Statistical guarantees for variational Bayes

Anirban Bhattacharya
Department of Statistics, Texas A&M University

Based on joint work with Debdeep Pati and Yun Yang

Acknowledgement: NSF CAREER, NSF DMS, and NSF CDSE-MSS

May 21, 2021

Workshop on High-dimensional Covariance Matrices, Networks, and Concentration Inequalities

Introduction and background

Bayesian inference

- Likelihood function $p(Y^n \mid \theta)$
- $\pi(\cdot)$ a prior distribution on parameter space Θ
- Posterior distribution

$$\pi\left(\theta \mid Y^n\right) = \frac{p(Y^n \mid \theta)\,\pi(\theta)}{\int_\Theta p(Y^n \mid \eta)\,\pi(d\eta)}$$

- All inference is based on the posterior distribution (point estimation, credible intervals, ...)
- Point estimate $\widehat{\theta}_B = \int \theta\,\pi(d\theta \mid Y^n)$; the posterior mean

Nonparametric Bayes

- Relax parametric assumptions on model/prior
- Infinite-dimensional parameter or parameter-dimension growing with sample size
- Highly flexible generative models for datasets with complex dependencies
- Bayes' formula automatically produces a point estimate (plus uncertainty quantification)

Example: Bayes deconvolution

- Consider $W_i = X_i + U_i$ with $U_i \overset{ind.}{\sim} \mathcal{N}(0, \sigma_i^2)$ for $i \in [n]$
- The unobserved $X_i \overset{ind.}{\sim} f$
- GWAS study: W_i the estimated marginal effect size of the ith SNP from a regression of the response on the ith standardized SNP, and X_i true effect size of ith SNP
- Goal: learn the true effect size distribution f (interest on exceedance probabilities)
- Shape constraints: f is *smooth, symmetric about zero, and unimodal*. Also, sharply peaked around zero and slowly decaying tails.

Motivating Example: Bayes deconvolution

- Flexible prior on f? Single Gaussian or t inadequate
- Want prior to have *large support* on the space of symmetric unimodal densities
- Exploit a representation theorem for such densities

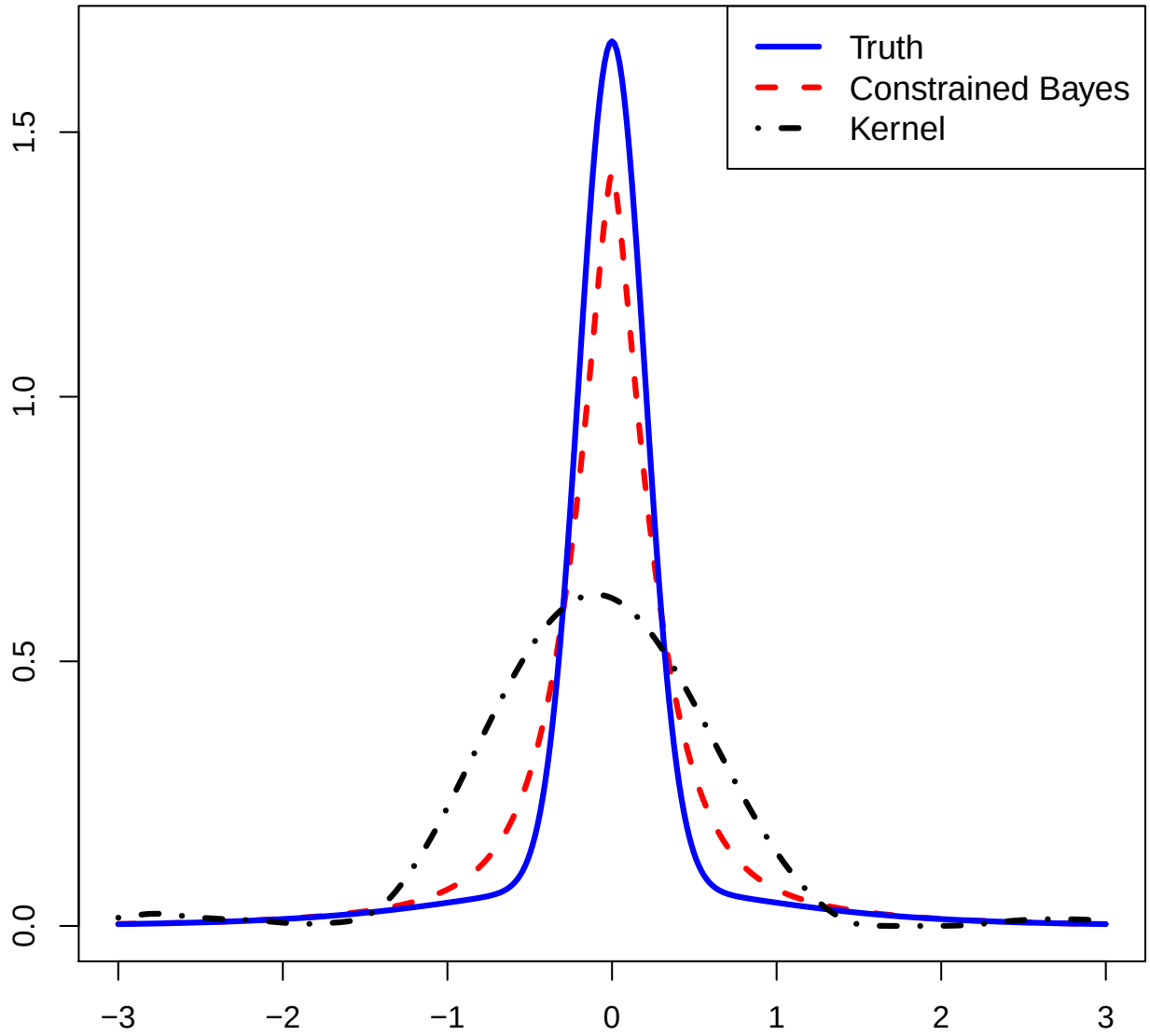
Motivating Example: Bayes deconvolution

- Want prior to have *large support* on the space of symmetric unimodal densities
- Representation theorem (Feller, 71): if f is symmetric, unimodal, and smooth (f' is finite a.e.), then there exists a density g on $[0, \infty)$ such that

$$f(x) = \int (2\theta)^{-1} \mathbb{1}_{(-\theta \leq x \leq \theta)} \, g(\theta) \, d\theta$$

- Hierarchical representation: $X \,|\, \theta \sim U(-\theta, \theta), \ \theta \sim g$
- Model g flexibly as a (Dirichlet process) mixture of gamma densities (Su, B., Zhang, Chatterjee, Carroll)

Motivating Example: Bayes deconvolution



Solid blue is the truth (a mixture of a t-density with 5 df and a normal density with sd 0.2), the dashed red is our constrained Bayesian method and the dash-dotted black is the deconvoluting kernel density estimator. $n = 5000$.

Hierarchical/Latent variable models

- Probit regression: $y_i \mid x_i, \beta \overset{ind.}{\sim} \text{Bernoulli}(\Phi(x_i'\beta))$ for $i \in [n]$
- LV representation: $y_i = \mathbb{1}(z_i > 0)$ and $z_i \overset{ind.}{\sim} N(x_i'\beta, 1)$ for $i \in [n]$
- Specify a prior distribution $\beta \sim \pi(\cdot)$
- Mixture model: $y_i \mid \theta \overset{iid}{\sim} \sum_{h=1}^{K} \pi_h N(\mu_h, \Sigma_h)$ where $\theta = \{(\pi_h, \mu_h, \Sigma_h)\}_{h=1}^{K}$
- LV representation: $y_i \mid z_i, \theta \overset{ind.}{\sim} N(\mu_{z_i}, \Sigma_{z_i})$ for $i \in [n]$, and $pr(z_i = h \mid \theta) = \pi_h$ for $h \in [K]$
- Specify a prior distribution on θ

General Framework

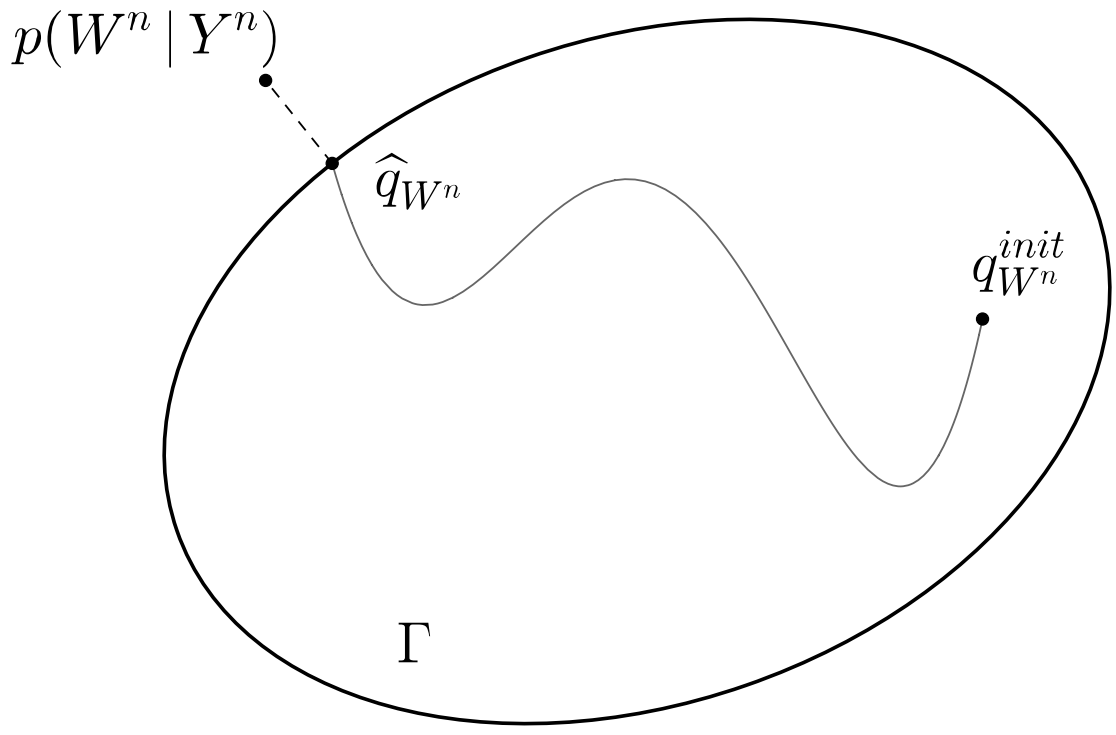
- Observations $Y^n = (y_1, \ldots, y_n)$
- Hidden variables $W^n = (\theta, Z^n)$
  - θ collects all parameters in the model
  - $Z^n = (z_1, \ldots, z_n)$ collects all latent variables
- Statistical model:
  - Observed-data likelihood function: $p(Y^n \mid Z^n, \theta)$
  - Latent variable distribution: $p(Z^n \mid \theta)$
  - Prior distribution on parameters: $\pi(\theta)$
- Conduct inference via the joint posterior distribution

$$p\left[\theta, Z^n \mid Y^n\right] = \frac{p(Y^n \mid Z^n, \theta)\, p(Z^n \mid \theta)\, \pi(\theta)}{\int_{\Theta \times \mathcal{Z}^n} p(Y^n \mid Z^n, \eta)\, p(Z^n \mid \eta)\, \pi(d\eta)}$$

How does one compute posterior quantities?

- Exact Bayesian inference: Markov Chain Monte Carlo (MCMC) sampling
- Approximate Bayesian inference: Laplace approximation and variants (INLA), Variational Bayes, Expectation propagation – approximate posterior functionals without full MCMC which can be expensive

Variational inference



- Let Γ denote a pre-specified family of distributions on Θ
- Idea: approximate the posterior $p(W^n \mid Y^n)$ by a closest member of this family in Kullback-Leibler (KL) (or another) divergence

$$\widehat{q}_{W^n} := \underset{q_{W^n} \in \Gamma}{\operatorname{argmin}} \; D_{\mathrm{KL}}\left[q_{W^n}(\cdot) \,\middle|\middle|\, p(\cdot \mid Y^n)\right]$$

Another perspective: ELBO decomposition

$$\log p(Y^n) = \underbrace{\int_{\mathcal{W}^n} q_{W^n}(w^n)\, \log \frac{p(Y^n \mid w^n)\, p_{W^n}(w^n)}{q_{W^n}(w^n)}\, dw^n}_{L(q_{W^n})} + \underbrace{\int_{\mathcal{W}^n} q_{W^n}(w^n)\, \log \frac{q_{W^n}(w^n)}{p(w^n \mid Y^n)}\, dw^n}_{D_{\mathrm{KL}}[\, q_{W^n}(\cdot)\, ||\, p(\cdot \mid Y^n)\,]} \geq L(q_{W^n})$$

- $L(q_{W^n})$ is called the Evidence Lower BOund (ELBO), since it provides a lower bound to the log evidence $\log p(Y^n)$
- KL minimization $\equiv$ ELBO maximization

Mean-field VI

- Mean-field variational family: consider all joint distributions over $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ that factorize as

$$q_\Theta(\theta) = \prod_{j=1}^{d} q_{\Theta_j}(\theta_j), \, \forall \, \theta = (\theta_1, \ldots, \theta_d)'$$

  - No constraint on forms of individual variational factors
  - Can be computed via the coordinate ascent variational inference (CAVI) algorithm in conditionally conjugate models [Bishop, 2006]
  - Similar to EM updates

Mean-field VI: Example

- Example 1: suppose $x_1, \ldots, x_n \mid \mu, \tau \overset{ind.}{\sim} N(\mu, \tau^{-1})$, and independent priors $\mu \sim N(0, \kappa^{-1})$ and $\tau \sim \text{Gamma}(a_0, b_0)$
- Mean-field decomposition $q(\mu, \tau) = q_{\boldsymbol{\mu}}(\mu)\, q_{\boldsymbol{\tau}}(\tau)$
- Iterative updates of (parallel) CAVI algorithm

$$q_{\boldsymbol{\mu}}^{(t)}(\mu) = N(\mu; m^{(t)}, 1/s^{(t)}), \quad q_{\boldsymbol{\tau}}^{(t)}(\tau) = \text{Gamma}(\tau; a^{(t)}, b^{(t)}),$$

with

$$s^{(t+1)} = nE_{q_{\boldsymbol{\tau}}^{(t)}}(\tau) + \kappa, \quad m^{(t+1)} = \frac{nE_{q_{\boldsymbol{\tau}}^{(t)}}(\tau)\,\bar{x}}{nE_{q_{\boldsymbol{\tau}}^{(t)}}(\tau) + \kappa},$$

and

$$a^{(t+1)} = \frac{n}{2} + a_0, \quad b^{(t+1)} = \left[\frac{1}{2}E_{q_{\boldsymbol{\mu}}^{(t)}}\|x - \mu 1_n\|_2^2 + b_0\right].$$

Mean-field VI: Example

- Example 2: suppose $x_i \overset{ind.}{\sim} N(\mu_{z_i}, \tau_{z_i}^{-1})$ for $i \in [n]$ with $\text{pr}(z_i = h) = p_h$ for $h \in [K]$
- Assume

$$q(\mu_{1:K}, \tau_{1:K}, p_{1:K}, z_{1:n}) = q_1(\mu_{1:K}, \tau_{1:K})\, q_2(p_{1:K})\, q_z(z_{1:n})$$

- CAVI Updates can be again derived in closed form

CAVI algorithm

- In general, with $q = q_1 \otimes \ldots \otimes q_d$, the CAVI algorithm cycles through

$$q_j \propto \exp\left(\int_{\Theta_{-j}} q_{-j} \log \pi_n\right), \quad j \in [d]$$

where $q_{-j} = \otimes_{k \neq j} q_k$ and π_n the posterior.

- Obtained by minimizing

$$D_{\mathrm{KL}}(q_j \otimes q_{-j} \,||\, \pi_n)$$

over q_j, keeping q_{-j} fixed.

Another example

- Suppose the target posterior is of the form

$$p(u_1, u_2) \propto e^{-n\, u_1^{k_1} u_2^{k_2}}\, u_1^{h_1} u_2^{h_2}, \quad u = (u_1, u_2) \in [0,1]^2,$$

  where k_1, k_2 and h_1, h_2 are non-negative integers.

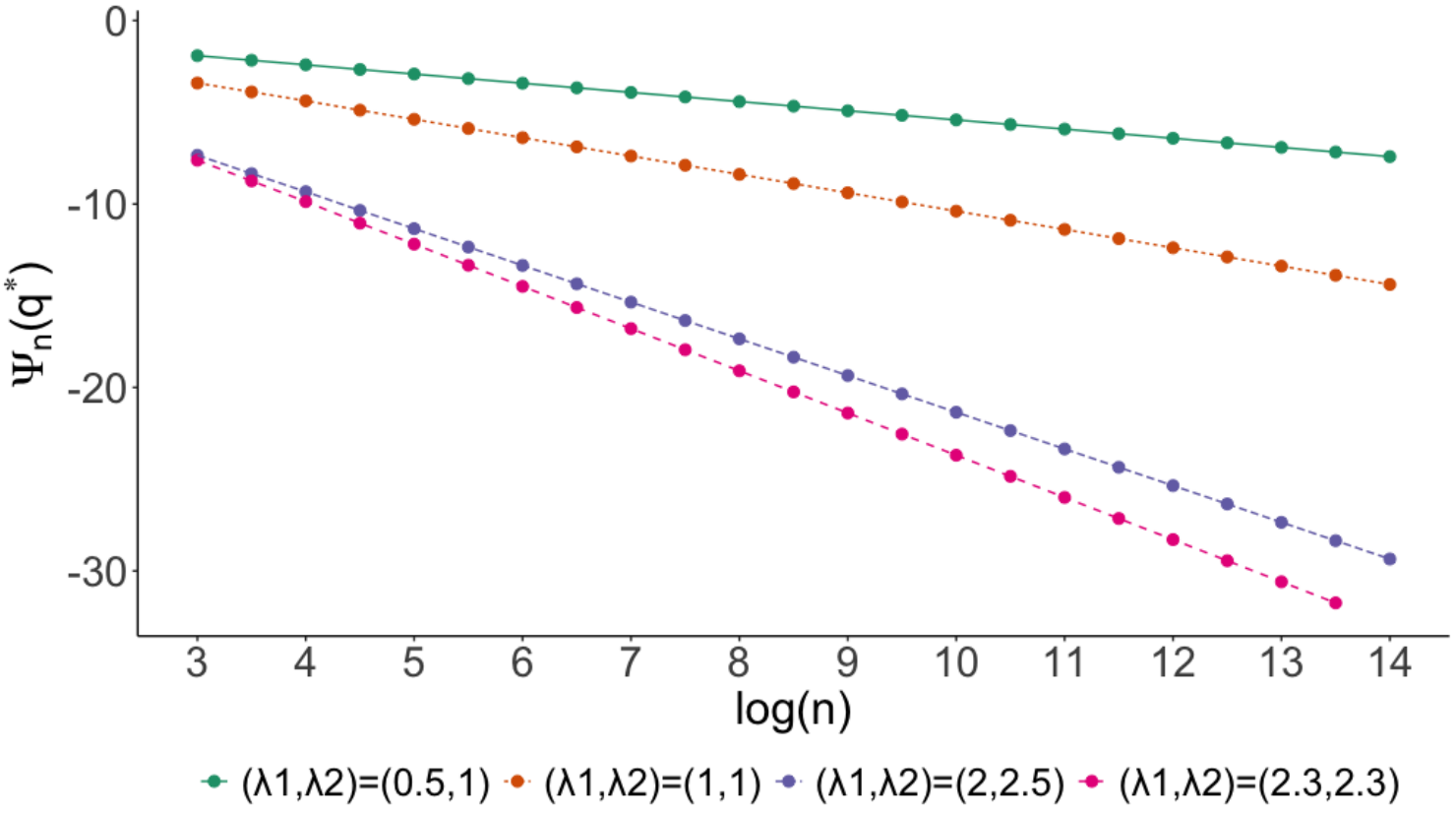
- The behavior of the normalizing constant as a function of n is well-understood: singular learning theory (Sumio Watanabe)

- For example,

$$\int_{[0,1]^2} e^{-nu_1^2 u_2^2} du \asymp \frac{C \log n}{\sqrt{n}}, \quad \int_{[0,1]^2} e^{-nu_1^2 u_2^4} du \asymp \frac{C}{n^{1/4}}.$$

- Consider a mean-field decomposition $q(u) = q_1(u_1) q_2(u_2)$.

Empirical demonstration



Other commonly used variational families

- Parametric family such as the exponential family

$$q_\Theta(\theta;\, \kappa) = h(\theta)\, \exp\left\{\langle \eta(\kappa),\, T(\theta)\rangle - A(\kappa)\right\}$$

  - Gaussian variational family
  - Copula VI [Tran et al 2015]
  - For non-conjugate models: tangent transformation [Jakkola and Jordan, 1999]
- More expressive variational distribution
  - Blackbox VI [Ranganath et al 2014]
  - Implicit VI [Huszár, 2017]
  - Variational Auto-Encoders [Kingma and Welling 2013]
  - Mixture of Gaussians [Zobay, 2014], implemented using variational boosting [Guo et al. 2016, Locatello et al. 2017, Miller et al. 2019, Campbell and Li, 2019]

More resources

- Chapter 10 of Bishop's Pattern Recognition & Machine Learning
- Variational Inference: A Review for Statisticians by Blei et al. (JASA)
- Advances in Variational Inference by Zhang et al. (IEEE Transactions on Pattern Analysis and Machine Intelligence)

Theory for VB: what to expect

Mean-field (and many other) variational families trade off between computational tractability and approximation capability. Can not expect full posterior approximation in general. What about point estimates?

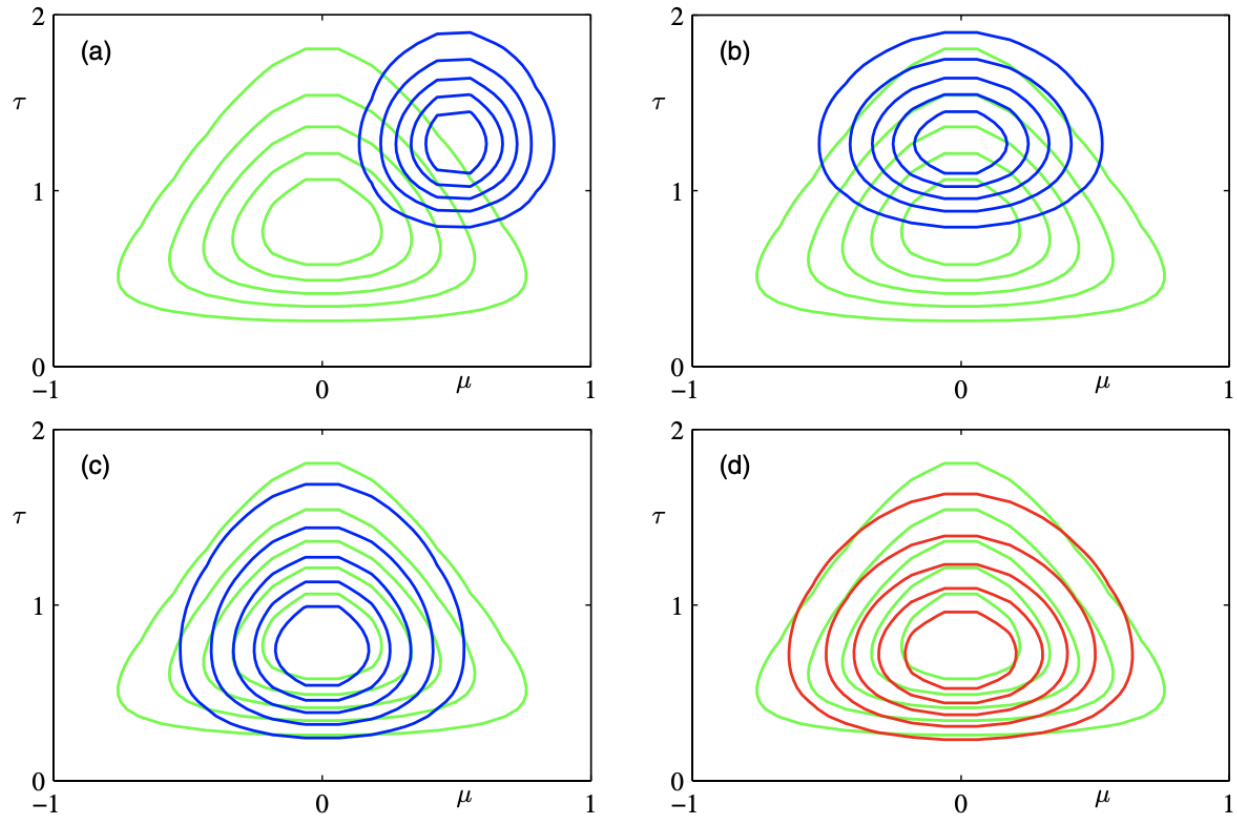


Figure 10.4 Illustration of variational inference for the mean μ and precision τ of a univariate Gaussian distribution. Contours of the true posterior distribution $p(\mu, \tau|D)$ are shown in green. (a) Contours of the initial factorized approximation $q_\mu(\mu)q_\tau(\tau)$ are shown in blue. (b) After re-estimating the factor $q_\mu(\mu)$. (c) After re-estimating the factor $q_\tau(\tau)$. (d) Contours of the optimal factorized approximation, to which the iterative scheme converges, are shown in red.

Picture credit: Bishop, PMLR

Questions of interest

- Statistical Accuracy: Is the center of $\widehat{q}_{\Theta}$ a good *proxy* for the posterior mean of the parameters? Can we use $\widehat{q}_{\Theta}$ to estimate the normalizing constant of the posterior, i.e., the marginal likelihood?

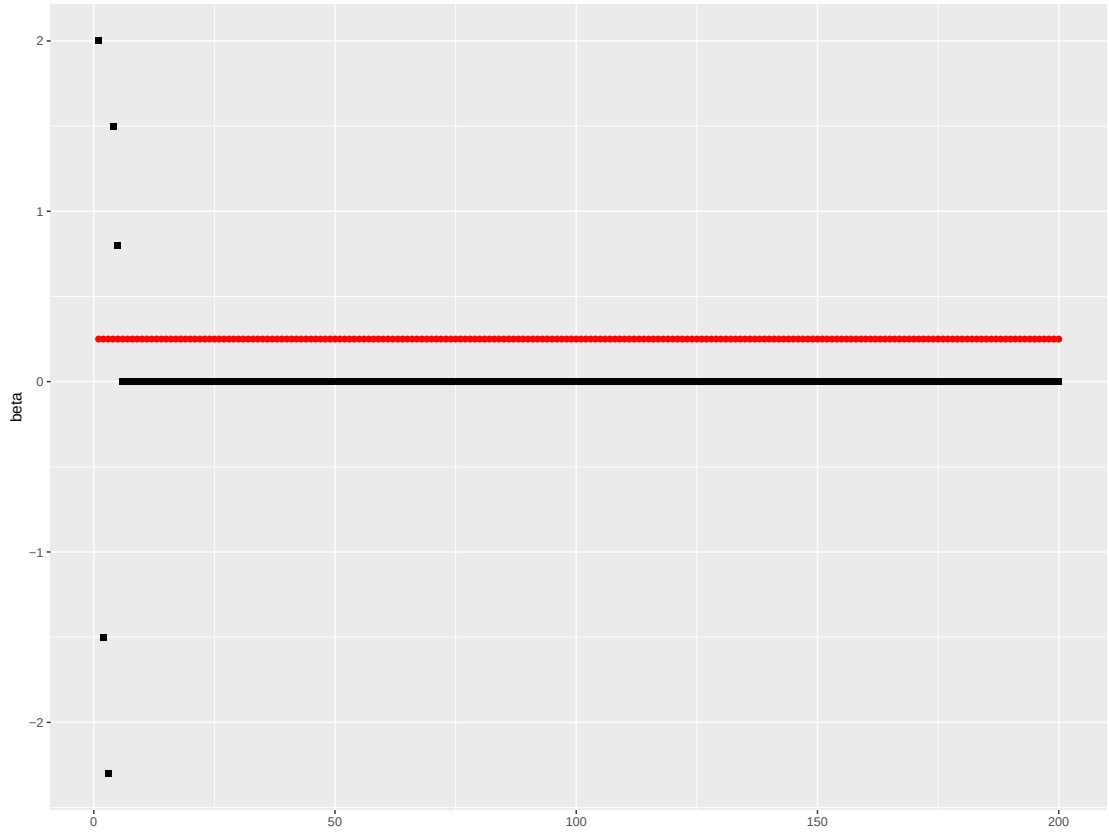
  Pati, B. and Yang, 2017; Yang, Pati, B. 2020; Wang & Blei, 2019a, 2019b; Zhang and Gao, 2020, Alquier and Ridgeway, 2020

- Computational guarantee: Does $\widehat{q}_{\Theta}^{init}$ converge to $\widehat{q}_{\Theta}$? Non-convex optimization. Case-by-case analysis available sporadically. Towards a general theory?

Fitted regression coefficients ($n = 100$, $d = 200$)

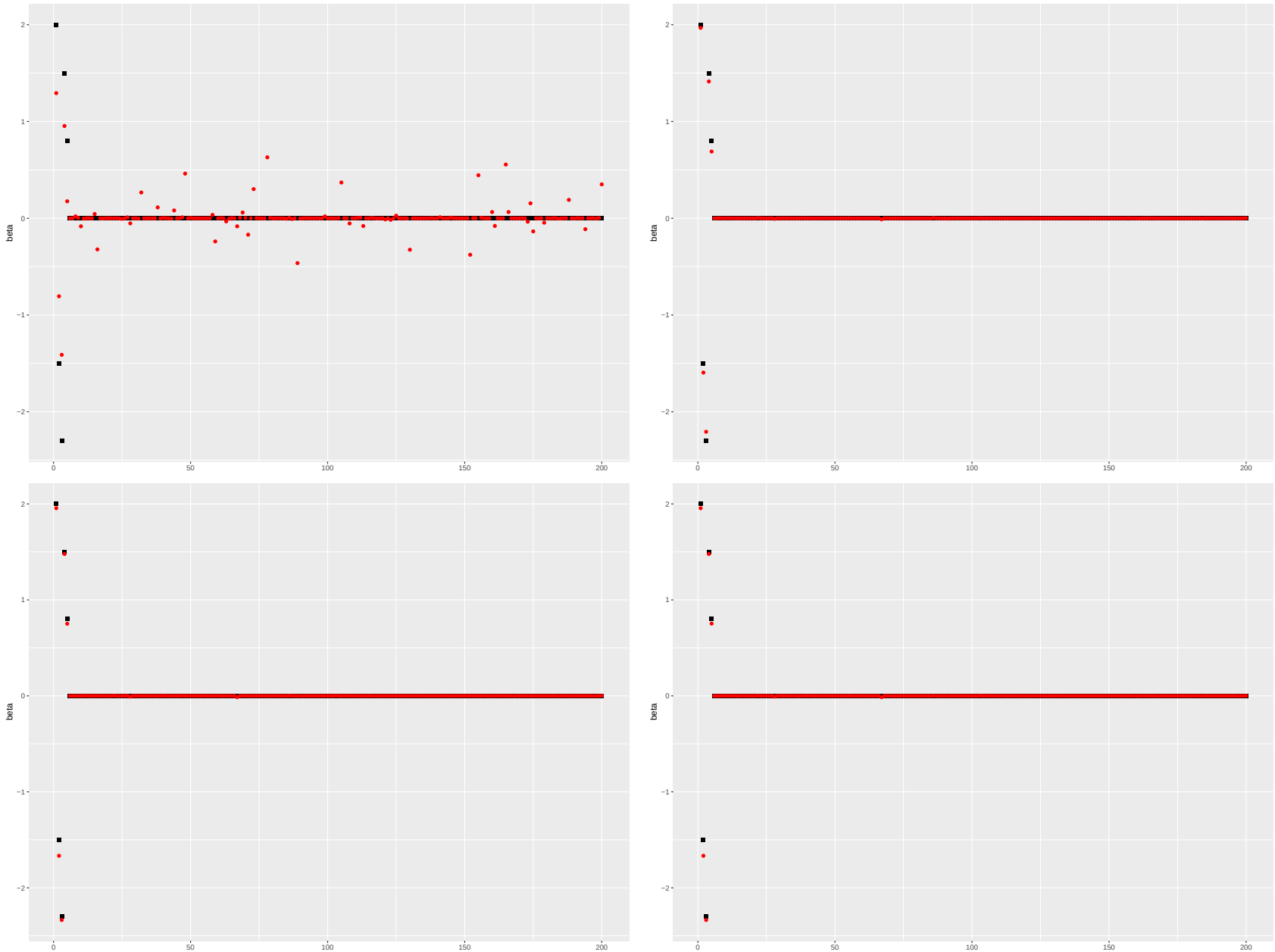
$$Y = X\beta^* + w, \quad w \sim \mathsf{N}(0, \sigma^2)$$

Variational estimate: $\hat{\beta}$ in Red and β^* in Black.



Initialization.

Fitted regression coefficients ($n = 100$, $d = 200$)



Iteration 2-5

Risk bounds for tempered posteriors

Posterior consistency

- Study of first order frequentist properties. Assume a frequentist setup, i.e., there is a true parameter/distribution generating the data
- Posterior consistency: the posterior assigns increasing (to one) probabilities to arbitrary fixed neighborhoods of the true parameter as sample size increases, almost surely w.r.t. data sequences generated by the true distribution
- Hellinger/L_1 distance most common choice of neighborhood
- Early results: Doob (1948), Schwartz (1978)
- Barron, Schervish, and Wasserman (1999); Ghosal, Ghosh, and Ramamoorthi (1999)

Posterior contraction

- Posterior contraction/concentration: allow the neighborhood size to shrink to zero. The fastest possible rate ε_n is called the posterior contraction rate
- Ghosal, Ghosh, and van der Vaart (2000); Shen and Wasserman (2001); Kleijn & van der Vaart (2006)
- Typically implies a central measure of the posterior, e.g., the posterior mean, converges to the true parameter at rate ε_n
- Benchmarked against the minimax optimal rate for the given problem
- Highly successful in delivering adaptive near-minimax (up to logarithmic terms) results for a broad variety of problems: review article by Rousseau et al. (2014)

Standard sufficient conditions

- The sufficient conditions for posterior contraction in correctly specified models are typically variants of

  (i) Prior mass condition: the prior has to assign sufficient mass to appropriate Kullback–Leibler neighborhoods of the truth

  (ii) Existence of sieves: one has to find a sequence of increasing subsets of the parameter space (sieves) with

  (a) appropriate control on their sizes (in terms of metric entropy)

  (b) exponentially small prior probability of their complements

Ghoshal, Ghosh, van der Vaart (2000)

THEOREM 2.1. *Suppose that for a sequence* ε_n *with* $\varepsilon_n \to 0$ *and* $n\varepsilon_n^2 \to \infty$, *a constant* $C > 0$ *and sets* $\mathscr{P}_n \subset \mathscr{P}$, *we have*

$$\log D(\varepsilon_n, \mathscr{P}_n, d) \leq n\varepsilon_n^2, \tag{2.2}$$

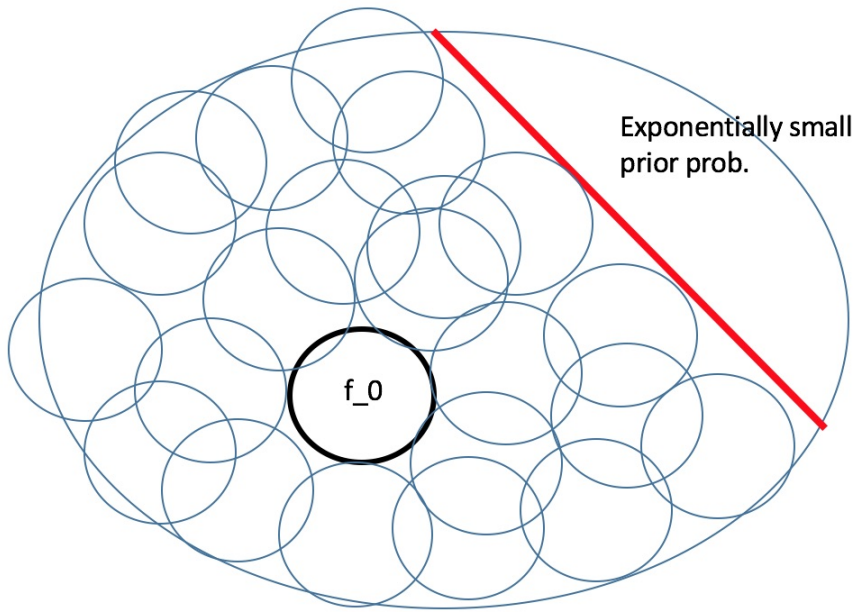
$$\Pi_n(\mathscr{P} \setminus \mathscr{P}_n) \leq \exp\bigl(-n\varepsilon_n^2(C+4)\bigr), \tag{2.3}$$

$$\Pi_n\Bigl(P\colon -P_0\Bigl(\log\frac{p}{p_0}\Bigr) \leq \varepsilon_n^2,\ P_0\Bigl(\log\frac{p}{p_0}\Bigr)^2 \leq \varepsilon_n^2\Bigr) \geq \exp\bigl(-n\varepsilon_n^2 C\bigr). \tag{2.4}$$

Then for sufficiently large M, *we have that* $\Pi_n(P\colon d(P, P_0) \geq M\varepsilon_n | X_1, \ldots, X_n) \to 0$ *in* P_0^n*-probability.*

Hypothesis testing and sieves

- The key idea is the construction of exponentially consistent test functions for the true density versus the complement of an ε_n-neighborhood of the truth (inside the sieve)
- Need control on the metric entropy of the sieve to bound the type-I error
- For Hellinger or L_1 distances, Birge-LeCam testing theory applies



Fractional posterior

- Raise the likelihood to a fractional power α and combine with the prior
- Fractional/tempered posterior

$$\pi_{n,\alpha}(\theta) := \frac{\{p(Y^n \mid \theta)\}^{\alpha}\, \pi(\theta)}{\int_{\Theta}\{p(Y^n \mid \theta)\}^{\alpha}\, \pi(d\theta)}$$

- Proceed as if regular posterior; e.g.,

$$\widehat{\theta}_{B,\alpha} = \int \theta\, \pi_{n,\alpha}(d\theta)$$

Walker & Hjort, 2002; Chernozhukhov & Hong, 2003; Jiang & Tanner, 2006; Zhang, 2006; Leung & Barron, 2006; Grunwald (2012, 2014); Miller & Dunson (2015); Atchadé, 2016; Bissiri et al., 2017, B, Pati, Yang, 2019.

Fractional posterior in the literature

- Gelman & Meng 1998, Neal 1998: posterior sampling
- Leung and Barron (2006); Dalalyan and Tsybakov (2008): aggregation with exponential weighting
- Grunwald (2012, 2014), Miller & Dunson (2015): robustness to misspecification
- Chernozhukhov & Hong, 2003; Jiang & Tanner, 2006; Atchadé, 2016; Martin & Walker, 2016: pseudo-posteriors/quasi-posteriors/Gibbs posteriors $\rightarrow$ replace negative log-likelihood by some other empirical risk measure

Remarks

- We shall work under the tempered posterior setup for this talk
- Substantially simplifies the analysis – only need the prior mass condition
- Treatment of correctly specified and misspecified models under a unified framework
- Variational risk bounds extend to $\alpha = 1$ with additional work
- Plan forward: we first derive risk bounds for a tempered posterior (no approximation) and then extend to its variational approximations

Remarks

- Let θ^* denote the (pseudo-)true parameter
- We shall derive Bayes risk bounds of the following flavor: with high probability under the data-generating distribution $\mathbb{P}_{\theta^*}$,

$$\int d(\theta, \theta^*) \pi_{n,\alpha}(d\theta) \leq C\varepsilon_n^2$$

  where d is a distance/divergence measure on the parameter space, and ε_n^2 typically corresponds to the minimax rate (possibly upto a logarithmic term) for the statistical problem.
- In particular, if d is convex in its first argument, then we immediately get

$$d\big(\widehat{\theta}_{B,\alpha}, \theta^*\big) \leq C\varepsilon_n^2 \text{ w.h.p.}$$

Variational characterization

- Fix any probability measure $q \ll \pi$ and consider

$$
\begin{aligned}
& D_{\mathrm{KL}}(q \,||\, \pi_{n,\alpha}) \\
&= -\int q(d\theta) \log \frac{\pi_{n,\alpha}(\theta)}{q(\theta)} \\
&= -\int q(d\theta) \log \frac{e^{\alpha \ell_n(\theta,\theta^*)}\, \pi(\theta)}{q(\theta)} + \log m_\alpha(Y^n) \\
&= \underbrace{-\alpha \int \ell_n(\theta,\theta^*)\, q(d\theta) + D_{\mathrm{KL}}(q \,||\, \pi)}_{\text{terms involving } q} + \log m_\alpha(Y^n).
\end{aligned}
$$

- Here θ^* is the (pseudo-)true parameter and

$$
\ell_n(\theta,\theta^*) := \ell_n(\theta) - \ell_n(\theta^*), \quad \ell_n(\theta) = \log p(Y^n \mid \theta).
$$

Variational Characterization

- Define

$$\Psi(q) = \underbrace{-\int \ell_n(\theta, \theta^*)\, q(d\theta)}_{\text{model fit}} + \underbrace{\alpha^{-1} D_{\mathrm{KL}}(q \,||\, \pi)}_{\text{penalty}}$$

- $\Psi(\cdot)$ is minimized at $\pi_{n,\alpha}$ over all $q \ll \pi$
- Variational/penalization characterization of (fractional) posterior
- α inverse temperature parameter
- For $\alpha = 1$, $\Psi(q)$ is simply the evidence lower bound in variational inference.

Variational Lemma

The above is equivalent to the well-known variational representation of the relative entropy.

Variational lemma: let μ be a probability measure and h a measurable function such that $e^h \in L_1(\mu)$. Then,

$$\log \int e^h d\mu = \sup_{\rho \ll \mu} \left[\int h d\rho - D_{\mathrm{KL}}(\rho \,||\, \mu) \right]$$

Risk bound for tempered posterior

- Theorem [B., Pati, Yang 2019]: With $\mathbb{P}_{\theta^*}$ probability at least $(1-\zeta)$,

$$(1-\alpha)\int_\Theta D_\alpha^{(n)}(\theta,\theta^*)\,\pi_{n,\alpha}(d\theta)$$
$$\leq \alpha\Bigg[\underbrace{-\int_\Theta \ell_n(\theta,\theta^*)\,\pi_{n,\alpha}(d\theta)+\alpha^{-1}D_{\mathrm{KL}}(\pi_{n,\alpha}\,||\,\pi)}_{\Psi(\pi_{n,\alpha})}\Bigg]+\log(1/\zeta).$$

- Here $D_\alpha^{(n)}(\theta,\theta^*)$ is the Rényi divergence of order $\alpha\in(0,1)$ between $p(Y^n\mid\theta)$ and $p(Y^n\mid\theta^*)$

More on the divergence measure

- Consider

$$\mathbb{E}_{\theta^*} e^{\alpha\, \ell_n(\theta,\theta^*)} = \mathbb{E}_{\theta^*} \left\{ \frac{p(Y^n \mid \theta)}{p(Y^n \mid \theta^*)} \right\}^{\alpha} = \underbrace{\int \left\{ \frac{p(y^n \mid \theta)}{p(y^n \mid \theta^*)} \right\}^{\alpha} p(dy^n \mid \theta^*)}_{A_\alpha^{(n)}(\theta,\theta^*)}$$

- $A_\alpha^{(n)}(\theta, \theta^*) \in (0, 1)$: Rényi affinity of order α
- If $\alpha = 1/2$: Hellinger affinity
- Rényi divergence of order $\alpha \in (0, 1)$

$$D_\alpha^{(n)}(\theta, \theta^*) = \frac{1}{\alpha - 1} \log A_\alpha^{(n)}(\theta, \theta^*) > 0$$

Example

- Suppose $p(\cdot \mid \theta) \equiv \mathcal{N}_n(\theta, \mathrm{I}_n)$ with $\theta \in \mathbb{R}^n$
- A simple calculation yields

$$D_\alpha^{(n)}(\theta, \theta^*) = \frac{\alpha \|\theta - \theta^*\|^2}{2(1-\alpha)}$$

- Equivalent to the squared Euclidean distance

Risk bound for tempered posterior

- Recall that

$$\Psi(q) = -\int_{\Theta} \ell_n(\theta, \theta^*)\, q(d\theta) + \alpha^{-1} D_{\mathrm{KL}}(q \,||\, \pi)$$

  is minimized at $\pi_{n,\alpha}$ among all $q \ll \pi$, and thus $\Psi(\pi_{n,\alpha}) \leq \Psi(q)$ for any $q \ll \pi$

- Thus, with $\mathbb{P}_{\theta^*}$ probability at least $(1-\zeta)$,

$$\underbrace{\int_{\Theta} D_{\alpha}^{(n)}(\theta, \theta^*)\, \pi_{n,\alpha}(d\theta)}_{\text{Bayes risk}} \leq \frac{\alpha}{1-\alpha} \inf_{q \ll \pi} \left[\Psi(q)\right] + \frac{\log(1/\zeta)}{1-\alpha}$$

- Choose a candidate q to strike a balance

Choice of q

- q places more mass near θ^* implies smaller $\left[-\int_\Theta \ell_n(\theta, \theta^*)\, q(d\theta) \right]$
- Also need control on the KL distance from π for the second term $\alpha^{-1} D_{\mathrm{KL}}(q \,||\, \pi)$
- To obtain the correct trade-off, set

$$q^{\mathrm{cand}}(\theta) = \frac{\pi(\theta)\mathbb{I}_{\mathcal{N}(\theta^*;\epsilon)}(\theta)}{\mathcal{Z}}, \quad \mathcal{Z} = \int_\Theta \pi(\theta)\mathbb{I}_{\mathcal{N}(\theta^*;\epsilon)}(\theta) d\theta$$

where $\mathcal{N}(\theta^*; \epsilon)$ is an appropriate neighborhood of θ^*

Choice of q

- Recall

$$q^{\text{cand}}(\theta) = \frac{\pi(\theta)\mathbb{I}_{\mathcal{N}(\theta^*;\epsilon)}(\theta)}{\mathcal{Z}}.$$

- Importantly, this implies,

$$D_{\text{KL}}(q^{\text{cand}} \,||\, \pi) = -\log\left[\pi(\mathcal{N}(\theta^*;\epsilon))\right]$$

- Need to choose the neighborhood so that the log-likelihood function can be controlled with high probability.
- With $V(g \,||\, h) = \int g \log^2(g/h) d\mu$, define

$$\mathcal{N}(\theta^*;\varepsilon) = \Big\{\theta \in \Theta \,:\, D_{\text{KL}}[p(\cdot \,|\, \theta^*) \,||\, p(\cdot \,|\, \theta)] \leq n\varepsilon^2;$$

$$V[p(\cdot \,|\, \theta^*) \,||\, p(\cdot \,|\, \theta)] \leq n\varepsilon^2 \Big\}$$

An explicit risk bound

Theorem

Suppose $\varepsilon \in (0,1)$ *satisfies* $n\,\varepsilon^2 > 2$ *and* $C > 1$. *With* $\mathbb{P}_{\theta^*}$ *probability at least* $1 - 2/\{(C-1)^2 n\varepsilon^2\}$,

$$\int \left\{ \frac{1}{n} D_\alpha^{(n)}(\theta, \theta^*) \right\} \pi_{n,\alpha}(d\theta)$$
$$\leq \frac{C\,\alpha}{1-\alpha}\,\varepsilon^2 + \left\{ -\frac{1}{n(1-\alpha)} \log \pi\big[B_n(\theta^*;\varepsilon)\big] \right\}.$$

In particular, if ε_n *satisfies the fixed point equation*

$$-\frac{\log \pi\big[B_n(\theta^*;\varepsilon)\big]}{n\,\varepsilon} = \varepsilon,$$

then the RHS becomes $\frac{C\,\alpha+1}{1-\alpha}\,\varepsilon_n^2$

Risk bound for tempered posterior: proof sketch

Recall the main result:

Theorem

With $\mathbb{P}_{\theta^*}$ probability at least $(1-\zeta)$,

$$(1-\alpha)\int_\Theta D_\alpha^{(n)}(\theta,\theta^*)\,\pi_{n,\alpha}(d\theta)$$
$$\leq \alpha\Bigg[\underbrace{-\int_\Theta \ell_n(\theta,\theta^*)\,\pi_{n,\alpha}(d\theta)+\alpha^{-1}D_{\mathrm{KL}}(\pi_{n,\alpha}\,||\,\pi)}_{\Psi(\pi_{n,\alpha})}\Bigg]+\log(1/\zeta).$$

Risk bound for tempered posterior: proof sketch

- We have, $\mathbb{E}_{\theta^*} e^{\alpha\, \ell_n(\theta,\theta^*)} = A_\alpha^{(n)}(\theta,\theta^*) = e^{-(1-\alpha)D_\alpha^{(n)}(\theta,\theta^*)}$
- For any $\zeta \in (0,1)$,

$$\mathbb{E}_{\theta^*}\ \exp\left[\alpha\, \ell_n(\theta,\theta^*) + (1-\alpha)D_\alpha^{(n)}(\theta,\theta^*) - \log(1/\zeta)\right] \leq \zeta$$

- Integrate w.r.t. prior + Fubini

$$\mathbb{E}_{\theta^*} \int_\Theta \exp\left[\alpha\, \ell_n(\theta,\theta^*) + (1-\alpha)D_\alpha^{(n)}(\theta,\theta^*) - \log(1/\zeta)\right] \pi(\mathsf{d}\theta) \leq \zeta$$

Recall Variational Lemma

Variational lemma: let μ be a probability measure and h a measurable function such that $e^h \in L_1(\mu)$. Then,

$$\log \int e^h d\mu = \sup_{\rho \ll \mu} \left[\int h d\rho - D_{\mathrm{KL}}(\rho \,||\, \mu) \right]$$

Back to proof sketch

- Variational lemma:

$$\mathbb{E}_{\theta^*} \exp \sup_{q \ll \pi} \left[\int_\Theta \left\{ \alpha \, \ell_n(\theta, \theta^*) + (1-\alpha) D_\alpha^{(n)}(\theta, \theta^*) - \log(1/\zeta) \right\} q(d\theta) - D_{\mathrm{KL}}(q \,||\, \pi) \right] \leq \zeta.$$

- Substitute $q = \pi_{n,\alpha}$,

$$\mathbb{E}_{\theta^*} \exp \left[\int_\Theta \left\{ \alpha \, \ell_n(\theta, \theta^*) + (1-\alpha) D_\alpha^{(n)}(\theta, \theta^*) - \log(1/\zeta) \right\} \pi_{n,\alpha}(d\theta) - D_{\mathrm{KL}}(\pi_{n,\alpha} \,||\, \pi) \right] \leq \zeta$$

Sketch contiued

- $\mathbb{E}\exp(Z) \leq \zeta$ implies $Z \leq 0$ with probability at least $(1-\zeta)$ by Markov
- With $\mathbb{P}_{\theta^*}$ probability at least $(1-\zeta)$,

$$
\begin{aligned}
&(1-\alpha)\int_\Theta D_\alpha^{(n)}(\theta,\theta^*)\,\pi_{n,\alpha}(d\theta) \\
&\leq -\alpha\int_\Theta \ell_n(\theta,\theta^*)\,\pi_{n,\alpha}(d\theta) + D_{\mathrm{KL}}(\pi_{n,\alpha}\,||\,\pi) + \log(1/\zeta) \\
&= \alpha\Bigg[\underbrace{-\int_\Theta \ell_n(\theta,\theta^*)\,\pi_{n,\alpha}(d\theta) + \alpha^{-1}D_{\mathrm{KL}}(\pi_{n,\alpha}\,||\,\pi)}_{\Psi(\pi_{n,\alpha})}\Bigg] + \log(1/\zeta).
\end{aligned}
$$

Remarks

- The risk bound above delivers near-optimal rate of contraction for tempered posteriors across variety of settings [B., Pati, Yang, 2019]
- Benchmarked against rates of posterior contraction in well-specified and mis-specified models

  Ghosal, Ghosh, and van der Vaart (2000); Shen and Wasserman (2001), Kleijn & van der Vaart (2006)
- Prior mass condition *alone* is sufficient to guarantee optimal concentration
- Additionally, provides insights into delivering similar results for variational approximations.

Risk bounds for variational approximations to tempered posteriors

Setup

- Our risk bounds for the tempered posterior use its variational characterization – natural to investigate similar ideas for variational approximations
- Derive Bayes risk bounds for the variational approximation

$$\int d(\theta, \theta^*) \widehat{q}_{\Theta,\alpha}(d\theta)$$

- Even when variational family Γ corresponds to a mean-field decomposition, we can obtain near-optimal performance in a fairly general class of models
- We separately deal the cases with and without latent variables

α-variational Bayes (no latent variable)

For a pre-specified variational family Γ, the α-variational approximation is given by

$$
\begin{aligned}
\widehat{q}_\Theta &:= \underset{q_\Theta \in \Gamma}{\operatorname{argmin}} \; KL\big[q_\Theta \,\big|\big|\, p_\alpha(\cdot \,|\, Y^n)\big] \\
&= \underset{q_\Theta \in \Gamma}{\operatorname{argmin}} \left\{ - \int_\Theta q_\Theta(\theta) \, \log \frac{p^\alpha(Y^n \,|\, \theta)\, \pi_\Theta(\theta)}{q_\Theta(\theta)} \, d\theta \right\} \\
&= \underset{q_\Theta \in \Gamma}{\operatorname{argmin}} \left\{ - \alpha \int_\Theta q_\Theta(\theta) \, \log p(Y^n \,|\, \theta) \, d\theta + D_{\mathrm{KL}}(q_\Theta \,||\, \pi_\Theta) \right\} \\
&= \underset{q_\Theta \in \Gamma}{\operatorname{argmin}} \left\{ - \alpha \int_\Theta q_\Theta(\theta) \, \ell_n(\theta, \theta^*) \, d\theta + D_{\mathrm{KL}}(q_\Theta \,||\, \pi_\Theta) \right\},
\end{aligned}
$$

Variational risk bound (no latent variable)

Theorem (Variational risk bound)

For any $\zeta \in (0,1)$ *and* $\alpha \in (0,1)$*, it holds with probability at least* $(1-\zeta)$ *that for any probability measure* $q_\Theta \in \Gamma$ *with* $q_\Theta \ll p_\Theta$*,*

$$\int_\Theta h^2\left[p(\cdot \mid \theta),\, p(\cdot \mid \theta^*)\right] \widehat{q}_\Theta(\theta)\, d\theta \quad \textit{(Variational Bayes risk)}$$

$$\leq \frac{\alpha}{n\,(1-\alpha)}\, \Psi(q_\Theta) + \frac{1}{n\,(1-\alpha)} \log(1/\zeta)$$

α-variational Bayes (no latent variable)

The objective function is equivalent to (up to division by α)

$$\Psi(q_\Theta) := \underbrace{-\int_\Theta q_\Theta(\theta)\,\ell_n(\theta,\theta^*)\,d\theta}_{\text{model fit}} + \underbrace{\alpha^{-1} D_{\mathrm{KL}}(q_\Theta \,||\, \pi_\Theta)}_{\text{regularization}}$$

- Model-fit term gets smaller as q_Θ places more mass near the true parameter θ^*
- With $q \equiv q^{\mathrm{opt}}$ with high prob.

$$\Psi(q_\Theta) = -\int_{\mathcal{N}(\theta^*;\epsilon)} \pi_\Theta(\theta)\,\ell_n(\theta,\theta^*)\,d\theta - \alpha^{-1}\log \Pi\{\mathcal{N}(\theta^*;\epsilon)\}.$$

- For parametric models, $\epsilon = \sqrt{d\log n/n}$ and

$$\Psi(q_\Theta) \le n\epsilon^2 + \underbrace{\frac{d}{2}\log(n/\epsilon)}_{\text{BIC penalty}}$$

.

Variational risk bound (no latent variable)

- Links the variational Bayes risk to the objective value $\Psi(q_\Theta)$ for any $q_\Theta \in \Gamma$ (e.g. data-dependent)
- Minimizing $\Psi(q_\Theta)$ within the variational family has the same effect as minimizing the variational Bayes risk
- Choose good $q_\Theta \in \Gamma$ to control $\Psi(q_\Theta)$

$$q^{\text{cand}}(\theta) = \frac{\pi(\theta)\mathbb{I}_{\mathcal{N}(\theta^*;\epsilon)}(\theta)}{\int_\Theta \pi(\theta)\mathbb{I}_{\mathcal{N}(\theta^*;\epsilon)}(\theta)d\theta}$$

Theorem

If $-\log \Pi\{\mathcal{N}(\theta^*;\delta)\} \leq n\epsilon^2(\delta)$ *and* Γ *is rich enough to contain* q^{cand}*, then* $\Psi(q_\Theta) \leq n\delta^2 + n\epsilon^2(\delta)$.

- Solving the F.P.E $\delta = \epsilon(\delta)$ gives the risk bound.

α-variational Bayes (with latent variable)

- Observation latent variable pair (Y_i, Z_i) are mutually independent across $i = 1, 2, \ldots, n$ [Wang and Titterington 2005]
- For simplicity, assume $Z_i \in \{1, 2, \ldots, K\}$ to be discrete
- $p(Z_i = z_i \mid \theta) = \pi_{z_i}$, $\pi_{z^n} = \prod_{i=1}^n \pi_{z_i}$
- Focus on mean-field approximation:

$$q_{W^n}(\theta, z^n) = q_\Theta(\theta) \prod_{i=1}^n q_{Z_i}(z_i), \quad \forall \theta \quad \text{and} \quad \forall z^n = (z_1, \ldots, z_n)^T$$

Usual VB objective function with latent variables

$$D_{\mathrm{KL}}(q_{W^n} \,||\, p(\cdot \mid Y^n)) = -\int \sum_{z^n} q_{Z^n}(z^n)\, q_\Theta(\theta)\, \log \frac{p(Y^n \mid \theta, z^n)\, \pi_{z^n}\, \pi_\theta(\theta)}{q_{Z^n}(z^n)\, q_\Theta}\, d\theta$$

$$= -\int \ell_n(\theta, \theta^*)\, q_\Theta(\theta)\, d\theta + \Delta_J + D_{\mathrm{KL}}(q_\Theta \,||\, \pi_\Theta) + \log p(Y^n)$$

The Jensen gap Δ_J can be written as

$$\Delta_J(q_\Theta, q_{Z^n}) = \int_\Theta q_\Theta(\theta) \Bigg\{ \underbrace{\sum_{i=1}^n \log p(Y_i \mid \theta)}_{\text{Log-marginal likelihood}}$$

$$\underbrace{- \sum_{i=1}^n \sum_{z_i} q_{Z_i}(z_i)\, \log \frac{p(Y_i \mid \theta, z_i)\, p(z_i \mid \theta)}{q_{Z_i}(z_i)}}_{\text{Its variational approximation}} \Bigg\}\, d\theta \geq 0$$

α-variational Bayes (with latent variable)

Define α-variational approximation to the fractional posterior

$$(\widehat{q}_\Theta, \widehat{q}_{Z^n}) = \underset{q_\Theta,\, q_{Z^n} = \prod_{i=1}^n q_{Z_i}}{\operatorname{argmin}} \Psi(q_\Theta, q_{Z^n}),$$

where $\Psi(q_\Theta, q_{Z^n}) :=$

$$\underbrace{-\int_\Theta q_\Theta(\theta)\, \ell_n(\theta, \theta^*)\, d\theta + \Delta_J(q_\Theta, q_{Z^n})}_{\text{model fit}} + \underbrace{\alpha^{-1} D_{\mathrm{KL}}(q_\Theta, \pi_\Theta)}_{\text{regularization}}.$$

When $\alpha = 1$, coincides with usual VB with latent variables.

Variational oracle inequality (with latent variable)

$$\Psi(q_\Theta, q_{Z^n}) := \underbrace{-\int_\Theta q_\Theta(\theta)\, \ell_n(\theta, \theta^*)\, d\theta + \Delta_J(q_\Theta, q_{Z^n})}_{\text{model fit}} + \underbrace{\alpha^{-1} D_{\mathrm{KL}}(q_\Theta\, \pi_\Theta)}_{\text{regularization}}.$$

Theorem (Variational oracle inequality)

For any $\zeta \in (0,1)$ and $\alpha \in (0,1)$, it holds with probability at least $(1-\zeta)$ that for any probability measure $q_\Theta \ll p_\theta$,

$$\int_\Theta D_\alpha\big[p(\cdot\,|\,\theta),\, p(\cdot\,|\,\theta^*)\big]\, \widehat{q}_\Theta(\theta)\, d\theta \quad \textit{(Variational Bayes risk)}$$

$$\leq \frac{\alpha}{n\,(1-\alpha)}\, \Psi(q_\Theta, q_{Z^n}) + \frac{1}{n\,(1-\alpha)} \log(1/\zeta)$$

Variational risk bound (with latent variable)

- Minimizing $\Psi(q_\Theta, q_{Z^n})$ within the variational family has the same effect as minimizing the variational Bayes risk
- As before, need good choices of q_Θ and q_{Z^n} to control $\Psi(q_\Theta, q_{Z^n})$
- Good choice of q_{Z^n} makes the Jensen gap Δ_J small

$$q_{Z_i} = \frac{p(Y_i \mid \theta^*, z_i)p(z_i \mid \theta^*)}{\sum_i p(Y_i \mid \theta^*, z_i)p(z_i \mid \theta^*)}$$

- Make Γ large enough so that it includes

$$q^{\text{opt}}(\theta) = \frac{\pi(\theta)\mathbb{I}_{\mathcal{N}(\theta^*;\epsilon)}(\theta)}{\int_\Theta \pi(\theta)\mathbb{I}_{\mathcal{N}(\theta^*;\epsilon)}(\theta)d\theta}$$

Example: Low-dimensional linear regression

- Low-dimensional linear model ($d \ll n$),

$$Y = X\beta + w, \qquad w \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

- Parameter of interest $\theta = (\beta, \sigma^2)$
- Variational family Γ:

$$q(\beta,\, \sigma^2) = q_\beta(\beta)\, q_{\sigma^2}(\sigma^2)$$

- Mean field approximation:

$$(\widehat{q}_\beta,\, \widehat{q}_{\sigma^2}) := \underset{q_\beta,\, q_{\sigma^2}}{\operatorname{argmin}} \left\{ KL(q_\beta \,||\, \pi_\beta) + KL(q_{\sigma^2} \,||\, \pi_{\sigma^2}) \right.$$
$$\left. - \alpha \iint q_\beta(\beta)\, q_{\sigma^2}(\sigma^2)\, \log p(Y^n \,|\, X,\, \beta,\, \sigma^2)\, d\beta\, d\sigma^2 \right\}$$

- "Conjugate" priors: Gaussian on β and inverse Gamma on σ^2

Low-dimensional linear regression

Assumption: The joint prior density of (β, σ^2) is continuous, and positive at the truth $\theta^* = (\beta^*, (\sigma^*)^2)$.

Corollary (Mean field approximation to low-dimensional linear regression)

If $d/n \to 0$ as $n \to \infty$, then it holds with probability tending to one as $n \to \infty$ that

$$\left\{ \int h^2 \left[p(\cdot \,|\, \theta) \,\middle|\middle|\, p(\cdot \,|\, \theta^*) \right] \widehat{q}_\theta(\theta)\, d\theta \right\}^{1/2} \lesssim \sqrt{\frac{d}{n \, \min\{\alpha, \, 1-\alpha\}} \, \log(d\, n)}.$$

High-dimensional sparse linear regression

- High-dimensional linear model ($d \gg n$),

$$Y = X\beta + w, \qquad w \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

- Point mass mixture prior on β:

$$z_j \overset{iid}{\sim} \left(1 - \frac{1}{d}\right)\delta_0 + \frac{1}{d}\,\delta_1, \quad \left[\beta_j \,|\, z_j\right] \sim \mathcal{N}(0,\, z_j\, \sigma_\beta^2)$$

- $z = (z_1, \ldots, z_d)^T$ and β together form the parameter $\theta = (z,\, \beta)$ of interest
- Block mean field family:

$$q(z,\, \beta) = \prod_{j=1}^{d} q_{z_j,\beta_j}(z_j,\, \beta_j)$$

High-dimensional sparse linear regression

- Block mean field approximation:

$$(\widehat{q}_{z_j,\beta_j},\, j \in [d]) := \underset{q_{z_j,\beta_j}, j\in[d]}{\operatorname{argmin}} \left\{ KL\Big[\bigotimes_{j=1}^{d} q_{z_j,\beta_j} \,\Big|\Big|\, \pi_{z,\,\beta} \Big] \right.$$
$$\left. - \alpha \sum_{z\in\{0,1\}^d} \int \prod_{j=1}^{d} q_{z_j,\beta_j}(z_j,\, \beta_j)\, \log p(Y^n \,|\, X,\, \beta,\, z)\, d\beta \right\}$$

- Each component $\widehat{q}_{z_j,\beta_j}$ is also point mass mixture ("conjugate" prior), i.e. admits form

$$(1-\phi_j)\;\; \delta_0 \;\otimes \underbrace{\delta_0}_{z_j=0} \;+\; \phi_j\;\; \delta_1 \;\otimes \underbrace{\mathcal{N}(\mu_j,\, \sigma_j^2)}_{z_j=1}$$

High-dimensional sparse linear regression

Assumption: $\pi_{\beta \mid z^*}$ is continuous and positive at β^*, and the truth β^* is s-sparse.

Corollary (Block mean field approximation to high-dimensional linear regression)

If $s \log d/n \to 0$ *as* $n \to \infty$*, then it holds with probability tending to one as* $n \to \infty$ *that*

$$\left\{ \int h^2 \left[p(\cdot \mid \beta) \,\middle|\middle|\, p(\cdot \mid \beta^*) \right] \widehat{q}_\beta(\beta)\, d\beta \right\}^{1/2} \lesssim \sqrt{\frac{s}{n \min\{\alpha,\, 1-\alpha\}} \log(d\, n)}.$$

Variational approximation in Gaussian mixture model

- Recall model

$$Y_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \mathcal{N}_d(\mu_k, I_d), \quad i = 1, 2, \ldots, n$$

- Using data augmentation, we can rewrite the model as

$$Z_i \overset{iid}{\sim} \text{Categorical}(\pi_1, \pi_2, \ldots, \pi_K), \quad \left[Y_i \,|\, Z_i = s\right] \sim \mathcal{N}_d(\mu_s, I_d),$$

Z_i : the latent class indicator variable for Y_i

- Mean field approximation

$$q(\pi, \mu, Z^n) = q_\pi(\pi)\, q_\mu(\mu)\, q_{Z^n}(z^n) = q_\pi(\pi)\, q_\mu(\mu) \prod_{i=1}^{n} q_{Z_i}(z_i)$$

Variational approximation in Gaussian mixture model

Assumption: There exists some constant $\delta_0 > 0$, such that each component of $\pi^* \in \mathcal{S}_K$ is at least δ_0; the prior densities p_μ and p_π are positive and continuous at μ^* and π^* respectively.

Corollary (Mean field approximation in Gaussian mixture model)

If $d\,K/n \to 0$ as $n \to \infty$, then it holds with probability tending to one as $n \to \infty$ that

$$\left\{ \int h^2 \left[p(\cdot \,|\, \theta) \,\middle|\middle|\, p(\cdot \,|\, \theta^*) \right] \widehat{q}_\theta(\theta)\, d\theta \right\}^{1/2} \lesssim \sqrt{\frac{d\,K}{n\, \min\{\alpha,\, 1-\alpha\}}\, \log(d\,n)}.$$

Summary and highlights

- Temperature parameter aids the theoretical analysis
- Main finding: point estimates obtained from α-VB have the same convergence rate as the true posterior mean in commonly used models without/with latent variables (Gaussian mixture models, Bayesian variable selection, Latent Dirichlet allocation, tangent transforms. . .) - minimal assumptions on prior
- Proof extends to $\alpha = 1$ with extra prior tail and entropy conditions
- Theory also extends to more expressive families

Ongoing work: convergence guarantees?

- Optimization in general mean-field is a non-convex optimization problem
- Existing analyses case-by-case
- Can we leverage concentration properties of the target to provide general-purpose sufficient conditions for (geometric) convergence of the CAVI iterates?
- An important quantity that shows up in our analysis is

$$A = \int (q_1^{(t+1)} - q_1^\star)\,(q_2^{(t)} - q_2^\star)\,\log \pi_n$$

  for a two-block mean-field decomposition $q = q_1 \otimes q_2$.

- Need appropriate control on A.

References

- Anirban Bhattacharya, Debdeep Pati, and Yun Yang. “Bayesian fractional posteriors.” The Annals of Statistics 47:1 (2019): 39-66.
- Yun Yang, Debdeep Pati, and Anirban Bhattacharya. “α-variational inference with statistical guarantees." Annals of Statistics 48:2 (2020): 886-905.
- Debdeep Pati, Anirban Bhattacharya, and Yun Yang. “On statistical optimality of variational Bayes.” International Conference on Artificial Intelligence and Statistics. 2018.
- Antik Charaborty, Anirban Bhattacharya, Bani K. Mallick. “Bayesian sparse multiple regression for simultaneous rank reduction and variable selection." Biometrika 107(1): 205-221.
- Anirban Bhattacharya, Debdeep Pati, Sean Plummer, and Yun Yang. “Evidence bounds in singular models: probabilistic and variational perspectives.” arXiv preprint arXiv:2008.04537 (2020) to be updated soon.

Thank You!