Split Variational Inference

Guillaume Bouchard
Onno Zoeter

Xerox Research Centre Europe

June 2009

High-Dimensional Integrals in Machine Learning

From a numerical point of view many core problems in machine learning are the computation of high-dimensional integrals

- Integrating over latent or nuisance parameters (E-step);
- Bayesian treatment of parameters;
- Computing general marginals;
- $\dots$

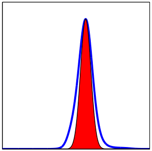
We will concentrate on the **general integral**

$$I = \int_{\mathcal{X}} f(x)dx \ .$$

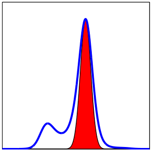
Variational approximations

Variational approximations (Mean Field, Belief Propagation, ...): approximate integration as optimization over a tractable family.

Fast and effective for "easy" models



Problematic for harder problems: asymmetry, multi-modality, ...



Split variational inference provides a general way to improve basic variational approaches using an any-time algorithm.

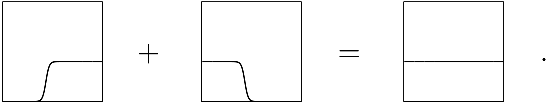
How to improve a basic variational approximation?

Basic idea: **soft-binning functions** $s_k : \mathcal{X} \times \mathcal{B} \mapsto [0,1]$ “split” and “focus” the integration problem.

A collection of K soft-binning functions satisfies

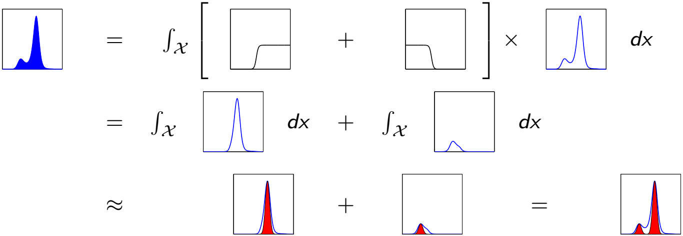
$$\forall_{x \in \mathcal{X}, \beta \in \mathcal{B}} \quad \sum_{k=1}^{K} s_k(x; \boldsymbol{\beta}) = 1 \ .$$

Example: a sigmoid function and its complement



Split an integral

$$
\begin{aligned}
I &= \int_{\mathcal{X}} f(x)dx \\
&= \int_{\mathcal{X}} [s_1(x) + s_2(x)] f(x) dx \\
&= \underbrace{\int_{\mathcal{X}} s_1(x) f(x) dx}_{\substack{I_1 \\ \approx \\ \tilde{I}_1}} + \underbrace{\int_{\mathcal{X}} s_2(x) f(x) dx}_{\substack{I_2 \\ \approx \\ \tilde{I}_2}}
\end{aligned}
$$



Split Mean Field

The soft-binning trick is simple, powerful, and very general.

This talk: **Split Mean Field**, a Split Variational Approach with Gaussian **Mean Field** approximations within each bin.

The Mean Field approximation is based on the positivity of the Kullback-Leibler divergence and yields a **lower bound**

$$\begin{aligned} I_k(\beta) &\equiv \int_{\mathcal{X}} s_k(x;\beta) f(x) dx \\ &\geq \exp\left(- \int_{\mathcal{X}} q_k(x) \log \frac{q_k(x)}{s_k(x;\beta) f(x)} \right) \\ &\equiv \underline{I_k}(\beta, q) \ . \end{aligned}$$

For SMF assume $\forall_x f(x) \geq 0$, i.e. a potential.

Optimizing Bins and Local Approximations

- The bins have free parameters β.
- Bound $I \geq \sum_{k=1}^{K} I_k(\beta)$ allows **principled maximization** over β.

A basic **coordinate ascent** approach works very well in practice.

With K fixed this alternates between

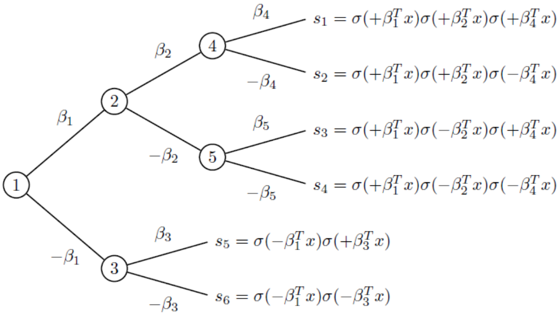
- Optimize bin parameters β with $\{q_k\}$ fixed.
- Optimize local approximations $\{q_k\}$ with β fixed.

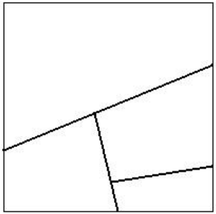
Flexible binning functions

Soft-max function $\frac{e^{\beta_k^T x}}{\sum_{k'} e^{\beta_{k'}^T x}}$ is notably hard to integrate.

A **product of sigmoids** is simple and effective.

A **hierarchy** is particularly flexible.

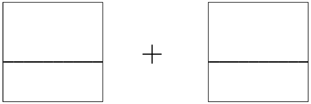




Incrementally growing K

Increase K iteratively.
When $\underline{L}$ plateaus add an extra split.

- Keep old tree fixed
- Decide on a leaf node to split
- Initialize the split with $\sigma(\beta^\top x + \alpha)$ with $\beta = 0$.



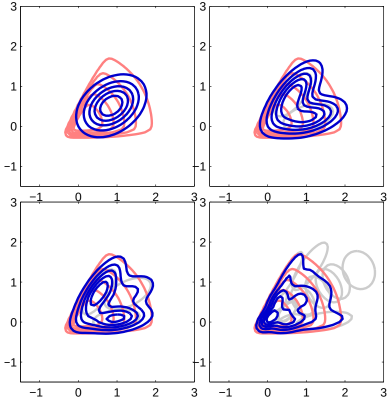
  Approximation does not change after split

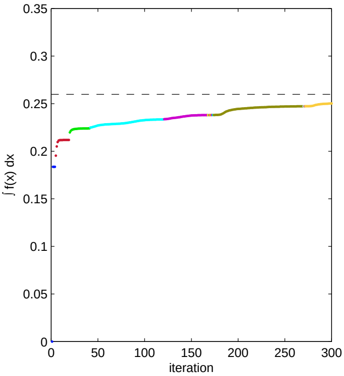
- Optimizing further can only improve approximation.

A two dimensional example

$$f(x) = \mathcal{N}(x)\sigma(20x_1 + 4)\sigma(20x_2 - 10x_1 + 4)$$

Exact integral $= 0.261$





Free energy computation

$$\underline{I_k}(\beta, q_k) \equiv \exp\left(-\int_{\mathcal{X}} q_k(x) \log \frac{q_k(x)}{s_k(x;\beta) f(x)}\right) .$$

Since q_k is Gaussian, the entropy term is easy.

The **energy term involving** $f(x)$ consists of

- Gaussian integral

$$\mathbb{E}_{\mathcal{N}(x;\mu,\Sigma)}[\log f(x)]$$

  - **Same as in standard MF**
  - Sometimes analytic
  - Otherwise based on an additional lower bound

Free energy computation

$$\underline{I}_k(\beta, q_k) \equiv \exp\left(-\int_{\mathcal{X}} q_k(x) \log \frac{q_k(x)}{s_k(x;\boldsymbol{\beta}) f(x)}\right) .$$

The energy term involving $s_k(x)$ has product of sigmoids in the log

- Separate Gaussian integrals

$$\mathbb{E}_{\mathcal{N}(x;\mu_k,\Sigma_k)}[\log \sigma(\beta_l^\top x + \alpha_l)] = \mathbb{E}_{\mathcal{N}(z;m_{kl},v_{kl})}[\log \sigma(z)] \quad \textbf{(1D!)}$$

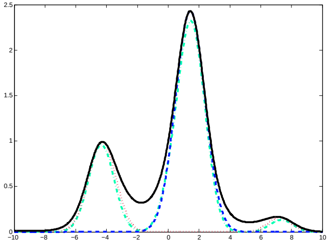
  - **Standard approach**: bound sigmoid by Gaussian [Jaakkola & Jordan 96].
    Fast updates (closed form), but loose since based on

$$\forall_x g(x) \geq \underline{g}(x) \Rightarrow \int_{\mathcal{X}} g(x)dx \geq \int_{\mathcal{X}} \underline{g}(x)dx .$$

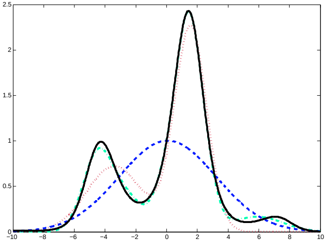
  - **Alternative**: Exact integration using special functions (as for *erf* function) or table indexed by (μ, σ).

Exact versus approximate treatment of bin terms

$$\mathbb{E}_{\mathcal{N}(x;\mu,\sigma)}[\underline{\log s_k}(x)]$$



$$\mathbb{E}_{\mathcal{N}(x;\mu,\sigma)}[\log s_k(x)]$$



Mixture Mean Field

Split Mean Field revisits the **Mixture Mean Field** idea. [Jaakkola&Jordan,1996;Lawrence et al., 1997]

MMF: a single mean field approximation with q **a mixture**

$$q(x) = \sum_{k=1}^{K} \pi_k q_k(x) \ ,$$

yielding the bound

$$\underline{I}(\{\pi_k, q_k\}) \equiv \exp\left(- \int_{\mathcal{X}} \sum_{k=1}^{K} \pi_k q_k(x) \log \frac{\sum_{k=1}^{K} \pi_k q_k(x)}{f(x)} \right) \ .$$

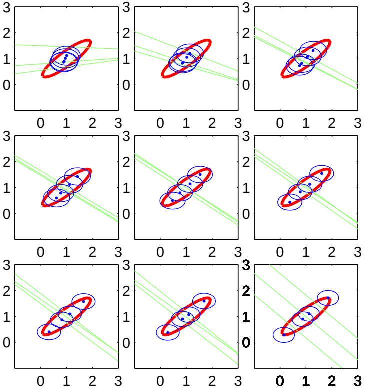
One can show that this is a special case of SMF with s_k **soft-max functions**:

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

Entropy term requires additional approximations.

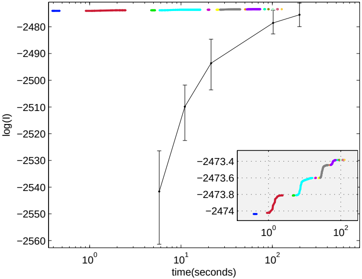
Correlated Gaussian example

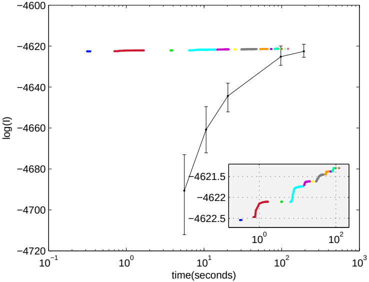
- Full covariance Gaussian distribution
- Approximated by a mixture of diagonal covariance Gaussian distributions



Bayesian inference

- logistic regression model $p(y|x, \theta) = \sigma(y\theta^T x)$
- posterior on 10 observations
- $I = \int_\Theta p_0(\theta) \prod_{i=1}^{10} p(y_i|x_i, \theta) d\theta$
- compared with classical Mean Field and Annealed Importance Sampling (AIS)
- relative bound improvement over Mean Field:
  - $2 \approx e^{0.7}$ in the Australian dataset
  - $3 \approx e^{1.2}$ in the Diabetes dataset





Conclusion

- **Split Variational Inference** = “divide and conquer” idea:
  1. take your favorite bounding technique
  2. choose a split function family
  3. alternatively optimize the splits and the bound in each bin
- improves Bayesian inference as the number of bins increases
- key messages:
  - sigmoid decision tree: flexible and efficient choice
  - exact sigmoid integrals instead of Jaakkola’s bound is much more accurate
- future research
  - use upper bounds, e.g. TRW
  - convergence analysis as K reaches infinity.