MANCHESTER 1824
The University of Manchester

On the asymptotic behaviour of α–divergence-based generalized posteriors

Ioar Casado-Telletxea

Centre for AI Fundamentals, University of Manchester

Joint work with Omar Rivasplata

BayesAI, Edinburgh 28 October 2026

Why generalize Bayesian inference?

Bayesian inference combines data and prior knowledge to produce a distribution over parameters:

$$Q_n^\star(\theta) \propto \pi(\theta) \prod_{i=1}^{n} p_\theta(X_i).$$

- A natural framework for uncertainty quantification.
- Flexibility through the choice of prior.

Why generalize Bayesian inference?

Bayesian inference combines data and prior knowledge to produce a distribution over parameters:

$$Q_n^\star(\theta) \propto \pi(\theta) \prod_{i=1}^{n} p_\theta(X_i).$$

- A natural framework for uncertainty quantification.
- Flexibility through the choice of prior.

Practical challenges motivate alternative constructions: misspecification, computational limits...

A large toolbox: tempered posteriors, Gibbs posteriors, variational methods...

How do these modifications affect posterior uncertainty?

Bayesian inference as an optimization problem

With empirical average loss ℓ_n and prior Π,

$$\boxed{Q_n^\star \in \operatorname*{arg\,min}_{Q \in \mathcal{P}(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \ell_n(\theta)\, dQ(\theta) + \frac{\eta}{n} D(Q\|\Pi) \right\}.}$$

For $D = D_{\mathrm{KL}}$,

$$\frac{dQ_n^\star}{d\Pi}(\theta) \propto \exp\left\{ -\frac{n}{\eta} \ell_n(\theta) \right\}.$$

Negative log-likelihood loss and $\eta = 1$ recover ordinary Bayes (variational representation of Bayesian inference).

Generalized Variational Inference (GVI) framework: Knoblauch, Jewson and Damoulas (2022).

Bayesian inference as an optimization problem II

We have a good understanding of the solutions to the GVI problem in relevant scenarios:

- Tempered/Gibbs posteriors: Arbitrary ℓ_n, $\eta \neq 1$ and $D = D_{\mathrm{KL}}$.

  Grünwald (2012), Holmes and Walker (2017), Martin and Syring (2022), Kasprzak et al. (2025)...

- Variational families: $\mathcal{Q} \subset \mathcal{P}(\mathbb{R}^d)$.

  Wang and Blei (2019), Zhang and Gao (2020), Alquier and Ridgway (2020)...

Bayesian inference as an optimization problem II

We have a good understanding of the solutions to the GVI problem in relevant scenarios:

- Tempered/Gibbs posteriors: Arbitrary ℓ_n, $\eta \neq 1$ and $D = D_{\mathrm{KL}}$.

  Grünwald (2012), Holmes and Walker (2017), Martin and Syring (2022), Kasprzak et al. (2025)...

- Variational families: $\mathcal{Q} \subset \mathcal{P}(\mathbb{R}^d)$.

  Wang and Blei (2019), Zhang and Gao (2020), Alquier and Ridgway (2020)...

However, we lack a good theoretical understanding of the $D \neq D_{\mathrm{KL}}$ case

Knoblauch (2019), Mildner, Giampouras and Damoulas (2025), Sklaviadis et al. (2025, 2026)

What do we mean by theoretical understanding?

- Bernstein–von Mises' (BvM) theorem

  Bernstein (1917), von Mises (1919), Le Cam (1953), Van der Vaart (2000) Ch. 10

  Our presentation follows Miller (2021)

  Let $\theta_n \to \theta_0$ and let $s_n \to 0$ be a concentration scale. If q_n denotes the density of

  $$\frac{\theta - \theta_n}{s_n}, \qquad \theta \sim Q_n^\star,$$

  a BvM-type result establishes convergence, typically in L^1 or total variation, $q_n \to q$, for some limiting density q.

  For regular, well-specified parametric models, classical BvM establishes $s_n = n^{-1/2}$ and $q = \mathcal{N}\left(0, H_0^{-1}\right)$ is a Gaussian where H_0 is the Hessian of the population loss.

What do we mean by theoretical understanding? II

- Connection with frequentist confidence intervals
  Fix a nominal level $1-\beta \in (0,1)$. An asymptotically credible sequence of regions is a sequence of data-dependent sets C_n whose posterior probability satisfies

$$Q_n^\star(C_n) \longrightarrow 1-\beta$$

  in probability under the sampling distribution. The frequentist coverage of C_n, however, is the probability that it contains the fixed population target over repeated realizations of the data: $\mathbb{P}(\theta_0 \in C_n)$.
  We say credible regions are asymptotically calibrated at level $1-\beta$ if

$$\mathbb{P}(\theta_0 \in C_n) \longrightarrow 1-\beta.$$

  In this case, they also constitute asymptotically exact frequentist confidence regions.

α–divergence GVI

We study the GVI problem with α–divergence regularizer:

$$Q_n^\star \in \underset{Q \in \mathcal{P}(\mathbb{R}^d)}{\arg\min} \left\{ \int_{\mathbb{R}^d} \ell_n(\theta)\, dQ(\theta) + \frac{\eta}{n} D_\alpha(Q \| \Pi) \right\},$$

where for $\alpha > 0$, $\alpha \neq 1$,

$$D_\alpha(Q\|\Pi) = \int f_\alpha\left(\frac{dQ}{d\Pi}\right) d\Pi, \qquad f_\alpha(t) = \frac{t^\alpha - \alpha t + \alpha - 1}{\alpha(\alpha - 1)}.$$

α–divergence GVI

We study the GVI problem with α–divergence regularizer:

$$Q_n^\star \in \operatorname*{arg\,min}_{Q \in \mathcal{P}(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \ell_n(\theta)\, dQ(\theta) + \frac{\eta}{n} D_\alpha(Q \| \Pi) \right\},$$

where for $\alpha > 0$, $\alpha \neq 1$,

$$D_\alpha(Q\|\Pi) = \int f_\alpha\left(\frac{dQ}{d\Pi}\right) d\Pi, \qquad f_\alpha(t) = \frac{t^\alpha - \alpha t + \alpha - 1}{\alpha(\alpha - 1)}.$$

- As $\alpha \to 1$, $f_\alpha(t) \to t \log t - t + 1$, giving D_{KL}.
- At $\alpha = \frac{1}{2}$, $f_{1/2}(t) = 2(\sqrt{t} - 1)^2$, the squared Hellinger distance.

Exact solutions to the GVI problem

Closed form solutions exist for the GVI problem with f–divergence regularizer.

Alquier (2021), Daunas, Esnaola, Perlaza and Poor (2024, 2026), Nguyen and Westerhout (2026)

Write $\Delta_n(\theta) = \ell_n(\theta) - \ell_n(\theta_n)$, $\kappa = (1-\alpha)^{-1}$ and $p = (\alpha - 1)^{-1}$. Under mild conditions,

$$\frac{dQ_n^\star}{d\Pi}(\theta) = \begin{cases} \left[\delta_n + \dfrac{(1-\alpha)n}{\eta}\Delta_n(\theta)\right]^{-\kappa}, & 0 < \alpha < 1, \\ \left[\delta_n - \dfrac{(\alpha-1)n}{\eta}\Delta_n(\theta)\right]_+^{p}, & \alpha > 1. \end{cases}$$

Here $\delta_n > 0$ enforces unit mass.

BvM-type theorem for α–divergences

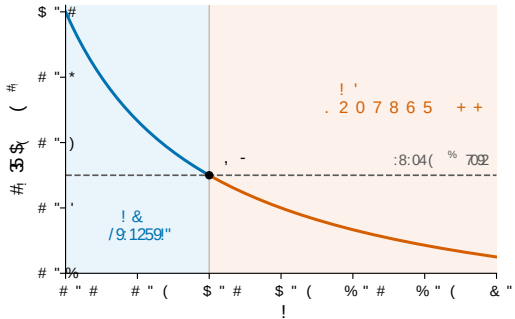
Let q_n be the density of $(\theta - \theta_n)/s_n$ under $Q_n^\star$, where $s_n := \sqrt{\frac{\eta\delta_n}{n}}$. Under standard regularity assumptions,

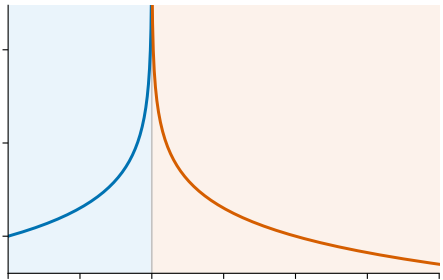
$$\|q_n - q_\alpha\|_{L^1(\mathbb{R}^d)} \longrightarrow 0, \qquad s_n \asymp n^{-b_\alpha}, \qquad b_\alpha = \frac{1}{2 + d(\alpha - 1)}.$$

Regime	Limiting density, up to normalization	Width exponent
$0 < \alpha < 1$	$q_\alpha \propto [1 + \frac{1-\alpha}{2} x^\top H_0 x]^{-\kappa}$	$b_\alpha > 1/2$
$\alpha > 1$	$q_\alpha \propto [1 - \frac{\alpha-1}{2} x^\top H_0 x]_+^p$	$b_\alpha < 1/2$

Proof technique: similar to Theorem 4 in Miller (2021), but controlling the scaling of δ_n.

Two qualitatively distinct regimes





Concentration exponent: $s_n \asymp n^{-b_\alpha}$, for $d = 1$.

Kernel exponent of the limiting Student-t and Pearson type-II laws.

$\alpha = 1$ separates two asymptotic regimes, in both limiting shape and concentration exponent.

Transition through $\alpha = 1$

Posterior credibility and sampling coverage separate

Under standard regularity conditions for M-estimators, we have $\sqrt{n}(\theta_n - \theta_0) \Rightarrow \mathcal{N}(0, \Sigma)$, for some $\Sigma \succ 0$.

See Van der Vaart (2000) Ch. 5

Posterior credibility and sampling coverage separate

Under standard regularity conditions for M-estimators, we have $\sqrt{n}(\theta_n - \theta_0) \Rightarrow \mathcal{N}(0, \Sigma)$, for some $\Sigma \succ 0$.

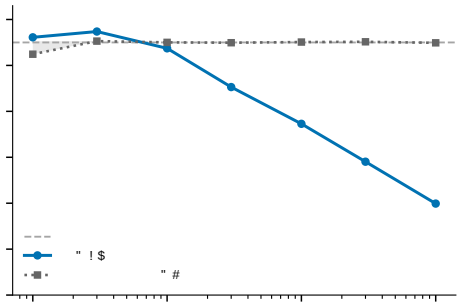
See Van der Vaart (2000) Ch. 5

Take a bounded measurable A containing a neighbourhood of 0, with $\int_A q_\alpha = 1 - \beta$, and set $C_n = \theta_n + s_n A$. Then

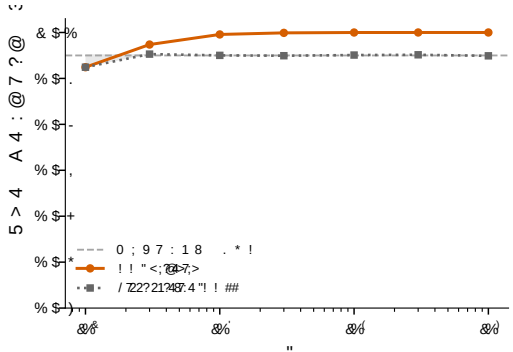
$$Q_n^\star(C_n) \longrightarrow 1 - \beta.$$

Regime	Relative width	Frequentist coverage
$0 < \alpha < 1$	$\sqrt{n}\, s_n \to 0$	$\mathbb{P}(\theta_0 \in C_n) \to 0$
$\alpha > 1$	$\sqrt{n}\, s_n \to \infty$	$\mathbb{P}(\theta_0 \in C_n) \to 1$

Frequentist coverage of credible intervals



$\alpha = 0.5$: undercoverage.



$\alpha = 2$: overcoverage.

Equal-tailed 95% posterior credible intervals.

Conclusions

- We obtain BvM-style theorems for α–divergence-based generalized posteriors.
- Non-Gaussian local limits and non $\sqrt{n}$ rescaling.
- The D_{KL} divergence $(\alpha \to 1)$ acts as a phase transition.
- Can have undesired consequences for the calibration of credible regions.

Conclusions

- We obtain BvM-style theorems for α–divergence-based generalized posteriors.
- Non-Gaussian local limits and non $\sqrt{n}$ rescaling.
- The D_{KL} divergence $(\alpha \to 1)$ acts as a phase transition.
- Can have undesired consequences for the calibration of credible regions.

A common philosophy when studying generalizations:

> *"Studying only the model theory of first-order logic would be analogous to the study of real analysis never knowing of any but the polynomial functions [...] It is only in the study of more general functions that one understands the importance of notions like continuity, differentiability or analyticity and their relations."*

Barwise (1985)

Scope and open questions

The results describe unrestricted α–divergence GVI posteriors in fixed dimension, with fixed $\eta > 0$ and standard but strong assumptions on ℓ_n.

Open questions:

- Calibration: Can $\eta = \eta_n$ correct the scale mismatch and calibrate credible regions?
- What changes for other f–divergence families or restricted variational families?

Thanks for your time!

QR for the preprint:



Personal webpage: icasado-telletxea.github.io
Email: ioar.casadotelletxea@manchester.ac.uk

Backup: precise regularity conditions

Local quadraticity. For all sufficiently large n and $\|h\| < \varepsilon_0$,

$$\Delta_n(\theta_n + h) = \tfrac{1}{2} h^\top H_n h + r_n(h), \qquad |r_n(h)| \le C \|h\|^3,$$

with $\theta_n \to \theta_0$ and symmetric $H_n \to H_0 \succ 0$.

Separation. For every $\varepsilon > 0$,

$$\liminf_{n\to\infty} \inf_{\|\theta - \theta_n\| \ge \varepsilon} \Delta_n(\theta) > 0.$$

Prior regularity. π is continuous at θ_0 and $\pi(\theta_0) > 0$.

Apply these conditions on a common probability-one event for random losses. For $0 < \alpha < 1$, also require $2/(1-\alpha) > d$.

Compare Miller (2021), Theorem 4 and Section 8.8.

Backup: parameters of the limiting distributions

Student case $\nu = 2\kappa - d$,

$$q_\alpha = t_\nu(0, S_\alpha), \qquad S_\alpha = \frac{2\kappa}{2\kappa - d} H_0^{-1}.$$

S_α is the scale matrix. If $\nu > 2$, the covariance is $\frac{2\kappa}{2\kappa - d - 2} H_0^{-1}$.

Pearson type–II case

$$\operatorname{supp}(q_\alpha) = \{x : x^\top H_0 x \le 2p\}, \qquad \operatorname{Cov}(q_\alpha) = \frac{2p}{2p + d + 2} H_0^{-1}.$$

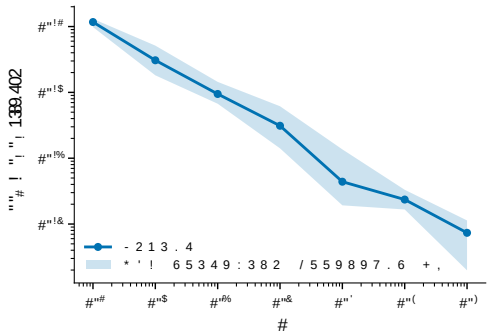
A regular example: intercept-only logistic regression

$$X_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p_0), \qquad p_0 = 0.7, \qquad \ell_n(\theta) = \log(1 + e^{\theta}) - \widehat{p}_n\theta.$$
$$\theta_n = \text{logit}(\widehat{p}_n), \quad \theta_0 = \text{logit}(p_0), \quad H_n = \widehat{p}_n(1 - \widehat{p}_n), \quad H_0 = 0.21.$$

- Prior: Student–t_5, location 0, scale 2. Set $\eta = 1$.
- $\alpha = 1/2$: Student–t_3 limit, $s_n \asymp n^{-2/3}$.
- $\alpha = 2$: Pearson type–II limit, $s_n \asymp n^{-1/3}$.

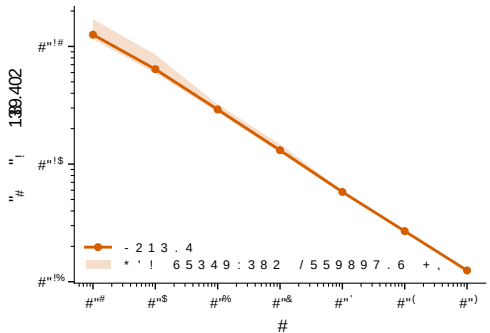
We compute each posterior from the empirical loss by solving its normalization equation numerically for δ_n.

Backup: coupled paths for $\alpha = 1/2$



Twelve independent paths, with prefix counts coupled across n.

Backup: coupled paths for $\alpha = 2$



Backup: selected references

This work
I. Casado-Telletxea and O. Rivasplata. *On the asymptotic behaviour of α–divergence generalized posteriors.* Manuscript, 2026.

Generalized variational inference
J. Knoblauch, J. Jewson and T. Damoulas (2022). *An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference.* JMLR 23(132), 1–109. Paper

Generalized-posterior asymptotics
J. W. Miller (2021). *Asymptotic Normality, Concentration, and Coverage of Generalized Posteriors.* JMLR 22(168), 1–53. Paper