Black-box α-divergence Minimization

José Miguel Hernádez–Lobato[1]

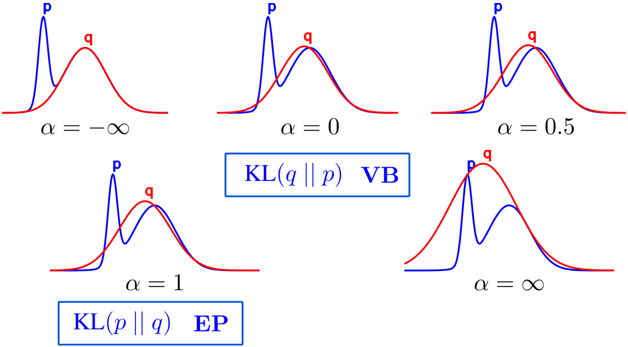
Joint work with Yingzhen Li, Daniel Hernández-Lobato,
Thang Bui and Richard Turner.

[1]Harvard Intelligent Probabilistic Systems Group
School of Engineering and Applied Sciences,
Harvard University
http://jmhl.org
jmh@seas.harvard.edu

α-Divergence

$$D_\alpha(p||q) = \frac{\int_x \alpha p(x) + (1-\alpha) q(x) - p(x)^\alpha q(x)^{1-\alpha}}{\alpha(1-\alpha)} \qquad \text{[Amari, 1985].}$$

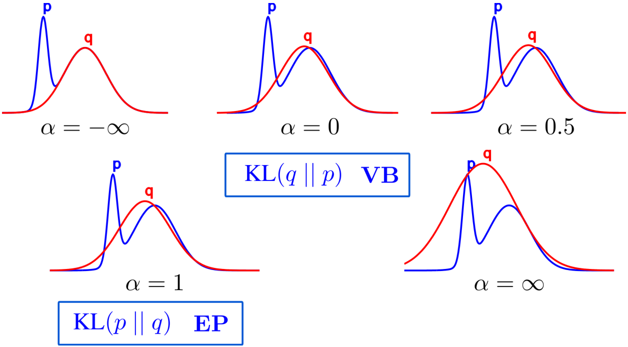
(Minka [2005]).



α-Divergence

$$D_\alpha(p||q) = \frac{\int_x \alpha p(x) + (1-\alpha) q(x) - p(x)^\alpha q(x)^{1-\alpha}}{\alpha(1-\alpha)} \qquad \text{[Amari, 1985].}$$

(Minka [2005]).

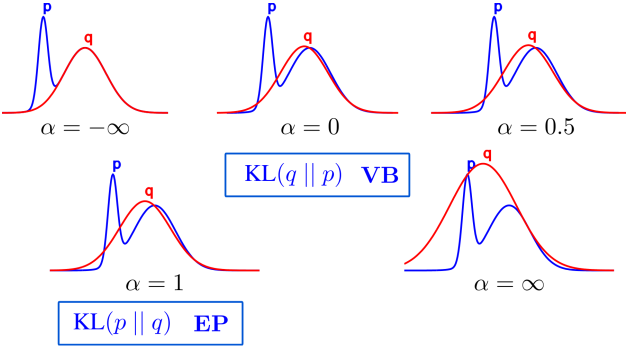


There are automatic tools for $\alpha = 0$ [Kucukelbir et al., 2015].

α-Divergence

$$D_\alpha(p||q) = \frac{\int_x \alpha p(x) + (1-\alpha) q(x) - p(x)^\alpha q(x)^{1-\alpha}}{\alpha(1-\alpha)} \qquad \text{[Amari, 1985].}$$

(Minka [2005]).



There are automatic tools for $\alpha = 0$ [Kucukelbir et al., 2015].

Can we have automatic tools for other values of α?

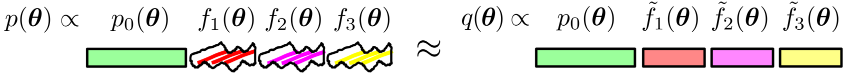
Local α-divergence minimization (Power EP)

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

$$p(\boldsymbol{\theta}) \propto p_0(\boldsymbol{\theta})\, f_1(\boldsymbol{\theta})\, f_2(\boldsymbol{\theta})\, f_3(\boldsymbol{\theta}) \approx q(\boldsymbol{\theta}) \propto p_0(\boldsymbol{\theta})\, \tilde{f}_1(\boldsymbol{\theta})\, \tilde{f}_2(\boldsymbol{\theta})\, \tilde{f}_3(\boldsymbol{\theta})$$

Local α-divergence minimization (Power EP)

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

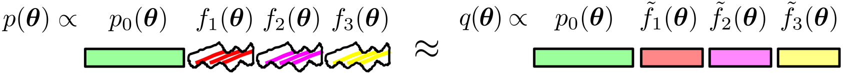


The Power-EP approximation to the evidence [Minka, 2005] is given by

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

Local α-divergence minimization (Power EP)

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

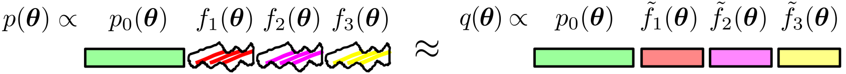
$$\log Z_{\mathrm{PEP}} = \log Z_q + \sum_{n=1}^{N} \frac{1}{\alpha_n} \log \mathbb{E}_q \left[\left(\frac{f_n(\boldsymbol{\theta})}{\tilde{f}_n(\boldsymbol{\theta})} \right)^{\alpha_n} \right] ,$$

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

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

Local α-divergence minimization (Power EP)

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

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

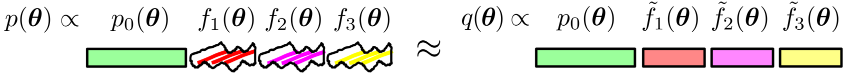
The power-EP solution for q can be obtained by solving

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

Solved with a double-loop algorithm [Heskes et al., 2002].

Local α-divergence minimization (Power EP)

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

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

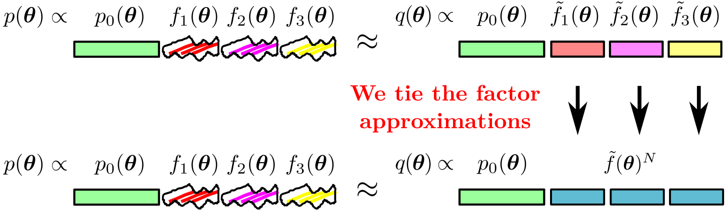
The power-EP solution for q can be obtained by solving

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

Solved with a double-loop algorithm [Heskes et al., 2002]. Too slow!

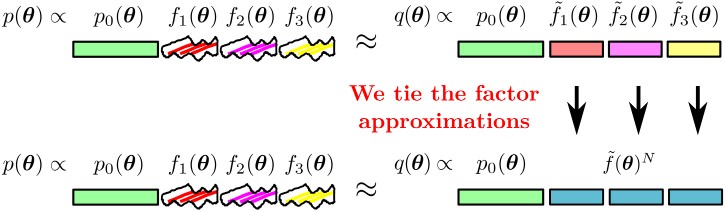
Optimization with tied approximate factors

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



Optimization with tied approximate factors

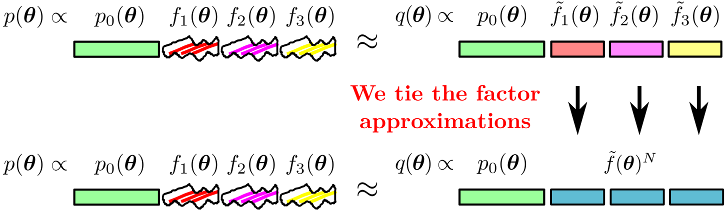
By following Li et al. [2015] (Stochastic Expectation Propagation):



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

Optimization with tied approximate factors

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



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

Noisy estimate of the evidence for automatic, scalable inference:

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

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

Experimental Results

We use stochastic optimization with minibatch size 100 and $K = 100$.

Experimental Results

We use stochastic optimization with minibatch size 100 and $K = 100$.

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

Experimental Results

We use stochastic optimization with minibatch size 100 and $K = 100$.

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

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	

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

Thank you for your attention!

I am in the job market!
Have a look at my website http://jmhl.org
jmh@seas.harvard.edu

References I

S. Amari. *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 *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, 2002.

A. Kucukelbir, R. Ranganath, A. Gelman, and D. Blei. Automatic variational inference in stan. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 568–576. Curran Associates, Inc., 2015.

Y. Li, J. M. Hernández-Lobato, and R. E. Turner. Stochastic expectation propagation. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., 2015.

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