Split Variational Inference

Guillaume Bouchard Onno Zoeter

Xerox Research Centre Europe
6 Chemin de Maupertuis
38240 Meylan, France
{Guillaume.Bouchard, Onno.Zoeter}@xrce.xerox.com

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 Overview

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 methods, loopy belief propagation, and expectation propagation 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 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 [2, 1]: 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.

Figure 1 shows results for Bayesian logistic regression.

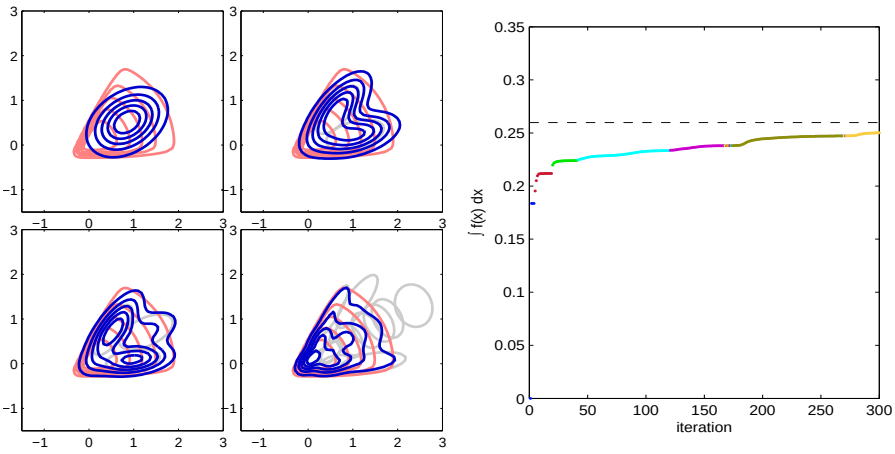

Figure 1: Bayesian logistic regression $f(x) = \mathcal{N}(x)\sigma(20x_1 + 4)\sigma(20x_2 - 10x_1 + 4)$. The left plot shows contour plots for the posteriors over parameters for a 2D dataset. Dark curves represent mean field and split mean field posteriors, light curve is exact. Individual ellipses are for the Gaussian components in the split mean field approximation. Iterations 1,2,3, and 300 are shown, there are 14 components at the 300th iteration. The right plot shows the exact evidence (dashed) and split mean field lower bound as function of iteration number. Every time the lower bound curve changes color a new split was introduced.

References

[1] Christopher M. Bishop, Neil Lawrence, Tommi Jaakkola, and Michael I. Jordan. Approximating posterior distributions in belief networks using mixtures. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10, pages 416–422. MIT Press, 1998.

[2] T. Jaakkola and M. Jordan. *Learning in Graphical Models*, chapter Improving the Mean Field Approximation via the use of Mixture Distributions, pages 163–173. MIT Press, 1999.