Black-box α-divergence minimization

Galina Boeva

MIPT, 2023

October 24, 2023

Motivation

Main idea

This is a general method for approximating α - divergence, which combines approaches such as variational Bayes and expectation propagation.

Definition

Posterior distribution

$$p(\theta|\mathcal{D}) \propto \left[\prod_{n=1}^{N} p(x_n|\theta)\right] p_0(\theta), \tag{1}$$

where $p(x_n|\theta)$ is a likelihood factor and $p_0(\theta)$ is the prior.

α-divergence

$$D_\alpha[p||q] = \frac{1}{\alpha(1\alpha)}\left(1 - \int p(\theta)^\alpha q(\theta)^{1\alpha} d\theta\right) \tag{2}$$

$$D_1[p||q] = \lim_{\alpha\to 1} D_\alpha[p||q] = KL[p||q]$$

$$D_0[p||q] = \lim_{\alpha\to 0} D_\alpha[p||q] = KL[q||p]$$

Definition

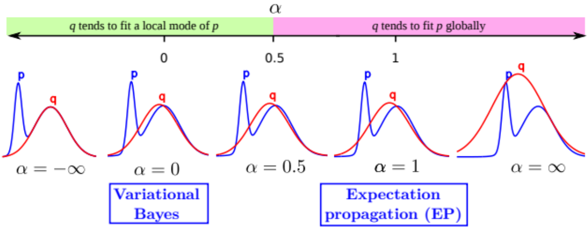

Figure: An illustration of approximating distributions by α-divergence minimization. Here p and q shown in the graphs are unnormalized probability densities.

Approximate local minimization of α-divergences

Power EP energy

$$E(\lambda_0, \lambda_n) = \log Z(\lambda_0) + \left(\frac{N}{\alpha} - 1\right) \log Z(\lambda_q) - \frac{1}{\alpha} \sum_{n=1}^{N} \log \int Zp(x_n|\theta)^\alpha \exp\{s(\theta)^T(\lambda_q - \alpha\lambda_n)\} d\theta \tag{3}$$

where

1. $p_0(\theta) = \exp\{s(\theta)^T \lambda_0 - \log Z(\lambda_0)\}$,
2. $f_n(\theta) = \exp\{s(\theta)^T \lambda_n\}$,
3. $q(\theta) \propto \exp\{s(\theta)^T (\sum_n \lambda_n + \lambda_0)\}$,
4. $\lambda_q = \sum_n \lambda_n + \lambda_0$.

Approximate local minimization of α-divergences

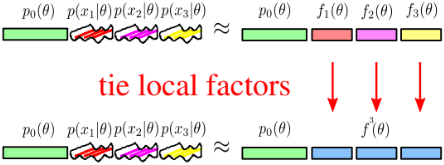

Figure: A cartoon for BB-α's factor tying constraint. Here we assume the dataset has N = 3 observations.

Approximate local minimization of α-divergences

Power EP energy

Let all the site parameters to be equal, i.e. $\lambda_n = \lambda \ \forall n$. Thus, $f_n(\theta) = f(\theta)$ $\forall n$. The cavity distributions with natural parameter:

$$\lambda = (N - \alpha)\lambda + \lambda_0, \lambda_q = N\lambda + \lambda_0$$

Thus, rewrite (3):

$$E(\lambda_0, \lambda) = \log Z(\lambda_0) - \log Z(\lambda_q) - \\ -\frac{1}{\alpha}\sum_{n=1}^{N} \log E_q \left(\left(\frac{p(x_n|\theta)}{f(\theta)} \right)^{\alpha} \right) \tag{4}$$

Empirical results

Dataset	Average Test Log-likelihood				Average Test Error			
	BB-α=1.0	**BB-α=0.5**	**BB-α=10^{-6}**	**BB-VB**	**BB-α=1.0**	**BB-α=0.5**	**BB-α=10^{-6}**	**BB-VB**
Ionosphere	-0.333±0.022	-0.333±0.022	-0.333±0.022	**-0.333±0.022**	0.124±0.008	0.124±0.008	**0.123±0.008**	0.123±0.008
Madelon	**-0.799±0.006**	-0.920±0.008	-0.953±0.009	-0.953±0.009	**0.445±0.005**	0.454±0.004	0.457±0.005	0.457±0.005
Pima	**-0.501±0.010**	-0.501±0.010	-0.501±0.010	-0.501±0.010	**0.234±0.006**	0.234±0.006	0.235±0.006	0.235±0.006
Colon Cancer	**-2.261±0.402**	-2.264±0.403	-2.268±0.404	-2.268±0.404	**0.303±0.028**	0.307±0.028	0.307±0.028	0.307±0.028
Avg. Rank	**1.895±0.097**	2.290±0.038	2.970±0.073	2.845±0.072	**2.322±0.048**	2.513±0.039	2.587±0.031	2.578±0.031

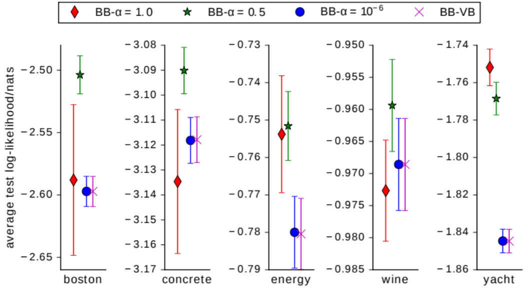

Figure: 1. Probit regression experiment results 2. Average test log-likelihood and the ranking comparisons.

Empirical results

CEP Dataset	BB-α=1.0	BB-α=0.5	BB-α=10^{-6}	BB-VB
Avg. Error	1.28±0.01	**1.08±0.01**	1.13±0.01	1.14±0.01
Avg. Rank	4.00±0.00	**1.35±0.15**	2.05±0.15	2.60±0.13
Avg. Log-likelihood	-0.93±0.01	**-0.74±0.01**	-1.39±0.03	-1.38±0.02
Avg. Rank	1.95±0.05	**1.05±0.05**	3.40±0.11	3.60±0.11

Figure: Average Test Error and Test Log-likelihood in CEP Dataset.

Literature

1. **Main article** Black-Box α-Divergence Minimization.