017)

$D =$ χ-divergence (Dieng et al., 2017)

$D =$ operators (Ranganath et al., 2016)

$D =$ scaled AB-divergence (Regli and Silva, 2018)

$D =$ Wasserstein distance (Ambrogioni et al., 2018)

...

(3) **Expectation Propagation (EP)** (Minka, 2001; Opper and Winther, 2000) and its variants (e.g. Hernández-Lobato et al., 2016).
Note: Particular type of **DVI**, with $D =$ (local) reverse KLD

3 Reinterpreting VI

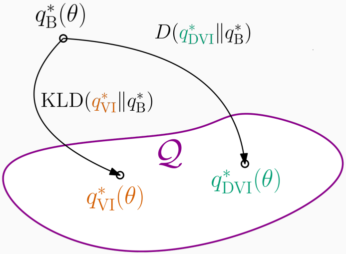
Purpose of part 3: Motivating **GVI**

(1) **Standard VI**: Optimality & reinterpretation

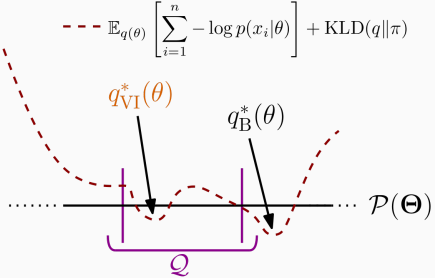
(2) **DVI**: "suboptimal" methods with better posteriors

(3) **GVI**: A modular alternative to **DVI**

Relationship between VI, GVI, DVI and exact inference?



(a) DVI (projection-centric) interpretation of **VI**



(b) GVI (= optimization-centric) Interpretation of **VI**

3.1 Optimality & reinterpretation of standard VI II/V

Standard VI: $q^*_{\mathrm{VI}} = \arg\min_{q \in \mathcal{Q}} \mathrm{KLD}(q \| q^*_B)$, q^*_B solves $P(\ell_n, \mathbf{KLD}, \mathcal{P}(\Theta))$

$$\mathrm{KLD}(q||q^*_B) =$$

$$\underbrace{\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log\left(\frac{q(\boldsymbol{\theta})}{\exp\left\{-\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\right\}\pi(\boldsymbol{\theta})}\right)\right]}_{\text{(Generalized) ELBO}} + \underbrace{\log\left(\int_{\boldsymbol{\theta}} \exp\left\{-\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\right\}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}\right)}_{\text{Generalized 'log evidence'}}$$

Inference = minimizing ELBO, which you can rewrite as

$$\mathrm{ELBO}(q) = \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\right] + \mathrm{KLD}(q||\pi). \tag{2}$$

...which is exactly the objective of $P(\ell_n, \mathbf{KLD}, \mathcal{Q})$

In other words, $P(\ell_n, \mathbf{KLD}, \mathcal{Q})$ $(= \mathrm{ELBO})$ is

$$q^*_{\mathrm{VI}}(\boldsymbol{\theta}) = \underset{q \in \mathcal{Q}}{\arg\min} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[\ell_n(\boldsymbol{\theta}, \boldsymbol{x}) \right] + \mathrm{KLD}\left(q || \pi \right) \right\},$$

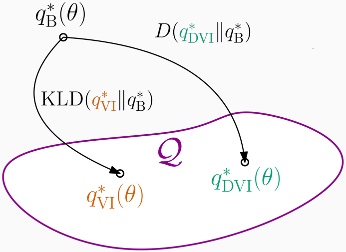
the $\mathcal{Q}$-constrained relaxation of $P(\ell_n, \mathbf{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$, whose objective is

$$q^*_B(\boldsymbol{\theta}) = \underset{q \in \mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[\ell_n(\boldsymbol{\theta}, \boldsymbol{x}) \right] + \mathrm{KLD}\left(q || \pi \right) \right\},$$

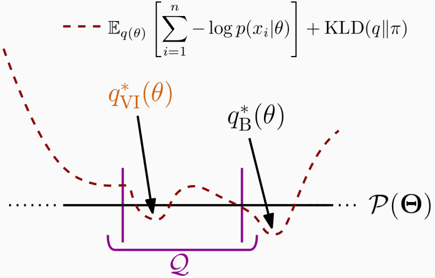
(which is the exact Bayesian objective).

$\Rightarrow$ Reinterpretation of **standard VI** as **Constrained optimization**!

3.1 Optimality & reinterpretation of standard VI IV/V



(a) **DVI** (projection-centric) interpretation of **VI**



(b) **GVI** (= optimization-centric) Interpretation of **VI**

Consequence I/II: VI-optimality

Theorem 4 (VI optimality)

For exact and coherent Bayesian posteriors solving $P(\ell_n, \mathrm{KLD}, \mathcal{P}(\Theta))$ and a fixed variational family $\mathcal{Q}$, standard **VI** produces the uniquely optimal $\mathcal{Q}$-constrained approximation to $P(\ell_n, \mathrm{KLD}, \mathcal{P}(\Theta))$ Having decided on approximating the Bayesian posterior with some $q \in \mathcal{Q}$, **VI** provides the uniquely optimal solution.

Consequences II/II: **DVI**-suboptimality. **Three** big disadvantages:

(1) If $F \neq \mathrm{KLD}$, **DVI** violates Axioms 1–3.

(2) **DVI** conflates ℓ_n and D (i.e., modularity of $P(\ell_n, D, \mathbf{\Pi})$ lost).

(3) Last Thm: **DVI** gives worse $\mathcal{Q}$-constrained posterior than standard **VI** (*relative to the standard Bayesian problem* $P(\ell_n, \mathrm{KLD}, \mathcal{P}(\mathbf{\Theta}))$)

Objection! **DVI** often produces better posteriors than **standard VI**!

3.2 Why does DVI produce better posteriors? II/II

Seeming contradiction:

(1) **VI** is the best approximation to the *standard* Bayesian posterior

(2) **DVI** often outperforms **VI** (e.g., on test scores)

Resolution:

> **DVI** outperforms **VI** by implicitly **defining a new, non-standard Bayesian problem***

$\Longrightarrow$ Inspires **Generalized Variational Inference (GVI)**

***non-standard Bayesian problem** = problem defining a belief distribution that can be understood as updating a prior belief with some data, but which is **not** obtained by Bayes' rule

3.3 Towards GVI I/II

GVI = combining advantages of **VI** and **DVI**:

(1) Has form $P(\ell_n, D, \mathcal{Q})$ Like **VI** i.e.

  (i) satisfies Axioms 1–3;

  (ii) provably interpretable modularity (**loss, uncertainty quantifier, admissible posteriors**

(2) Derives different & more appropriate posteriors like **DVI** but

  (i) without conflating ℓ_n and D

  (ii) with **explicit** rather than **implicit** changes .

Definition 1 (GVI)

Any Bayesian inference method solving $P(\ell_n, D, \mathcal{Q})$ with admissible choices ℓ_n, D and $\mathcal{Q}$ is a Generalized Variational Inference (**GVI**) method (and satisfies Axioms 1–3 by definition).

Illustration: DVI aims for D, but changes ℓ_n – GVI doesn't

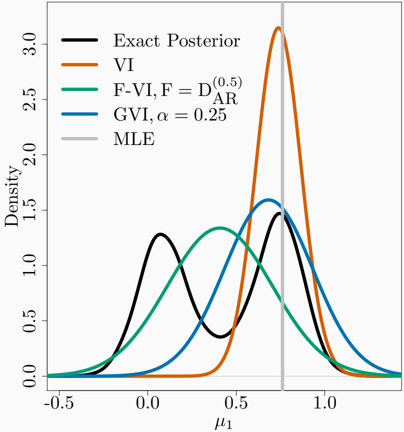


Figure 4 – **Exact**, VI, DVI ($F = D_{AR}^{(0.5)}$) and $P(\ell_n, D_{AR}^{(\alpha)}, \mathcal{Q})$ based GVI marginals of the location in a 2 component mixture model. Respecting ℓ_n, VI and GVI provide uncertainty quantification around the most likely value $\widehat{\boldsymbol{\theta}}_n$ via D. In contrast, DVI implicitly changes the loss and has a mode at the locally most *unlikely* value of $\boldsymbol{\theta}$.

4 GVI: What does it do?

Purpose of part 4: Exploring **GVI**'s relationship to M-open world & study three use cases

(1) Embed **GVI** into the M-open world

(2) Robust alternatives to $\ell(\boldsymbol{\theta}, x_i) = -\log(p(x_i|\boldsymbol{\theta}))$

(3) Prior-robust uncertainty quantification and adjusting marginal variances via D

4.1 M-open vs M-closed inference

M-closed view: Model and prior are correct, i.e.

$$\exists \boldsymbol{\theta}^* \in \boldsymbol{\Theta} \text{ s.t. } x_i \sim p(x_i|\boldsymbol{\theta}^*)$$

Q: Why do purist Bayesians love **exact** inference (MCMC, SMC, ...)?
A: With M-closed view, can focus on computation of Bayes posterior

$$q^*(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \quad = \text{solution of } P\left(\sum_{i=1}^{n} -\log p(x_i|\boldsymbol{\theta}), D, \mathcal{P}(\boldsymbol{\Theta})\right)$$

M-open view: Model and prior are **not** correct, i.e.

$$\nexists \boldsymbol{\theta}^* \in \boldsymbol{\Theta} \text{ s.t. } x_i \sim p(x_i|\boldsymbol{\theta}^*)$$

Traditional stats: Devise better models until they are 'close enough'
$\Longrightarrow$ Focus on inference still useful!
Modern stats/ML: Use some black box model (BNN, DGP, ...)
$\Longrightarrow$ Do we even want standard posterior!?

4.1 Consequences for inference

Conclusion 1: If your model is pretty good, use exact inference or **VI**.
Conclusion 2: If **DVI** works better than **VI**, your model is not great.
Conclusion 3: If you know why model isn't great, address it with **GVI**.

M-closed assumptions are very far from the truth sometimes:

- Time Series/on-line inference:
  - (i) Stationarity
  - (ii) Short-term memory
  - (iii) Noise level constant/non-erratic
- Bayesian Neural Networks (BNNs):
  - (i) There is a true weight parameter $\boldsymbol{\theta}^*$ indexing the network
  - (ii) All priors on the thousands of entries of $\boldsymbol{\theta}$ is well-specified
- (Deep) Gaussian Processes (DGPs):
  - (i) (Conditionally on latent space) correct likelihood is specified
  - (ii) DGP prior and kernel choice generate correct latent spaces

$\Longrightarrow$ Clearly, **M-open** world more appropriate $\Longrightarrow$ **GVI**

GVI modularity: The loss ℓ_n

Q1: Why use $\ell_n(\boldsymbol{\theta}, \boldsymbol{x}) = \sum_{i=1}^n -\log(p(x_i|\boldsymbol{\theta}))$?
A: Assuming that the true data-generating mechanism is $\boldsymbol{x} \sim g$,

$$\arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} -\log(p(x_i|\boldsymbol{\theta})) \approx \arg\min_{\boldsymbol{\theta}} \mathbb{E}_g\left[-\log(p(\boldsymbol{x}|\boldsymbol{\theta}))\right]$$
$$= \arg\min_{\boldsymbol{\theta}} \mathbb{E}_g\left[-\log\left(p(\boldsymbol{x}|\boldsymbol{\theta})\right) + \log(g(\boldsymbol{x}))\right] = \arg\min_{\boldsymbol{\theta}} \mathrm{KLD}(g\|p(\cdot|\boldsymbol{\theta}))$$

Interpretation: $\boldsymbol{-\log\left(p(x_i|\theta)\right)}$ = targeting KLD-minimizing $p(\cdot|\boldsymbol{\theta})$

Q2: Are there other $\boldsymbol{\mathcal{L}^D(p(x_i|\theta))}$ for divergence $\boldsymbol{D}$?
A: Yes! (e.g. Jewson et al., 2018; Futami et al., 2017; Ghosh and Basu, 2016; Hooker and Vidyashankar, 2014)

4.2 GVI: The losses II/III

Q3: Why use other $\mathcal{L}^{D}(p(x_i|\theta))$?
A: Robustness (for D = a robust divergence) [log/KLD non-robust!]

Robustness recipe: $\alpha/\beta/\gamma$-divergences using generalized log functions
E.g.: β indexes β-divergence ($D_B^{(\beta)}$) via

$$\log_\beta(x) = \frac{1}{(\beta-1)\beta}\left[\beta x^{\beta-1} - (\beta-1)x^{\beta}\right]$$
$$D_B^{(\beta)}(g||p(\cdot|\boldsymbol{\theta})) = \mathbb{E}_g\left[\log_\beta(p(\boldsymbol{x}|\boldsymbol{\theta})) - \log_\beta(g(\boldsymbol{x}))\right]$$

Note 1: $D_B^{(\beta)} \to \mathrm{KLD}$ as $\beta \to 1$!
Note 2: Admits $D_B^{(\beta)}$-targeting loss as

$$\mathcal{L}_p^{\beta}(\theta, x_i) = -\frac{1}{\beta-1}p(x_i|\boldsymbol{\theta})^{\beta-1} + \frac{I_{p,\beta}(\boldsymbol{\theta})}{\beta}, \quad I_{p,c}(\boldsymbol{\theta}) = \int p(x|\boldsymbol{\theta})^c dx$$

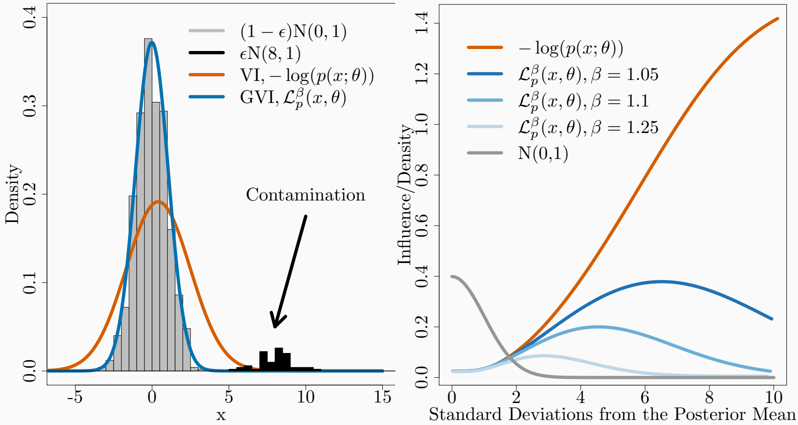


Figure 5 – Left: Robustness against model misspecification. Depicted are posterior predictives under $\varepsilon = 5\%$ outlier contamination using **VI** and $P(\sum_{i=1}^n \mathcal{L}_p^\beta(\boldsymbol{\theta}, x_i), \mathrm{KLD}, \mathcal{Q})$, $\beta = 1.5$. **Right:** From Knoblauch et al. (2018). Influence of x_i on exact posteriors for different losses.

GVI modularity: The uncertainty quantifier D

Q: Which VI drawbacks can be addressed via D?

A: Any **uncertainty quantification** properties, e.g.

- Over-concentration (= underestimating marginal variances)
- Sensitivity to badly specified priors
- ...

Example 1: GVI can fix over-concentrated posteriors

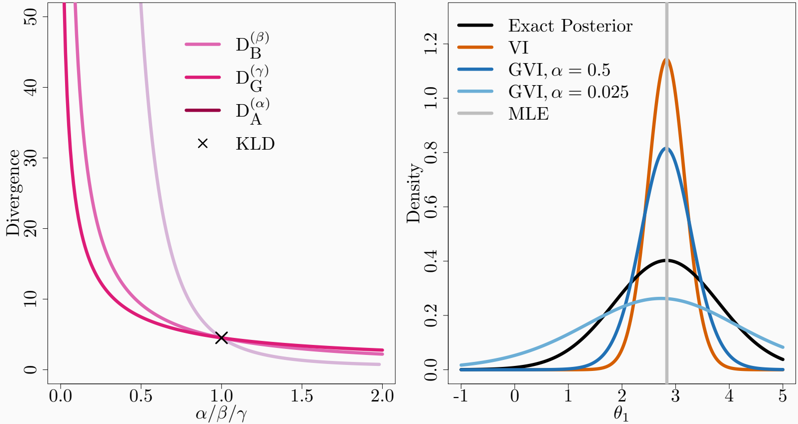


Figure 6 – Left: Magnitude of the penalty incurred by $D(q||\pi)$ for different uncertainty quantifiers D and fixed densities π, q. **Right**: Using $D_{AR}^{(\alpha)}$ with different choices of α to "customize" uncertainty.

Example 2: Avoiding prior sensitivity

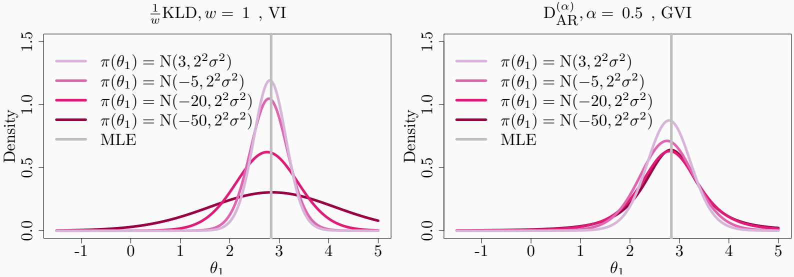


Figure 7 – Prior sensitivity with **VI** (**left**) vs. prior robustness with **GVI** (**right**). Priors are more badly specified for **darker** shades.

4.4 GVI: Three use cases

Summary: **GVI** is natural in the M-open (i.e. real) world. Applications include

(1) Robustness to model misspecification ($=$ adapting ℓ_n)

(2) "Customized" marginal variances ($=$ adapting D)

(3) Prior robustness ($=$ adapting D)

Purpose of part 5: GVI inference & experiments

(1) How/when can we "black box" **GVI**?

(2) A case study in robustness with Bayesian On-line Changepoint Detection

(3) **DVI** vs **GVI** & changes in D (on Bayesian Neural Nets)

(4) **VI** vs **GVI** & changes in ℓ_n (on Deep Gaussian Processes)

Setup: $\mathcal{Q} = \{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) : \boldsymbol{\kappa} \in K\}$ variational family s.t.

(i) one can sample $\boldsymbol{\theta}^{(1:S)} \sim q(\boldsymbol{\theta}|\boldsymbol{\kappa})$;

(ii) derivative $\nabla_{\boldsymbol{\kappa}} \log(q(\boldsymbol{\theta}|\boldsymbol{\kappa}))$ exists.

Case 1: Closed form for $\nabla_{\boldsymbol{\kappa}} D(q||\pi)$ $\rightarrow$ unbiased estimate:

$$\nabla_{\boldsymbol{\kappa}} \hat{L}(q|\ell_n, D) = \frac{1}{S} \sum_{s=1}^{S} \left\{ \ell_n(\boldsymbol{\theta}^{(s)}, \boldsymbol{x}) \cdot \nabla_{\boldsymbol{\kappa}} \log(q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa})) \right\} + \nabla_{\boldsymbol{\kappa}} D(q||\pi)$$

Thm. 7: Closed forms for most $\alpha/\beta/\gamma$- & Rényi-divergence.

Case 2: $D(q||\pi) = \mathbb{E}_q[\ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta})]$ (e.g., f-divs) $\rightarrow$ unbiased estimate:

$$\nabla_{\boldsymbol{\kappa}} \hat{L}(q|\ell_n, D) = \frac{1}{S} \sum_{s=1}^{S} \left\{ \left[\ell_n(\boldsymbol{\theta}^{(s)}, \boldsymbol{x}) + \ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)}) \right] \cdot \nabla_{\boldsymbol{\kappa}} \log(q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa})) + \nabla_{\boldsymbol{\kappa}} \ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)}) \right\}.$$

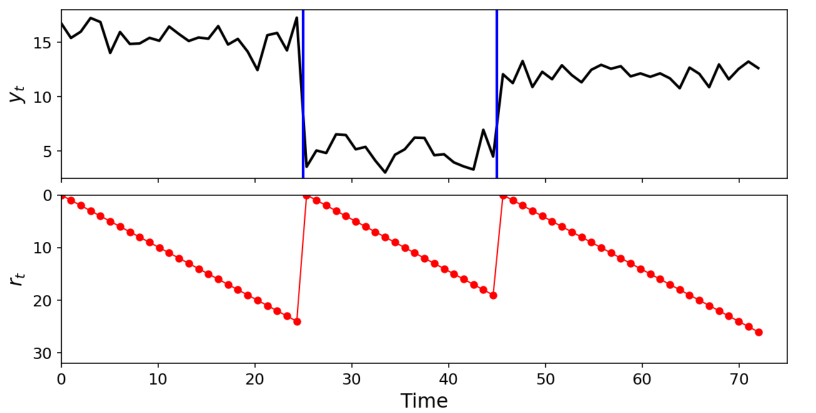
5.2 Standard BOCPD I/XI

Idea due to Adams and MacKay (2007) and Fearnhead and Liu (2007):

(1) Define **Run-length at** $t = r_t \iff$ there was a CP at time $t - r_t$.

(2) **Inference on last CP** via $p(r_t|y_{1:t})$ rather than on *all* CPs

(3) Resulting complexity: $\mathcal{O}(t)$ **rather than** $\mathcal{O}(\prod_{i=1}^{t} i)$.



5.2 BOCPD + model selection (Knoblauch and Damoulas, 2018) II/XI

Idea: Multiple models, different between segments
New Random Variable: m_t, the model at time t

$$r_t|r_{t-1} \sim H(r_t, r_{t-1}) \quad \text{[conditional \textbf{CP} prior]} \tag{3a}$$

$$m_t|m_{t-1}, r_t \sim q(m_t|m_{t-1}, r_t) \quad \text{[conditional \textbf{model} prior]} \tag{3b}$$

$$\boldsymbol{\theta}_m|m_t \sim \pi_{m_t}(\boldsymbol{\theta}_{m_t}) \quad \text{[parameter prior]} \tag{3c}$$

$$\boldsymbol{y}_t|m_t, \boldsymbol{\theta}_{m_t} \sim f_{m_t}(\boldsymbol{y}_t|\boldsymbol{\theta}_{m_t}) \quad \text{[\textbf{observation} density]} \tag{3d}$$

where $q(m_t|m_{t-1}, r_t) = \mathbb{1}_{\{r_t>0\}}\delta(m_{t-1}) + \mathbb{1}_{\{r_t=0\}}q(m_t)$.

Recursion:

$$p(\boldsymbol{y}_1, r_1 = 0, m_1) = q(m_1)\int_{\Theta_{m_1}} f_{m_1}(\boldsymbol{y}_1|\boldsymbol{\theta}_{m_1})\pi_{m_1}(\boldsymbol{\theta}_{m_1})d\boldsymbol{\theta}_{m_1} = q(m_1)f_{m_1}(\boldsymbol{y}_1|\boldsymbol{y}_0)$$

$$p(\boldsymbol{y}_{1:t}, r_t, m_t) = \sum_{m_{t-1}, r_{t-1}} \Big\{ f_{m_t}(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}, r_{t-1})q(m_t|\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) H(r_t, r_{t-1})p(\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) \Big\}$$

5.2 Standard BOCPD with model selection III/XI

$$p(\boldsymbol{y}_{1:t}, r_t, m_t) = \sum_{m_{t-1}, r_{t-1}} \Big\{ f_{m_t}(\boldsymbol{y}_t | \boldsymbol{y}_{1:(t-1)}, r_{t-1}) q(m_t | \boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) \\ H(r_t, r_{t-1}) p(\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) \Big\}$$

Inference:

(1) Evidence: $p(\boldsymbol{y}_{1:t}) = \sum_{r_t, m_t} p(\boldsymbol{y}_{1:t}, r_t, m_t)$

(2) run-length & model posterior: $p(r_t, m_t | \boldsymbol{y}_{1:t}) = p(\boldsymbol{y}_{1:t}, r_t, m_t) / p(\boldsymbol{y}_{1:t})$

(3) Prediction: $p(\boldsymbol{y}_{t+1} | \boldsymbol{y}_{1:t}) = \sum_{r_t, m_t} f_{m_t}(\boldsymbol{y}_{t+1} | \boldsymbol{y}_{1:t}, r_t) p(r_t, m_t | \boldsymbol{y}_{1:t})$

(4) Run-length marginal posterior: $p(r_t | \boldsymbol{y}_{1:t}) = \sum_{m_t} p(r_t, m_t | \boldsymbol{y}_{1:t})$

(5) Model marginal posterior: $p(m_t | \boldsymbol{y}_{1:t}) = \sum_{r_t} p(r_t, m_t | \boldsymbol{y}_{1:t})$.

(6) MAP segmentation:
$MAP_t = \max_{r,t} \{ MAP_{t-r-1} \cdot p(r_t = r, m_t = m | \boldsymbol{y}_{1:t}) \}$

Last plot: Clear that model selection non-robust! Why?

- On-line processing
- Moderate/high dimensions for $\boldsymbol{y}_t$

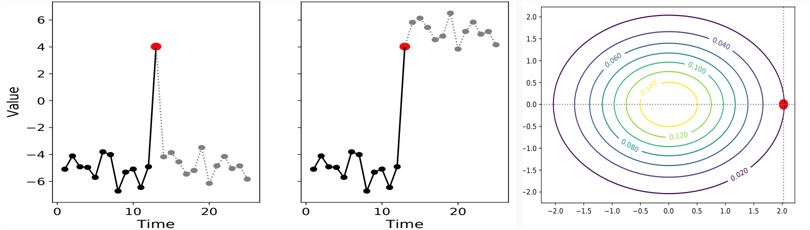
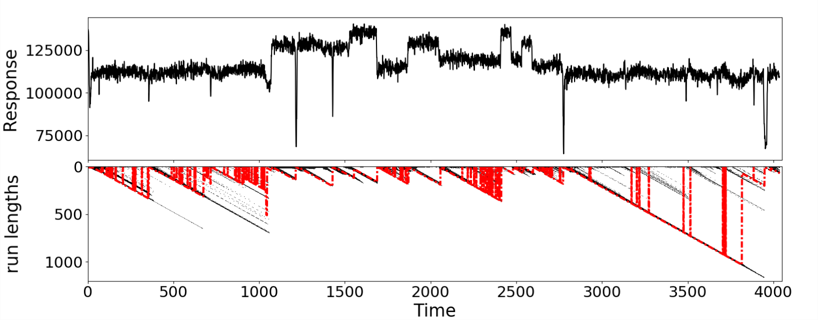


Figure 8 – Left, Center: Price for on-line processing is that outliers are confused with changepoings. **Right:** Multivariate densities become very small even if outliers occur only in a single dimension.

5.2 Issue: Outliers & misspecification V/X



Five Autoregressive processes with two CPs

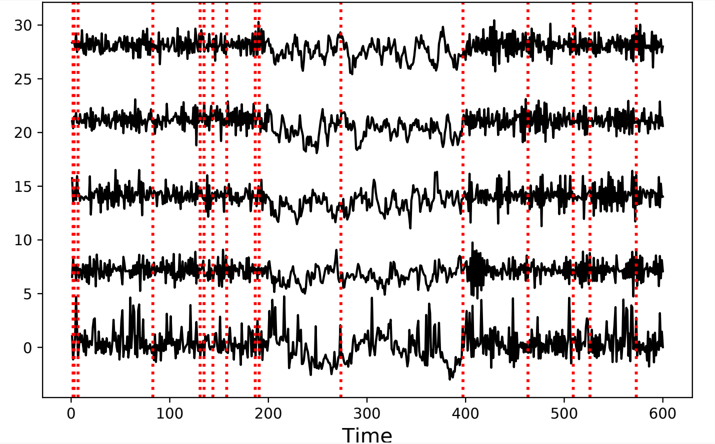
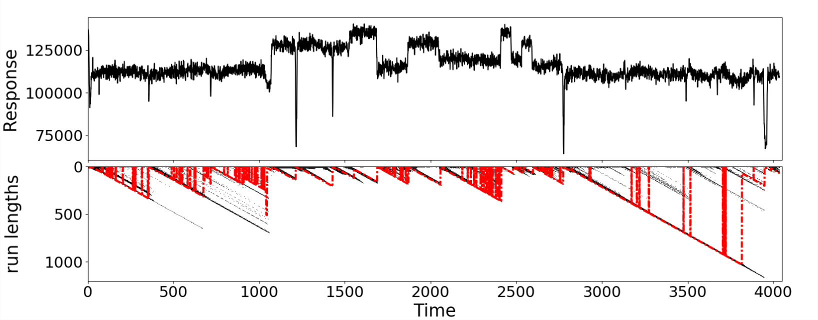


Figure 9 – Maximum A Posteriori (MAP) CPs of standard BOCPD shown as dashed vertical lines. True CPs at $t = 200, 400$.

5.2 Fix: Adapt loss (Knoblauch et al., 2018) VII/XI



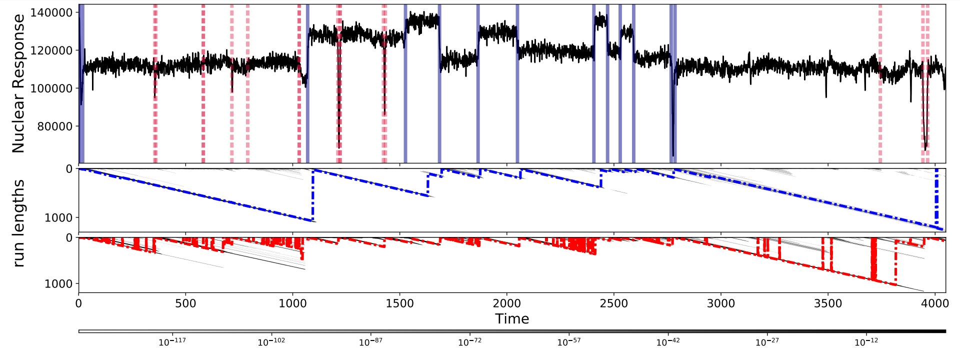


Figure 10 – Robust segmentation and run-length distribution and **additionally found CPs with non-robust run-length distribution**

[$\text{FDR}: > 99\% \Longrightarrow 8\%$ and reduction in MSE (MAE) by 10% (6%)]

Five Autoregressive processes with two CPs

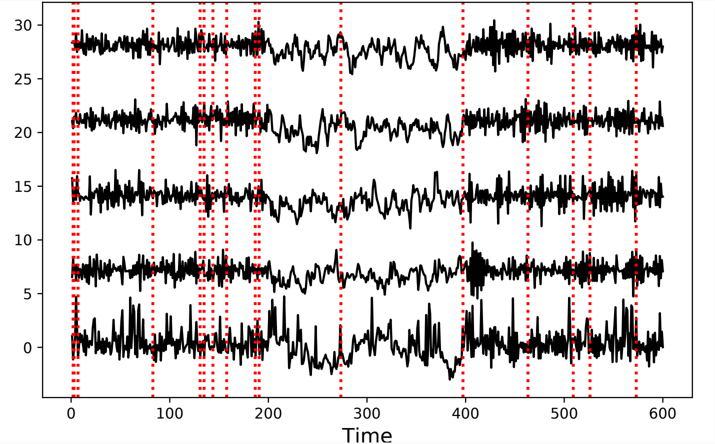


Figure 11 – Maximum A Posteriori (MAP) CPs of standard BOCPD shown as dashed vertical lines. True CPs at $t = 200, 400$.

Five Autoregressive processes with two CPs

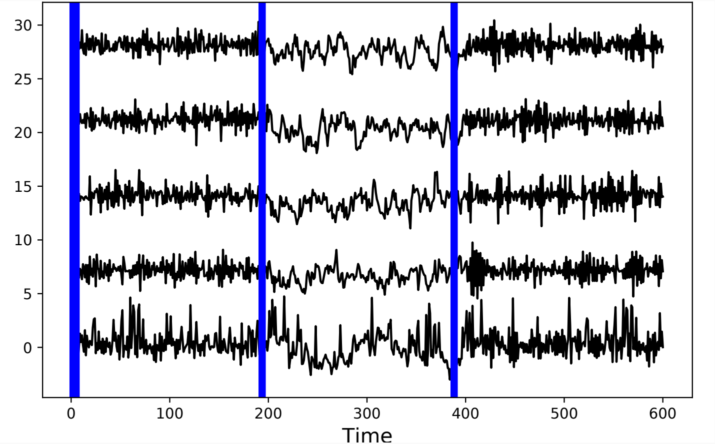


Figure 12 – Maximum A Posteriori (MAP) CPs of **robust** BOCPD shown as solid vertical lines. True CPs at $t = 200, 400$.

5.2 Fix: Robustness by adapting the loss XI/XI

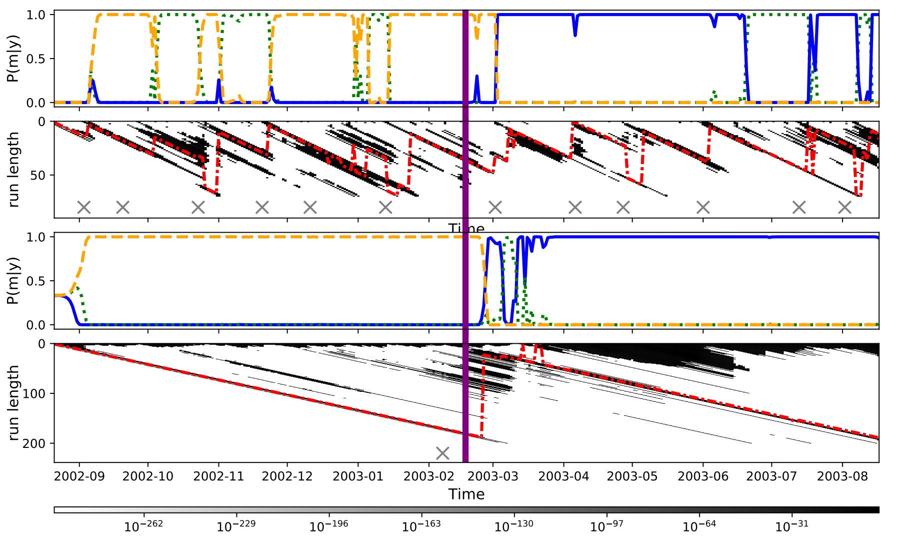


Figure 13 – Top & bottom two panels: standard & robust BOCPD.

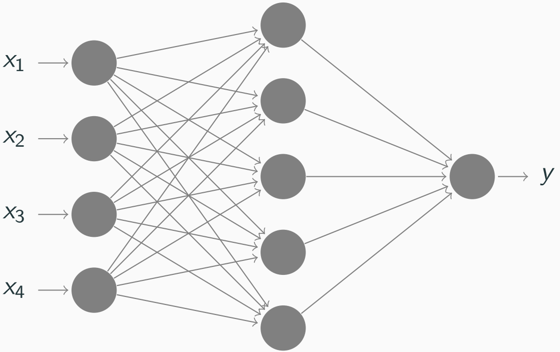
5.3 Experiments with Bayesian Neural Nets (BNNs) I/IV

BNNs are intractable Bayesian regression models with

$$y|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{y}; F_{\boldsymbol{\theta}}(\boldsymbol{x}), \sigma^2),$$

with $F_{\boldsymbol{\theta}}(\boldsymbol{x})$ defining a non-linear transform of $\boldsymbol{x}$ parameterized by $\boldsymbol{\theta}$.
(**Note:** Our experiments use one hidden layer with 50 ReLu neurons.)

$F_{\boldsymbol{\theta}}(\boldsymbol{x})$



5.3 Experiments with Bayesian Neural Nets (BNNs) II/IV

Methods: Comparison of black box approximate Bayesian methods:

- **VI**
- **DVI** based on $F = D_{AR}^{(\alpha)}$ (Li and Turner, 2016)
- **DVI** based on $F = D_{A}^{(\alpha)}$ (Hernández-Lobato et al., 2016)
- **GVI** with $D = D_{AR}^{(\alpha)}$.

Note: Everything run with settings of Li and Turner (2016) and Hernández-Lobato et al. (2016)

- Variational family $\mathcal{Q}$: A fully factorized normal
- Optimization of σ^2 (i.e., point estimation akin to type-II ML)
- ADAM (Kingma and Ba, 2014) with default settings and 500 epochs
- 50 Random splits with 90:10 training:test ratio
- benchmark UCI (Lichman et al., 2013) datasets

...

5.3 Experiments with Bayesian Neural Nets (BNNs) III/IV

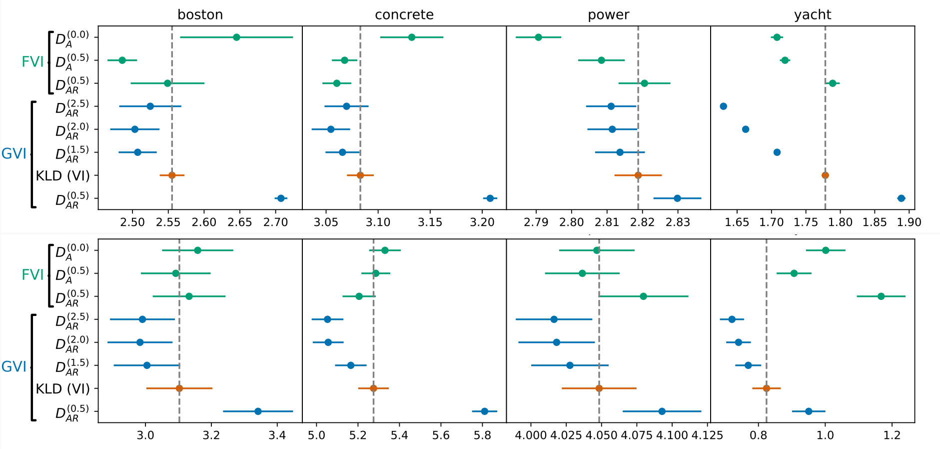


Figure 14 – Performance on BNNs: **DVI**, **GVI** with $D = D_{AR}^{(\alpha)}$, and **VI**. **Top**: Negative test log likelihoods. **Bottom row**: Test RMSE.

Observation: **GVI** outperforms **VI** for **over-concentrated** posteriors (i.e. $\alpha > 1$)! So how does **under-concentrated** **DVI** outperform **VI**?!?

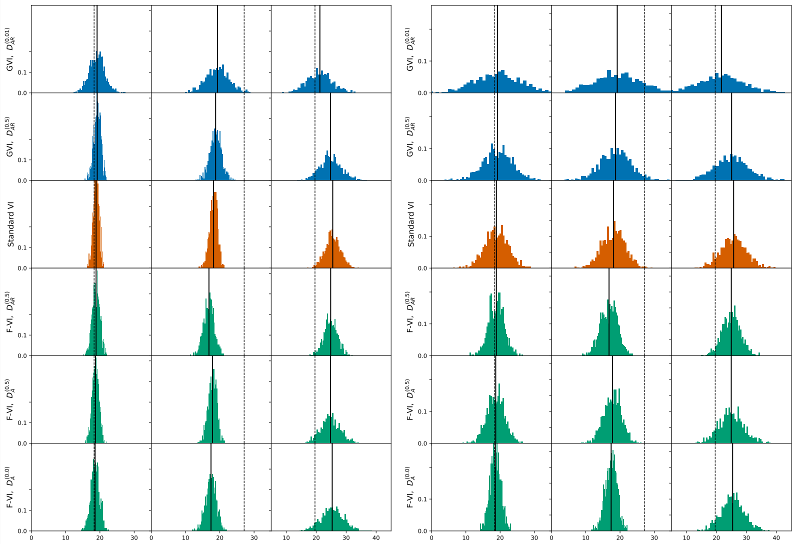


Figure 15 – Left: Parameter posteriors (**DVI** as expected). **Right:** Posterior predictives (**DVI not** as expected)

5.3 Experiments with Bayesian Neural Nets (BNNs) V/V

Q: Why does this happen for **DVI** and not for **GVI**?!

A: **DVI** does not distinguish uncertainty quantification & loss!

DVI objective: σ^2 affects **target (!)**

$$\widehat{\sigma}^2, q^*(\boldsymbol{\theta}|\widehat{\sigma}^2, \boldsymbol{\kappa}) = \arg\min_{\sigma^2} \left\{ \arg\min_{q \in \mathcal{Q}} F\left(q(\boldsymbol{\theta}|\sigma^2, \boldsymbol{\kappa}) \| \underbrace{\widetilde{q}(\boldsymbol{\theta}|\sigma^2, \boldsymbol{x}, \boldsymbol{y})}_{\text{i.e., } \widetilde{q} = \widetilde{q}^{\sigma}} \right) \right\}$$

$\Rightarrow$ optimizing for $\sigma^2 =$ **changing the target $\widetilde{\boldsymbol{q}}^{\boldsymbol{\sigma}}$**!

GVI objective: σ^2 indexes the **loss** only

$$\widehat{\sigma}^2, q^*(\boldsymbol{\theta}|\widehat{\sigma}^2, \boldsymbol{x}, \boldsymbol{y}) = \arg\min_{\sigma^2} \left\{ \arg\min_{q \in \mathcal{Q}} \left\{ \mathbb{E}_q \left[\underbrace{\boldsymbol{\ell_n(\theta, x|y, \sigma^2)}}_{\text{i.e., } \ell_n = \ell_n^{\sigma}} \right] + D(q\|\pi) \right\} \right\}$$

$\Rightarrow$ optimizing for $\sigma^2 =$ finding **optimal loss $\boldsymbol{\ell_n^{\sigma}}$**

5.4 Experiments with Deep Gaussian Processes (DGPs) I/II

Principal idea: Use the BNN architecture with GP priors on $F_{\boldsymbol{\theta}}(\cdot)$:

$$
\begin{aligned}
y|\boldsymbol{F}^L &\sim p\left(y\middle|\ \boldsymbol{F}^L\right) \\
\boldsymbol{F}^L|\boldsymbol{F}^{L-1} &\sim \mathrm{GP}\left(\mu^L(\boldsymbol{F}^{L-1}), \mathsf{K}^L(\boldsymbol{F}^{L-1}, \boldsymbol{F}^{L-1})\right) \\
\boldsymbol{F}^{L-1}|\boldsymbol{F}^{L-2} &\sim \mathrm{GP}\left(\mu^{L-1}(\boldsymbol{F}^{L-2}), \mathsf{K}^{L-1}(\boldsymbol{F}^{L-2}, \boldsymbol{F}^{L-2})\right) \\
&\dots \\
\boldsymbol{F}^1|\boldsymbol{x} &\sim \mathrm{GP}\left(\mu^1(\boldsymbol{x}), \mathsf{K}^1(\boldsymbol{x}, \boldsymbol{x})\right),
\end{aligned}
$$

Methods: Comparison of black box approximate Bayesian methods:

- State of the art **VI** (Salimbeni and Deisenroth, 2017) (comprehensively beat competing **DVI** methods (Bui et al., 2016))
- **GVI** with $\ell_n = \sum_{i=1}^n \mathcal{L}_p^{\gamma}(\boldsymbol{\theta}, x_i)$.

Note: Everything run with settings of Salimbeni and Deisenroth (2017)

[Derivations for **DGP-GVI**: `https://arxiv.org/abs/1904.02303`.]

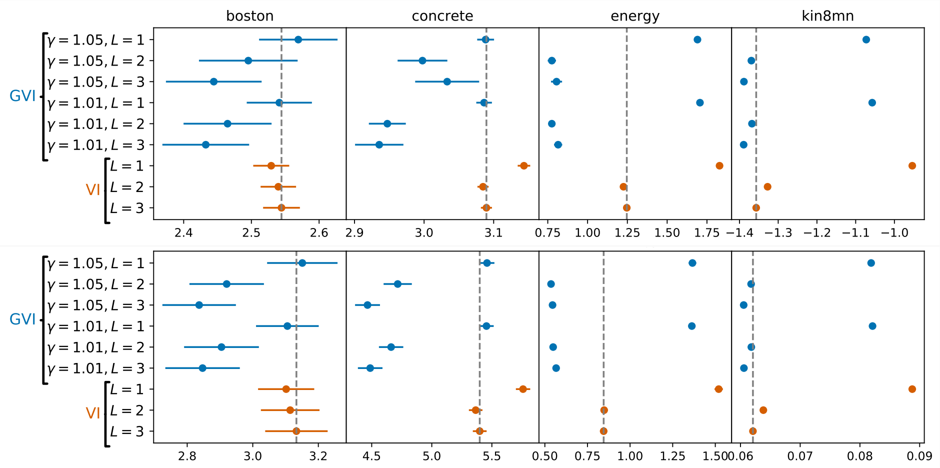


Figure 16 – DGP performance with L layers, **GVI** with & **VI**. **Top row**: Negative test log likelihoods. **Bottom row**: Test RMSE.

Summary & Conclusion

Summary:

Part 1: Ways to look at Bayesian inference: belief updates (about arbitrary parameters) & optimization over $\mathcal{P}(\boldsymbol{\Theta})$

Part 2: Bayesian inference as a modular & interpretable triplet $P(\ell_n, D, \boldsymbol{\Pi})$: **loss**, **uncertainty quantifier** & **admissible posteriors**.

Part 3: Fallout of $P(\ell_n, D, \boldsymbol{\Pi})$: **VI** optimality & **DVI** suboptimality $\rightarrow$ **GVI**

Part 4: Some of **GVI**'s use cases: Robust losses, alternative ways of quantifying uncertainty. Also: its upper bound interpretation

Part 5: Black box methods with **GVI** & empirical performance.

Main Conclusions:

(I) **GVI**: **principled & explicit** design of $\mathcal{Q}$-constrained posteriors

(II) **GVI**: **tackles drawbacks of VI** (e.g., robustness, marginals)

(III) **GVI**: **State of the art** $\mathcal{Q}$-constrained posteriors **on BNNs & DGPs**

Adams, R. P. and MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

Alquier, P. and Guedj, B. (2018). Simpler pac-bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902.

Alquier, P., Ridgway, J., and Chopin, N. (2016). On the properties of variational approximations of gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414.

Ambrogioni, L., Güçlü, U., Güçlütürk, Y., Hinne, M., van Gerven, M. A. J., and Maris, E. (2018). Wasserstein variational inference. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 2478–2487. Curran Associates, Inc.

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

Bui, T., Hernández-Lobato, D., Hernandez-Lobato, J., Li, Y., and Turner, R. (2016). Deep gaussian processes for regression using approximate expectation propagation. In *International Conference on Machine Learning*, pages 1472–1481.

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. (2017). Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, pages 2732–2741.

Fearnhead, P. and Liu, Z. (2007). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605.

Futami, F., Sato, I., and Sugiyama, M. (2017). Variational inference based on robust divergences. *arXiv preprint arXiv:1710.06595*.

Ganchev, K., Gillenwater, J., Taskar, B., et al. (2010). Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11(Jul):2001–2049.

Ghosh, A. and Basu, A. (2016). Robust bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437.

Grünwald, P., Van Ommen, T., et al. (2017). Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103.

Hernández-Lobato, J. M., Li, Y., Rowland, M., Hernández-Lobato, D., Bui, T. D., and Turner, R. E. (2016). Black-box α-divergence minimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 1511–1520.

Holmes, C. and Walker, S. (2017). Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503.

Hooker, G. and Vidyashankar, A. N. (2014). Bayesian model robustness via disparities. *Test*, 23(3):556–584.

Huang, C.-W., Tan, S., Lacoste, A., and Courville, A. C. (2018). Improving explorability in variational inference with annealed variational objectives. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 9724–9734. Curran Associates, Inc.

Main References ii

Jewson, J., Smith, J., and Holmes, C. (2018). Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Knoblauch, J. and Damoulas, T. (2018). Spatio-temporal Bayesian on-line changepoint detection with model selection. In *Proceedings of the 27th International Conference on Machine Learning (ICML-18)*.

Knoblauch, J., Jewson, J., and Damoulas, T. (2018). Doubly robust Bayesian inference for non-stationary streaming data using β-divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 64–75.

Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081.

Lichman, M. et al. (2013). Uci machine learning repository.

Miller, J. W. and Dunson, D. B. (2018). Robust bayesian inference via coarsening. *Journal of the American Statistical Association*, (just-accepted):1–31.

Minka, T. P. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc.

Opper, M. and Winther, O. (2000). Gaussian processes for classification: Mean-field algorithms. *Neural computation*, 12(11):2655–2684.

Ranganath, R., Tran, D., Altosaar, J., and Blei, D. (2016). Operator variational inference. In *Advances in Neural Information Processing Systems*, pages 496–504.

Regli, J.-B. and Silva, R. (2018). Alpha-beta divergence for variational inference. *arXiv preprint arXiv:1805.01045*.

Saha, A., Bharath, K., and Kurtek, S. (2017). A geometric variational approach to bayesian inference. *arXiv preprint arXiv:1707.09714*.

Salimbeni, H. and Deisenroth, M. (2017). Doubly stochastic variational inference for deep gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599.

Yang, Y., Pati, D., and Bhattacharya, A. (2017). α-variational inference with statistical guarantees. *arXiv preprint arXiv:1710.03266*.

Zellner, A. (1988). Optimal information processing and bayes's theorem. *The American Statistician*, 42(4):278–280.

Zhu, J., Chen, N., and Xing, E. P. (2014). Bayesian inference with posterior regularization and applications to infinite latent svms. *The Journal of Machine Learning Research*, 15(1):1799–1847.

4.2 GVI: Choosing loss hyperparameters

$$\mathcal{L}_p^{\beta}(\boldsymbol{\theta}, x_i) = -\frac{1}{\beta - 1} p(x_i|\boldsymbol{\theta})^{\beta-1} + \frac{I_{p,\beta}(\boldsymbol{\theta})}{\beta}$$

$$\mathcal{L}_p^{\gamma}(\boldsymbol{\theta}, x_i) = -\frac{1}{\gamma - 1} p(x_i|\boldsymbol{\theta})^{\gamma-1} \frac{\gamma}{I_{p,\gamma}(\boldsymbol{\theta})^{\frac{\gamma-1}{\gamma}}}$$

$$I_{p,c}(\boldsymbol{\theta}) = \int p(x|\boldsymbol{\theta})^c dx$$

where $I_{p,c}(\boldsymbol{\theta}) = \int p(x|\boldsymbol{\theta})^c dx$.

Note 1: $\mathcal{L}_p^{\gamma}(\boldsymbol{\theta}, x_i)$ multiplicative & always $< 0 \rightarrow$ store as log!
Note 2: Conditional independence $\neq$ additive for $\mathcal{L}_p^{\beta}(\boldsymbol{\theta}, x_i), \mathcal{L}_p^{\gamma}(\boldsymbol{\theta}, x_i)$
Note 3: In practice, usually best to choose $\beta/\gamma = 1 + \varepsilon$ for some small ε

Appendix: Choosing hyperparameters

Q: Any principled way of choosing hyperparameters?

A: Very much unsolved problem, solutions so far:

- D: brute force (CV) (Regli and Silva, 2018) [slow/expensive]
- ℓ_n: Via *points of highest influence* (Knoblauch et al., 2018)
- ℓ_n: on-line updates using loss-minimization (Knoblauch et al., 2018)

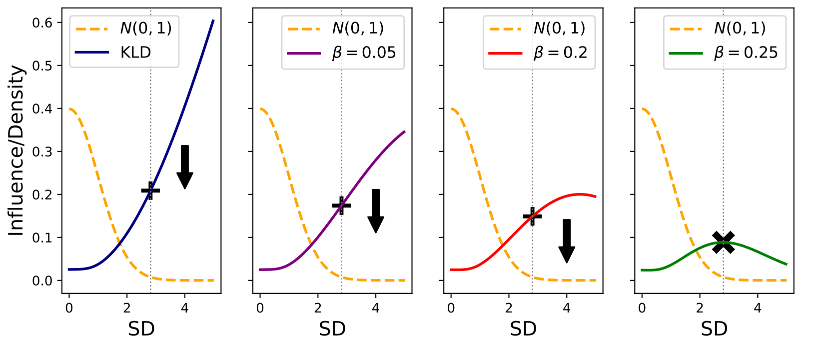


Figure 17 – Illustration of the initialization procedure using *points of highest influence* logic, from left to right.

Appendix: Choosing D for conservative marginals I/II

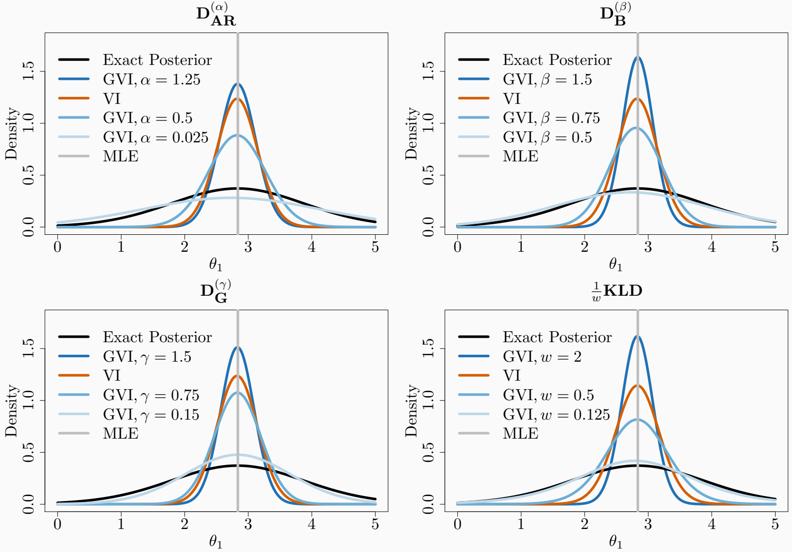


Figure 18 – Marginal VI and GVI posterior for a Bayesian linear model under the $D_{AR}^{(\alpha)}$, $D_{B}^{(\beta)}$, $D_{G}^{(\gamma)}$ and $\frac{1}{w}\mathrm{KLD}$ uncertainty quantifier for different values of the divergence hyperparameters.

Appendix: Choosing D for conservative marginals II/II

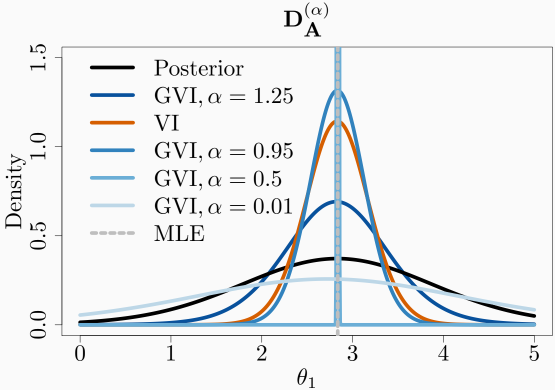


Figure 19 – Marginal **VI** and **GVI** posterior for a Bayesian linear model under the $D_A^{(\alpha)}$ uncertainty quantifier. The boundedness of the $D_A^{(\alpha)}$ causes **GVI** to severely over-concentrate if α is not carefully specified.

Appendix: Choosing D for prior robustness I/IV

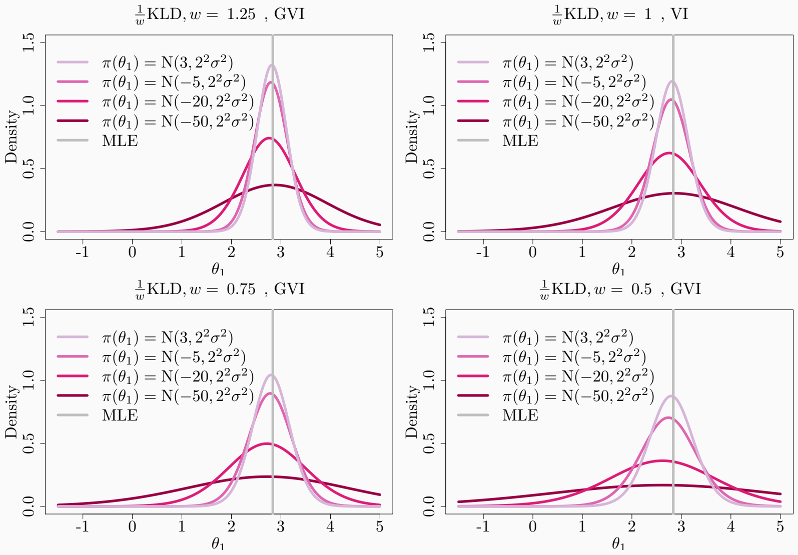


Figure 20 – Marginal **VI** and **GVI** posterior for a Bayesian linear model under different priors, using $D = \frac{1}{w}\mathrm{KLD}$ as the uncertainty quantifier.

Appendix: Choosing D for prior robustness II/IV

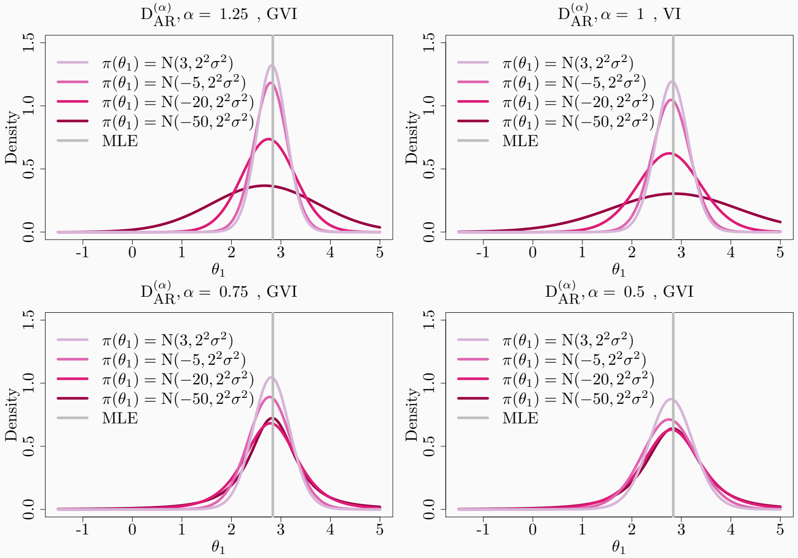


Figure 21 – Marginal **VI** and **GVI** posterior for a Bayesian linear model under different priors, using $D = D_{AR}^{(\alpha)}$ as the uncertainty quantifier.

Appendix: Choosing D for prior robustness III/IV

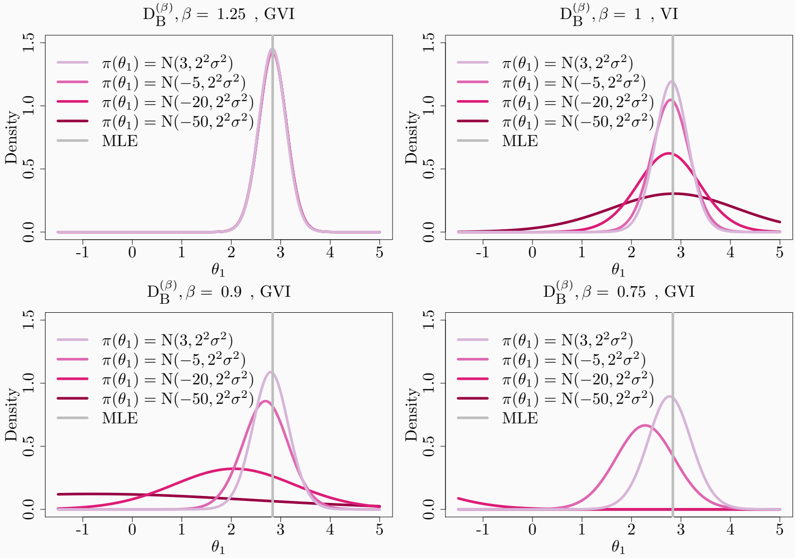


Figure 22 – Marginal **VI** and **GVI** posterior for a Bayesian linear model under different priors, using $D = D_B^{(\beta)}$ as the uncertainty quantifier.

Appendix: Choosing D for prior robustness IV/IV

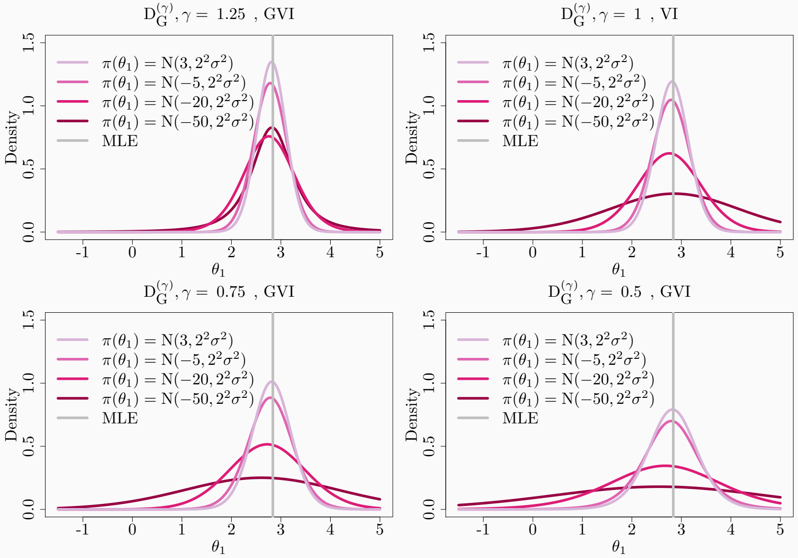


Figure 23 – Marginal **VI** and **GVI** posterior for a Bayesian linear model under different priors, using $D = D_G^{(\gamma)}$ as the uncertainty quantifier.

Appendix: GVI lower bound interpretation I/II

Question: VI is also interpretable as optimizing a lower bound on the evidence! Is there anything comparable for GVI?

Answer: Yes, e.g. for $D_B^{(\beta)}$, $D_G^{(\gamma)}$, $D_{AR}^{(\alpha)}$: Consider *generalized* evidence:

Recall: Generalized Bayes posterior (Bissiri et al., 2016) is

$$q^*_{\ell_n}(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \exp\{-\ell_n(\boldsymbol{\theta}, \boldsymbol{x})\} \quad \text{and so } p_{\ell_n}(\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} q^*_{\ell_n}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

GVI's objectives $L(q|\boldsymbol{x}, D, \ell_n)$ will optimize

$$L(q|\boldsymbol{x}, D, \ell_n) \geq g^D(\underbrace{-\log p_{f^D(\ell_n)}(\boldsymbol{x})}_{\substack{\text{negative log evidence;} \\ f^D(\ell_n) \text{ maps } \ell_n \text{ into a new loss}}}) + \underbrace{T^D(q)}_{\text{Approximate target}}$$

(**Note**: VI is special case where this holds with *equality* (so that the approximate target is the exact target) and where $g^{\mathrm{KLD}}(x) = x$, $L(q|\boldsymbol{x}, D, \ell_n) = \mathrm{ELBO}(q)$, $T^{\mathrm{KLD}}(q) = \mathrm{KLD}(q||q^*_{\ell_n})$, $f^{\mathrm{KLD}}(\ell_n) = \ell_n$.)

Appendix: GVI lower bound interpretation II/II

GVI's objectives $L(q|\boldsymbol{x}, D, \ell_n)$ will optimize

$$L(q|\boldsymbol{x}, D, \ell_n) \geq g^D\Big(\underbrace{-\log p_{f^D(\ell_n)}(\boldsymbol{x})}_{\substack{\text{negative log evidence;} \\ f^D(\ell_n) \text{ maps } \ell_n \text{ into a new loss}}} \Big) + \underbrace{T^D(q)}_{\text{Approximate target}}$$

Example: Rényi's α-divergence ($D_{AR}^{(\alpha)}$) for $\alpha > 1$ gives

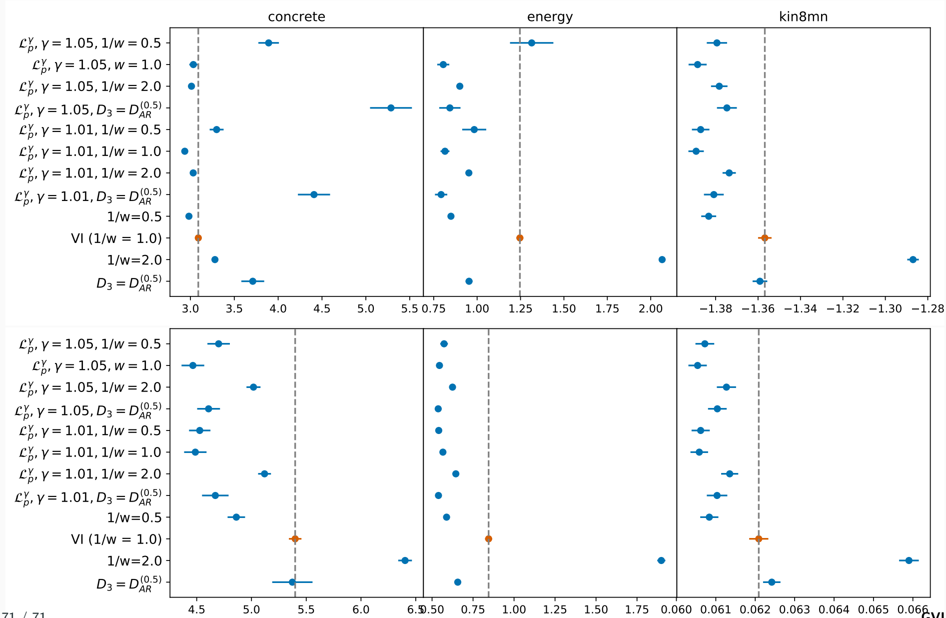
$$g^{D_{AR}^{(\alpha)}}(x) = \tfrac{1}{\alpha}x,$$
$$f^{D_{AR}^{(\alpha)}}(\ell_n) = \alpha\ell_n,$$
$$T_{D_{AR}^{(\alpha)}}(q) = \tfrac{1}{\alpha}\mathrm{KLD}(q||q^*_{\alpha\ell_n}),$$

so putting it together one finds that for $D = D_{AR}^{(\alpha)}$ with $\alpha > 1$,

$$L(q|\boldsymbol{x}, D, \ell_n) \geq -\frac{1}{\alpha}\log p_{\alpha\ell_n}(\boldsymbol{x}) + \frac{1}{\alpha}\mathrm{KLD}(q||q^*_{\alpha\ell_n})$$

(Which is just a $\frac{1}{\alpha}$-scaled version of the ELBO for the loss $\alpha\ell_n$!)

5.4 Experiments with Deep Gaussian Processes (DGPs) III/IV



GVI