Universal Boosting Variational Inference

Trevor Campbell
Department of Statistics
University of British Columbia
Vancouver, BC V6T 1Z4
trevor@stat.ubc.ca

Xinglong Li
Department of Statistics
University of British Columbia
Vancouver, BC V6T 1Z4
xinglong.li@stat.ubc.ca

Abstract

Boosting variational inference (BVI) approximates an intractable probability density by iteratively building up a mixture of simple component distributions one at a time, using techniques from sparse convex optimization to provide both computational scalability and approximation error guarantees. But the guarantees have strong conditions that do not often hold in practice, resulting in degenerate component optimization problems; and we show that the ad-hoc regularization used to prevent degeneracy in practice can cause BVI to fail in unintuitive ways. We thus develop *universal boosting variational inference* (UBVI), a BVI scheme that exploits the simple geometry of probability densities under the Hellinger metric to prevent the degeneracy of other gradient-based BVI methods, avoid difficult joint optimizations of both component and weight, and simplify fully-corrective weight optimizations. We show that for any target density and any mixture component family, the output of UBVI converges to the best possible approximation in the mixture family, even when the mixture family is misspecified. We develop a scalable implementation based on exponential family mixture components and standard stochastic optimization techniques. Finally, we discuss statistical benefits of the Hellinger distance as a variational objective through bounds on posterior probability, moment, and importance sampling errors. Experiments on multiple datasets and models show that UBVI provides reliable, accurate posterior approximations.

1 Introduction

Bayesian statistical models provide a powerful framework for learning from data, with the ability to encode complex hierarchical dependence structures and prior domain expertise, as well as coherently capture uncertainty in latent parameters. The two predominant methods for Bayesian inference are Markov chain Monte Carlo (MCMC) [1, 2]—which obtains approximate posterior samples by simulating a Markov chain—and variational inference (VI) [3, 4]—which obtains an approximate distribution by minimizing some divergence to the posterior within a tractable family. The key strengths of MCMC are its generality and the ability to perform a computation-quality tradeoff: one can obtain a higher quality approximation by simulating the chain for a longer period [5, Theorem 4 & Fact 5]. However, the resulting Monte Carlo estimators have an unknown bias or random computation time [6], and statistical distances between the discrete sample posterior approximation and a diffuse true posterior are vacuous, ill-defined, or hard to bound without restrictive assumptions or a choice of kernel [7–9]. Designing correct MCMC schemes in the large-scale data setting is also a challenging task [10–12]. VI, on the other hand, is both computationally scalable and widely applicable due to advances from stochastic optimization and automatic differentiation [13–17]. However, the major disadvantage of the approach—and the fundamental reason that MCMC remains the preferred method in statistics—is that the variational family typically does not contain the posterior, fundamentally limiting the achievable approximation quality. And despite recent results in the asymptotic theory of variational methods [18–22], it is difficult to assess the effect of the chosen

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

family on the approximation for finite data; a poor choice can result in severe underestimation of posterior uncertainty [23, Ch. 21].

Boosting variational inference (BVI) [24–26] is an exciting new approach that addresses this fundamental limitation by using a nonparametric mixture variational family. By adding and reweighting only a single mixture component at a time, the approximation may be iteratively refined, achieving the computation/quality tradeoff of MCMC and the scalability of VI. Theoretical guarantees on the convergence rate of Kullback-Leibler (KL) divergence [24, 27, 28] are much stronger than those available for standard Monte Carlo, which degrade as the number of estimands increases, enabling the practitioner to confidently reuse the same approximation for multiple tasks. However, the bounds require the KL divergence to be sufficiently smooth over the class of mixtures—an assumption that does not hold for many standard mixture families, e.g. Gaussians, resulting in a degenerate procedure in practice. To overcome this, an ad-hoc entropy regularization is typically added to each component optimization; but this regularization invalidates convergence guarantees, and—depending on the regularization weight—sometimes does not actually prevent degeneracy.

In this paper, we develop *universal boosting variational inference* (UBVI), a variational scheme based on the Hellinger distance rather than the KL divergence. The primary advantage of using the Hellinger distance is that it endows the space of probability densities with a particularly simple unit-spherical geometry in a Hilbert space. We exploit this geometry to prevent the degeneracy of other gradient-based BVI methods, avoid difficult joint optimizations of both component and weight, simplify fully-corrective weight optimizations, and provide a procedure in which the normalization constant of f does not need to be known, a crucial property in most VI settings. It also leads to the universality of UBVI: we show that for *any* target density and *any* mixture component family, the output of UBVI converges to the best possible approximation in the mixture family, even when the mixture family is misspecified. We develop a scalable implementation based on exponential family mixture components and standard stochastic optimization techniques. Finally, we discuss other statistical benefits of the Hellinger distance as a variational objective through bounds on posterior probability, moment, and importance sampling errors. Experiments on multiple datasets and models show that UBVI provides reliable, accurate posterior approximations.

2 Background: variational inference and boosting

Variational inference, in its most general form, involves approximating a probability density p by minimizing some divergence $\mathrm{D}\left(\cdot||\cdot\right)$ from ξ to p over densities ξ in a family $\mathcal{Q}$,

$$q = \arg\min_{\xi \in \mathcal{Q}} \mathrm{D}\left(\xi||p\right).$$

Past work has almost exclusively involved parametric families $\mathcal{Q}$, such as mean-field exponential families [4], finite mixtures [29–31], normalizing flows [32], and neural nets [16]. The issue with these families is that typically $\min_{\xi \in \mathcal{Q}} \mathrm{D}\left(\xi||p\right) > 0$—meaning the practitioner cannot achieve arbitrary approximation quality with more computational effort—and a priori, there is no way to tell how poor the best approximation is. To address this, *boosting variational inference* (BVI) [24–26] proposes the use of the nonparametric family of all finite mixtures of a component density family $\mathcal{C}$,

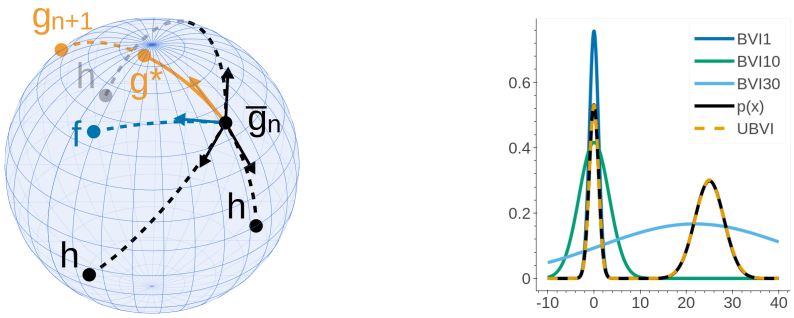
$$\mathcal{Q} = \operatorname{conv} \mathcal{C} := \left\{ \sum_{k=1}^{K} w_k \xi_k : K \in \mathbb{N}, w \in \Delta^{K-1},\ \forall k \in \mathbb{N}\ \xi_k \in \mathcal{C} \right\}.$$

Given a judicious choice of $\mathcal{C}$, we have that $\inf_{\xi \in \mathcal{Q}} \mathrm{D}\left(\xi||p\right) = 0$; in other words, we can approximate any continuous density p with arbitrarily low divergence [33]. As optimizing directly over the nonparametric $\mathcal{Q}$ is intractable, BVI instead adds one component at a time to iteratively refine the approximation. There are two general formulations of BVI; Miller et al. [26] propose minimizing KL divergence over both the weight and component simultaneously,

$$q_n = \sum_{k=1}^{n} w_{nk} \xi_k \quad \xi_{n+1}, \omega = \arg\min_{\xi \in \mathcal{C}, \rho \in [0,1]} \mathrm{D}_{\mathrm{KL}}\left(\rho\xi + (1-\rho) q_n||p\right) \quad w_{n+1} = \left[(1-\omega) w_n\ \omega\right]^T,$$

while Guo et al. and Wang [24, 25] argue that optimizing both simultaneously is too difficult, and use a gradient boosting [34] formulation instead,

$$\xi_{n+1} = \arg\min_{\xi \in \mathcal{C}} \left\langle \xi, \nabla \mathrm{D}_{\mathrm{KL}}\left(\cdot||p\right)\big|_{q_n} \right\rangle \quad w_{n+1} = \arg\min_{\omega = \left[(1-\rho) w_n\ \rho\right]^T, \rho \in [0,1]} \mathrm{D}_{\mathrm{KL}}\left(\sum_{k=1}^{n+1} \omega_k \xi_k || p \right).$$



(a) (b)

Figure 1: (1a): Greedy component selection, with target f, current iterate $\bar{g}_n$, candidate components h, optimal component g_{n+1}, the closest point $g^\star$ to f on the $\bar{g}_n \to g_{n+1}$ geodesic, and arrows for initial geodesic directions. The quality of g_{n+1} is determined by the distance from f to $g^\star$, or equivalently, by the alignment of the initial directions $\bar{g}_n \to g_{n+1}$ and $\bar{g}_n \to f$. (1b): BVI can fail even when p is in the mixture family. Here $p = \frac{1}{2}\mathcal{N}(0,1) + \frac{1}{2}\mathcal{N}(25,5)$, and UBVI finds the correct mixture in 2 iterations. BVI (with regularization weight in $\{1, 10, 30\}$) does not converge. For example, when the regularization weight is 1 the first component will have variance < 5, and the second component optimization diverges since the target $\mathcal{N}(25,5)$ component has a heavier tail. Upon reweighting the second component is removed, and the approximation will never improve.

Both algorithms attain $\mathrm{D}_{\mathrm{KL}}\left(q_N||p\right) = O(1/N)$[1]— the former by appealing to results from convex functional analysis [35, Theorem II.1], and the latter by viewing BVI as functional Frank-Wolfe optimization [27, 36, 37]. This requires that $\mathrm{D}_{\mathrm{KL}}\left(q||p\right)$ is strongly smooth or has bounded curvature over $q \in \mathcal{Q}$, for which it is sufficient that densities in $\mathcal{Q}$ are bounded away from 0, bounded above, and have compact support [27], or have a bounded parameter space [28]. However, these assumptions do not hold in practice for many simple (and common) cases, e.g., where $\mathcal{C}$ is the class of multivariate normal distributions. Indeed, gradient boosting-based BVI methods all require some ad-hoc entropy regularization in the component optimizations to avoid degeneracy [24, 25, 28]. In particular, given a sequence of regularization weights $r_n > 0$, BVI solves the following component optimization [28]:

$$\xi_{n+1} = \arg\min_{\xi \in \mathcal{C}} \left\langle \xi, \log \frac{\xi^{r_{n+1}} q_n}{p} \right\rangle. \tag{1}$$

This addition of regularization has an adverse effect on performance in practice as demonstrated in Fig. 1b, and can lead to unintuitive behaviour and nonconvergence—even when $p \in \mathcal{Q}$ (Proposition 1) or when the distributions in $\mathcal{C}$ have lighter tails than p (Proposition 2).

Proposition 1. *Suppose $\mathcal{C}$ is the set of univariate Gaussians with mean 0 parametrized by variance, let $p = \mathcal{N}(0,1)$, and let the initial approximation be $q_1 = \mathcal{N}(0,\tau^2)$. Then BVI in Eq. (1) with regularization $r_2 > 0$ returns a degenerate next component ξ_2 if $\tau^2 \leq 1$, and iterates infinitely without improving the approximation if $\tau^2 > 1$ and $r_2 > \tau^2 - 1$.*

Proposition 2. *Suppose $\mathcal{C}$ is the set of univariate Gaussians with mean 0 parametrized by variance, and let $p = \mathrm{Cauchy}(0,1)$. Then BVI in Eq. (1) with regularization $r_1 > 0$ returns a degenerate first component ξ_1 if $r_1 \geq 2$.*

3 Universal boosting variational inference (UBVI)

3.1 Algorithm and convergence guarantee

To design a BVI procedure without the need for ad-hoc regularization, we use a variational objective based on the *Hellinger distance*, which for any probability space $(\mathcal{X}, \Sigma, \mu)$ and densities p, q is

$$\mathrm{D}^2_{\mathrm{H}}\left(p,q\right) := \frac{1}{2}\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mu(\mathrm{d}x).$$

[1]We assume throughout that nonconvex optimization problems can be solved reliably.

Algorithm 1 The universal boosting variational inference (UBVI) algorithm.

1: **procedure** UBVI(p, $\mathcal{H}$, N)
2: $f \overset{\propto}{\leftarrow} \sqrt{p}$
3: $\bar{g}_0 \leftarrow 0$
4: **for** $n = 1, \ldots, N$ **do**
▷ Find the next component to add to the approximation using Eq. (5)
5: $g_n \leftarrow \arg\max_{h \in \mathcal{H}} \langle f - \langle f, \bar{g}_{n-1} \rangle \bar{g}_{n-1}, h \rangle \Big/ \sqrt{1 - \langle h, \bar{g}_{n-1} \rangle^2}$
▷ Compute pairwise normalizations using Eq. (2)
6: **for** $i = 1, \ldots, n$ **do**
7: $Z_{n,i} = Z_{i,n} \leftarrow \langle g_n, g_i \rangle$
8: **end for**
▷ Update weights using Eq. (7)
9: $d = \left(\langle f, g_1 \rangle, \ldots, \langle f, g_n \rangle \right)^T$
10: $\beta = \arg\min_{b \in \mathbb{R}^n, b \geq 0} b^T Z^{-1} b + 2 b^T Z^{-1} d$
11: $(\lambda_{n,1}, \ldots, \lambda_{n,n}) = \frac{Z^{-1}(\beta + d)}{\sqrt{(\beta+d)^T Z^{-1} (\beta+d)}}$
▷ Update boosting approximation
12: $\bar{g}_n \leftarrow \sum_{i=1}^n \lambda_{ni} g_i$
13: **end for**
14: **return** $q = \bar{g}_N^2$
15: **end procedure**

Our general approach relies on two facts about the Hellinger distance. First, the metric $\mathrm{D}_{\mathrm{H}}(\cdot, \cdot)$ endows the set of μ-densities with a simple geometry corresponding to the nonnegative functions on the unit sphere in $L^2(\mu)$. In particular, if $f, g \in L^2(\mu)$ satisfy $\|f\|_2 = \|g\|_2 = 1$, $f, g \geq 0$, then $p = f^2$ and $q = g^2$ are probability densities and

$$\mathrm{D}_{\mathrm{H}}^2(p, q) = \frac{1}{2} \|f - g\|_2^2 .$$

One can thus perform Hellinger distance boosting by iteratively finding components that minimize geodesic distance to f on the unit sphere in $L^2(\mu)$. Like the Miller et al. approach [26], the boosting step directly minimizes a statistical distance, leading to a nondegenerate method; but like the Guo et al. and Wang approach [24, 25], this avoids the joint optimization of both component and weight; see Section 3.2 for details. Second, a conic combination $g = \sum_{i=0}^N \lambda_i g_i$, $\lambda_i \geq 0$, $\|g_i\|_2 = 1$, $g_i \geq 0$ in $L^2(\mu)$ satisfying $\|g\|_2 = 1$ corresponds to the mixture model density

$$q = g^2 = \sum_{i,j=1}^N Z_{ij} \lambda_i \lambda_j \left(\frac{g_i g_j}{Z_{ij}} \right) \qquad Z_{ij} := \langle g_i, g_j \rangle \geq 0. \tag{2}$$

Therefore, if we can find a conic combination satisfying $\|f - g\|_2 \leq \sqrt{2}\epsilon$ for $p = f^2$, we can guarantee that the corresponding mixture density q satisfies $\mathrm{D}_{\mathrm{H}}(p, q) \leq \epsilon$. The mixture will be built from a family $\mathcal{H} \subset L^2(\mu)$ of component functions for which $\forall h \in \mathcal{H}$, $\|h\|_2 = 1$ and $h \geq 0$. We assume that the target function $f \in L^2(\mu)$, $\|f\|_2 = 1$, $f \geq 0$ is known up to proportionality. We also assume that f is not orthogonal to $\operatorname{span} \mathcal{H}$ for expositional brevity, although the algorithms and theoretical results presented here apply equally well in this case. We make no other assumptions; in particular, we do not assume f is in $\operatorname{cl} \operatorname{span} \mathcal{H}$.

The universal boosting variational inference (UBVI) procedure is shown in Algorithm 1. In each iteration, the algorithm finds a new mixture component from $\mathcal{H}$ (line 5; see Section 3.2 and Fig. 1a). Once the new component is found, the algorithm solves a convex quadratic optimization problem to update the weights (lines 9–11). The primary requirement to run Algorithm 1 is the ability to compute or estimate $\langle h, f \rangle$ and $\langle h, h' \rangle$ for $h, h' \in \mathcal{H}$. For this purpose we employ an exponential component family $\mathcal{H}$ such that Z_{ij} is available in closed-form, and use samples from h^2 to obtain estimates of $\langle h, f \rangle$; see Appendix A for further implementation details.

The major benefit of UBVI is that it comes with a computation/quality tradeoff akin to MCMC: for any target p and component family $\mathcal{H}$, (1) there is a unique mixture $\hat{p} = \hat{f}^2$ minimizing $\mathrm{D}_{\mathrm{H}}(\hat{p}, p)$ over the closure of finite mixtures $\operatorname{cl} \mathcal{Q}$; and (2) the output q of UBVI$(p, \mathcal{H}, N)$ satisfies $\mathrm{D}_{\mathrm{H}}(q, \hat{p}) = O(1/N)$ with a dimension-independent constant. No matter how coarse the family $\mathcal{H}$ is, the output of UBVI will converge to the best possible mixture approximation. Theorem 3 provides the precise result.

Theorem 3. *For any density p there is a unique density $\hat{p} = \hat{f}^2$ satisfying $\hat{p} = \arg\min_{\xi \in \mathrm{cl}\,\mathcal{Q}} \mathrm{D}_{\mathrm{H}}(\xi, p)$. If each component optimization Eq.* (5) *is solved with a relative suboptimality of at most $(1-\delta)$, then the variational mixture approximation q returned by* UBVI*(p, $\mathcal{H}$, N) satisfies*

$$\mathrm{D}_{\mathrm{H}}(\hat{p}, q)^2 \leq \frac{J_1}{1 + \left(\frac{1-\delta}{\tau}\right)^2 J_1(N-1)} \quad J_1 := 1 - \left\langle \hat{f}, g_1 \right\rangle^2 \in [0,1) \quad \tau := \textit{Eq.}\ (3) < \infty.$$

The proof of Theorem 3 may be found in Appendix B.3, and consists of three primary steps. First, Lemma 9 guarantees the existence and uniqueness of the convergence target $\hat{f}$ under possible misspecification of the component family $\mathcal{H}$. Then the difficulty of approximating $\hat{f}$ with conic combinations of functions in $\mathcal{H}$ is captured by the basis pursuit denoising problem [38]

$$\tau := \inf_{\substack{h_i \in \mathrm{cone}\,\mathcal{H} \\ x \in [0,1)}} (1-x)^{-1} \sum_{i=1}^{\infty} \|h_i\|_2 \quad \text{s.t.} \quad \|\hat{f} - \sum_{i=1}^{\infty} h_i\|_2 \leq x, \quad \forall i,\ h_i \geq 0. \tag{3}$$

Lemma 10 guarantees that τ is finite, and in particular $\tau \leq \frac{\sqrt{1-J_1}}{1-\sqrt{J_1}}$, which can be estimated in practice using Eq. (9). Finally, Lemma 11 develops an objective function recursion, which is then solved to yield Theorem 3. Although UBVI and Theorem 3 is reminiscent of past work on greedy approximation in a Hilbert space [34, 39–46], it provides the crucial advantage that the greedy steps do not require knowledge of the normalization of p. UBVI is inspired by a previous greedy method [46], but provides guarantees with an arbitrary, potentially misspecified infinite dictionary in a Hilbert space, and uses quadratic optimization to perform weight updates. Note that both the theoretical and practical cost of UBVI is dominated by finding the next component (line 5), which is a nonconvex optimization problem. The other expensive step is inverting Z; however, incremental methods using block matrix inversion [47, p. 46] reduce the cost at iteration n to $O(n^2)$ and overall cost to $O(N^3)$, which is not a concern for practical mixtures with $\ll 10^3$ components. The weight optimization (line 10) is a nonnegative least squares problem, which can be solved efficiently [48, Ch. 23].

3.2 Greedy boosting along density manifold geodesics

This section provides the technical derivation of UBVI (Algorithm 1) by expoiting the geometry of square-root densities under the Hellinger metric. Let the conic combination in $L^2(\mu)$ after initialization followed by $N-1$ steps of greedy construction be denoted

$$\bar{g}_n := \sum_{i=1}^{n} \lambda_{ni} g_i, \quad \|\bar{g}_n\|_2 = 1,$$

where $\lambda_{ni} \geq 0$ is the weight for component i at step n, and g_i is the component added at step i. To find the next component, we minimize the distance between $\bar{g}_{n+1}$ and f over choices of $h \in \mathcal{H}$ and position $x \in [0,1]$ along the $\bar{g}_n \to h$ geodesic,[2]

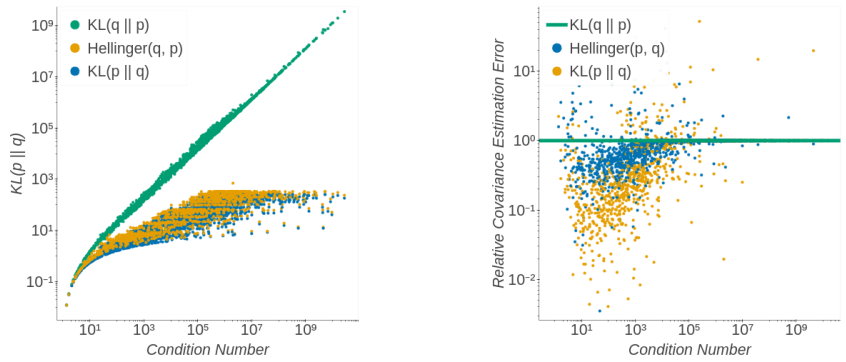
$$\begin{aligned} \bar{g}_0 = 0 \qquad g_{n+1}, x^\star &= \operatorname*{arg\,min}_{h \in \mathcal{H}, x \in [0,1]} \left\| f - \left(x \frac{h - \langle h, \bar{g}_n \rangle \bar{g}_n}{\|h - \langle h, \bar{g}_n \rangle \bar{g}_n\|_2} + \sqrt{1-x^2}\, \bar{g}_n \right) \right\|_2 \\ &= \operatorname*{arg\,max}_{h \in \mathcal{H}, x \in [0,1]} x \left\langle f, \frac{h - \langle h, \bar{g}_n \rangle \bar{g}_n}{\|h - \langle h, \bar{g}_n \rangle \bar{g}_n\|_2} \right\rangle + \sqrt{1-x^2}\, \langle f, \bar{g}_n \rangle . \end{aligned} \tag{4}$$

Noting that $h - \langle h, \bar{g}_n \rangle \bar{g}_n$ is orthogonal to $\bar{g}_n$, the second term does not depend on h, and $x \geq 0$, we avoid optimizing the weight and component simultaneously and find that

$$g_{n+1} = \operatorname*{arg\,max}_{h \in \mathcal{H}} \left\langle \frac{f - \langle f, \bar{g}_n \rangle \bar{g}_n}{\|f - \langle f, \bar{g}_n \rangle \bar{g}_n\|_2}, \frac{h - \langle h, \bar{g}_n \rangle \bar{g}_n}{\|h - \langle h, \bar{g}_n \rangle \bar{g}_n\|_2} \right\rangle = \operatorname*{arg\,max}_{h \in \mathcal{H}} \frac{\langle f - \langle f, \bar{g}_n \rangle \bar{g}_n, h \rangle}{\sqrt{1 - \langle h, \bar{g}_n \rangle^2}}. \tag{5}$$

Intuitively, Eq. (5) attempts to maximize alignment of g_{n+1} with the residual $f - \langle f, \bar{g}_n \rangle \bar{g}_n$ (the numerator) resulting in a ring of possible solutions, and among these, Eq. (5) minimizes alignment

[2] Note that the $\arg\max$ may not be unique, and when $\mathcal{H}$ is infinite it may not exist; Theorem 3 still holds and UBVI works as intended in this case. For simplicity, we use $(\dots) = \arg\max(\dots)$ throughout.



(a) (b)

Figure 2: Forward KL divergence—which controls worst-case downstream importance sampling error—and importance-sampling-based covariance estimation error on a task of approximating $\mathcal{N}(0, A^T A)$, $A_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ with $\mathcal{N}(0, \sigma^2 I)$ by minimizing Hellinger, forward KL, and reverse KL, plotted as a function of condition number $\kappa(A^T A)$. Minimizing Hellinger distance provides significantly lower forward KL divergence and estimation error than minimizing reverse KL.

with the current iterate $\bar{g}_n$ (the denominator). The first form in Eq. (5) provides an alternative intuition: g_{n+1} achieves the maximal alignment of the *initial geodesic directions* $\bar{g}_n \to f$ and $\bar{g}_n \to h$ on the sphere. See Fig. 1a for a depiction. After selecting the next component g_{n+1}, one option to obtain $\bar{g}_{n+1}$ is to use the optimal weighting $x^\star$ from Eq. (4); in practice, however, it is typically the case that solving Eq. (5) is expensive enough that finding the optimal set of coefficients for $\{g_1, \ldots, g_{n+1}\}$ is worthwhile. This is accomplished by maximizing alignment with f subject to a nonnegativity and unit-norm constraint:

$$(\lambda_{(n+1)1}, \ldots, \lambda_{(n+1)(n+1)}) = \underset{x \in \mathbb{R}^{n+1}}{\arg\max} \quad \left\langle f, \sum_{i=1}^{n+1} x_i g_i \right\rangle \quad \text{s.t.} \quad x \geq 0, \quad x^T Z x \leq 1, \tag{6}$$

where $Z \in \mathbb{R}^{N+1 \times N+1}$ is the matrix with entries Z_{ij} from Eq. (2). Since projection onto the feasible set of Eq. (6) may be difficult, the problem may instead be solved using the dual via

$$\begin{aligned} (\lambda_{(n+1)1}, \ldots, \lambda_{(n+1)(n+1)}) &= \frac{Z^{-1}(\beta + d)}{\sqrt{(\beta + d)^T Z^{-1} (\beta + d)}} \\ d = (\langle f, g_1 \rangle, \ldots, \langle f, g_{n+1} \rangle)^T \qquad \beta &= \underset{b \in \mathbb{R}^{n+1}, b \geq 0}{\arg\min} \; b^T Z^{-1} b + 2 b^T Z^{-1} d. \end{aligned} \tag{7}$$

Eq. (7) is a nonnegative linear least-squares problem—for which very efficient algorithms are available [48, Ch. 23]—in contrast to prior variational boosting methods, where a fully-corrective weight update is a general constrained convex optimization problem. Note that, crucially, none of the above steps rely on knowledge of the normalization constant of f.

4 Hellinger distance as a variational objective

While the Hellinger distance has most frequently been applied in asymptotic analyses (e.g., [49]), it has seen recent use as a variational objective [50] and possesses a number of particularly useful properties that make it a natural fit for this purpose. First, $\mathrm{D}_{\mathrm{H}}(\cdot, \cdot)$ applies to any arbitrary pair of densities, unlike $\mathrm{D}_{\mathrm{KL}}(p||q)$, which requires that $p \ll q$. Minimizing $\mathrm{D}_{\mathrm{H}}(\cdot, \cdot)$ also implicitly minimizes error in posterior probabilities and moments—two quantities of primary importance to practitioners—via its control on total variation and ℓ-Wasserstein by Propositions 4 and 5. Note that the upper bound in Proposition 4 is typically tighter than that provided by the usual $\mathrm{D}_{\mathrm{KL}}(q||p)$ variational objective via Pinsker's inequality (and at the very least is always in $[0, 1]$), and the bound in Proposition 5 shows that convergence in $\mathrm{D}_{\mathrm{H}}(\cdot, \cdot)$ implies convergence in up to ℓ^{th} moments [51, Theorem 6.9] under relatively weak conditions.

Proposition 4 (e.g. [52, p. 61])**.** *The Hellinger distance bounds total variation via*

$$\mathrm{D}_{\mathrm{H}}^2(p, q) \leq \mathrm{D}_{\mathrm{TV}}(p, q) := \frac{1}{2}\|p - q\|_1 \leq \mathrm{D}_{\mathrm{H}}(p, q) \sqrt{2 - \mathrm{D}_{\mathrm{H}}^2(p, q)}.$$

Proposition 5. *Suppose $\mathcal{X}$ is a Polish space with metric $d(\cdot,\cdot)$, $\ell \geq 1$, and p, q are densities with respect to a common measure μ. Then for any x_0,*

$$W_\ell(p,q) \leq 2\mathrm{D}_{\mathrm{H}}(p,q)^{1/\ell}\left(\mathbb{E}\left[d(x_0,X)^{2\ell}\right]+\mathbb{E}\left[d(x_0,Y)^{2\ell}\right]\right)^{1/2\ell},$$

where $Y \sim p(y)\mu(\mathrm{d}y)$ and $X \sim q(x)\mu(\mathrm{d}x)$. In particular, if densities $(q_N)_{N\in\mathbb{N}}$ and p have uniformly bounded $2\ell^{th}$ moments, $\mathrm{D}_{\mathrm{H}}(p,q_N) \to 0 \implies W_\ell(p,q_N) \to 0$ as $N \to \infty$.

Once a variational approximation q is obtained, it will typically be used to estimate expectations of some function of interest $\phi(x) \in L^2(\mu)$ via Monte Carlo. Unless q is trusted entirely, this involves importance sampling—using $I_n(\phi)$ or $J_n(\phi)$ in Eq. (8) depending on whether the normalization of p is known—to account for the error in q compared with the target distribution p [53],

$$I_n(\phi) := \frac{1}{N}\sum_{n=1}^{N}\frac{p(X_i)}{q(X_i)}\phi(X_i) \qquad J_n(\phi) := \frac{I_n(\phi)}{I_n(1)} \qquad X_i \overset{\text{i.i.d.}}{\sim} q(x)\mu(\mathrm{d}x). \tag{8}$$

Recent work has shown that the error of importance sampling is controlled by the intractable forward KL-divergence $\mathrm{D}_{\mathrm{KL}}(p||q)$ [54]. This is where the Hellinger distance shines; Proposition 6 shows that it penalizes both positive and negative values of $\log p(x)/q(x)$ and thus provides moderate control on $\mathrm{D}_{\mathrm{KL}}(p||q)$—unlike $\mathrm{D}_{\mathrm{KL}}(q||p)$, which only penalizes negative values. See Fig. 2 for a demonstration of this effect on the classical correlated Gaussian example [23, Ch. 21]. While the takeaway from this setup is typically that minimizing $\mathrm{D}_{\mathrm{KL}}(q||p)$ may cause severe underestimation of variance, a reasonable practitioner should attempt to use importance sampling to correct for this anyway. But Fig. 2 shows that minimizing $\mathrm{D}_{\mathrm{KL}}(q||p)$ doesn't minimize $\mathrm{D}_{\mathrm{KL}}(p||q)$ well, leading to poor estimates from importance sampling. Even though minimizing $\mathrm{D}_{\mathrm{H}}(p,q)$ also underestimates variance, it provides enough control on $\mathrm{D}_{\mathrm{KL}}(p||q)$ so that importance sampling can correct the errors. Direct bounds on the error of importance sampling estimates are also provided in Proposition 7.

Proposition 6. *Define $R := \log\frac{p(X)}{q(X)}$ where $X \sim p(x)\mu(\mathrm{d}x)$. Then*

$$\mathrm{D}_{\mathrm{H}}(p,q) \geq \frac{1}{2}\sqrt{\mathbb{E}\left[R^2\left(\frac{1+\mathbb{1}[R\leq 0]R}{1+R}\right)^2\right]} \geq \frac{\mathrm{D}_{\mathrm{KL}}(p||q)}{2\sqrt{1+\mathbb{E}\left[\mathbb{1}[R>0](1+R)^2\right]}}.$$

Proposition 7. *Define $\alpha := \left(N^{-1/4}+2\sqrt{\mathrm{D}_{\mathrm{H}}(p,q)}\right)^2$. Then the importance sampling error with known normalization is bounded by*

$$\mathbb{E}\left[|I_n(\phi)-I(\phi)|\right] \leq \|\sqrt{p}\phi\|_2\alpha,$$

and with unknown normalization by

$$\forall t>0 \qquad \mathbb{P}(|J_n(\phi)-I(\phi)| > \|\sqrt{p}\phi\|_2 t) \leq \left(1+4t^{-1}\sqrt{1+t}\right)\alpha.$$

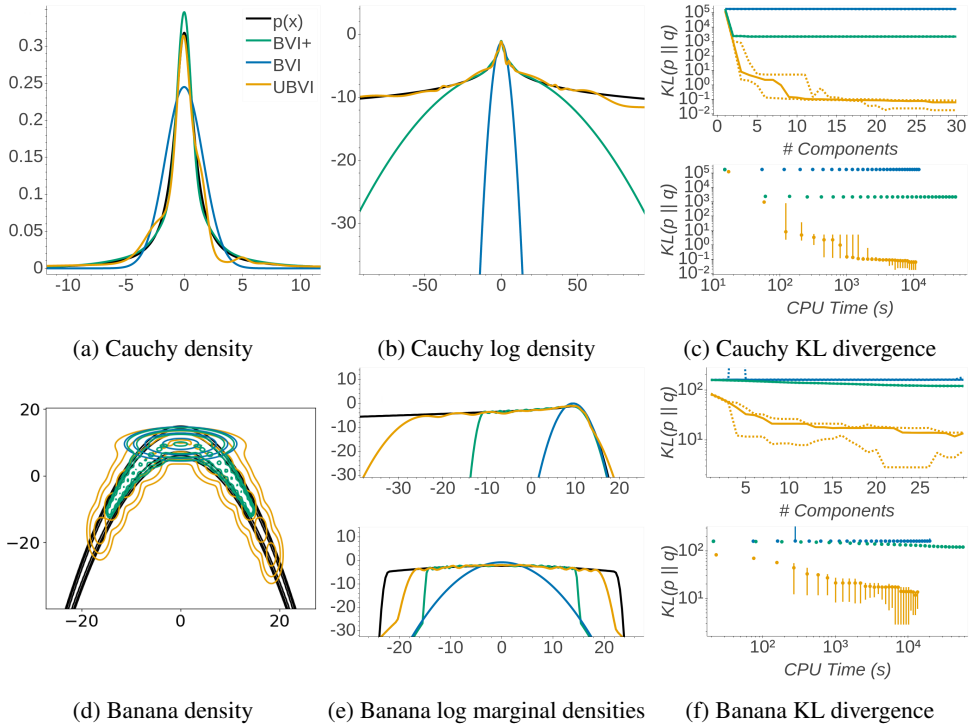
Next, the Hellinger distance between densities q, p can be estimated with high relative accuracy given samples from q, enabling the use of the above bounds in practice. This involves computing either $\widehat{\mathrm{D}^2_{\mathrm{H}}(p,q)}$ or $\widetilde{\mathrm{D}^2_{\mathrm{H}}(p,q)}$ below, depending on whether the normalization of p is known. The expected error of both of these estimates relative to $\mathrm{D}_{\mathrm{H}}(p,q)$ is bounded via Proposition 8.

$$\widehat{\mathrm{D}^2_{\mathrm{H}}(p,q)} := 1-\frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{p(X_n)}{q(X_n)}}, \quad \widetilde{\mathrm{D}^2_{\mathrm{H}}(p,q)} := 1-\frac{\frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{p(X_n)}{q(X_n)}}}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}\frac{p(X_n)}{q(X_n)}}}, \quad X_n \overset{\text{i.i.d.}}{\sim} q(x)\mu(\mathrm{d}x). \tag{9}$$

Proposition 8. *The mean absolute difference between the Hellinger squared estimates is*

$$\mathbb{E}\left[\left|\widehat{\mathrm{D}^2_{\mathrm{H}}(p,q)}-\mathrm{D}_{\mathrm{H}}(p,q)^2\right|\right] \leq \frac{\mathrm{D}_{\mathrm{H}}(p,q)\sqrt{2-\mathrm{D}_{\mathrm{H}}(p,q)^2}}{\sqrt{N}}$$
$$\mathbb{E}\left[\left|\widetilde{\mathrm{D}^2_{\mathrm{H}}(p,q)}-\mathrm{D}_{\mathrm{H}}(p,q)^2\right|\right] \leq \sqrt{2}\left(1+\sqrt{N}^{-1}\right)\mathrm{D}_{\mathrm{H}}(p,q).$$

It is worth pointing out that although the above statistical properties of the Hellinger distance make it well-suited as a variational objective, it does pose computational issues during optimization. In particular, to avoid numerically unstable gradient estimation, one must transform Hellinger-based objectives such as Eq. (5). This typically produces biased and occasionally higher-variance Monte Carlo gradient estimates than the corresponding KL gradient estimates. We detail these transformations and other computational considerations in Appendix A.



(a) Cauchy density (b) Cauchy log density (c) Cauchy KL divergence

(d) Banana density (e) Banana log marginal densities (f) Banana KL divergence

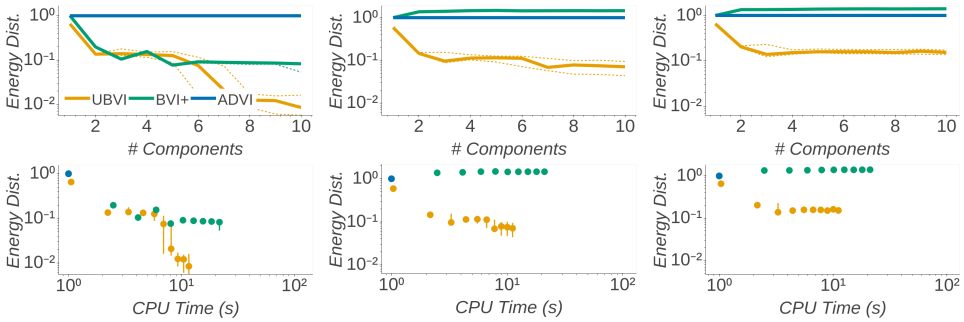
Figure 3: Results on the Cauchy and banana distributions; all subfigures use the legend from Fig. 3a. (Figs. 3a and 3d): Density approximation with 30 components for Cauchy (3a) and banana (3d). BVI has degenerate component optimizations after the first, while UBVI and BVI+ are able to refine the approximation. (Figs. 3b and 3e): Log density approximations for Cauchy (3b) and banana marginals (3e). UBVI provides more accurate approximation of distribution tails than the KL-based BVI(+) algorithms. (Figs. 3c and 3f): The forward KL divergence vs. the number of boosting components and computation time. UBVI consistently improves its approximation as more components are added, while the KL-based BVI(+) methods improve either slowly or not at all due to degeneracy. Solid lines / dots indicate median, and dashed lines / whiskers indicate 25^{th} and 75^{th} percentile.

5 Experiments

In this section, we compare UBVI, KL divergence boosting variational inference (BVI) [28], BVI with an ad-hoc stabilization in which q_n in Eq. (1) is replaced by $q_n + 10^{-3}$ to help prevent degeneracy (BVI+), and standard VI. For all experiments, we used a regularization schedule of $r_n = 1/\sqrt{n}$ for BVI(+) in Eq. (1). We used the multivariate Gaussian family for $\mathcal{H}$ parametrized by mean and log-transformed diagonal covariance matrix. We used 10,000 iterations of ADAM [55] for optimization, with decaying step size $1/\sqrt{1+i}$ and Monte Carlo gradients based on 1,000 samples. Fully-corrective weight optimization was conducted via simplex-projected SGD for BVI(+) and nonnegative least squares for UBVI. Monte Carlo estimates of $\langle f, g_n \rangle$ in UBVI were based on 10,000 samples. Each component optimization was initialized from the best of 10,000 trials of sampling a component (with mean m and covariance Σ) from the current mixture, sampling the initialized component mean from $\mathcal{N}(m, 16\Sigma)$, and setting the initialized component covariance to $\exp(Z)\Sigma$, $Z \sim \mathcal{N}(0,1)$. Each experiment was run 20 times with an Intel i7 8700K processor and 32GB of memory. Code is available at `www.github.com/trevorcampbell/ubvi`.

5.1 Cauchy and banana distributions

Fig. 3 shows the results of running UBVI, BVI, and BVI+ for 30 boosting iterations on the standard univariate Cauchy distribution and the banana distribution [56] with curvature $b = 0.1$. These distributions were selected for their heavy tails and complex structure (shown in Figs. 3b and 3e), two features that standard variational inference does not often address but boosting methods should



(a) Synthetic (b) Chemical Reactivity (c) Phishing

Figure 4: Results from Bayesian logistic regression posterior inference on the synthetic (4a), chemical (4b), and phishing (4c) datasets, showing the energy distance [57] to the posterior (via NUTS [58]) vs. the number of components and CPU time. Energy distance and time are normalized by the VI median. Solid lines / dots indicate median, and dashed lines / whiskers indicate 25^{th} / 75^{th} percentile.

handle. However, BVI particularly struggles with heavy-tailed distributions, where its component optimization objective after the first is degenerate. BVI+ is able to refine its approximation, but still cannot capture heavy tails well, leading to large forward KL divergence (which controls downstream importance sampling error). We also found that the behaviour of BVI(+) is very sensitive to the choice of regularization tuning schedule r_n, and is difficult to tune well. UBVI, in contrast, approximates both heavy-tailed and complex distributions well with few components, and involves no tuning effort beyond the component optimization step size.

5.2 Logistic regression with a heavy-tailed prior

Fig. 4 shows the results of running 10 boosting iterations of UBVI, BVI+, and standard VI for posterior inference in Bayesian logistic regression. We used a multivariate $\mathcal{T}_2(\mu, \Sigma)$ prior, where in each trial, the prior parameters were set via $\mu = 0$ and $\Sigma = A^T A$ for $A_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. We ran this experiment on a 2-dimensional synthetic dataset generated from the model, a 10-dimensional chemical reactivity dataset, and a 10-dimensional phishing websites dataset, each with 20 subsampled datapoints.[3] The small dataset size and heavy-tailed prior were chosen to create a complex posterior structure better-suited to evaluating boosting variational methods. The results in Fig. 4 are similar to those in the synthetic test from Section 5.1; UBVI is able to refine its posterior approximation as it adds components without tuning effort, while the KL-based BVI+ method is difficult to tune well and does not reliably provide better posterior approximations than standard VI. BVI (no stabilization) is not shown, as its component optimizations after the first are degenerate and it reduces to standard VI.

6 Conclusion

This paper developed universal boosting variational inference (UBVI). UBVI optimizes the Hellinger metric, avoiding the degeneracy, tuning, and difficult joint component/weight optimizations of other gradient-based BVI methods, while simplifying fully-corrective weight optimizations. Theoretical guarantees on the convergence of Hellinger distance provide an MCMC-like computation/quality tradeoff, and experimental results demonstrate the advantages over previous variational methods.

7 Acknowledgments

T. Campbell and X. Li are supported by a National Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and an NSERC Discovery Launch Supplement.

[3] Real datasets available online at `http://komarix.org/ac/ds/` and `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`.

References

[1] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov chain Monte Carlo*. CRC Press, 2011.

[2] Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian data analysis*. CRC Press, 3rd edition, 2013.

[3] Michael Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[4] Martin Wainwright and Michael Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

[5] Gareth Roberts and Jeffrey Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

[6] Pierre Jacob, John O'Leary, and Yves Atchadé. Unbiased Markov chain Monte Carlo with couplings. *arXiv:1708.03625*, 2017.

[7] Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, 2015.

[8] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. In *International Conference on Machine Learning*, 2016.

[9] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, 2016.

[10] Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18:1–43, 2017.

[11] Steven Scott, Alexander Blocker, Fernando Bonassi, Hugh Chipman, Edward George, and Robert McCulloch. Bayes and big data: the consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11:78–88, 2016.

[12] Michael Betancourt. The fundamental incompatibility of scalable Hamiltonian Monte Carlo and naive data subsampling. In *International Conference on Machine Learning*, 2015.

[13] Matthew Hoffman, David Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14:1303–1347, 2013.

[14] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18:1–45, 2017.

[15] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *International Conference on Artificial Intelligence and Statistics*, 2014.

[16] Diedrik Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[17] Aılım Güneş Baydin, Barak Pearlmutter, Alexey Radul, and Jeffrey Siskind. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18:1–43, 2018.

[18] Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 2018 (to appear).

[19] Yixin Wang and David Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 0(0):1–15, 2018.

[20] Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. *The Annals of Statistics*, 2018 (to appear).

[21] Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12:2995–3035, 2018.

[22] Guillaume Dehaene and Simon Barthelmé. Expectation propagation in the large data limit. *Journal of the Royal Statistical Society, Series B*, 80(1):199–217, 2018.

[23] Kevin Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, 2012.

[24] Fangjian Guo, Xiangyu Wang, Kai Fan, Tamara Broderick, and David Dunson. Boosting variational inference. In *Advances in Neural Information Processing Systems*, 2016.

[25] Xiangyu Wang. Boosting variational inference: theory and examples. Master's thesis, Duke University, 2016.

[26] Andrew Miller, Nicholas Foti, and Ryan Adams. Variational boosting: iteratively refining posterior approximations. In *International Conference on Machine Learning*, 2017.

[27] Francesco Locatello, Rajiv Khanna, Joydeep Ghosh, and Gunnar Rätsch. Boosting variational inference: an optimization perspective. In *International Conference on Artificial Intelligence and Statistics*, 2018.

[28] Francesco Locatello, Gideon Dresdner, Rajiv Khanna, Isabel Valera, and Gunnar Rätsch. Boosting black box variational inference. In *Advances in Neural Information Processing Systems*, 2018.

[29] Tommi Jaakkola and Michael Jordan. Improving the mean field approximation via the use of mixture distributions. In *Learning in graphical models*, pages 163–173. Springer, 1998.

[30] O. Zobay. Variational Bayesian inference with Gaussian-mixture approximations. *Electronic Journal of Statistics*, 8:355–389, 2014.

[31] Samuel Gershman, Matthew Hoffman, and David Blei. Nonparametric variational inference. In *International Conference on Machine Learning*, 2012.

[32] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, 2015.

[33] Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[34] Robert Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[35] Tong Zhang. Sequential greedy approximation for certain convex optimization problems. *IEEE Transactions on Information Theory*, 49(3):682–691, 2003.

[36] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110, 1956.

[37] Martin Jaggi. Revisiting Frank-Wolfe: projection-free sparse convex optimization. In *International Conference on Machine Learning*, 2013.

[38] Scott Chen, David Donoho, and Michael Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[39] Yoav Freund and Robert Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

[40] Andrew Barron, Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Approximation and learning by greedy algorithms. *The Annals of Statistics*, 36(1):64–94, 2008.

[41] Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In *Uncertainty in Artificial Intelligence*, 2010.

[42] Stéphane Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[43] Sheng Chen, Stephen Billings, and Wan Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50(5):1873–1896, 1989.

[44] Scott Chen, David Donoho, and Michael Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 1999.

[45] Joel Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[46] Trevor Campbell and Tamara Broderick. Bayesian coreset construction via greedy iterative geodesic ascent. In *International Conference on Machine Learning*, 2018.

[47] Kaare Petersen and Michael Pedersen. *The Matrix Cookbook*. Online: https://www.math.uwaterloo.ca/ hwolkowi/matrixcookbook.pdf, 2012.

[48] Charles Lawson and Richard Hanson. *Solving least squares problems*. Society for Industrial and Applied Mathematics, 1995.

[49] Subhashis Ghosal, Jayanta Ghosh, and Aad van der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

[50] Yingzhen Li and Richard Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, 2016.

[51] Cédric Villani. *Optimal transport: old and new*. Springer, 2009.

[52] David Pollard. *A user's guide to probability theory*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 7$^{\text{th}}$ edition, 2002.

[53] Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work? Evaluating variational inference. In *International Conference on Machine Learning*, 2018.

[54] Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018.

[55] Diederik Kingma and Jimmy Ba. ADAM: a method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[56] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, pages 223–242, 2001.

[57] Gábor Szekely and Maria Rizzo. Energy statistics: statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

[58] Matthew Hoffman and Andrew Gelman. The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1351–1381, 2014.

[59] Trevor Campbell and Tamara Broderick. Automated scalable Bayesian inference via Hilbert coresets. *Journal of Machine Learning Research*, 20(15):1–38, 2019.