VARIATIONAL CONSENSUS MONTE CARLO

By Maxim Rabinovich, Elaine Angelino, and Michael I. Jordan

UC Berkeley

Practitioners of Bayesian statistics have long depended on Markov chain Monte Carlo (MCMC) to obtain samples from intractable posterior distributions. Unfortunately, MCMC algorithms are typically serial, and do not scale to the large datasets typical of modern machine learning. The recently proposed consensus Monte Carlo algorithm removes this limitation by partitioning the data and drawing samples conditional on each partition in parallel [22]. A fixed aggregation function then combines these samples, yielding approximate posterior samples. We introduce *variational consensus Monte Carlo* (VCMC), a variational Bayes algorithm that optimizes over aggregation functions to obtain samples from a distribution that better approximates the target. The resulting objective contains an intractable entropy term; we therefore derive a relaxation of the objective and show that the relaxed problem is blockwise concave under mild conditions. We illustrate the advantages of our algorithm on three inference tasks from the literature, demonstrating both the superior quality of the posterior approximation and the moderate overhead of the optimization step. Our algorithm achieves a relative error reduction (measured against serial MCMC) of up to 39% compared to consensus Monte Carlo on the task of estimating 300-dimensional probit regression parameter expectations; similarly, it achieves an error reduction of 92% on the task of estimating cluster comembership probabilities in a Gaussian mixture model with 8 components in 8 dimensions. Furthermore, these gains come at moderate cost compared to the runtime of serial MCMC—achieving near-ideal speedup in some instances.

1. Introduction. Modern statistical inference demands scalability to massive datasets and high-dimensional models. Innovation in distributed and stochastic optimization has enabled parameter estimation in this setting, e.g. via stochastic [3] and asynchronous [20] variants of gradient descent. Achieving similar success in Bayesian inference – where the target is a posterior distribution over parameter values, rather than a point estimate – remains computationally challenging.

Two dominant approaches to Bayesian computation are variational Bayes and Markov chain Monte Carlo (MCMC). Within the former, scalable algorithms like stochastic variational inference [11] and streaming variational Bayes [4] have successfully imported ideas from optimization. Within MCMC, adaptive subsampling procedures [2, 14], stochastic gradient Langevin dynamics [25], and Firefly Monte Carlo [16] have applied similar ideas, achieving computational gains by operating only on data subsets. These algorithms are serial, however, and thus cannot take advantage of multicore and multi-machine architectures. This motivates data-parallel MCMC algorithms such as asynchronous variants of Gibbs sampling [1, 8, 12].

Our work belongs to a class of *communication-avoiding* data-parallel MCMC algorithms. These algorithms partition the full dataset $X_{1:N}$ into K disjoint subsets $X_{I_{1:K}}$ where X_{I_k} denotes the data associated with core k. Each core samples from a *subposterior* distribution,

$$p_k\left(\theta_k\right) \propto p\left(X_{I_k} \mid \theta_k\right) p\left(\theta_k\right)^{1/K}, \tag{1}$$

and then a centralized procedure combines the samples into an approximation of the full posterior. Due to their efficiency, such procedures have recently received substantial attention [18, 22, 24].

One of these algorithms, *consensus Monte Carlo* (CMC), requires communication only at the start and end of sampling [22]. CMC proceeds from the intuition that subposterior samples, when aggregated correctly, can approximate full posterior samples. This is formally backed by the factorization

$$p(\theta \mid x_{1:N}) \propto p(\theta) \prod_{k=1}^{K} p(X_{I_k} \mid \theta) = \prod_{k=1}^{K} p_k(\theta). \tag{2}$$

If one can approximate the subposterior densities p_k, using kernel density estimates for instance [18], it is therefore possible to recombine them into an estimate of the full posterior.

Unfortunately, the factorization does not make it immediately clear how to aggregate on the level of *samples* without first having to obtain an estimate of the densities p_k themselves. CMC alters (2) to untie the parameters across partitions and plug in a deterministic link F from the θ_k to θ:

$$p(\theta \mid x_{1:N}) \approx \prod_{k=1}^{K} p_k(\theta_k) \cdot \delta_{\theta = F(\theta_1, \ldots, \theta_K)}. \tag{3}$$

This approximation and an aggregation function motivated by a Gaussian approximation lie at the core of the CMC algorithm [22].

The introduction of CMC raises numerous interesting questions whose answers are essential to its wider application. Two among these stand out as particularly vital. First, how should the aggregation function be chosen to achieve the closest possible approximation to the target posterior? Second, when model parameters exhibit structure or must conform to constraints — if they are, for example, positive semidefinite covariance matrices or labeled centers of clusters — how can the weighted averaging strategy of Scott et al. [22] be modified to account for this structure?

In this paper, we propose *variational consensus Monte Carlo* (VCMC), a novel class of data-parallel MCMC algorithms that allow both questions to be addressed. By formulating the choice of aggregation function as a variational Bayes problem, VCMC makes it possible to adaptively choose the aggregation function to achieve a closer approximation to the true posterior. The flexibility of VCMC likewise supports nonlinear aggregation functions, and we exploit this versatility to introduce structured aggregation functions applicable to inference problems that are not purely vectorial.

An appealing benefit of the VCMC point of view is a clarification of the untying step leading to (3). In VCMC, the approximate factorization corresponds to a variational approximation to the true posterior. This approximation can be viewed as the joint distribution of $(\theta_1, \ldots, \theta_K)$ and θ in an augmented model that assumes conditional independence between the data partitions and posits a deterministic mapping from partition-level parameters to the single global parameter. The added flexibility of this point-of-view makes it possible to move beyond subposteriors and include alternative forms of (3) within the CMC framework. In particular, it is possible to define $p_k(\theta_k) = p(\theta_k)\, p(X_{I_k} \mid \theta_k)$, using partial posteriors in place of subposteriors (cf. [23]). Although extensive investigation of this issue is beyond the scope of this paper, we provide some evidence in Section 6 that partial posteriors are a better choice in some circumstances and demonstrate that VCMC can provide substantial gains in both the partial posterior and subposterior settings.

Before proceeding, we outline the remainder of this paper. Below, in §2, we review CMC and related data-parallel MCMC algorithms. Next, we cast CMC as a variational Bayes problem in §3.

We define the variational optimization objective in §4, addressing the challenging entropy term by relaxing it to a concave lower bound, and give conditions for which this leads to a blockwise concave maximization problem. In §5, we define several classes of aggregation functions, and design novel classes that enable *structured* aggregation of constrained and structured samples—e.g. positive semidefinite matrices and mixture model parameters. In §6, we evaluate the performance of VCMC and CMC relative to serial MCMC. We replicate experiments carried out by Scott et al. [22] and execute more challenging experiments in higher dimensions and with more data. Finally in §7, we summarize our approach and discuss several open problems generated by this work.

2. Related work. We focus on data-parallel MCMC algorithms for large-scale Bayesian posterior sampling. Several recent research threads propose schemes in the setting where the posterior factors as in (2). In general, these parallel strategies are approximate relative to serial procedures, and the specific algorithms differ in terms of the approximations employed and amount of communication required.

At one end of the communication spectrum are algorithms that fit into the MapReduce model [7]. First, K parallel cores sample from K subposteriors, defined in (1), via any Monte Carlo sampling procedure. The subposterior samples are then aggregated to obtain approximate samples from the full posterior. This leads to the challenge of designing proper and efficient aggregation procedures.

Scott et al. [22] propose *consensus Monte Carlo* (CMC), which constructs approximate posterior samples via weighted averages of subposterior samples; our algorithms are motivated by this work. Let $\theta_{k,t}$ denote the t-th subposterior sample from core k. In CMC, the aggregation function averages across each set of K samples $\{\theta_{k,t}\}_{k=1}^{K}$ to produce one approximate posterior sample $\hat{\theta}_t$. Uniform averaging is a natural but naïve heuristic that can in fact be improved upon via a weighted average,

$$\hat{\theta} = F\left(\theta_{1:K}\right) = \sum_{k=1}^{K} W_k \theta_k, \tag{4}$$

where in general, θ_k is a vector and W_k can be a matrix. The authors derive weights motivated by the special case of a Gaussian posterior, where each subposterior is consequently also Gaussian. Let Σ_k be the covariance of the k-th subposterior. This suggests weights $W_k = \Sigma_k^{-1}$ equal to the subposteriors' inverse covariances. CMC treats arbitrary subpostertiors as Gaussians, aggregating with weights given by empirical estimates of $\hat{\Sigma}_k^{-1}$ computed from the observed subposterior samples.

Neiswanger et al. [18] propose aggregation at the level of distributions rather than samples. Here, the idea is to form an approximate posterior via a product of density estimates fit to each subposterior, and then sample from this approximate posterior. The accuracy and computational requirements of this approach depend on the complexity of these density estimates. Wang and Dunson [24] develop alternate data-parallel MCMC methods based on applying a Weierstrass transform to each subposterior. These *Weierstrass sampling* procedures introduce auxiliary variables and additional communication between computational cores.

3. Consensus Monte Carlo as variational inference. Given the distributional form of the CMC framework (3), we would like to choose F so that the induced distribution on θ is as close as possible to the true posterior. This is precisely the problem addressed by variational Bayes, which approximates an intractable posterior $p\left(\theta \mid \mathrm{X}\right)$ by the solution q^* to the constrained optimization problem

$$\min D_{\mathrm{KL}}\left(q \mid\mid p\left(\cdot \mid \mathrm{X}\right)\right) \quad \text{subject to} \quad q \in \mathcal{Q},$$

where $\mathcal{Q}$ is the family of variational approximations to the distribution, usually chosen to make both optimization and evaluation of target expectations tractable. We thus view the aggregation problem in CMC as a variational inference problem, with the variational family given by all distributions $\mathcal{Q} = \mathcal{Q}_{\mathcal{F}} = \{q_F \colon F \in \mathcal{F}\}$, where each F is in some function class $\mathcal{F}$ and defines a density

$$q_F(\theta) = \int_{\Omega^K} \prod_{k=1}^{K} p_k(\theta_k) \cdot \delta_{\theta = F(\theta_1, \ldots, \theta_K)} \, \mathrm{d}\theta_{1:K}.$$

In practice, we parameterize $\mathcal{F}$ by a finite dimensional set of parameters and optimize over it using projected stochastic gradient descent (SGD).

4. The variational optimization problem. Standard optimization of the variational Bayes objective uses the evidence lower bound (ELBO)

$$\begin{aligned} \log p(\mathrm{X}) = \log \mathbb{E}_q \left[\frac{p(\theta,\ \mathrm{X})}{q(\theta)} \right] &\geq \mathbb{E}_q \left[\log \frac{p(\theta,\ \mathrm{X})}{q(\theta)} \right] \\ &= \log p(\mathrm{X}) - D_{\mathrm{KL}}\left(q \,||\, p(\cdot \mid \mathrm{X})\right) =: \mathcal{L}_{\mathrm{VB}}(q) \,. \end{aligned} \tag{5}$$

We can therefore recast the variational optimization problem in an equivalent form as

$$\max \mathcal{L}_{\mathrm{VB}}(q) \quad \text{subject to} \quad q \in \mathcal{Q}.$$

Unfortunately, the variational Bayes objective $\mathcal{L}_{\mathrm{VB}}$ remains difficult to optimize. Indeed, by writing

$$\mathcal{L}_{\mathrm{VB}}(q) = \mathbb{E}_q \left[\log p(\theta,\ \mathrm{X})\right] + \mathrm{H}[q]$$

we see that optimizing $\mathcal{L}_{\mathrm{VB}}$ requires computing an entropy $\mathrm{H}[q]$ and its gradients. We can deal with this issue by deriving a lower bound on the entropy that relaxes the objective further.

Concretely, suppose that every $F \in \mathcal{F}$ can be decomposed as $F(\theta_{1:K}) = \sum_{k=1}^{K} F_k(\theta_k)$, with each F_k a differentiable bijection. Since the θ_k come from subposteriors conditioning on different segments of the data, they are independent. The entropy power inequality [6] therefore implies

$$\begin{aligned} \mathrm{H}[q] &\geq \max_{1 \leq k \leq K} \mathrm{H}\left[F_k(\theta_k)\right] = \max_{1 \leq k \leq K} \left(\mathrm{H}[p_k] + \mathbb{E}_{p_k}\left[\log \det \left[J(F_k)(\theta_k)\right]\right] \right) \\ &\geq \min_{1 \leq k \leq K} \mathrm{H}[p_k] + \max_{1 \leq k \leq K} \mathbb{E}_{p_k}\left[\log \det \left[J(F_k)(\theta_k)\right]\right] \end{aligned} \tag{6}$$

$$\geq \min_{1 \leq k \leq K} \mathrm{H}[p_k] + \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{p_k}\left[\log \det \left[J(F_k)(\theta_k)\right]\right] =: \tilde{\mathrm{H}}[q] \,, \tag{7}$$

where $J(f)(\theta)$ denotes the Jacobian of the function f evaluated at θ. The proof of this inequality can be found in the supplement.

This approach gives an explicit, easily computed approximation to the entropy—and this approximation is a lower bound, allowing us to interpret it simply as a further relaxation of the original inference problem. Furthermore, and crucially, it decouples p_k and F_k, thereby making it possible to optimize over F_k without estimating the entropy of any p_k. We note additionally that if we are willing to sacrifice concavity, we can use the tighter lower bound on the entropy given by (6).

Putting everything together, we can define our relaxed variational objective as

$$\mathcal{L}(q) = \mathbb{E}_q \left[\log p(\theta,\ \mathrm{X})\right] + \tilde{\mathrm{H}}[q] \,. \tag{8}$$

Maximizing this function is the variational Bayes problem we consider in the remainder of the paper.

Conditions for concavity. Under certain conditions, the problem posed above is blockwise concave. To see when this holds, we use the language of graphical models and exponential families. To derive the result in the greatest possible generality, we decompose the variational objective as

$$\mathcal{L}_{\mathrm{VB}} = \mathbb{E}_q\left[\log p\left(\theta,\ \mathrm{X}\right)\right] + \mathrm{H}\left[q\right] \geq \tilde{\mathcal{L}} + \tilde{\mathrm{H}}\left[q\right]$$

and prove concavity directly for $\tilde{\mathcal{L}}$, then treat our choice of relaxed entropy (7). We emphasize that while the entropy relaxation is only defined for decomposed aggregation functions, concavity of the partial objective holds for arbitrary aggregation functions. All proofs are in the supplement.

Suppose the model distribution is specified via a graphical model G, so that $\theta = \left(\theta^u\right)_{u \in V(G)}$, such that each conditional distribution is defined by an exponential family

$$\log p\left(\theta^u \mid \theta^{\mathrm{par}(u)}\right) = \log h^u\left(\theta^u\right) + \sum_{u' \in \mathrm{par}(u)} \left(\theta^{u'}\right)^T T^{u' \to u}\left(\theta^u\right) - \log A^u\left(\theta^{\mathrm{par}(u)}\right).$$

If each of these log conditional density functions is log-concave in θ^u, we can guarantee that the log likelihood is concave in each θ^u individually.

Theorem 4.1 (Blockwise concavity of the variational cross-entropy). *Suppose that the model distribution is specified by a graphical model G in which each conditional probability density is a log-concave exponential family. Suppose further that the variational aggregation function family satisfies $\mathcal{F} = \prod_{u \in V(G)} \mathcal{F}^u$ such that we can decompose each aggregation function across nodes via*

$$F\left(\theta\right) = \left(F^u\left(\theta^u\right)\right)_{u \in V(G)}, \quad F \in \mathcal{F} \quad \textit{and} \quad F^u \in \mathcal{F}^u.$$

If each $\mathcal{F}^u$ is a convex subset of some vector space $\mathcal{H}^u$, then the variational cross-entropy $\tilde{\mathcal{L}}$ is concave in each F^u individually.

Assuming that the aggregation function can be decomposed into a sum over functions of individual subposterior terms we can also prove concavity of our entropy relaxation (7).

Theorem 4.2 (Concavity of the relaxed entropy). *Suppose $\mathcal{F} = \prod_{k=1}^K \mathcal{F}_k$, with each function $F \in \mathcal{F}$ decomposing as $F\left(\theta_1, \ldots, \theta_K\right) = \sum_{k=1}^K F_k\left(\theta_k\right)$ for unique bijective $F_k \in \mathcal{F}_k$. Then the relaxed entropy (7) is concave in F.*

As a result, we derive concavity of the variational objective in a broad range of settings.

Corollary 4.1 (Concavity of the variational objective). *Under the hypotheses of Theorems 4.1 and 4.2, the variational Bayes objective $\mathcal{L} = \tilde{\mathcal{L}} + \tilde{\mathrm{H}}$ is concave in each F^u individually.*

5. Variational aggregation function families. The performance of our algorithm depends critically on the choice of aggregation function family $\mathcal{F}$. The family must be sufficiently simple to support efficient optimization, expressive to capture the complex transformation from the set of subposteriors to the full posterior, and structured to preserve structure in the parameters. We now illustrate some aggregation functions that meet these criteria.

5.1. *Vector aggregation.* In the simplest case, $\theta \in \mathbb{R}^d$ is an unconstrained vector. Then, a linear aggregation function $F_W = \sum_{k=1}^K W_k \theta_k$ makes sense, and it is natural to impose constraints to make this sum behave like a weighted average—i.e., each $W_k \in \mathcal{S}_+^d$ is a positive semidefinite (PSD) matrix and $\sum_{k=1}^K W_k = I_d$. For computational reasons, it is often desirable to restrict to diagonal W_k.

5.2. *Spectral aggregation.* Cases involving structure exhibit more interesting behavior. Indeed, if our parameter is a PSD matrix $\Lambda \in \mathcal{S}_+^d$, applying the vector aggregation function above to the flattened vector form $\mathrm{vec}\left(\Lambda\right)$ of the parameter does not suffice. Denoting elementwise matrix product as $\circ$, we note that this strategy would in general lead to $F_W\left(\Lambda_{1:m}\right) = \sum_{k=1}^K W_k \circ \Lambda_k \notin \mathcal{S}_+^d$.

We therefore introduce a more sophisticated aggregation function that preserves PSD structure. For this, given symmetric $A \in \mathbb{R}^{d\times d}$, define $R\left(A\right)$ and $D\left(A\right)$ to be orthogonal and diagonal matrices, respectively, such that $A = R\left(A\right)^T D\left(A\right) R\left(A\right)$. Impose further—and crucially—the canonical ordering $D\left(A\right)_{11} \geq \cdots \geq D\left(A\right)_{dd}$. We can then define our spectral aggregation function by

$$F_W^{\mathrm{spec}}\left(\Lambda_{1:K}\right) = \sum_{k=1}^K R\left(\Lambda_k\right)^T \left[W_k D\left(\Lambda_k\right)\right] R\left(\Lambda_k\right).$$

Assuming $W_k \in \mathcal{S}_+^d$, the output of this function is guaranteed to be PSD, as required. As above we restrict the set of W_k to the matrix simplex $\{\left(W_k\right)_{k=1}^K : W_k \in \mathcal{S}_+^d,\ \sum_{k=1}^K W_k = I\}$.

5.3. *Combinatorial aggregation.* Additional complexity arises with unidentifiable latent variables and, more generally, models with multimodal posteriors. Since this class encompasses many popular algorithms in machine learning, including factor analysis, mixtures of Gaussians and multinomials, and latent Dirichlet allocation (LDA), we now show how our framework can accommodate them.

For concreteness, suppose now that our model parameters are given by $\theta \in \mathbb{R}^{L\times d}$, where L denotes the number of global latent variables (e.g. cluster centers). We introduce discrete alignment parameters a_k that indicate how latent variables associated with partitions map to global latent variables. Each a_k is thus a one-to-one correspondence $[L] \to [L]$, with $a_{k\ell}$ denoting the index on worker core k of cluster center ℓ. For fixed a, we then obtain the variational aggregation function

$$F_a\left(\theta_{1:K}\right) = \left(\sum_{k=1}^K W_{k\ell}\theta_{k a_{k\ell}(\ell)}\right)_{\ell=1}^L.$$

Optimization can then proceed in an alternating manner, switching between the alignments a_k and the weights W_k, or in a greedy manner, fixing the alignments at the start and optimizing the weight matrices. In practice, we do the latter, aligning using a simple heuristic objective $\mathcal{O}\left(a\right) = \sum_{k=2}^K \sum_{\ell=1}^L \left\|\bar{\theta}_{k a_{k\ell}} - \bar{\theta}_{1\ell}\right\|_2^2$, where $\bar{\theta}_{k\ell}$ denotes the mean value of cluster center ℓ on partition k. As $\mathcal{O}$ suggests, we set $a_{1\ell} = \ell$. This is permitted because the global order is only determined up to a permutation. We find that minimizing $\mathcal{O}$ via the Hungarian algorithm [15] leads to good alignments.

6. Empirical evaluation. We now evaluate VCMC on three inference problems, in a range of data and dimensionality conditions. In the vector parameter case, we compare directly to the simple weighting baselines corresponding to previous work on CMC [22]; in the other cases, we compare to structured analogues of these weighting schemes. Our experiments demonstrate the advantages of VCMC across the whole range of model dimensionality, data quantity, and availability of parallel resources.

Baseline weight settings. Scott et al. [22] studied linear aggregation functions with fixed weights,

$$W_k^{\mathrm{unif}} = \frac{1}{K}\cdot I_d \qquad \text{and} \qquad W_k^{\mathrm{gauss}} \propto \mathrm{diag}\left(\hat{\Sigma}_k\right)^{-1}, \tag{9}$$

corresponding to uniform averaging and Gaussian averaging, respectively, where $\hat{\Sigma}_k$ denotes the standard empirical estimate of the covariance. These are our baselines for comparison.

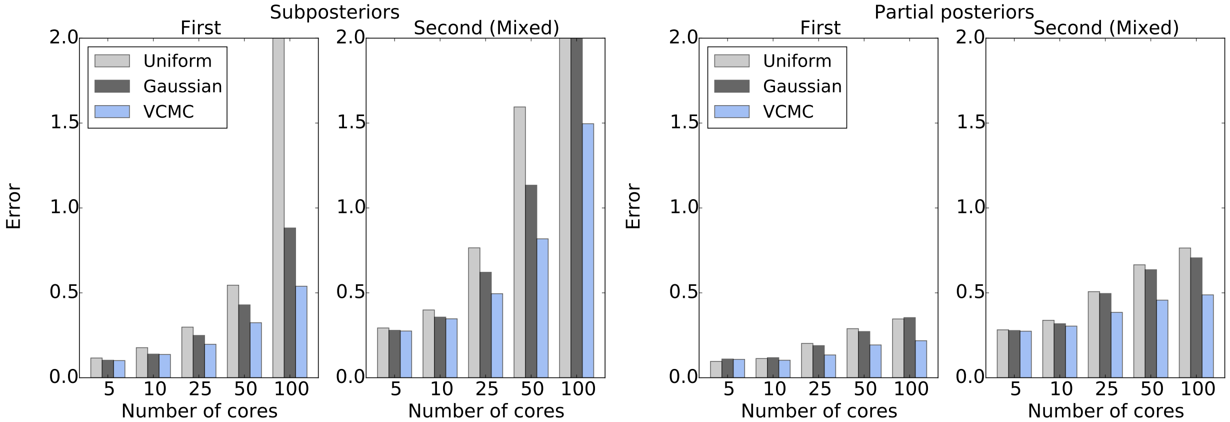

Fig 1: High-dimensional probit regression ($d = 300$). Moment approximation error for the uniform and Gaussian averaging baselines and VCMC, relative to serial MCMC, for *(left)* subposteriors and *(right)* partial posteriors. We assessed three groups of functions: first moments, with $f(\beta) = \beta_j$ for $1 \leq j \leq d$; pure second moments, with $f(\beta) = \beta_j^2$ for $1 \leq j \leq d$; and mixed second moments, with $f(\beta) = \beta_i \beta_j$ for $1 \leq i < j \leq d$. For brevity, results for pure second moments are relegated to Figure 6 in the supplement. Relative errors are truncated to 2 in all cases.

Evaluation metrics. Since the goal of MCMC is usually to estimate event probabilities and function expectations, we evaluate algorithm accuracy for such estimates, relative to serial MCMC output. For each model, we consider a suite of test functions $f \in \mathcal{F}$ (e.g. low degree polynomials, cluster comembership indicators), and we assess the error of each algorithm $\mathcal{A}$ using the metric

$$\epsilon_{\mathcal{A}}(f) = \frac{|\mathbb{E}_{\mathcal{A}}[f] - \mathbb{E}_{\mathrm{MCMC}}[f]|}{|\mathbb{E}_{\mathrm{MCMC}}[f]|}.$$

In the body of the paper, we report median values of $\epsilon_{\mathcal{A}}$, computed within each test function class. The supplement expands on this further, showing quartiles for the differences in ϵ_{VCMC} and ϵ_{CMC}.

6.1. *Bayesian probit regression.* We consider the nonconjugate probit regression model. In this case, we use linear aggregation functions as our function class. For computational efficiency, we also limit ourselves to diagonal W_k. We use Gibbs sampling on the following augmented model:

$$\begin{aligned} \beta &\sim \mathcal{N}(0,\ \sigma^2 I_d) \\ Z_n \mid \beta,\ x_n &\sim \mathcal{N}(\beta^T x_n,\ 1) \\ Y_n \mid Z_n,\ \beta,\ x_n &= \begin{cases} 1 \text{ if } Z_n > 0, \\ 0 \text{ otherwise.} \end{cases} \end{aligned}$$

This augmentation allows us to implement an efficient and rapidly mixing Gibbs sampler, where

$$\begin{aligned} \beta \mid x_{1:N} &= \mathrm{X} \\ z_{1:N} &= z \sim \mathcal{N}\left(\Sigma \mathrm{X}^T z,\ \Sigma\right) \\ \Sigma &= \left(\sigma^{-2} I_d + \mathrm{X}^T \mathrm{X}\right)^{-1}. \end{aligned}$$

We run two experiments: the first using a data generating distribution from Scott et al. [22], with $N = 8500$ data points and $d = 5$ dimensions, and the second using $N = 10^5$ data points and $d = 300$ dimensions. As shown in Figure 1 and, in the supplement,[1] Figures 5 and 6, VCMC decreases the

[1] Due to space constraints, we relegate results for $d = 5$ to the supplement.

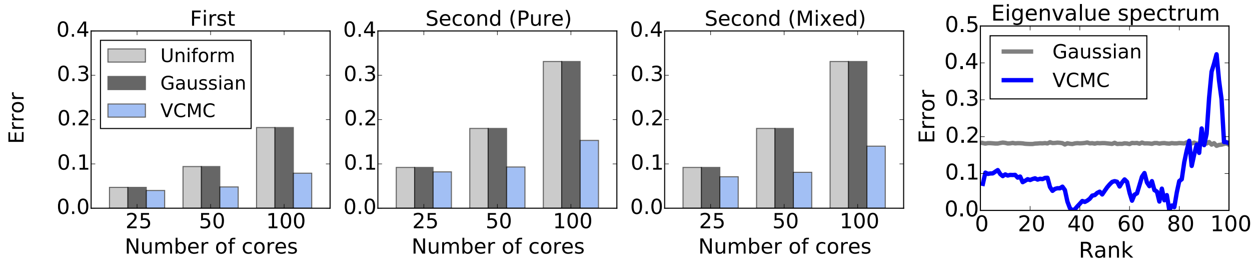

Fig 2: High-dimensional normal-inverse Wishart model ($d = 100$). *(Far left, left, right)* Moment approximation error for the uniform and Gaussian averaging baselines and VCMC, relative to serial MCMC. Letting ρ_j denote the j^{th} largest eigenvalue of Λ^{-1}, we assessed three groups of functions: first moments, with $f(\Lambda) = \rho_j$ for $1 \leq j \leq d$; pure second moments, with $f(\Lambda) = \rho_j^2$ for $1 \leq j \leq d$; and mixed second moments, with $f(\Lambda) = \rho_i \rho_j$ for $1 \leq i < j \leq d$. *(Far right)* Graph of error in estimating $\mathbb{E}[\rho_j]$ as a function of j (where $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_d$).

error of moment estimation compared to the baselines, with substantial gains starting at $K = 25$ partitions (and increasing with K). We also run the high-dimensional experiment using partial posteriors [23] in place of subposteriors, and observe substantially lower errors in this case.

6.2. *Normal-inverse Wishart model.* To compare directly to prior work [22], we consider the normal-inverse Wishart model

$$\Lambda \sim \text{Wishart}(\nu,\ V)$$
$$X_n \mid \mu,\ \Lambda \sim \mathcal{N}\left(\mu,\ \Lambda^{-1}\right).$$

Here, we use spectral aggregation rules as our function class, restricting to diagonal W_k for computational efficiency. We run two sets of experiments: one using the covariance matrix from Scott et al. [22], with $N = 5000$ data points and $d = 5$ dimensions, and one using a higher-dimensional covariance matrix designed to have a small spectral gap and a range of eigenvalues, with $N = 10^5$ data points and $d = 100$ dimensions. In both cases, we use a form of projected SGD, using 40 samples per iteration to estimate the variational gradients and running 25 iterations of optimization. We note that because the mean μ is treated as a point-estimated parameter, one could sample Λ exactly using normal-inverse Wishart conjugacy [10]. As Figure 2 shows,[2] VCMC improves both first and second posterior moment estimation as compared to the baselines. Here, the greatest gains from VCMC appear at large numbers of partitions ($K = 50, 100$). We also note that uniform and Gaussian averaging perform similarly because the variances do not differ much across partitions.

6.3. *Mixture of Gaussians.* A substantial portion of Bayesian inference focuses on latent variable models and, in particular, mixture models. We therefore evaluate VCMC on the mixture of Gaussians model defined by

$$\theta_{1:L} \sim \mathcal{N}\left(0,\ \tau^2 I_d\right)$$
$$Z_n \sim \text{Cat}(\pi)$$
$$X_n \mid Z_n = z \sim \mathcal{N}\left(\theta_z,\ \sigma^2 I_d\right),$$

where the mixture weights π and the prior and likelihood variances τ^2 and σ^2 are assumed known. We use the combinatorial aggregation functions defined in Section 5; we set $L = 8$, $\tau = 2$, $\sigma = 1$,

[2] Due to space constraints, we compare to the $d = 5$ experiment of Scott et al. [22] in the supplement.

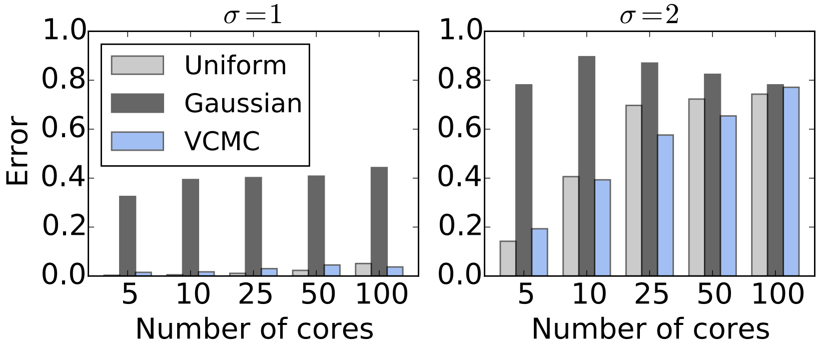

Fig 3: Mixture of Gaussians ($d = 8, L = 8$). Expectation approximation error for the uniform and Gaussian baselines and VCMC. We report the median error, relative to serial MCMC, for cluster comembership probabilities of pairs of test data points, for *(left)* $\sigma = 1$ and *(right)* $\sigma = 2$, where we run the VCMC optimization procedure for 50 and 200 iterations, respectively. When $\sigma = 2$, some comembership probabilities are estimated poorly by all methods; we therefore only use the 70% of comembership probabilities with the smallest errors across all the methods.

and π uniform and generate $N = 5 \times 10^4$ data points in $d = 8$ dimensions, using the model from Nishihara et al. [19]. The resulting inference problem is therefore $L \times d = 64$-dimensional. All samples were drawn using the PyStan implementation of Hamiltonian Monte Carlo (HMC).

As Figure 3 shows, VCMC drastically improves moment estimation compared to the baseline Gaussian averaging (9). To assess how VCMC influences estimates in cluster membership probabilities, we generated 100 new test points from the model and analyzed cluster comembership probabilities for all pairs in the test set. Concretely, for each x_i and x_j in the test data, we estimated $\mathbb{P}\left[x_i \text{ and } x_j \text{ belong to the same cluster}\right]$. Figure 3 shows the resulting boost in accuracy: when $\sigma = 1$, VCMC delivers estimates close to those of serial MCMC, across all numbers of partitions; the errors are larger for $\sigma = 2$. Unlike previous models, uniform averaging here outperforms Gaussian averaging, and indeed is competitive with VCMC.

6.4. *Assessing computational efficiency.* The efficiency of VCMC depends on that of the optimization step, which depends on factors including the step size schedule, number of samples used per iteration to estimate gradients, and size of data minibatches used per iteration. Extensively assessing the influence of all these factors is beyond the scope of this paper, and is an active area of research both in general and specifically in the context of variational inference [13, 17, 21]. Here, we provide an initial assessment of the computational efficiency of VCMC, taking the probit regression and Gaussian mixture models as our examples, using step sizes and sample numbers from above, and eschewing minibatching on data points.

Figure 4 shows timing results for both models. For the probit regression, while the optimization cost is not negligible, it is significantly smaller than that of serial sampling, which takes over 6000 seconds to produce 1000 effective samples.[3] Across most numbers of partitions, approximately 25 iterations—corresponding to less than 1500 seconds of wall clock time—suffices to give errors close to those at convergence. For the mixture, on the other hand, the computational cost of optimization is minimal compared to serial sampling. We can see this in the overall speedup of VCMC relative to serial MCMC: for sampling and optimization combined, low numbers of partitions ($K \leq 25$) achieve

[3] We ran the sampler for 5100 iterations, including 100 burnin steps, and kept every fifth sample.

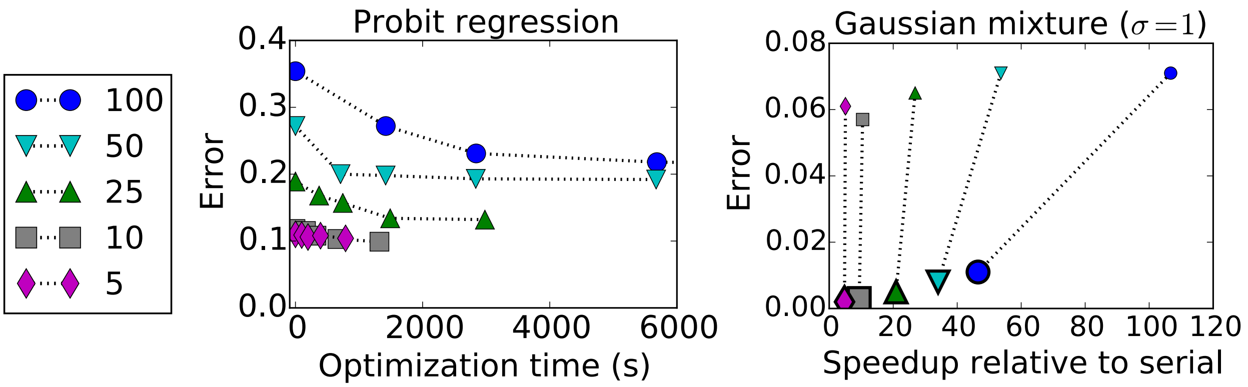

Fig 4: Error versus timing and speedup measurements. *(Left)* VCMC error as a function of number of seconds of optimization. The cost of optimization is nonnegligible, but still moderate compared to serial MCMC—particularly since our optimization scheme only needs small batches of samples and can therefore operate concurrently with the sampler. *(Right)* Error versus speedup relative to serial MCMC, for both CMC with Gaussian averaging (small markers) and VCMC (large markers).

speedups close to the ideal value of K, and large numbers ($K = 50,\ 100$) still achieve good speedups of about $K/2$. The cost of the VCMC optimization step is thus moderate—and, when the MCMC step is expensive, small enough to preserve the linear speedup of embarrassingly parallel sampling. Moreover, since the serial bottleneck is an optimization, we are optimistic that performance, both in terms of number of iterations and wall clock time, can be significantly increased by using techniques like data minibatching [9], adaptive step sizes [21], or asynchronous updates [20].

7. Conclusion and future work. The flexibility of variational consensus Monte Carlo (VCMC) opens several avenues for further research. Following previous work on data-parallel MCMC, we used the subposterior factorization. Our variational framework can accomodate more general factorizations that might be more statistically or computationally efficient – e.g. the factorization used by Broderick et al. [4]. We also introduced structured sample aggregation, and analyzed some concrete instantiations. Complex latent variable models would require more sophisticated aggregation functions – e.g. ones that account for symmetries in the model [5] or lift the parameter to a higher dimensional space before aggregating. Finally, recall that our algorithm – again following previous work – aggregates in a sample-by-sample manner, cf. (4). Other aggregation paradigms may be useful in building approximations to multimodal posteriors or in boosting the statistical efficiency of the overall sampler.

Acknowledgments. We thank R.P. Adams, N. Altieri, T. Broderick, R. Giordano, M.J. Johnson, and S.L. Scott for for helpful discussions. Support for this project was provided by ONR under the Multidisciplinary University Research Initiative (MURI) program (N00014-11-1-0688). E.A. is supported by the Miller Institute for Basic Research in Science, University of California, Berkeley. M.R. is supported by a Fannie and John Hertz Foundation Fellowship and an NSF Graduate Research Fellowship. This research is supported in part by NSF CISE Expeditions Award CCF-1139158, LBNL Award 7076018, and DARPA XData Award FA8750-12-2-0331, and gifts from Amazon Web Services, Google, SAP, The Thomas and Stacey Siebel Foundation, Adatao, Adobe, Apple, Inc., Blue Goji, Bosch, C3Energy, Cisco, Cray, Cloudera, EMC2, Ericsson, Facebook, Guavus, HP, Huawei, Informatica, Intel, Microsoft, NetApp, Pivotal, Samsung, Schlumberger, Splunk, Virdata and VMware.

References.

[1] A. U. Asuncion, P. Smyth, and M. Welling. Asynchronous distributed learning of topic models. In *Advances in Neural Information Processing Systems 21*, NIPS '08, pages 81–88, 2008.

[2] R. Bardenet, A. Doucet, and C. Holmes. Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. In *Proceedings of the 31st International Conference on Machine Learning*, ICML '14, 2014.

[3] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 2nd edition, 1990.

[4] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan. Streaming variational Bayes. In *Advances in Neural Information Processing Systems 26*, NIPS '13, pages 1727–1735, 2013.

[5] T. Campbell and J. P. How. Approximate decentralized Bayesian inference. In *30th Conference on Uncertainty in Artificial Intelligence*, UAI '14, 2014.

[6] T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing).* Wiley-Interscience, 2006.

[7] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, Jan. 2008.

[8] F. Doshi-Velez, D. A. Knowles, S. Mohamed, and Z. Ghahramani. Large scale nonparametric Bayesian inference: Data parallelisation in the Indian buffet process. In *Advances in Neural Information Processing Systems 22*, NIPS '09, pages 1294–1302, 2009.

[9] J. C. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[10] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis, Third Edition.* Chapman and Hall/CRC, 2013.

[11] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, May 2013.

[12] M. Johnson, J. Saunderson, and A. Willsky. Analyzing Hogwild parallel Gaussian Gibbs sampling. In *Advances in Neural Information Processing Systems 26*, pages 2715–2723, 2013.

[13] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, NIPS '13, pages 315–323, 2013.

[14] A. Korattikara, Y. Chen, and M. Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *Proceedings of the 31st International Conference on Machine Learning*, ICML '14, 2014.

[15] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2): 83–97, 1955.

[16] D. Maclaurin and R. P. Adams. Firefly Monte Carlo: Exact MCMC with subsets of data. In *Proceedings of 30th Conference on Uncertainty in Artificial Intelligence*, UAI '14, 2014.

[17] S. Mandt and D. M. Blei. Smoothed gradients for stochastic variational inference. In *Advances in Neural Information Processing Systems 27*, NIPS '14, pages 2438–2446, 2014.

[18] W. Neiswanger, C. Wang, and E. Xing. Asymptotically exact, embarrassingly parallel MCMC. In *30th Conference on Uncertainty in Artificial Intelligence*, UAI '14, 2014.

[19] R. Nishihara, I. Murray, and R. P. Adams. Parallel MCMC with generalized elliptical slice sampling. *Journal of Machine Learning Research*, 15:2087–2112, 2014.

[20] F. Niu, B. Recht, C. Ré, and S. Wright. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems 24*, NIPS '11, pages 693–701, 2011.

[21] R. Ranganath, C. Wang, D. M. Blei, and E. P. Xing. An adaptive learning rate for stochastic variational inference. In *Proceedings of the 30th International Conference on Machine Learning*, ICML '13, pages 298–306, 2013.

[22] S. L. Scott, A. W. Blocker, and F. V. Bonassi. Bayes and big data: The consensus Monte Carlo algorithm. In *Bayes 250*, 2013.

[23] H. Strathmann, D. Sejdinovic, and M. Girolami. Unbiased Bayes for big data: Paths of partial posteriors. *arXiv preprint arXiv:1501.03326*, 2015.

[24] X. Wang and D. B. Dunson. Parallel MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*, 2013.

[25] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, ICML '11, 2011.

APPENDIX A: PROOFS

Proof of entropy relaxation. We apply the entropy power inequality [6], which asserts that for independent d-dimensional random vectors $\psi_{1:K}$, the sum

$$\psi = \sum_{k=1}^{K} \psi_k$$

satisfies

$$e^{\frac{2h(\psi)}{d}} \geq \sum_{k=1}^{K} e^{\frac{2h\left(\psi_k\right)}{d}} \geq \max_{1\leq k\leq K} e^{\frac{2h\left(\psi_k\right)}{d}}, \tag{10}$$

where h denotes differential entropy.

In our case, we have

$$\psi_k = F_k\left(\theta_k\right)$$

and

$$\psi = \theta = F\left(\theta_1, \ldots, \theta_K\right)$$

Since

$$\mathrm{H}\left[q\right] = h\left(\psi\right),$$

equation (10) implies

$$\mathrm{H}\left[q\right] \geq \max_{1\leq k\leq K} h\left(\psi_k\right) = \max_{1\leq k\leq K} \left(\mathrm{H}\left[p_k\right] + \mathbb{E}_{p_k}\left[\log\det J\left(F_k\right)\left(\theta_k\right)\right]\right).$$

Defining

$$\tilde{\mathrm{H}}\left[q\right] = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{p_k}\left[\log\det J\left(F_k\right)\left(\theta_k\right)\right] + \min_{1\leq k\leq K} \mathrm{H}\left[p_k\right],$$

we immediately see that

$$\mathrm{H}\left[q\right] \geq \tilde{\mathrm{H}}\left[q\right],$$

as required. □

Proof of Theorem 4.1. We first define

$$\mathcal{L}_0\left(q\right) = \mathbb{E}_q\left[\log p\left(\theta,\ X\right)\ \mid \theta_{1:K}\right] = \log p\left(F\left(\theta_{1:K}\right),\ X\right).$$

Since $\mathcal{L}\left(q\right) = \mathbb{E}_{p_{1:K}}\left[\mathcal{L}_0\left(q\right)\right]$, where the expectation is taken with respect to the subposteriors, which do not vary with q, it suffices to show that $\mathcal{L}_0$ is concave in each F^u individually for each fixed $\theta_{1:K}$. Furthermore, since $F\left(\theta_{1:K}\right)$ is linear in F by the definition of function addition, it actually suffices to show $\ell\left(\theta\right) = \log p\left(F\left(\theta_{1:K}\right),\ X\right)$ in each θ^u individually. To see why this holds, first observe that for each $u \in V(G)$, we have

$$\ell\left(\theta\right) = \log h^u\left(\theta^u\right) + \sum_{u'\in\mathrm{par}(u)} \left(\theta^{u'}\right)^T T^{u'\to u}\left(\theta^{u'}\right) \tag{11}$$

$$+ \sum_{v\in\mathrm{ch}(u)} \left[\left(\theta^u\right)^T T^{u\to v}\left(\theta^v\right) - \log A^v\left(\theta^{\mathrm{par}(v)}\right)\right] + c_u, \tag{12}$$

where c_u is a function of θ that is constant in θ^u. By the log-concavity assumption, the sum of the first two terms of $\ell(\theta)$ in (12) is concave in θ^u. On the other hand, by basic properties of exponential families, each $\log A^v\left(\theta^{\mathrm{par}(v)}\right)$ is convex in $\theta^{\mathrm{par}(v)}$ and hence in θ^u, making its negative concave. Since the remaining terms are linear or constant, ℓ is in fact concave in θ^u. The claim follows. □

Proof of Theorem 4.2. Clearly it suffices to show that each $\mathbb{E}_{p_k}\left[\log\det J\left(F_k\right)\left(\theta_k\right)\right]$ is concave and for this it suffices to show that for fixed θ_k, $\log\det J\left(F_k\right)\left(\theta_k\right)$ is concave. This is immediate, however, since the Jacobian is a linear function and log det is a concave function. □

APPENDIX B: VARIATIONAL OBJECTIVE FUNCTIONS

We derive the variational objectives and gradients for the models we analyze. Throughout, we make the convention that for $A,\ B \in \mathbb{R}^{d\times d}$,

$$\langle\langle A,\ B\rangle\rangle = \mathrm{Tr}\left(AB\right)$$

denotes the trace inner product.

B.1. Bayesian probit regression. In this section, we compute the variational objective for the Bayesian probit regression model. For convenience, we define

$$\mu_k = \mathbb{E}_{p_k}\left[\beta_k\right] \quad\text{and}\quad S_k = \mathbb{E}_{p_k}\left[\beta_k\beta_k^T\right].$$

In this notation, the variational objective takes the simple form

$$\begin{aligned}\mathcal{L}(W) = -\frac{1}{2\sigma^2}\sum_{k=1}^{K}&\left[\langle\langle S_k,\ W_k^T W_k\rangle\rangle + 2\sum_{\ell\neq k}\langle\langle \mu_k\mu_\ell^T W_\ell^T,\ W_k\rangle\rangle\right]\\ &+\sum_{n=1}^{N}\left[y_n\cdot\mathbb{E}_q\left[\log\Phi_n\right] + \left(1-y_n\right)\cdot\mathbb{E}_q\left[\log\left(1-\Phi_n\right)\right]\right]\\ &+\frac{1}{K}\sum_{k=1}^{K}\log\det\left(W_k\right)\end{aligned}$$

where $\Phi_n = \Phi\left(\sum_k\langle\langle W_k,\ \beta_k x_n^T\rangle\rangle\right)$.

This leads to the gradients

$$\begin{aligned}\nabla_{W_k}\mathcal{L} = \frac{1}{\sigma^2}&\left[S_k W_k^T + \sum_{\ell\neq k}\left(\mu_k\mu_\ell^T W_\ell^T + W_\ell\mu_\ell\mu_k^T\right)\right]\\ &+\sum_{n=1}^{N}\mathbb{E}_q\left[\left(\frac{\phi_n}{\Phi_n\left(1-\Phi_n\right)}\cdot\left(y_n-\Phi_n\right)\right)\cdot\beta_k\right]x_n^T\\ &+\frac{W_k^{-1}}{K},\end{aligned}$$

where we have additionally defined $\phi_n = \phi\left(\sum_{k=1}^{K}\langle\langle W_k,\ \beta_k x_n^T\rangle\rangle\right)$ and

$$\beta = \sum_{k=1}^{K} W_k\beta_k.$$

B.2. Normal-inverse Wishart model. The variational objective for the normal-inverse Wishart model takes the form

$$\mathcal{L}(W) = \mathbb{E}_q\left[\mathcal{L}_0\left(W,\ \Lambda_{1:K}\right)\right] + \tilde{\mathrm{H}}[q],$$

where

$$\begin{aligned}\mathcal{L}_0(W) = &-\frac{1}{2}\sum_{k=1}^{K}\left\langle\left\langle R_k\left(V^{-1}+X^TX\right)R_k^T,\ W_kD_k\right\rangle\right\rangle \\ &+\frac{N}{2}\sum_{k=1}^{K}\left\langle\left\langle R_k\left(\mu\bar{x}^T+\bar{x}\mu^T\right),\ W_kD_k\right\rangle\right\rangle - \frac{N}{2}\sum_{k=1}^{K}\left\langle\left\langle\left(R_k\mu\right)\left(R_k\mu\right)^T,\ W_kD_k\right\rangle\right\rangle \\ &+\frac{\nu+N-d-1}{2}\cdot\log\det\left(\sum_{k=1}^{K}R_k^T\left[W_kD_k\right]R_k\right),\end{aligned}$$

and we have compressed our notation by setting $\mu=\sum_k A_k\mu_k$, $\bar{x}=\frac{1}{N}\sum_n x_n$, $R_k=R\left(\Lambda_k\right)$, and $D_k=D\left(\Lambda_k\right)$. As before, we have

$$\tilde{\mathrm{H}}[q] = \frac{1}{K}\sum_{k=1}^{K}\log\det\left(W_k\right),$$

where we have suppressed the constant depending on the $p_{1:K}$ since it does not vary with W_k.

Recalling that W_k is diagonal, we can obtain the gradients by first computing

$$\begin{aligned}\nabla_{W_k}\mathcal{L}_0(W) = &\,D_k\cdot\mathrm{diag}\left[R_k\left(V^{-1}+X^TX\right)R_k^T\right] \\ &+\frac{N}{2}\cdot D_k\left(R_k\mu\circ\bar{x}+R_k\bar{x}\circ\mu\right) - \frac{N}{2}\cdot D_k\left(R_k\mu\right)\circ\left(R_k\mu\right) \\ &+\frac{\nu+N-d-1}{2}\cdot D_k\cdot\mathrm{diag}\left[R_k\left(\sum_{\ell=1}^{K}R_\ell^T\left[W_\ell D_\ell\right]R_\ell\right)^{-1}R_k^T\right],\end{aligned}$$

where we have used $\circ$ to denote elementwise vector products. We then find

$$\nabla_{W_k}\mathcal{L} = \mathbb{E}_q\left[\nabla_{W_k}\mathcal{L}_0(W)\right] + \frac{W_k^{-1}}{K}.$$

B.3. Mixture of Gaussians. Per the description of aggregation in Section 5, we define merged samples in the mixture of Gaussians model by the equations

$$\theta_\ell^* = F_{a\ell}\left(\theta_{1:K,1:L}\right) = \sum_{k=1}^{K}W_{k\ell}\theta_{ka_{k\ell}},$$

where $\ell = 1,\ldots,L$ denotes the cluster index and a_k denotes the alignment mapping indices on the master core to indices on worker core k. Throughout this section, we treat the alignment variables as fixed.

Using this notation, we define

$$\mathcal{L}_0\left(W,\ \theta_{1:K,1:L}\right) = -\frac{1}{2\tau^2}\sum_{\ell=1}^{L}||\theta_\ell^*||_2^2 - \frac{1}{2\sigma^2}\sum_{\ell=1}^{L}\sum_{i=1}^{n}\gamma_{i\ell}(W)\,||\theta_\ell^*-x_i||_2^2,$$

where

$$\gamma_{n\ell} = \frac{\tilde{\gamma}_{n\ell}}{\sum_{\ell'=1}^{L} \tilde{\gamma}_{n\ell'}}$$

and

$$\tilde{\gamma}_{n\ell} = \exp\left(-\frac{1}{2\sigma^2}\left|\left|\theta_\ell^* - x_n\right|\right|_2^2\right).$$

The variational objective then takes the form

$$\mathcal{L}\left(W\right) = \mathbb{E}_{p_{1:K}}\left[\mathcal{L}_0\left(W,\ \theta_{1:K,1:L}\right)\right] + \tilde{\mathrm{H}}\left[q\right],$$

with the usual equation

$$\tilde{\mathrm{H}}\left[q\right] = \frac{1}{K}\sum_{k=1}^{K}\sum_{\ell=1}^{L}\log\det\left(W_{k\ell}\right).$$

Some calculation then shows that the gradients with respect to the various $W_{k\ell}$ are given by

$$\begin{aligned}\nabla_{k\ell}\mathcal{L}_0\left(W,\ \theta_{1:K,1:L}\right) = {} & \frac{1}{2\sigma^4}\sum_{n=1}^{N}\gamma_{n\ell}\left(1-\gamma_{i\ell}\right)\left|\left|\theta_\ell^* - x_n\right|\right|_2^2 \cdot \theta_{ka_{k\ell}}\left(\theta_\ell^* - x_n\right)^T \\ & - \left(\frac{1}{\tau^2} + \frac{\sum_{n=1}^{N}\gamma_{n\ell}}{\sigma^2}\right) \cdot \theta_{ka_{k\ell}}\left(\theta_\ell^* - \tilde{x}_\ell\right)^T,\end{aligned}$$

where

$$\tilde{x}_\ell = \left(\frac{1}{\tau^2} + \frac{\sum_{n=1}^{N}\gamma_{n\ell}}{\sigma^2}\right)^{-1}\sum_{n=1}^{N}\frac{\gamma_{n\ell}}{\sigma^2}\cdot x_n.$$

This covers the case of general PSD matrices $W_{k\ell}$. When the matrices are restricted to be diagonal, we get the simplified gradient

$$\begin{aligned}\nabla_{k\ell}\mathcal{L}_0\left(W,\ \theta_{1:K,1:L}\right) = {} & \frac{1}{2\sigma^4}\sum_{n=1}^{N}\gamma_{i\ell}\left(1-\gamma_{n\ell}\right)\left|\left|\theta_\ell^* - x_n\right|\right|_2^2 \cdot \theta_{ka_{k\ell}} \circ \left(\theta_\ell^* - x_n\right) \\ & - \left(\frac{1}{\tau^2} + \frac{\sum_{n=1}^{N}\gamma_{n\ell}}{\sigma^2}\right) \cdot \theta_{ka_{k\ell}} \circ \left(\theta_\ell^* - \tilde{x}_\ell\right),\end{aligned}$$

where $\circ$ denotes elementwise multiplication of vectors.

Since

$$\nabla_{k\ell}\mathcal{L}\left(W\right) = \mathbb{E}_{p_{1:K}}\left[\nabla_{k\ell}\mathcal{L}\left(W,\ \theta_{1:K,1:L}\right)\right] + \frac{W_{k\ell}^{-1}}{K},$$

this gives us all the information we need to implement an optimization procedure for the objective.

APPENDIX C: EXTENDED EMPIRICAL EVALUATION

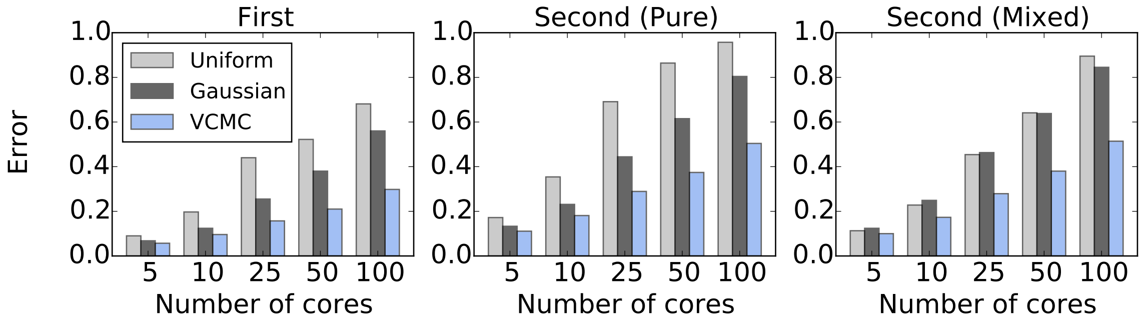

Fig 5: Five-dimensional probit regression ($d = 5$). Moment approximation error for the uniform and Gaussian averaging baselines and VCMC, relative to serial MCMC. We assessed three groups of functions: *(left)* first moments, with $f(\beta) = \beta_j$ for $1 \le j \le d$; *(center)* pure second moments, with $f(\beta) = \beta_j^2$ for $1 \le j \le d$; and *(right)* mixed second moments, with $f(\beta) = \beta_i\beta_j$ for $1 \le i < j \le d$.

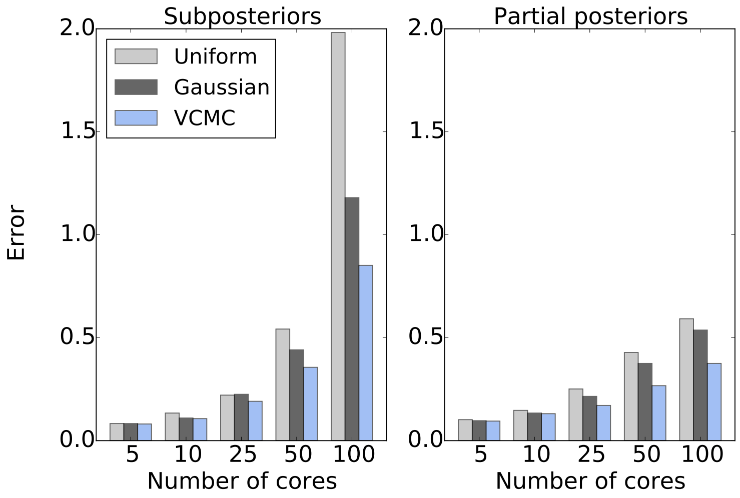

Fig 6: High-dimensional probit regression ($d = 300$). Moment approximation error for the uniform and Gaussian averaging baselines and VCMC, relative to serial MCMC, for subposteriors *(left)* and partial posteriors *(right)*. Here we show the pure second moments.

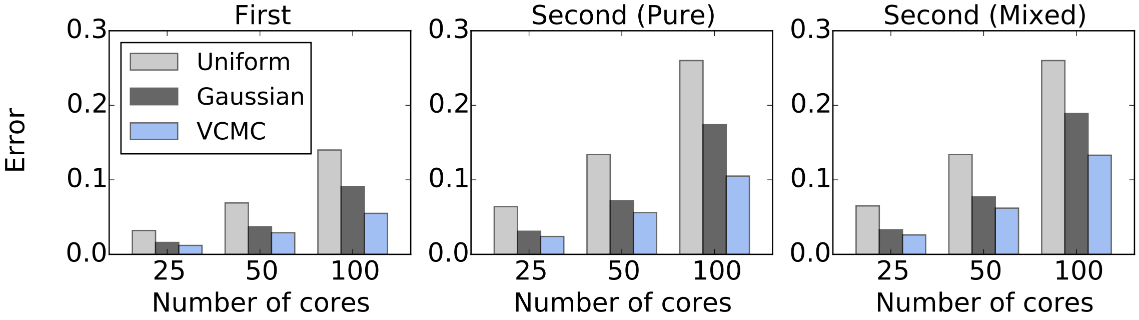

Fig 7: Five-dimensional normal-inverse Wishart model ($d = 5$). Moment approximation error for the uniform and Gaussian averaging baselines and VCMC, relative to serial MCMC. Letting ρ_j denote the j^{th} largest eigenvalue of Λ^{-1}, we assessed three groups of functions: *(left)* first moments, with $f(\Lambda) = \rho_j$ for $1 \leq j \leq d$; *(center)* pure second moments, with $f(\Lambda) = \rho_j^2$ for $1 \leq j \leq d$; and *(right)* mixed second moments, with $f(\Lambda) = \rho_i \rho_j$ for $1 \leq i < j \leq d$.

Computer Science Division
UC Berkeley
Berkeley, CA
E-mail: rabinovich@eecs.berkeley.edu
elaine@eecs.berkeley.edu
jordan@eecs.berkeley.edu