Black-box α-divergence minimization

José Miguel Hernández-Lobato[1,*], Yingzhen Li,[2,*], Mark Rowland[2], Daniel Hernández-Lobato[3], Thang Bui[2], Richard E. Turner[2]

Equal contributors*, Harvard University[1], University of Cambridge[2], Universidad Autónoma de Madrid[3]

1 - Minimizing α-divergences

The α-divergence between two distributions p and q is defined as [Amari, 1985]

$$D_\alpha(p||q) = \frac{\int_x \alpha p(x) + (1-\alpha)q(x) + p(x)^\alpha q(x)^{1-\alpha}}{\alpha(1-\alpha)}$$

Example of solution minimizing $D_\alpha(p||q)$:

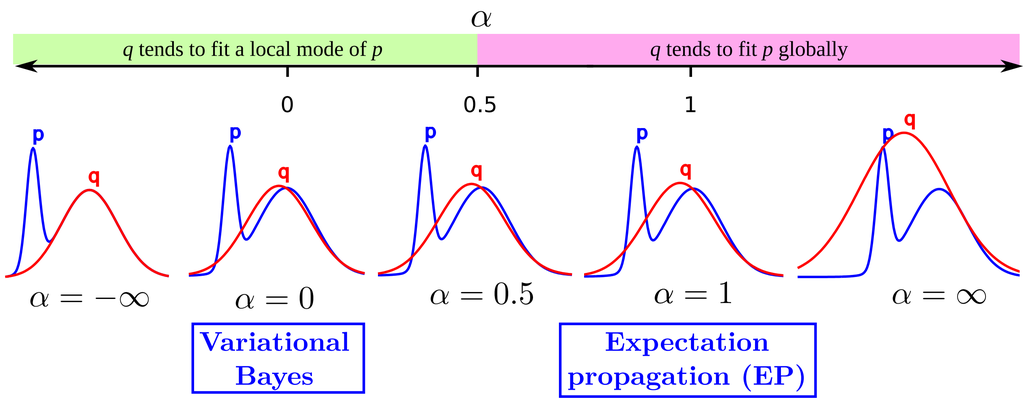
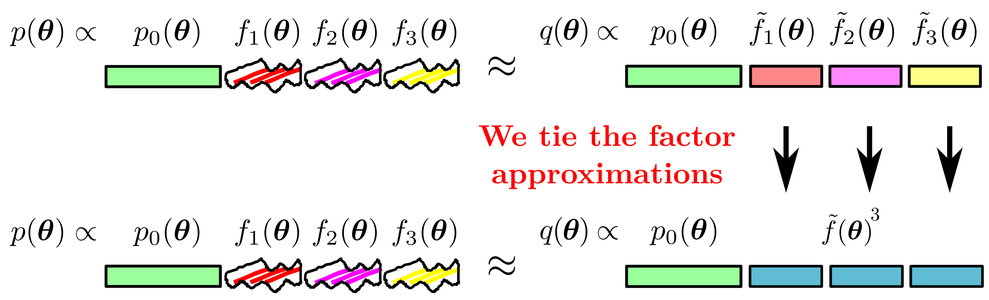


Figure source: [Minka, 2005].

- There are black-box and automatic methods for Variational Bayes ($\alpha = 0$) based on **stochastic optimization** and **automatic differentiation** approaches.
- Can we obtain similar methods for any α?

3 - Optimization with tied approximate factors

We simplify the optimization problem by following [Li et al. 2015] and tying the $\tilde{f}_n$:



The **reparameterization trick** [Kingma et al., 2014] is used to optimize a **stochastic** estimate of $\log Z_{\text{PEP}}$:

$$\log \hat{Z}_{\text{PEP}} = \log Z_q + \frac{N}{|\mathbf{S}|}\sum_{n\in\mathbf{S}}\frac{1}{\alpha}\log\frac{1}{K}\sum_{k=1}^{K}\left(\frac{f_n(\theta_k)}{\tilde{f}(\theta_k)}\right)^\alpha,$$

with minibatch $\mathbf{S}$ and K samples $\theta_1, \ldots, \theta_K \sim q$.

We obtain an **automatic, scalable** inference method that minimizes **local α-divergences**.

6 - Results with Bayesian neural networks

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

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

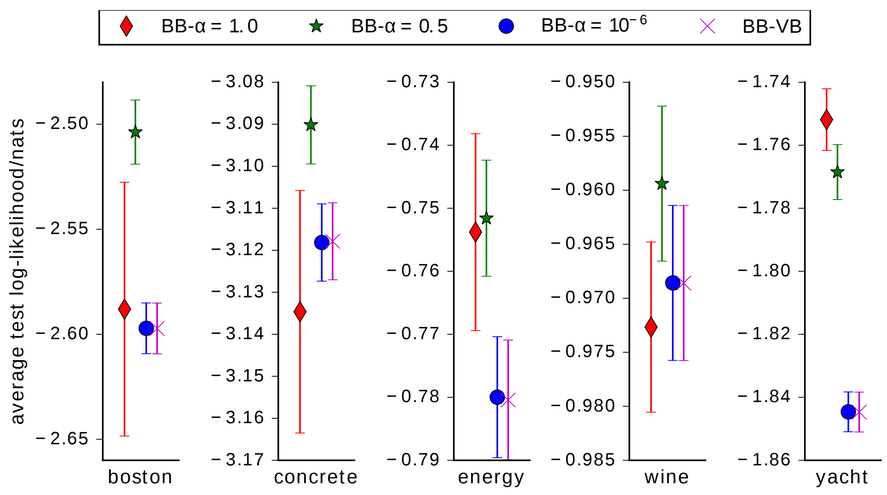


Figure: Average test log-likelihood.

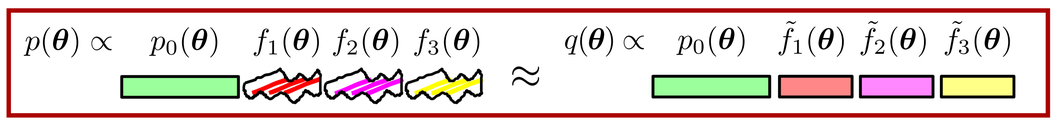
Table: Average Test Error and Test Log-likelihood in MNIST.

MNIST Dataset	BB-α=1.0	BB-α=0.5	BB-α=10^{-6}	BB-VB	BB-α=-1.0
Avg. Error	0.0151±0.0001	0.0144±0.0001	0.0136±0.0001	0.0136±0.0001	**0.0133±0.0001**
Avg. Rank	4.975±0.025	4.025±0.025	2.150±0.115	2.125±0.153	**1.725 ±0.210**
Avg. Log-likelihood	-0.0551±0.0004	-0.0509±0.0003	-0.0468±0.0002	-0.0468±0.0002	**-0.0447±0.0002**
Avg. Rank	5.0000±0.0000	4.0000±0.0000	2.5500±0.1141	2.4500±0.1141	**1.0000±0.0000**

2 - Power EP: local α-divergence minimization

Power EP approximates $p(\theta) \propto p_0(\theta)\prod_{i=1}^N f_n(\theta)$ with $q(\theta) \propto p_0(\theta)\prod_{n=1}^N \tilde{f}_n(\theta)$ by minimizing the local α-divergences

$$D_\alpha[p_n||q] \quad \text{for } n = 1, \ldots, N, \quad \text{where} \quad p_n(\theta) \propto \frac{f_n(\theta)q(\theta)}{\tilde{f}_n(\theta)}.$$



The Power-EP approximation to the evidence [Minka, 2005] is given by

$$\log Z_{\text{PEP}} = \log Z_q + \sum_{n=1}^{N}\frac{1}{\alpha}\log\mathbb{E}_q\left[\left(\frac{f_n(\theta)}{\tilde{f}_n(\theta)}\right)^\alpha\right],$$

where $Z_q = \int p_0(\theta)\prod_{n=1}^N \tilde{f}_n(\theta)\, d\theta$.

The power-EP solution for q can be obtained by solving the optimization problem

$$\max_q \min_{\tilde{f}_1,\ldots,\tilde{f}_N} \log Z_{\text{PEP}} \quad \text{subject to} \quad q(\theta) = p_0(\theta)\prod_{n=1}^{N}\tilde{f}_n(\theta),$$

- Can be solved with a double-loop algorithm [Heskes and Zoeter, 2002].
- At convergence, the local α-divergences are minimized.
- **Convergence is too slow to be useful in practice!**

4 - Stationary conditions of the objective function

- **With non-tied factors.**
  We get the well-known matching of expected sufficient statistics by power-EP:

$$\mathbf{E}_q[s(\theta)] = \mathbf{E}_{\tilde{p}_n}[s(\theta)], \quad \text{for} \quad n = 1, \ldots, N,$$

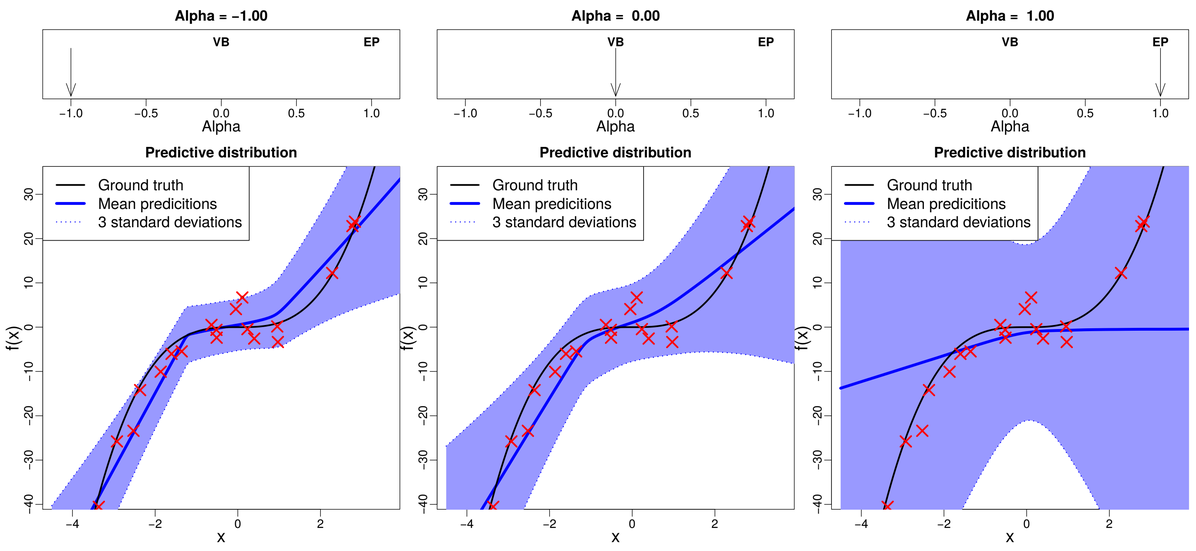
  where $\tilde{p}_n \propto f_n(\theta)^\alpha \tilde{f}_k(\theta)^{-\alpha} q(\theta)$ is the n-th **tilted distribution** and $s(\theta)$ is the vector of **sufficient statistics**.

- **With tied factors.**
  The expectation of $s(\theta)$ with respect to q are equal to the average of the expectations of $s(\theta)$ over the tilted distributions:

$$\mathbf{E}_q[s(\theta)] = \frac{1}{N}\sum_{n=1}^{N}\mathbf{E}_{\tilde{p}_n}[s(\theta)],$$

- With lots of data, the tied-factor solution is expected to converge to the non-tied one.

5 - Toy example with Bayesian neural networks



Negatives values of α produce better results.

7 - Bias-variance analysis

Table: Average Bias in the Gradient.

	BB-α=1.0	BB-α=0.5	BB-α=10^{-6}
$K = 1$	0.2774	0.1214	0.0460
$K = 5$	0.0332	0.0189	0.0162
$K = 10$	0.0077	0.0013	0.0001

Table: Average Standard Deviation in the Gradient.

	BB-α=1.0	BB-α=0.5	BB-α=10^{-6}
$K = 1$	14.1209	14.0159	13.9109
$K = 5$	12.7953	12.8418	12.8984
$K = 10$	12.3203	12.4101	12.5058

8 - Summary

Black-box α-divergence minimization:

- generalizes variational Bayes and an expectation-propagation-like algorithm.
- better than any of them in regression problems with neural networks.
- straightforward to apply in very complex probabilistic models.