Supplemental document On Statistical Optimality of Variational Bayes

Debdeep Pati
Texas A&M University

Anirban Bhattacharya
Texas A&M University

Yun Yang
Florida State University

1 Proofs of results in the main document

1.1 Convention

Equations in the main document are cited as (1), (20 etc., retaining their numbers, while new equations defined in this document are numbered (S1), (S2) etc.

1.2 Proof of Theorem 3.1

As in the proof sketch in the main document, our first step is to show that under the testing assumption **T**,

$$\int_\Theta \xi(\theta,\theta^*)\, p_\theta(d\theta) \le e^{Cn\varepsilon_n^\kappa}, \tag{1}$$

w.h.p. (w.r.t. $\mathbb{P}_{\theta^*}^{(n)}$), where recall $\log \xi(\theta,\theta^*) = \ell_n(\theta,\theta^*) + nd^\kappa(\theta,\theta^*)$. We first establish (1). Define

$$T_1 = \int_{d(\theta,\theta^*)\le\varepsilon_n} \xi(\theta,\theta^*)\, p_\theta(d\theta),$$
$$T_2 = \int_{d(\theta,\theta^*)>\varepsilon_n} \xi(\theta,\theta^*)\, p_\theta(d\theta).$$

Let us first tackle T_1. Since $\mathbb{E}_{\theta^*}[e^{\ell_n(\theta,\theta^*}] = 1$, we have,

$$\mathbb{E}_{\theta^*} T_1 = \int_{d(\theta,\theta^*)\le\varepsilon_n} e^{nd^\kappa(\theta,\theta^*)}\, p_\theta(d\theta) \le e^{n\varepsilon_n^\kappa}.$$

Hence, by Markov's inequality, $T_1 \le e^{Cn\varepsilon_n^\kappa}$ with probability at least $1 - e^{-Cn\varepsilon_n^\kappa}$.

Let us now focus on T_2. Write $T_2 = T_{21} + T_{22}$, where

$$T_{21} = \int_{d(\theta,\theta^*)>\varepsilon_n} (1-\phi_n)\, \xi(\theta,\theta^*)\, p_\theta(d\theta),$$
$$T_{22} = \int_{d(\theta,\theta^*)>\varepsilon_n} \phi_n\, \xi(\theta,\theta^*)\, p_\theta(d\theta),$$

where ϕ_n is the test function from Assumption **T**. Focus on T_{21} first. Observe

$$\mathbb{E}_{\theta^*} T_{21} = \int_{d(\theta,\theta^*)>\varepsilon_n} \mathbb{E}_\theta[1-\phi_n]\, e^{nd^\kappa(\theta,\theta^*)}\, p_\theta(d\theta) \le e^{-Cn\varepsilon_n^\kappa}.$$

This implies, by Markov's inequality, than $T_{21} \le e^{-Cn\varepsilon_n^\kappa}$ with probability at least $1 - e^{-Cn\varepsilon_n^\kappa}$.

Finally, focus on T_{22}. Since $\mathbb{E}_{\theta^*}[\phi_n] \le e^{-n\varepsilon_n^\kappa}$, it follows from Markov's inequality that $\phi_n \le e^{-Cn\varepsilon_n^\kappa}$ with probability at least $1 - e^{-Cn\varepsilon_n^\kappa}$. Hence, $T_{22} \le e^{-Cn\varepsilon_n^\kappa} T_2$ w.h.p. Adding the w.h.p. bound for T_{21}, we obtain, w.h.p.,

$$T_2 \le e^{-Cn\varepsilon_n^\kappa} T_2 + e^{-Cn\varepsilon_n^\kappa}.$$

Rearranging, $T_2 \le e^{-Cn\varepsilon^\kappa}$ with probability at least $1 - e^{-Cn\varepsilon_n^\kappa}$. Combining with the bound for T_1, (1) is established.

Once (1) is established, the next step is to link the integrand in (1) with the latent variables. To that end, observe that

$$\xi(\theta,\theta^*) = \sum_{s^n} \exp\{h(\theta,s^n)\}\, \widehat{q}_{S^n}(s^n),$$

where

$$h(\theta,s^n) = \log \frac{p(Y^n \mid \mu, s^n)\, \pi_{s^n}}{p(Y^n \mid \theta^*)\, \widehat{q}_{S^n}(s^n)} + nd^\kappa(\theta,\theta^*).$$

Combining the above with (10), we have, w.h.p.,

$$\int_\Theta \sum_{s^n} \exp\big\{h(\theta,s^n)\big\}\, \widehat{q}_{S^n}(s^n)\, p_\theta(d\theta) \le e^{Cn\varepsilon_n^\kappa}. \tag{2}$$

Next, use a well-known variational/dual representation of the KL divergence (see, e.g., Corollary 4.15 of [1]) which states that for any probability measure μ and any measurable function h with $e^h \in L_1(\mu)$,

$$\log \int e^{h(\eta)}\, \mu(d\eta) = \sup_\rho \Big[\int h(\eta)\, \rho(d\eta) - D(\rho \,||\, \mu) \Big], \tag{3}$$

where the supremum is over all probability measures $\rho \ll \mu$. In the present context, setting $\eta = (\theta, s^n)$, $\mu := \widehat{q}_{S^n} \otimes p_\theta$, and $\rho = \widehat{q}_\theta \otimes \widehat{q}_{S^n}$, it follows from the variational lemma (3) and some rearrangement of terms that w.h.p.

$$n \int_\Theta d^\kappa(\theta,\theta^*)\, \widehat{q}_\theta(d\theta) \le n\varepsilon_n^\kappa + D(\widehat{q}_\theta \,||\, p_\theta) - \int_\Theta \sum_{s^n} h(\theta,s^n)\, \widehat{q}_\theta(d\theta).$$

From (7)–(9) (in the main document), it follows that the right hand side of the above display equals

$n\varepsilon_n^{\kappa} + \Omega(\widehat{q}_\theta, \widehat{q}_{S^n})$. The proof of the theorem then follows, since by definition, $\Omega(\widehat{q}_\theta, \widehat{q}_{S^n}) \le \Omega(q_\theta, q_{S^n})$ for any (q_θ, q_{S^n}) in the variational family Γ.

1.3 Proof of Lemma 4.3

Since $W_1(P^*, P) < \varepsilon$, there exists a coupling q such that $\sum_{k,k'} q_{kk'} \|\mu_k^* - \mu_{k'}\| < \varepsilon$. Then $\sum_k \pi_k^* \inf_{k'} \|\mu_k^* - \mu_{k'}\| < \varepsilon$. Since $\pi_k^* \ge \delta$, we have $\inf_{k'} \|\mu_k^* - \mu_{k'}\| \le \varepsilon/\delta$ for all $k = 1, \ldots, K$. This means for any k, there exists a k' such that $\|\mu_k^* - \mu_{k'}\| < \varepsilon/\delta$. Without loss of generality, let $k' = k$. This proves the first part of the assertion. To prove the second part, observe that for $k \ne k'$, $\|\mu_k^* - \mu_{k'}\| \ge \zeta - \|\mu_{k'}^* - \mu_{k'}\| \ge \kappa - \varepsilon/\delta$. Then

$$
\begin{aligned}
\varepsilon > W_1(P^*, P) &\ge \inf_q \sum_{k \ne k'} q_{kk'} \|\mu_k^* - \mu_{k'}\| \\
&\ge (\zeta - \varepsilon/\delta) \inf_{C \in C_{XY}} \mathbb{P}(X \ne Y) \\
&= (\zeta - \varepsilon/\delta) \sum_{k=1}^K |\pi_k^* - \pi_k|,
\end{aligned}
$$

implying $\sum_{k=1}^K |\pi_k^* - \pi_k| \le \varepsilon/(\zeta - \varepsilon/\delta)$.

1.4 Proof of Theorem 4.2

We first ensure the existence of the test functions Φ_n, and Ψ_n as described in (20)-(23). First, we find find the covering numbers $N(\varepsilon, \mathcal{P}, W_1)$ and $N(\varepsilon, \mathcal{F}, h)$ to upper bound the Type I and II errors of the test functions Φ_n and Ψ_n. Note that

$$
h^2[f(\cdot \,|\, P_1) \,||\, f(\cdot \,|\, P_2)] \le \sum_{k=1}^K |\pi_{1,k} - \pi_{2,k}| + \sum_{k=1}^k \pi_{1,k} \|\mu_{1,k} - \mu_{2,k}\|.
$$

Hencd $N(\varepsilon, \mathcal{F}, h) \le N(\varepsilon^2/2, \mathcal{S}^{K-1}, ||\cdot||_1) \times \{N(\varepsilon^2/2, C_\mu, ||\cdot||)\}^K$ where $||\cdot||_1$ denotes the L_1 norm between two probability vectors and $||\cdot||$ denotes the Euclidean norm. From Lemma A.4 of [2], we obtain $N(\varepsilon^2/2, \mathcal{S}^{K-1}, ||\cdot||_1) \le (10/\varepsilon^2)^{K-1}$. Also, $\{N(\varepsilon^2/2, C_\mu, ||\cdot||)\} \le (2C_U/\varepsilon^2)^d$ for a global constant C_U is the diameter of the set C_μ. Then $N(\varepsilon, \mathcal{F}, h) \le (C/\varepsilon^2)^{dK}$ for some constant $C > 0$. To obtain an upper bound for $N(\varepsilon, \mathcal{P}, W_1)$, we note that

$$
W_1(P_1, P_2) \le \sum_{k=1}^K \max\{\pi_{1,k}, \pi_{2,k}\} \|\mu_{1,k} - \mu_{2,k}\| + C_U \sum_{k=1}^K |\pi_{1,k} - \pi_{2,k}|.
$$

Hence $N(\varepsilon, \mathcal{P}, W_1) \le N(\varepsilon/(2C_U), \mathcal{S}^{K-1}, ||\cdot||_1) \times \{N(\varepsilon/(2K), C_\mu, ||\cdot||)\}^K \le (CK/\varepsilon)^{dK}(10/\varepsilon)^{K-1}$. Hence $\log N(\varepsilon, \mathcal{F}, h) \lesssim dK \log(1/\varepsilon)$ and $\log N(\varepsilon, \mathcal{P}, W_1) \lesssim dK \log(K/\varepsilon)$. Then, we have from (20)-(21)

$$
\mathbb{E}_{P^*} \Phi_n \le e^{-C_1 n\varepsilon^2 + dK \log(1/\varepsilon)} \tag{4}
$$
$$
\mathbb{E}_P[1 - \Phi_n] \le e^{-C_2 n h^2[f(\cdot \,|\, P) \,||\, f(\cdot \,|\, P^*)]}, \tag{5}
$$

for any P with $h[f(\cdot \,|\, P) \,||\, f(\cdot \,|\, P^*)] > \varepsilon$. In this case, we choose $\varepsilon \equiv \varepsilon_n$ to be as constant multiple of $\{(dK/n)\log n\}^{1/2}$. Also, we have from (22)–(23)

$$
\mathbb{E}_{P^*} \Psi_n \le e^{-C_1 n\varepsilon^2 + dK \log(K/\varepsilon)} \tag{6}
$$
$$
\mathbb{E}_P[1 - \Psi_n] \le e^{-C_2 n W_1^2(P, P^*)}, \tag{7}
$$

for any P with $W_1(P, P^*) > \varepsilon$. In this case, we choose $\varepsilon \equiv \varepsilon_n$ to be as constant multiple of $\{(dK/n)\log(Kn)\}^{1/2}$.

Recall the two KL neighborhoods around (π^*, μ^*) with radius $(\varepsilon_\pi, \varepsilon_\mu)$ as

$$
\begin{aligned}
\mathcal{B}_n(\pi^*, \varepsilon_\pi) &= \Big\{ D(\pi^* \,||\, \pi) \le \varepsilon_\pi^2, \quad V(\pi^* \,||\, \pi) \le \varepsilon_\pi^2 \Big\}, \\
\mathcal{B}_n(\mu^*, \varepsilon_\mu) &= \Big\{ \sup_s D\big[p(\cdot \,|\, \mu^*, s) \,\big|\big|\, p(\cdot \,|\, \mu, s)\big] \le \varepsilon_\mu^2, \\
&\qquad \sup_s V\big[p(\cdot \,|\, \mu^*, s) \,\big|\big|\, p(\cdot \,|\, \mu, s)\big] \le \varepsilon_\mu^2 \Big\},
\end{aligned}
$$

where we used the shorthand $D(\pi^* \,||\, \pi) = \sum_s \pi_s^* \log(\pi_s^*/\pi_s)$ to denote the KL divergence between multinomial distributions with parameters $\pi^*, \pi \in \mathcal{S}_K$. We choose q_θ as the restriction of p_θ into $\mathcal{B}_n(\pi^*, \varepsilon_\pi) \times \mathcal{B}_n(\mu^*, \varepsilon_\mu)$.

It is easy to verify that under Assumption **R**, there exists some constant C_1 depending only on δ_0 such that $\mathcal{B}_n(\pi^*, \sqrt{K}\,\varepsilon) \supset \{\pi : \max_k |\pi_k - \pi_k^*| \le C_1\,\varepsilon\}$ (by using the inequality $D(p \,||\, q) \ge 2\,h^2(p \,||\, q)$). In addition, for Gaussian mixture model, it is easy to verify that the KL neighborhood $\mathcal{B}_n(\mu^*, \varepsilon)$ contains the set $\{\mu : \max_k \|\mu_k - \mu_k^*\| \le 2\,\varepsilon\}$. As a consequence, with $\varepsilon_\pi = \sqrt{K}\,\varepsilon$ and $\varepsilon_\mu = \varepsilon$ yields (using the prior thickness assumption and the fact that the volumes of $\{\pi : \max_k |\pi_k - \pi_k^*| \le C_1\,\varepsilon\}$ and $\{\mu : \max \|\mu_k - \mu_k^*\| \le C_2\,\varepsilon\}$ are at least $\mathcal{O}(\varepsilon^{-K})$ and $\mathcal{O}\big((\sqrt{d}/\varepsilon)^{dK}\big)$ respectively). Then we have from Theorem 3.2,with probability tending to one as $n \to \infty$,

$$
\int \Big\{ h^2\big[f(\cdot \,|\, \theta) \,\big|\big|\, f(\cdot \,|\, \theta^*)\big] \Big\} \widehat{q}_\theta(\theta)\, d\theta \lesssim \frac{d\,K}{n} \log n + K\,\varepsilon^2 + \frac{d\,K}{n}\, \log \frac{d}{\varepsilon}.
$$

Choosing $\varepsilon = \sqrt{d/n}$ in the above display yields the claimed bound.

Also, we have with high probability

$$\int \Big\{ W_1^2 \big[f(\cdot \,|\, \theta) \,||\, f(\cdot \,|\, \theta^*) \big] \Big\} \, \widehat{q}_\theta(\theta)\, d\theta \lesssim \frac{d\,K}{n} \log(Kn) + K\,\varepsilon^2 + \frac{d\,K}{n} \, \log \frac{d}{\varepsilon}.$$

Choosing $\varepsilon = \sqrt{d/n}$ in the above display yields the claimed bound noting that the first term in the right hand side of the preceding display is dominant.

1.5 Proof of Theorem 4.1

Under the notation in the paper, for each $n = 1, \ldots, N$, the latent variable $S_n = \{z_{dn} : d = 1, \ldots, D\}$. We use Theorem 3.2 with $d = h$ (Hellinger metric) and view each latent variable S_n per observation in the theorem as a block of D independent latent variable per observation. The existence of the test is automatic [3] with the Hellinger metric (parameter space is compact). This leads to that with probability tending to one as $N \to \infty$,

$$\int \sum_{d=1}^{D} h^2 \big[p_d(\cdot \,|\, \theta) \,||\, p_d(\cdot \,|\, \theta^*) \big] \, d\theta \leq \bigg(\sum_{d=1}^{D} \varepsilon_{\gamma_d}^2 + \sum_{k=1}^{K} \varepsilon_{\beta_k}^2 \bigg) + \Big\{ - \frac{1}{N} \sum_{d=1}^{D} \log P_{\gamma_d} \big[\mathcal{B}_N(\gamma_d^*,\, \varepsilon_{\gamma_d}) \big] \Big\} + \Big\{ - \frac{1}{N} \sum_{k=1}^{K} \log P_{\beta_k} \big[B_N(\beta_k^*,\, \varepsilon_{\beta_k}) \big] \Big\},$$

where KL neighborhoods $\mathcal{B}_N(\gamma_d^*;\, \varepsilon_{\gamma_d}) := \big\{ D(\gamma_d^* \,||\, \gamma_d) \leq \varepsilon_{\gamma_d}^2,\ V(\gamma_d^* \,||\, \gamma_d) \leq \varepsilon_{\gamma_d}^2 \big\}$, for $d = 1, \ldots, D$, and $B_N(\beta_k^*,\, \varepsilon_{\beta_k}) = \big\{ \max_k D\big[p(\cdot \,|\, \beta_k,\, k) \,||\, p(\cdot \,|\, \beta_k,\, k) \big] \leq \varepsilon_{\beta_k}^2,\ \max_{S_n} V\big[p(\cdot \,|\, \beta_k,\, k) \,||\, p(\cdot \,|\, \beta_k,\, k) \big] \leq \varepsilon_{\beta_k}^2 \big\}$.

Let S_k^β denote the index set corresponding to the non-zero components of β_k for $k = 1, \ldots, K$, and S_d^γ the index set corresponding to the non-zero components of γ_d for $d = 1, \ldots, D$. Under Assumption **S**, it is easy to verify that for some sufficiently small constants $c_1, c_2 > 0$, it holds for all $d = 1, \ldots, D$ that $\mathcal{B}_N(\gamma_d^*,\, \varepsilon_{\gamma_d}) \supset \big\{ \|(\gamma_d)_{(S_d^\gamma)^c}\|_1 \leq c_1\, \varepsilon_{\gamma_d},\ \|(\gamma_d)_{S_d^\gamma} - (\gamma_d^*)_{S_d^\gamma}\|_\infty \leq c_1\, \varepsilon_{\gamma_d} \big\}$, and for all $k = 1, \ldots, K$ that $\mathcal{B}_N(\beta_k^*,\, \varepsilon_{\beta_k}) \supset \big\{ \|(\beta_k)_{(S_k^\beta)^c}\|_1 \leq c_2\, \varepsilon_{\beta_k},\ \|(\beta_k)_{S_k^\beta} - (\beta_k^*)_{S_d^\beta}\|_\infty \leq c_2\, \varepsilon_{\beta_d} \big\}$. Applying Theorem 2.1 in [4], we obtain the following prior concentration bounds for high-dimensional Dirichlet priors

$$P_{\gamma_d}\big\{ \|(\gamma_d)_{(S_d^\gamma)^c}\|_1 \leq c_1\, \varepsilon_{\gamma_d},\ \|(\gamma_d)_{S_d^\gamma} - (\gamma_d^*)_{S_d^\gamma}\|_\infty \leq c_1\, \varepsilon_{\gamma_d} \big\} \gtrsim \exp\Big\{ - C\, e_d \, \log \frac{K}{\varepsilon_{\gamma_d}} \Big\},\ d = 1, \ldots, D;$$

$$P_{\beta_k}\big\{ \|(\beta_k)_{(S_k^\beta)^c}\|_1 \leq c_2\, \varepsilon_{\beta_k},\ \|(\beta_k)_{S_k^\beta} - (\beta_k^*)_{S_k^\beta}\|_\infty \leq c_2\, \varepsilon_{\beta_d} \big\} \gtrsim \exp\Big\{ - C\, d_k \, \log \frac{V}{\varepsilon_{\beta_k}} \Big\},\ k = 1, \ldots, K,$$

for some constant $C > 0$.

Putting pieces together, we obtain

$$\int \sum_{d=1}^{D} h^2 \big[p_d(\cdot \,|\, \theta) \,||\, p_d(\cdot \,|\, \theta^*) \big] \, d\theta \lesssim \bigg(\sum_{d=1}^{D} \varepsilon_{\gamma_d}^2 + \sum_{k=1}^{K} \varepsilon_{\beta_k}^2 \bigg) + \frac{1}{N} \sum_{d=1}^{D} e_d \, \log \frac{K}{\varepsilon_{\gamma_d}} + \frac{1}{N} \sum_{k=1}^{K} d_k \, \log \frac{V}{\varepsilon_{\beta_k}},$$

which leads to the desired bound by optimally choosing ε_{γ_d}'s and ε_{β_k}'s.

References

[1] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[2] Subhashis Ghosal and Aad W van der Vaart. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Annals of Statistics*, pages 1233–1263, 2001.

[3] Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

[4] Yun Yang and David B Dunson. Minimax optimal Bayesian aggregation. *arXiv preprint arXiv:1403.1345*, 2014.