Variational Refinement for Importance Sampling
Using the Forward Kullback-Leibler Divergence

Ghassen Jerfel[*,1,2,3]   **Serena Wang**[*,1,2]   **Clara Wong-Fannjiang**[2]   **Katherine A. Heller**[1,3]   **Yian Ma**[1,2,4]

Michael I. Jordan[2]

[1]Google Research, Mountain View, CA, USA
[2]University of California Berkeley, Berkeley, CA, USA
[3]Duke University, Durham, NC, USA
[4]University of California San Diego, San Diego, CA, USA

Abstract

Variational Inference (VI) is a popular alternative
to asymptotically exact sampling in Bayesian in-
ference. Its main workhorse is optimization over a
reverse Kullback-Leibler divergence (RKL), which
typically underestimates the tail of the posterior
leading to miscalibration and potential degeneracy.
Importance sampling (IS), on the other hand, is
often used to fine-tune and de-bias the estimates of
approximate Bayesian inference procedures. The
quality of IS crucially depends on the choice of the
proposal distribution. Ideally, the proposal distri-
bution has heavier tails than the target, which is
rarely achievable by minimizing the RKL. We thus
propose a novel combination of optimization and
sampling techniques for approximate Bayesian in-
ference by constructing an IS proposal distribution
through the minimization of a forward KL (FKL)
divergence. This approach guarantees asymptotic
consistency and a fast convergence towards both
the optimal IS estimator and the optimal varia-
tional approximation. We empirically demonstrate
on real data that our method is competitive with
variational boosting and MCMC.

1 INTRODUCTION

Bayesian analysis provides a principled framework to en-
code complex hierarchical structures and prior beliefs in
order to capture posterior uncertainty about latent variables
θ given observed data x via the posterior $p(\theta|x)$. The in-
ferential goal often involves computing expectations over
this posterior distribution, $\mathbb{E}_{\theta \sim p(\theta|x)}[f(\theta)]$, which is typi-
cally accomplished by sampling. Unfortunately, sampling
directly from the posterior is usually intractable. Computing
posterior functionals thus requires approximate inference
methods such as variational inference (VI) [Jordan et al.,

1999, Wainwright and Jordan, 2008], Markov Chain Monte
Carlo (MCMC) [Brooks et al., 2011, Andrieu et al., 2003],
and importance sampling (IS) [Gelman and Meng, 1998],
among others.

Recently, variational inference has grown in popularity be-
cause it recasts the inference problem as an optimization
problem that can leverage recent advances in stochastic
optimization [Bottou, 2010, Hoffman et al., 2013] and auto-
matic differentiation [Maclaurin et al., 2015]. Specifically,
VI poses a tractable family of distributions $\mathcal{Q}$ and minimizes
the reverse Kullback-Leibler divergence (RKL) between q
and p, i.e. $\mathrm{KL}(q||p)$. However, $\mathcal{Q}$ is often misspecified lead-
ing to unknown bias in VI solutions [Blei et al., 2016]. It
is also generally difficult to assess the quality of a VI ap-
proximation on downstream tasks based on the value of the
RKL divergence [Yao et al., 2018, Campbell and Li, 2019,
Rainforth et al., 2018, Huggins et al., 2020]. This motivates
the use of importance sampling (IS) to de-bias posterior
summaries regardless of the misspecification of $\mathcal{Q}$ [Gelman
and Meng, 1998, Owen, 2013]. Unlike the RKL, the perfor-
mance of q as an IS proposal is indicative of its quality as an
approximation of p on downstream tasks [Yao et al., 2018].

This workflow capitalizes on complementary strengths: the
computational efficiency of VI can sidestep the challenge
of selecting a proposal distribution for IS which, in turn,
ensures the consistency of the refined posterior expectation.
However, RKL minimization typically results in light tails
which can cause instability for importance sampling [Dieng
et al., 2017, Yao et al., 2018]. On the other hand, the forward
KL divergence (FKL or $\mathrm{KL}(p||q)$) is known to control the
estimation error of importance sampling [Chatterjee and
Diaconis, 2018] but is rarely used due to its intractability.

In this paper, we propose to replace reverse KL with forward
KL as the variational objective whose minimization yields
an optimal IS proposal distribution. We make four distinct
contributions in this vein:

1. We derive a self-normalized importance sampling esti-
   mate for the intractable FKL divergence.

2. We demonstrate how FKL-based boosting can combine IS and VI for multimodal target distributions.

3. We show that FKL boosting is guaranteed to converge at a rate of $O(\frac{1}{K})$, where K is the number of boosting iterations, to the best approximation from a family of mixture distributions. This immediately guarantees convergence to the optimal proposal distribution as per the results of [Chatterjee and Diaconis, 2018].

4. We demonstrate empirically that our approach is competitive with state-of-the-art VI and Hamiltonian Monte Carlo on regression tasks over real datasets using Bayesian neural networks (BNNs) and Bayesian linear regression (BLR).

Our proposed algorithm is thus a principled inference technique, with a well-defined computation-quality trade-off, that can be used independently or as a refining step to correct for the error in a given approximation.
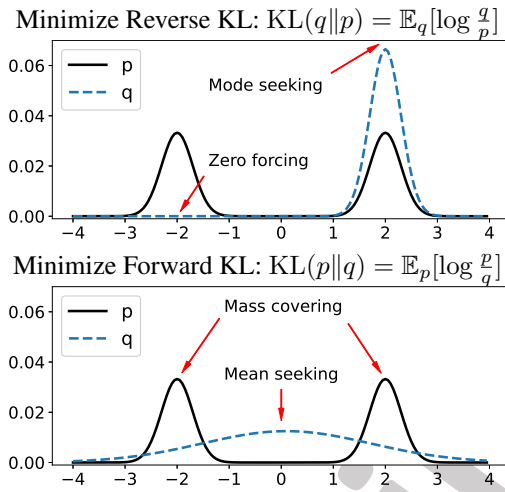

Figure 1: Results of minimizing RKL vs. FKL for a Gaussian variational approximation on a bimodal target.

2 BACKGROUND

Let θ denote the variable of interest with probability density p, and let f denote a function of θ. Our goal is to estimate an expectation

$$\mathbb{E}_{\theta \sim p(\theta|x)}\left[f(\theta)\right]. \tag{1}$$

In Bayesian inference, θ generally represents a latent variable to be integrated over in the posterior distribution, $p(\theta|x)$, conditioned on the observed data x.

If one can draw S samples $\{\theta_s\}_{s=1}^S$ from the target distribution $p(\theta|x)$ then Eq. (1) can be estimated by simple Monte Carlo integration. However, $p(\theta|x)$ is typically only known up to a normalization constant and thus cannot be readily sampled from. Instead, given samples from an approximation q of the posterior, we can estimate the expectation as:

$$\mathbb{E}_{p(\theta|x)}[f(\theta)] \approx \frac{\sum_{i=1}^S f(\theta_s)w_s}{\sum_{i=1}^S w_s}. \tag{2}$$

If the samples are weighted equally (i.e., $w_s = 1$), Eq. (2) is equivalent to the VI estimate, which has low variance but can be biased and inconsistent [Owen, 2013].

If instead we weigh by the importance ratios, $w_s = \frac{p(\theta_s|x)}{q(\theta_s)}$, we recover the IS estimate:

$$\mathbb{E}_{\theta \sim p(\theta|x)}\left[f(\theta)\right] = \mathbb{E}_{\theta \sim q(\theta)}\left[\frac{p(\theta|x)}{q(\theta)}f(\theta)\right], \tag{3}$$

which is consistent (bias=$\mathcal{O}(1/S)$) but with potentially large or infinite variance [Owen, 2013].

2.1 VI WITH REVERSE KL MINIMIZATION

Variational inference posits a family $\mathcal{Q}$ of distributions that are easy to evaluate or sample from, and defines the variational approximation $q^* \in \mathcal{Q}$ as the distribution that minimizes the reverse Kullback-Leibler (RKL) divergence to the posterior p: $q^* = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \quad \mathrm{KL}(q\|p)$.

Unless the choice of $\mathcal{Q}$ specifically includes the target distribution, minimizing the reverse KL divergences leads to a q^* that underestimates the target covariance. The decomposition of this divergence sheds light on the cause:

$$\mathrm{KL}(q\|p) = \mathbb{E}_q[\log q] - \mathbb{E}_q[\log p] \tag{4}$$

The first term in Eq. (4) is the entropy of q whose penalization is known to cause light tails. Furthermore, the second term is minimized when $q = 0$ for $p > 0$, leading to *zero forcing* or over-pruning [Higgins et al., 2016]. The RKL divergence thus favors a single mode (*mode seeking*), is biased towards avoiding false positives, and poses difficulties when approximating heavy-tailed [Guo et al., 2016, Li and Turner, 2016, Dieng et al., 2017] or multimodal targets [Miller et al., 2017]. This effect is illustrated in Figure 1.

2.2 IMPORTANCE SAMPLING AND FKL

Due to the unknown bias of an RKL-minimizing q^*, refining through importance weighting is recommended to de-bias the estimate of Eq. (1) [Yao et al., 2018, Vehtari et al., 2015]. However, the light tails of this q^* can lead to high or even infinite variance for an IS estimator which limits the efficiency of q^* as an IS proposal distribution.

As an alternative to RKL, minimizing the forward KL divergence can mitigate the issue of covariance underestimation and tail undersampling. Consider the decomposition of the forward KL divergence:

$$\mathrm{KL}(p\|q) = \mathbb{E}_p[\log p] - \mathbb{E}_p[\log q]. \tag{5}$$

In Eq. (5), the cross-entropy is minimized by setting $q > 0$ whenever $p > 0$, which leads to *mass covering* as illustrated in Figure 1. With better tail coverage, an FKL-minimizing q^* can yield an IS estimate (Eq. (3)) with lower variance than an approximation generated by RKL. Indeed, Chatterjee and Diaconis [2018] demonstrate that the variance of an importance sampling estimate scales as $O\left(\frac{e^{\mathrm{KL}(p\|q)}}{\sqrt{S}}\right)$, where the number of samples S required for IS to provide accurate mean estimates scales exponentially with FKL.

3 RELATED WORK

We present related work on improving estimates of the expectation over intractable target distributions (e.g. high dimensional, heavy tailed, or multimodal).

3.1 VARIATIONAL INFERENCE DIVERGENCES

Prior work has addressed the covariance underestimation and light tails pathologies of reverse KL minimization while seeking to improve the quality of the approximation as an IS proposal through the minimization of alternative divergences such as (reversed) Renyi-α [Li and Turner, 2016, Hernández-González and Cerquides, 2020], Chi-square ($\alpha = 2$) [Dieng et al., 2017], or Hellinger ($\alpha = 1/2$) [Campbell and Li, 2019] divergences. FKL can be seen as a special case of α divergences when $\alpha \to 1$ which is not considered in any of these prior VI works.

3.2 FKL FOR APPROXIMATE INFERENCE

While the forward KL's computational inconvenience has limited its use for variational inference, inference techniques such as Belief Propagation (BP) and Expectation Propagation (EP) can be regarded as performing FKL minimization locally [Minka et al., 2005]: KL is minimized one data partition at a time instead of globally as in VI. Minka [2001] demonstrates that this local minimization procedure is not guaranteed to converge and may not result in representative posteriors. Another set of techniques that utilize variants of the FKL divergence includes reweighted wake-sleep [Bornschein and Bengio, 2014] which alternates minimizing an approximation of FKL during the sleep phase while minimizing an approximation of RKL during the wake phase. However, this is known to lead to a biased estimator [Bornschein and Bengio, 2014].

3.3 VARIATIONAL BOOSTING

Variational boosting (VB) [Miller et al., 2017, Guo et al., 2016, Locatello et al., 2018a, Jerfel, 2017] has been suggested in various forms to address the multimodality challenge for variational inference. Variational boosting posits a family of mixture distributions $\mathcal{Q}_k$:

$$\mathcal{Q}_k = \left\{ q : q(\theta) = \sum_{i=1}^{k} \lambda_i f_i(\theta), \lambda \in \Delta_k \right\}, \quad (6)$$

and sequentially constructs a variational mixture approximation by adding and re-weighting one (typically Gaussian) mixture component at a time to minimize the KL objective:

$$\{\mu_i\}_{i=1}^{k}, \{\Sigma_i\}_{i=1}^{k}, \{\lambda_i\}_{i=1}^{k}$$
$$\leftarrow \underset{\mu,\Sigma,\lambda}{\operatorname{argmin}} \, \mathrm{KL}\left(\sum_{i=1}^{k} \lambda_i f_i(\theta; \mu_i, \Sigma_i) \middle\| p(\theta|x) \right). \quad (7)$$

This form of reverse KL-based boosting is known to struggle with degeneracy where the optimization at certain boosting iterations can lead to point-mass mixture components [Campbell and Li, 2019]. Ad-hoc regularization techniques are often needed in practice [Locatello et al., 2018a], but are not necessarily sufficient [Campbell and Li, 2019]. To address this pathology, Campbell and Li [2019] proposed a boosting algorithm, based on the Hellinger divergence, which does not not guarantee scalability with dimensions.

3.4 ADAPTIVE IMPORTANCE SAMPLING

Adaptive IS (AIS) methods such as Adaptive multiple IS [Cornuet et al., 2012] and incremental mixture IS [Raftery and Bao, 2010] are designed for multimodal targets. However, none of the existing works directly optimize for the FKL divergence which controls the worst case IS estimation error. For example, Cappé et al. [2008] minimize an entropy criterion whereas Douc et al. [2007] minimize the empirical variance of the importance weights, which does not necessarily correlate with the quality of IS estimation [Vehtari et al., 2015].

3.5 COMBINING IS AND VI

As outlined in Table 1, VI and VB suffer from covariance underestimation, and can struggle to approximate heavy-tailed distributions [Blei et al., 2016]. AIS, on the other hand, can approximate heavy-tailed targets but cannot scale efficiently in dimensionality [Owen, 2013]. A combination of these two lines of research may benefit from their complementary strengths while sidestepping shared weaknesses (e.g., multimodality). However, it is often difficult to combine optimization-based and sampling-based inference techniques. This is because sampling methods such as MCMC define the approximate distribution implicitly such that its density cannot be evaluated. This has driven the development of alternatives to the KL divergence such as the variational contrastive divergence [Ruiz and Titsias, 2019]. However, we are not aware of similar work for IS that leverages the computational efficiency of VI through a unifying loss such as the FKL divergence.

Other prior proposals combining IS and VI [Domke and Sheldon, 2018] have focused on minimizing the RKL, and thus do not inherently capture heavier tails of the target distribution. Prangle [2019] recently presented concurrent work on combining IS and VI. However, their method relies on normalizing flows for constructing the proposal distribution such that it does not guarantee a multimodal approximation. Ramos-López et al. [2018] uses a stream of IS weights to fit parameters for Gaussian mixture posteriors to minimize the FKL divergence. However, unlike our work, they assume access to normalized importance weights instead of samples.

Failure Mode	VI	IS	VB	AIS	Ours
Multimodality	✗	✗	✓	✓	✓
Heavy tails	✗	✓	✗	✓	✓
Cov estimation	✗	✓	✗	✓	✓
High dimensions	✓	✗	✓	✗	✓*

Table 1: Comparing approximate inference techniques.

4 METHODOLOGY

We develop our novel approach to integrate variational inference and importance sampling using the forward KL divergence with a focus on multimodal targets.

Note that we assume the target density p and the approximation q share the same support which can be $\mathbb{R}^d$ or a subset thereof. This guarantees that p is absolutely-continuous with respect to q (noted as $p \ll q$) and vice-versa which is necessary for the definition of the KL divergence.

4.1 FORWARD KL VARIATIONAL APPROXIMATION FOR IS

A theoretically reasonable desire is for the proposal distribution to minimize the forward KL divergence. However, we cannot compute FKL exactly for unnormalized target distributions. This stems from the expectation under the target p in Eq. (5).

By contrast, the reverse KL takes the expectation in Eq. (4)) under a normalized approximation q. This can be seen as a tractable approximation to FKL with unknown bias [Yao et al., 2018]. In fact, for misspecified choices of $\mathcal{Q}$ (i.e. when $\mathrm{KL}(q^*\|p) \neq 0$), Campbell and Li [2019] demonstrate that minimizing the RKL is not guaranteed to minimize the FKL divergence. Furthermore, the light tails of the RKL minimization solution renders it inadequate for IS.

Alternatively, towards deriving a consistent approximation of the forward KL divergence, we rearrange densities inside

the expectation as follows:

$$\mathrm{KL}(p\|q) = \mathbb{E}_p\left[\log\left(\frac{p}{q}\right)\right] = \mathbb{E}_q\left[\frac{p}{q}\log\left(\frac{p}{q}\right)\right]. \quad (8)$$

FKL can then be approximated through self-normalized importance sampling (SNIS) [Murphy, 2012] which is known to be consistent:

$$\theta_s \sim q(\theta) \quad r_s = \frac{p(\theta_s|x)}{q(\theta_s)} \quad w_s = \frac{r_s}{\sum_{s=1}^S r_s}$$

$$\mathrm{KL}(p\|q) = \sum_{s=1}^S w_s \cdot \log\left(\frac{p(\theta_s|x)}{q(\theta_s)}\right). \quad (9)$$

The SNIS estimation can have arbitrarily high variance depending on q. Gradients of Eq. (8) with respect to the parameters of q_i can also have high variance since we optimize over the same distribution from which samples are drawn. For certain distributions, the reparametrization trick can reduce this variance [Kingma and Welling, 2013]. Furthermore, the sequential setting described in section 4.3 contributes to reducing the gradient variance.

4.2 FORWARD KL BOOSTING

We sequentially construct a proposal mixture distribution q that is both easy to sample from (lower SNIS bias) and minimizes the FKL objective for efficient IS estimation.

Given the computational convenience of Gaussian distributions, we set $\mathcal{Q}$ to be the family of Gaussian mixtures in Section 7 such that the proposal at the Kth boosting iteration is $q_K(\theta) := \sum_{i=1}^K \lambda_i f_i(\theta; \phi_i)$ where $f_i = \mathcal{N}(\mu_i, \Sigma_i)$.

Our framework also supports mixtures of heavy-tailed distributions which are desirable for adaptive IS [Geweke, 1989].

At each boosting iteration, we minimize the FKL between the mixture q_i and the target p by fitting a new component f_i and a mixture weight $\lambda_i = \gamma$ while holding the parameters of previously-learned mixture components fixed:

$$\operatorname*{argmin}_{q_i \in \mathcal{Q}} \mathrm{KL}(p\|q_i) = \operatorname*{argmin}_{f_i, \gamma} \mathrm{KL}\left(p\|\gamma f_i + (1-\gamma)q_{i-1}\right)$$

This is known to be more efficient and stable than the joint optimization of all mixture components at each iteration [Locatello et al., 2018a]. The mixture weights for the fixed mixture components $\{\lambda_j\}_{j=1}^{i-1}$ are re-scaled by $(1-\gamma)$ and can be further adjusted with a fully-corrective weight search [Locatello et al., 2018b] using the gradient derived in Eq. (5). Fitting the mixture weights is a convex problem that can benefit from higher-order optimizers such as Newton's. We provide an alternative approach to building q_K using the remainder distribution in Appendix C.

4.3 LOWER VARIANCE SNIS WITH BOOSTING

The main computational concern in estimating the FKL divergence with SNIS is the variance of importance sampling in Eq. (8). However, the sequential construction described above enables further reformulations of the SNIS approximation which can lead to a lower variance than the naive approximation of Eq. (8). In fact, the global objective can be re-written as:

$$\underset{q_i \in \mathcal{Q}}{\arg\min} \quad \mathrm{KL}(p\|q_i) = \underset{q_i \in \mathcal{Q}}{\arg\min} \quad \mathbb{E}_p \left[\log \frac{p}{q_i} \right] \quad (10)$$

$$= \underset{f_i,\gamma}{\arg\min} \quad \mathbb{E}_{q_i} \left[\frac{p}{q_i} \log \frac{p}{\gamma f_i + (1-\gamma)q_{i-1}} \right] \quad (11)$$

$$= \underset{f_i,\gamma}{\arg\min} \quad \mathbb{E}_{q_{i-1}} \left[\frac{p}{q_{i-1}} \log \frac{p}{\gamma f_i + (1-\gamma)q_{i-1}} \right] \quad (12)$$

Eq. (12), which is only possible in a sequential setting such as ours, reduces the gradient variance since the component being estimated, f_i, is independent of the distribution q_{i-1} from which samples are drawn to approximate the FKL. Furthermore, we can draw a large number of samples from q_{i-1} a single time at the start of each boosting iteration.

The SNIS approximation of Eq. (12) given samples θ_s from q_{i-1} can then be computed as:

$$\sum_{s=1}^{S} w_s [\log p(\theta_s) - \log(\lambda f_i(\theta_s) + (1-\lambda)q_{i-1}(\theta_s))],$$

where w_s are computed as in Eq. (9).

Trade-off of sampling q_i or q_{i-1}: While q_i is expected to handle well-separated modes better, in our experiments (7 we found that using q_{i-1} is sufficient for capturing distant modes (e.g. Fig. 4) while reducing the SNIS variance. Variance reduction was especially crucial on higher dimensional applications such as Bayesian NNs in Table. 3.

4.4 INITIALIZE WITH RKL, REFINE WITH FKL

For the first boosting iteration, we do not have an existing approximation to sample from. Instead, Eq. (12) can be re-written as follows:

$$\underset{f_i,\gamma}{\arg\min} \, \mathbb{E}_{f_i} \left[\frac{p}{f_i} \log \frac{p}{\gamma f_i + (1-\gamma)q_{i-1}} \right].$$

This exacerbates the sensitivity to the initialization of f_i. In fact, a sharply peaked initialization centered could limit the SNIS estimate from properly capturing the tail behavior. Therefore, a diffuse initialization is likely to be beneficial.

An even more practical initialization would use RKL-based VI to identify the mode of the target and provide a computationally efficient approximation that can be refined by

FKL boosting in later iterations. As such, this FKL-boosting workflow can be considered as general framework for the iterative refinement of any given posterior approximation to be used for the estimation of expectations of interest. This best combines the strengths of VI and IS as the first iteration of RKL minimization can reduce the variance of the SNIS approximation of FKL for the following iterations.

5 ANALYSIS

We provide theoretical analysis of the proposed method of performing importance sampling with a proposal distribution constructed from FKL-based boosting.

5.1 FKL CONTROLS MOMENT ESTIMATION ERROR

Minimizing the FKL implicitly minimizes the error in posterior probabilities and moments via its control on total variational (through Pinsker's inequality[Tsybakov, 2008]) and l-Wasserstein as follows [Bolley and Villani, 2005]: Assume that the probability density q is n-exponentially integrable. Then for target distribution p such that $p \ll q$:

$$W_n(q,p) \le C_n^{EI}(q) \left[\mathrm{KL}(p\|q)^{\frac{1}{n}} + \frac{\mathrm{KL}(p\|q)^{\frac{1}{2n}}}{2} \right], \quad (13)$$

where

$$C_n^{EI}(q) = \inf_{x_0 \in \mathbb{R}^d, \alpha > 0} \left(\frac{3}{\alpha} + \frac{2}{\alpha} \log \int e^{\alpha\|x-x_0\|^n} q(x) \mathrm{d}x \right).$$

This implies the convergence of the first n moments.

Note that because of the symmetry of Wasserstein distances, we can switch probability densities p and q in the above inequalities and obtain bounds in terms of RKL. However, that would incur the n-exponential integrability condition on the target probability p, which boils down to generalized sub-gaussianity of its tail.

The n-exponential integrability assumption is not required of the target density p in the case of Eq. (13). Instead, it is only required of the family of variational approximations q which is easier to enforce and verify (and is automatically satisfied by the mixture of Gaussians). A smaller constant C_n^{EI} can also be achieved by the same reasoning.

5.2 FKL BOOSTING CONVERGES AT $O(1/K)$

Assuming $p \ll q$, which can be verified by the design of $\mathcal{Q}$, the functional gradient of the forward KL divergence is derived in Appendix B.5 as $\frac{\delta KL(p\|q)}{\delta q} = -\frac{p}{q}$. The convexity of $\mathrm{KL}(p\|q)$ in q is well established in the literature (proven with the log-sum inequality). Furthermore, we show in Appendix 5.2 that the FKL functional is also β-smooth in q

where β depends on the range of the values that the density q can take. This requires bounding q away from zero and from above which is typical in prior theoretical work [Guo et al., 2016] and aligns with practice as it can translate to a bounded parameter space for a given family of distributions.

For a convex and strongly smooth functional, the greedy sequential approximation framework of Zhang [2003] provides an asymptotic guarantee for the convergence to a target distribution in the convex hull of a given base family such that the approximation error at the Kth iteration is $\mathrm{KL}(p \| q_K) = \mathrm{KL}\left(p \left\| \sum_{i=1}^{K} \lambda_i f_i \right.\right) = O(1/K)$. This framework does not require each iteration to exactly solve for the optimal mixture component which can be difficult for the non-convex optimization sub-problems.

5.3 COMPUTATION-QUALITY TRADE-OFF

While our iterative algorithm is guaranteed to converge asymptotically to the optimal proposal distribution, we can identify three sources of approximation error: the variational inference error, the SNIS approximation bias, and the greedy sequential approximation error which depend on the number of VI iterations, IS samples, and boosting iterations, respectively. These three errors thus finely control the compute-quality trade-off of our framework. We analyze these tradeoffs in more depth through simulation experiments below.

5.4 THE DEPENDENCE OF THE IS PROPOSAL ON THE INTEGRAND

Similarly to prior empirical [Owen, 2013] and theoretical [Chatterjee and Diaconis, 2018] works, we do not address any assumptions about the function f being integrated. However, the optimal proposal $q^* \propto |f| \cdot p$ when p is normalizable and $q^* \propto |f - I| \cdot p$ otherwise [Kahn and Marshall, 1953]. Nonetheless, we only require that the integrand f does not contain any singularities and shares the same support as p and q. In this case, a simple rearrangement of the terms inside the expectation (1) would imply that q should approximate $f \cdot p$ instead of p where f is the integrand. Therefore, characteristics of f have no effect on our asymptotic guarantees or general methodology.

6 SIMULATION EXPERIMENTS

Using two illustrative simulation experiments, we provide further intuition for the behavior and performance of the proposed methodology using the FKL.

6.1 SIMULATION 1: CAUCHY

First, we demonstrate the aforementioned computation-quality trade-offs using a standard Cauchy target distribution. For intuition on the behavior of boosting using the FKL, Figure 3 illustrates the density plots of the target p and the boosting approximation q after various iterations.

From previous sections, we observe a trade-off between sample complexity of SNIS and the optimization complexity of variational boosting. Inclusion of more mixture components K and more accurate optimization in the variational boosting steps can save exponentially many samples in SNIS. However, there is a diminishing gain in increasing K. We demonstrate in Fig. 2 this effect: both FKL boosting and RKL boosting decreases forward KL divergence as more variational components are added. The decrease slows down significantly after inclusion of 3 mixture components. We therefore introduce in the experiments up to 3 mixture components and select the best performance on validation data set. From Fig. 2 and the experimental results, we observe uniform improvements of FKL over RKL methods.
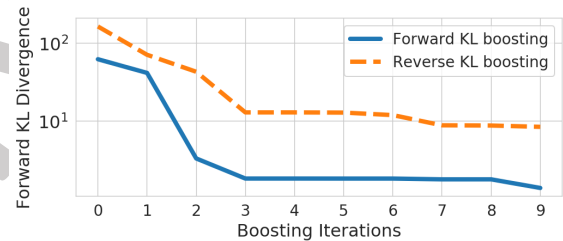

Figure 2: A comparison in terms of the FKL divergence to a Cauchy target distribution over the course of variational boosting using the FKL and RKL divergences.

6.2 SIMULATION 2: WELL-SEPARATED MODES

We next demonstrate the performance of boosting with the FKL on a distribution with a large number of well-separated modes. We set as the target distribution a 2-dimensional Gaussian mixture model (GMM) with 20 components, previously used by Ma et al. [2019].

The log-residual in Fig. 4 demonstrates the sequential improvement of our approximation. Moment estimation results for this experiment can be found in Fig. 1 in the Appendix. Overall, boosting with the FKL effectively improves both moment estimation and the actual FKL, and outperforms the RKL as the number of boosting iterations increases.

7 REAL DATA EXPERIMENTS

We evaluate the performance of the proposed method when applied to Bayesian linear regression (BLR) and Bayesian
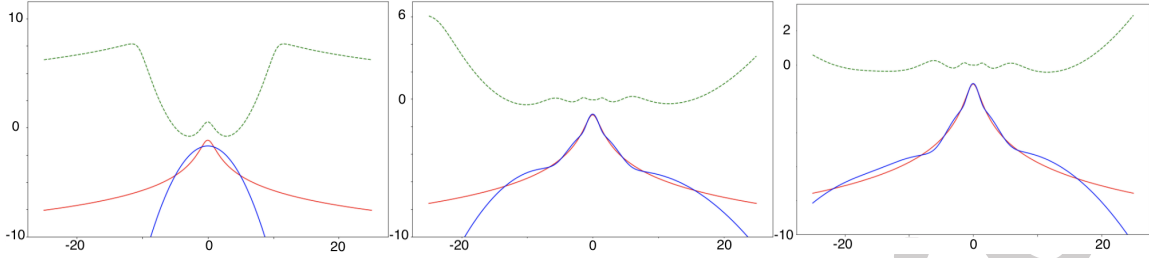
Figure 3: Log density plots of the result $\log q_k$ (blue, solid) of FKL boosting on the Cauchy target ($\log p$) (red, solid) . The log-residual ($\log p/q$)(green, dashed) is indicative of the IS variance (from left to right: boosting iterations 1, 5, and 10).
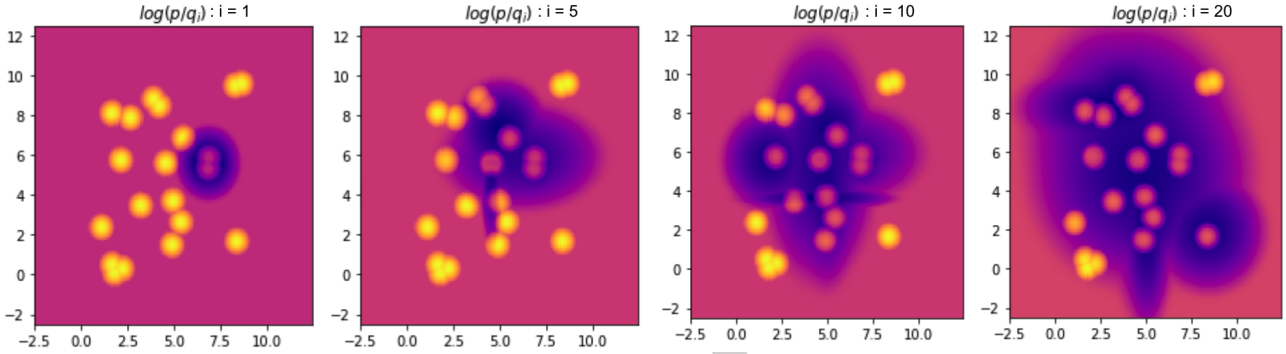

Figure 4: Log residual ($\log p/q_k$) plots of for FKL boosting on a 2-dimensional GMM of 20 components [Ma et al., 2019].

neural networks (BNNs) using a Gaussian prior and a heavy tailed prior. We use four datasets from UCI [Dua and Graff, 2017] (Table 1 in Appendix E). We split each dataset into twenty randomly drawn 90%/10% train/test splits, which we denote $\mathcal{D}_{\text{train}} = \{x_i, y_i\}_{i=1}^{N_{\text{train}}}$ and $\mathcal{D}_{\text{test}} = \{x_i, y_i\}_{i=1}^{N_{\text{test}}}$, with input x_i and output y_i. We report the mean and std. dev. of results over all splits.

We demonstrate our proposed method with $K = 1, 2$, and 3. We refer to fitting a single Gaussian, or $K = 1$, as FKL VI. For runs with more than one boosting iteration, we initialize the first component using the RKL, and optimize subsequent iterations using the FKL (as described in Section 4.4).

Optimization details: We use the ADAM optimizer [Kingma and Ba, 2014] for each boosting iteration with a fixed learning rate and compute gradients based on a fixed number of samples using Autograd [Maclaurin et al., 2015]. At the end of each boosting iteration, mixture weights are fully re-optimized using simplex-projected gradient descent [Bubeck, 2014] based on the analytical gradient in Eq. (5). Details about practical considerations and hyperparameters can be found in Appendix E, and we include our code with the submission.

Parameter initialization: We follow standard practice with similar experiments (see e.g. [Miller et al., 2017]) where the means of each component are initialized at zero and the diagonal elements of the initial covariance matrix are drawn from $\mathcal{N}(0, \sigma^2)$ (σ is tuned as a hyperparameter, see

Appendix E.2). At the start of each boosting iteration, we further apply an initialization heuristic which approximates the mode of the residual by gradient descent on the remainder density. This requires a single sample per gradient step and is run for 400 steps. Note that this initialization is significantly simpler and less compute-intensive than prior approaches which include a weighted EM algorithm [Miller et al., 2017] and a random search over $10,000$ samples or more [Campbell and Li, 2019].

Comparisons: We compare our approach to variational inference (RKL VI) and variational boosting with 2 and 3 boosting iterations (RKL VB) [Miller et al., 2017]. For the comparison to RKL VI and RKL VB, we use the same parametrization, initialization, and optimization techniques as for FKL VI and FKL VB. This might lead to discrepancies compared to the published results [Miller et al., 2017]; however, keeping these details consistent better disentangles the effect of the RKL vs. FKL optimization.

We also compare to directly sampling from the target distribution using Hamiltonian Monte Carlo (HMC) [Neal, 2011], implemented using the TensorFlow Probability library [Dillon et al., 2017]. We additionally ran 3 HMC chains in parallel and averaged the results, similar to [Hoffman and Ma, 2020]. Results were comparable between 1 and 3 HMC chains, and we include the results for 3 chains in Appendix E.

Method	Wine ($d = 14$)	Boston ($d = 16$)	Concrete ($d = 11$)	Power ($d = 7$)
HMC	-1.002 ($\pm$ 0.012)	-2.923 ($\pm$ 0.035)	-3.780 ($\pm$ 0.013)	-2.923 ($\pm$ 0.006)
RKL VI	-1.013 ($\pm$ 0.013)	-2.947 ($\pm$ 0.035)	-3.798 ($\pm$ 0.013)	-2.921 ($\pm$ 0.006)
RKL VB 2	-1.014 ($\pm$ 0.011)	-2.924 ($\pm$ 0.033)	-3.787 ($\pm$ 0.014)	-3.067 ($\pm$ 0.003)
RKL VB 3	-1.007 ($\pm$ 0.012)	-2.945 ($\pm$ 0.039)	-3.800 ($\pm$ 0.015)	-3.073 ($\pm$ 0.003)
FKL VI	**-0.998** ($\pm$ 0.012)	-2.905 ($\pm$ 0.036)	-3.775 ($\pm$ 0.014)	**-2.920** ($\pm$ 0.005)
FKL VB 2	**-0.998** ($\pm$ 0.012)	-2.906 ($\pm$ 0.036)	**-3.762** ($\pm$ 0.013)	-2.921 ($\pm$ 0.006)
FKL VB 3	**-0.998** ($\pm$ 0.013)	**-2.904** ($\pm$ 0.036)	**-3.762** ($\pm$ 0.013)	-2.921 ($\pm$ 0.006)

Table 2: Predictive log probabilities on test for BLR with Gaussian prior (mean $\pm$ standard error over 20 train/test splits).

7.1 BLR AND BNNS WITH GAUSSIAN PRIORS

We follow the experimental setup of [Miller et al., 2017] for both BNNs and BLR. We place a Gaussian prior over each weight in the model, and an inverse Gamma prior on the variances:

$$\alpha \sim \text{Gamma}(1, 0.1); \quad \tau \sim \text{Gamma}(1, 0.1);$$
$$w_i \sim \mathcal{N}(0, 1/\alpha); \quad y|x, w, \tau \sim \mathcal{N}(\phi(x, w), 1/\tau),$$

where w is the set of weights, and $\phi(x, w)$ is either a linear function of x (BLR) or a multi-layer perception (BNN). For our BNNs, we set ϕ to be a one-hidden layer neural network with 50 hidden units and ReLU activation function, as done by [Miller et al., 2017, Hernández-Lobato and Adams, 2015]. The full set of parameters that we sample is $\theta = (w, \alpha, \tau)$. We use the posterior predictive distribution to compute the distribution for a given new input x:

$$p(y|x, \mathcal{D}_{\text{train}}) = \int p(y|x, \theta) p(\theta|\mathcal{D}_{\text{train}}) d\theta. \quad (14)$$

We use importance sampling to estimate this posterior predictive distribution given S samples $\theta_s \sim q(\theta)$:

$$p(y|x, \mathcal{D}_{\text{train}}) \approx \frac{1}{S} \sum_{i=1}^{S} \frac{p(\theta_s|\mathcal{D}_{\text{train}})}{q(\theta_s)} p(y|x, \theta_s), \quad (15)$$

where $q(\theta)$ is the proposal distribution fit to $p(\theta|\mathcal{D}_{\text{train}})$ using either forward KL refinement (our method, FKL VB) or reverse KL refinement (RKL VB, [Miller et al., 2017]).

Note that Miller et al. [2017] does not use importance sampling to estimate the posterior predictive distribution. We add importance weights here as an ablation to limit our analysis to the difference between FKL and RKL optimization. For completness, we report the estimates without IS using RKL VB in Appendix E. For comparing to HMC, we do not use importance sampling due to lack of an explicit density function, and instead compute (Eq. 14) by averaging over direct HMC samples from the posterior distribution $p(\theta|\mathcal{D}_{\text{train}})$ (Eq. (13) in Appendix E).

As our final evaluation metric, we report the average predictive log probabilities on held-out test data:

$$\frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{x,y \in \mathcal{D}_{\text{test}}} \log p(y|x, \mathcal{D}_{\text{train}}). \quad (16)$$

7.2 BLR WITH HEAVY TAILED PRIORS

In addition to the Gaussian prior, we also perform Bayesian linear regression with a heavy tailed prior. Following [Campbell and Li, 2019] we place a $\mathcal{T}_2$ prior on the weights. We use the same inverse Gamma prior on the variance:

$$\tau \sim \text{Gamma}(1, 0.1); \quad w \sim \mathcal{T}_2(0, A^T A);$$
$$y|x, w, \tau \sim \mathcal{N}(\phi(x, w), 1/\tau),$$

where A is fixed, and each entry is drawn i.i.d. before the optimization process: $A_{ij} \sim \mathcal{N}(0, 1)$. For these BLR experiments, $\phi(x, w)$ is a linear function of x with weight parameters w. The full set of parameters that we sample is $\theta = (w, \tau)$. We estimate the same posterior predictive distribution in Eq. (14) using IS in Eq. (15), and report the average predictive log probabilities from Eq. (16).

8 DISCUSSION

Tables 2, 3, and 4 present the results on the UCI datasets for BLR with a Gaussian prior, BNNs with a Gaussian prior, and BLR with a heavy tailed prior, respectively. The lowest mean predictive log probability is highlighted in bold.

Minimizing the FKL divergence outperforms RKL across all four datasets and three experimental settings. This demonstrates the inadequacy of RKL-based VI for the construction of IS proposals. This should incentivize wider adoption of the FKL divergence, especially when the downstream task extends beyond simple prediction and requires a calibrated estimation of the posterior predictive distribution. FKL outperforming HMC, the gold standard of Bayesian inference, on three of the four datasets across all settings is another promising result.

However, we observe a decay in the held-out log-likelihood as more components are added for certain datasets (e.g. FKL VB with BNN on Boston). This is consistent with prior variational boosting results on the same datasets [Miller et al., 2017] and typically signals an over-fitting problem. Therefore, it is worth emphasizing that the convergence analysis of Section 5 is limited to optimization guarantees and does not extend to learning or generalization guarantees.

Method	Wine ($d = 653$)	Boston ($d = 753$)	Concrete ($d = 503$)	Power ($d = 303$)
HMC	-0.990 ($\pm$ 0.014)	-2.709 ($\pm$ 0.101)	-3.281 ($\pm$ 0.017)	**-2.817** ($\pm$ 0.007)
RKL VI	-0.993 ($\pm$ 0.014)	-2.858 ($\pm$ 0.020)	-3.230 ($\pm$ 0.015)	-2.851 ($\pm$ 0.008)
RKL VB 2	-0.990 ($\pm$ 0.015)	-2.832 ($\pm$ 0.018)	-3.253 ($\pm$ 0.015)	-2.945 ($\pm$ 0.009)
RKL VB 3	-0.981 ($\pm$ 0.012)	-2.744 ($\pm$ 0.011)	-3.255 ($\pm$ 0.017)	-3.002 ($\pm$ 0.012)
FKL VI	-0.991 ($\pm$ 0.015)	**-2.677** ($\pm$ 0.011)	-3.328 ($\pm$ 0.019)	-2.872 ($\pm$ 0.012)
FKL VB 2	-0.979 ($\pm$ 0.017)	-2.779 ($\pm$ 0.012)	-3.193 ($\pm$ 0.016)	-2.870 ($\pm$ 0.008)
FKL VB 3	**-0.967** ($\pm$ 0.014)	-2.801 ($\pm$ 0.012)	**-3.192** ($\pm$ 0.016)	-2.851 ($\pm$ 0.009)

Table 3: Predictive log probabilities on test for BNNs with Gaussian prior (mean $\pm$ standard error over 20 train/test splits).

Method	Wine ($d = 13$)	Boston ($d = 15$)	Concrete ($d = 10$)	Power ($d = 6$)
HMC	-1.004 ($\pm$ 0.012)	-2.962 ($\pm$ 0.033)	-3.808 ($\pm$ 0.026)	**-2.916** ($\pm$ 0.005)
RKL VI	-1.011 ($\pm$ 0.013)	-2.924 ($\pm$ 0.037)	-3.789 ($\pm$ 0.013)	-2.924 ($\pm$ 0.006)
RKL VB 2	-1.007 ($\pm$ 0.014)	-2.944 ($\pm$ 0.031)	-3.788 ($\pm$ 0.012)	-3.028 ($\pm$ 0.003)
RKL VB 3	-1.008 ($\pm$ 0.012)	-2.940 ($\pm$ 0.035)	-3.796 ($\pm$ 0.014)	-3.005 ($\pm$ 0.004)
FKL VI	-0.993 ($\pm$ 0.013)	**-2.904** ($\pm$ 0.036)	-3.775 ($\pm$ 0.014)	-2.940 ($\pm$ 0.005)
FKL VB 2	**-0.973** ($\pm$ 0.009)	-2.907 ($\pm$ 0.035)	**-3.773** ($\pm$ 0.015)	-2.921 ($\pm$ 0.006)
FKL VB 3	-0.975 ($\pm$ 0.009)	-2.906 ($\pm$ 0.036)	-3.774 ($\pm$ 0.015)	-2.922 ($\pm$ 0.006)

Table 4: Predictive log probabilities on test for BLR with heavy tailed prior (mean $\pm$ standard error over 20 train/test splits).

Computational considerations One main advantage of RKL methods is the low computational overhead, especially as compared to MCMC methods. In our experiments we observed that, even in high dimensions, there do not seem to be significant computational differences between optimizing the FKL objective and optimizing the RKL objective. In fact, our reported results compare RKL and FKL methods for the same number of IS samples and optimization iterations. See Appendix E for exact hyperparameter values and wall clock times.

9 CONCLUSION

Overall, we propose a principled algorithm that combines the strengths of importance sampling and variational inference to efficiently approximate multimodal and possibly heavy-tailed targets. Unlike prior work that relies on RKL, our minimization of FKL aligns with the analysis of the variance of IS [Chatterjee and Diaconis, 2018] which guarantees an optimal proposal distribution asymptotically. One challenge for this approach is the variance of the SNIS estimate of the forward KL divergence. Developing variance-reduction schemes for these types of objectives is an open research problem. Nonetheless, existing techniques for resampling or smoothing the importance weights can immediately apply to our proposed method.

References

Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 2003.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*, 2016.

François Bolley and Cédric Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la faculté des sciences de Toulouse*, 13(3):331–352, 2005.

Jörg Bornschein and Yoshua Bengio. Reweighted wake-sleep. *arXiv preprint arXiv:1406.2751*, 2014.

Léon Bottou. *Large-scale Machine Learning with Stochastic Gradient Descent*. Springer, 2010.

Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

Trevor Campbell and Xinglong Li. Universal boosting variational inference. In *Advances in Neural Information Processing Systems*, pages 3484–3495, 2019.

Olivier Cappé, Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18(4):447–459, 2008.

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

Jean-Marie Cornuet, Jean-Michel Marin, Antonietta Mira, and Christian P Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, 2012.

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, pages 2732–2741, 2017.

Joshua V. Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A. Saurous. Tensorflow distributions. *arXiv preprint arXiv:1711.10604, 2017*, 2017.

Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. In *Advances in neural information processing systems*, pages 4470–4479, 2018.

Randal Douc, Arnaud Guillin, J-M Marin, and Christian P Robert. Minimum variance importance sampling via population Monte Carlo. *ESAIM: Probability and Statistics*, 11:427–447, 2007.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

John Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339, 1989.

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

Jerónimo Hernández-González and Jesús Cerquides. A robust solution to variational importance sampling of minimum variance. *Entropy*, 22:1405, 2020.

José Miguel Hernández-Lobato and Ryan P Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. *International Conference on Machine Learning (ICML)*, 2015.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. 2016.

Matthew Hoffman and Yi-An Ma. Black-box variational inference as distilled Langevin dynamics. In *International Conference on Machine Learning*, pages 9267–9277, 2020.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research (JMLR)*, 14, 2013.

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

Ghassen Jerfel. Boosted stochastic backpropagation for variational inference. 2017.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

H. Kahn and A. W. Marshall. Methods of reducing sample size in monte carlo computations. *Operations Research*, 1(5):263–278, 1953.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

Francesco Locatello, Rajiv Khanna, Joydeep Ghosh, and Gunnar Ratsch. Boosting variational inference: an optimization perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 464–472. PMLR, 2018a.

Francesco Locatello, Rajiv Khanna, Joydeep Ghosh, and Gunnar Rätsch. Boosting variational inference: An optimization perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 464–472, 2018b.

Yi-An Ma, Emily B Fox, Tianqi Chen, and Lei Wu. Irreversible samplers from jump and continuous markov processes. *Statistics and Computing*, 29(1):177–202, 2019.

Dougal Maclaurin, David Duvenaud, and Ryan P. Adams. Autograd: Reverse-mode differentiation of native python. *ICML workshop on Automatic Machine Learning*, 2015.

Andrew Miller, Nicholas Foti, and Ryan Adams. Variational boosting: Iteratively refining posterior approximations. *ICML*, 2017.

Thomas Peter Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

Tom Minka et al. Divergence measures and message passing. Technical report, Citeseer, 2005.

Kevin P Murphy. *Machine Learning: a Probabilistic Perspective*. MIT press, 2012.

Radford Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov chain Monte Carlo*, 2011.

Art B Owen. *Monte Carlo Theory, Methods and Examples*. 2013.

Dennis Prangle. Distilling importance sampling. *arXiv preprint arXiv:1910.03632*, 2019.

Adrian E Raftery and Le Bao. Estimating and projecting trends in HIV/AIDS generalized epidemics using incremental mixture importance sampling. *Biometrics*, 66(4): 1162–1173, 2010.

Tom Rainforth, Adam R Kosiorek, Tuan Anh Le, Chris J Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. *arXiv preprint arXiv:1802.04537*, 2018.

Darío Ramos-López, Andrés R. Masegosa, Antonio Salmerón, Rafael Rumí, Helge Langseth, Thomas D. Nielsen, and Anders L. Madsen. Scalable importance sampling estimation of gaussian mixture posteriors in bayesian networks. *International Journal of Approximate Reasoning*, 100:115–134, 2018.

Francisco JR Ruiz and Michalis K Titsias. A contrastive divergence for combining variational inference and MCMC. *arXiv preprint arXiv:1905.04062*, 2019.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends®in Machine Learning*, 1:1–305, 2008.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work?: Evaluating variational inference. *arXiv preprint arXiv:1802.02538*, 2018.

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