Black-box α-divergence minimization

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

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)]

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

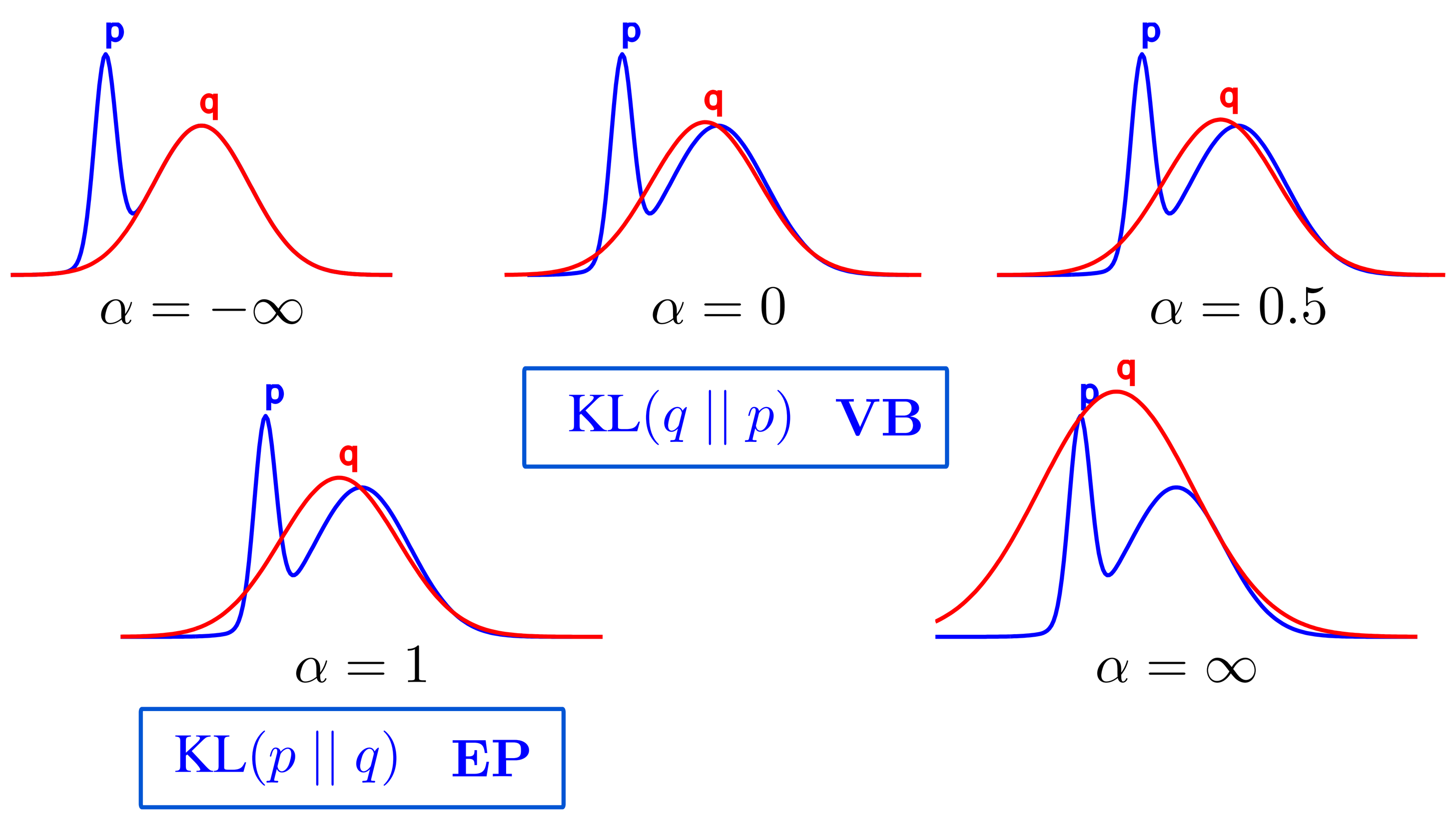
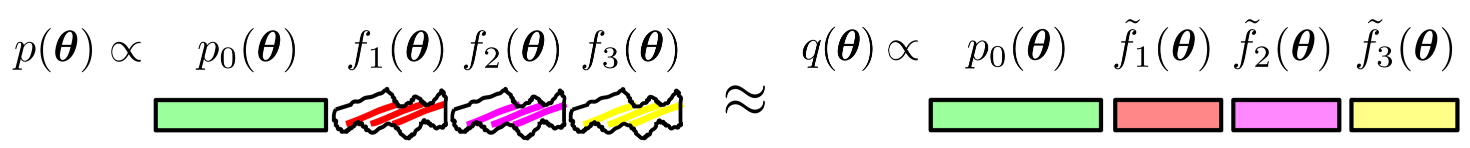


Figure source: [Minka (2005)].

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

2 - Local α-divergence minimization (Power EP)

We approximate $p(\boldsymbol{\theta}) \propto p_0(\boldsymbol{\theta}) \prod_{i=1}^N f_n(\boldsymbol{\theta})$ with $q(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta}) \prod_{n=1}^N \tilde{f}_n(\boldsymbol{\theta})$.



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

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

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

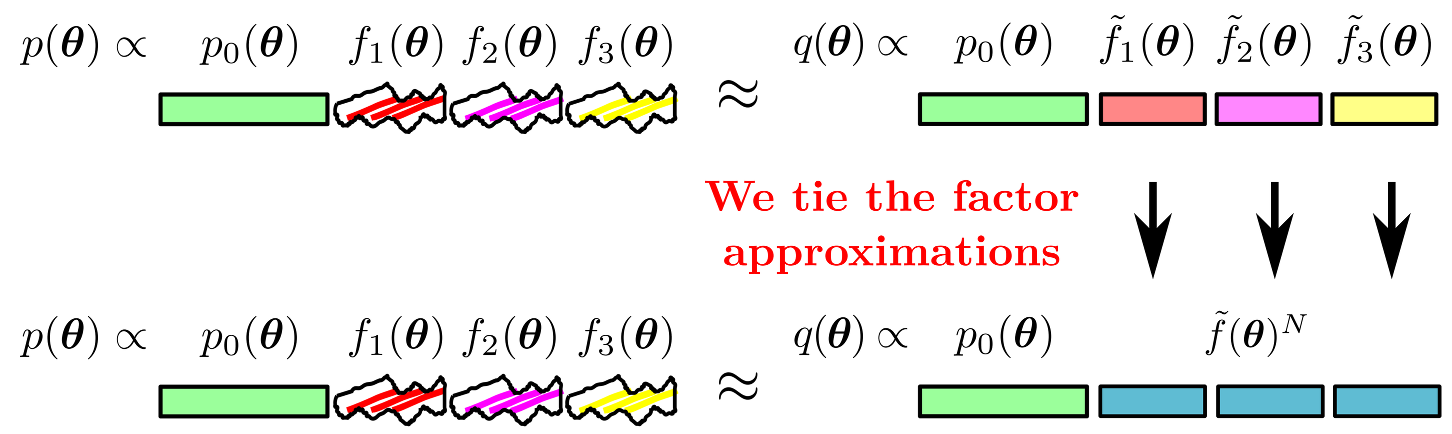
The power-EP solution for q can be obtained by solving the optimization problem

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

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

Optimization with tied approximate factors

By following [Li et al. (2015)]:



No double-loop needed. Memory saving scales as $\mathcal{O}(N)$.

Stochastic estimate of the evidence for automatic, scalable inference:

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

for minibatch $\mathbf{S}$ and K samples $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K \sim q$.

Stationary conditions for tied and non-tied factors when $\alpha = 1$ (EP)

- **With non-tied factors:**

  We get the well-known matching of expected sufficient statistics:

$$\boxed{\mathbf{E}_q[s(\boldsymbol{\theta})] = \mathbf{E}_{f_n q^{\setminus n}}[s(\boldsymbol{\theta})], \quad \text{for} \quad n = 1, \ldots, N,}$$

  where $f_n q^{\setminus n} \propto p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta}) \prod_{k \neq n}^N \tilde{f}_k(\boldsymbol{\theta})$ is called the n-th tilted distribution and $s(\boldsymbol{\theta})$ are the sufficient statistics.

- **With tied factors:**

  The expectation of $s(\boldsymbol{\theta})$ with respect to q are equal to the average of the expectations of $s(\boldsymbol{\theta})$ over the tilted distributions:

$$\boxed{\mathbf{E}_q[s(\boldsymbol{\theta})] = \frac{1}{N} \sum_{n=1}^N \mathbf{E}_{f_n q^{\setminus n}}[s(\boldsymbol{\theta})], \quad \text{for} \quad n = 1, \ldots, N,}$$

With a lot of data, the solution with tied factors is expected to converge to the solution with non-tied factors.

Stationary conditions for the prior hyper-parameters

The Z_{PEP} is maximized with respect to the prior hyper-parameters when

$$\boxed{\mathbf{E}_q[s(\boldsymbol{\theta})] = \mathbf{E}_{p_0}[s(\boldsymbol{\theta})], \quad \text{for} \quad n = 1, \ldots, N,}$$

Experiments and results

Stochastic optimization with minibatch size 100 and $K = 100$. q is factorized Gaussian.

We use **autograd** for automatic gradient computation.

Bayesian probit regression:

Table : Average Test Log-likelihood and Standard Errors, Probit Regression.

Dataset	**WB-α=1.0**	**BB-α=1.0**	**BB-α=10^{-6}**	**BB-VB**
Ionosphere	-0.3211±0.0134	-0.3206±0.0134	**-0.3204±0.0134**	-0.3204±0.0134
Madelon	-0.6771±0.0021	-0.6764±0.0019	-0.6763±0.0012	**-0.6763±0.0012**
Pima	**-0.4993±0.0098**	-0.4997±0.0099	-0.5001±0.0099	-0.5001±0.0099
Avg. Rank	2.5510±0.1110	**2.3810 ±0.0854**	2.5170±0.0967	2.5510±0.0717

Bayesian neural networks with 100 hidden units and one single hidden layer:

We tune α, learning rates and prior variance with Bayesian optimization.

Table : Average Test Log-likelihood and Standard Errors, Neural Networks.

Dataset	**BB-α=BO**	**BB-α=1**	**BB-α=10^{-6}**	**BB-VB**	**Avg. α**
Boston	**-2.549±0.019**	-2.621±0.041	-2.614±0.021	-2.578±0.017	0.45 ±0.04
Concrete	**-3.104±0.015**	-3.126±0.018	-3.119±0.010	-3.118±0.010	0.72 ±0.03
Energy	-0.979±0.028	-1.020±0.045	**-0.945±0.012**	-0.994±0.014	0.72 ±0.03
Wine	-0.949±0.009	**-0.945±0.008**	-0.967±0.008	-0.964±0.007	0.86 ±0.04
Yacht	**-1.102±0.039**	-2.091±0.067	-1.594±0.016	-1.646±0.017	0.48 ±0.01
Avg. Rank	**1.835 ±0.065**	2.504±0.080	2.766±0.061	2.895±0.057	

Conclusions

- By using tied factors, we **avoid double loop algorithms** in Power-EP.
- The approximation of the evidence can then be optimized using **stochastic methods**. These can be combined with **automatic differentiation tools**.
- By doing so, we enable **black-box automatic α-divergence minimization**.
- **Tuning α** seems to produce gains in complex posterior distributions such as those in Bayesian neural networks.

Future work

- Bayesian neural networks for classification. Experiments on MNIST.
- How to tune each α_n optimally to its corresponding exact factor f_n?
- Analysis of the amount of bias and variance in the stochastic gradients.

References

Amari, S. *Differential-geometrical methods in statistics*, volume 28. Springer Science & Business Media, 1985.

Heskes et al. Expectation propagation for approximate inference in dynamic bayesian networks. In *UAI 18*, 2002.

Kucukelbir, Alp, Ranganath, Rajesh, Gelman, Andrew, and Blei, David. Automatic variational inference in stan. In *NIPS 28*, pp. 568–576. 2015.

Li, Yingzhen, Hernández-Lobato, José Miguel, and Turner, Richard E. Stochastic expectation propagation. In *NIPS 28*. 2015.

Minka, Tom. Divergence measures and message passing. Technical report, Microsoft Research, 2005.