Workshop on Generalized Variational Inference (GVI)

Posterior beliefs with The Rule of Three

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

July 19, 2019

[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 workshop

1. Theoretical foundations & The Rule of Three (10:00–11:00)
   - 1.1 Basics: losses & divergences
   - 1.2 Bayesian inference: updating view vs. optimization view
   - 1.3 The rule of three: special cases, modularity & axiomatic foundations
   - 1.4 Q&A/Discussion
2. Optimality of **VI**, **F-VI** suboptimality & **GVI**'s motivation (11:30–12:30)
   - 2.1 VI interpretations: discrepancy-minimization vs constrained optimization
   - 2.2 VI optimality & sub-optimality of F-VI
   - 2.3 GVI as modular and explicit alternative to F-VI
   - 2.4 GVI use cases
   - 2.5 GVI's lower bound interpretation
   - 2.6 Q&A/Discussion
3. **GVI** Applications (14:00 – 15:00)
   - 3.1 Robust Bayesian On-line Changepoint Detection
   - 3.2 Bayesian Neural Networks
   - 3.3 Deep Gaussian Processes
   - 3.4 Other work
   - 3.5 Q&A/Discussion
4. Chalk talk: Consistency & Concentration rates for **GVI** (15:30 – 16:30)
   - 4.1 The role of Γ-convergence
   - 4.2 GVI procedures as ε-optimizers
   - 4.3 Central results
   - 4.4 Q&A/Discussion

Part 1

Part 1: Theoretical foundations & The Rule of Three

1. Basics & Preliminaries
    - 1.1 Notation & color code
    - 1.2 Divergences & losses
    - 1.3 Robust divergences & the KLD
2. Perspectives on Bayesian inference
    - 2.1 The traditional perspective
    - 2.2 The optimization perspective
    - 2.3 The loss-minimization perspective
3. The Rule of Three
    - 3.1 Extending the Bayesian paradigm
    - 3.2 Provable modularity
    - 3.3 Axiomatic derivation
    - 3.4 Relationship to existing methods
4. Discussion

1. Basics & Preliminaries: Purpose

Purpose of section 1: Setting the scene.

(1) What are divergences & why should you care?

(2) What is the relationship between divergences & losses?

(3) What are robust divergences & 'generalized log functions'?

1.1 Notation & color code

Color code:

- Relating to **losses**
- Relating to **divergences**
- Relating to (sub)spaces on $\mathcal{P}(\boldsymbol{\Theta})$

Notation:

(i) $\boldsymbol{\Theta}$ = parameter space, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ = parameter value

(ii) q, π are **densities** on $\boldsymbol{\Theta}$, i.e. $q, \pi : \boldsymbol{\Theta} \to \mathbb{R}_+$

π = **prior** (i.e., known before data is seen)

q = **posterior** (i.e., known after data is seen)

(iii) $\mathcal{P}(\boldsymbol{\Theta})$ = set of all probability measures on $\boldsymbol{\Theta}$

(**Note**: Will abuse notation and write/treat $q, \pi \in \mathcal{P}(\boldsymbol{\Theta})$)

(iv) $\mathcal{Q}$ = **parameterized** subset of $\mathcal{P}(\boldsymbol{\Theta})$, i.e. $\mathcal{Q}$ = **variational family**

(v) $g(x_i)$ = density for iid observations $x_{1:n} \overset{iid}{\sim} g$

(vi) $p(x_i|\boldsymbol{\theta})$ = likelihood model for x_i indexed by $\boldsymbol{\theta}$

1.2 Divergences: Definition

Definition 1 (Statistical Divergence)

A statistical divergence $D(q||\pi)$ is a measure of discrepancy between two probability densities q and π on the same space Θ with the following two properties

1. $D(q||\pi) \geq 0$, $\forall q, \pi \in \mathcal{P}(\Theta)$
2. $D(q||\pi) = 0$ if and only if $q = \pi$ (a.e.).

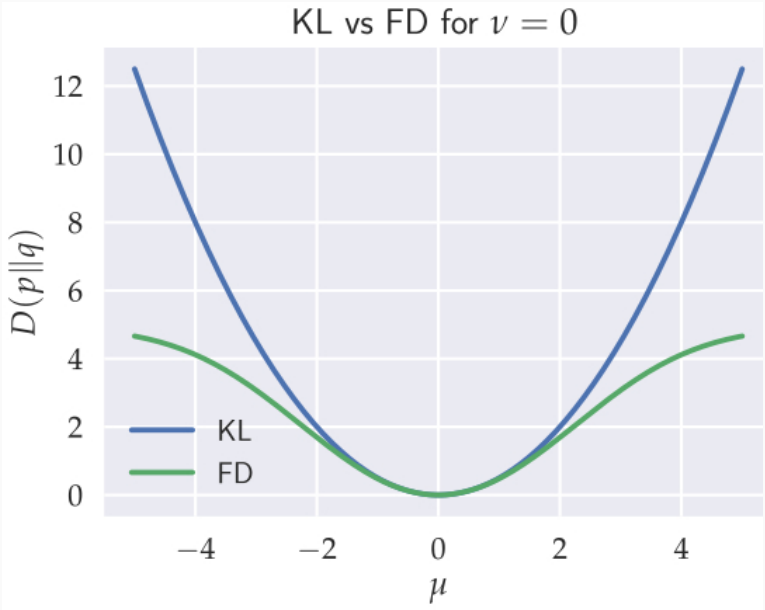


Figure 1 – Fisher vs Kullback-Leibler, courtesy of Boris Belousov's blog.

1.2 Divergences

Implications:

(1) Divergence D maps from ∞-dim spaces to $\mathbb{R}$, i.e. $D : \mathcal{P}(\boldsymbol{\Theta})^2 \to \mathbb{R}_+$;

(2) A divergence D need *not* be a distance metric as it can

- be asymmetric
- not be satisfying the triangle inequality;

(3) Any distance metric D on $\mathcal{P}(\boldsymbol{\Theta})$ is also a divergence (e.g., TVD).

Some examples:

Kullback-Leibler Divergence (KLD)

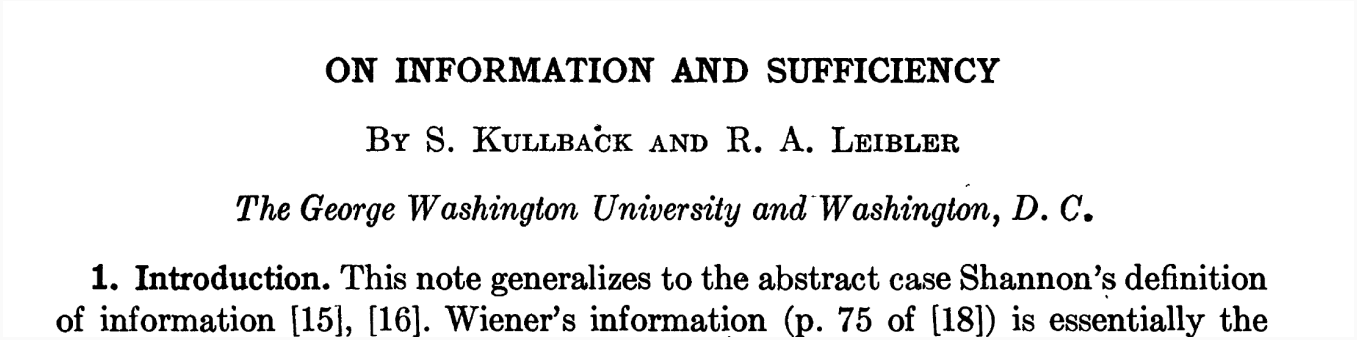
Rényi's α-divergence ($D_{AR}^{(\alpha)}$)

α-divergence ($D_{A}^{(\alpha)}$)

β-divergence ($D_{B}^{(\beta)}$)

...

1.2 Divergences: KLD



ON INFORMATION AND SUFFICIENCY

By S. Kullback and R. A. Leibler

The George Washington University and Washington, D. C.

1. Introduction. This note generalizes to the abstract case Shannon's definition of information [15], [16]. Wiener's information (p. 75 of [18]) is essentially the

Definition:

$$\mathrm{KLD}(q||\pi) = \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right)\right] = \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log q(\boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta})\right]$$

Interpretation due to Kullback and Leibler (1951):

- Two hypotheses about which population $\boldsymbol{\theta}$ comes from:
  (i) $\mathcal{H}_\pi$ (i.e., hypothesis is that $\theta \sim \pi$) and
  (ii) $\mathcal{H}_q$ (i.e., hypothesis is that $\theta \sim q$).
- $\log\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right)$ = **information in $\boldsymbol{\theta}$ for discriminating $\mathcal{H}_\pi$ and $\mathcal{H}_q$**
- KLD = **arithmetic mean** information for discriminating $\mathcal{H}_\pi$ and $\mathcal{H}_q$ (mean relative to $\boldsymbol{\theta} \sim q$)

1.2 Divergences: $D_{AR}^{(\alpha)}$

Definition (different from the one used later!):

$$D_{AR}^{(\alpha)}(q||\pi) = \frac{1}{(\alpha-1)} \log\left(\mathbb{E}_{q(\boldsymbol{\theta})}\left[\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right)^{\alpha-1}\right]\right)$$

Interpretation due to Rényi et al. (1961):

- $\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right)^{\alpha-1}$ = **information** in $\boldsymbol{\theta}$ for discriminating $\mathcal{H}_\pi$ and $\mathcal{H}_q$
- $D_{AR}^{(\alpha)}$ = **geometric mean** information for discriminating $\mathcal{H}_\pi$ and $\mathcal{H}_q$ (mean relative to $\boldsymbol{\theta} \sim q$)
- Easiest to see if $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$, i.e. $|\boldsymbol{\Theta}| = M$:

$$D_{AR}^{(\alpha)}(q||\pi) = \frac{1}{(\alpha-1)} \log\left(\sum_{i=1}^{M} q(\boldsymbol{\theta}_i)\left(\frac{q(\boldsymbol{\theta}_i)}{\pi(\boldsymbol{\theta}_i)}\right)^{\alpha-1}\right).$$

$$e^{D_{AR}^{(\alpha)}(q||\pi)} = \left(\sum_{i=1}^{M} q(\boldsymbol{\theta}_i)\left(\frac{q(\boldsymbol{\theta}_i)}{\pi(\boldsymbol{\theta}_i)}\right)^{\alpha-1}\right)^{\frac{1}{\alpha-1}},$$

1.2 Divergences: f-divergences

Definition (for convex function f s.t. $f(1) = 0$):

$$D^f(q||\pi) = \mathbb{E}_{\pi(\boldsymbol{\theta})}\left[f\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right)\right]$$

Information interpretation:

- $f\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right)$ = **information** in $\boldsymbol{\theta}$ for discriminating $\mathcal{H}_\pi$ and $\mathcal{H}_q$
- $D^f(q||\pi)$ = **arithmetic mean** information for discriminating $\mathcal{H}_\pi$, $\mathcal{H}_q$ (mean relative to $\boldsymbol{\theta} \sim \pi$)
- Important examples:
  - (i) KLD for $f(z) = z\log(z)$
  - (ii) $D_A^{(\alpha)}$ for $f(z) = z\log_\alpha(z)$ with $\log_\alpha(z) = \frac{1}{\alpha(\alpha-1)}(z^{\alpha-1} - 1)$
    Note: $\lim_{\alpha\to 1}\log_\alpha(z) = \log(z)$, i.e. generalized log function!

1.2 Divergences & losses: KLD

Definition:

$$\mathrm{KLD}(g||p) = \mathbb{E}_{g(\boldsymbol{x})}\left[\log\left(\frac{g(\boldsymbol{x})}{p(\boldsymbol{x}|\boldsymbol{\theta})}\right)\right] = \mathbb{E}_{g(\boldsymbol{x})}\left[\log g(\boldsymbol{x}) - \log p(\boldsymbol{x}|\boldsymbol{\theta})\right]$$

Intepretation in e.g. Dawid et al. (2016), Jewson et al. (2018):

- With sample $x_{1:n} \overset{iid}{\sim} g$, find best $\boldsymbol{\theta}^*$ indexing likelihood $p(x|\boldsymbol{\theta})$
  $\Longrightarrow$ **Scoring rule inference:** Solve (e.g. for $f = -\log p(\boldsymbol{x}|\boldsymbol{\theta})$)

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} f(x_i, \boldsymbol{\theta}). \tag{1}$$

- f is a **proper** scoring rule (= form of loss) if

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}\sim g}\left[f(\boldsymbol{x}, \boldsymbol{\theta})\right] \tag{2}$$

  (1) is minimized for $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ and
  (2) $p(\boldsymbol{x}|\boldsymbol{\theta}^*) = g(\boldsymbol{x})$ if p is the correct likelihood model for g.
- $\min_{\boldsymbol{\theta}} \mathrm{KLD}(g||p) = \min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}\sim g}\left[f(\boldsymbol{x}, \boldsymbol{\theta})\right]$ for $f = -\log p(\boldsymbol{x}|\boldsymbol{\theta})$
  $\Longrightarrow$ MLE = minimizer of the KLD (i.e., a divergence)!

1.2 Divergences & losses: $D_B^{(\beta)}$

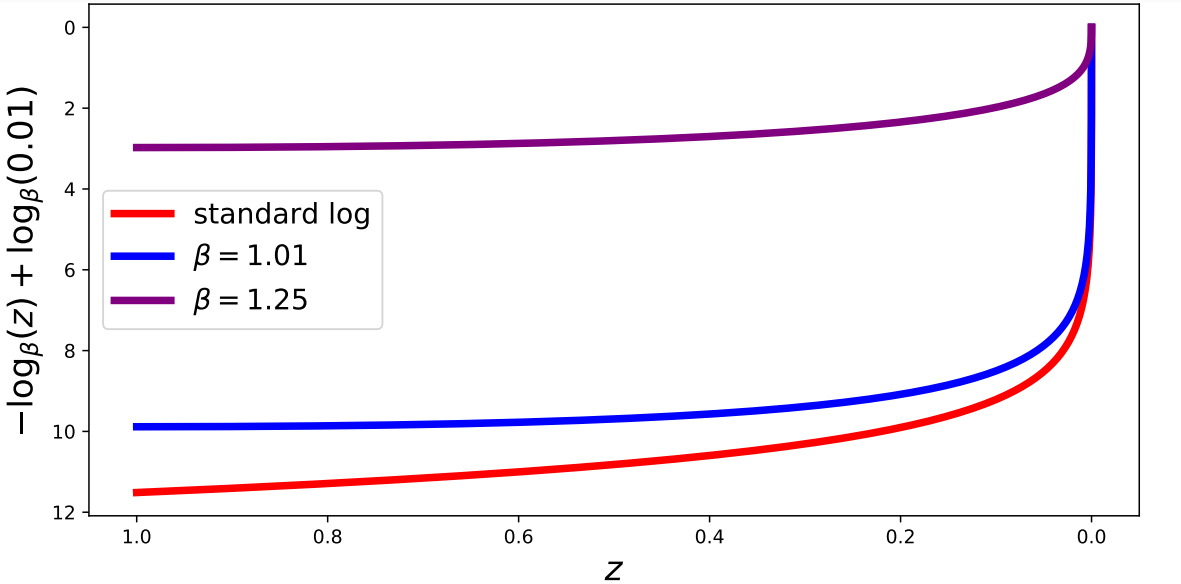
Definition:

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

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

Interpretation (e.g., Basu et al. (1998) & Jewson et al. (2018)):

- $\log_\beta$ is a generalized log function: $\lim_{\beta\to 1}\log_\beta(x) = \log(x)$ for all x
- If $\beta > 1$, $\log_\beta$ is a **robust** alternative for the log **score**



1.2 Divergences & losses: Bregman Divergences

Note: $D_B^{(\beta)}$ is a special case of Bregman (ϕ-) divergences (D^ϕ):

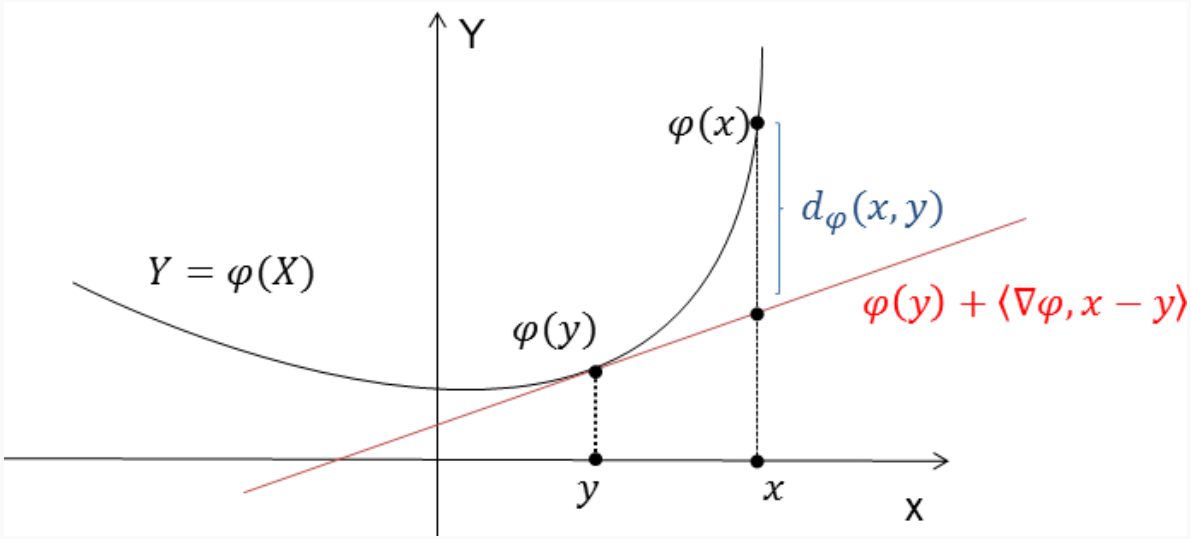
$$D^\phi(g||p) = \int_{\mathcal{X}} [\phi(g(\boldsymbol{x})) - \phi(p(\boldsymbol{x}|\boldsymbol{\theta})) - (g(\boldsymbol{x}) - p(\boldsymbol{x}|\boldsymbol{\theta})) \cdot \phi'(p(\boldsymbol{x}|\boldsymbol{\theta}))]\, d\boldsymbol{x},$$

$$D_B^{(\beta)}(g||p) = \mathbb{E}_{g(\boldsymbol{x})}\left[\log_\beta g(\boldsymbol{x}) - \log_\beta p(\boldsymbol{x}|\boldsymbol{\theta})\right] = D^{\phi_\beta}(g||p),$$

$$\phi_\beta(z) = \frac{1}{\beta(\beta-1)}(z^\beta - \beta z + \beta - 1)$$

Interpretation (e.g., Ferreira et al. (2015)):

- How good is 1st-order approximation of $\phi(g)$ expanded around p?



1.3 Robust divergences & the KLD

Useful observation: 'Robust divergences' = generalizing the log that KLD is based on (in different ways, see Cichocki and Amari (2010))

Examples:

- β-divergence ($D_B^{(\beta)}$), see previous slide
- Rényi's α-divergence ($D_{AR}^{(\alpha)}$)
- γ-divergence ($D_G^{(\gamma)}$)
- α-divergence ($D_A^{(\alpha)}$)
- ...

Convention: All generalizations recover KLD for $\alpha/\beta/\gamma \to 1$

Summary: Basics & Preliminaries

(1) Divergence $=$ discrepancy measure in infinite-dimensional space

(2) 'Holy grail' of divergences $=$ KLD $=$ mean log-information for discriminating two hypotheses ($\boldsymbol{\theta} \sim \pi$ and $\boldsymbol{\theta} \sim q$)

(3) Robust alternatives $=$ generalizations of log function

(4) Divergences can be used to derive losses/score functions

2 The form of the Bayesian problem: Purpose

Purpose of section 2: Laying groundwork for The Rule of Three

(1) Bayesian inference minimizes losses

(2) Bayesian inference regularizes with the prior

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

2.1 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:

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

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**)

2.2 The Bayesian problem: The optimization perspective

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

$$q^*(\boldsymbol{\theta}) = \underset{q \in \mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \underbrace{\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} -\log(p(x_i|\boldsymbol{\theta}))\right]}_{\text{minimized by } q(\boldsymbol{\theta}) = \delta_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{\theta}),\ \hat{\boldsymbol{\theta}}_n = \text{MLE}} + \underbrace{\text{KLD}\left(q||\pi\right)}_{\text{minimized by } q = \pi} \right\}, \quad (3)$$

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 $\text{KLD}(q||\pi)$

2.3 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

2.3 The Bayesian problem: The loss-minimization perspective

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

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

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

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

Inference interpretation = regularized loss-minimization:

- $\ell(\boldsymbol{\theta}, x_i)$ = **loss** of $\boldsymbol{\theta}$ for x_i
- Inference = regularizing loss-minimizer $\hat{\boldsymbol{\theta}}_n$ with $\text{KLD}(q||\pi)$

Summary: Perspectives on Bayesian inference

(1) Bayesian inference is an **optimization problem** over $\mathcal{P}(\Theta)$ about a parameter θ specified via a loss and regularized by the KLD

(2) Solution of problem $=$ multiplicative **belief updates**

(3) Theoretically coherent to form beliefs for **any** parameter θ occurring in a loss function

3. The Rule of Three: Purpose

Purpose of section 3: Derive generalized Bayesian inference

(1) What is $P(\ell_n, D, \Pi)$/The Rule of Three?

(2) What is each of the three component's functionality?

(3) Can you derive this representation axiomatically?

(4) What is its relationship to existing Bayesian inference methods?

3.1 Recap: Bayesian inference solves optimization problem

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

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

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

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

Inference interpretation = regularized loss-minimization:

- $\ell(\boldsymbol{\theta}, x_i)$ = **loss** of $\boldsymbol{\theta}$ for x_i
- Inference = regularizing loss-minimizer $\hat{\boldsymbol{\theta}}_n$ with $\text{KLD}(q||\pi)$

3.1 The Rule of Three: Extending the Bayesian Paradigm

Our **generalized representation of Bayesian inference**:

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

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
- $\mathbf{D}$ = **divergence**, acting as uncertainty quantifier/regularizer
- $\boldsymbol{\Pi}$ = set of **admissible posterior** beliefs
- **Inference** = constrained, regularized optimization

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

3.2 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.

Note: Proof less trivial than it may look: Take $\ell_n^{(1)}$ and $\ell_n^{(2)} = 2 \cdot \ell_n^{(1)}$?!

3.3 Generalized Bayesian problem: Axiomatic derivation I/II

Axiom 1 (Representation)

Bayesian inference infers posteriors q on Θ by (i) measuring how q fits a sample $\boldsymbol{x}$ via the expectation of a loss $\ell_n(\boldsymbol{\theta}, \boldsymbol{x})$, (ii) quantifying uncertainty about $\boldsymbol{\theta}^*$ via a divergence D between prior π and q, (iii) optimizing q over a space of probability distributions Π on Θ.

Axiom 2 (Information Difference)

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

Axiom 3 (Prior Regularization)

q is regularized against π by penalizing the divergence $D(q||\pi)$.

Axiom 4 (Translation Invariance)

For constant C and $\ell'_n = \ell_n + C$, $P(\ell'_n, D, \Pi) = P(\ell_n, D, \Pi)$.

3.3 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 $L(q|\boldsymbol{x}, \ell_n, D) = f\left(\mathbb{E}_{q(\boldsymbol{\theta})}[\ell_n(\boldsymbol{\theta}, \boldsymbol{x})], D(q||\pi)\right)$, 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 3 and 4 only if $\circ = +$.

Implications/relevance:

- Bayesian inference = constrained, regularized optimization
- Objective only depends on $\mathbb{E}_{q(\boldsymbol{\theta})}[\ell_n(\boldsymbol{\theta}, \boldsymbol{x})]$ and $D(q||\pi)$
- For elementary $f\left(\mathbb{E}_{q(\boldsymbol{\theta})}[\ell_n(\boldsymbol{\theta}, \boldsymbol{x})], D(q||\pi)\right)$, f must be addition.

  (**Note:** Axiom 4 excludes most non-elementary f)

3.4 Generalized Bayesian problem & existing methods I/III

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

$P(\ell_n, D, \boldsymbol{\Pi})$ 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)

3.4 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]	$-\log(p(\boldsymbol{\theta}\|x_i))$	$\frac{1}{w}$KLD, $w > 1$	$\mathcal{P}(\boldsymbol{\Theta})$
Divergence Bayes[3]	divergence-based ℓ	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Standard VI	$-\log(p(\boldsymbol{\theta}\|x_i))$	**KLD**	$\mathcal{Q}$
Power VI[4]	$-\log(p(\boldsymbol{\theta}\|x_i))$	$\frac{1}{w}$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 ℓ**	**any 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)

3.4 Generalized Bayesian problem & existing methods III/III

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

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

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

- F = Rényi's α-divergence (Li and Turner, 2016; Saha et al., 2017)
- F = χ-divergence (Dieng et al., 2017)
- F = operators (Ranganath et al., 2016)
- F = scaled AB-divergence (Regli and Silva, 2018)
- F = 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 **F-VI**, with F = (local) reverse KLD

3. Summary: The Rule of Three

(1) The Rule of Three, i.e. $P(\ell_n, D, \Pi)$ is a natural generalization of Bayes rule

(2) $P(\ell_n, D, \Pi)$ is **modular**!

(3) While intuitively appealing, $P(\ell_n, D, \Pi)$ can also be derived **axiomatically**!

(4) Existing methods are recovered – even approximate ones! This does include **VI**, but **not** **F-VI** (Part 2 for more on this)

Discussion / Q&A

If you have questions that you think are stupid, I am sure they are not – here are some questions to compare against that people actually asked on Reddit:

- *I was bitten by a turtle as a young lad, can I still drink orange juice?*
- *I made Jesus-shaped pancakes but burnt them – am I going to hell?*
- *Wtf is Obama's last name? Does anyone know?*
- *Is an egg a fruit or a vegetable?*
- ...

Main References i

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.

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559.

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.

Cichocki, A. and Amari, S.-i. (2010). Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568.

Dawid, A. P., Musio, M., and Ventura, L. (2016). Minimum scoring rule inference. *Scandinavian Journal of Statistics*, 43(1):123–138.

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.

Ferreira, D. P. L., Backes, A. R., and Barcelos, C. A. Z. (2015). Bregman divergence applied to hierarchical segmentation problems. In *Iberoamerican Congress on Pattern Recognition*, pages 493–500. Springer.

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.

Main References ii

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.

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

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

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

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*.

Rényi, A. et al. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

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

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.