STATISTICAL GUARANTEES AND ALGORITHMIC CONVERGENCE ISSUES OF VARIATIONAL BOOSTING

Biraj Subhra Guha
PhD Candidate, Department of Statistics
Texas A & M University

Anirban Bhattacharya
Associate Professor, Department of Statistics
Texas A & M University

Debdeep Pati
Associate Professor, Department of Statistics
Texas A & M University

October 23, 2020

ABSTRACT

We provide statistical guarantees for Bayesian variational boosting by proposing a novel small bandwidth Gaussian mixture variational family. We employ a functional version of Frank-Wolfe optimization as our variational algorithm and study frequentist properties of the iterative boosting updates. Comparisons are drawn to the recent literature on boosting, describing how the choice of the variational family and the discrepancy measure affect both convergence and finite-sample statistical properties of the optimization routine. Specifically, we first demonstrate stochastic boundedness of the boosting iterates with respect to the data generating distribution. We next integrate this within our algorithm to provide an explicit convergence rate, ending with a result on the required number of boosting updates.

1 INTRODUCTION

Variational Bayes has gained popularity in recent years as an alternative to Markov chain Monte Carlo procedures to approximate analytically intractable posterior distributions; refer to Blei et al. [2017] for a comprehensive overview. Variational inference formulates the problem of approximating the posterior as an optimization routine by minimizing a measure of discrepancy between probability densities in an approximating class and the posterior density. The variational solution refers to the closest member of the approximating class to the posterior, with closeness measured through divergences or metrics, usually Kullback–Leibler divergence. Other discrepancy measures for approximating the posterior have been studied, like the Wasserstein distance and Rényi divergence in Huggins et al. [2020], Fisher distance in Huggins et al. [2018] and Hellinger metric in Campbell and Li [2019].

The approximating class or the domain of optimization, commonly referred to as the variational family, plays a central role in these methods. It is chosen to strike a balance between computational tractability and approximation power. A richer, more flexible family allows better approximation of the posterior, while a simpler class of distributions facilitate calculations and computation speed. The Gaussian family is a popular example of a parametric variational family, where the optimization effectively takes place over a finite-dimensional parameter space. For a semi-parametric approach, one can use the popular mean-field family, which only assumes that the variational density factorizes over pre-specified sub-blocks of the parameter, with the factors otherwise unrestricted.

Statistical guarantees, frequentist validation as well as convergence issues focusing on mean-field appear in works like Yang et al. [2017], Pati et al. [2018], Alquier and Ridgway [2017], Zhang and Gao [2017], Wang and Blei [2018], Mukherjee et al. [2018], Yang et al. [2018] and Huggins et al. [2020]. However, mean-field approximations can only hope to recover the center of the posterior and fails to capture posterior co-dependence, so need for more general

families arise. Copula modelling has been used in Tran et al. [2015] and Han et al. [2016], while implicit distribution families have been used in Huszár [2017], Han et al. [2016], Titsias and Ruiz [2018], Yin and Zhou [2018], Molchanov et al. [2018] and Shi et al. [2017]. Another recent approach to gain modelling flexibility is to use mixture distributions as variational families, which is the focus of this paper. Wang et al. [2006] is an early theoretical work on Gaussian mixtures as variational family, focusing on conjugate priors. As proposed in Guo et al. [2016] and Miller et al. [2017], the structure of mixture families naturally gives rise to the idea of variational boosting. This computation method iteratively builds a mixture distribution approximation to the posterior by adding simple, new components and re-weighting them. The components of the mixture can thus be considered weak learners in this boosting framework, which are averaged in a weighted, sequential fashion to produce a mixture, the strong learner.

Variational boosting offers better computational efficacy due to iterative fitting, while simultaneously improving approximation prowess owing to the more flexible mixture distribution class. Guo et al. [2016] modify the boosting method based on L_2–regularized variational objective. Miller et al. [2017] incorporate covariance structure to modify the variational family. Locatello et al. [2017] provide some theoretical basis of this computational method using truncated densities as mixture components. However, their result is limited to compactly supported densities only, thus technically not including even Gaussian distributions. This idea is extended in Locatello et al. [2018] for black box variational inference; refer to Ranganath et al. [2014] for the original work on black box variational inference. Wang [2016] uses gradient boosting technique, and suffers from a drawback similar to Locatello et al. [2017]. Campbell and Li [2019] note that the domain of mixture families does not allow Kullback–Leibler divergence to be sufficiently smooth, hence switches to Hellinger metric to provide algorithm and theoretical study for boosting. We address the above problem with the boosting method by providing a pathway to work with mixture families and simultaneously maintaining the use of Kullback–Leibler divergence.

Our contribution to variational boosting revolves around frequentist properties of the variational solution. We study this by proposing a small bandwidth Gaussian mixture variational family and using a functional version of the Frank–Wolfe algorithm (refer to Frank and Wolfe [1956] for the original formulation) for the variational optimization routine. Our method relaxes the assumption in Locatello et al. [2017] regarding compact support of variational distributions, allows working with the standard choice of Kullback–Leibler divergence in contrast to Campbell and Li [2019], as well as makes assumptions that are strictly milder than Local Asymptotic Normality (LAN) type assumptions in Wang and Blei [2018]. Our first result is on understanding statistical properties of the global optimizer of the boosting algorithm. In particular, we show that the Kullback–Leibler divergence of the posterior from the optimal variational solution is bounded in probability, a phenomenon that is similar to what is observed in Bernstein-von Mises theorems for regular parametric models. Our second result pertains to convergence analysis of the algorithm. Our findings are less than encouraging, much along the conjecture of Campbell and Li [2019]. Specifically, we show that the number of iterations required for the boosting algorithm to converge is exponential in the inverse bandwidth, which is a parameter crucial to the definition of our small bandwidth mixture variational family. We provide intutive justification for this in sections 4.2 and 4.3.

2 BACKGROUND AND TARGET

We start with $\mathbb{R}^p$-valued data points $X_1 \ldots, X_n$ which are independently and identically distributed according to density $f(x; \theta)$, where $\theta \in \mathbb{R}^d$, the parameter space. Given a prior density $\pi(\theta)$ on parameter θ, we denote the posterior of θ as

$$\pi_n(\theta) = \frac{\prod_{i=1}^n f(X_i; \theta)\pi(\theta)}{\int \prod_{i=1}^n f(X_i; \theta)\pi(\theta)d\theta}.$$

Variational boosting works with the following type of distribution family on $\mathbb{R}^d$:

$$\mathcal{Q} = \text{conv}(\Gamma) = \left\{ \sum_{k=1}^K \beta_k \phi_k : \phi_k \in \Gamma, \ \boldsymbol{\beta} \in \Delta^K, \ K \geq 1 \right\},$$

where Γ is any family of simple distributions on $\mathbb{R}^d$ and Δ^K denotes the unit simplex in $\mathbb{R}^K$. The variational family for boosting framework is thus a flexible mixture family. The main aim is to find $q_n^*(\mathcal{Q}) \in \mathcal{Q}$ such that

$$q_n^*(\mathcal{Q}) = \text{argmin}\{q \in \mathcal{Q} \mid KL(q||\pi_n)\}, \quad m_n^*(\mathcal{Q}) = KL(q_n^*(\mathcal{Q})||\pi_n), \tag{1}$$

where, $KL(a||b) = \int a(\theta) \log(a(\theta)/b(\theta))d\theta$ denotes the Kullback–Leibler divergence of density b from density a, both defined on the parameter space. This proceeds through an optimization routine, called the boosting technique. To describe the algorithm, let $\psi_n^{(k)}$ denote the k-th iterate in the algorithm for $k \geq 0$. Given $\psi_n^{(k)}$, the next iterate $\psi_n^{(k+1)}$ is obtained by

$$\psi_n^{(k+1)} = (1 - \gamma_K)\psi_n^{(k)} + \gamma_k \phi_n^{(k+1)},$$

where the weight $\gamma_k \in [0,1]$ and $\phi_n^{(k+1)} \in \Gamma$ depend on the boosting approach employed. Locatello et al. [2017] proposed the use of Frank–Wolfe algorithm to tackle boosting technique iterates and our setup bears similarity to theirs. Iterates of this kind are also used in Guo et al. [2016] and Locatello et al. [2018]. Observe that, if $\psi_n^{(k)}$ is already a mixture distribution, every iteration just adds a new component, namely $\phi_n^{(k+1)}$, to the mixture. The Frank–Wolfe algorithm (see appendix; also refer to Jaggi [2013] and Frank and Wolfe [1956]) handles optimization by proceeding exactly in this fashion, and hence is a natural choice as variational algorithm in this case. A quantity $\mathcal{C}_n$ called curvature (see appendix and (15) in section 4.3 for definition), that depends only on the the objective map $q \mapsto KL(q||\pi_n)$ and domain of optimization $\mathcal{Q}$, plays a crucial role in this algorithm. After initializing with some $\phi_n^{(0)} \in \Gamma$, the k-th step of the routine involves finding the new component $\phi^{(k+1)}$ to be added, through a linear minimization routine called linear minimization oracle (LMO). This intermediate step is carried out by solving the LMO approximately in terms of the derivative of the objective map and curvature $\mathcal{C}_n$. Now note that, $q_n^*(\mathcal{Q})$ defined in (1) is a random quantity with respect to data $X_1, \ldots X_n$, and so is each iterate of the boosting routine. We aim to provide statistical properties of the random quantities $m_n^*(\mathcal{Q})$ and $\psi_n^{(k)}$ with respect to the true data generating distribution. As a first step, we show stochastic boundedness of $m_n^*(\mathcal{Q})$ in our theorem 1. We next use theorem 1 from Jaggi [2013] and our result on upper bounding the curvature $\mathcal{C}_n$ (theorem 2) to upper bound the decrements of objective value in the boosting algorithm in terms of the variational family hyper-parameters and sample size n. Finally, we tie up the above two results to gain parity of the theoretical minimum and the algorithm. We end with a result (corollary 2) on the order of the required number of boosting updates for a certain degree of error.

3 STATISTICAL PROPERTIES OF THE VARIATIONAL OPTIMIZER

3.1 The small bandwidth mixture Gaussian family

Recall the definition of $m_n^*(\mathcal{Q})$, the minimum of the objective map $q \mapsto KL(q||\pi_n)$ over domain $\mathcal{Q}$. Since the function $a \mapsto KL(a||b)$ has closed sub-level sets, this minimum is attained, i.e. $m_n^*(\mathcal{Q}) = KL(q_n^*(\mathcal{Q})||\pi_n)$ is the minimum corresponding to the domain $\mathcal{Q}$. $m_n^*(\mathcal{Q})$ is a random quantity with respect to data $X_1 \ldots X_n$, so we can make probability statements about it with respect to the true data generating distribution. Before we state our stochastic boundedness theorem (section 3.3), we discuss our setup and introduce our variational family. Consider the following restricted Gaussian family for a fixed c_0 with $1 < c_0 < 2$, and some $M, \sigma_n > 0$:

$$\Gamma_n = \left\{ N\left(\mu, \sigma^2 I_d\right) : \|\mu\|_2 \leq M, \ 0 < \sigma_n \leq \sigma \leq c_0^{1/2}\sigma_n \right\}. \tag{2}$$

Denoting by $\mathrm{conv}(\Gamma_n)$ the set of all finite affine combinations of members of Γ_n, we define $\mathcal{Q}_n = \mathrm{conv}(\Gamma_n)$ as the following restricted mixture Gaussian family:

$$\mathcal{Q}_n = \left\{ \sum_{k=1}^K \beta_k \phi_k : \phi_k \in \Gamma_n, \ \boldsymbol{\beta} \in \Delta^K, \ K \geq 1 \right\}. \tag{3}$$

This domain $\mathcal{Q}_n$ is our variational family of choice, which we call the small bandwidth mixture Gaussian family. Observe that, the components of the mixtures are isotropic Gaussians with means lying in the radius-M compact Euclidean ball around zero in $\mathbb{R}^d$, while the variance σ is constrained to be of the same order as σ_n, the bandwidth parameter. The specific constraint on the constant c_0 plays a crucial role in our analysis and will be justified when we study the employed algorithm and its convergence in detail in section 4.

3.2 Comparison of divergences and Bernstein-von Mises Phenomenon

The Bernstein-von Mises (BvM) theorem is a well-known frequentist phenomenon for Bayesian posteriors; refer to Van der Vaart [2000] for an overview of the BvM phenomenon. It encompasses results about the asymptotic normality of appropriately scaled posterior distributions under regularity conditions on the likelihood and the prior. Local Asymptotic Normality (LAN) assumptions on the likelihood is a pathway to achieving BvM, and Wang and Blei [2018] employ it in the context of variational inference. However, BvM results use total variation distance (d_{TV}) as the metric, and we wish to focus on Kullback–Leibler discrepancy. In this context, we give a brief comparison of Kullback–Leibler divergence (KL) with total variation distance and Hellinger distance (d_H). We choose to work with a simplified setting in order to help emphasize our point. Suppose we use a single normal distribution to approximate the posterior. Say $X_1, \ldots, X_n$ are d-vectors, which are independently and identically distributed as $N(\theta, \Sigma)$, with $\theta \sim N(\mu_0, \Sigma_0)$ and μ_0, Σ, Σ_0 known. Let $\overline{X}_n$ denote the sample mean, and μ_n, Σ_n the posterior normal's mean and variance respectively. Let θ_0 denote the true parameter. Then the following statements follow from straightforward calculations for Gaussians,

Results:

$$KL\left(N\left(\theta_0, n^{-1}\Sigma\right) || N\left(\mu_n, \Sigma_n\right)\right) \rightsquigarrow \frac{1}{2}\chi_d^2, \tag{4}$$

$$d_{TV}\left(N\left(\mu_n, \Sigma_n\right), N\left(\theta_0, n^{-1}\Sigma\right)\right) \to 0 \quad \text{a.s.,} \tag{5}$$

$$d_H\left(N\left(\mu_n, \Sigma_n\right), N\left(\theta_0, n^{-1}\Sigma\right)\right) \to 0 \quad \text{a.s.,} \tag{6}$$

$$KL\left(N\left(\overline{X}_n, n^{-1}\Sigma\right) || N\left(\mu_n, \Sigma_n\right)\right) \to 0 \quad \text{a.s.,} \tag{7}$$

where '$\rightsquigarrow$' denotes weak convergence and a.s stands for almost sure validity with respect to the data generating distribution. Refer to the appendix for proofs of these statements. Since d_{TV} and d_H are equivalent distances, (5) and (6) imply each other, so we just compare (4) with (5). The result in (5) says the posterior comes close to a single Gaussian distribution in total variation, a.s with respect to the data. However the very same distributions are not close in Kullback–Leibler divergence, even in the simplest Gaussian case, as pointed out by (4). This suggests that under Kullback–Leibler divergence, which is a stronger measure of discrepancy, the divergence between the posterior and a deterministic approximator of it should not go down to zero. We shall see this property in play in our theorem 1.

Now note the comparison of (4) and (7), where, just by changing the centering from the truth (a deterministic quantity) to the sample mean (a random data-dependent quantity), we achieve convergence to zero under Kullback–Leibler divergence. However, this phenomenon is very special to this case, as the correct centering may be computationally impossible to find for complicated posteriors. Hence, (4) is of more practical importance to us than (7) as a statistical statement. In the next section, we present a result similar in flavor to (4), but milder and applicable much more generally.

In general, for the BvM phenomenon to hold, strong regularity conditions are required, which guarantee posterior shape (Gaussian) with high probability with respect to data. However, we wish to include those posteriors in our analysis as well whose shapes are non-Gaussian, making our analysis more general. Works in Kruijer et al. [2010] and Shen et al. [2013] establish approximations of deterministic densities, suitably smooth and exponentially tailed, using Gaussian mixtures. However, if we wish to apply such results to the posterior, which is a random density, we need high probability statements about the smoothness and tail of the posterior, which might tantamount to using hypotheses that the BvM phenomenon demands.

3.3 Stochastic Boundedness of the Kullback–Leibler Discrepancy

We now state a theorem about the theoretical minimum Kullback–Leibler divergence $m_n^*(\mathcal{Q}_n)$; see section 2 for definition of m_n^* and section 3.1 for definition of $\mathcal{Q}_n$. Recall that $X_1, \ldots X_n$ are independently and identically distributed data points following density $x \mapsto f(x; \theta)$ and $\pi(\theta)$ is the prior on the parameter $\theta \in \mathbb{R}^d$. With $\theta_0 \in \mathbb{R}^d$ denoting the true parameter value, define the log-likelihood ratio $\ell_i(\theta, \theta_0) = \log\left(f(X_i; \theta)/f(X_i; \theta_0)\right)$ for $i = 1, \ldots, n$, $L_n(\theta, \theta_0) = \sum_{i=1}^n \ell_i(\theta, \theta_0)$, $KL(\theta_0||\theta) = -E(\ell_1(\theta, \theta_0))$, $\mu_2(\theta_0||\theta) = E(\ell_1(\theta, \theta_0))^2$ and $U(\theta) = -\log(\pi(\theta))$ Also, denote by $KL^{(j)}(\theta_0||\theta)$ and $\mu_2^{(j)}(\theta_0||\theta)$ for $j = 1, 2$ the respective derivatives of the maps with respect to the second argument. Let $s_{\max}(A)$ stand for the highest singular value of square matrix A. $\|.\|_2$ stands for the l_2 norm on $\mathbb{R}^d$. Denote by $\lesssim, \gtrsim$ the corresponding inequalities up to absolute constants. The following assumptions will be required for the theorem:

Assumption 1: The truth θ_0 satisfies $\|\theta_0\|_2 \le M$.

Assumption 2: The variance bound σ_n satisfies $\sigma_n \le n^{-1/2} \le c_0^{1/2}\sigma_n$ for all $n \ge 1$.

Assumption 3: The quantities $KL(\theta_0||\theta), \mu_2(\theta_0||\theta)$ are finite for every $\theta \in \mathbb{R}^d$.

Assumption 4: Matrices $KL^{(2)}(\theta_0||\theta), \mu_2^{(2)}(\theta_0||\theta)$ and $U_2^{(2)}(\theta)$ exist on $\mathbb{R}^d$ and satisfy for any $\theta, \theta' \in \mathbb{R}^d$:

$$s_{max}\left(KL^{(2)}(\theta_0||\theta) - KL^{(2)}(\theta_0||\theta')\right) \lesssim \|\theta - \theta'\|_2^{\alpha_1},$$

$$s_{max}\left(\mu_2^{(2)}(\theta_0||\theta) - \mu_2^{(2)}(\theta_0||\theta')\right) \lesssim \|\theta - \theta'\|_2^{\alpha_2}, \tag{8}$$

$$s_{max}\left(U_2^{(2)}(\theta) - U_2^{(2)}(\theta')\right) \lesssim \|\theta - \theta'\|_2^{\alpha_3},$$

for some $\alpha_1, \alpha_2, \alpha_3 \ge 0$.

Assumption 5: $KL(\theta_0||\theta) \gtrsim \|\theta - \theta_0\|_2^2$.

4

Theorem 1: *Under assumptions 1-5, it holds that $m_n^*(\mathcal{Q}_n)$ is bounded in probability with respect to the data generating distribution, i.e. given any $\epsilon \in (0,1)$, there exists $M_\epsilon, N_\epsilon > 0$ such that for all $n \geq N_\epsilon$, we have with probability greater than $1 - \epsilon$*

$$m_n^*(\mathcal{Q}_n) < M_\epsilon. \tag{9}$$

Remark 1: Assumption 2 dictates the exact order of σ_n, and also allows $N_\epsilon = 1$. It is at par with the order of the bandwidth expected in parametric estimation. Finiteness of $\mu_2(\theta_0||\theta)$ in assumption 3 is crucial for concentration inequalities to be applied. The smoothness assumption i.e. assumption 4, helps dictate posterior moments, but not the shape of the entire posterior. The final assumption is the standard identifiability condition for using Kullback–Leibler divergence. Assumptions of these types are quite common in the literature; refer to the moment assumptions for the posterior in Huggins et al. [2020] in the context of distributional bounds in variational inference and section 5 in Ghosal et al. [2000] in the context of posterior contraction. An important observation is the fact that we do not intend to recover the posterior mean as in Pati et al. [2018], where assumptions are aimed at studying variational point estimates.

Remark 2: The hypothesis for the theorem is milder than what can guarantee a weak convergence type result, like the (4) in section 3.2. There are no assumptions that allow local quadratic nature of the posterior, as we impose conditions only on the expected log-likelihood. This makes our assumptions more general than a BvM type setup. Contrast this with the stochastic LAN assumption in Wang and Blei [2018], which approximates the likelihood with a stochastic linear term and a deterministic quadratic term.

Remark 3: Theorem 1 establishes $m_n^*(\mathcal{Q}_n)$ to be an $O_p(1)$ quantity with respect to the true data generating distribution. From the proof in appendix, one can further conclude $M_\epsilon \gtrsim \epsilon^{-1/2}$ for small ϵ.

We shall state a corollary which is a simplification of theorem 1 in the case the density $x \mapsto f(x:\theta)$ belongs to the K-parameter exponential family on $\mathbb{R}^p$. Let $\theta \in \mathbb{R}^d$ be the canonical parameter, $T_l : \mathbb{R}^p \mapsto \mathbb{R}$, $l = 1, \ldots K$ be the sufficient statistics and $A : \mathbb{R}^d \mapsto \mathbb{R}$ be the log-partition function. The form of the density is given by

$$f(x; \theta) = \exp\left(\sum_{l=1}^K \theta_l T_l(x) - A(\theta)\right).$$

Let $A^{(j)}(\theta), j = 1, 2$ denote the respective derivatives of $A(\theta)$. We shall use the notion of strong convexity in the corollary that follows; refer to the appendix for a general definition. We also need to note down the definition of α-Lipschitz functions:

Definition: Vector or square-matrix valued functions f defined on $D \subset \mathbb{R}^d$ are said to be α-Lipschitz for an $\alpha > 0$, if there exists a constant $C > 0$ such that for all $x, y \in D$

1. $\|f(x) - f(y)\|_2 \leq C\|x - y\|_2^\alpha$ for vector valued functions, and
2. $s_{\max}(f(x) - f(y)) \leq C\|x - y\|_2^\alpha$ for square-matrix valued functions.

Corollary 1: *Assume that $A(\theta)$ is twice differentiable and strongly convex on $\mathbb{R}^d$. Also assume that the vectors $A^{(1)}(\theta), A^{(2)}(\theta)\theta$ and the square matrices $A^{(1)}(\theta)A^{(1)}(\theta)^T, A^{(2)}(\theta)$ are α-Lipschitz functions of $\theta \in \mathbb{R}^d$ for some $\alpha \geq 0$. Under these conditions, assumptions on the expected likelihood (assumption 3 and 4) in theorem 1 hold.*

3.4 Sketch of Proof of Theorem 1

We now briefly discuss the proof technique of theorem 1, shedding more light on the importance of the assumptions made; refer to appendix for a detailed proof of theorem 1. Note that $m_n^*(\mathcal{Q}_n)$ is bounded above by the objective map $q \mapsto KL(q||\pi_n)$ evaluated at any member of $\mathcal{Q}_n$. We choose that member to be q_0, the d-dimensional Gaussian density centered at the truth θ_0 and variance $\sigma_n^2 I_d$. Here, I_d stands for the d-dimensional identity matrix and σ_n satisfies assumption 2. Along with assumption 1, we have, $q_0 \in \mathcal{Q}_n$ and hence $m_n^*(\mathcal{Q}_n) \leq KL(q_0||\pi_n)$. Thus it is enough to show $KL(q_0||\pi_n)$ is bounded in probability. This Kullback–Leibler discrepancy can now be broken down in a sum to give two deterministic and two random quantities. The stochastic part of the sum is given by

$$\log\left(m(X_n)\right) - \left(\int L_n(\theta, \theta_0)q_0(\theta)d\theta\right),$$

where

$$m(X_n) = \int \exp\left(L_n(\theta, \theta_0)\right)\pi(\theta)d\theta.$$

Algorithm 1 Functional Frank–Wolfe Algorithm with small bandwidth mixture Gaussian variational family

1. Initialize with $\psi_n^{(0)} = \phi_n^{(0)} \in \Gamma_n$.

2. At k-th step, calculate the subgradient $s_n^{(k)} = \log(\psi_n^{(k)}) - \log(\pi_n)$ at the k-th iterate $\psi_n^{(k)}$.

3. Set $\gamma_k = 2/(k+2)$, solve LMO approximately i.e. find $\phi_n^{(k+1)} \in \Gamma_n$ such that $\int s_n(\theta) \phi_n^{(k+1)} d\theta \leq \min \left\{ \phi \in \Gamma_n \mid \int s_n^{(k)} \phi(\theta) d\theta \right\} + \gamma_k \mathcal{C}_n/2$.

4. Update $\psi_n^{(k+1)} = (1 - \gamma_k) \psi_n^{(k)} + \gamma_k \phi_n^{(k+1)}$ to get the $(k+1)$-th iterate.

Under true data generating distribution, the integrand defining $m(X_n)$ has expectation 1, which helps upper bound in probability the first stochastic term above. For the second, we notice that $L_n(\theta, \theta_0)$ has expectation $-nKL(\theta_0||\theta)$ under true θ_0. We then utilize the smoothness and identifiability assumptions on $KL(\theta_0||\theta)$ to obtain its Taylor series expansion around θ_0, that helps lower bound in probability the second stochastic term through Chebyshev's inequality. We can now conclude our result by noting that sum of $O_p(1)$ quantities is again $O_p(1)$.

4 CONVERGENCE ANALYSIS OF THE ALGORITHM

4.1 Steps of the Algorithm

We opt for a functional version of the Frank–Wolfe algorithm as our variational boosting algorithm (algorithm 1), which bears analogy to variant 0 of algorithm 1 in Locatello et al. [2017]. Refer to Jaggi [2013] for the general algorithm and our discussion in the appendix for a brief overview and notations. Recall that we aim to perform the optimization (1) with domain $\mathcal{Q} = \mathcal{Q}_n = \text{conv}(\Gamma_n)$. Thus our objective map is $q \mapsto KL(q||\pi_n), q \in \mathcal{Q}_n$, which is convex on its domain. Let us generically denote members of Γ_n, which are single Gaussians, by ϕ and those of $\mathcal{Q}_n$, which are mixture Gaussians, by ψ. Superscripts stand for iterate numbers and the subscript n denotes dependence on sample size. Say $\psi_n^{(k)}$ is the mixture obtained as the k-th step iterate. We use the notation $\mathcal{D}_n$ to denote the Bregman divergence of our convex objective map (refer to appendix for general definition of Bregman divergence). For $\psi_1, \psi_2 \in \mathcal{Q}_n$, the Bregman divergence $\mathcal{D}_n$ of ψ_1 from ψ_2, at ψ_1, under the objective map, is given by

$$\mathcal{D}_n(\psi_2||\psi_1) = KL(\psi_2||\pi_n) - KL(\psi_1||\pi_n) - \int (\psi_2 - \psi_1) \left(\log(\psi_1) - \log(\pi_n) \right) d\theta. \tag{10}$$

The last term above derives from the fact that our domain of optimization, $\mathcal{Q}_n$, lies embedded in the L_2 inner product function space on $\mathbb{R}^d$. The term $\log(\psi_1) - \log(\pi_n)$, appearing within the integrand in (10), is the sub-gradient (see appendix for general definition of subgradient) of the objective map at ψ_1, and $\psi_n^{(k)}$ plays the role of ψ_1 at the k-th step of the algorithm. We use the notation $\mathcal{C}_n$ to denote the curvature of the objective map for the domain $\mathcal{Q}_n$ (see section 4.3 for details on $\mathcal{C}_n$).

That $s_n^{(k)} = \log(\psi_n^{(k)}) - \log(\pi_n)$ in algorithm 1 is indeed a valid subgradient of the objective map at $\psi_n^{(k)}$, follows from the following lemma on Bregman divergence $\mathcal{D}_n$ and the non-negativity of Kullback–Leibler divergence. It basically says calculating $\mathcal{D}_n$ and Kullback–Leibler divergence of the objective map are one and the same.

Lemma 1: *For any densities ψ_1 and ψ_2 defined on $\mathbb{R}^d$, we have*

$$\mathcal{D}_n(\psi_2||\psi_1) = KL(\psi_2||\psi_1).$$

In algorithm 1, the target is to greedily fit single Gaussian components to build a mixture of Gaussians, that is close to the posterior in Kullback–Leibler divergence. Let the optimal approximating mixture be denoted by $\psi_n^* = q_n^*(\mathcal{Q}_n)$ (see (1) for definition of q_n^*). Note that, both $\psi_n^{(k)}$ and ψ_n^* are random with respect to data. But the practitioner is given the data $X_1, \ldots X_n$, hence she runs algorithm 1 deterministically and upper bounds, point-wise on the sample space, the random quantity

$$KL(\psi_n^{(k)}||\pi_n) - KL(\psi_n^*||\pi_n).$$

We want to find the aforementioned upper bound in terms of number of iterations k and sample size n.

Step 3 of algorithm 1 is an approximate linear minimization routine for which we shall use the parametric structure of Gaussians and optimize over the parameters. The practitioner starts with an initial guess of $\mu^{(0)}, \sigma^{(0)}$ such that $N\left(\mu^{(0)}, (\sigma^{(0)})^2 I_d\right) \in \Gamma_n$. This gives her the Gaussian $\phi_n^{(0)} = N\left(\mu^{(0)}, (\sigma^{(0)})^2 I_d\right)$. At the beginning of the k-th

step, she has the mixture $\psi_n^{(k-1)} = \sum_{j=1}^{k-1} \beta_j \phi_n^{(j)}$, where the vector $\beta \in \Delta^{k-1}$ and are positive functions of γ_l's with $\gamma_l = 2/(l+2), l = 1 \ldots k-1$. In order to obtain the Gaussian $\phi_n^{(k)}$, she finds approximate $\mu^{(k)}, \sigma^{(k)}$ through the LMO optimization routine

$$\operatorname{argmin}\left\{ \|\mu\|_2 \leq M, \ 0 < \sigma_n \leq \sigma \leq \sqrt{c_0}\sigma_n : \int \phi(\theta; \mu, \sigma^2) \log\left(\frac{\psi_n^{(k-1)}(\theta)}{\pi_n(\theta)}\right) d\theta \right\}. \tag{11}$$

By lemma 5 of Jaggi [2013], the practitioner is allowed to use any algorithm at her disposal for the above routine, as long as she is able to perform the optimization of this k-th step with error $\leq \gamma_k \mathcal{C}_n/2$. A very important observation is that knowing the normalizing constant of the posterior is not necessary for the above routine. Thus, using the $\phi_n^{(k)}$ obtained, we update $\psi_n^{(k)} = (1 - \gamma_k)\psi_n^{(k-1)} + \gamma_k \phi_n^{(k)}$.

4.2 Rate of Convergence of Algorithm 1

We now state our main theorem on rate of convergence:

Theorem 2:

$$KL(\psi_n^{(k)}||\pi_n) - KL(\psi_n^*||\pi_n) \leq \frac{8(2 - c_0)^{-d/2} \exp\left(\frac{2M^2}{(2-c_0)\sigma_n}\right)}{k + 2}. \tag{12}$$

This theorem upper bounds the gap in value of the Kullback–Leibler objective map, between the k-th boosting iterate and the optimal approximator to the posterior. It depends upon the sample size, number of iterations, dimension of parameter space and hyper-parameters of the domain $\mathcal{Q}_n$. We make the following important remarks:

Remark 4: The above convergence-rate holds point-wise with respect to the data generating distribution, and the upper bound is deterministic.

Remark 5: The rate is sub-linear in the number of iterates k and exponential in inverse bandwidth σ_n^{-1}. Sub-linearity follows from the use of Frank–Wolfe, while exponentiality is an artifact of enforcing Kullback–Leibler divergence to be strongly smooth over our small bandwidth domain.

We now combine theorems 1 and 2 to obtain a novel probability statement about the random k-th iterate of our boosting algorithm. It is important to know how many iterations we need in order to obtain a certain error in the algorithm, and that depends on the sample size n. So we let k vary with sample size n, i.e. take $k \equiv k_n$.

Corollary 2: *If $k_n \gtrsim \exp(n^{1/2})$, then for any $\epsilon \in (0,1)$, there exists constant $C > 0$, such that with probability greater than $1 - \epsilon$ we have*

$$KL(\psi_n^{(k_n)}||\pi_n) < C\epsilon^{-1/2}. \tag{13}$$

Remark 6: Corollary 2 shows the the required order of k_n, the number of iterations, in terms of sample size n that maintains order parity while combining theorems 1 and 2. It shows after how many runs of the boosting algorithm the random iterates are guaranteed to be bounded in probability with respect to the data generating distribution.

4.3 Sketch of Proof of Theorem 2

We briefly discuss how we arrived at theorem 2. For k-th step, theorem 1 of Jaggi [2013] allows the following upper bound regarding our objective map:

$$KL(\psi_n^{(k)}||\pi_n) - KL(\psi_n^*||\pi_n) \leq \frac{4\mathcal{C}_n}{k + 2}, \tag{14}$$

where $\mathcal{C}_n$ denotes the curvature of the objective map $q \mapsto KL(q||\pi_n)$ over domain of optimization $\mathcal{Q}_n$. It is given by

$$\mathcal{C}_n = \sup\left\{\frac{2}{\alpha^2}\mathcal{D}_n(\psi_2||\psi_1) : \psi_1 \in \mathcal{Q}_n, \phi \in \Gamma_n, \alpha \in [0,1], \psi_2 = \psi_1 + \alpha(\phi - \psi_1)\right\}. \tag{15}$$

Curvature is thus essentially the maximum scaled Bregman divergence between densities in domain $\mathcal{Q}_n$ and their perturbations through mixtures. It is entirely determined by n, σ_n and the variational family hyper-parameters M, c_0, d. Refer to the appendix for a general definition of curvature. We prove the following lemma that upper bounds $\mathcal{C}_n$ on the small bandwidth mixture Gaussian domain $\mathcal{Q}_n$:

Lemma 2:

$$\mathcal{C}_n \leq 2(2 - c_0)^{-d/2} \exp\left(\frac{2M^2}{(2 - c_0)\sigma_n}\right).$$
(16)

Remark 6: Note the exponential dependence of the bound on inverse bandwidth σ_n^{-1}. Whether $\mathcal{C}_n$ is a finite quantity depends on the constraints in the definition of Γ_n (see section 3.1) through which it is defined. We are enforcing finiteness of $\mathcal{C}_n$ through utilizing proposed small bandwidth mixture Gaussian family. which is the essence of the constraint $1 < c_0 < 2$ in the definition of Γ_n.

Remark 7: Locatello et al. [2017] and our work share the central motive of ensuring strong smoothness (see definition in appendix) of Kullback–Leibler divergence, which essentially dictates finiteness of $\mathcal{C}_n$. It is a property of both the objective function to be minimized, and the domain over which it is minimized. However, Locatello et al. [2017] assume Γ_n to consist of truncated, lower bounded densities, which although works from a practical point of view, excludes the very basic Gaussian distribution. Their theory only accommodates compactly supported densities, which is too restrictive and is remedied through the above lemma. Thus lemma 2 is a cardinal contribution of this work; it shows that the curvature of Kullback–Leibler discrepancy is bounded over the domain of small bandwidth Gaussian mixture family.

Remark 8: If $c_0 \geq 2$, then upper bound on $\mathcal{C}_n$ improves with smaller bandwidth σ_n, which is evidently untrue, as smaller bandwidth Gaussian mixtures are spikier and worse approximators through Kullback–Leibler divergence. If $c_0 < 1$, the bound improves for higher dimension, which contradicts curse of dimensionality. This intuitively justifies the technical importance of the constraint: $1 < c_0 < 2$.

We now state a straightforward formula that shall play a crucial role in the proof of lemma 2, as well as provide insight into why the proposed family of small bandwidth Gaussian mixtures is a good choice as a variational family.

Lemma 3: *Whenever $2\sigma_1^2 > \sigma_2^2 > 0$, χ^2 distance between two Gaussians is given by*

$$\chi^2\left(\mathrm{N}\left(\mu_2, \sigma_2^2\mathbf{I}_d\right) \,||\, \mathrm{N}\left(\mu_1, \sigma_1^2\mathbf{I}_d\right)\right) = -1 + \left(\frac{\sigma_1^2}{\sigma_2\sqrt{2\sigma_1^2 - \sigma_2^2}}\right)^d \exp\left(\frac{\|\mu_2 - \mu_1\|_2^2}{2\sigma_1^2 - \sigma_2^2}\right).$$
(17)

Remark 8: The χ^2 distance (refer to appendix for definition) arises during the calculations of curvature of Kullback–Leibler divergence, and one can clearly see the right hand side above is defined as a real number only for a certain interval of values of the variances. This calculation was an important motivator for the small bandwidth variational family. It is also the reason for the exponential dependence of the upper bound, in theorem 2, on the inverse bandwidth.

We can now directly plug into lemma 3 the domain parameters of $\mathcal{Q}_n$ to calculate the proposed upper bound in lemma 2. In turn, lemma 2 directly gives us theorem 2 in combination with (14).

5 CONCLUDING REMARKS

To the best of our knowledge, we provide, for the first time, statistical properties of the iterates in a variational boosting algorithm. As part of frequentist validation of this variational method, we assume regularity conditions similar to ones widely used in Bayesian contraction literature. Our hypotheses are general enough to include most likelihoods and priors; they do not demand Gaussian like posteriors. Regarding the boosting algorithm itself, we employ the non-compact domain of mixture Gaussian densities as our variational family, much in contrast to the use of compact domain in Locatello et al. [2017]. Convergence analysis of our algorithm shows how smoothness properties of Kullback–Leibler discrepancy lead to exponential dependence of iterate number on the inverse bandwidth, which can be contrasted to faster convergence rates for weaker metrics like Hellinger distance [Campbell and Li, 2019].

6 SUPPLEMENTARY MATERIALS (APPENDIX)

REVISITING FRANK–WOLFE ALGORITHM

We note down the basics of Frank–Wolfe algorithm in this section. The reader is referred to Jaggi [2013] for further details and to Frank and Wolfe [1956] for the original formulation. We start by reviewing the notation. In what follows, Y is an inner product space with $\langle y_1, y_2 \rangle$ denoting the inner product of $y_1, y_2 \in Y$ and $\|y\| = \langle y, y \rangle$ the norm induced by the inner product. D shall denote a compact, convex subset of Y, which is our optimization domain. We shall work with a convex function f defined on D, which is our objective function for the optimization routine. We

---

Algorithm 2 Frank–Wolfe algorithm with approximate Linear Minimization Oracle:

1. Initialize with $x^{(0)} \in D$.

2. For the k-th step, set $\gamma_k = 2/(k+2)$ and calculate sub-gradient $\nabla f(x^{(k)})$.

3. Solve the linear minimization oracle (LMO) approximately, i.e. find $y^{(k+1)} \in D$ such that $\langle \nabla f(x^{(k)}), y^{(k+1)} \rangle \leq \min \left\{ y \in D \mid \langle \nabla f(x^{(k)}), y \rangle \right\} + \gamma_k \mathcal{C}_{f,D}/2$.

4. Update $x^{(k+1)} = (1 - \gamma_k)x^{(k)} + \gamma_k y^{(k+1)}$.

---

start by revising the notion of a subgradient.

Definition 1: A sub-gradient of f at $x \in D$, denoted by $\nabla f(x)$, is a member of $\partial f(x) \subset D$, given by

$$\partial f(x) = \{ y \in D \mid f(z) - f(x) - \langle y, z - x \rangle \geq 0, \ \forall z \in D \}.$$

It is easy to note that, if $Y = \mathbb{R}^d$, and convex function f is differentiable at $x \in Y$, then the gradient at x satisfies $f'(x) \in \partial f(x)$. Subgradients are useful when the notion of differentiability is untenable. Next, we note down the definition of Bregman divergence.

Definition 2: For any $x, y \in D$, the Bregman divergence of y from x under function f is defined as

$$\mathcal{D}_f(y\|x) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle. \tag{18}$$

With this definition in hand, we now define the curvature of f on domain D.

Definition 3: The curvature $\mathcal{C}_{f,D}$ of f on domain D is given by

$$\mathcal{C}_{f,D} = \sup \left\{ \frac{2}{\alpha^2} \mathcal{D}_f(x_2\|x_1) : y, x_1 \in D, \alpha \in [0,1], x_2 = x_1 + \alpha(y - x_1) \right\}. \tag{19}$$

One can think of the curvature as the maximum scaled Bregman divergence between points in D and their perturbations through mixtures. We now recall the definition and significance of strong smoothness and strong convexity of f.

Definition 4: If for any $x, y \in D$ and some $C_1, C_2 > 0$ (possibly depending on f and D)

1. $\mathcal{D}_f(y\|x) \leq C_1 \|y - x\|^2$, then f is strongly smooth on D,

2. $\mathcal{D}_f(y\|x) \geq C_2 \|y - x\|^2$, then f is strongly convex on D.

Convex functions f on D that satisfy strong smoothness allow calculations of rate of convergence. The most basic Frank–Wolfe algorithm minimizes convex function f, defined on domain D. We note down a version with approximately solved subproblem, following Jaggi [2013].

Let x^* denote the minimum point in domain D. The above algorithm, by theorem 1 in Jaggi [2013], satisfies

$$f(x^{(k)}) - f(x^*) \leq 4C_{f,D}/(k+2) \tag{20}$$

This gives us the rate of convergence of this algorithm, in terms of the curvature $\mathcal{C}_{f,D}$. Note that such a rate of convergence with respect of number of iterations k is called sub-linear. For statistical problems, $\mathcal{C}_{f,D}$ is typically a function of the sample size and the parameter dimension, and can be quite large for densities having non-compact support.

PROOF OF THEOREM 1

In what follows, $\gtrsim, \lesssim$ respectively stand for greater than and less than up to an absolute constant and $s_{\max}(A)$ denotes the highest singular value of square matrix A. Let q_0 denote the d-dimensional Gaussian density, centered at the truth θ_0, and variance $\sigma_n^2 I_d$, where σ_n satisfies assumption 2. Along with assumption 1, we have $q_0 \in \mathcal{Q}_n$ and

hence $m_n^*(\mathcal{Q}_n) \leq KL(q_0||\pi_n)$, so that it is enough to show $KL(q_0||\pi_n)$ is bounded in probability. We decompose $KL(q_0||\pi_n)$ as

$$\int q_0(\theta) \log \frac{q_0(\theta)}{\pi_n(\theta)} d\theta = -d\left[\log(\sqrt{2\pi}) + \frac{1}{2}\right] + \log(m(X_n)) - \left(\int L_n(\theta, \theta_0) q_0(\theta) d\theta\right) + \int U(\theta) q_0(\theta) d\theta. \quad (21)$$

Since sum of $O_p(1)$ quantities are again $O_p(1)$, we can stochastically bound the above expression term by term. The first and last terms are already constants. We will also prove $\int U(\theta) q_0(\theta) d\theta \lesssim U(\theta_0)$. So, it is enough to show $\log(m(X_n))$ and $\int L_n(\theta, \theta_0) q_0(\theta) d\theta$ are stochastically bounded from above and below, respectively. We have

$$\mathrm{pr}\left(\log(m(X_n)) > -\log \epsilon\right) = \mathrm{pr}\left((m(X_n)) > \frac{1}{\epsilon}\right) \leq \epsilon . E\left(m(X_n)\right) = \epsilon \quad \forall n. \quad (22)$$

We now employ Taylor expansion around θ_0. Using assumption 4 for $U(\theta)$, observe that

$$\int q_0(\theta) U(\theta) d\theta \lesssim U(\theta_0) + s_{max}^2 \left(U^{(2)}(\theta_0)\right) \left(\int \|\theta - \theta_0\|^2 q_0(\theta) d\theta\right) + \left(\int \|\theta - \theta_0\|_2^{2+2\alpha_3} q_0(\theta) d\theta\right) \lesssim U(\theta_0), \quad (23)$$

and for $\mu_2(\theta_0||\theta)$, observe that

$$\int q_0(\theta) \mu_2(\theta_0||\theta) d\theta \leq s_{max}^2 \left(\mu_2^{(2)}(\theta_0||\theta_0)\right) \left(\int \|\theta - \theta_0\|^2 q_0(\theta) d\theta\right) + \left(\int \|\theta - \theta_0\|_2^{2+2\alpha_2} q_0(\theta) d\theta\right) \leq C_1 \sigma_n^2, \quad (24)$$

for constant $C_1 > 0$. Again, by assumption 4 for $KL(\theta_0||\theta)$, we get the upper bound (similar to previous step)

$$\int q_0(\theta) KL(\theta_0||\theta) d\theta \leq C_u \sigma_n^2, \quad (25)$$

and using assumption 5, we get the lower bound

$$C_l \sigma_n^2 \leq \int q_0(\theta) KL(\theta_0||\theta) d\theta, \quad (26)$$

where $0 < C_l < C_u$ are absolute constants. Now, for $\delta > 0$ to be chosen later, we have

$$\begin{aligned}
&\mathrm{pr}\left[\int L_n(\theta, \theta_0) q_0(\theta) d\theta \leq -C_u(1+\delta) n \sigma_n^2\right] \\
&\leq \mathrm{pr}\left[\int L_n(\theta, \theta_0) q_0(\theta) d\theta \leq -(1+\delta) n \int KL(\theta_0||\theta) q_0(\theta) d\theta\right] \\
&\leq \mathrm{pr}\left[\int \frac{1}{\sqrt{n}} \zeta_n(\theta, \theta_0) q_0(\theta) d\theta \leq -\delta \sqrt{n} \int KL(\theta_0||\theta) q_0(\theta) d\theta\right] \\
&\leq \frac{E\left(\int l(\theta, \theta_0) q_0(\theta) d\theta\right)^2}{\delta^2 n \left(\int KL(\theta_0||\theta) q_0(\theta) d\theta\right)^2} \leq \frac{\int q_0(\theta) \mu_2(\theta_0||\theta) d\theta}{\delta^2 n \left(\int KL(\theta_0||\theta) q_0(\theta) d\theta\right)^2} \leq \frac{C_1 \sigma_n^2}{\delta^2 C_l^2 n \sigma_n^4} = \frac{C_1}{\delta^2 C_l n \sigma_n^2}.
\end{aligned} \quad (27)$$

From assumption 2, we have $c_0^{-1/2} n^{-1/2} \leq \sigma_n \leq n^{-1/2}$ and hence, given $\epsilon > 0$, we can choose $\delta := \left(C_1 c_0^{1/2}/\epsilon C_l\right)^{1/2}$ to get

$$\mathrm{pr}\left[\int L_n(\theta, \theta_0) q_0(\theta) d\theta \leq -C_u \left(1 + \left(C_1 c_0^{1/2}/\epsilon C_l\right)^{1/2}\right)\right] \leq \epsilon, \quad (28)$$

and the stochastic boundedness result is complete.

PROOF OF COROLLARY 1

For this specific family of densities, let $\nu_l = E(T_l|\theta_0)$, $l = 1 \ldots K$ denote the expectations of the sufficient statistics, and let σ_{ℓ_1, l_2} denote $Cov(T_{\ell_1}, T_{l_2}|\theta_0)$ for $\ell_1, l_2 = 1, \ldots K$. Let $\nu_0 = (\nu_1, \ldots \nu_K)$ and $\Sigma_0 = ((\sigma_{\ell_1, l_2}))_{\ell_1, l_2}$, so that for the K-vector $T = (T_1, \ldots T_K)$, $E(T|\theta_0) = \nu_0$ and $Cov(T|\theta_0) = \Sigma_0$. Direct calculations shows

$$KL(\theta_0||\theta) = A(\theta) - A(\theta_0) - (\theta - \theta_0)^T \nu_0, \quad (29)$$

and we know that for Exponential families, $\nu_0 = A^{(1)}(\theta_0)$. Thus (12) becomes (see Definition 2)

$$KL(\theta_0||\theta) = D_A(\theta_0||\theta). \tag{30}$$

It is worthwhile to observe the analogy to lemma 2. $KL(\theta_0||\theta)$ is finite for all $\theta \in \Theta$ as $D_A(\theta_0||\theta)$ is. Now by hypothesis, $A(\theta)$ is strongly convex, and hence by part 2 of Definition 4 and (13), we have $KL(\theta_0||\theta) \geq C||\theta - \theta_0||_2^2$ for some constant $C > 0$. Next, calculation shows

$$\mu_2(\theta_0||\theta) = (\theta - \theta_0)^T \Sigma_0 (\theta - \theta_0) + (D_A(\theta_0||\theta))^2, \tag{31}$$

and we know that for Exponential families, $\Sigma_0 = A^{(2)}(\theta_0)$. Thus (14) becomes

$$\mu_2(\theta_0||\theta) = (\theta - \theta_0)^T A^{(2)}(\theta_0)(\theta - \theta_0) + (D_A(\theta_0||\theta))^2. \tag{32}$$

Since, by hypothesis, $A^{(2)}(\theta)$ exists and $D_A(\theta)$ is finite for all $\theta \in \Theta$, we conclude $\mu_2(\theta_0||\theta)$ is finite for all $\theta \in \Theta$. This verifies assumptions 3 and 5 starting from the hypothesis. Using (13), we see that the first and second derivatives of D_A, with respect to the second argument, satisfy

$$D_A^{(1)}(\theta_0||\theta) = A^{(1)}(\theta) - A^{(1)}(\theta_0), \quad D_A^{(2)}(\theta_0||\theta) = A^{(2)}(\theta). \tag{33}$$

Thus, from (30), (32) and (33), we have

$$KL^{(2)}(\theta_0||\theta) = A^{(2)}(\theta),$$
$$\mu_2^{(2)}(\theta_0||\theta) = 2\left(A^{(2)}(\theta_0) + D_A(\theta_0||\theta)A^{(2)}(\theta)\right) + \left(A^{(1)}(\theta) - A^{(1)}(\theta_0)\right)\left(A^{(1)}(\theta) - A^{(1)}(\theta_0)\right)^T. \tag{34}$$

Since sum of α-Lipschitz functions is again α-Lipschitz, (17) shows why the hypothesis of the corollary suffices to conclude that assumption 4 of theorem 1 holds.

PROOF OF LEMMA 1

$KL(\psi_2||\pi_n) = \int \psi_2 \log \psi_2 - \int \psi_2 \log \pi_n$, $KL(\psi_1||\pi_n) = \int \psi_1 \log \psi_1 - \int \psi_1 \log \pi_n$, so that $KL(\psi_2||\pi_n) - KL(\psi_1||\pi_n) = KL(\psi_2||\psi_1) + \int (\psi_2 - \psi_1) \log \psi_1 - \int (\psi_2 - \psi_1) \log \pi_n = KL(\psi_2||\psi_1) + \int (\psi_2 - \psi_1)(\log \psi_1 - \log \pi_n)$ and we are done.

PROOF OF LEMMA 2

Let $\phi, \phi_1, \ldots \phi_k \in \Gamma_n, \psi_1 = \sum_{j=1}^n \beta_j \phi_j$ for $\boldsymbol{\beta} \in \Delta^k$ and $\psi_2 = \psi_1 + \alpha(\phi - \psi_1)$ for $\alpha \in [0, 1]$. Starting with lemma 1, we have the Taylor expansion

$$\begin{aligned}
\mathcal{D}_n(\psi_2||\psi_1) &= KL(\psi_1 + \alpha(\phi - \psi_1)||\psi_1) \\
&= \alpha \frac{\partial}{\partial \alpha'}\bigg|_{\alpha'=0} KL(\psi_1 + \alpha'(\phi - \psi_1)||\psi_1) + \frac{\alpha^2}{2} \frac{\partial^2}{\partial \alpha'^2}\bigg|_{\alpha'=\beta} KL(\psi_1 + \alpha'(\phi - \psi_1)||\psi_1) \\
&=: \alpha T_1 + \frac{\alpha^2}{2} T_2,
\end{aligned} \tag{35}$$

where β lies in between 0 and α, and T_1, T_2 denote the first and second derivatives respectively. For T_1 we have

$$\begin{aligned}
\frac{\partial}{\partial \alpha} KL(\psi_1 + \alpha(\phi - \psi_1)||\psi_1) &= \frac{\partial}{\partial \alpha}\left[\int (\psi_1 + \alpha(\phi - \psi_1)) \log\left(1 + \alpha\left(\frac{\phi}{\psi_1} - 1\right)\right)\right] \\
&= \int (\phi - \psi_1) \log\left(1 + \alpha\left(\frac{\phi}{\psi_1} - 1\right)\right),
\end{aligned} \tag{36}$$

where in the last step we have used the fact $\int \psi_1 = \int \phi = 1$. This directly shows $T_1 = 0$. Also, for any $\alpha \in [0, 1]$, we have

$$\frac{\partial^2}{\partial \alpha^2} KL(\psi_1 + \alpha(\phi - \psi_1)||\psi_1) = \int \frac{(\phi - \psi_1)^2}{\psi_1 + \alpha(\phi - \psi_1)} \leq \chi^2(\psi_1||\phi) + \chi^2(\phi||\psi_1). \tag{37}$$

11

Now, by Cauchy–Schwartz inequality,

$$\chi^2(\psi_1||\phi) = \int \frac{\left(\sum_{j=1}^k \beta_j(\phi_j - \phi)\right)^2}{\sum_{j=1}^k \beta_j \phi} \le \int \sum_{j=1}^k \beta_j \frac{(\phi_j - \phi)^2}{\phi} = \sum_{j=1}^k \beta_j \chi^2(\phi_j||\phi),$$

$$\chi^2(\phi||\psi_1) = \int \frac{\left(\sum_{j=1}^k \beta_j(\phi - \phi_j)\right)^2}{\sum_{j=1}^k \beta_j \phi_j} \le \int \sum_{j=1}^k \beta_j \frac{(\phi - \phi_j)^2}{\phi_j} = \sum_{j=1}^k \beta_j \chi^2(\phi||\phi_j).$$

(38)

Adding the above equations and then using (37), we have $T_2 \le \sum_{j=1}^k \beta_j \left(\chi^2(\phi||\phi_j) + \chi^2(\phi_j||\phi)\right)$. Now from (35) and statement following (36), we can conclude that $\mathcal{D}_n(\psi_2||\psi_1) \le \frac{\alpha^2}{2} \sum_{j=1}^k \beta_j \left(\chi^2(\phi||\phi_j) + \chi^2(\phi_j||\phi)\right)$. This gives

$$\mathcal{C}_n \le \sup\left\{ \sum_{j=1}^k \beta_j \left(\chi^2(\phi||\phi_j) + \chi^2(\phi_j||\phi)\right) : \phi, \phi_1 \ldots \phi_k \in \Gamma_n, \boldsymbol{\beta} \in \Delta^k \right\},$$

(39)

where this upper bound is a function of only M, c_0 and σ_n, appearing in the definition of Γ_n. We have reduced the upper bound calculation to that of χ^2 divergence of single Gaussians, so we can now apply lemma 3. For any two members of Γ_n, say ϕ_2 and ϕ_1, we get

$$\chi^2(\phi_2||\phi_1) \le (2 - c_0)^{-d/2} \exp\left(\frac{2M^2}{(2 - c_0)\sigma_n}\right).$$

(40)

Identical bound holds if we swap ϕ_1 and ϕ_2. Now this translates to

$$\mathcal{C}_n \le 2(2 - c_0)^{-d/2} \exp\left(\frac{2M^2}{(2 - c_0)\sigma_n}\right).$$

(41)

PROOF OF LEMMA 3

We first define the χ^2 divergence between densities, which is a discrepancy measure comparable to Kullback–Leibler divergence and stronger than it. Refer to **?** for more details.

Definition 5: For densities ϕ_1, ϕ_2 on $\mathbb{R}^d$, the χ^2 divergence of ϕ_2 from ϕ_1 is defined as

$$\chi^2(\phi_2||\phi_1) = -1 + \int \frac{\phi_2^2}{\phi_1}.$$

(42)

Choosing $\phi_i = N\left(\mu_i, \sigma_i^2 \mathbf{I}_d\right), i = 1, 2$, we have

$$\frac{\phi_2^2}{\phi_1}(y) = \left(\sqrt{2\pi}\frac{\sigma_2^2}{\sigma_1}\right)^{-d} \exp\left[\left(-\frac{1}{2}\right)\left(\frac{2\|y - \mu_2\|_2^2}{\sigma_2^2} - \frac{\|y - \mu_1\|_2^2}{\sigma_1^2}\right)\right].$$

(43)

Let the term inside the exponent in (26) be $-B(y)/2$. We re-write

$$B(y) = \left(\frac{2\sigma_1^2 - \sigma_2^2}{\sigma_1^2 \sigma_2^2}\right) \left\| y - \frac{\frac{2\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}}{\frac{2\sigma_1^2 - \sigma_2^2}{\sigma_1^2 \sigma_2^2}} \right\|_2^2 - \frac{2\|\mu_2 - \mu_1\|_2^2}{2\sigma_1^2 - \sigma_2^2},$$

(44)

which shows

$$\int_{y \in \mathbb{R}^d} \left(\sqrt{2\pi}\right)^{-d} \exp\left[\left(-\frac{1}{2}\right)\left(\frac{2\|y - \mu_2\|_2^2}{\sigma_2^2} - \frac{\|y - \mu_1\|_2^2}{\sigma_1^2}\right)\right] dy = \exp\left(\frac{\|\mu_2 - \mu_1\|_2^2}{2\sigma_1^2 - \sigma_2^2}\right) \left(\frac{\sqrt{2\sigma_1^2 - \sigma_2^2}}{\sigma_1 \sigma_2}\right)^{-d}.$$

(45)

Combining (26) and (28), and then using (25), we get lemma 3.

AUXILIARY RESULTS

In this section we provide proofs of the results $(4) - (7)$ in the main article. Since (5) follows directly from Bernstein-von-Mises theorem and its equivalence with (6) follows from

$$d_H^2 \le d_{TV} \le \sqrt{2} d_H, \tag{46}$$

we only focus on

Proposition 1:

$$
KL \left(N \left(\theta_0, n^{-1} \Sigma \right) || N \left(\mu_n, \Sigma_n \right) \right) \rightsquigarrow \frac{1}{2} \chi_d^2,
$$
$$
KL \left(N \left(\overline{X}_n, n^{-1} \Sigma \right) || N \left(\mu_n, \Sigma_n \right) \right) \to 0 \quad \text{a.s.,}
\tag{47}
$$

where '$\rightsquigarrow$' denotes weak convergence and a.s stands for almost sure validity with respect to the data generating distribution.

PROOF OF PROPOSITION 1

We start with noting that

$$\mu_n = \frac{n \overline{X}_n + \mu_0}{n+1}, \quad \Sigma_n^{-1} = n \Sigma^{-1} + \Sigma_0^{-1}, \quad v_{1,n} := \mu_n - \theta_0, \quad w_{1,n} := \mu_n - \overline{X}_n. \tag{48}$$

Observe the difference between $v_{1,n}$ and $w_{1,n}$ by noting

$$
\sqrt{n} v_{1,n} = \frac{\sqrt{n} \left(\overline{X}_n - \theta_0 \right)}{1 + \frac{1}{n}} + \frac{\sqrt{n}}{n+1} \left(\mu_0 - \theta_0 \right) =: v_{2,n} + v_{3,n},
$$
$$
\sqrt{n} w_{1,n} = \frac{\sqrt{n}}{n+1} \left(\mu_0 - \overline{X}_n \right) =: w_{2,n}.
\tag{49}
$$

We now have from (48) and (49)

$$
\begin{aligned}
v_{1,n}^T \Sigma_n^{-1} v_{1,n} &= (v_{2,n} + v_{3,n})^T \left[\Sigma^{-1} + n^{-1} \Sigma_0^{-1} \right] (v_{2,n} + v_{3,n}) \\
&= (v_{2,n} + v_{3,n})^T \Sigma^{-1} (v_{2,n} + v_{3,n}) + n^{-1} (v_{2,n} + v_{3,n})^T \Sigma_0^{-1} (v_{2,n} + v_{3,n}) =: E_{1,n} + n^{-1} E_{2,n}.
\end{aligned}
\tag{50}
$$

Let us deal with $E_{2,n}$ first. Breaking down further, we see

$$E_{2,n} = v_{2,n}^T \Sigma_0^{-1} v_{2,n} + 2 v_{2,n}^T \Sigma_0^{-1} v_{3,n} + v_{3,n}^T \Sigma_0^{-1} v_{3,n}. \tag{51}$$

The last term in the right-hand-side of (51) is non-stochastic, and Σ_0, μ_0 are free of n. Hence, we get

$$v_{3,n}^T \Sigma_0^{-1} v_{3,n} = \frac{n}{(n+1)^2} \left(\mu_0 - \theta_0 \right)^T \Sigma_0^{-1} \left(\mu_0 - \theta_0 \right) \to 0. \tag{52}$$

The second term in the right-hand-side of (51) satisfies

$$v_{2,n}^T \Sigma_0^{-1} v_{3,n} = \left(\frac{n}{n+1} \right)^2 \left[\left(\mu_0 - \theta_0 \right)^T \Sigma_0^{-1} \right] \left(\overline{X}_n - \theta_0 \right) \to 0 \quad a.s., \tag{53}$$

using Strong Law of Large Numbers(SLLN). Now for the first term on the right-hand-side of (51), we have

$$n^{-1} v_{2,n}^T \Sigma_0^{-1} v_{2,n} = \left(\frac{n}{n+1} \right)^2 \left(\overline{X}_n - \theta_0 \right)^T \Sigma_0^{-1} \left(\overline{X}_n - \theta_0 \right) \to 0 \quad a.s., \tag{54}$$

where we have again used SLLN. Putting together (51)–(54), we get that

$$n^{-1} E_{2,n} \to 0 \quad a.s. \tag{55}$$

We now deal with $E_{1,n}$ in the right-hand-side of (50), observing

$$E_{1,n} = v_{2,n}^T \Sigma^{-1} v_{2,n} + 2 v_{2,n}^T \Sigma^{-1} v_{3,n} + v_{3,n}^T \Sigma^{-1} v_{3,n}. \tag{56}$$

Similar to (52)–(53), we have

$$v_{3,n}^T \Sigma^{-1} v_{3,n} \to 0, \quad v_{2,n}^T \Sigma^{-1} v_{3,n} \to 0 \; a.s., \tag{57}$$

while the first term on the right-hand-side of (56) satisfies

$$v_{2,n}^T \Sigma^{-1} v_{2,n} = \left(\frac{n}{n+1}\right)^2 \left\| \Sigma_0^{-1/2} \sqrt{n} \left(\overline{X}_n - \theta_0\right) \right\|_2^2 \rightsquigarrow \chi_d^2. \tag{58}$$

by an application of Slutsky's theorem, as the random vector within norm signs in (58) follows a standard Gaussian distribution in d-dimensions under the true vector θ_0. Another application of Slutsky's theorem allows us to get from (50)

$$v_{1,n}^T \Sigma_n^{-1} v_{1,n} \rightsquigarrow \chi_d^2. \tag{59}$$

We next turn our attention to

$$w_{1,n}^T \Sigma_n^{-1} w_{1,n} = w_{2,n}^T \left[\Sigma^{-1} + n^{-1} \Sigma_0^{-1}\right] w_{2,n} = w_{2,n}^T \Sigma^{-1} w_{2,n} + n^{-1} w_{2,n}^T \Sigma_0^{-1} w_{2,n} := F_{1,n} + n^{-1} F_{2,n}, \tag{60}$$

to get

$$
\begin{aligned}
F_{1,n} &= \frac{n}{(n+1)^2} \left(\overline{X}_n - \mu_0\right)^T \Sigma^{-1} \left(\overline{X}_n - \mu_0\right) \to 0 \quad a.s., \\
F_{2,n} &= \frac{n}{(n+1)^2} \left(\overline{X}_n - \mu_0\right)^T \Sigma_0^{-1} \left(\overline{X}_n - \mu_0\right) \to 0 \quad a.s.
\end{aligned}
\tag{61}
$$

Note that the limit is driven by the preceding factor of $n/(n+1)^2$ rather than SLLN. We thus have

$$w_{1,n}^T \Sigma_n^{-1} w_{1,n} \to 0 \quad a.s. \tag{62}$$

We are now prepared for the final part of the proof where we use the well-known formula for KL divergence between Gaussians to get

$$
\begin{aligned}
KL\left(N\left(\theta_0, n^{-1}\Sigma\right) \| N\left(\mu_n, \Sigma_n\right)\right) &= \frac{1}{2}\left[\operatorname{trace}\left(n^{-1}\Sigma_n^{-1}\Sigma\right) - d - \log\det\left(n^{-1}\Sigma_n^{-1}\Sigma\right) + v_{1,n}^T \Sigma_n^{-1} v_{1,n}\right] \\
KL\left(N\left(\overline{X}_n, n^{-1}\Sigma\right) \| N\left(\mu_n, \Sigma_n\right)\right) &= \frac{1}{2}\left[\operatorname{trace}\left(n^{-1}\Sigma_n^{-1}\Sigma\right) - d - \log\det\left(n^{-1}\Sigma_n^{-1}\Sigma\right) + w_{1,n}^T \Sigma_n^{-1} w_{1,n}\right]
\end{aligned}
\tag{63}
$$

Most of the expressions are common for both the equations in (63), and we evaluate them term by term. We have

$$
\begin{aligned}
\operatorname{trace}\left(n^{-1}\Sigma_n^{-1}\Sigma\right) - d &= \operatorname{trace}\left(I_d + n^{-1}\Sigma_0^{-1}\Sigma\right) - d \to 0, \\
\log\det\left(n^{-1}\Sigma_n^{-1}\Sigma\right) &= \log\det\left(I_d + n^{-1}\Sigma_0^{-1}\Sigma\right) \to 0.
\end{aligned}
\tag{64}
$$

Proposition 1 now follows by combining (59) and (62)–(64), along with an application of Slutsky's theorem.

References

Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *arXiv preprint arXiv:1706.09293*, 2017.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Trevor Campbell and Xinglong Li. Universal boosting variational inference. *arXiv preprint arXiv:1906.01235*, 2019.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3 (1-2):95–110, 1956.

Fangjian Guo, Xiangyu Wang, Kai Fan, Tamara Broderick, and David B Dunson. Boosting variational inference. *arXiv preprint arXiv:1611.05559*, 2016.

Shaobo Han, Xuejun Liao, David Dunson, and Lawrence Carin. Variational gaussian copula inference. In *Artificial Intelligence and Statistics*, pages 829–838, 2016.

Jonathan Huggins, Mikolaj Kasprzak, Trevor Campbell, and Tamara Broderick. Validated variational inference via practical posterior error bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 1792–1802. PMLR, 2020.

Jonathan H Huggins, Trevor Campbell, Mikołaj Kasprzak, and Tamara Broderick. Practical bounds on the error of bayesian posterior approximations: A nonasymptotic approach. *arXiv preprint arXiv:1809.09505*, 2018.

Ferenc Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

Willem Kruijer, Judith Rousseau, Aad Van Der Vaart, et al. Adaptive bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.

Francesco Locatello, Rajiv Khanna, Joydeep Ghosh, and Gunnar Rätsch. Boosting variational inference: an optimization perspective. *arXiv preprint arXiv:1708.01733*, 2017.

Francesco Locatello, Gideon Dresdner, Rajiv Khanna, Isabel Valera, and Gunnar Rätsch. Boosting black box variational inference. In *Advances in Neural Information Processing Systems*, pages 3401–3411, 2018.

Andrew C Miller, Nicholas J Foti, and Ryan P Adams. Variational boosting: Iteratively refining posterior approximations. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2420–2429. JMLR. org, 2017.

Dmitry Molchanov, Valery Kharitonov, Artem Sobolev, and Dmitry Vetrov. Doubly semi-implicit variational inference. *arXiv preprint arXiv:1810.02789*, 2018.

Soumendu Sundar Mukherjee, Purnamrita Sarkar, YX Rachel Wang, and Bowei Yan. Mean field for the stochastic blockmodel: optimization landscape and convergence issues. In *Advances in Neural Information Processing Systems*, pages 10694–10704, 2018.

Debdeep Pati, Anirban Bhattacharya, and Yun Yang. On statistical optimality of variational bayes. In *International Conference on Artificial Intelligence and Statistics*, pages 1579–1588, 2018.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

Weining Shen, Surya T Tokdar, and Subhashis Ghosal. Adaptive bayesian multivariate density estimation with dirichlet mixtures. *Biometrika*, 100(3):623–640, 2013.

Jiaxin Shi, Shengyang Sun, and Jun Zhu. Kernel implicit variational inference. *arXiv preprint arXiv:1705.10119*, 2017.

Michalis K Titsias and Francisco JR Ruiz. Unbiased implicit variational inference. *arXiv preprint arXiv:1808.02078*, 2018.

Dustin Tran, David Blei, and Edo M Airoldi. Copula variational inference. In *Advances in Neural Information Processing Systems*, pages 3564–3572, 2015.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Bo Wang, DM Titterington, et al. Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650, 2006.

Xiangyu Wang. *Boosting variational inference: theory and examples*. PhD thesis, PhD thesis, Duke University, 2016.

Yixin Wang and David M Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, pages 1–15, 2018.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. *arXiv preprint arXiv:1802.05438*, 2018.

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. *arXiv preprint arXiv:1710.03266*, 2017.

Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. *arXiv preprint arXiv:1805.11183*, 2018.

Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *arXiv preprint arXiv:1712.02519*, 2017.