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

Anonymous Authors
Anonymous Institution

Abstract

Variational Inference (VI) is a popular alternative to asymptotically exact sampling in Bayesian inference. Its main workhorse is optimization over a reverse Kullback-Leibler divergence (RKL), which typically underestimates the tail of the posterior and causes miscalibration and potential degeneracy (over-pruning). Importance sampling (IS), on the other hand, is often used to fine-tune and debias the estimates of approximate Bayesian inference procedures. The quality of IS crucially depends on the choice of the proposal distribution. Ideally, the proposal distribution has heavier tails than the target, which is unachievable by minimizing the RKL. We thus propose a novel combination of optimization and sampling techniques for approximate Bayesian inference 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 variational approximation.

1. Introduction

Bayesian analysis provides a powerful framework to encode complex hierarchical structures and prior beliefs and capture posterior uncertainty about latent variables θ given observed data x via the posterior $p(\theta|x)$. The inferential goal often involves computing expectations over this posterior distribution, $\mathbb{E}_{\theta\sim p(\theta|x)}[f(\theta)]$, for some function f which is generally accomplished by sampling. Unfortunately, the posterior distribution is often difficult to sample from directly, and approximate inference methods are necessary to estimate posterior functionals. These methods include optimization-based techniques such as variational inference (VI) (Jordan et al., 1999; Wainwright and Jordan, 2008), and sampling-based techniques such as Markov Chain Monte Carlo (MCMC) (Brooks et al., 2011; Andrieu et al., 2003) and importance sampling (IS) (Gelman and Meng, 1998).

Variational inference has grown in popularity among these methods due to the computational convenience of minimizing the reverse (exclusive) Kullback-Leibler divergence (RKL or $KL(q||p)$) over a tractable family of distributions. Furthermore, formulating inference as divergence minimization enables the application of recent advances in stochastic optimization (Bottou, 2010; Hoffman et al., 2013) and automatic differentiation (Maclaurin et al., 2015). However, the quality of VI approximations, when the variational family does not exactly match the target, is of major concern, especially when the approximation is used to estimate posterior expectations (Yao et al., 2018; Campbell and Li, 2019; Rainforth et al., 2018; Huggins et al., 2020).

With importance sampling (Gelman and Meng, 1998; Owen, 2013), one can re-weigh posterior summaries from a variational approximation $q(\theta)$ in order to correct for the error in q as compared to the target at the cost of additional variance relative to a simple Monte

Carlo estimate. The performance of q when used as an IS proposal can also be a practical diagnostic for whether $p(\theta|x)$ should be approximated by $q(\theta)$ (Yao et al., 2018).

Importance sampling is thus still a relevant tool especially in the modern era of approximate inference where the family of approximate distributions is often mis-specification and refinement might be needed to de-bias posterior summaries.However, practical applications of importance sampling are challenged by the choice of proposal distribution, which is particularly difficult in multimodal and high-dimensional settings. One potential way out of this conundrum is to use of variational inference to construct a tractable proposal distribution for IS and thereby capitalize on the complementary strengths of IS and VI. Unfortunately, the VI approximations obtained by RKL minimization are not ideal proposal distributions for IS due to the light tails resulting from entropy penalization (Minka et al., 2005). This typically leads to heavy-right tails in the distribution of importance weights, which can lead to unstable IS estimates, sometimes with infinite variance (Yao et al., 2018). On the other hand, forward (inclusive) KL divergence (FKL) (Minka et al., 2005) controls the worst-case estimation error of importance sampling (Chatterjee and Diaconis, 2018). Furthermore, forward KL does not suffer from issues of covariance underestimation (Minka et al., 2005; Campbell and Li, 2019) or zero forcing (Hoffman, 2017).

We propose to replace reverse KL with forward KL as the variational objective that supplies an IS proposal distribution. We make the following contributions in this vein:

1. We derive a self-normalized importance sampling estimate for the FKL divergence.
2. We show how FKL-based boosting can be used to better harness the combination of IS and VI to address the challenge of estimating 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 implies convergence to the optimal proposal distribution as per the results of Chatterjee and Diaconis (2018). Our proposed algorithm is 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 a given approximation's error.

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 $\mathbb{E}_{\theta \sim p(\theta|x)}[f(\theta)]$. 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 the expectation over the target distribution 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 $q(\theta)$ of the posterior, we can estimate the expectation as:

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

If the samples are weighted equally (i.e., $w_s = 1$), Eq. (1) 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:

$$\mathbb{E}_{\theta \sim p(\theta|x)} [f(\theta)] = \mathbb{E}_{\theta \sim q(\theta)} \left[\frac{p(\theta|x)}{q(\theta)} f(\theta) \right], \tag{2}$$

which is asymptotically consistent ($O(1/S)$ bias) but with possibly infinite variance.

2.1. Variational Inference 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 $\mathcal{Q}$ exactly includes the target distribution, minimizing the reverse KL divergences leads to an q^* that underestimates the target co-variance. The decomposition of this divergence sheds light on the cause:

$$\mathrm{KL}(q\|p) = \mathbb{E}_q[\log q] - \mathbb{E}_q[\log p] \tag{3}$$

The first term in Eq. (3) 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*) and is biased towards avoiding false positives, which can cause severe under-estimation of the covariance (Blei et al., 2016; Murphy, 2012) 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 F.

2.2. Importance Sampling and Forward KL

Since the the variational family $\mathcal{Q}$ rarely includes the true posterior, refining posterior approximations through importance weighting is often recommended to de-bias expectations of interest (Yao et al., 2018; Vehtari et al., 2015). However, the light tails of VI approximations using RKL can lead to high or even infinite variance for the IS estimator.

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

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

In Eq. (4), the cross-entropy is minimized by setting $q > 0$ whenever $p > 0$, which leads to *mass covering* as illustrated in Figure F. With tails that are more likely to cover the target distribution p, an approximation q^* generated by minimizing the FKL can lead to an IS estimate (Eq. (2)) with lower variance. Chatterjee and Diaconis (2018) show that the forward KL divergence controls the downstream worst-case estimation error of IS.

3. 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.

3.1. Forward KL-based Variational Approximation for IS

By Theorem 1.1 of Chatterjee and Diaconis (2018), we know 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 importance sampling to provide accurate mean estimates scales exponentially with the KL divergence $\mathrm{KL}(p\|q)$. Therefore, an optimal proposal distribution should minimize the forward KL divergence which is computationally difficult given the need for samples from an un-normalized target p. Alternatively, the reverse KL divergence, which utilizes samples from q instead, can be regarded as a tractable approximation with unknown bias (Yao et al., 2018). However, for tractable choices of $\mathbb{Q}$ (i.e. when $KL(q||p) \neq 0$) it has been demonstrated that minimizing the RKL is not guaranteed to minimize the FKL divergence (Campbell and Li, 2019). Furthermore, the light tails of the RKL minimization solution can lead to unstable IS estimates (Yao et al., 2018).

Therefore, we seek to directly minimize the forward KL objective instead of the RKL to estimate the optimal IS proposal with improved covariance and tail estimation. From the IS point of view, this combination with VI enables a computationally efficient and principled construction of optimal proposal distributions.

3.2. Self-Normalized IS Approximation of Forward KL

We cannot compute the forward KL divergence exactly for unnormalized target distributions. This stems from the expectation under the target p in Eq. (4)). By contrast, the RKL takes the expectation in Eq. (3)) under a normalized approximation q. Therefore, to approximate FKL, we can rearrange densities inside the expectation as follows:

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

Eq. (5) can then be approximated through self-normalized importance sampling (SNIS) (Murphy, 2012) which is known to generally be consistent:

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

The SNIS estimation can have arbitrarily high variance depending on q. Gradients of Eq. (5) 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). Regardless, gradient variance is less of a concern in the following sequential approximation framework.

3.3. Forward KL-Based Boosting

We seek to sequentially construct a proposal mixture distribution q that is both easy to sample from and minimizes the FKL divergence to ensure efficient IS estimation.

The design of a variational family $\mathcal{Q}$ can be challenging a priori which often leads to model misspecification in VI estimates. Our experiments in App. D rely on the family of Gaussian mixtures due to the low variance of reparametrization gradients (Kingma and Welling, 2013). Our framework also supports mixtures of heavy-tailed distributions which are desirable for adaptive IS (Geweke, 1989).

Our proposal distribution at the K^{th} boosting iteration is $q_K(\theta) := \sum_{i=1}^{K} \lambda_i f_i(\theta; \phi_i)$. At each boosting iteration, we minimize the forward KL between the mixture q_i and the target p while holding the parameters of previously-learned mixture components fixed:

$$\underset{q_i \in \mathcal{Q}}{\text{argmin}} \quad \text{KL}(p \| q_i) = \underset{f_i, \gamma}{\text{argmin}} \quad \text{KL}\left(p \| \gamma f_i + (1-\gamma) q_{i-1}\right). \tag{6}$$

This is known to be more efficient and more stable than the joint optimization of all mixture components at each iteration (Locatello et al., 2018a). The mixture weight for the new component is then set to $\lambda_i = \gamma$, and all other weights $\{\lambda_j\}_{j=1}^{i-1}$ are re-scaled by $(1-\gamma)$. Mixture weights can be further adjusted with a fully-corrective weight search using the gradient derived in Eq. (20) (Guo et al., 2016; Locatello et al., 2018b). We provide an alternative approach to building q_K using the remainder in App. B.

3.4. Lower Self-Normalized IS Variance with Boosting

The main computational concern in estimating the FKL divergence with SNIS is the variance of IS in Eq. (5). However, the sequential construction described above enables further rearrangements of the SNIS approximation which can lead to a lower variance than the naive approximation of Eq. (5). For the first approach, the global objective can be re-written as:

$$\underset{q_i \in \mathcal{Q}}{\text{argmin}} \quad \text{KL}(p \| q_i) = \underset{q_i \in \mathcal{Q}}{\text{argmin}} \quad \mathbb{E}_p \left[\log \frac{p}{q_i}\right] = \underset{f_i, \gamma}{\text{argmin}} \quad \mathbb{E}_{q_i} \left[\frac{p}{q_i} \log \frac{p}{\gamma f_i + (1-\gamma) q_{i-1}}\right] \tag{7}$$

$$= \underset{f_i, \gamma}{\text{argmin}} \quad \mathbb{E}_{q_{i-1}} \left[\frac{p}{q_{i-1}} \log \frac{p}{\gamma f_i + (1-\gamma) q_{i-1}}\right] \tag{8}$$

Eq. (8), 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. (8) given samples θ_s from q_{i-1} can be computed as:

$$\sum_{s=1}^{S} w_s [\log p(\theta_s) - \log\left(\lambda f_i(\theta_s) + (1-\lambda) q_{i-1}(\theta_s)\right)], \tag{9}$$

where w_s are the self-normalized importance weights computed as in Eq. (6).

3.5. Initialize with Reverse KL, Refine with Forward L

Our sequential construction of the proposal can reduce the variance of the SNIS approximation of the FKL divergence by drawing samples from the previously-estimated and fixed mixture distribution q_{i-1}. However, in the first boosting iteration, we do not have an existing approximation to sample from. Instead, Eq. (9) can be re-written as follows:

$$\underset{q_i \in \mathcal{Q}}{\text{argmin}}\, \mathbb{E}_p \left[\log \frac{p}{q_i}\right] = \underset{f_i, \gamma}{\text{argmin}}\, \mathbb{E}_{f_i} \left[\frac{p}{f_i} \log \frac{p}{\gamma f_i + (1-\gamma) q_{i-1}}\right]. \tag{10}$$

This exacerbates the sensitivity to the initialization of f_i. In fact, if f_i were initialized to be sharply peaked at zero, for example, then our SNIS estimate of FKL cannot properly capture the tails. Therefore, a diffuse initialization is recommended.

However, 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.

4. Analysis

We provide theoretical analysis of the proposed method of performing IS with a proposal distribution constructed from FKL-based boosting (further analysis in App. C).

4.1. Forward KL 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 n-Wasserstein as follows (Bolley and Villani, 2005): Assume that the probability density q is n-exponentially integrable. Then for target distribution p such that $p \ll q$:

$$W_n(q,p) \leq C_n^{EI}(q) \left[\mathrm{KL}(p\|q)^{1/n} + \frac{\mathrm{KL}(p\|q)}{2}^{1/2n} \right], \tag{11}$$

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 convergence in up to n^{th} moments (Bolley and Villani, 2005).

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 does not need to be satisfied by the target density p in the case of Eq. (11). 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.

4.2. Forward KL 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 App. E.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 App. 4.2 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 K^{th} iteration is : $\mathrm{KL}(p\|q_K) = \mathrm{KL}\left(p \middle\| \sum_{i=1}^{K} \lambda_i f_i \right) = O(1/K)$. 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.

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.

F. Bolley and C. 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.

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

S. Brooks, A. Gelman, G. Jones, and X.L 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.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

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.

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.

M D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research (JMLR)*, 14, 2013.

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. Learning deep latent Gaussian models with Markov chain Monte Carlo. In *International conference on machine learning*, pages 1510–1519, 2017.

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.

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.

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

Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

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

Tom Minka et al. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 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.

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.

Appendix A. Related Work

We present related work on improving estimates of the expectation over intractable target distributions that are high dimensional, heavy tailed, and/or multimodal.

A.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), χ^2 ($\alpha = 2$) (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.

A.2. Forward KL 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). This is different from how a KL divergence is typically used in VI since EP/BP do not optimize a global objective that is related to the KL divergence and are thus not guaranteed to converge. Furthermore, EP/BP does not provide an easy estimate or bound on the marginals. 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 leading to systemic bias.

A.3. Variational Boosting

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

$$\mathcal{Q}_k = \left\{ q : q(\theta) = \sum_{i=1}^{k} \lambda_i f_i(\theta), \lambda \in \Delta_k \right\}, \tag{12}$$

and iteratively refines a variational mixture approximation by adding and re-weighting one (typically Gaussian) mixture component at a time in order to minimize the global objective:

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

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 components (Campbell and Li, 2019). Ad-hoc regularization techniques are often deployed in practice (Locatello et al., 2018a), but are not necessarily sufficient (Campbell and Li, 2019). To address this pathology of RKL-based boosting, Campbell and Li (2019) proposed a boosting algorithm, based on the Hellinger divergence, which is still not guaranteed to scale well with dimension.

A.4. Adaptive Importance Sampling

The adaptive IS (AIS) literature presents a range of techniques that involve adapting proposal distributions for multimodal targets. Adaptive multiple IS (Cornuet et al., 2012) and incremental mixture IS (Raftery and Bao, 2010) are two of the most popular methods. 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).

A.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. AIS, on the other hand, can approximate heavy-tailed targets but cannot scale efficiently in dimensionality. 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 importance weighting and VI (Domke and Sheldon, 2018) have still focused on minimizing the reverse KL, 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.

Failure Mode	VI	IS	VB	AIS	Ours
Multimodality	✗	✗	✓	✓	✓
Heavy-tailed target	✗	✓	✗	✓	✓
Cov underestimation	✗	✓	✗	✓	✓
High-dim scalability	✓	✗	✓	✗	✓*

Table 1: Comparing approximate inference techniques.

Appendix B. An Alternative approach to FKL-Based Boosting: Minimizing the Remainder

As we seek to construct an optimal proposal through the minimization of an SNIS approximation of FKL, a trade-off arises: "should we make the distribution easier to sample from in order to minimize the SNIS variance or should we bring it closer to the target in order to improve the worst-case IS estimation error?" In particular, the closer the proposal gets to a multimodal target, the harder it may be to sample from. Therefore, this trade-off translates to two distinct approaches for the greedy additive construction of an optimal proposal mixture distribution.

The first approach described in Section 3.3 is the most straightforward as it minimizes the forward KL between the mixture q_i and the target p while holding the parameters of previously-learned mixture components fixed.

Alternatively, define the remainder distribution at iteration i as $r_i(\theta) = \frac{p(\theta|x)}{q_i(\theta)}$. A second approach is to minimize the FKL between each new component and r_i, which may be simpler with fewer modes than p:

$$\underset{f_i}{\operatorname{argmin}}\ \mathrm{KL}(r_i \| f_i) = \underset{f_i}{\operatorname{argmin}}\ \mathrm{KL}\left(\frac{p_i}{q_{i-1}} \middle\| f_i\right).$$

The mixture weight can be estimated in this scenario for each mixture component by gradient descent using the gradient with respect to FKL (see App. E.4).

This second approach is appealing because, at each boosting iteration, r_i becomes shallower with fewer modes which makes it easier to sample from than the multimodal proposal q_i. This approach can also be motivated by gradient boosting (Friedman, 2001) or matching pursuit (Mallat and Zhang, 1993) where one seeks to identify the mixture component that best fits the functional residual. Furthermore, this approach might be less prone to degeneracy. In fact, a derivation of this approach for the reverse KL, in App. E.2, identifies intrinsic entropy regularization which is often incorporated ad-hoc in similar objectives.

B.1. Lower Variance SNIS with Boosting

The remainder-based objective can also be re-written:

$$\underset{f_i}{\operatorname{argmin}} \quad \mathrm{KL}(r_i \| f_i) = \underset{f_i}{\operatorname{argmin}} \quad \mathbb{E}_{f_i}\left[\frac{r_i}{f_i} \log \frac{r_i}{f_i}\right].$$

This objective also enjoys lower gradient variance since the distribution being sampled from is a single Gaussian component and not a multimodal proposal. However, the light tails of a single Gaussian component can lead to high variance SNIS estimates in the case of severe mismatch with the target distribution. Therefore, the choice of either should be guided by assumptions about the tails of the target. However, given comparable empirical results of the two approaches, we limit the analysis of Section 4 and the experiments of Section D to the first and more straightforward approach.

Appendix C. Additional Analysis

C.1. Sample Complexity of SNIS

By Theorem 1.1 of Chatterjee and Diaconis (2018), we know 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 importance sampling to provide accurate mean estimates scales exponentially with the KL divergence $\mathrm{KL}(p\|q)$.

C.2. Computation-Quality Trade-Off

While our our iterative algorithm which 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 which depends on the number of samples, and the greedy sequential approximation error which depend on the number of VI iterations, IS samples, and boosting iterations, respectively. These three hyperparameters thus finely control the compute-quality trade-off of our framework.

From previous sections, we observe a trade-off between sample complexity of SNIS versus 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. 4 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. 4 and the experimental results, we observe uniform improvements of FKL over RKL methods.

Appendix D. Experiments

We evaluate the performance of the proposed method when applied to Bayesian linear regression (BLR) and Bayesian neural networks (BNNs) using a Gaussian prior and a heavy tailed prior. We use four datasets from UCI Dua and Graff (2017). We split each dataset into five 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 boosting iterations. 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 3.5.

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. (20). Details about practical considerations and hyperparameters can be found in App.D, and we include our code with the submission.

Comparisons: We compare our approach to variational boosting using reverse KL Miller et al. (2017). First, we compare to fitting a single Gaussian distribution to the target to minimize the RKL (RKL VI). We also compare to variational boosting with two and three components, also minimizing RKL (RKL VB). 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 App. D.

D.1. Toy Data

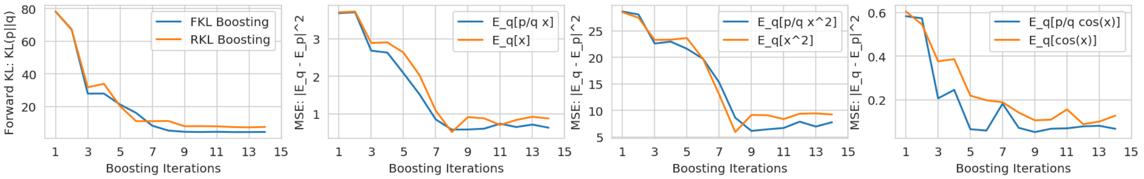

Figure 1: Evolution of (exact) FKL divergence and the mean-squared error of moment estimation (using samples from the FKL solution) on the task of estimating a 2D GMM of 20 components (Ma et al., 2018).

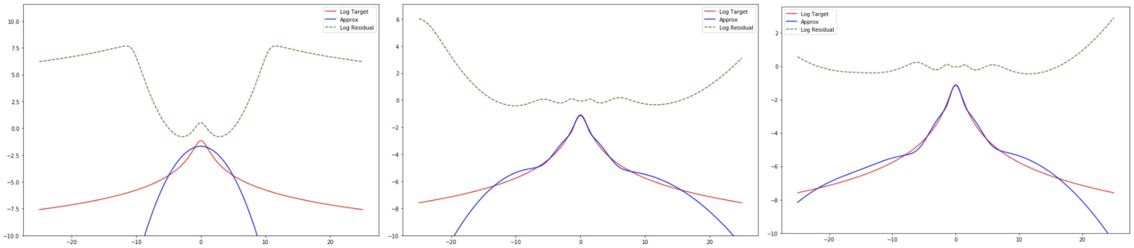

Figure 2: Density plots of FKL boosting on Cauchy (from left to right: iterations 1, 5, and 10).

D.2. BLR and BNNs with Gaussian Priors

Method	Wine (d=14)	Boston (d=16)	Concrete (d=11)	Power (d=7)
HMC	-0.998 (± 0.067)	-2.822 (± 0.084)	-3.755 (± 0.051)	**-2.942** (± 0.039)
RKL VI	-1.004 (± 0.057)	-2.809 (± 0.088)	-3.766 (± 0.045)	-2.956 (± 0.034)
RKL VB 2	-1.001 (± 0.053)	-2.825 (± 0.081)	-3.764 (± 0.052)	-3.275 (± 0.029)
RKL VB 3	-0.999 (± 0.058)	-2.832 (± 0.085)	-3.763 (± 0.047)	-3.215 (± 0.027)
FKL VI	**-0.985** (± 0.072)	-2.791 (± 0.091)	-3.762 (± 0.056)	-2.954 (± 0.0031)
FKL VB 2	-0.991 (± 0.076)	**-2.779** (± 0.090)	**-3.761** (± 0.051)	-2.955 (± 0.034)
FKL VB 3	-0.990 (± 0.080)	-2.796 (± 0.092)	-3.765 (± 0.047)	-2.956 (± 0.033)

Table 2: Predictive log probabilities on test for BLR with Gaussian prior (5 train/test splits).

We first apply the same Gaussian prior as 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

Method	Wine (d=653)	Boston (d=753)	Concrete (d=503)	Power (d=303)
HMC	-0.982 (± 0.078)	-2.498 (± 0.096)	-3.271 (± 0.076)	**-2.835** (± 0.033)
RKL VI	-0.992 (± 0.066)	-2.858 (± 0.077)	-3.230 (± 0.047)	-2.951 (± 0.037)
RKL VB 2	-0.994 (± 0.054)	-2.832 (± 0.055)	-3.253 (± 0.058)	-3.363 (± 0.041)
RKL VB 3	-0.981 (± 0.057)	-2.769 (± 0.091)	-3.294 (± 0.042)	-3.522 (± 0.054)
FKL VI	-0.981 (± 0.067)	**-2.477** (± 0.055)	**-2.897** (± 0.060)	-2.848 (± 0.037)
FKL VB 2	-0.972 (± 0.082)	-2.551 (± 0.066)	-3.469 (± 0.069)	-2.946 (± 0.037)
FKL VB 3	**-0.967** (± 0.069)	-2.881 (± 0.048)	-3.449 (± 0.066)	-3.227 (± 0.039)

Table 3: Predictive log probabilities on test for BNNs with Gaussian prior (5 train/test splits).

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:

$$p(y|x, \mathcal{D}_{\text{train}}) = \int p(y|x, \theta)p(\theta|\mathcal{D}_{\text{train}})d\theta. \tag{13}$$

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

$$p(y|x, \mathcal{D}_{\text{train}}) \approx \frac{1}{S}\sum_{i=1}^{S} \frac{p(\theta_s|\mathcal{D}_{\text{train}})}{q(\theta_s)} p(y|x, \theta_s), \tag{14}$$

where $q(\theta)$ is the proposal distribution fit to the posterior $p(\theta|\mathcal{D}_{\text{train}})$ 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 App. D.

For the comparison with HMC, we do not use importance sampling due to lack of an explicit density function, and instead compute (13) by averaging over direct HMC samples from the posterior distribution $p(\theta|\mathcal{D}_{\text{train}})$:

$$p(y|x^*, \mathcal{D}_{\text{train}}) \approx \frac{1}{L}\sum_{l=1}^{L} p(y|x^*, \theta^{(l)}), \quad \theta^{(l)} \sim p(\theta|\mathcal{D}_{\text{train}}) \tag{15}$$

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

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

D.3. BLR with Heavy Tailed Priors

Method	Wine (d=13)	Boston (d=15)	Concrete (d=10)	Power (d=6)
HMC	-0.999 (± 0.064)	-2.908 (± 0.082)	-3.754 (± 0.052)	**-2.942** (± 0.039)
RKL VI	-0.997 (± 0.070)	-2.847 (± 0.076)	-3.766 (± 0.052)	-2.951 (± 0.031)
RKL VB 2	-0.992 (± 0.075)	-2.856 (± 0.067)	-3.765 (± 0.047)	-3.317 (± 0.039)
RKL VB 3	-0.999 (± 0.064)	-2.878 (± 0.070)	-3.773 (± 0.055)	-3.372 (± 0.036)
FKL VI	**-0.987** (± 0.072)	-2.900 (± 0.064)	**-3.750** (± 0.052)	-2.950 (± 0.032)
FKL VB 2	-1.090 (± 0.115)	-2.858 (± 0.090)	-3.769 (± 0.039)	-3.912 (± 0.023)
FKL VB 3	-1.098 (± 0.102)	**-2.845** (± 0.098)	-3.767 (± 0.054)	-4.482 (± 0.014)

Table 4: Predictive log probabilities on test for BLR with heavy tailed prior (5 train/test splits).

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 again estimate the same posterior predictive distribution in Eq. (13) using importance sampling in Eq. (14), and report the average predictive log probabilities from Eq. (16).

Appendix E. Derivations

E.1. Connecting Forward KL to other metrics used for VI

By the monotonicity of Renyi-α divergences, given that $\lim_{\alpha \to 1} D_\alpha(p, q) = KL(p\|q)$, and from Dieng et al. (2017) :

$$\text{KL}(p\|q) \leq D_2(p, q) \leq \chi^2(p, q) \tag{17}$$

E.2. Reverse KL remainder: intrinsic entropy regularization

Computing the remainder-reverse KL objective using our approach leads to the well-known although usually ad-hoc entropy regularization (e.g. Locatello et al. (2018a)).

$$\text{KL}(f_i\|r_i) = \text{KL}(f_i\|\frac{p}{q_{i-1}}) = \mathbb{E}_{f_i}[\log \frac{f_i q_{i-1}}{p}] = \mathbb{E}_{f_i}[\log \frac{q_{i-1}}{p}] + \mathbb{E}_{f_i}[\log f_i] \tag{18}$$

As we can see, while the first term is the mean of the log-residual under the new component f_i, the typical objective for gradient boosting, the second term is the entropy of f_i.

E.3. SNIS derivation

Since we do not assume to know the normalization constant of p, we shall approximate the above quantities by self-normalized importance sampling while making the distinction between the normalized p and the un-normalized $\hat{p}$:

$$\theta_s \sim q_{i-1}, \quad w^s = \frac{p(\hat{\theta}_s)}{q_{i-1}(\theta_s)}, \quad w^s_{norm} = \frac{w^s}{\sum_s w^s} \tag{19}$$

$$\mathbb{E}_{q_{i-1}}[\frac{p}{q_{i-1}} \log \frac{p}{\lambda f_i + (1-\lambda) q_{i-1}}] = \frac{\mathbb{E}_{q_{i-1}}[\frac{\hat{p}}{q_{i-1}} \log \frac{p}{\lambda f_i + (1-\lambda) q_{i-1}}]}{\mathbb{E}_{q_{i-1}}[\frac{\hat{p}}{q_{i-1}}]}$$

$$\approx \sum_s \frac{\frac{p(\hat{\theta}_s)}{q_{i-1}(\theta_s)}}{\sum_s \frac{p(\hat{\theta}_s)}{q_{i-1}(\theta_s))}} [\log \frac{p(\theta_s)}{\lambda f_i(\theta_s) + (1-\lambda) q_{i-1}(\theta_s)}] = \sum_s w^s_{norm} [\log p(\theta_s) - \log\left(\lambda f_i(\theta_s) + (1-\lambda) q_{i-1}(\theta_s)\right)]$$

E.4. Gradients of mixture weights

For forward KL:

$$\nabla_{\lambda_i} \mathbb{E}_p \left[\log p - \log \sum_j^K \lambda_j q_j \right] = -\mathbb{E}_p \left[\nabla_{\lambda_i} \log \sum_j^K \lambda_j q_j \right]$$

$$= -\mathbb{E}_p \left[\frac{q_i}{\sum_j^K \lambda_j q_j} \right] = -\mathbb{E}_{q_i} \left[\frac{p}{q} \right]. \tag{20}$$

For reverse KL:

$$\nabla_{\lambda_i} \mathbb{E}_q [\log q - \log p] = \mathbb{E}_{\sum_j^K \lambda_j q_j} [\nabla_{\lambda_i} \log (\sum_j^K \lambda_j q_j)]$$

$$= \mathbb{E}_{q_i} [\log q - \log p]$$

E.5. The functional gradient of the forward KL divergence

We assume $\operatorname{supp} p \subseteq \operatorname{supp} q$: that is, p is absolutely continuous with respect to the variational approximation q_i which can be ensured by the design of the variational family $\mathcal{Q}$.

Let $D(q) = \mathrm{KL}(p \| q)$. Functional gradient $\frac{\delta D}{\delta q}$ can be computed from the Taylor expansion of the KL functional Friedman (2001) as follows:

$$\lim_{\epsilon \to 0} \frac{D(q + \epsilon \cdot h) - D(q)}{\epsilon} = \int \frac{\partial D}{\partial q} h dx \tag{21}$$

As detailed in the App. E.5

$$\begin{aligned}\frac{D(q+\epsilon\cdot h)-D(q)}{\epsilon} &= \frac{1}{\epsilon}\int p\log p - p\log(q+\epsilon h)\\ &+ p\log p - p\log q\\ &= -\frac{1}{\epsilon}\int p\log(q+\epsilon h) - p\log q\\ &= -\frac{1}{\epsilon}\int p\log(1+\epsilon\frac{h}{q})\end{aligned}$$

We have the logarithmic inequality $\frac{x}{x+1} \leq \log(1+x) \leq x \forall x > -1$ where we can substitute $\epsilon\frac{h}{q} > 0$ for x and arrive at

$$-\frac{h}{q} \leq -\frac{1}{\epsilon}\log(1+\epsilon\frac{h}{q}) \leq -\frac{\frac{h}{q}}{1+\epsilon\frac{h}{q}}$$

By the monotone convergence theorem we can take the limit inside the integral and arrive at

$$\lim_{\epsilon\to 0}\int -p\frac{1}{\epsilon}\log(1+\epsilon\frac{h}{q}) = \int -p\frac{h}{q} \tag{22}$$

$$\frac{\delta D(q)}{\delta q} = -\frac{p}{q}$$

E.6. Boosting convergence analysis

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 the base family at a rate of $O(1/K)$ where K is the number of boosting iterations. This framework does not require each iteration to exactly solve for the optimal mixture component which can be difficult in variational inference.

While the convexity of $\mathrm{KL}(p\|q)$ in q is well established in the literature (proven with the log-sum inequality) for the forward KL divergence functional, we can show that FKL is also β-smooth in q where β depends on the maximum and minimum values that the density q can take. To establish strong smoothness, on the other hand, stricter assumptions about the densities are necessary. If we assume that all densities are bounded away from 0 and from above q_1 then for any pair of densities q_1 and q_2 there exists a $\beta = \sup \frac{p}{q_1 * q_2} \geq 0$ such that the functional gradient $\frac{\delta D}{\delta q}$ is β- Lipschitz, that is $\left|\frac{\delta D}{\delta q}(q_2) - \frac{\delta D}{\delta q}(q_1)\right| \leq \beta|q_2 - q_1|$. We can verify this choice of β:

$$\left|\frac{\delta D}{\delta q}(q_2) - \frac{\delta D}{\delta q}(q_1)\right| = \left|\frac{-p}{q_2} - \frac{-p}{q_1}\right| \tag{23}$$

$$= \left|\frac{p(q_2-q_1)}{q_2 q_1}\right| \tag{24}$$

$$= \frac{p}{q_2 q_1}|q_2 - q_1| \tag{25}$$

$$\leq \beta|q_2 - q_1| \tag{26}$$

Note that the boundedness assumptions are not unrealistic in practice and can translate to a bounded parameter space for a given family of distributions.

$$\mathrm{KL}(p\|q_i) = \mathrm{KL}(p\|\sum_{i}^{k}\lambda_i f_i) = O(1/k) \tag{27}$$

Appendix F. Supporting Figures

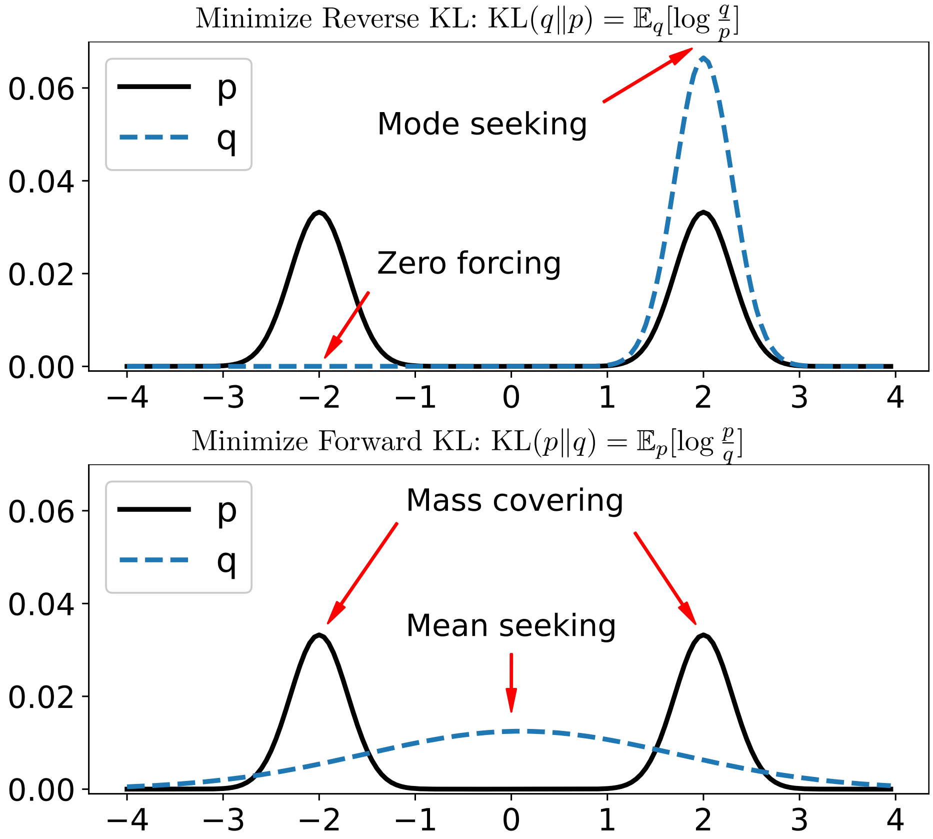

Figure 3: Results of minimizing RKL vs. FKL for a Gaussian variational approximation.

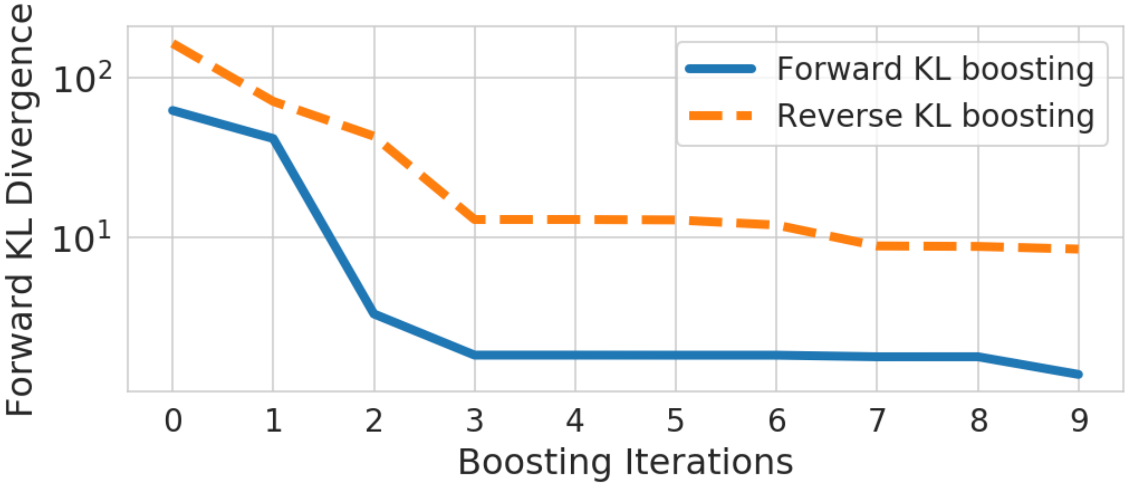

Figure 4: A comparison in terms of the FKL divergence to a Cauchy target distribution over the course of variational boosting using the FKL and RKL divergences.