# Optimization-centric Generalizations of Bayesian Inference

Generalized Variational Inference & beyond

November 5, 2021

Biometrika Fellow @ Department of Statistical Science, UCL
Visiting Researcher @ The Alan Turing Institute for Data Science and AI

# What you should take away from today

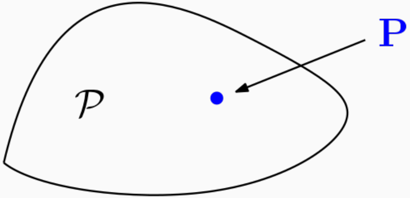
(1) Bayesian inference = a **single** optimization problem $\boldsymbol{P} \in \mathcal{P}$

(2) Optimization problem $\boldsymbol{P} \in \mathcal{P}$ needs **strict assumptions**

(3) These assumptions are often violated in Machine Learning (ML)

(4) **Many** other optimization problems $\boldsymbol{P}' \in \mathcal{P}$ you could solve instead

(5) In ML, we would often do better to solve $\boldsymbol{P}' \in \mathcal{P}$ rather than $\boldsymbol{P}$



# Structure

(1) **All about that Bayes**
- (1.1) Bayes' rule
- (1.2) Assumptions of Bayesian inference
- (1.3) Machine Learning, Assumptions & Bayesian inference

(2) **An optimization-centric generalization**
- (2.1) Bayesian inference **is** optimization
- (2.2) Modularity & Interpretation
- (2.3) A natural generalization
- (2.4) Special cases of note
- (2.5) Generalized Variational Inference

(3) **Properties of this generalization**
- (3.1) 'Sanity checks': existence, uniqueness, consistency
- (3.2) Axiomatic foundations
- (3.3) Relation to VI, PAC-Bayes bounds, power posteriors, ...
- (3.4) Closed Forms

(4) **Applications**
- (4.1) Robustness: prior misspecification
- (4.2) Robustness: likelihood misspecification
- (4.3) Simplifying Bayesian computation

# Literature I will talk about

**Main focus (parts 1+2):**

JK, Jack Jewson, & Theo Damoulas; *Generalized Variational Inference: Three Arguments for deriving new Posteriors*, currently minor revisions at JMLR https://arxiv.org/abs/1904.02063

**literature I touch upon (parts 3+4):**

Takuo Matsubara, JK, Francois-Xavier Briol, & Chris Oates; *Generalised Bayesian Inference with Stein Discrepancies: Robust Bayes for Models with an Intractable Likelihood*, submitted to JRSS-B 2021 https://arxiv.org/abs/2104.07359

JK; *Frequentist Consistency of Generalized Variational Inference*, 2019 https://arxiv.org/abs/1904.04946

JK; *Robust Deep Gaussian Processes*, 2019 https://arxiv.org/abs/1904.02303

JK, Jack Jewson, & Theo Damoulas; *Doubly Robust Bayesian Inference for Non-Stationary Streaming Data using $\beta$-Divergences*, NeurIPS 2018 https://arxiv.org/abs/1806.02261

Sebastian Schmon, Patrick Cannon, & JK; *Generalized Posteriors in Approximate Bayesian Computation*, AABI 2021 https://arxiv.org/abs/2011.08644

Pierre Alquier; *Non-exponentially weighted aggregation: regret bounds for unbounded losses*, 2020 https://arxiv.org/abs/2009.03017

**Non-exhaustive list of important literature I do not touch upon:**

Alexander A. Alemi; *Variational Predictive Information Bottleneck*; AABI 2019 https://arxiv.org/abs/1910.10831

Jeffrey Miller; *Asymptotic normality, concentration, and coverage of generalized posteriors*, working paper 2019 (arxiv 1907.09611)

Badr-Eddine Chérief-Abdellatif; *Contributions to the theoretical study of variational inference and robustness*, PhD thesis 2020

...

# (1) All about that Bayes

Not actually Rev. Thomas Bayes

Not actually the form of the result in Bayes (1763), but in Laplace (1774)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

# (1.1) Bayes' Rule: Interpretation as belief updates

**Ingredients**:

- $n$ observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$,
- **prior** $\pi(\boldsymbol{\theta})$,
- **likelihood terms** $\{p(x_i|\boldsymbol{\theta})\}_{i=1}^n$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

**Output (via Bayes' Rule) = posterior belief**:

$$\underbrace{q_n^*(\boldsymbol{\theta})}_{\textbf{posterior}} \propto \underbrace{\pi(\boldsymbol{\theta})}_{\textbf{prior}} \prod_{i=1}^n \underbrace{p(x_i|\boldsymbol{\theta})}_{\textbf{likelihood terms}} \tag{1}$$

**Inference interpretation = belief updates**:

$$\pi(\boldsymbol{\theta}) \xrightarrow{\text{Update with Bayes' rule via } \{p(x_i|\boldsymbol{\theta})\}_{i=1}^n} q_n^*(\boldsymbol{\theta})$$

# (1.2) Three assumptions underlying Bayesian inference

1.) Likelihoods are correct, i.e. $p(x_i|\theta^*) \overset{!}{=} d\mathbb{P}(x_i)$ for some unknown $\theta^*$

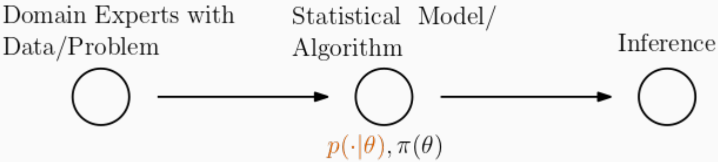
2.) Prior encodes the complete prior information on $\theta^*$

3.) Closed forms/Infinite computational power,
so that one can work with the posterior.

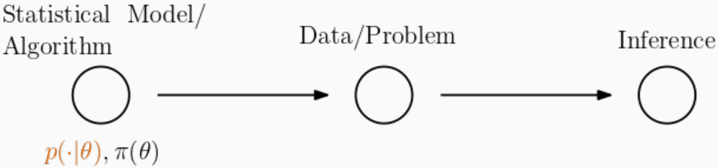
$$q_n^*(\theta) \propto \pi(\theta) \prod_{i=1}^n p(x_i|\theta)$$

# (1.3) Machine Learning & Bayesian inference

## Traditional role of Statistics in science:




## Modern Machine Learning:



# (1.3) Machine Learning & Bayesian inference

**Conclusion I:** Assumptions of Bayesian inference fit traditional science

(1) **Correct Likelihoods:** Domain Experts help design these

(2) **Priors encoding complete information:** Experiments/studies build on **prior** work, which can often be summarized in distributions.

(Prior elicitation is an entire field within Bayesian statistics.)

(3) **Closed forms/infinite computational power:** Data collection is expensive; computational cost much less important

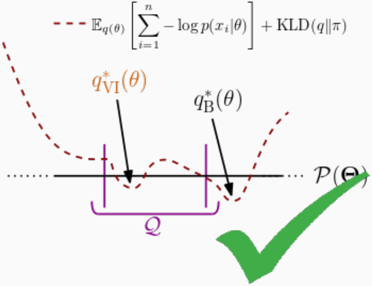
**Conclusion II:** ML violates assumptions of Bayesian Inference

(1) **Correct Likelihoods:** Likelihoods are defined before data is seen

(2) **Priors encoding complete information:** Priors often impossible to elicit (e.g., Bayesian Neural Networks)

(3) **Closed forms/infinite computational power:** Violated by virtually any ML application


Posterior
Prior
Likelihood


$\mathbb{E}_{q(\theta)}\left[\sum_{i=1}^{n} -\log p(x_i|\theta)\right] + \text{KLD}(q\|\pi)$
$q^*_{\text{VI}}(\theta)$
$q^*_{\text{B}}(\theta)$
$\mathcal{P}(\Theta)$
$\mathcal{Q}$

The Bayes posterior $q_n^*(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^n p(x_i|\boldsymbol{\theta})$ uniquely solves

$$q_n^*(\boldsymbol{\theta}) = \underset{q \in \mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \underbrace{\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^n -\log(p(x_i|\boldsymbol{\theta}))\right]}_{\text{minimized by } q(\boldsymbol{\theta}) = \delta_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{\theta}),\ \hat{\boldsymbol{\theta}}_n = \text{MLE}} + \underbrace{\text{KLD}\left(q||\pi\right)}_{\text{minimized by } q = \pi} \right\}, \quad (2)$$

**Notation:**

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

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

# (2.2) Modularity and Interpretation

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

3.) Reduce computational cost by optimizing over a smaller set $\Pi \subset \mathcal{P}(\Theta)$

(= Variational Inference!)

2.) Guard against likelihood misspecification by using robust model scoring rules $\mathcal{L}(\cdot)$ instead of $-\log(\cdot)$

1.) Suppress ill-informed priors by choosing a weaker regularizing divergence $D(\cdot\|\pi)$ instead of $\text{KLD}(\cdot\|\pi)$

# (2.3) A natural generalization

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

**An optimization-centric generalization of Bayesian inference:**

(1) $\boldsymbol{D}(\cdot||\pi)$ = **any divergence regularizer** penalizing deviations from $\pi$

(2) $\ell(\boldsymbol{\theta}, \boldsymbol{x})$ = **any loss** assessing how well $\boldsymbol{\theta}$ and $\boldsymbol{x}$ fit together

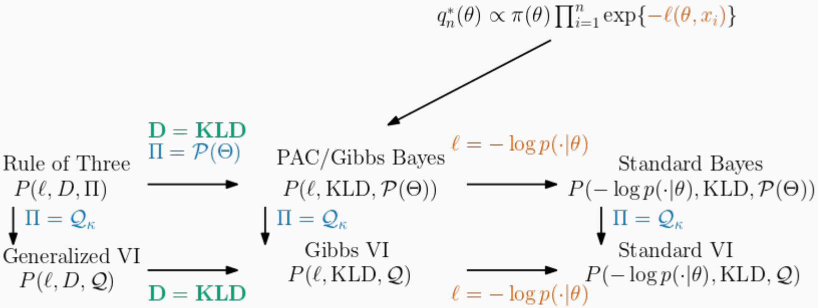
(3) $\boldsymbol{\Pi} \subseteq \mathcal{P}(\boldsymbol{\Theta})$ = **some subset** of all probability distributions on $\boldsymbol{\Theta}$

$\Longrightarrow$ **Shorthand Notation:** $P(\boldsymbol{\ell}, \boldsymbol{D}, \boldsymbol{\Pi})$

# (2.4) Special cases of note

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

**(Some) special cases of interest:**



# (2.5) Generalized Variational Inference

**Terminology:**

- Presented so far: a conceptual generalization of Bayesian inference
- **Generalized Variational Inference (GVI)** $=$ computable inference algorithms based on that generalization
- $\implies$ means that $\Pi = \mathcal{Q}$ for some variational family $\mathcal{Q}$.

**Open questions:**

- General properties of this generalisation?
- Applications/use cases?

$\implies$ answered next

# (3) General features of the generalization

**Answers to three questions:**

(3.1) 'Is this even reasonable?' (sanity checks)

(3.2) 'Okay, but under which conditions is it reasonable?' (axiomatic justification)

(3.3) 'Fine, but can it help me understand existing methods?' (VI, PAC-Bayes bounds, power posteriors, ...)

# (3.1) 'Sanity checks': existence, uniqueness, consistency

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

**Existence:** If $\boldsymbol{\Pi}$ is convex and chosen so that $q \mapsto \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \sum_{i=1}^{n} \boldsymbol{\ell}(\boldsymbol{\theta}, x_i) \right]$ and $q \mapsto \boldsymbol{D}(q||\pi)$ are continuous on $\boldsymbol{\Pi}$, then the minimum exists whenever $q \mapsto \boldsymbol{D}(q||\pi)$ is convex.
$\Longrightarrow$ basic convex analysis on Banach spaces

**Uniqueness:** Guaranteed if $q_n^*$ exists and $q \mapsto \boldsymbol{D}(q||\pi)$ is *strictly* convex
$\Longrightarrow$ basic convex analysis on Banach spaces

# (3.1) 'Sanity checks': existence, uniqueness, consistency

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

**Beyond continuity & convexity** of $q \mapsto \boldsymbol{D}(q||\pi)$? You can still show existence and uniqueness without convex regularizers [need that losses are norm-coercive or $\Theta$ is compact; Arguments not that basic; see Lemma 1 in Knoblauch (2019)]

**Consistency:** Guaranteed under relatively mild regularity conditions; arguments are unfortunately quite complicated and rely on $\Gamma$-convergence, see Knoblauch (2019).
$\Longrightarrow$ Idea of $\Gamma$-convergence $\approx$ 'if a sequence of objectives $\Gamma$-converge, then their minimizers also converge (in a suitable sense, usually weakly)'

# (3.2) Axiomatic justification

**Q: Does this generalisation result from 'reasonable' axioms?**

**Axiom 1 (Variational representation)**

The posterior $q^* \in \mathcal{P}(\boldsymbol{\Theta})$ solves an optimization problem over some space $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$. For any finite sample $\{x_i\}_{i=1}^n$, the optimization problem seeks to jointly minimize two criteria:

(i) An 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 some statistical divergence $D$.

**Theorem 1 (Form 1)**

*Under Axiom 1, posterior belief distributions can be written as*

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

*where $f : \mathbb{R}^2 \to \mathbb{R}$ is some function that may depend on $\pi, \Pi, \ell, \{x_i\}_{i=1}^n$, or $D$.*

**Axiom 2 (Recovers Bayesian Posteriors)**

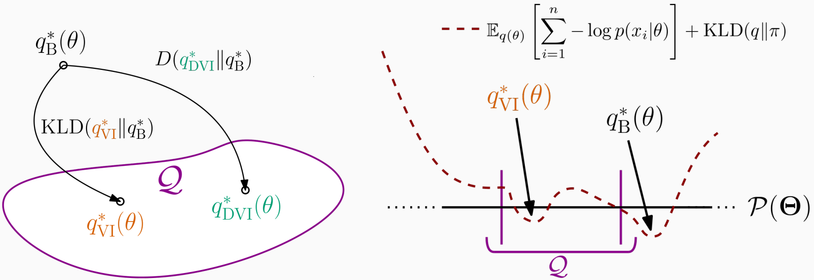
Function $f$ in Theorem 1 does not depend on $\pi, \Pi, \ell, \{x_i\}_{i=1}^n$, or $D$. Further, $q^*$ is the Gibbs posterior if $D = \mathrm{KLD}$, $\Pi = \mathcal{P}(\boldsymbol{\Theta})$.

**Theorem 2**

*Suppose the posterior belief $q^* \in \mathcal{P}(\boldsymbol{\Theta})$ satisfies Axioms 1 and 2. Then the objective of Theorem 1 is uniquely identified as $f(x, y) = x + y$ so that*

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

# (3.3) Relation to VI / PAC-Bayes / power posteriors




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

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

**Proposition 1 (Optimality of standard VI)**

Relative to the infinite-dimensional optimization problem over $\mathcal{P}(\boldsymbol{\Theta})$ characterizing the Gibbs posterior and a fixed variational family $\boldsymbol{\Pi}$, standard VI produces the optimal solution (i.e. posterior belief) in $\boldsymbol{\Pi}$.

**Proof.**

VI posteriors are minima of the same objective as the full Bayesian posterior—but constrained to some subset $\boldsymbol{\Pi}$. □

**Proposition 2 (Suboptimality of alternative methods)**

Relative to the infinite-dimensional problem over $\mathcal{P}(\boldsymbol{\Theta})$ characterizing Gibbs posteriors, and relative to a fixed finite-dimensional variational family $\boldsymbol{\Pi}$, non-standard VI methods produce sub-optimal solutions (i.e., posterior beliefs).

# (3.3) Relation to VI / PAC-Bayes / power posteriors

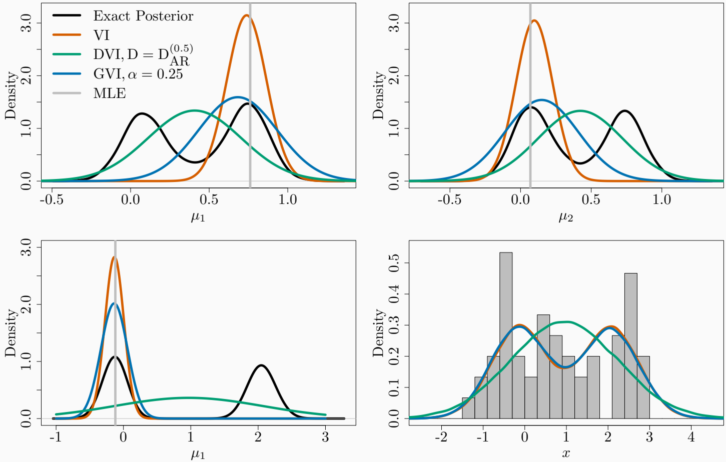



**Figure 2 – Top row:** marginal posteriors for location parameters $\mu_1, \mu_2$ in 2D Bayesian Gaussian Mixture Model. **Bottom left:** same marginal for $\mu_1$ as we increase $|\mu_1 - \mu_2|$. **Bottom right:** posterior predictives.

# (3.3) Relation to VI / PAC-Bayes / power posteriors

**E.g., McAllester's (original) PAC-Bayes bound:** if $x_i \overset{iid}{\sim} \mu$ so that true risk is $R(\boldsymbol{\theta}) = \mathbb{E}_{x\sim\mu}[\ell(\boldsymbol{\theta}, x)]$ and $a \leq \ell \leq b$, then uniformly for all $q \in \mathcal{P}(\Theta)$ with probability at least $1 - \varepsilon$,

$$\mathbb{E}_{q(\boldsymbol{\theta})}\left[R(\boldsymbol{\theta})\right] \leq \mathbb{E}_{q(\boldsymbol{\theta})}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, x_i)\right] + \sqrt{\frac{\mathrm{KLD}(q, \pi) + \log\frac{2\sqrt{n}}{\varepsilon}}{2n}}.$$

I.e., we can minimize the righthand side over $q \in \mathcal{P}(\Theta)$ to find that **the tightest generalisation bound** has a solution of the form $P(\boldsymbol{\ell}, \boldsymbol{D}_{McA}, \mathcal{P}(\Theta))$ with

$$D_{\mathrm{McA}}(q\|\pi) = \sqrt{n}\cdot\left(\sqrt{\frac{\mathrm{KLD}(q, \pi) + \log\frac{2\sqrt{n}}{\varepsilon}}{2}} - \sqrt{\frac{\log\frac{2\sqrt{n}}{\varepsilon}}{2}}\right).$$

$\Longrightarrow$ PAC-Bayes $\overset{!}{=}$ a way to justify non-standard prior regularization, e.g. Bégin et al. (2016) [Rényi's $\alpha$-divergence], Alquier & Guedj (2018), Ohnishi & Honorio (2020) [$f$-divergences]

# (3.3) Relation to VI / PAC-Bayes / power posteriors

**Power posteriors/'cold posteriors':** For some $\beta > 0$, given by

$$q^*(\theta) \propto \pi(\theta) \prod_{i=1}^{n} p(x_i|\theta)^\beta = P(-\log p(\cdot|\boldsymbol{\theta}) \cdot \beta, \mathrm{KLD}, \mathcal{P}(\Theta)) \qquad (3)$$

$$= P\left(-\log p(\cdot|\boldsymbol{\theta}), \mathrm{KLD} \cdot \frac{1}{\beta}, \mathcal{P}(\Theta)\right)$$

**Cold posteriors:** $\beta > 1$, i.e. more weight on data rather than prior
$\Longrightarrow$ often used for NNs, where our priors are extremely poor.
**Power posteriors:** $\beta < 1$, i.e. more weight on prior rather than data
$\Longrightarrow$ often used when likelihoods are poorly misspecified and advertised as 'robust' (... it's always more robust as $n \to \infty$, but if $n$ is small, your robustness properties will depend a lot on the prior)

# (3.3) Relation to VI / PAC-Bayes / power posteriors /...

| Method | $\ell(\boldsymbol{\theta}, x_i)$ | $D$ | $\Pi$ |
|---|---|---|---|
| Standard Bayes | $-\log p(x_i \mid \boldsymbol{\theta})$ | KLD | $\mathcal{P}(\boldsymbol{\Theta})$ |
| Power Likelihood Bayes | $-\log p(x_i \mid \boldsymbol{\theta})$ | $\frac{1}{w}$ KLD, $w < 1$ | $\mathcal{P}(\boldsymbol{\Theta})$ |
| Composite Likelihood Bayes | $-w_i \log p(x_i \mid \boldsymbol{\theta})$ | KLD | $\mathcal{P}(\boldsymbol{\Theta})$ |
| Divergence-based Bayes | divergence-based $\ell$ | KLD | $\mathcal{P}(\boldsymbol{\Theta})$ |
| PAC/Gibbs Bayes | any $\ell$ | any $D$ | $\mathcal{P}(\boldsymbol{\Theta})$ |
| VAE | $-\log p_{\boldsymbol{\zeta}}(x_i \mid \boldsymbol{\theta})$ | KLD | $\mathcal{Q}$ |
| $\beta$-VAE | $-\log p_{\boldsymbol{\zeta}}(x_i \mid \boldsymbol{\theta})$ | $\beta \cdot$ KLD, $\beta > 1$ | $\mathcal{Q}$ |
| Bernoulli-VAE | continuous Bernoulli | KLD | $\mathcal{Q}$ |
| **Standard VI** | $-\log p(x_i \mid \boldsymbol{\theta})$ | KLD | $\mathcal{Q}$ |
| Power VI | $-\log p(x_i \mid \boldsymbol{\theta})$ | $\frac{1}{w}$ KLD, $w < 1$ | $\mathcal{Q}$ |
| Utility VI | $-\log p(x_i \mid \boldsymbol{\theta}) + \log u(h, x_i)$ | KLD | $\mathcal{Q}$ |
| Regularized Bayes | $-\log p(x_i \mid \boldsymbol{\theta}) + \phi(\boldsymbol{\theta}, x_i)$ | KLD | $\mathcal{Q}$ |
| Gibbs VI | any $\ell$ | KLD | $\mathcal{Q}$ |
| **Generalized VI** | any $\ell$ | any $D$ | $\mathcal{Q}$ |

# (3.4) Closed Forms

**Well-known:** Gibbs posteriors

$$P(\boldsymbol{\ell}, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta})) \propto \pi(\boldsymbol{\theta}) \cdot \exp\left\{-\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\right\}$$

**Unknown (until a few months ago):** What if $D \neq$ KLD?

$\Longrightarrow$ We now know the general form if $D$ an $\phi$-divergence

**Proposition 3.1.** *Assume that $\phi$ is differentiable, strictly convex and define $\tilde{\phi}$ on $\mathbb{R}$ by $\tilde{\phi}(x) = \phi(x)$ if $x \geq 0$ and $\tilde{\phi}(x) = +\infty$ otherwise. Then*

$$\tilde{\phi}^* = \sup_{x \in \mathbb{R}}[xy - \tilde{\phi}(x)] = \sup_{x \geq 0}[xy - \phi(x)] \tag{3.1}$$

*is differentiable and for any $y \in \mathbb{R}$,*

$$\nabla\tilde{\phi}^*(y) = \operatorname*{argmax}_{x \geq 0}\{xy - \phi(x)\}. \tag{3.2}$$

*Assume moreover than $\tilde{\phi}^*(\lambda - a) - \lambda \to \infty$ when $\lambda \to \infty$, for any $a \geq 0$. Then*

$$\lambda_t \in \operatorname*{argmin}_{\lambda \in \mathbb{R}}\left\{\int \tilde{\phi}^*\left(\lambda - \eta\sum_{s=1}^{t-1}\ell_s(\theta)\right)\pi(\mathrm{d}\theta) - \lambda\right\} \tag{3.3}$$

*exists, and*

$$\rho^t(\mathrm{d}\theta) = \nabla\tilde{\phi}^*\left(\lambda_t - \eta\sum_{s=1}^{t-1}\ell_s(\theta)\right)\pi(\mathrm{d}\theta) \tag{3.4}$$

*minimizes* (1.4).

## Two examples:

**Example 3.2** ($\chi^2$-divergence). *We come back to the example* $\phi(x) = x^2 - 1$, $D_\phi(\rho||\pi) = \chi^2(\rho||\pi)$ *the chi-squared divergence. In this case,* $\phi^*(y) = (y^2/4)\mathbf{1}_{\{y\geq 0\}}$ *and so* $\nabla\tilde{\phi}^*(y) = (y/2)_+$. *This leads to*

$$\rho^t(\mathrm{d}\theta) = \left[\frac{\lambda_t - \eta\sum_{s=1}^{t-1}\ell_s(\theta)}{2}\right]_+ \pi(\mathrm{d}\theta). \tag{3.7}$$

*In this case,* $\lambda_t$ *is not available in closed form, but it is the only constant that will make the above sum to* 1.

**Example 3.3** ($p$-power divergence). *More generally, consider* $\phi(x) = x^p - 1$. *In this case* $\nabla\tilde{\phi}^*(y) = (y/p)_+^{1/(p-1)}$. *This leads to*

$$\rho^t(\mathrm{d}\theta) = \left[\frac{\lambda_t - \eta\sum_{s=1}^{t-1}\ell_s(\theta)}{p}\right]_+^{\frac{1}{p-1}} \pi(\mathrm{d}\theta). \tag{3.8}$$

# (4) Applications

**Problems that can be tackled:**

- Robustness to poor priors
- Robustness to poor likelihoods
- Simplified computation
- ...

# (4.1) Robustness to prior misspecification

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

**What we want:** $D$ that behaves like $\text{KLD}$ if $\pi$ is reasonable, but ignores it if the data don't fit the prior at all

**First idea:** down-weight $D = \text{KLD}$ like in cold posteriors

$\Longrightarrow$ Problem: Now, not being certain of your prior amounts to being more certain in your posterior....

**Q:** Is there an alternative?

$\Longrightarrow$ For reasons we don't understand fully$^{\dagger}$, Rényi's $\alpha$-divergence seems to do behave exactly as we want! (small loss of efficiency observed if prior is well-specified)

$^{\dagger}$some limited theoretical results in Theorem 14 of Knoblauch et al. (2019)

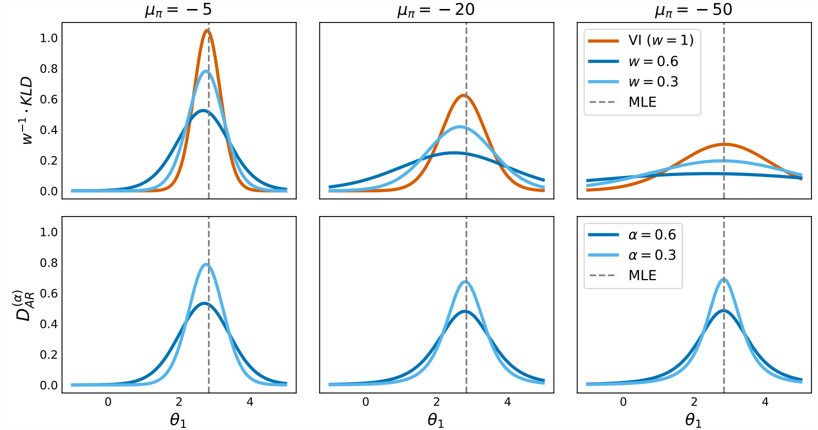



**Figure 3** – The prior for the coefficients is a Normal Inverse Gamma distribution given by $\boldsymbol{\mu} \sim \mathcal{NI}^{-1}(\mu_\pi \cdot 1_d, v_\pi \cdot I_d, a_\pi, b_\pi)$ with $v_\pi = 4 \cdot I_d$, $a_\pi = 3$, $b_\pi = 5$ and various values for $\mu_\pi$. For all posteriors, the loss $\ell$ is the correctly specified negative log likelihood; all posteriors lie inside a mean field normal family $\mathcal{Q}$.

# (4.1) Robustness to prior misspecification

**Example:** Bayesian Neural Networks

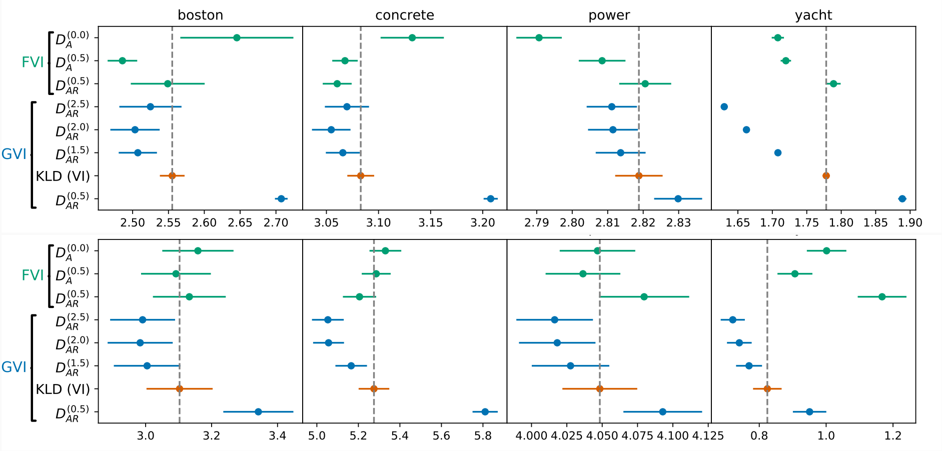



**Figure 4 – Top**: Negative test log likelihoods. **Bottom row**: Test RMSE.

# (4.2) Robustness to likelihood misspecification

**What it means:** [Taking definition from Hooker & Vidyashankar]

Consider $\varepsilon$*-contamination model* of size $\varepsilon \in (0, 1)$

$$\mathbb{P}_{n,\epsilon,y} = (1-\epsilon)\mathbb{P}_n + \epsilon\delta_y; \quad y \in \mathcal{X}$$

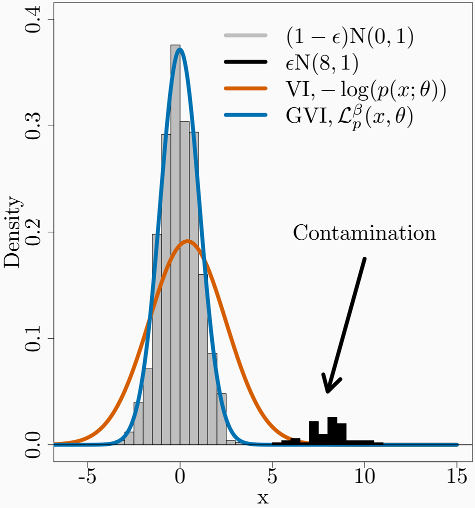
and write $\ell_{\varepsilon,n}(\boldsymbol{\theta}) = \mathbb{E}_{x\sim\mathbb{P}_{n,\epsilon,y}}[\ell(\boldsymbol{\theta}, x)]$. Define the *posterior influence function* as

$$\text{PIF}(y, \theta, \mathbb{P}_n) := \frac{\mathrm{d}}{\mathrm{d}\epsilon} P(\ell_{\varepsilon,n}(\boldsymbol{\theta}), \boldsymbol{D}, \boldsymbol{\Pi})|_{\epsilon=0}.$$

The posterior $P(\ell_{\varepsilon,n}(\boldsymbol{\theta}), \boldsymbol{D}, \boldsymbol{\Pi})$ is called *globally bias-robust* if $\sup_{\theta\in\Theta}\sup_{y\in\mathcal{X}} |\text{PIF}(y, \theta, \mathbb{P}_n)| < \infty$, meaning that the sensitivity of the generalised posterior to the contaminant $y$ is limited.

# (4.2) Robustness to likelihood misspecification

**What it means:**



# (4.2) Motivation for Discrepancy-based losses

**General Observation: log likelihoods $\approx$ minimize the KLD**

$$
\begin{aligned}
q_n^*(\boldsymbol{\theta}) &= \underset{q \in \boldsymbol{\Pi}}{\arg\min} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \sum_{i=1}^{n} - \log p(x_i|\boldsymbol{\theta}) \right] + \mathrm{KLD}\left(q||\pi\right) \right\} \\
&= \underset{q \in \boldsymbol{\Pi}}{\arg\min} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \frac{1}{n} \sum_{i=1}^{n} - \log \frac{p(x_i|\boldsymbol{\theta})}{p_0(x_i)} \right] - \frac{1}{n} \sum_{i=1}^{n} \log p_0(x_i) + \frac{1}{n} \mathrm{KLD}\left(q||\pi\right) \right. \\
&= \underset{q \in \boldsymbol{\Pi}}{\arg\min} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \underbrace{\left[ \frac{1}{n} \sum_{i=1}^{n} - \log \frac{p(x_i|\boldsymbol{\theta})}{p_0(x_i)} \right]}_{\approx \mathrm{KLD}(p_0||p(\cdot|\boldsymbol{\theta}))} + \frac{1}{n} \mathrm{KLD}\left(q||\pi\right) \right\}
\end{aligned}
$$

**Obvious question:** What are the (dis)advantages of minimizing other discrepancies between $p_0$ and $p(\cdot|\boldsymbol{\theta})$ instead?

**General Answer:** Decreases statistical efficiency, increases robustness

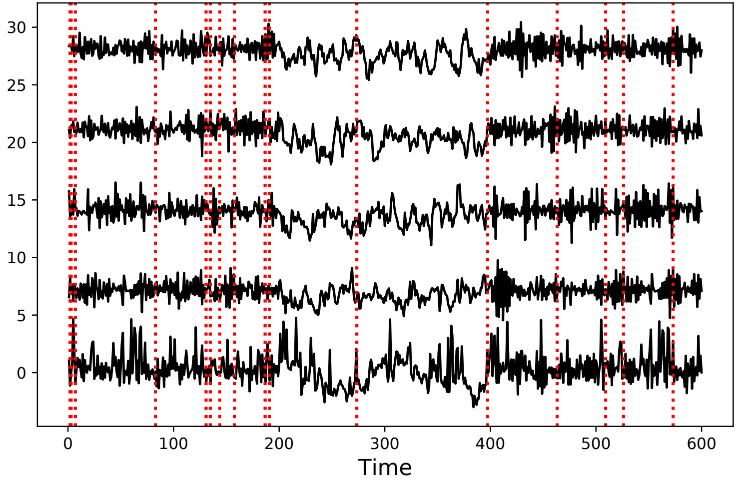
# (4.2) Robustness to likelihood misspecification

**Some applications of robust divergences as losses:**

- Reducing false detection of changepoints
- Improving performance of deep Gaussian Processes
- Graphical models
- ...

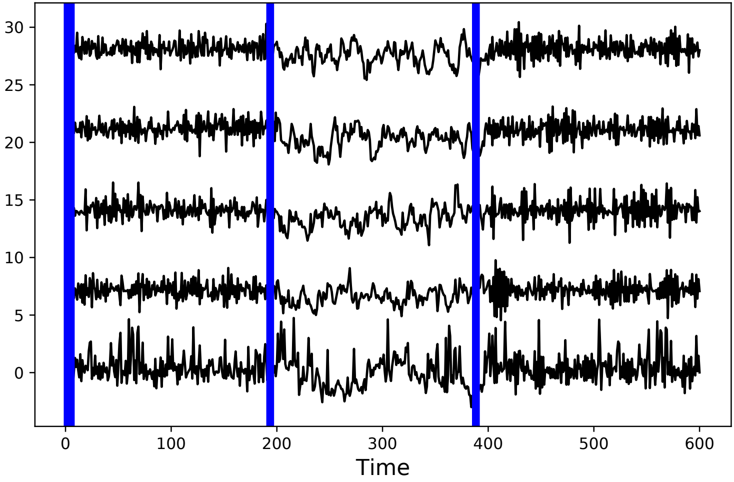
Many, many more... — worth considering every time your likelihood is at best a reasonable guess.

## Using standard Bayesian On-line Changepoint Detection



# (4.2) Outlier-Robust Changepoint Detection

## Using the $\beta$-divergence for Robust Changepoint Detection



# (4.2) Likelihoods with Deep Gaussian Processes

## $\gamma$-divergences for DGP regression[3]

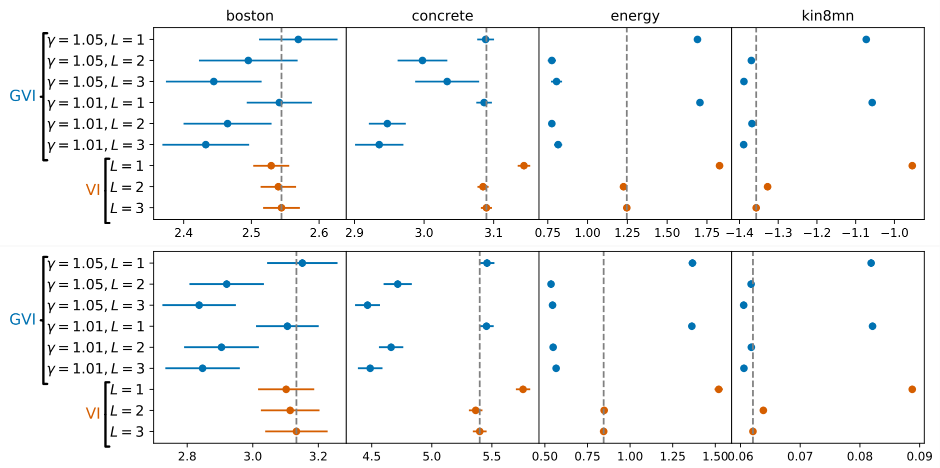



**Figure 5 – Top row**: Negative test log likelihoods. **Bottom row**: Test RMSE.

# (4.3) Improving Bayesian computation

**Idea:** Find losses $\tilde{\ell} \neq \ell$ s.t.
$\text{CompTime}\left[P(\tilde{\ell}, \boldsymbol{D}, \boldsymbol{\Pi})\right] << \text{CompTime}\left[P(\ell, \boldsymbol{D}, \boldsymbol{\Pi})\right]$

**Example 1:** Approximate Bayesian Computation (ABC).
**Step 1:** $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$
**Step 2:** sample $x_{\text{fake}} \sim p(x_{\text{fake}}|\boldsymbol{\theta})$
**Step 3:** keep $\boldsymbol{\theta}$ if $D(x_{\text{fake}}, x_{\text{observed}}) < \varepsilon$; discard otherwise.
$\Longrightarrow$

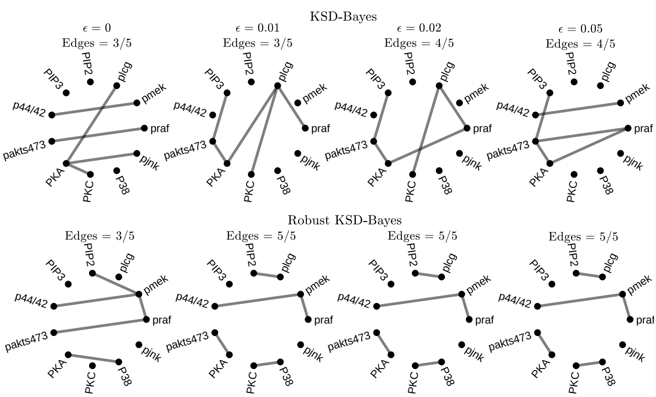
$$q(\theta|x_{\text{observed}}) \approx q_{\text{abc}}(\theta|x_{\text{observed}}) \propto \int 1_{[D(x_{\text{fake}}, x_{\text{observed}})<\varepsilon]} p(x_{\text{fake}}|\theta)\pi(\theta) dx_{\text{fake}}.$$

**Usually:** $q_{\text{abc}}(\theta|x_{\text{observed}})$ interpreted as approximation to $q(\theta|x_{\text{observed}})$.

**Alternative:** $q_{\text{abc}}(\theta|x_{\text{observed}}) \propto \pi(\theta) \exp\{-\underbrace{L_n(\theta, x_{\text{observed}})}_{=\int 1_{[D(x_{\text{fake}}, x_{\text{observed}})<\varepsilon]} p(x_{\text{fake}}|\theta)}\}$

$\Longrightarrow$ simulator error model $err(x_{\text{fake}}, x_{\text{observed}}) = \int 1_{[D(x_{\text{fake}}, x_{\text{observed}})<\varepsilon]}$
$\Longrightarrow$ 'smoother' error models more sample-efficient (e.g., $err =$normal)

# (4.3) Improving Bayesian computation



# (4.3) Improving Bayesian computation

**Idea:** Find losses $\tilde{\ell} \neq \ell$ s.t.
$\text{CompTime}\left[P(\tilde{\ell}, \boldsymbol{D}, \boldsymbol{\Pi})\right] << \text{CompTime}\left[P(\ell, \boldsymbol{D}, \boldsymbol{\Pi})\right]$

**Example 2:** Intractable Likelihoods/Energy-based models
**Challenge:** Likelihoods with unknown normalisers, i.e.

$$p(x|\boldsymbol{\theta}) = \underbrace{\widehat{p}(x|\boldsymbol{\theta})}_{\text{known}} \cdot \underbrace{Z(\boldsymbol{\theta})}_{\text{unknown}}$$

$\Longrightarrow$ standard Bayesian posterior 'doubly intractable'

**Solution:** loss depending on $p(x|\boldsymbol{\theta})$ *only via* $\nabla_x p(x|\boldsymbol{\theta}) \stackrel{!}{=} \nabla_x \widehat{p}(x|\boldsymbol{\theta})$
$\Longrightarrow$ Stein's method! Operationalisable via Kernel Stein Discerpancies
$\Longrightarrow$ Makes posteriors 'singly intractable'

. . . **Question:** but can we do even better?!

# (4.3) Improving Bayesian computation

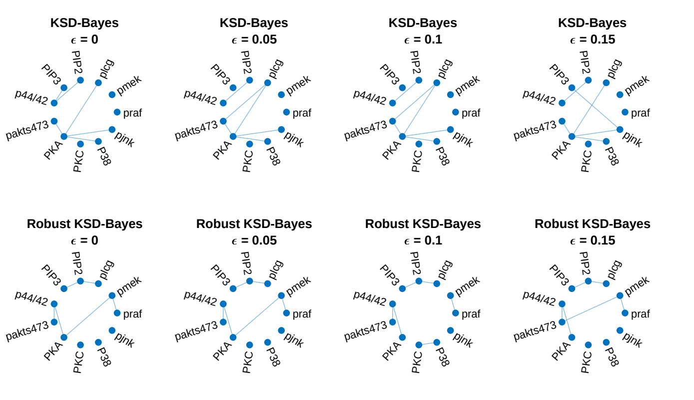
**Answer:** Yes, whenever $p(x|\boldsymbol{\theta})$ is part of the exponential family!

$\Longrightarrow$ Then, we get **closed forms** if $\pi$ is normal!

$\Longrightarrow$ Closed forms instead of 'doubly intractable' posteriors

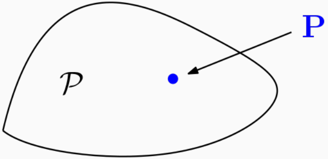
$\Longrightarrow$ Added bonus: robustness to model misspecification!

$$p(x|\theta) \propto \exp\left(-\sum_i \theta_i \exp(x_i) - \sum_{i<j} \theta_{i,j} \exp(x_i)\exp(x_j)\right) \times \exp\left(\sum_i x_i\right)$$



# Summary

(1) Bayesian inference = a **single** optimization problem $\boldsymbol{P} \in \mathcal{P}$
(2) Optimization problem $\boldsymbol{P} \in \mathcal{P}$ needs **strict assumptions**
(3) These assumptions are often violated in Machine Learning (ML)
(4) **Many** other optimization problems $\boldsymbol{P'} \in \mathcal{P}$ you could solve instead
(Even though we still know relatively little about them!)
(5) In ML, we would often do better to solve $\boldsymbol{P'} \in \mathcal{P}$ rather than $\boldsymbol{P}$




**Also—And perhaps even more importantly:**

- The study of $\boldsymbol{P'} \in \mathcal{P}$ has just begun! Get involved! :)
- If I managed to inspire you to work on these problems, get in touch! I love to collaborate, and there are enough open problems for a lifetime! :)

# Contact me/follow my research at

**E-mail address:** j.knoblauch@warwick.ac.uk

**Personal Website:** https://jeremiasknoblauch.github.io/

**Twitter Handle:** @LauchLab