Split Variational Inference

Guillaume Bouchard GUILLAUME.BOUCHARD@XEROX.COM
Onno Zoeter ONNO.ZOETER@XEROX.COM
Xerox Research Center Europe, 6, chemin de Maupertuis, 38240 Meylan, France

Abstract

We propose a deterministic method to evaluate the integral of a positive function based on soft-binning functions that smoothly cut the integral into smaller integrals that are easier to approximate. In combination with mean-field approximations for each individual sub-part this leads to a tractable algorithm that alternates between the optimization of the bins and the approximation of the local integrals. We introduce suitable choices for the binning functions such that a standard mean field approximation can be extended to a split mean field approximation without the need for extra derivations. The method can be seen as a revival of the ideas underlying the mixture mean field approach. The latter can be obtained as a special case by taking soft-max functions for the binning.

1. Introduction

Many methods in (Bayesian) machine learning and optimal control have at their heart a large-scale integration problem. For instance the computation of the data log-likelihood in the presence of nuisance parameters, prediction in the presence of missing data, and the computation of the posterior distribution over parameters all can be simply expressed as integration problems.

In this paper we will look at the computation of the integral of a positive function f:

$$I = \int_{\mathcal{X}} f(x)dx, \quad \forall_{x\in\mathcal{X}} f(x) \geq 0 \ . \tag{1}$$

The integrals encountered in real world applications are often of a very high dimension, of a particularly unpleasant form not amenable to analytic solutions, or both.

Recent advances in variational approaches such as mean-field (Opper & Saad, 2001) methods, loopy belief propagation (Frey & Mackay, 1998), and expectation propagation (Minka, 2001) have provided useful approximations for many interesting models. Although they are relatively fast to compute and accurate for some models they can yield poor results if the shape of the function $f(x)$ cannot be accurately captured by the variational distribution. For instance a Gaussian approximation to a multi-modal, an asymmetric, or a heavy-tailed function $f(x)$ will yield coarse results.

A simple but powerful idea that is at the basis of the techniques developed in this paper is to choose *soft-binning functions* $\mathcal{S} = \{s_1, \cdots s_K\}$, such that the original objective function $f(x)$ can be split into K functions, that individually are easier to approximate.

The parametric functions $s_k : \mathcal{X} \times \mathcal{B} \mapsto [0, 1]$ are binning functions on the space $\mathcal{X}$ if

$$\forall_{x\in\mathcal{X},\beta\in\mathcal{B}} \quad \sum_{k=1}^{K} s_k(x;\boldsymbol{\beta}) = 1 \ . \tag{2}$$

Using such binning functions, the original objective can be written in terms of K integrals

$$I_k(\boldsymbol{\beta}) = \int_{\mathcal{X}} s_k(x;\boldsymbol{\beta}) f(x)dx \ ,$$

as

$$I = \sum_{k=1}^{K} I_k(\boldsymbol{\beta}) \ .$$

To estimate I, any form of s_k can be chosen and any method can be used to approximate the I_k's. For instance with s_k "hard" binning functions and constant (resp. affine) functions to approximate $f(x)$ on the support of $s_k(.;\boldsymbol{\beta})$ one obtains the classic rectangular rule (resp. trapezoidal rule). These classic rules work well for low-dimensional integrals and are based on

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

binning functions that divide $\mathcal{X}$ into non-overlapping intervals. We use the term soft-bins to emphasize that it is useful to look at "bins" that have full support on $\mathcal{X}$ and aim to alter the shape of the original function f to make it more amenable to a variational approximation. A second difference from the classical trapezoidal rule is that the presence of the parameter $\boldsymbol{\beta}$ makes it possible to improve the approximation by optimizing over the binning. To this end it will be interesting to consider bounds

$$\underline{I_k}(q_k, \boldsymbol{\beta}) \leq I_k(\boldsymbol{\beta}) \ ,$$

with variational parameters q_k. Bounds allow the use of coordinate ascent style algorithms to optimize both over $\boldsymbol{\beta}$ and the q_k's. In addition, perhaps more importantly, they ensure guaranteed improvements as the number of bins K is increased.

Split variational inference is a generally applicable method and could also be used to construct upper bounds. To demonstrate some of its potential we will focus in this paper on a combination with mean-field techniques. Such a split mean field approach can be seen as a revisit of mixture mean field (Jaakkola & Jordan, 1999; Bishop et al., 1998): the methods share the idea of introducing extra components in the variational approximation with the aim of increasing the lower bound. The main difference is that the multiple components are by construction introduced as a mixture of Gaussians in mixture mean field, whereas in split mean field any choice for the binning functions can be made to introduce extra components in the approximation in more flexible ways.

2. Mean Field Bounds

The local integrals I_k, can be lower bounded using standard mean field approximations. The mean-field bound (Parisi, 1987; Opper & Saad, 2001) can be derived using the fact that the Kullback-Leibler (KL) divergence between two distributions p and q

$$\mathrm{KL}(p||q) \equiv \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \ ,$$

is never negative, and 0 if and only if $p = q$ (See e.g. (Cover & Thomas, 1991)). Using a KL as a slack term with variational parameter q_k we obtain

$$\begin{aligned} I_k(\boldsymbol{\beta}) &\geq I_k(\boldsymbol{\beta}) \times \exp\left(-\mathrm{KL}\left(q_k || \frac{f_k(x;\boldsymbol{\beta})}{I_k(\boldsymbol{\beta})}\right)\right) \\ &= \exp\left(-\int_{\mathcal{X}} q_k(x) \log \frac{q_k(x)}{s_k(x;\boldsymbol{\beta}) f(x)}\right) \qquad (3) \\ &\equiv \underline{I_k}(\boldsymbol{\beta}) \ . \end{aligned}$$

The KL slack term ensures that the bound is tight in the sense that if no further restrictions are placed on $\mathcal{Q}$, the family from which q_k is chosen, at $q_k^* = \frac{f_k}{I_k}$ the bound touches. This also implies that if we assume that the original integral (1) is not tractable to compute, the unrestricted optimization is also not tractable to perform. In mean-field approximations the family $\mathcal{Q}$ is restricted to tractable distributions such as fully factorized distributions, multivariate Gaussians, or tree structured graphical models.

3. Binning Functions: Product of Sigmoids

A product of sigmoids proves to be a flexible and powerful choice for s_k. In its simplest form, without the product, two sigmoidal functions

$$s_1(x;\boldsymbol{\beta}) = \sigma(\beta^T x + \alpha) \qquad (4)$$
$$s_2(x;\boldsymbol{\beta}) = \sigma(-\beta^T x - \alpha) \ , \qquad (5)$$

with $\sigma(x) \equiv \frac{1}{1+e^{-x}}$ form a binning of the space $\mathcal{X}$ since

$$s_1(x;\boldsymbol{\beta}) = 1 - s_2(x;\boldsymbol{\beta}) \ .$$

We can think of this as a soft partioning of $\mathcal{X}$ into two "soft" half spaces.

Multiple splits can be obtained by taking products of sigmoids. By organizing the sigmoidal functions into a tree as in Figure 1 a flexible structure for the bins is obtained: each soft-bin can be split independently. As a comparison, a straightforward use of a product of sigmoids is a special case of the tree construction where all splits in a single level share the same parameters. Formally, each bin s_k in a tree with K leaves is the product of $K-1$ sigmoidal functions, i.e.:

$$s_k(x;\boldsymbol{\beta}) = \prod_{\ell=1}^{K-1} \sigma\left(d_{k\ell}(\beta_\ell^T x + \alpha_\ell)\right)^{|d_{k\ell}|} \ ,$$

where $d_{k\ell} \in \{-1, 0, 1\}$ are constants that follow from the path from the root to leave k. A simple recursion argument shows that this construction satisfies the binning property $\sum_{k=1}^{K} s_k(x, \boldsymbol{\beta}) = 1$ for any $x \in \mathcal{X}$ and any $\beta \in \mathcal{B}$. The key interest is that the product is transformed into a sum in (3) so that expectations are decoupled.

It is instructive to look at the relationship between mixture mean field and split mean field in more detail. Mixture mean field starts with the original objective (1) and introduces a mean field approximation analogous to (3) only once, and directly to I. The family of $\mathcal{Q}$ is subsequently restricted to a *mixture* of

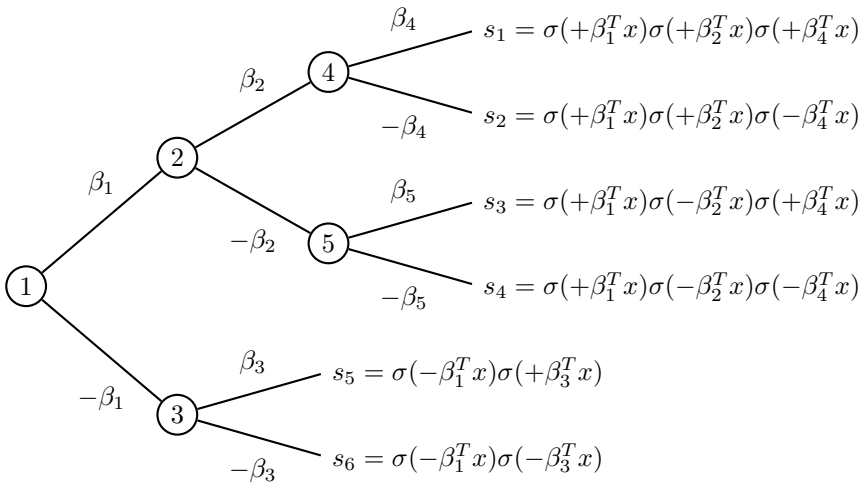

Figure 1. An example of a soft-binning of $\mathcal{X}$ based on products of sigmoids. The tree ensures that the bins defined by level l are independently sub-divided in level $l+1$.

tractable distributions. For easy of notation we consider in this section $\log I$, yielding the mixture mean field lower bound

$$\log I \geq \left(- \int_{\mathcal{X}} \sum_{k=1}^{K} \pi_k q_k(x) \log \frac{\sum_{k'=1}^{K} \pi_{k'} q_{k'}(x)}{f(x)} dx \right) . \tag{6}$$

If we introduce additional variational parameters w_k with $\sum_{k=1}^{K} w_k = 1$ into the split mean field objective:

$$\log \sum_{k=1}^{K} \underline{I_k} \geq \sum_{k=1}^{K} w_k \log \underline{I_k} + \sum_{k=1}^{K} w_k \log w_k \tag{7}$$

$$= \sum_{k=1}^{K} w_k \int_{\mathcal{X}} q_k(x) \log \frac{w_k q_k(x)}{s_k(x) f(x)} dx, \tag{8}$$

and restrict the binning function s_k to be the soft-max

$$s_k(x) = \frac{\pi_k q_k(x)}{\sum_{k'=1}^{K} \pi_{k'} q_{k'}(x)} , \tag{9}$$

we see that we recover mixture mean field (6) if we identify $w_k = \pi_k$.

Note that the additional bound introduced in (7) can be derived using Jensen's inequality and is tight: with $w_k \propto \underline{I_k}$ an equality is obtained.

In mixture mean field, and in split mean field with the above choice for the binning function, a problem remains that the sum inside the log in (6) is hard to deal with. The reinterpretation of mixture mean field as a split mean field algorithm allows for approximations based on changes of s_k: for example the product of sigmoids outlined in Section 3.

Algorithm 1 Split Mean Field

q = InitializeFactoredDistributions
do
  $\boldsymbol{\beta}^{\text{new}}$ = OptimizeBins($q, \boldsymbol{\beta}$)
  for $k = 1, \ldots, K$
    q_k^{new} = UpdateBound($f, s(\cdot, \boldsymbol{\beta}^{\text{new}}), q_k$)
  end
while $\underline{I_k}(\boldsymbol{\beta}^{\text{new}}, q_k^{\text{new}})$ not converged

4. Split Mean Field

The general split mean field algorithm outlined in Algorithm 1 can be used in many different settings. Depending on the characteristics of f, and trade-offs between speed and accuracy one obtains slight variants of the basic algorithm. In this section we discuss several cases. Table 1 gives an overview.

4.1. Continuous Problems

Let us first consider continuous spaces $\mathcal{X} = \mathbb{R}^D$ (the left half in Table 1). For these cases we will concentrate on Gaussian mean field approximations. Inspecting (3) we see that to evaluate the lower bound we need to evaluate $\langle \log s_k(x; \boldsymbol{\beta}) \rangle_{q_k(x)}$ and $\langle \log f(x) \rangle_{q_k(x)}$. If we can compute these integrals and the derivatives with respect to their (variational) parameters, we can optimize the lower bound using a suitable unconstrained gradient based method such as BFGS. Since s_k and f appear as separate terms in (3) due to the log we can consider them one by one. For some $f(x)$ the required expectation and derivatives can be computed analytically. This is for instance the case if f is Gaussian and for the Gumbel, expsin, and products of sigmoids examples from Section 5.1. For other important classes

Table 1. Implementation details for different problem characteristics and different speed-accuracy trade-offs.

	$\mathcal{X} = \mathbb{R}^D$			$\mathcal{X} = \{1, \ldots, M\}^D$
	$f(x)$ Gaussian	$f(x)$ non-Gaussian		
		$\mathbb{E}_{\mathcal{N}(x;\mu,\Sigma)}[\log f(x)]$ Analytic	Gaussian $\underline{f}(x,\eta)$	
Special function for $\mathbb{E}_{\mathcal{N}(x;0,1)}[\log s_k(x)]$	Gradient Ascent	Gradient Ascent	Gradient Ascent	Intractable
Gaussian $\underline{s_k}(\xi_k, \boldsymbol{\beta})$	Section 4.1.2	Gradient Ascent	Section 4.1.2	Section 4.2

of $f(x)$ useful Gaussian lower bounds are known based on local variational parameters η

$$\underline{f}(x;\eta) \equiv \exp\left(-\frac{1}{2}x^T A(\eta)x + b(\eta)^T x + c(\eta)\right) \leq f(x) \; .$$

This is for instance the case for the Cauchy example in Section 5.1 and in general for many heavy-tailed distributions. If Gaussian lower bounds to $f(x)$ are used, additional coordinate ascent steps will be made in η.

The second term to evaluate is $\langle \log s_k(x;\boldsymbol{\beta})\rangle_{q_k(x)}$. With the choice of s_k as a product of sigmoids we only need to consider the case of a single sigmoid again due to the log in (3). One way to proceed is to construct special functions which correspond to 1D integrals

$$\begin{aligned}
\gamma(\mu,\sigma) &= \int \mathcal{N}(x;\mu,\sigma)\log\frac{1}{1+\mathrm{e}^{-x}}dx \; , \\
\nabla_\mu\gamma(\mu,\sigma) &= \int \frac{d}{d\mu}\mathcal{N}(x;\mu,\sigma)\log\frac{1}{1+\mathrm{e}^{-x}}dx \; , \\
\nabla_\sigma\gamma(\mu,\sigma) &= \int \frac{d}{d\sigma}\mathcal{N}(x;\mu,\sigma)\log\frac{1}{1+\mathrm{e}^{-x}}dx \; ,
\end{aligned}$$

either by tabulating or finding solutions akin to the numerical approximation of the erf function[1]. Note that the fact that the Gaussian family is closed under linear transformations implies that a special function which takes only two parameters suffices for the problems where there are two relevant parameters in q_k and two in $\boldsymbol{\beta}$.

If D is of medium size, i.e. such that inverting a $D \times D$ matrix is feasible, the Gaussian variational distributions q_k can have a full covariance matrix. For larger D, to obtain a tractable distribution the q_k's can be restricted to fully factorized distributions $q_k(x) = \prod_{d=1}^D q_{kd}(x_d)$ or to form tree structured models.

[1]We use the trapezoidal methods for ease of implementation in the illustrations. The error can be made as small as machine precision, but formally speaking, the use of the trapezoidal method implies a loss of the bound property.

This completes the description of the top row of the continuous problems in Table 1. It is important to note that the optimization problem factorizes by construction and that the derivations and implementations required to handle the binning functions s_k are independent of the form of f and can be done once. That means that if there already is a standard Gaussian mean field approximation for f no additional work is needed.

4.1.1. Bounding the Bins

The bottom row in Table 1 denotes a different treatment of s_k: instead of evaluating $\langle \log s_k(x;\boldsymbol{\beta})\rangle_{q_k(x)}$ exactly and relying on gradient steps, it is also possible to construct a Gaussian lower bound on s_k. A useful Gaussian lower bound to the logistic sigmoid is provided in (Jaakkola & Jordan, 1996) which is based on the fact that the log is upper bounded by any of its tangents. For the product of sigmoids in s_k this gives the following lower bound

$$\begin{aligned}
s_k(x;\boldsymbol{\beta}) &\geq \exp\left(-\frac{1}{2}x^T A_k x + b_k^T x + c_k\right) \; , &(10)\\
&\equiv \underline{s_k}(x;\xi_k,\boldsymbol{\beta}) \; , &(11)
\end{aligned}$$

where

$$\begin{aligned}
A_k(\{\xi_{k\ell}\},\boldsymbol{\beta}) &= 2\sum_{\ell=1}^{K-1}|d_{k\ell}|\lambda(\xi_{k\ell})(\beta_\ell\beta_\ell^T) \\
b_k(\{\xi_{k\ell}\},\boldsymbol{\beta}) &= \sum_{\ell=1}^{K-1}|d_{k\ell}|(\frac{d_{k\ell}}{2} - 2\alpha_\ell\lambda(\xi_{k\ell}))\beta_\ell \\
c_k(\{\xi_{k\ell}\},\boldsymbol{\beta}) &= \sum_{\ell=1}^{K-1}|d_{k\ell}|\Big(\lambda(\xi_{k\ell})(\xi_{k\ell}^2 - \alpha_\ell^2) \\
&\qquad + \frac{d_{k\ell}\alpha_\ell + \xi_{k\ell}}{2} - \log(1+e^{\xi_{k\ell}})\Big)
\end{aligned}$$

and

$$\lambda(y) = \frac{1}{2y}\left(\sigma(y) - \frac{1}{2}\right) \; .$$

The bound holds for any $\xi_{k\ell} \in \mathbb{R}$.

The lower bound (10) is a Gaussian potential: its precision matrix is of rank at most $K-1$ and its scaling factor is such that it is below the product of sigmoids everywhere.

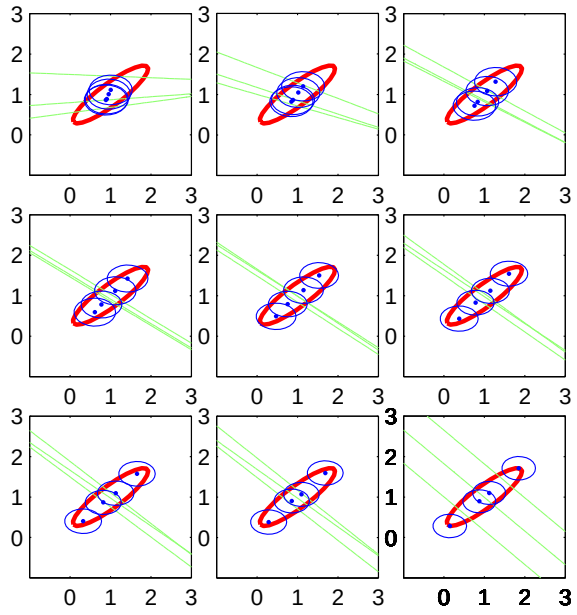

Figure 2. First 9 iterations of the 4-components split mean field algorithm. The red ellipse is the true function. The blue circles are the factored distribution approximating it. The green lines represent the component splits.

In the case of a Gaussian lower bound on s_k and a Gaussian lower bound or exact Gaussian $f(x)$ we obtain closed form updates for all parameters in a coordinate ascent algorithm. This makes it possible to work with very high numbers of bins, K. Using a second underline to denote the introduction of additional local variational parameters, the objective we aim to optimize can be expressed as

$$\underline{\underline{I}}(\boldsymbol{\beta}, \eta, \xi, \{q_k\}) = \sum_{k=1}^{K} \exp\left(- \int_{\mathcal{X}} q_k(x) \log \frac{q_k(x)}{\underline{s_k}(x;\xi,\boldsymbol{\beta}) \underline{f}(x;\eta_k)} dx \right) \quad (12)$$

Update of $\boldsymbol{\beta}$. The Gaussian lower bound $\underline{s_k}$ is a product of Gaussian potentials. The log in (12) makes that we can treat each independently. On inspection of (12) and the Gaussian form of $\underline{s_k}$ we notice that the optimization problem with respect to parameters (β_l, α_l) for a single Gaussian potential, with all other parameters fixed, is an unconstrained quadratic function.

Update of η. The optimization of η is just as for $\boldsymbol{\beta}$ an unconstrained quadratic optimization problem.

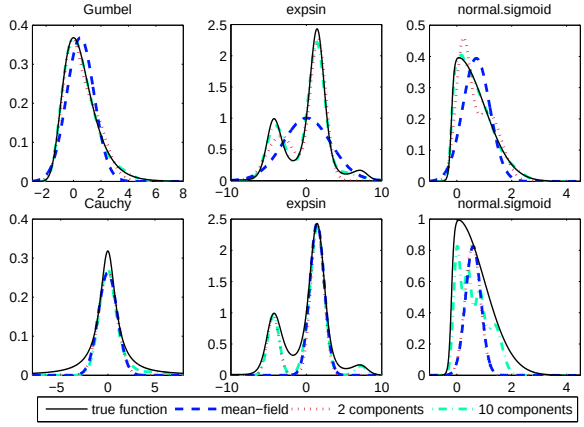

Figure 3. Split mean-field applied to 1D functions. The bottom row shows the result of the algorithm based on quadratic lower bounds to the splits $\log s_k$ and the log-functions $\log f$. The top row is based on the exact computation of $\langle \log s_k \rangle_{q_k}$.

Update of ξ. Perhaps more surprisingly than for $\boldsymbol{\beta}$ and η, a closed form solution for ξ exists as well (Jaakkola & Jordan, 1996).

Update for q_k. We observe that (12) has a KL form. Hence q_k is minimized if we take $q_k = \underline{s_k f_k}$. If q_k is restricted to be a fully factorized distribution we have the similar arguments for each of the marginals q_{kd}.

To complete the discussion of the continuous cases in Table 1: if s_k is lower bounded, but f_k is non-linear, optimization of $\boldsymbol{\beta}$ and ξ can be performed in closed form as outlined above, but the optimization with respect to q_k will require gradient based methods.

4.2. Discrete Problems

The binned mean field algorithm is not restricted to continuous $\mathcal{X}$. In a discrete setting, where we interpret the integral in (1) as a high-dimensional sum, we can obtain useful bounds based on factorized discrete distributions $q_k(x) = \prod_{i=1}^{d} \prod_{j=1}^{M} \pi_{ij}^{x_{ij}}$ where $x_{ij} \in \{0,1\}$ and $\sum_j x_{ij} = 1$ for all variable i and π_{ij} are the variational parameters (in the simplex). The update for q_k is

$$q_k(x) = \text{Discrete}(x|\pi_{ij}) \ ,$$

where

$$\pi_{ij} \propto \exp\{A_{ij}(\chi_k) + \Lambda_{kij} + b_{ij}(\chi_k) + \nu_{kij}\} \ .$$

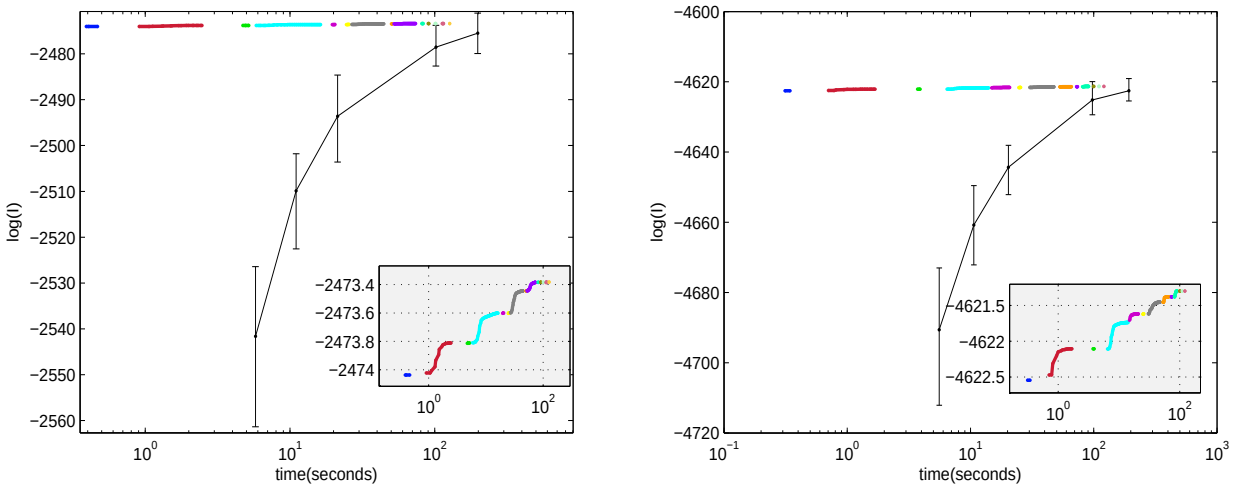

Figure 5. Bayesian logistic regression computation on Australian (13 dimensions) and Diabetes (9 dimensions) datasets. The $y-$axis indicates the value of $\log \int_\Theta p_0(\theta) \prod_i p(y_i|x_i,\theta) d\theta$. The black error bars indicates the 80% confidence interval from AIS. The colored points indicate the values of the split mean field bound, where component additions correspond to a change of color. The small plot at the bottom right is a zoom of the large axis to see the bound improvement as a function of the CPU time.

5. Experiments

5.1. Toy examples

In Figure 3, we show several example of functions whose integrals can be efficiently approximated using split mean-field. Standard algebra enable use to compute a quadratic lower bound to the Cauchy pdf, the expsin function ($f(x) = e^{-\frac{x^2}{20}+\sin(x)}$) and the sigmoid and therefore apply algorithm 2. In the top row, we applied the approximate techniques based on the lower bound of the binning function. We see that even in the 10-component case, the approximation is still far from optimal, but the algorithm is very fast in practice and allows us to work with several thousands of components. One the bottom row, the use of exact integrals of the sigmoids enable a better fit to the functions, with a nearly perfect approximation in the 10-component case.

Correlated Gaussian The method is illustrated in Figure 2. In this example, we used a 2D function $f(x) = \mathcal{N}\left(x| \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.6 & -0.3 \\ -0.3 & 0.7 \end{bmatrix}^{-1}\right)$. We use uncorrelated Gaussian distributions as mean field components: $q_k(x) = \mathcal{N}(x|m_k, S_k)$ where S_k is a diagonal matrix. For standard mean field, this example was described in detail in Bishop (2006), Chapter 10. We used exactly the same updates for the optimization of the factored distributions. In this example, the use of a 2-components (resp 4-components) mixture reduces the relative error of factored mean field by more than 40% (resp 55%) relative to the exact value of the 2D integral.

A non-trivial 2D integral is shown on Figure 4. It corresponds to the product of a normal distribution with two sigmoids. We used the function γ previously defined to do exact computation of the integrals, both for the means and the splits. To have an idea of the improvement over standard mean-field, we plotted the evolution of the free energies during the iterations of the algorithm.

The mixture-based approach consistently outperforms the standard mean-field. We also see that the inclusion of new splits does not decrease the likelihood, because the mixture components are initialized based on the previous solution. The two-dimensional contours of the function show that the orientation of the mixture components is aligned with the sharp angles of the original function (see e.g. the two-component mixture in the top right panel), showing that complex dependencies can be models through this approach. The last contour plot containing 14 components is not perfect but the left curve shows that its integral is very close to the optimal one.

Figure 6 shows a comparison of sigmoidal split functions and the softmax from Eq. 9 (MMF). As discussed in Section 3 the sum of the softmax in the entropy term of the objective requires additional approximations. The experiments are based on the advanced ad-

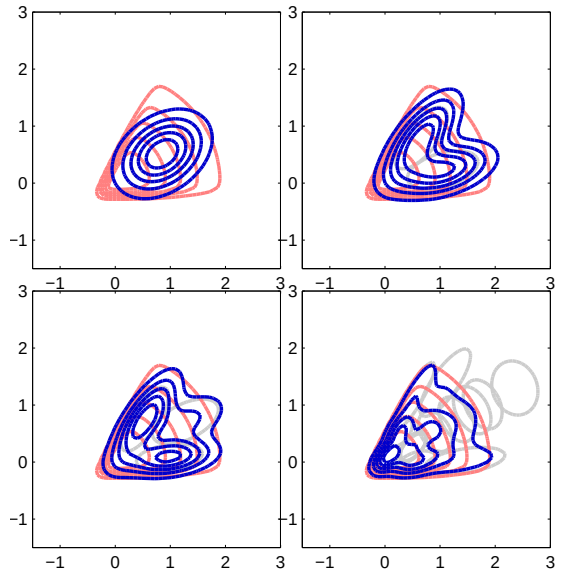

Figure 4. A 2D example of split Mean-Field for $f(x) = \mathcal{N}(x)\sigma(20x_1 + 4)\sigma(20x_2 - 10x_1 + 4)$. The true function is given by light contours, the mixture approximation by dark contours and the individual Gaussian components by light elliptic contours. Iterations 1, 2, 3 and 300 are shown. There were 14 components at the 300th iteration.

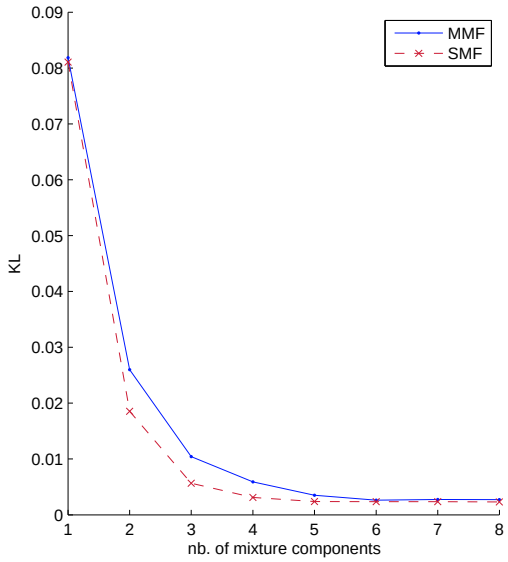

Figure 6. A comparison between two estimates of the Gumbel distribution. One is based on sigmoidal soft-binning functions, the other on the soft-max (MMF). Only the best restart of MMF is shown.

ditional bounds from (Jaakkola & Jordan, 1999). For small K we see a relative increased performance of the sigmoidal split functions. We believe that for this example this is largely due to the absence of the additional bound. The most important difference observed in our experiments is that the soft-max split function is very sensitive to initialization in contrast to the tree of sigmoids (only the best result for the soft-max split function is shown).

5.2. Bayesian inference

We tested our method on the popular logistic regression model $p(y|x,\theta) = \sigma(y\theta^T x)$ where $x \in \Re^D$ is the input vector and $y \in \{-1, 1\}$ is the output label. The goal is to have a accurate approximation of the posterior distribution whose normalization factor is a D-dimensional integral $I = \int_\Theta p_0(\theta) \prod_{i=1}^n p(y_i|x_i,\theta) d\theta$. For every dataset, we choose the parameters of the Gaussian prior p_0 using a simple heuristic: on half of the dataset we randomly build 100 datasets (with replication) of 10 data points. For each of these small datasets, we learned the MAP estimator of the logistic regression with unit-variance prior. Finally, the mean and variance of p_0 were set to the empirical mean and variance of the estimators. On the second half of the dataset, we randomly chose 10 observation points for the likelihood $\prod_{i=1}^n p(y_i|x_i,\theta)$. The integral I is therefore a product between a Gaussian and 10 sigmoids, which can be approximated using the split mean-field approach or by Annealed Importance Sampling (AIS), as described in (Neal, 1998).

Typical examples on the public datasets with binary labels are shown in Figure 5. We see that during the first iterations, the evaluation of the integral by AIS is inaccurate compared to split mean-field which gives very stable values with a constant improvement over time. However, asymptotically, the AIS tends to an unbiased estimate of the integral, that is larger than the best value of our algorithm. As shown in the zoomed axes, there is a relative bound improvement of $0.7 \approx \log(2)$ in the Australian dataset (resp $1.2 \approx \log(3)$ in the Diabetes dataset), which means that the integral given by split mean field is two times bigger (resp. three times bigger) than the standard mean-field approach. Such differences are likely to create strong biases in the estimates of the Bayes factor when doing model selection (Beal & Ghahramani, 2003).

6. Discussion

In this paper we have revived the mixture mean field idea of improving lower bounds to difficult high-dimensional integrals by increasing the number of components in the approximation. The interesting twist

is that the additional components are not introduced by working with variational mixtures directly. Instead a suitable set of soft-binning functions s_k are chosen such that the original integral can be split into a sum of integrals. The hope is that even if $f(x)$ is hard to approximate directly, the soft-binning functions can be chosen such that the individual $s_k(x)f(x)$ *can* be approximated accurately and efficiently.

This approach is very general, and the use of mean field approximations for the K individual integrals is only one possibility. A benefit of the mean field choice is that the lower bound ensures that the approximation improves as K is increased. Also, as discussed in Section 4.1.2, if a standard Gaussian mean field implementation exists for a particular problem, the split mean field algorithm can be used without any additional implementation overhead. Lastly, by choosing soft-max functions for the binning functions s_k we find that we can retrieve the established mixture mean field approach. Other choices for the local approximations are a worthwhile pursuit.

The insight that multi-component approximations can be created by suitable choices for the binning function introduces many degrees of freedom over a mixture of Gaussians choice. A flexible, powerful, and relatively efficient choice is a product of sigmoids assembled in a decision tree such that half spaces can be split independently. For small examples where very accurate brute-force estimates of the objective function can be found, we observe that even the introduction of a single extra component reduces the error (gap between lower bound and the exact integral) by typically 40%. Although the algorithm is fast enough to handle a large number of bins (hundreds), we found that in a time that would be reasonable for practical use a gap will always persist. Reductions of 40% of the error in 10 times the computation time of standard mean field and more than 60% in 100 times are typical. This is observed for multi-modal examples we have tried, that are arguably particularly suited to the method, but also for heavy-tailed and asymmetric examples.

For large examples from the UCI dataset we see similar increases in the estimate of the log-likelihood as the number of bins increases (an increase of the likelihood by a factor of 2 or even 3). It is for these larger problems impossible to accurately assess the relative reduction in error, since annealed importance sampling was not able to give reliable estimates of the exact integral. Annealed importance sampling is considered to be among the state of the art in settings where accurate estimates are of more concern than efficiency. We have not tried other methods.

Split mean field with the choices made here has proven to be an effective improvement upon standard mean field approximations in time critical applications. Many generalizations and alternative uses of split variational inference remain to be explored. Also of great interest is a careful study of the behavior of the approximation as K reaches infinity.

References

Beal, M. J., & Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics 7* (pp. 453–464). Oxford University Press.

Bishop, C. M., Lawrence, N., Jaakkola, T., & Jordan, M. I. (1998). Approximating posterior distributions in belief networks using mixtures. *Advances in Neural Information Processing Systems* (pp. 416–422). MIT Press.

Cover, T., & Thomas, J. (1991). *Elements of information theory.* Series in Telecommunications. John Wiley & Sons. 1st edition.

Frey, B. J., & Mackay, D. J. C. (1998). A revolution: Belief propagation in graphs with cycles. *Advances in Neural Information Processing Systems* (pp. 479–485). MIT Press.

Jaakkola, T., & Jordan, M. (1996). A variational approach to Bayesian logistic regression problems and their extensions. *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics.*

Jaakkola, T., & Jordan, M. (1999). *Learning in graphical models*, chapter Improving the Mean Field Approximation via the use of Mixture Distributions, 163–173. MIT Press.

Minka, T. (2001). Expectation propagation for approximate bayesian inference. *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)* (pp. 362–369). San Francisco, CA: Morgan Kaufmann Publishers.

Neal, R. M. (1998). Annealed importance sampling. *Statistics and Computing*, *11*, 125–139.

Opper, M., & Saad, D. (Eds.). (2001). *Advanced mean field methods.* MIT Press.

Parisi, G. (1987). *Statistical field theory.* Addison-Wesley.