STOCHASTIC EXPECTATION PROPAGATION

Yingzhen Li[1], José Miguel Hernández-Lobato[2] and Richard E. Turner[1]
[1]University of Cambridge, [2]Harvard University
{yl494, ret26}@cam.ac.uk, jmh@seas.harvard.edu

Neural Information Processing Systems Foundation

FROM EP TO STOCHASTIC EP

Goal: approximate the true posterior $q(\theta) \approx p(\theta|\mathcal{D})$

$p(\theta|\mathcal{D}) \propto p_0(\theta)\, p(x_1|\theta)p(x_2|\theta)p(x_3|\theta) \approx q(\theta) \propto p_0(\theta)\, f_1(\theta)\, f_2(\theta)\, f_3(\theta)$

idealised

capture affect of $p(x_3|\theta)$ on posterior

$p(\theta|\mathcal{D}) \propto p_0(\theta)\, p(x_1|\theta)p(x_2|\theta)\, p(x_3|\theta) \approx q^{new}(\theta) \propto p_0(\theta)\, p(x_1|\theta)p(x_2|\theta)\, f_3^{new}(\theta)$

EP — **approximate** — **intractable**

$\tilde{p}(\theta) \propto p_0(\theta)\, f_1(\theta)\, f_2(\theta)\, p(x_3|\theta) \approx q_{EP}^{new}(\theta) \propto p_0(\theta)\, f_1(\theta)\, f_2(\theta)\, f_3^{new}(\theta)$

SEP — **tie local factors** — **memory** $\mathcal{O}(ND^2)$ (Gaussian approx.)

$p(\theta|\mathcal{D}) \propto p_0(\theta)\, f(\theta)^{N-1}\, p(x_3|\theta) \approx q_{SEP}^{new}(\theta) \propto p_0(\theta)\, f^{new}(\theta)^N$

memory $\mathcal{O}(D^2)$

AN ALTERNATIVE VIEW OF SEP

Equivalent factorisation: $\overline{p(x|\theta)} = \left(\prod_n p(x_n|\theta)\right)^{1/N}$

$p(\theta|\mathcal{D}) \propto p_0(\theta)\, p(x_1|\theta)p(x_2|\theta)p(x_3|\theta)$

EP on the equivilent factorisation will converge to $f_n(\theta) = f(\theta)$ that captures the **average affect** of likelihood terms on posterior!

$\approx p_0(\theta)\, \overline{p(x|\theta)}^N \approx q(\theta) \propto p_0(\theta)\, f(\theta)^N$

EP (on the equivalent factorisation)

$p(\theta|\mathcal{D}) \propto p_0(\theta)\, f(\theta)^{N-1}\, \overline{p(x|\theta)} \approx q^{new}(\theta) \propto p_0(\theta)\, f^{new}(\theta)^N$

intractable

SEP

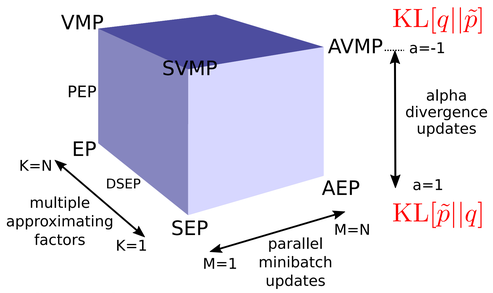
$p(\theta|\mathcal{D}) \propto p_0(\theta)\, f(\theta)^{N-1}\, \overline{p(x_3|\theta)} \approx q_{SEP}^{new}(\theta) \propto p_0(\theta)\, f^{new}(\theta)^N$

stochastic approximation

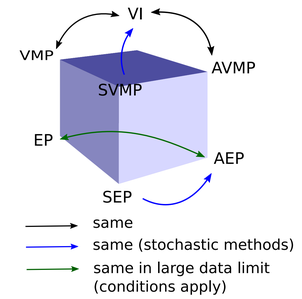
Work in progress:
guarantees of lower bound
analysis of approximation error

RELATED ALGORITHMS

A) Relationships between algorithms



B) Fixed points properties



DSEP (K=2)

$p(\theta|\mathcal{D}) \propto p_0(\theta)\, p(x_1|\theta)p(x_2|\theta)p(x_3|\theta)p(x_4|\theta) \approx q_{DSEP}(\theta) \propto p_0(\theta)\, f_1(\theta)^{N_1}\, f_2(\theta)^{N_2}$

SEP FOR LATENT VARIABLE MODELS

Model: $p(\boldsymbol{x}_n|\boldsymbol{h}_n, \boldsymbol{\theta})$, $\boldsymbol{h}_n \sim p_0(\boldsymbol{h}_n)$ i.i.d., posterior

$$p(\boldsymbol{\theta}, \{\boldsymbol{h}_n\}|\mathcal{D}) \approx q(\boldsymbol{\theta}, \{\boldsymbol{h}_n\}) \propto p_0(\boldsymbol{\theta}) f(\boldsymbol{\theta})^N \prod_{n=1}^N g_n(\boldsymbol{h}_n)$$

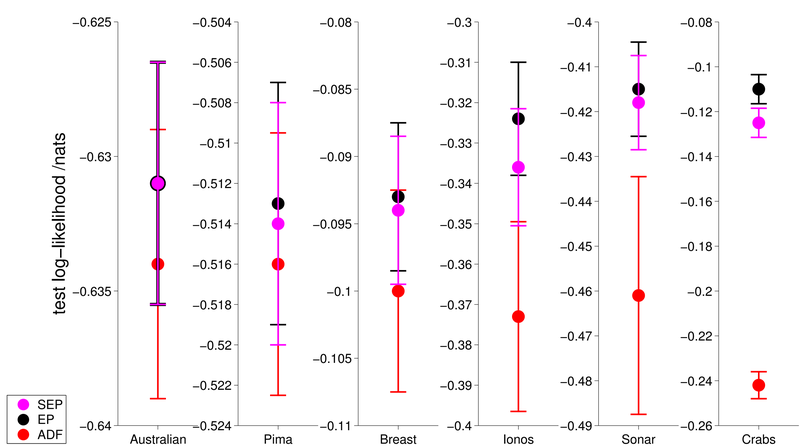
Update $f(\boldsymbol{\theta})g_n(\boldsymbol{h}_n)$ such that

$$p_0(\boldsymbol{\theta}) f(\boldsymbol{\theta})^{N-1} \prod_{m\neq n} g_m(\boldsymbol{h}_m) p(\boldsymbol{x}_n|\boldsymbol{h}_n, \boldsymbol{\theta}) p_0(\boldsymbol{h}_n)$$

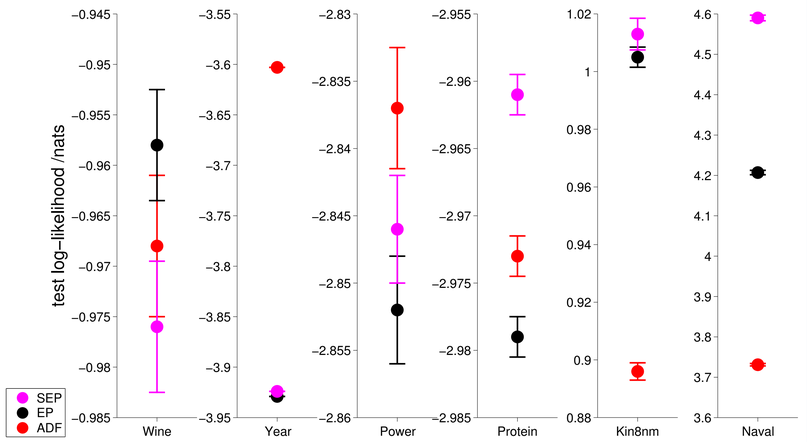
$$\approx \quad p_0(\boldsymbol{\theta}) f(\boldsymbol{\theta})^{N-1} \prod_{m\neq n} g_m(\boldsymbol{h}_m) f^{new}(\boldsymbol{\theta}) g_n^{new}(\boldsymbol{h}_n)$$

No need to maintain g_n if updates can be computed fast!

BAYESIAN PROBIT CLASSIFICATION



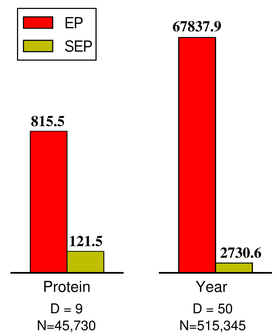
BAYESIAN NEURAL NETWORK



- **Prior:** $p_0(\lambda|\gamma) = Gam(\alpha_0^\lambda, \beta_0^\lambda)$, $p_0(\boldsymbol{w}|\lambda) = \prod_{ijl} \mathcal{N}(w_{ij,l}; 0, \lambda^{-1})$
- **Likelihood:** $p(\gamma) = Gam(\alpha_0^\gamma, \beta_0^\gamma)$, $p(y|\boldsymbol{x}, \boldsymbol{w}, \gamma) = \mathcal{N}(y; f(\boldsymbol{x}; \boldsymbol{w}), \gamma^{-1})$
- $f(\boldsymbol{x}; \boldsymbol{w})$ is the output of a neural network with ReLU units.
- **Inference:** PBP[a] works better when under-estimating uncertainty.

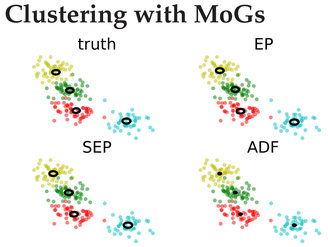
[a] Hernández-Lobato J M, Adams R P. Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks. In *ICML* 2015.

Memory savings (in MB)

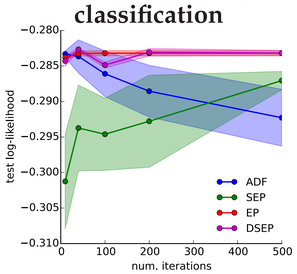


MORE EXAMPLES

1. Clustering with MoGs



2. Odd-vs-even digit classification



CONCLUSIONS

- We scaled EP to large datasets with nearly identical performances
- We extended stochastic EP and related it to variational Bayes
- SEP is well suited to "big model, big data" settings