Generalized Variational Inference (GVI)

Posterior beliefs with the rule of three

Jeremias Knoblauch\textsuperscript{1,3}, Jack Jewson\textsuperscript{1,3},
Theodoros Damoulas\textsuperscript{1,2,3}

April 18, 2019

\textsuperscript{1}University of Warwick, Department of Statistics
\textsuperscript{2}University of Warwick, Department of Computer Science
\textsuperscript{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 F-VI produce better posteriors?
   3.3 Towards GVI

4. GVI: What does it do?
   4.1 The losses
   4.2 Uncertainty Quantification
   4.3 Three GVI use cases

5. GVI: Inference & Experiments
   5.1 black box GVI
   5.2 Bayesian Neural Networks
   5.3 Deep Gaussian Processes
Purpose of part 1: Motivate the rule of three

(1) Bayesian inference minimizes **losses**
(2) Bayesian inference **regularizes** with the prior
(3) Bayesian inference $= \text{optimization over (sub)spaces of probability measures}$
1.1 The Bayesian problem: Traditional perspective

**Ingredients** (for the simplest case) are:

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

**Output = posterior belief**:

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

**Inference interpretation = belief updates**:

- likelihoods $\{p(x_i|\theta)\}_{i=1}^{n_1+n_2}$ update prior about $\theta$
- Old posterior $\tilde{\pi}(\theta) =$ new prior (coherence/Bayesian additivity)
1.2 The Bayesian problem: The optimization perspective

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

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

Notation:

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

Inference interpretation = regularized loss-minimization:

- $-\log(p(x_i|\theta)) =$ loss of $\theta$ for $x_i$
- Inference = regularizing MLE $\hat{\theta}_n$ with $\text{KLD}(q||\pi)$
1.3 The Bayesian problem: The loss-minimization perspective

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

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

for \( \tilde{\pi}(\theta) = \pi(\theta) \exp \left\{ -\sum_{i=1}^{n_1} \ell(\theta, x_i) \right\} \)

Inference interpretation = belief updates:

- Again: losses \( \{\ell(\theta, x_i)\}_{i=1}^{n_1+n_2} \) update prior about \( \theta \)
- Again: Old posterior \( \tilde{\pi}(\theta) = \) new prior (coherence)
- Difference: \( \theta \) arbitrary, e.g. \( \ell(\theta, x_i) = |x_i - \theta| \) admissible
Easy to show: Zellner’s representation valid for any $\ell(\theta, x_i)$:

$$q^*(\theta) = \arg\min_{q \in \mathcal{P}(\Theta)} \left\{ \mathbb{E}_{q(\theta)} \left[ \sum_{i=1}^{n} \ell(\theta, x_i) \right] + \text{KLD} (q||\pi) \right\}$$

minimized by $\delta_{\hat{\theta}_n}(\theta)$

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

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

Two more generalizations (break coherence):

- Replacing $\mathcal{P}(\Theta)$ with $\mathcal{Q} \subset \mathcal{P}(\Theta)$ ($=\text{VI}$)
- Replacing KLD with inference-problem specific regularizers
Our generalized representation of Bayesian inference:

\[ q^*(\theta) = \arg \min_{q \in \Pi} \left\{ \mathbb{E}_{q(\theta)} \left[ \sum_{i=1}^{n} \ell(\theta, x_i) \right] + D(q||\pi) \right\} \]

minimized by \( \delta_{\theta_n}(\theta) \)

minimized by \( q = \pi \)

Notation:

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

Inference interpretation = regularized & constrained minimization:

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

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

(1) Interpretations & modularity of $\ell_n$, $D$ and $\Pi$?
(2) Is there an axiomatic justification?
(3) Which existing methods does this (not) encompass?
2.1 Generalized Bayesian problem: provable modularity

\[ q^*(\theta) = \arg \min_{q \in \Pi} \left\{ \mathbb{E}_{q(\theta)} \left[ \sum_{i=1}^{n} \ell(\theta, x_i) \right] + D(q || \pi) \right\} \]

minimized by \( \delta_{\hat{\theta}_n}(\theta) \)

minimized by \( q = \pi \)

Roles of \( \ell_n, D, \Pi \):

- \( \ell_n \): which parameter \( \theta \) do we care about?
- \( D \): How is uncertainty quantified/what does \( q^* \) look like?
- \( \Pi \): Which beliefs are allowed?

\( \Rightarrow \) (provable) modularity of \( P(\ell_n, D, \Pi) \)!

Theorem 1 (GVI modularity)

For Bayesian inference with \( P(\ell_n, D, \Pi) \), making it robust to model misspecification amounts to changing \( \ell_n \). Conversely, adapting uncertainty quantification (fixing \( \Pi, \pi, \theta^*, \hat{\theta}_n \)) amounts to changing \( D \).
2.2 Generalized Bayesian problem: Axiomatic derivation I/II

Axiom 1 (Representation)
Bayesian inference infers posteriors $q$ on $\Theta$ by (i) measuring how $q$ fits a sample $x$ via the expectation of a loss $\ell_n(\theta, x)$, (ii) quantifying uncertainty about $\theta^*$ via a divergence $D$ between prior $\pi$ and $q$, (iii) optimizing $q$ over a space of probability distributions $\Pi$ on $\Theta$.

Axiom 2 (Information Difference)
$P(\ell_n, D, \Pi)$ produces different posteriors for $x = x_{1:n}$ and $x' = x_{1:n+m}$ if there is an information difference, i.e. if $\ell_n(\theta, x) \neq \ell_{n+m}(\theta, x')$.

Axiom 3 (Prior Regularization)
$q$ is regularized against $\pi$ by penalizing the divergence $D(q||\pi)$.

Axiom 4 (Translation Invariance)
For constant $C$ and $\ell'_n = \ell_n + C$, $P(\ell'_n, D, \Pi) = P(\ell_n, D, \Pi)$. 
Theorem 2 (Form 1)
If Axiom 1 holds, \( P(\ell_n, D, \Pi) \) has form arg min\(_{q\in\Pi}\) \( \{L(q|x, \ell_n, D)\} \) for \( L(q|x, \ell_n, D) = f(\mathbb{E}_{q(\theta)}[\ell_n(\theta, x)], D(q||\pi)) \), for some \( f : \mathbb{R}^2 \to \mathbb{R} \).

Theorem 3 (Form 2)
For \( P(\ell_n, D, \Pi) \) being arg min\(_{q\in\Pi}\) \( \{L(q|x, \ell_n, D)\} \) and \( \circ \) an elementary operation on \( \mathbb{R} \), \( L(q|x, \ell_n, D) = \mathbb{E}_{q(\theta)}[\ell_n(\theta, x)] \circ D(q||\pi) \) satisfies Axioms 3 and 4 only if \( \circ = + \).

Implications/relevance:
- Bayesian inference = constrained, regularized optimization
- Objective only depends on \( \mathbb{E}_{q(\theta)}[\ell_n(\theta, x)] \) and \( D(q||\pi) \)
- For elementary \( f(\mathbb{E}_{q(\theta)}[\ell_n(\theta, x)], D(q||\pi)) \), \( f \) must be addition.
  (Note: Axiom 4 excludes most non-elementary \( f \))
2.3 Generalized Bayesian problem & existing methods I/III

\[ q^*(\theta) = \arg \min_{q \in \Pi} \left\{ \mathbb{E}_q(\theta) [\ell_n(\theta, x)] + D(q \| \pi) \right\} \]

\( P(\ell_n, D, \Pi) \) covers & gives insight into existing methods, e.g.

- **Power Bayes**: \( P(w\ell_n, D, \Pi) = P(\ell_n, \frac{1}{w}D, \Pi) \).
  \( \implies w\)-power likelihood = \( \frac{1}{w} \times \) more trust in your prior.

- **Regularized Bayes**: Adding \( \Phi(q(\theta, x)) = \mathbb{E}_{q(\theta,x)} [\phi(\theta, x)] \) into the objective corresponds to \( P(\ell_n + \phi, D, \Pi) \).
  \( \implies \) RegBayes = a form of GVI that changes \( \ell_n \)
### 2.3 Generalized Bayesian problem & existing methods II/III

<table>
<thead>
<tr>
<th>Method</th>
<th>$\ell(\theta, x_i)$</th>
<th>$D$</th>
<th>$\Pi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Bayes</td>
<td>$-\log(p(\theta</td>
<td>x_i))$</td>
<td>KLD</td>
</tr>
<tr>
<td>Generalized Bayes$^1$</td>
<td>any $\ell$</td>
<td>KLD</td>
<td>$\mathcal{P}(\Theta)$</td>
</tr>
<tr>
<td>Power Bayes$^2$</td>
<td>$-\log(p(\theta</td>
<td>x_i))$</td>
<td>$\frac{1}{w}$KLD, $w &gt; 1$</td>
</tr>
<tr>
<td>Divergence Bayes$^3$</td>
<td>divergence-based $\ell$</td>
<td>KLD</td>
<td>$\mathcal{P}(\Theta)$</td>
</tr>
<tr>
<td><strong>Standard VI</strong></td>
<td>$-\log(p(\theta</td>
<td>x_i))$</td>
<td><strong>KLD</strong></td>
</tr>
<tr>
<td>Power VI$^4$</td>
<td>$-\log(p(\theta</td>
<td>x_i))$</td>
<td>$\frac{1}{w}$KLD, $w &gt; 1$</td>
</tr>
<tr>
<td>Regularized Bayes$^5$</td>
<td>$-\log(p(\theta</td>
<td>x_i)) + \phi(\theta, x_i)$</td>
<td>KLD</td>
</tr>
<tr>
<td>Gibbs VI$^6$</td>
<td>any $\ell$</td>
<td>KLD</td>
<td>$Q$</td>
</tr>
<tr>
<td><strong>Generalized VI</strong></td>
<td>any $\ell$</td>
<td>any $D$</td>
<td>$Q$</td>
</tr>
</tbody>
</table>

**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)
Not everything fits $P(\ell_n, D, \Pi)$:

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

2. **F-Variational inference (F-VI)**: VI based on discrepancy $F \neq \text{KLD}$ (locally) solving $\hat{q}^* = \arg\min_{q \in Q} F(q \| \tilde{q})$ for $\tilde{q} =$ standard Bayesian posterior, e.g.
   
   \[ F = \text{Rényi’s } \alpha\text{-divergence} \text{ (Li and Turner, 2016; Saha et al., 2017)} \]
   \[ F = \chi\text{-divergence} \text{ (Dieng et al., 2017)} \]
   \[ F = \text{operators} \text{ (Ranganath et al., 2016)} \]
   \[ F = \text{scaled AB-divergence} \text{ (Regli and Silva, 2018)} \]
   \[ F = \text{Wasserstein distance} \text{ (Ambrogioni et al., 2018)} \]
   
   …

3. **Expectation Propagation (EP)** (Minka, 2001; Opper and Winther, 2000) and its variants (e.g. Hernández-Lobato et al., 2016).

**Note:** Particular type of **F-VI**, with $F = (\text{local})$ reverse KLD
### Purpose of part 3: Motivating GVI

1. **Standard VI**: Optimality & reinterpretation
2. **F-VI**: “suboptimal” methods with better posteriors
3. **GVI**: A modular alternative to F-VI
3.1 Optimality & reinterpretation of standard VI I/V

Relationship between VI and exact inference?

Traditional view: Discrepancy-minimization, i.e. $\text{VI} = \text{approximation minimizing the KLD to } \tilde{q}$. (Inspiration for F-VI methods)

[From Variational Inference: Foundations and Innovations (Blei, 2019)]
3.1 Optimality & reinterpretation of standard VI II/V

**Standard VI**: \( q^* = \arg \min_{q \in Q} \text{KLD}(q \| \tilde{q}) \), \( \tilde{q} \) solves \( P(\ell_n, \text{KLD}, \mathcal{P}(\Theta)) \)

\[
\text{KLD}(q \| \tilde{q}) = \mathbb{E}_q(\theta) \left[ \log \left( \frac{q(\theta)}{\exp \left\{ - \sum_{i=1}^{n} \ell(\theta, x_i) \right\} \pi(\theta)} \right) \right] + \log \left( \int_\theta \exp \left\{ - \sum_{i=1}^{n} \ell(\theta, x_i) \right\} \pi(\theta) d\theta \right)
\]

(Generalized) ELBO

Generalized 'log evidence'

Inference = minimizing ELBO, which you can rewrite as

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

... which is exactly the objective of \( P(\ell_n, \text{KLD}, \mathcal{Q}) \)
In other words, $P(\ell_n, \text{KLD}, Q)$ (\(=\ \text{ELBO}\)) is
\[
q^*(\theta) = \arg\min_{q \in Q} \left\{ \mathbb{E}_{q(\theta)} [\ell_n(\theta, x)] + D (q||\pi) \right\},
\]
the $Q$-constrained relaxation of $P(\ell_n, \text{KLD}, \mathcal{P}(\Theta))$, whose objective is
\[
q^*(\theta) = \arg\min_{q \in \mathcal{P}(\Theta)} \left\{ \mathbb{E}_{q(\theta)} [\ell_n(\theta, x)] + D (q||\pi) \right\},
\]
(which is the exact Bayesian objective).

⇒ Reinterpretation of standard VI as Constrained optimization!
3.1 Optimality & reinterpretation of standard VI IV/V

Alternative view: $\text{VI} = \mathcal{Q}$-constrained version of exact Bayes problem

Figure 1 – Left: Unconstrained (i.e. exact) Bayesian inference. Right: Constrained (i.e. standard variational) Bayesian inference
Consequence I/II: VI-optimality

Theorem 4 (VI optimality)

For exact and coherent Bayesian posteriors solving $P(\ell_n, KLD, P(\Theta))$ and a fixed variational family $Q$, standard VI produces the uniquely optimal $Q$-constrained approximation to $P(\ell_n, KLD, P(\Theta))$. Having decided on approximating the Bayesian posterior with some $q \in Q$, VI provides the uniquely optimal solution.
3.2 Why does F-VI produce better posteriors? I/II

Consequences II/II: F-VI-suboptimality. Three big disadvantages:

(1) If $F \neq \text{KLD}$, F-VI violates Axioms 1–4.

(2) F-VI conflates $\ell_n$ and $D$ (i.e., modularity of $P(\ell_n, D, \Pi)$ lost).

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

Objection! F-VI often produces better posteriors than standard VI!
3.2 Why does F-VI produce better posteriors? II/II

Seeming contradiction:

(1) VI is the best approximation to the standard Bayesian posterior
(2) F-VI often outperforms VI (e.g., on test scores)

Resolution:

F-VI outperforms VI by implicitly targeting a non-standard Bayesian problem that is more appropriate than $P(\ell_n, \text{KLD}, P(\Theta))$

$\implies$ Inspires Generalized Variational Inference (GVI)
3.3 Towards GVI I/II

**GVI** = combining advantages of **VI** and **F-VI**:

(1) Has form $P(\ell_n, D, Q)$ Like **VI** i.e.
   - (i) satisfies Axioms 1–4;
   - (ii) provably interpretable modularity (**loss, uncertainty quantifier, admissible posteriors**)

(2) Derives different & more appropriate posteriors like **F-VI** but
   - (i) without conflating $\ell_n$ and $D$
   - (ii) with explicit rather than implicit changes.

**Definition 1 (GVI)**

Any Bayesian inference method solving $P(\ell_n, D, Q)$ with admissible choices $\ell_n$, $D$ and $Q$ is a Generalized Variational Inference (**GVI**) method satisfying Axioms 1 – 4.
Illustration: **F-VI** aims for $D$, but changes $\ell_n$ – **GVI** doesn’t

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

**Exact Posterior**

**VI**

**F-VI, F = D_{AR}^{(0.5)}**

**GVI, $\alpha = 0.25$**

**MLE**
Purpose of part 4: Exploring three use cases of GVI

(1) Robust alternatives to $\ell(\theta, x_i) = -\log(p(x_i|\theta))$

(2) Prior-robust uncertainty quantification via $D$

(3) Adjusting marginal variances via $D$
4.1 **GVI: The losses I/III**

**GVI modularity: The loss $\ell_n$**

**Q1:** Why use $\ell_n(\theta, x) = \sum_{i=1}^{n} -\log(p(x_i|\theta))$?

**A:** Assuming that the true data-generating mechanism is $x \sim g$,

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

**Interpretation:** $-\log(p(x_i|\theta)) = \text{targeting KLD-minimizing } p(\cdot|\theta)$

**Q2:** Are there other $\mathcal{L}^D(p(x_i|\theta))$ for divergence $D$?

**A:** Yes! (e.g. Jewson et al., 2018; Futami et al., 2017; Ghosh and Basu, 2016; Hooker and Vidyashankar, 2014)
Q3: Why use other $\mathcal{L}^D(p(x; \theta))$?

A: Robustness (for $D$ = a robust divergence) [log/KLD non-robust!]

Robustness recipe: $\alpha/\beta/\gamma$-divergences using generalized log functions

E.g.: $\beta$ indexes $\beta$-divergence ($D_B^{(\beta)}$) via

$$
\log_\beta(x) = \frac{1}{(\beta - 1)\beta} \left[ \beta x^{\beta - 1} - (\beta - 1)x^\beta \right]
$$

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

Note 1: $D_B^{(\beta)} \rightarrow$ KLD as $\beta \rightarrow 1$!

Note 2: Admits $D_B^{(\beta)}$-targeting loss as

$$
\mathcal{L}_p^{\beta}(\theta, x_i) = -\frac{1}{\beta - 1} p(x_i|\theta)^{\beta - 1} + \frac{l_p,\beta(\theta)}{\beta}, \quad l_p,\cdot(\theta) = \int p(x|\theta)^c dx
$$
4.1 **GVI**: The losses \(III/III\)

**Figure 3 – Left**: Robustness against model misspecification. Depicted are posterior predictives under \(\epsilon = 5\%\) outlier contamination using \(\text{VI}\) and \(P(\sum_{i=1}^{n} \mathcal{L}_p^\beta(\theta, x_i), \text{KLD}, Q)\), \(\beta = 1.5\). **Right**: From Knoblauch et al. (2018). Influence of \(x_i\) on exact posteriors for different losses.
4.2 GVI: Uncertainty Quantification I/III

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
- ...
4.2 GVI: Uncertainty Quantification II/III

Example 1: **GVI** can fix over-concentrated posteriors

![Graph showing divergence and density](image)

**Figure 4 – Left:** Magnitude of the penalty incurred by $D(q||\pi)$ for different uncertainty quantifiers $D$ and fixed densities $\pi, q$. **Right:** Using $D^{(\alpha)}_{AR}$ with different choices of $\alpha$ to “customize” uncertainty.
Example 2: Avoiding prior sensitivity

Figure 5 – Prior sensitivity with VI (left) vs. prior robustness with GVI (right). Priors are more badly specified for darker shades.
Summary: some GVI applications include

(1) Robustness to model misspecification \( (= \text{adapting } \ell_n) \)
(2) “Customized” marginal variances \( (= \text{adapting } D) \)
(3) Prior robustness \( (= \text{adapting } D) \)
Purpose of part 5: \textbf{GVI} inference & experiments

(1) How/when can we “black box” \textbf{GVI}?

(2) \textbf{F-VI} vs \textbf{GVI} & changes in $D$ (on Bayesian Neural Nets)

(3) \textbf{VI} vs \textbf{GVI} & changes in $\ell_n$ (on Deep Gaussian Processes)
5.1 Black Box GVI

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

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

(ii) derivative \( \nabla_\kappa \log(q(\theta | \kappa)) \) exists.

**Case 1:** Closed form for \( \nabla_\kappa D(q || \pi) \rightarrow \) unbiased estimate:

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

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

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

\[
\nabla_\kappa \hat{L}(q | \ell_n, D) = \frac{1}{S} \sum_{s=1}^{S} \left\{ \left[ \ell_n(\theta^{(s)}, x) + \ell^D_{\kappa, \pi}(\theta^{(s)}) \right] \cdot \nabla_\kappa \log(q(\theta^{(s)} | \kappa)) \right. \\
+ \left. \nabla_\kappa \ell^D_{\kappa, \pi}(\theta^{(s)}) \right\}.
\]
BNNs are intractable Bayesian regression models with

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

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

$$F_\theta(x)$$
Methods: Comparison of black box approximate Bayesian methods:

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

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

- Variational family $Q$: A fully factorized normal
- Optimization of $\sigma^2$ (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
5.2 Experiments with Bayesian Neural Nets (BNNs) III/IV

Figure 6 – Performance on BNNs: F-VI, GVI with $D = D^{(\alpha)}_{AR}$, 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 F-VI outperform VI?!!
5.2 Experiments with Bayesian Neural Nets (BNNs) IV/V

**Figure 7** – Left: Parameter posteriors (**F-VI** as expected). **Right**: Posterior predictives (**F-VI not** as expected)
Q: Why does this happen for F-VI and not for GVI?!

A: F-VI does not distinguish uncertainty quantification & loss!

**F-VI objective**: $\sigma^2$ affects target (!)

$$\hat{\sigma}^2, q^*(\theta|\hat{\sigma}^2, \kappa) = \arg\min\sigma^2 \left\{ \arg\min\limits_{q \in Q} F\left(q(\theta|\sigma^2, \kappa)\mid \tilde{q}(\theta|\sigma^2, x, y)\right) \right\}$$

i.e., $\tilde{q} = \tilde{q}^\sigma$

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

**GVI objective**: $\sigma^2$ indexes the loss only

$$\hat{\sigma}^2, q^*(\theta|\hat{\sigma}^2, x, y) = \arg\min\sigma^2 \left\{ \arg\min\limits_{q \in Q} \left\{ \mathbb{E}_q [\ell_n(\theta, x|y, \sigma^2)] + D(q||\pi) \right\} \right\}$$

i.e., $\ell_n = \ell^\sigma_n$

$\Rightarrow$ optimizing for $\sigma^2 =$ finding **optimal loss** $\ell^\sigma_n$
Principal idea: Use the BNN architecture with GP priors on $F_\theta(\cdot)$:

\[
y | F^L \sim p(y | F^L)
\]

\[
F^L | F^{L-1} \sim \text{GP} \left( \mu^L(F^{L-1}), K^L(F^{L-1}, F^{L-1}) \right)
\]

\[
F^{L-1} | F^{L-2} \sim \text{GP} \left( \mu^{L-1}(F^{L-2}), K^{L-1}(F^{L-2}, F^{L-2}) \right)
\]

\[\ldots\]

\[
F^1 | x \sim \text{GP} \left( \mu^1(x), K^1(x, x) \right)
\]

Methods: Comparison of black box approximate Bayesian methods:

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

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

5.3 Experiments with Deep Gaussian Processes (DGPs) II/II

Figure 8 – DGP performance with $L$ layers, **GVI** with $\ell_n (\theta, x) = \sum_{i=1}^n \mathcal{L}_p (\theta, x_i)$ & **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 \( \mathcal{P}(\Theta) \)

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

Part 3: Fallout of \( P(\ell_n, D, \Pi) \): VI optimality & F-VI 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 & DGP


Appendix: Choosing robust $\ell_n$

\begin{align*}
\mathcal{L}_{p}^\beta(\theta, x_i) &= -\frac{1}{\beta - 1} p(x_i | \theta)^{\beta - 1} + \frac{l_{p, \beta}(\theta)}{\beta} \\
\mathcal{L}_{p}^\gamma(\theta, x_i) &= -\frac{1}{\gamma - 1} p(x_i | \theta)^{\gamma - 1} \frac{\gamma}{l_{p, \gamma}(\theta)^{\gamma - 1}} \\
l_{p, c}(\theta) &= \int p(x | \theta)^c \, dx
\end{align*}

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

**Note 1:** $\mathcal{L}_{p}^\gamma(\theta, x_i)$ multiplicative & always $< 0 \rightarrow$ store as log!

**Note 2:** Conditional independence $\neq$ additive for $\mathcal{L}_{p}^\beta(\theta, x_i), \mathcal{L}_{p}^\gamma(\theta, x_i)$

**Note 3:** In practice, usually best to choose $\beta / \gamma = 1 + \varepsilon$ for some small $\varepsilon$
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]
- \( \ell_n \): Via \textit{points of highest influence} (Knoblauch et al., 2018)
- \( \ell_n \): on-line updates using loss-minimization (Knoblauch et al., 2018)

\[\text{Figure 9} \quad \text{Illustration of the initialization procedure using \textit{points of highest influence} logic, from left to right.}\]
Appendix: Choosing $D$ for conservative marginals I/II

Figure 10 — Marginal VI and GVI posterior for a Bayesian linear model under the $D_{AR}^{(\alpha)}$, $D_{B}^{(\beta)}$, $D_{G}^{(\gamma)}$ and $\frac{1}{w}KLD$ uncertainty quantifier for different values of the divergence hyperparameters.
Appendix: Choosing $D$ for conservative marginals II/II

Figure 11 – 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 $\alpha$ is not carefully specified.
Appendix: Choosing $D$ for prior robustness I/IV

![Graphs showing marginal VI and GVI posterior for different priors using $D = \frac{1}{w} \text{KLD}$ as uncertainty quantifier.]

**Figure 12** — Marginal VI and GVI posterior for a Bayesian linear model under different priors, using $D = \frac{1}{w} \text{KLD}$ as the uncertainty quantifier.
Appendix: Choosing $D$ for prior robustness II/IV

Figure 13 – 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 14 – Marginal VI and GVI posterior for a Bayesian linear model under different priors, using $D = D^{(β)}_B$ as the uncertainty quantifier.
Figure 15 – 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

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

GVI’s objectives $L(q|x, D, \ell_n)$ will optimize

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

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

54 / 56
Appendix: GVI lower bound interpretation II/II

GVI’s objectives $L(q|x, D, \ell_n)$ will optimize

$$L(q|x, D, \ell_n) \geq g^D(-\log p_{f^D(\ell_n)}(x)) + T^D(q)$$

negative log evidence; $f^D(\ell_n)$ maps $\ell_n$ into a new loss

Approximate target

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

$$g_{AR}^{D^{(\alpha)}}(x) = \frac{1}{\alpha} x,$$

$$f_{AR}^{D^{(\alpha)}}(\ell_n) = \alpha \ell_n,$$

$$T_{D_{AR}^{(\alpha)}}(q) = \frac{1}{\alpha} \text{KLD}(q||q^*_{\alpha \ell_n}),$$

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

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

(Which is just a $\frac{1}{\alpha}$-scaled version of the ELBO for the loss $\alpha \ell_n$!)
5.3 Experiments with Deep Gaussian Processes (DGPs) III/IV

[Graph showing experimental results for concrete, energy, and kin8mn datasets with different parameters and metrics.]