Boosting Variational Inference: an Optimization Perspective

Francesco Locatello
MPI for Intelligent Systems
ETH Zurich

Rajiv Khanna
UT Austin

Joydeep Ghosh
UT Austin

Gunnar Rätsch
ETH Zurich

Abstract

Variational inference is a popular technique to approximate a possibly intractable Bayesian posterior with a more tractable one. Recently, boosting variational inference [20, 4] has been proposed as a new paradigm to approximate the posterior by a mixture of densities by greedily adding components to the mixture. However, as is the case with many other variational inference algorithms, its theoretical properties have not been studied. In the present work, we study the convergence properties of this approach from a modern optimization viewpoint by establishing connections to the classic Frank-Wolfe algorithm. Our analyses yields novel theoretical insights regarding the sufficient conditions for convergence, explicit rates, and algorithmic simplifications. Since a lot of focus in previous works for variational inference has been on tractability, our work is especially important as a much needed attempt to bridge the gap between probabilistic models and their corresponding theoretical properties.

1 Introduction

Variational inference [1] is a method to approximate complicated probability distributions with simpler ones. In many applications, calculating the exact posterior distribution is intractable, and methods like MCMC while being flexible can also be prohibitively expensive. Variational inference restricts the posterior to be a member of a simpler and more tractable family of distributions, and the inference problem reduces to finding this member that can "closely" represent the true underlying posterior. The closeness is typically measured in the KL sense.

One of the most commonly used family of distributions for the tractable set is the so called *mean field family*, which assumes a factored structure. An example of such a family is the set of Gaussian distributions with diagonal covariance matrices. While the inference is computationally efficient due to the properties of Gaussian distributions, this family can be too restrictive. As such, the approximated distribution is often not a good representation of the true posterior. A simple example is a multi-modal distribution. The mean field family will be able to only capture one of the modes.

There have been a number of efforts to improve the approximation while retaining the simplicity of Gaussian distributions. For example, one could consider approximating by a mixture of Gaussian distributions and allowing more than just isotropic structures. A mixture of isotropic Gaussian distributions is already a much more powerful and flexible model than a single isotropic Gaussian. In fact, it is flexible enough to model any distribution arbitrarily closely [22]. While there has been significant algorithmic and empirical development for studying variational inference using mixture models [20, 4, 16, 17], there have been limited theoretical studies. In this work, our aim is to bridge this gap.

We study, from an optimization perspective, the approximation of a posterior by iteratively adding simpler distributions, not necessarily Gaussians, *greedily* [4]. Given that one can find the components of the mixtures, building a mixture is a convex problem which we show have efficient algorithms converging to the global optimum. On the other hand, finding these individual components is non-convex and is known to exhibit several local optima [19, 1]. However, we show that one does not need to solve the inner non-convex problem exactly to achieve the same strong convergence guarantees. The key to our analyses is establishing connections with a functional variant of the well known Frank-Wolfe Algorithm [6]. This connection helps us provide the convergence rate of the greedy variational boosting algorithm with explicit constants in terms of the properties of the distributions.

To the best of our knowledge, these explicit rates have not been known before in the context of variational inference. Moreover, we are also able to provide novel insights, including sufficient conditions for a linear convergence as opposed

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

to the previously conjectured sublinear $\mathcal{O}(1/T)$ rates where T is the number of iterations. Our contributions are both algorithmic and theoretical:

- We connect boosting variational inference (Algorithm 2 in [4]) with the Frank-Wolfe framework [7] enabling us to carefully analyze its convergence. We also thoroughly analyze the assumptions essential to ensure global convergence and present an explicit rate (with constants) for their conjectured $\mathcal{O}(1/T)$ rate.
- We propose simpler variants of the same algorithm that retain the same strong theoretical properties (fixed step size and closed-form line search in Algorithm 1).
- We provide sufficient conditions under which greedy algorithms achieve linear ($\mathcal{O}(e^{-T})$) convergence and therefore are much faster than what was previously conjectured.
- We revisit the Norm-Corrective Frank-Wolfe in Algorithm 3 and give linear convergence guarantees at the cost of a slightly larger computational cost. This algorithm allows one to selectively reoptimize all the weights of the mixture efficiently at every iteration resulting in much faster convergence in practice.

1.1 Related work

Variational approximations by using mixture models has been extensively studied and applied. Perhaps the closest algorithmic setup to our work is that of [4]. They iteratively add components to the mixture greedily, similar to gradient boosting. They require the boosting subroutine to return the optimal density but as we show, this is not required for obtaining their conjectured convergence rate of $O(1/T)$, where T is the number of components added. [20] also use a very similar algorithm in their setup.

Traditional approaches directly target the non-convex problem of finding exactly the first density of the mixture. For this problem, some convergence analysis was carried out by [10], but their rates are only applicable *locally*, as they depend on a smoothness assumption of the KL divergence which does not hold globally unless the iterate is close to the optimum [21]. As we will see, greedy methods have the clear advantage that one does not need to perfectly find the best approximating distribution in the family as previously considered by [4]. A rough approximate solution is enough to ensure convergence.

The Frank-Wolfe Algorithm [6] is a popular algorithm for convex constrained minimization, and is specially attractive because of its cheap projection-free iterations. The algorithm is well studied both theoretically and empirically [14, 9, 8], and has even been applied to non-euclidean spaces. For example, [13] consider a variational objective for approximate marginal inference over the marginal polytope.

The rest of the paper is organized as follows. We review the variational inference problem from an optimization perspective in Section 2 and the necessary and sufficient assumptions that are required to show convergence in Section 3. We present our further algorithmic contributions for the framework in Section 4. We conclude the paper with an experimental proof of concept showing that the proposed methods converge as expected.

Notation. We represent vectors by small letters bold, e.g. $\mathbf{x}$ and matrices by capital bold, e.g., $\mathbf{X}$. For a non-empty subset $\mathcal{A}$ of some Hilbert space $\mathcal{H}$, let $\mathrm{conv}(\mathcal{A})$ denote its convex hull. $\mathcal{A}$ is often called *atom set* in the literature, and its elements are called *atoms*. Given a closed set $\mathcal{A}$, we call its diameter $\mathrm{diam}(\mathcal{A}) = \max_{\mathbf{z}_1,\mathbf{z}_2\in\mathcal{A}} \|\mathbf{z}_1 - \mathbf{z}_2\|$ and its radius $\mathrm{radius}(\mathcal{A}) = \max_{\mathbf{z}\in\mathcal{A}} \|\mathbf{z}\|$. The support of a density function q is a measurable set denoted by capital letters sans serif i.e. Z. Sometimes, we write the domain of a density function with the same notation, but if the domain and the support do not coincide it would be made explicit. The inner product between two density functions $p, q : \mathsf{Z} \to \mathbb{R}$ in L^2 is defined as $\langle p, q\rangle := \int_{\mathsf{Z}} p(z)q(z)dz$.

2 Variational Inference Problem Setting

Say, we observe N data points $\mathbf{x}$ from some space. The Bayesian modelling approach consists of specifying a prior $\pi(\mathbf{z})$ on the data and the likelihood $p(\mathbf{x}|\mathbf{z})$ for some parameter vector $\mathbf{z} \in \mathsf{Z}$ where Z is a measurable set, for example $\mathbb{R}^D$ [1]. One of the challenges of Bayesian inference is that the posterior, obtained through Bayes theorem could be intractable because of a hard to calculate normalization constant. Instead, the joint distribution is usually easier to evaluate i.e. $p(\mathbf{x}, \mathbf{z})$. From a functional perspective, the posterior can be written as $p_{\mathbf{x}}(\mathbf{z}) : \mathsf{Z} \to \mathbb{R}^+_{>0}$. We assume that $p_{\mathbf{x}}(\mathbf{z}) \neq 0\ \forall \mathbf{z} \in \mathsf{Z}$. We use $p_{\mathbf{x}}$ to represent the posterior and p for the joint distribution. The goal of variational inference is to find a density from a constrained set of tractable densities $\mathcal{Q}$ with support Q, $q : \mathsf{Q} \to (0, \infty), q \in \mathcal{Q}$ that is close in the KL sense to the true posterior. The respective optimization problem is:

$$\min_{q\in\mathcal{Q}} D^{KL}(q\|p_{\mathbf{x}}). \tag{1}$$

Note that an unconstrained minimization would yield q to be equal to the true posterior. Thus, one would ideally want the set $\mathcal{Q}$ to be able to represent the parameter space Z well, while still retaining tractability. The objective in Equation (1) is not computable as it requires access to $p_{\mathbf{x}}(\mathbf{z})$ [1]. Instead, it is common practice to maximize the so called the evidence lower bound (ELBO), given by:

$$-\mathbb{E}\left[\log q(\mathbf{z})\right] + \mathbb{E}\left[\log p(\mathbf{x}, \mathbf{z})\right] \tag{2}$$

It is easy to see that maximizing the ELBO, is equivalent to solving the following optimization problem:

$$\min_{q \in \mathcal{Q}} D^{KL}(q||p) \tag{3}$$

While it is well known that D^{KL} is strictly convex in q, its smoothness and strong convexity depends on the choice of $\mathcal{Q}$. [25, 4] showed that the smoothness constant can be bounded by the minimal value obtained by all pdf functions of the densities in $\mathcal{Q}$ in their domain and [25] showed that the strong convexity constant is equal to the respective maximal value.

3 Domain Restricted Densities for Variational Inference

For simplicity in the following we write $D^{KL}(q)$ instead of $D^{KL}(q||p_{\mathbf{x}})$. A sufficient condition for smoothness of the $D^{KL}(q)$ is that the density q is bounded away from zero [4]. We extend this result, showing the necessary condition for global smoothness of $D^{KL}(q)$ to hold on the entire support Q.

Lemma 1. *$D^{KL}(q)$ is Lipschitz smooth on $\mathcal{Q}$ with constant $L = \frac{1}{\epsilon}$ if and only if $q/p_{\mathbf{x}} : \mathsf{Q} \to [\epsilon, \infty)$ with $\epsilon > 0$ i.e. is bounded away from zero in Q. A sufficient condition for smoothness of $D^{KL}(q)$ is $q : \mathsf{Q} \to [\epsilon, \infty)$ with $\epsilon > 0$ i.e. is bounded away from zero in Q.*

Smoothness is a typical assumption which is useful to measure the convergence of optimization algorithms and was employed also in the variational inference setting [10]. Lemma 1 entails that the proofs based on smoothness are valid only in some regions of the space.

Lemma 1 states that if q is a good approximation for $p_{\mathbf{x}}$ (i.e. their ratio is bounded away from zero) then D^{KL} is smooth. If one consider a general density q, a simple way to ensure smoothness is to bound q away from zero. Therefore, we restrict the support of the approximating densities to compact sets. In practice, if the algorithms are initialized well enough, $q/p_{\mathbf{x}}$ can be bounded away from zero. As an example, consider a mixture of two Gaussians with mean in $\mathbb{R}^1$ sufficiently far apart. The boosting approach place a density on one of the modes first and then to the other. Therefore, the gradient of the D^{KL} at the second iteration – $\log(q_1/p_{\mathbf{x}})$ – is arbitrarily large in some parts of the domain depending on how far are the modes and the covariance matrix of q_1. Unfortunately, those are precisely the parts the method targets. Thus, we need to ensure that a significant mass is placed on the second mode as well. For a family of densities which is not bounded away from zero, truncating the support can be seen as a smoothing condition. Initializing with the solution of the mean field variational inference would place some mass on both the modes, so the D^{KL} would appear smooth to the algorithm and truncation might not be necessary. While this is valid in practice, we focus on truncated densities as we need to ensure that the rates we present in this work are valid for any density in the set $\mathcal{A}$ independently of $p_{\mathbf{x}}$ and any initial approximation. Following the line of work of [12, 11] we introduce the information projection from $\mathcal{Q}$ to another set $\mathcal{A}$ where all the densities $q \in \mathcal{A}$ are obtained by truncating densities from $\mathcal{Q}$ and therefore have bounded support A.

Intuitively, variational inference aims at projecting the true posterior on the set of tractable densities $\mathcal{Q}$ (for example factorial in the mean field case). Instead, the boosting variational inference considers mixtures of densities from the set $\mathcal{Q}$, i.e., the optimization is constrained to $\mathrm{conv}(\mathcal{Q})$. The underlying intuition is that $\mathrm{conv}(\mathcal{Q})$ is more expressive than $\mathcal{Q}$. For example, any density can be approximated with a mixture of Gaussian distributions with some appropriate covariance matrix. In order to comment about the rates of convergence, we further restrict the densities in $\mathcal{Q}$ to have a truncated support $\mathsf{A} \subseteq \mathsf{Q}$ and we call this set $\mathcal{A}$. Therefore, $q(\mathbf{z}) : \mathsf{A} \to [\epsilon, \infty)$ with $\epsilon > 0$ and $q(\mathbf{z}) = 0 \; \forall \mathbf{z} \in \mathsf{Q} \setminus \mathsf{A}$. To distinguish a density in $\mathcal{Q}$ and its truncated version in $\mathcal{A}$ we write $q_{\mathsf{Q}} \in \mathcal{Q}$ for the former and $q \in \mathcal{A}$ for the latter.

Therefore, we solve the following optimization problem:

$$\arg\min_{q \in \mathrm{conv}(\mathcal{A})} D^{KL}(q||p_{\mathbf{x}}). \tag{4}$$

As the original posterior $p_{\mathbf{x}}$ has support Z, the choice of $\mathrm{conv}(\mathcal{A})$ as optimization domain is suboptimal wrt $\mathcal{Q}$ or $\mathrm{conv}(\mathcal{Q})$ as its support is a subset $\mathsf{A} \subseteq \mathsf{Q} \subseteq \mathsf{Z}$. We now measure exactly the error which is introduced truncating the support.

Let us first consider the projection of $p_{\mathbf{x}}$ onto A (i.e. restrict the support of p from Z to A). We then have that:

$$p_{\mathsf{A}}(\mathbf{z}) = \begin{cases} \frac{p_{\mathbf{x}}(\mathbf{z})}{\int_{\mathsf{Z}} p_{\mathbf{x}}(\mathbf{z}) \delta_{\mathsf{A}}(\mathbf{z}) d\mathbf{z}}, & \text{if } \mathbf{z} \in \mathsf{A} \\ 0, & \text{otherwise} \end{cases}$$

Where $\delta_{\mathsf{A}}(\mathbf{z})$ is the delta set function. Using the definition of $p_{\mathsf{A}}(\mathbf{z})$ we have that:

$$\begin{aligned} D^{KL}(p_{\mathsf{A}}||p_{\mathbf{x}}) &= \int_{\mathsf{A}} p_{\mathsf{A}} \log \frac{p_{\mathsf{A}}}{p_{\mathbf{x}}} d\mathbf{z} \\ &= \int_{\mathsf{A}} p_{\mathsf{A}} \log \frac{p_{\mathbf{x}}}{p_{\mathbf{x}} \cdot p_{\mathsf{Z}}(\mathbf{z} \in \mathsf{A})} d\mathbf{z} \\ &= -\log p_{\mathsf{Z}}(\mathbf{z} \in \mathsf{A}) \end{aligned} \tag{5}$$

This error represent a tradeoff between the smoothness of the objective (and therefore the rate of the boosting algorithm) and the quality of the approximation. The hope, is that $\mathrm{conv}(\mathcal{A})$ is a richer family of distributions than $\mathcal{Q}$ (i.e. mean field variational inference) and is more tractable than both $\mathcal{Q}$ and $\mathrm{conv}(\mathcal{Q})$ from the optimization perspective. Note that p_{A} does not have to be in $\mathrm{conv}(\mathcal{A})$. If $\mathcal{A}$ contains non-degenerate truncated Gaussian distributions with some appropriate covariance matrix then $\mathrm{conv}(\mathcal{A})$ contains p_{A} which becomes the minimizer $q^\star$ of Equation (4).

In the rest of the paper, we consider the set $\mathcal{A}$ as the set of non degenerate truncated distributions (upper and lower bound on the determinant of the covariance matrix). We assume that the elements in $\mathcal{A}$ have all the following:

A1. truncated densities with bounded support A

A2. $q(\mathbf{z}) \geq \epsilon > 0 \; \forall \; \mathbf{z} \in \mathsf{A}$ and q is bounded from above by M

Under these assumption, we can analyze some of the properties of the optimization domain.

Theorem 2. *The set $\mathcal{A}$ of non degenerate truncated distributions bounded from above and compact support A is a compact subset of $\mathcal{H}$.*

The proof is deferred to the Appendix B. Due to the convenient form of $\mathcal{A}$ we can also compute its diameter as:

Corollary 3. *Given a distribution $q \in \mathcal{A}$, it holds that $\operatorname{diam}(\mathcal{A})^2 \leq \max_{q \in \mathcal{A}} 4\|q\|^2 \leq 4M^2\mathcal{L}(\mathsf{A})$ where $\mathcal{L}(\mathsf{A})$ is the Lebesgue measure of the support A, which is bounded under the assumptions of Theorem 2.*

We will extensively discuss the impact of these assumptions on both the convergence and the approximation quality in Section 4.

4 Functional Frank-Wolfe For Density Functions

In this section, we explain the foundations of boosting via Frank-Wolfe in function spaces. In the analysis of [24], the authors enforce a bounded polytope using functions in L^1 with bounded L_∞ norm. Following the more traditional approaches of [7, 14, 18], we further assume that the functions must have bounded L_2 norm.

The optimization problem we want to solve is:

$$\min_{q \in \operatorname{conv}(\mathcal{A})} f(q). \tag{6}$$

where $\mathcal{A} \subset L^2$ is compact (see Theorem 2) and f is a convex functional over $\operatorname{conv}(\mathcal{A})$ with bounded curvature over the same domain. The curvature is defined as in [7]:

$$C_{f,\mathcal{A}} := \sup_{\substack{s \in \mathcal{A},\, q \in \operatorname{conv}(\mathcal{A}) \\ \gamma \in [0,1] \\ y = q + \gamma(s-q)}} \frac{2}{\gamma^2} D(y, q), \tag{7}$$

where

$$D(y, q) := f(y) - f(q) - \langle y - q, \nabla f(q) \rangle.$$

It is known that $C_{f,\mathcal{A}} \leq L \operatorname{diam}(\mathcal{A})^2$ if f is L-smooth over $\operatorname{conv}(\mathcal{A})$. Due to Lemma 1, we know that the $D^{KL}(q)$ with $q \in \mathcal{A}$ is smooth which implies that the curvature is bounded. Therefore, $D^{KL}(q)$ is a valid objective for the FW framework. In each iteration, the FW algorithm queries a so-called linear minimization oracle (LMO) which solves the optimization problem:

$$\mathrm{LMO}_{\mathcal{A}}(y) := \arg\min_{s \in \mathcal{A}} \langle y, s \rangle \tag{8}$$

for a given $y \in \mathcal{H}$ and $\mathcal{A} \subset \mathcal{H}$. As computing an exact solution of (8), depending on $\mathcal{A}$, is often hard in practice, it is desirable to rely on an approximate LMO that returns an approximate minimizer $\tilde{s}$ of (8) for some accuracy parameter δ and the current iterate q^t such that:

$$\langle y, \tilde{s} - q^t \rangle \leq \delta \min_{s \in \mathcal{A}} \langle y, s - q^t \rangle \tag{9}$$

The LMO is, in general, a hard optimization problem. Therefore, an approximate solution is commonly employed. We discuss a simple algorithm to implement the LMO in Section 4.1. The Frank-Wolfe algorithm is depicted in Algorithm 1. Note that Algorithm 2 in [4] is a variant of Algorithm 1.

Algorithm 1 Affine Invariant Frank-Wolfe

init $q^0 \in \operatorname{conv}(\mathcal{A})$
for $t = 0 \dots T$
  Find $s^t := (\text{Approx-})\mathrm{LMO}_{\mathcal{A}}(\nabla f(q^t))$
  Variant 0: $\gamma = \frac{2}{t+2}$
  Variant 1: $\gamma = \min\left\{1, \frac{\langle -\nabla f(q^t), s^t - q^t \rangle}{C_{f,\mathcal{A}}}\right\}$
  Update $q^{t+1} := (1-\gamma)q^t + \gamma s^t$
end for

Algorithm 1 is known to converge sublinearly with the following rate.

Theorem 4 ([7])**.** *Let $\mathcal{A} \subset \mathcal{H}$ be a compact set and let $f \colon \mathcal{H} \to \mathbb{R}$ be a convex function with bounded curvature $C_{f,\mathcal{A}}$ over $\mathcal{A}$. Then, the Affine Invariant Frank-Wolfe algorithm (Algorithm 1) converges for $t \geq 0$ as*

$$f(q^t) - f(q^\star) \leq \frac{2\left(\frac{1}{\delta} C_{f,\mathcal{A}} + \varepsilon_0\right)}{\delta t + 2}$$

where $\varepsilon_0 := f(q^0) - f(q^\star)$ is the initial error in objective, and $\delta \in (0, 1]$ is the accuracy parameter of the employed approximate LMO.

In some cases convergence might actually be faster (i.e. linear), as stated below.

Theorem 5 ([3])**.** *Let $\mathcal{A} \subset \mathcal{H}$ be a compact set and let $f \colon \mathcal{H} \to \mathbb{R}$ be a strongly convex function with bounded curvature $C_{f,\mathcal{A}}$ over $\mathcal{A}$. Further, assume $q^\star$ lies within relative interior of $\operatorname{conv}(\mathcal{A})$. Then, the Affine Invariant Frank-Wolfe algorithm (Algorithm 1) produces a sequence of iterates that converges goemetrically to $q^\star$*

Discussion: Recall that $C_{f,\mathcal{A}} \leq L \operatorname{diam}(\mathcal{A})^2$. In Theorem 2 we showed that the set of non degenerate truncated

distributions is bounded and in Lemma 1 we showed that the D^{KL} exhibits bounded curvature on $\mathcal{A}$. These results are important as they theoretically justify why we can successfully build a mixture of distributions approximating the posterior in a boosting-like approach. These optimization subtleties were not addressed in [4, 20] but are essential for the convergence of Algorithm 1. In Theorem 5 we introduce the idea that greedily adding a density in a boosting fashion is converging linearly under some additional assumptions. As one can not check whether the optimum is in the relative interior or not, we now focus on the sublinear rate, trying to understand how the assumptions which are made on the target family of distributions influence the convergence.

We now characterize the constants in Theorem 4 for the boosting variational inference problem.

Theorem 6. *Let the set $\mathcal{A}$ satisfy A1 and A2. Then, it holds that:*

$$C_{f,\mathcal{A}} \leq L \operatorname{diam}(\mathcal{A})^2 \leq 4\frac{M^2}{\epsilon}\mathcal{L}(\mathsf{A})$$

Corollary 7. *Under the assumption of Theorem 6, the Affine Invariant Frank-Wolfe algorithm (Algorithm 1) converges for $t \geq 0$ as*

$$f(q^t) - f(q^\star) \leq 8\frac{M^2\mathcal{L}(\mathsf{A})}{\epsilon(\delta^2 t + 2)} + \frac{2\varepsilon_0}{\delta t + 2}$$

where $\varepsilon_0 := f(q^0) - f(q^\star)$ is the initial error in objective, and $\delta \in (0, 1]$ is the accuracy parameter of the employed approximate LMO.

Discussion: As expected, the rate depends on the two main assumptions we introduced: compact support and non degenerate distributions. The support and covariance matrix directly influence the values of ϵ and M. This is substantially different to what is presented in [4, 20] as the explicit assumptions we make allows us to understand how the choices in the distribution family influences the rate. In particular, [4, 20] did not consider the importance of bounded supports, and as we show that it is vital for their conjecture of $O(1/t)$ to hold. Similarly, the sublinear convergence analysis of variational inference of [10] only holds where the ratio $q/p_{\mathbf{x}}$ is bounded (recall from Lemma 1).

If the set $\mathcal{A}$ contains truncated Gaussian distributions with non-degenerate covariance matrix but with small enough determinant to perfectly approximate any density defined on a bounded support it also satisfies A1 and A2. We can now write the suboptimality of the boosting approach, making the tradeoff between the support and the approximation error in term of D^{KL} explicit. Indeed, in Equation (5) we compute the information lost in the projection on a compact support. On the other hand, $q^\star$ represent the projection of p onto the support A as well. Therefore, we can finally give the Theorem that measures the total information loss of boosting variational inference via Frank-Wolfe.

Theorem 8. *Let the set $\mathcal{A}$ of non degenerate truncated Gaussian distribution have compact support $\mathsf{A} \in \mathbb{R}^d$. Further assume that their means are in A and their covariance matrix before truncation is given by $\sigma^2\mathbf{I}$ with $\sigma \geq \sigma_{\min} > 0$ with $\sigma_{\min}$ being small enough such that $p_{\mathsf{A}} \in \operatorname{conv}(\mathcal{A})$. Let $\mathbf{a}$ and $\mathbf{b}$ be the vertices of the diameter of A. Then, the information loss of the Affine Invariant Frank-Wolfe algorithm (Algorithm 1) with some choice of the compact support A converges for $t \geq 0$ as*

$$D^{KL}(q^t||p) \leq \frac{4P(\mathcal{N}(\mathbf{a}, \sigma^2_{\min}\mathbf{I}) \in \mathsf{A})}{\sigma_{\min}^{\frac{d}{2}} 2^{\frac{d}{2}} K^2} \exp\left(\frac{1}{2}\frac{\operatorname{diam}(\mathsf{A})^2}{\sigma^2_{\min}}\right) \frac{1}{\delta^2 t + 2} + \frac{2\varepsilon_0}{\delta t + 2} - \log p(\mathbf{z}_{\mathsf{Z}\setminus\mathsf{A}} = 0)$$

where $\varepsilon_0 = D^{KL}(q^0||p) - D^{KL}(q^\star||p)$, $\delta \in (0, 1]$ is the accuracy parameter of the employed approximate LMO*, p is the true posterior distribution and $K := min_{\boldsymbol{\mu}\in\mathsf{A}} P(\mathcal{N}(\mathbf{z}, \boldsymbol{\mu}, \sigma^2_{max} I) \in \mathcal{A})$. Note that K is bounded away from zero.*

Discussion: Note that the diameter of $\mathcal{A}$ is related to the L2 norm of its elements. If the dimensionality increases, this notion of distance loses meaning (curse of dimensionality). This explicit dependency in the rate is an artifact of the proof technique as a consequence of using the L2 norm. Note that K depends implicitly on d and it decreases whenever d increases and the support of $\mathcal{A}$ remains fixed [5]. Understanding whether the rate is meaningful in high dimensions is a challenging question. Better rates might be achieved with a different notion of distance and are left as future work.

4.1 Implementing the LMO

To solve the LMO problem we revisit a technique well known in the stochastic variational inference framework [23, 10] to account for our constrained scenario. Let us rewrite the optimization problem of Equation (8) exploiting the parametric form of the distributions in $\mathcal{A}$ as:

$$\arg\min_{\theta: s(\theta)\in\mathcal{A}} \langle s(\theta), \nabla f(q^t)\rangle = \arg\min_{\theta: s(\theta)\in\mathcal{A}} \mathbb{E}_{\mathbf{z}\sim s(\theta)}\left[\nabla f(q^t(\mathbf{z}))\right]$$

In order to obtain a valid solution of the LMO problem, we perform projected gradient descent on the parameters of $s(\mathbf{z};\theta)$ with a stochastic approximation of the gradient. Let $\operatorname{proj}_{\mathcal{A}}$ be an operator such that $proj_{\mathsf{A}}[s(\mathbf{z})] \in \mathcal{A}$ holds. This operator is easy to implement in the Gaussian case, as it is reduced to a box constraint for the mean, a constraint on the eigenvalues of the covariance matrix and a truncation. We therefore sample S points from $s(\mathbf{z};\theta)$ and use the

following estimator for the gradient:

$$\begin{aligned}\nabla_\theta \mathbb{E}_{\mathbf{z}\sim s(\mathbf{z};\theta)}\left[\nabla f(q^t(\mathbf{z}))\right] &= \int_{\mathcal{D}} \nabla f(q^t(\mathbf{z}))\nabla_\theta s(\mathbf{z};\theta) d\mathbf{z} \\ &= \int_{\mathcal{D}} \nabla f(q^t(\mathbf{z})) s(\mathbf{z};\theta)\cdot \\ &\quad \nabla_\theta \log s(\mathbf{z};\theta) d\mathbf{z} \\ &\approx \frac{1}{S}\sum_{s=1}^{S} \nabla f(q^t(\mathbf{z}^{(s)}))\cdot \\ &\quad \nabla_\theta \log s(\mathbf{z}^{(s)};\theta) \\ &=: \hat{\nabla}_\theta \mathbb{E}_{\mathbf{z}\sim s(\mathbf{z};\theta)}\left[\nabla f(q^t(\mathbf{z}))\right]\end{aligned} \tag{10}$$

where the $\mathbf{z}^{(s)}$ are sampled from $s(\mathbf{z};\theta)$. This stochastic approximation of the gradient is known to suffer from high variance. Any of the known variance reduction techniques known can be used, e.g., see [23].

We now perform a projected gradient step as:

$$\begin{aligned}s^{l+1}(\mathbf{z};\theta) = proj_{\mathcal{A}}\Big[&s^l(\mathbf{z};\theta) \\ &-\eta\cdot \hat{\nabla}_\theta \mathbb{E}_{\mathbf{z}\sim s^l(\mathbf{z};\theta)}\left[\nabla f(q^t(\mathbf{z}))\right]\Big]\end{aligned} \tag{11}$$

for some stepsize η. Note that $\hat{\nabla}_\theta \mathbb{E}_{\mathbf{z}\sim s^l(\mathbf{z};\theta)}\left[\nabla f(q^t(\mathbf{z}))\right]$ is an unbiased estimator for the gradient as showed in [15]. Further approximation is possible in the data domain as the sampling process is i.i.d. and $\nabla f(q^t) = \log \frac{q^t(\mathbf{z})}{p(\mathbf{x},\mathbf{z})}$. The stochastic LMO algorithm is depicted in Algorithm 2. Notably, an approximate solution of the LMO is sufficient to ensure convergence, even if it is δ-approximate only in expectation [7]. Therefore, relying on cheap estimates of the gradient is well posed in this framework.

Note that the linear problem of Equation (8) without the constraints would be trivially solved by a degenerate distribution placed on the minimum value of the gradient. Therefore, if the set $\mathcal{A}$ contains truncated normal distributions there is a local minimum with covariance $\sigma_{\min}\mathbf{I}$. Therefore, in the experiments we do not learn the covariance matrix. Recall that an approximate solution for the LMO problem is enough to converge.

Algorithm 2 stochastic LMO

init $s^0(\mathbf{z};\theta) \in \mathcal{A}$
for $l = 0$ to L
  Compute $\hat{\nabla}_\theta \mathbb{E}_{\mathbf{z}\sim s(\mathbf{z};\theta)}\left[\nabla f(q^t(\mathbf{z}))\right]$ using Equation (10)
  Compute $s^{l+1}(\mathbf{z};\theta)$ from Equation 11
end while
return $s^L(\mathbf{z})$

4.2 Implementing Line Search

While one can always perform line search on the original objective, we propose a cheaper alternative which still exhibits the same convergence guarantees. Our alternative can become attractive whenever line search on the D^{KL} is too expensive computationally. Let us consider the smoothness quadratic upper bound:

$$f(q^{t+1}) \leq \min_{\gamma\in[0,1]} f(q^t) + \gamma\langle s - q^t, \nabla f(q^t)\rangle + \frac{\gamma^2}{2} C_{f,\mathcal{A}}$$

Instead of performing line search on the original function we compute the stepsize on the quadratic upper bound, which in turns yields a close form solution:

$$\gamma = clip_{[0,1]} \frac{\langle s - q^t, -\nabla f(q^t)\rangle}{C_{f,\mathcal{A}}}$$

This quantity can be efficiently estimated via Monte-Carlo sampling as both s and q^t are easy to sample. To sample from q^t one can first sample one of the distribution forming the ensemble and then sample a point from that distribution.

4.3 Norm-Corrective Frank-Wolfe

In this section, we review the norm-corrective Frank-Wolfe [18] which is presented in Algorithm 3. The main limitation of Algorithm 1 is that each iteration uniformly reduces the weights of all the atoms that are active (i.e. the densities with non zero weight in the mixture). This is undesirable especially in the variational inference setting where the first approximating densities carries a lot of the information. On the other hand, in the early iterations, suboptimal choices can be made as they are considered optimal by the greedy strategy but lose significance as the optimization proceeds. Therefore, it is useful to selectively update all the weights of the mixtures at the same time. For efficiency reasons, we update all the weights at every iteration but rather than minimizing the D^{KL} directly we target its quadratic upper bound as we did in the previous section. This results in a quadratic programming problem on the probability simplex (recall that weights sums to one) for which many efficient solutions are known as T is typically small. The

Algorithm 3 Norm-Corrective Frank-Wolfe

init $q^0 \in \text{conv}(\mathcal{A})$, and $\mathcal{S} := \{q^0\}$
for $t = 0 \dots T$
  Find $z_t := (\text{Approx-})\text{LMO}_{\mathcal{A}}(\nabla f(q^t))$
  $\mathcal{S} := \mathcal{S} \cup \{z_t\}$
  Let $b := q^t - \frac{1}{L}\nabla f(q^t)$
  Variant 0: Update $q^{t+1} := \arg\min_{z\in\text{conv}(\mathcal{S})} \|z - b\|_2^2$
  Variant 1: Update $q^{t+1} := \arg\min_{z\in\text{conv}(\mathcal{S})} f(z)$
  Optional: Correction of some/all atoms $z_{0\dots t}$
end for

name "norm-corrective" is used to illustrate that the algorithm relies on a simple quadratic surrogate function (or upper bound on f), which only depends on the smoothness

constant L. This procedure allows for efficient optimization using standard convex solvers. Finding the closest point in norm can typically be performed much more efficiently than solving a general optimization problem on the D^{KL} over the same domain, which is what the "fully-corrective" algorithm variants require in each iteration (Variant 1). Variant 0 of Algorithm 3 is the equivalent of Variant 1 of Algorithm 1 where the line search on the quadratic upper bound is performed on all the active atoms rather than just the one added in the current iteration, hence the name corrective.

In [18], the authors showed sublinear convergence of Algorithm 3. In this work, we show that under some additional assumptions the convergence is actually linear.

Theorem 9 ([14]). *Let $\mathcal{A} \subset \mathcal{H}$ be a compact set and let $f\colon \mathcal{H} \to \mathbb{R}$ be both L-smooth and μ-strongly convex over the optimization domain. Then, the suboptimality of the iterates of Variant 1 of Algorithm 3 decreases geometrically at each step as:*

$$\varepsilon_{t+1} \leq (1-\beta)\,\varepsilon_t, \tag{12}$$

where $\beta := \delta^2 \frac{\mu \text{PWidth}^2}{L\,\text{diam}(\mathcal{A})^2} \in (0,1]$, $\varepsilon_t := f(q^t) - f(q^\star)$ is the suboptimality at step t and $\delta \in (0,1]$ is the relative accuracy parameter of the employed approximate LMO.

In Theorem 9 we used the notion of pyramidal width:

$$\text{PWidth}(\mathcal{A}) := \min_{\substack{\mathcal{K} \in \text{faces}(\text{conv}(\mathcal{A})) \\ q \in \mathcal{K} \\ r \in \text{cone}(\mathcal{K}-q)\setminus\{\mathbf{0}\}}} PdirW(\mathcal{K} \cap \mathcal{A}, r, q).$$

For an in depth description of the PWidth, see [14]. In the continuous setting, the pyramidal width can be arbitrarily small. For such a reason, quantization of the mean vector is sufficient to ensure that the pyramidal width is bounded away from zero. To obtain a linear convergence rate for Variant 0 of Algorithm 3 one needs to upper-bound the number of "bad steps". This notion comes from the Pairwise and Away step Frank-Wolfe [14]. Let $\mathbf{v}_t$ be the away vertex $v_t = LMO_{\mathcal{S}}(-\nabla f(q^t))$, the exponential decay is not guaranteed when we remove all the weight from $\mathbf{v}_t$ but $|\mathcal{S}_t| = |\mathcal{S}_{t+1}|$. Unfortunately, the tightest known bound for Variant 0 on the number of good steps is $k(t) \geq t/(3|\mathcal{A}|! + 1)$. The rate of Variant 0 is given in the Appendix. While this approach is unsatisfactory, the linear convergence of Frank-Wolfe is an active field of research beyond the scope of this paper. In any case, Algorithm 3 is potentially much faster than Algorithm 1 at the cost of a greater computation complexity per iteration. Furthermore, Algorithm 1 is already linearly convergent if the optimum lies in the relative interior of $\text{conv}(A)$ as shown in [3]. Therefore, in practice, the norm corrective variant can achieve linear convergence and in general converges faster than Algorithm 1.

Discussion In other words, we showed that with the standard assumptions necessary to show sublinear convergence of FW on the variational inference problem, one can use the full FW framework allowing for potentially globally linearly convergent algorithms. After a quantization of the mean values, the convergence is linear as $\text{conv}(\mathcal{A})$ has a finite number of faces. To the best of our knowledge, our results are the first linearly convergent algorithms on the boosting variational inference problem. Furethermore, we identify which assumptions depends on the development of the Frank-Wolfe analysis (bounded pyramidal width for Algorithm 3 or optimum in the relative interior of $\text{conv}(\mathcal{A})$ for Algorithm 1). The relation between PWidth and diam is also known as *condition number* of a set and is related to its eccentricity. Intuitively, a smaller diameter helps the optimization by reducing the size of the search space. On the other hand, in the continuous setting the set $\mathcal{S}$ can contain atoms forming a very narrow pyramid which in the limit gives vanishing pyramidal width. Unfortunately, computing this constant is challenging and it is known only for few examples, see [14].

5 Experimental Proof of Concept

Synthetic data In this section we empirically observe the convergence of Algorithms 1 and 3 on a toy task verifying that the convergence follows our analysis. In particular, we consider two simple forms for the posterior distribution in 1 dimension, a heavy tailed Cauchy distribution and a mixture of Gaussian distributions. We approximate both distributions using the line search and the fully corrective variants of FW. As expected, even after the rough approximations we performed, the fully corrective perfectly fits the target distribution in a very limited number of iterations. To ensure linear convergence we performed quantization of the mean vectors (stride of 0.0001). In both examples we used $L = 15$ and $L = 5$ for line search and the fully corrective respectively. To find the weight in the fully corrective we used standard semidefinite-quadratic programming (cvx solver). As expected, while being more expensive per iteration, Algorithm 3 converges much faster in terms of number of iterations. Therefore, we showed that linear convergence is achievable using Algorithm 3 while minimizing the D^{KL}.

Discussion In [4] the authors perform an extensive experimental evaluation showing the remarkable practical performances of Algorithm 1. On the other hand, they do not truncate the Gaussian distributions in the experiments and still observe excellent convergence properties. Note that, provided that the algorithm is initialized well enough, q/p can be bounded away from zero which entails that there exist a finite L which upper bounds the smoothness constant for a fixed and finite number of iterations. As they regularize the LMO with the log of the determinant of the covariance matrix their set $\mathcal{A}$ has bounded diameter. Therefore, their algorithm is linearly convergent whenever the true posterior is in the relative interior of $\text{conv}(\mathcal{A})$ and sublinear otherwise.

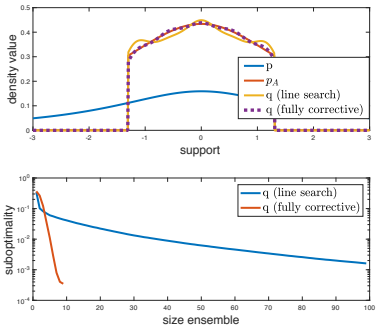

Figure 1: Convergence of Algorithm 3 compared to 1 on a truncated cauchy distribution

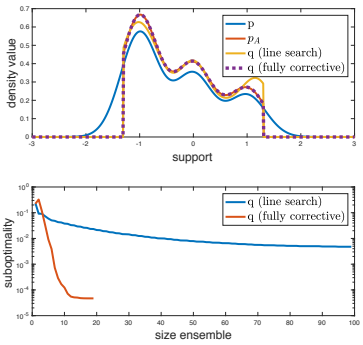

Figure 2: Convergence of Algorithm 3 compared to 1 on a truncated mixture of Gaussian distributions

Real Data To illustrate the practical utility of the boosting framework, we implement the algorithm for the real world application of predicting whether a chemical is reactive or not (i.e. the response vector $\mathbf{y}$ is binary valued) from its features $\mathbf{X}$. We use the CHEMREACT dataset which contains 26733 chemicals, each with 100 features. The training data contains 24059 points, while the rest forms the testing dataset. For the prediction task, we employ the use of Bayesian Logistic Regression with a spherical prior on the regression coefficients $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. If $\mathbf{x}_i \in \mathbb{R}^{100}$ and $y_i \in \{0, 1\}$ are the i^{th} feature vector and response value respectively, then the logistic likelihood function can be written as:

$$\log p(\mathbf{y}|\mathbf{w}; \mathbf{X}) := \sum_i y_i \text{sigmoid}(\mathbf{x}_i^\top \mathbf{w}) + (1 - y_i)[1 - \text{sigmoid}(\mathbf{x}_i^\top \mathbf{w})],$$

where we represent $\mathbf{X}$ as the feature matrix formed by stacking $\mathbf{x}_i$, $\mathbf{y}$ is the response vector, and the sigmoid function is $\text{sigmoid}(\alpha) = \frac{1}{1+\exp(-\alpha)}$. It is straightforward to see that the posterior for the above model does not have a closed form expression, nor is it easy to sample from it. Typically, even for such a relatively simple model, MCMC techniques can be prohibitively slow, and so mean field variational inference is often used.

We use the mean field variational inference to initialize our boosting algorithm, and we show that the mixture of gaussians from the mean field field family gives a better training fit and testing accuracy than the vanilla mean field inference. We reduce the variance of the gradient estimator with the Rao-Blackwellization [2]. To illustrate the importance of the connections with the Frank Wolfe algorithm, we implement three different methods of optimizing over the weights of the mixture. First of all, we implement the line search technique minimizing the original objective already proposed in [4]. However, a simpler fixed step size also guarantees convergence as per the FW analysis, and so does the fully corrective step that optimizes over all the previous weights. This is illustrated in Figure 3. Specifically, we report the training data log-likelihood values to show that the three different techniques offer varying rates of training data fit as expected. The training data fit also translates to the test data accuracy, which we present as the area under the curve (AUC) of the receiver operator characteristic.

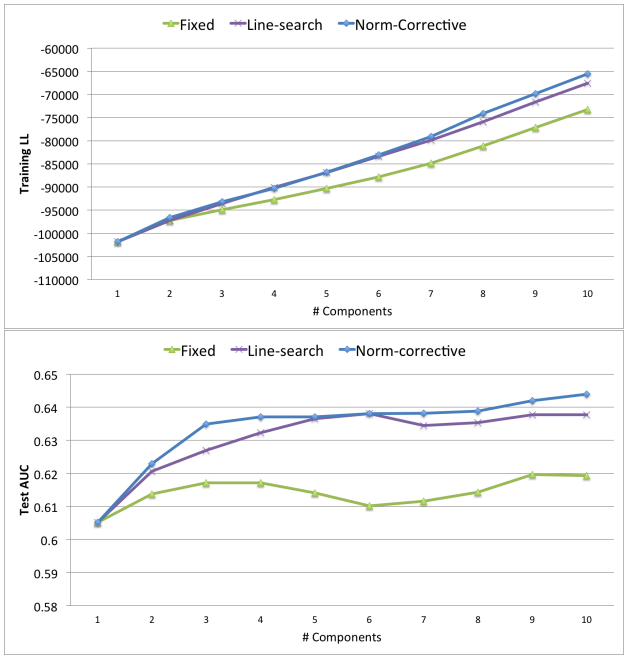

Figure 3: Application of different weights optimization techniques for *ChemReact* dataset: norm corrective (Algorithm 3), line search [4] and decaying fixed step size (Algorithm 1 variant 0)

6 Conclusion

We have presented an in-depth theoretical convergence analysis of the boosting variational inference paradigm, delineating explicitly the rates and assumptions that are required for the previously conjectured sublinear and the presented linear convergence rates.

Acknowledgments: We thank Sahand N. Negahban for the useful discussion. FL is supported by the Max-Planck ETH Center for Learning Systems. RK is supported by NSF Grant IIS 1421729.

References

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

[2] George Casella and Christian P Robert. Rao-blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

[3] Jacques Guélat and Patrice Marcotte. Some comments on Wolfe's away step. *Mathematical Programming*, 35(1):110–119, 1986.

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

[5] John Hopcroft and Ravi Kannan. Foundations of data science. 2014.

[6] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML*, volume 28, pages 427–435, 2013.

[7] Martin Jaggi. Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. In *ICML 2013 - Proceedings of the 30th International Conference on Machine Learning*, 2013.

[8] P. Jain, N. Rao, and I. Dhillon. Structured sparse regression via greedy hard thresholding. In *Advances in Neural Information Processing Systems*, pages 1516–1524, 2016.

[9] A. Joulin, K. Tang, and L. Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *European Conference on Computer Vision (ECCV)*, 2014.

[10] Mohammad Emtiyaz Khan, Reza Babanezhad, Wu Lin, Mark Schmidt, and Masashi Sugiyama. Faster stochastic variational inference using proximal-gradient methods with general divergence functions. *arXiv preprint arXiv:1511.00146*, 2015.

[11] Rajiv Khanna, Joydeep Ghosh, Russell A Poldrack, and Oluwasanmi Koyejo. Sparse submodular probabilistic pca. In *AISTATS*, 2015.

[12] Oluwasanmi O Koyejo, Rajiv Khanna, Joydeep Ghosh, and Russell Poldrack. On prior distributions and approximate inference for structured variables. In *Advances in Neural Information Processing Systems*, pages 676–684, 2014.

[13] Rahul G. Krishnan, Simon Lacoste-Julien, and David Sontag. Barrier frank-wolfe for marginal inference. pages 532–540, 2015.

[14] Simon Lacoste-Julien and Martin Jaggi. On the Global Linear Convergence of Frank-Wolfe Optimization Variants. In *NIPS 2015*, pages 496–504, 2015.

[15] Pierre L'Ecuyer. Note: On the interchange of derivative and expectation for likelihood ratio derivative estimators. *Management Science*, 41(4):738–747, 1995.

[16] Jonathan Q Li and Andrew R Barron. Mixture density estimation. *NIPS - Advances in Neural Information Processing Systems 12*, 1999.

[17] Q. Li. Phd thesis, yale university, 1998.

[18] Francesco Locatello, Rajiv Khanna, Michael Tschannen, and Martin Jaggi. A unified optimization view on generalized matching pursuit and frank-wolfe. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[19] Stephan Mandt, James McInerney, Farhan Abrol, Rajesh Ranganath, and David Blei. Variational tempering. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 704–712, 2016.

[20] Andrew C Miller, Nicholas Foti, and Ryan P Adams. Variational boosting: Iteratively refining posterior approximations. *arXiv preprint arXiv:1611.06585*, 2016.

[21] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2009.

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

[23] Rajesh Ranganath, Sean Gerrish, and David M Blei. Black box variational inference. In *AISTATS*, pages 814–822, 2014.

[24] Chu Wang, Yingfei Wang, Robert Schapire, et al. Functional frank-wolfe boosting for general loss functions. *arXiv preprint arXiv:1510.02558*, 2015.

[25] Xiangyu Wang. *Boosting Variational Inference: Theory and Examples*. PhD thesis, Duke University, 2016.