Boosting Variational Inference: an Optimization Perspective

Francesco Locatello
MPI for Intelligent Systems - ETH Zurich
flocatello@tuebingen.mpg.de

Rajiv Khanna
UT Austin
rajivak@utexas.edu

Joydeep Ghosh
UT Austin
jghosh@utexas.edu

Gunnar Rätsch
ETH Zurich
raetsch@inf.ethz.ch

Abstract

Variational inference is a popular technique to approximate a possibly intractable Bayesian posterior with a more tractable one. Recently, boosting variational inference [11, 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 sublinear/linear 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 set 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 counter-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 [12]. While there has been

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

significant algorithmic and empirical development for studying variational inference using mixture models [11, 4, 8, 9], there have been limited theoretical studies. In this work, our aim is to bridge this gap.

Our contributions are both algorithmic and theoretical:

- We connect boosting variational inference (Algorithm 2 in [4]) with the Frank-Wolfe framework [5] 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 present the Norm-Corrective Frank-Wolfe in Algorithm 2 which enjoys linear convergence (Theorem 5) 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 general.

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 equivalent to maximizing the ELBO, is 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 $\mathcal{Q}$. [14, 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 [14] showed that the strong convexity constant is equal to the respective maximal value. For simplicity in the following we write $D^{KL}(q)$ instead of $D^{KL}(q || p_{\mathbf{x}})$.

3 Domain Restricted Densities for Variational Inference

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 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 mild assumption which is useful to measure the convergence of optimization algorithms and was employed also in the variational inference setting [6]. Lemma 1 entails that the proofs based on smoothness are valid only in some regions of the space. Therefore, we solve the following optimization problem:

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

Where $\mathcal{A}$ contains truncated densities from $\mathcal{Q}$. 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 $\mathsf{A} \subseteq \mathsf{Q} \subseteq \mathsf{Z}$. 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 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. 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

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 D. 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 $\mathrm{diam}(\mathcal{A})^2 \leq \max_{q \in \mathcal{A}} 4\|q\|^2 \leq 4M^2\mathcal{L}(\mathsf{A})$ 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 foundation of boosting via Frank-Wolfe in function spaces. In the analysis of [13], the authors enforce a bounded polytope using functions in L^1 with bounded L_∞ norm. Instead, following the more traditional approaches of [5, 7, 10], we make no assumption on the polytope other than being a compact subset of a Hilbert space $\mathcal{H}$ i.e. the functions must have bounded L_2 norm.

The curvature of a function f is defined as in [5]:

$$C_{f,\mathcal{A}} := \sup_{\substack{s \in \mathcal{A},\, q \in \mathrm{conv}(\mathcal{A}) \\ \gamma \in [0,1] \\ y = q + \gamma(s-q)}} \frac{2}{\gamma^2} D(y, q), \tag{5}$$

where

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

It is known that $C_{f,\mathcal{A}} \leq L\,\mathrm{diam}(\mathcal{A})^2$ if f is L-smooth over $\mathrm{conv}(\mathcal{A})$. 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:

$$\mathrm{LMO}_{\mathcal{A}}(y) := \arg\min_{s \in \mathcal{A}} \langle y, s \rangle \tag{6}$$

for a given $y \in \mathcal{H}$ and $\mathcal{A} \subset \mathcal{H}$. As computing an exact solution of (6), 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 (6) for some accuracy parameter δ and the current iterate q^t such that:

$$\langle y, \tilde{s} - q^t \rangle \leq \delta \min_{s \in \mathcal{A}} \langle y, s - q^t \rangle \tag{7}$$

The Frank-Wolfe algorithm is depicted in Algorithm 1 and the Norm Correctvie variant in Algorithm 2. Note that Algorithm 2 in [4] is a variant of Algorithm 1.

Algorithm 1 Affine Invariant Frank-Wolfe

init $q^0 \in \mathrm{conv}(\mathcal{A})$
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 2 Norm-Corrective Frank-Wolfe

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

Theorem 4. *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 $p_{\mathsf{A}} \in \mathrm{conv}(\mathcal{A})$. 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*

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

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 $K := min_{\boldsymbol{\mu}\in\mathsf{A}} P(\mathcal{N}(\mathbf{z}, \boldsymbol{\mu}, \sigma_{min}^2 I) \in \mathcal{A})$. Note that K is bounded away from zero.*

Theorem 5. *Let $\mathcal{A} \subset \mathcal{H}$ be a compact set and let $f : \mathcal{H} \to \mathbb{R}$ be both L-smooth and μ-strongly convex over the optimization domain. Then, the suboptimality of the iterates of Variant 1 of Algorithm 2 decreases geometrically at each step as:*

$$\varepsilon_{t+1} \leq (1-\beta)\,\varepsilon_t, \tag{8}$$

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

In Theorem 5 we used the notion of pyramidal width:

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

For an in depth description of the PWidth, see [7]. 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 2 one needs to upper-bound the number of "bad steps". This notion comes from the Pairwise and Away step Frank-Wolfe [7]. 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 2 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 $\mathrm{conv}(A)$ as shown in [3]. Therefore, in practice, the norm corrective variant can achieve linear convergence and in general converges faster than Algorithm 1. We provide experimental validation of this claim in the Appendix.

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

[6] 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.

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

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

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

[10] 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.

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

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

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

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

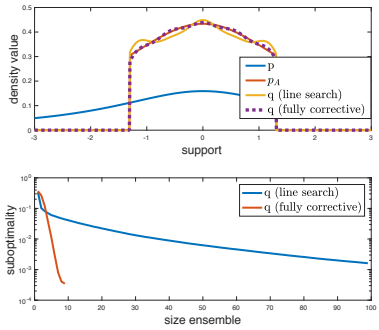

Figure 1: Convergence of Algorithm 2 compared to 1 on a truncated cauchy distribution

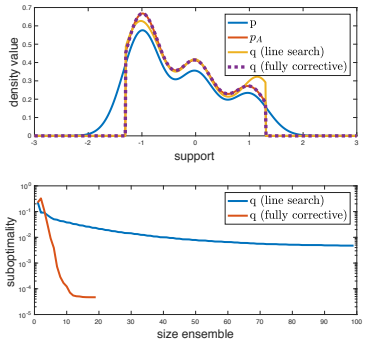

Figure 2: Convergence of Algorithm 2 compared to 1 on a truncated mixture of Gaussian distributions

A Notation.

We represent vectors by small letters bold, e.g. $\mathbf{x}$ and matrices by capital bold, e.g., $\mathbf{X}$. Given a non-empty subset $\mathcal{A}$ of some Hilbert space $\mathcal{H}$ and 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 $\langle p, q\rangle := \int_{\mathsf{Z}} p(z)q(z)dz$.

B Experimental Proof of Concept

Synthetic data In this section we empirically observe the convergence of Algorithms 1 and 2 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[1]). As expected, while being more expensive per iteration, Algorithm 2 converges much faster in terms of number of iterations. Therefore, we showed that linear convergence is achievable using Algorithm 2 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. 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 $\mathrm{conv}(\mathcal{A})$ and sublinear otherwise.

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

[1] http://cvxr.com/cvx/

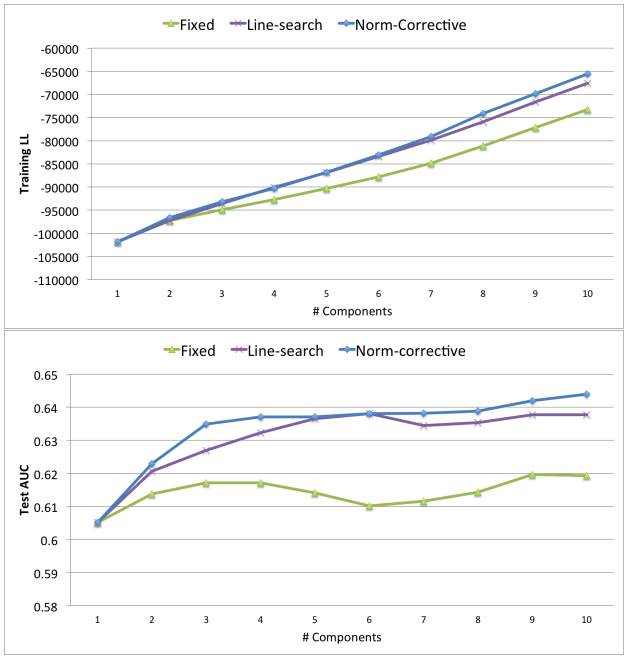

Figure 3: Application of different weights optimization techniques for *ChemReact* dataset: norm corrective (Algorithm 2), line search [4] and decaying fixed step size (Algorithm 1 variant 0)

form 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 ith 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.

C Proof of Lemma 1

If q/p is bounded away from zero, $D^{KL}(q)$ is trivially smooth as it's gradient has bounded norm.

Viceversa, we need to show that if $D^{KL}(q)$ is smooth then q/p is bounded away from zero. Since $D^{KL}(q)$ is smooth, its gradient is absolutely continuous and therefore differentiable almost every-

where with bounded norm. Now, $\nabla D^{KL}(q) = \log \frac{q}{p}$ and its derivative exists everywhere and is bounded except for a ball around the origin with arbitrary small radius. If by contradiction this ball is in the domain $\mathcal{Z}$ (i.e. q/p is not bounded away from zero) this set does not have Lebesgue measure zero and thus $D^{KL}(q)$ is not smooth as its gradient is not absolutely continuous. Note that the D^{KL} can be locally smooth if p is arbitrarily small in the same region of q and they both decrease equally fast.

A sufficient condition is q bounded away from zero everywhere in its support as it would imply $q/p \geq \varepsilon > 0$.

D Proof of Main Results:

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

Proof.

$$\begin{aligned}\operatorname{diam}(\mathcal{A})^2 &= \max_{p,q\in\mathcal{A}} \|p-q\|^2 \\ &\leq \max_{p,q\in\mathcal{A}} (\|p\|+\|q\|)^2 \\ &\leq \max_{q\in\mathcal{A}} 4\|q\|^2\end{aligned}$$

$q \in \mathcal{A}$ is defined everywhere in A and is bounded in infinity norm by assumption. The result of the integral is bounded as A is compact. In particular:

$$\begin{aligned}\|q\|^2 &= \int_{\mathsf{A}} q(\mathbf{z})^2 d\mathbf{z} \\ &\leq M^2 \int_{\mathsf{A}} 1 d\mathbf{z}\end{aligned}$$

Now, $\int_{\mathsf{A}} 1 d\mathbf{z}$ is the Lebesgue measure of the set A which is finite as A is compact and non zero as A is full-dimensional.

For truncated gaussian distributions with diagonal covariance matrix we compute a tighter diameter:

$$\begin{aligned}\|q\|^2 &\leq \int_{\mathsf{A}} q(\mathbf{z})^2 d\mathbf{z} \\ &\leq \int_{\mathsf{A}} \frac{\mathcal{N}(\mathbf{z},\boldsymbol{\mu},\sigma^2 I)^2 \delta_{\mathsf{A}}(\mathbf{z})}{P(\mathcal{N}(\mathbf{z},\boldsymbol{\mu},\sigma^2 I)\in\mathcal{A})^2} d\mathbf{z} \\ &\leq \int_{\mathsf{Z}} \frac{\mathcal{N}(\mathbf{z},\boldsymbol{\mu},\sigma^2 I)^2}{P(\mathcal{N}(\mathbf{z},\boldsymbol{\mu},\sigma^2 I)\in\mathcal{A})^2} d\mathbf{z} \\ &\leq \frac{1}{P(\mathcal{N}(\mathbf{z},\boldsymbol{\mu},\sigma^2 I)\in\mathcal{A})^2} \int_{\mathsf{Z}} \mathcal{N}(\mathbf{z},\boldsymbol{\mu},\sigma^2 I)^2 d\mathbf{z}\end{aligned}$$

and

$$\int_{\mathsf{Z}} \mathcal{N}(\mathbf{z},\boldsymbol{\mu},\sigma^2 I)^2 d\mathbf{z} = \frac{1}{\sigma^d (2\sqrt{\pi})^d}$$

Therefore, the maximum norm is:

$$\frac{1}{\sigma_{min}^d (2\sqrt{\pi})^d min_{\boldsymbol{\mu}\in\mathsf{A}} P(\mathcal{N}(\mathbf{z},\boldsymbol{\mu},\sigma_{min} I)\in\mathcal{A})^2}$$

We call $K^2 := min_{\boldsymbol{\mu}\in\mathsf{A}} P(\mathcal{N}(\mathbf{z},\boldsymbol{\mu},\sigma_{min}^2 I)\in\mathcal{A})^2$. and write:

$$\operatorname{diam}(\mathcal{A})^2 \leq \frac{4}{\sigma_{min}^d (2\sqrt{\pi})^d K^2}$$

□

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

$$C_{f,\mathcal{A}} \leq L \operatorname{diam}(\mathcal{A})^2 \leq \frac{4}{\sigma_{min}^d (2\sqrt{\pi})^d K^2 \epsilon}$$

Proof. The proof if trivial after showing Theorem 2 and recalling that $L = \frac{1}{\epsilon}$ □

Theorem' 4. *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*

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

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 $K := min_{\boldsymbol{\mu} \in \mathsf{A}} P(\mathcal{N}(\mathbf{z}, \boldsymbol{\mu}, \sigma_{min}^2 I) \in \mathcal{A})$. Note that K is bounded away from zero.*

Proof. To show the result we essentially need to compute $C_{f,\mathcal{A}}$ for the particular choice in the theorem statement. Let $\mathbf{a}, \mathbf{b}$ be two points A such that the minimum value of any $q \in \mathcal{A}$ is attained in $\mathbf{b}$ by a density centered in $\mathbf{a}$ (wlog). It is trivial to show that these points are the vertices of the diameter of the support A.

First of all, recall that:

$$\operatorname{diam}(\mathcal{A})^2 \leq \frac{4}{\sigma_{\min}^d (2\sqrt{\pi})^d K^2}$$

The minimal value of any $q \in \mathcal{A}$ can be computed explicitly as by assumption is reached in b:

$$\epsilon = \frac{\mathcal{N}(\mathbf{b}; \mathbf{a}, \sigma_{\min}^2 \mathbf{I})}{P(\mathcal{N}(\mathbf{a}, \sigma_{\min}^2 \mathbf{I}) \in \mathsf{A})} = \frac{1}{L}.$$

Therefore:

$$\begin{aligned} L \operatorname{diam}(\mathcal{A})^2 &\leq \frac{P(\mathcal{N}(\mathbf{a}, \sigma_{\min}^2 \mathbf{I}) \in \mathsf{A})}{\mathcal{N}(\mathbf{b}; \mathbf{a}, \sigma_{\min}^2 \mathbf{I})} \cdot \frac{4}{\sigma_{\min}^d (2\sqrt{\pi})^d K^2} \\ &= \frac{4P(\mathcal{N}(\mathbf{a}, \sigma_{\min}^2 \mathbf{I}) \in \mathsf{A})}{\sigma_{\min}^d (2\sqrt{\pi})^d \mathcal{N}(\mathbf{b}; \mathbf{a}, \sigma_{\min}^2 \mathbf{I}) K^2} \\ &= \frac{4P(\mathcal{N}(\mathbf{a}, \sigma_{\min}^2 \mathbf{I}) \in \mathsf{A})}{\sigma_{\min}^{\frac{d}{2}} 2^{\frac{d}{2}} K^2} \exp\left(\frac{1}{2} \frac{\|\mathbf{a} - \mathbf{b}\|^2}{\sigma_{\min}^2}\right) \\ &= \frac{4P(\mathcal{N}(a, \sigma_{\min}^2 \mathbf{I}) \in \mathsf{A})}{\sigma_{\min}^{\frac{d}{2}} 2^{\frac{d}{2}} K^2} \exp\left(\frac{1}{2} \frac{\operatorname{diam}(\mathsf{A})^2}{\sigma_{\min}^2}\right) \end{aligned}$$

As we assumed that σ_{min} is small enough to approximate perfectly p_{A} the proof is concluded. □

Theorem' 5. *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 0 of Algorithm 2 decreases geometrically at each "good step" as:

$$\varepsilon_{t+1} \leq (1 - \beta)\, \varepsilon_t, \tag{9}$$

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

Proof. The proof is a trivial extension of the one presented in [7]. It only differs in the use of the smoothness upper bound. Let $v_t = LMO_{\mathcal{S}}(-\nabla f(q^t))$ The update of Algorithm 2 yields:

$$\begin{aligned} f(q^{t+1}) &= \min_{q^{t+1} \in \mathrm{conv}(\mathcal{S})} f(q^t) + \gamma \langle \nabla f(q^t), q^{t+1} - q^t \rangle \\ &+ \frac{\gamma^2}{2} L \|q^{t+1} - q^t\|^2 \\ &\leq \min_{\gamma \in [0,1]} f(q^t) + \gamma \langle \nabla f(q^t), \tilde{z}_t - v_t \rangle \\ &+ \frac{\gamma^2}{2} L \|\tilde{z}_t - v_t\|^2 \\ &= f(q^t) - \frac{\langle \nabla f(q^t), \tilde{z}_t - v_t \rangle^2}{2L\|\tilde{z}_t - v_t\|^2}. \end{aligned}$$

This upper bound holds for Algorithm 2 as minimizing the RHS of the first equality coincides with the update of Algorithm 2. The last equality comes from the assumption that we are performing a good step. Using $\varepsilon_t = f(q^\star) - f(q^t)$, we can lower bound the error decay as

$$\varepsilon_t - \varepsilon_{t+1} \geq \frac{\langle \nabla f(q^t), \tilde{z}_t - v_t \rangle^2}{2L\|\tilde{z}_t - v_t\|^2}. \tag{10}$$

The rest of the proof is identical to the one in [7] for the Pairwise Frank-Wolfe. □