Expectation Propagation: why use it, when to use it

Simon Barthelmé, CNRS, Gipsa-lab (Grenoble). Joint work with Guillaume Dehaene (Univ. Genève), Nicolas Chopin & Vincent Cottet (ENSAE)

November 2, 2015

Outline

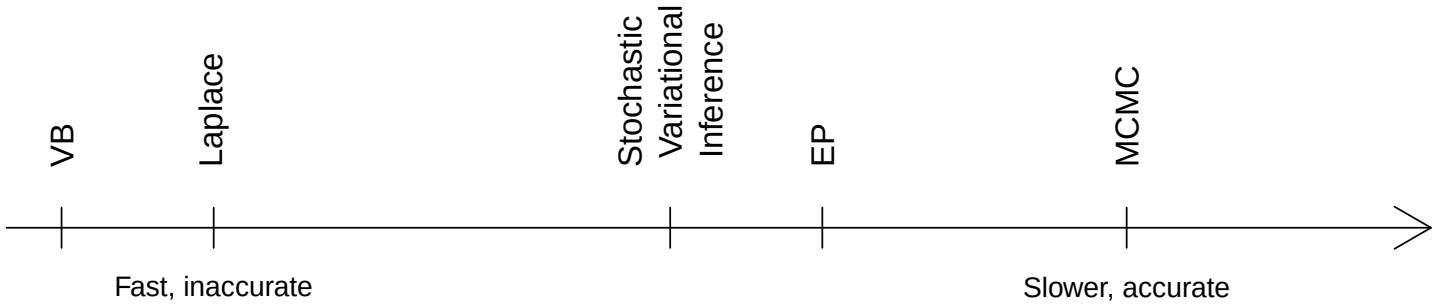
- Part I: The algorithm, some applications
- Part II: Some theoretical guarantees
- Conclusion, new directions for EP

Variational inference

- MCMC works fine in many cases, but it is slow
- Variational inference algorithms try to speed things up by giving up on exactness: replace true posterior with a tractable approximation
- Very popular in Machine Learning
- In statistics: became more popular with INLA (Rue et. al, 2009)

Tradeoffs in variational inference

- Variational Inference methods are on an axis that goes from "fast and inaccurate" to "slower but more accurate"



Expectation Propagation

- EP was introduced by Tom Minka (2001)
- EP is known to be very accurate in many empirical cases
  - Gaussian processes
  - Logistic regression
- EP is very easy to parallelise (Barthelmé, Chopin, Cottet, 2015. Cseke & Heskes, 2011)
- EP is fast when implemented properly

Objective

We have a posterior distribution $\pi(\theta)$, we wish to approximate it with a Gaussian q such that

$$\underset{q \in \mathcal{Q}}{\operatorname{argmin}} KL\left(\pi || q\right)$$

$$\mathrm{KL}(\pi||q) = \int \pi(\theta) \log \frac{\pi(\theta)}{q(\theta)} \, d\theta$$

Properties of the KL objective

The solution of:

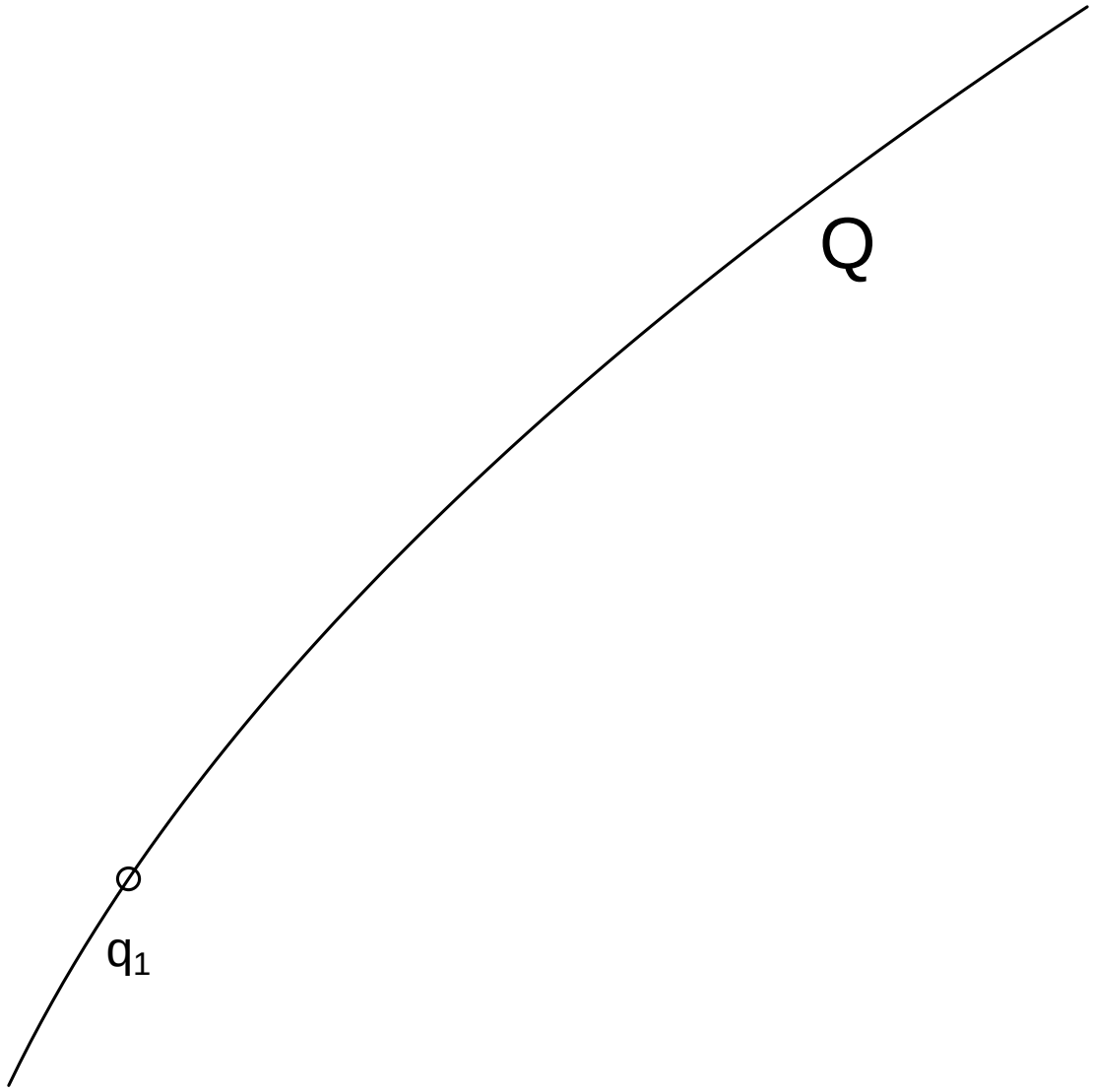
$$\underset{q \in \mathcal{Q}}{\operatorname{argmin}} KL\left(\pi || q\right)$$

has a "closed form" of sorts. It is the Gaussian $q^\star$ with mean $E(\pi)$ and variance $Var(\pi)$.

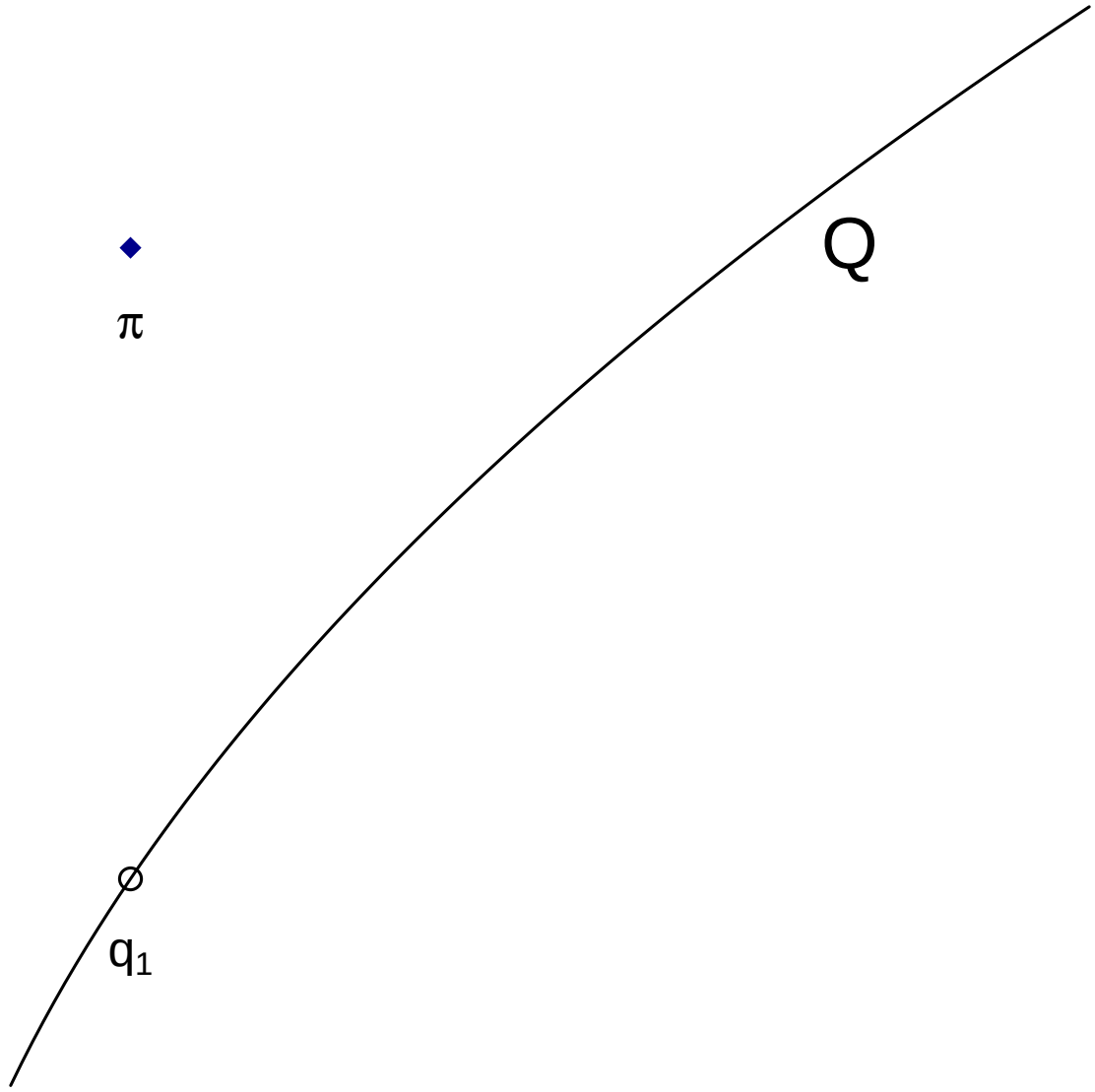
Obviously we have no hope of optimising the objective *exactly*.

In EP we will replace it with simpler, local problems we can actually solve.

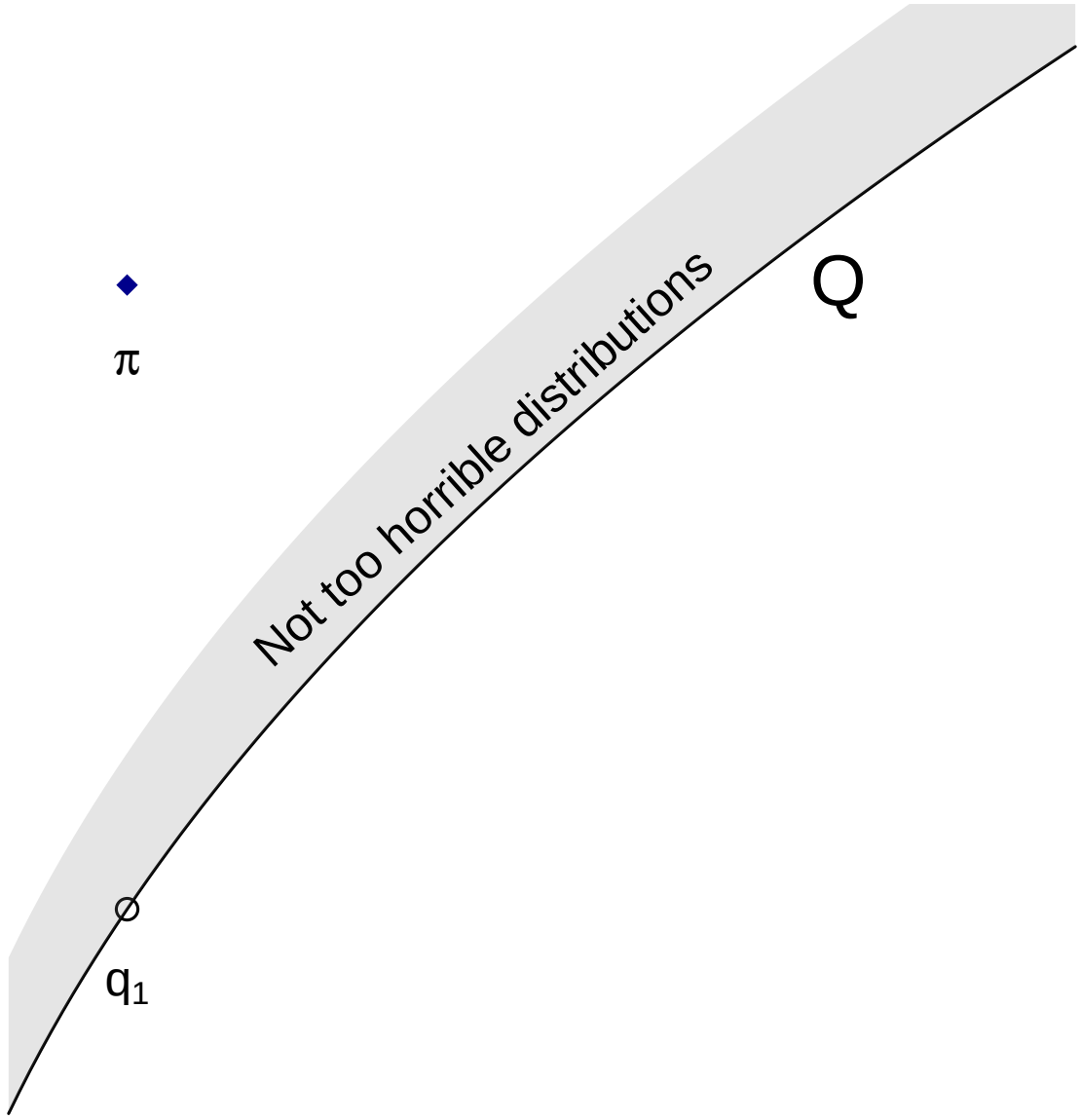
EP: the big picture



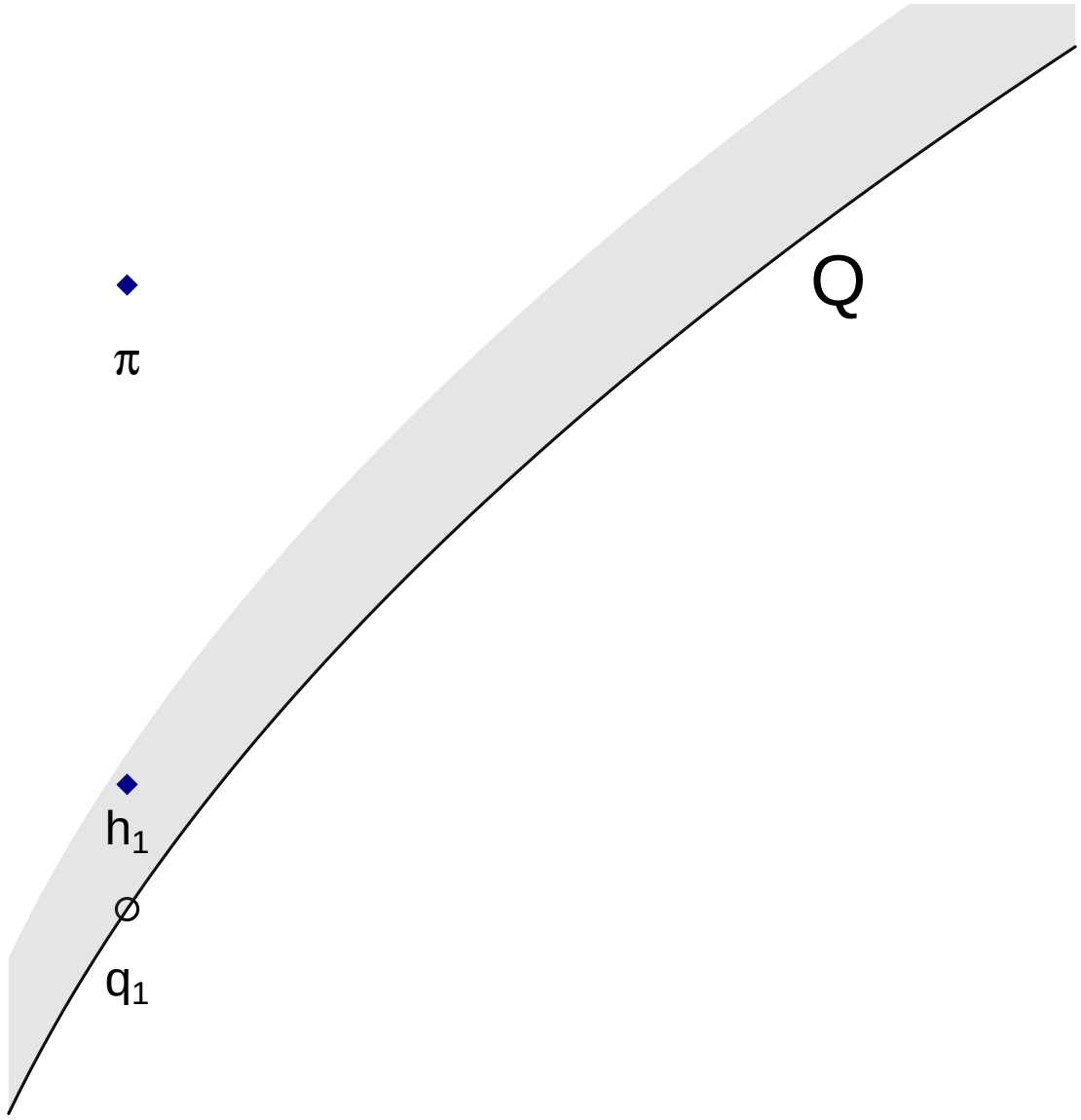
EP: the big picture



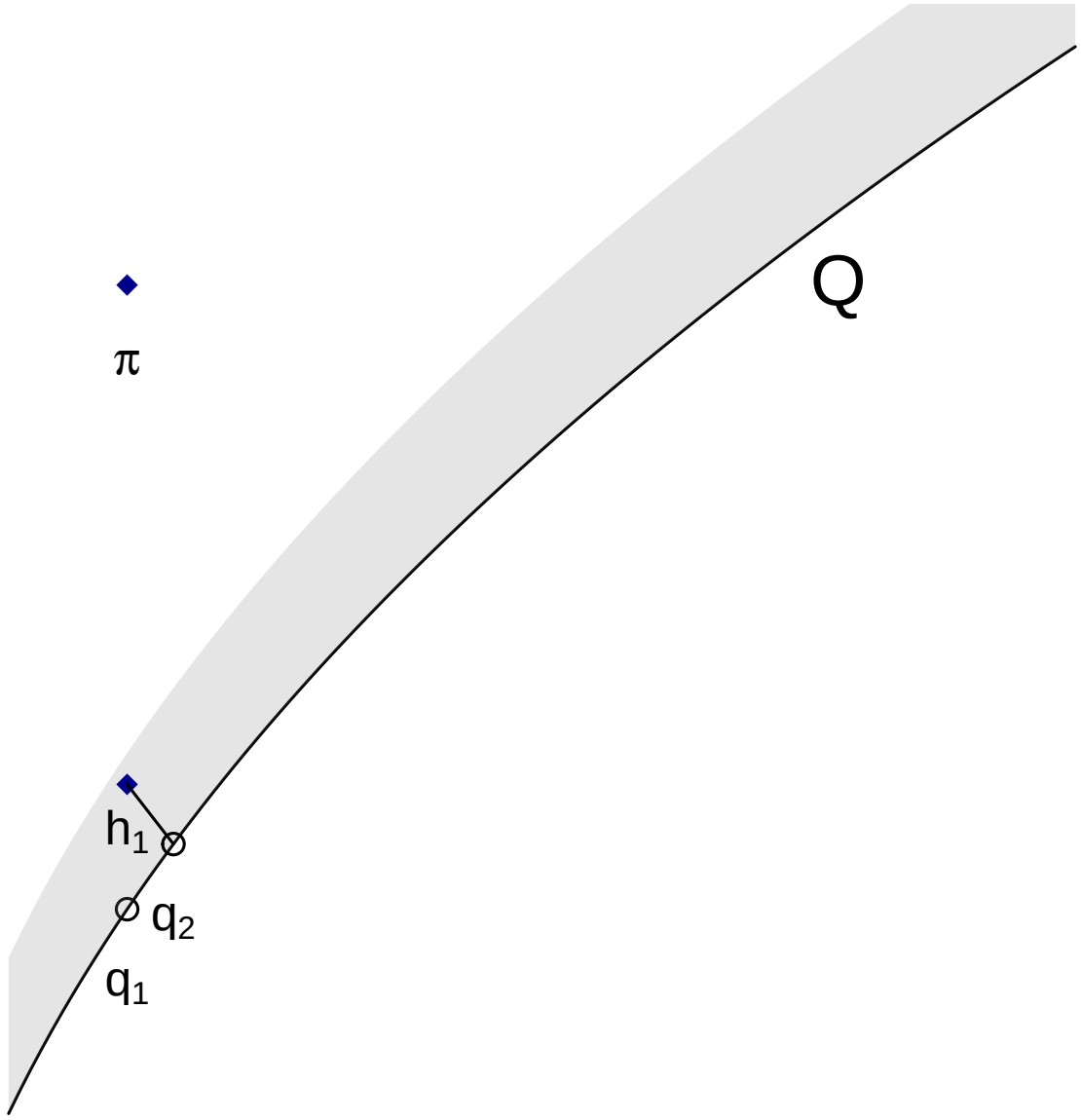
EP: the big picture



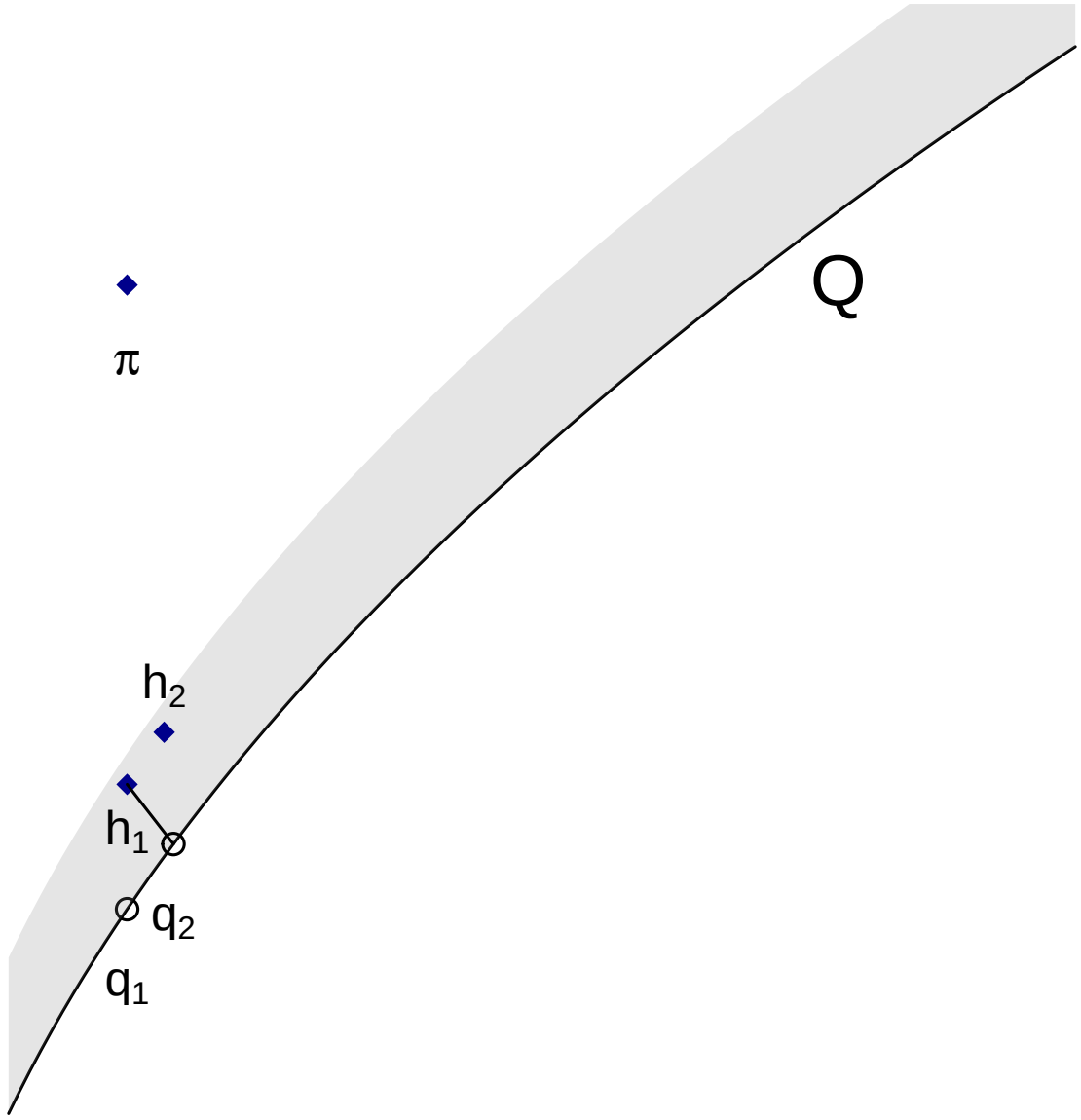
EP: the big picture



EP: the big picture



EP: the big picture



How EP works (I): write the posterior as a product

Consider a posterior distribution with independent datapoints:

$$\pi(\theta) = p(\theta|\mathbf{y}) \propto p\,(\theta) \prod_{i=1}^{n} l_i\,(\theta)$$

It can be written as a product of factors:

$$\pi(\theta) \propto \prod_{i=0}^{n} l_i(\theta)$$

How EP works (II): take a product of Gaussians

We will approximate the posterior:

$$\pi(\theta) \propto \prod_{i=0}^{n} l_i(\theta)$$

with a product of a Gaussian factors:

$$q(\theta) \propto \prod_{i=0}^{n} q_i(\theta)$$

How EP works (II): take a product of Gaussians

Each Gaussian factor equals:

$$q_i(\theta) = \exp(-\frac{1}{2}\theta^t \mathbf{A}_i \theta + \mathbf{r}_i \theta)$$

And so the approximation is a Gaussian too:

$$q(\theta) = \prod_{i=0}^{n} q_i(\theta) = \exp(-\frac{1}{2}\theta^t \sum\{\mathbf{A}_i\}\theta + \sum\{\mathbf{r}_i\}\theta)$$

How EP works (III): hybridise the true and approximate distribution

You can form a *hybrid* between the true and the approximate distribution by replacing one of the approximate factors with one of the true factors:

1. Take out the approximate factor

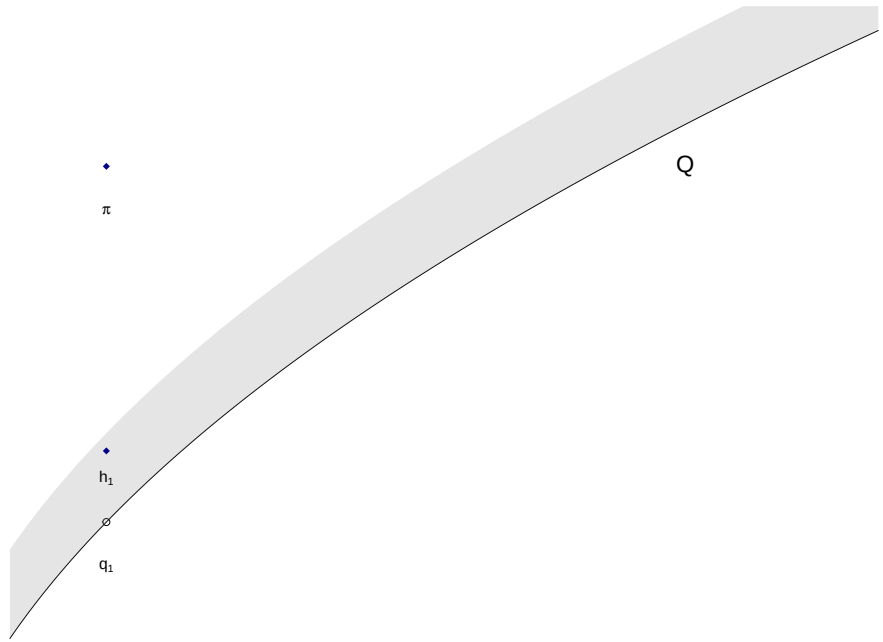
$$q_{-i}(\theta) = \prod_{i \neq j}^{n} q_i(\theta)$$

2. Insert the true factor

$$h_i(\theta) = l_i(\theta) q_{-i}(\theta)$$

How EP works (III): project the hybrid

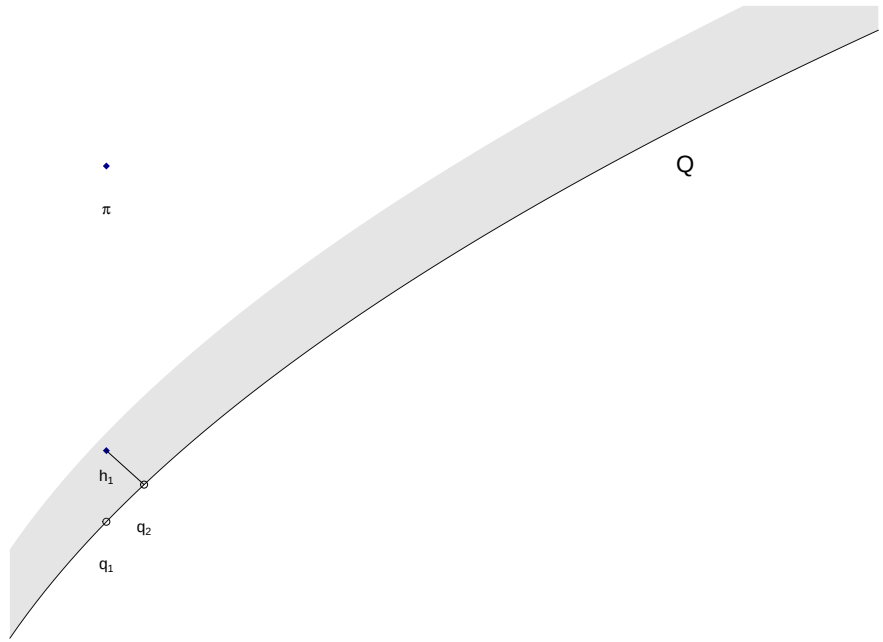
Hopefully the hybrid is in some sense closer to the true distribution



How EP works (III): project the hybrid

Now we need to project the hybrid, i.e. solve:

$$\underset{q \in \mathcal{Q}}{\operatorname{argmin}} KL\left(h_i || q\right)$$



How EP works (III): project the hybrid

That's just equivalent to computing the moments:

$$z = \int h_i(\theta) d\theta$$

$$E(\theta) = z^{-1} \int \theta h_i(\theta) d\theta$$

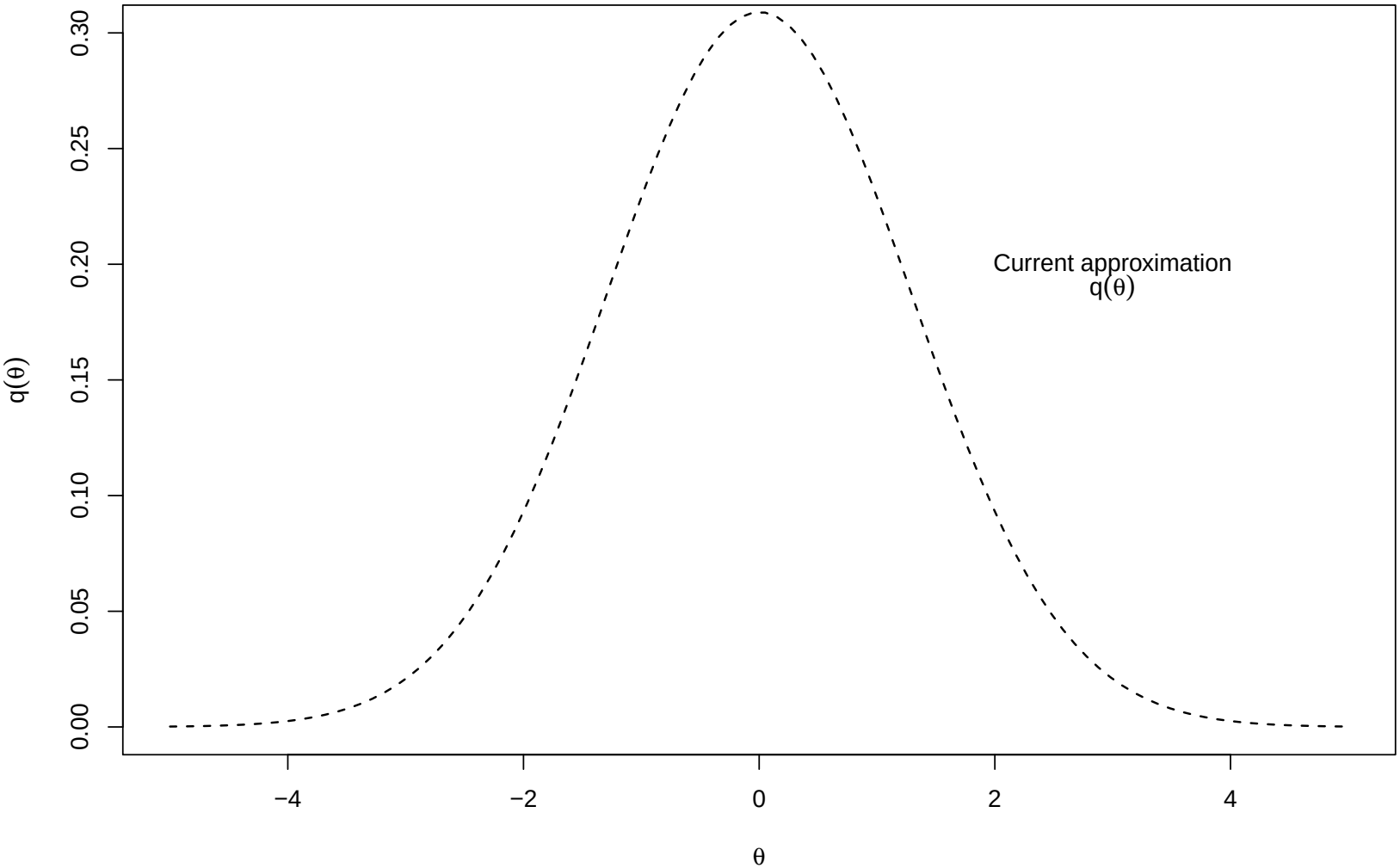
$$\Sigma = z^{-1} \int (\theta - E(\theta))(\theta - E(\theta))^t h_i(\theta) d\theta$$

Our new global approximation q' is a Gaussian with mean and covariance as above.

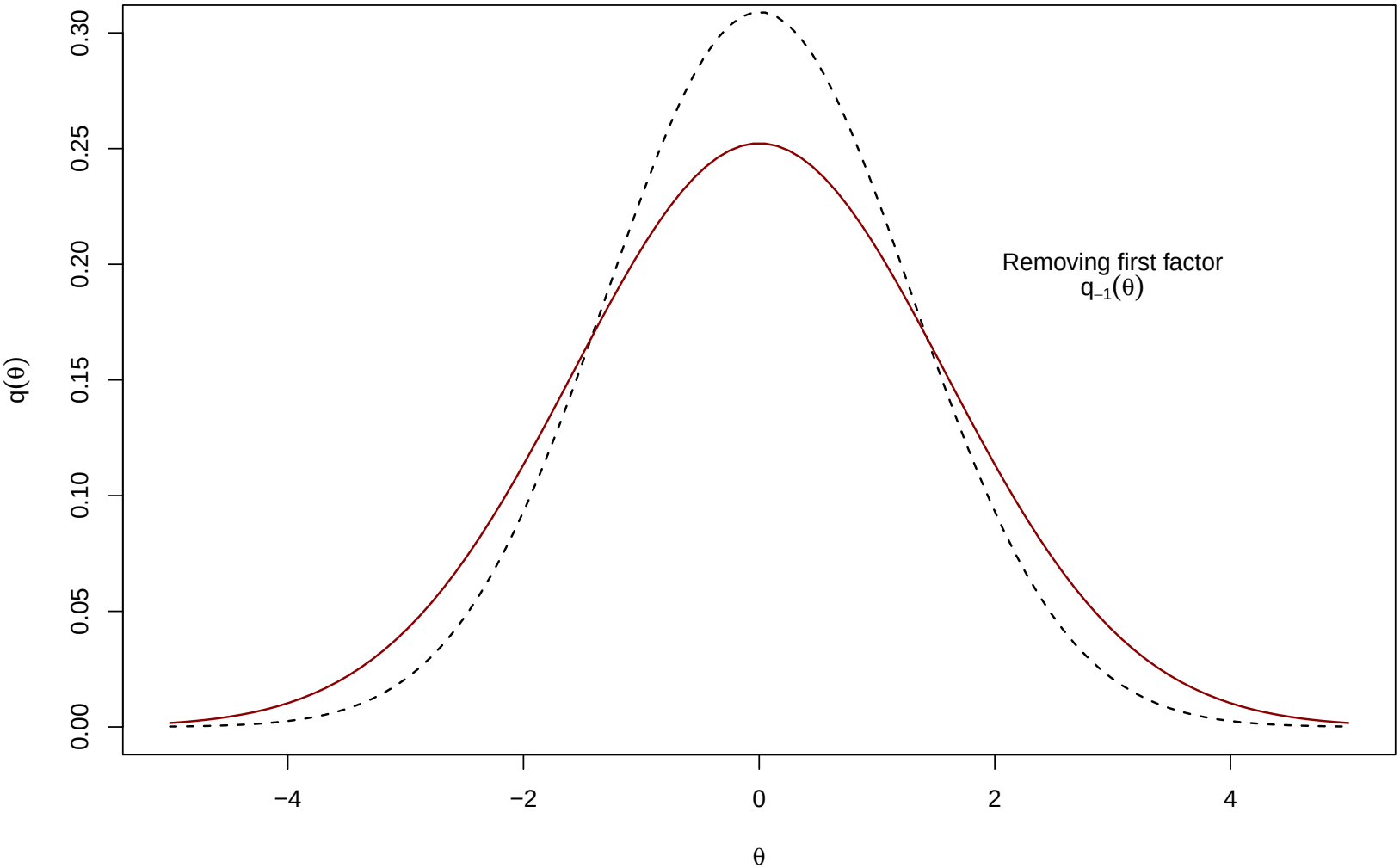
How EP works (IV): update the approximate factor

- The last step is to update q_i, the Gaussian approximation of the factor we've just updated.
- Find the Gaussian q_i such that $q_i q_{-i}$ has the same moments as the hybrid.
- It's a simple linear operation in the natural parameters

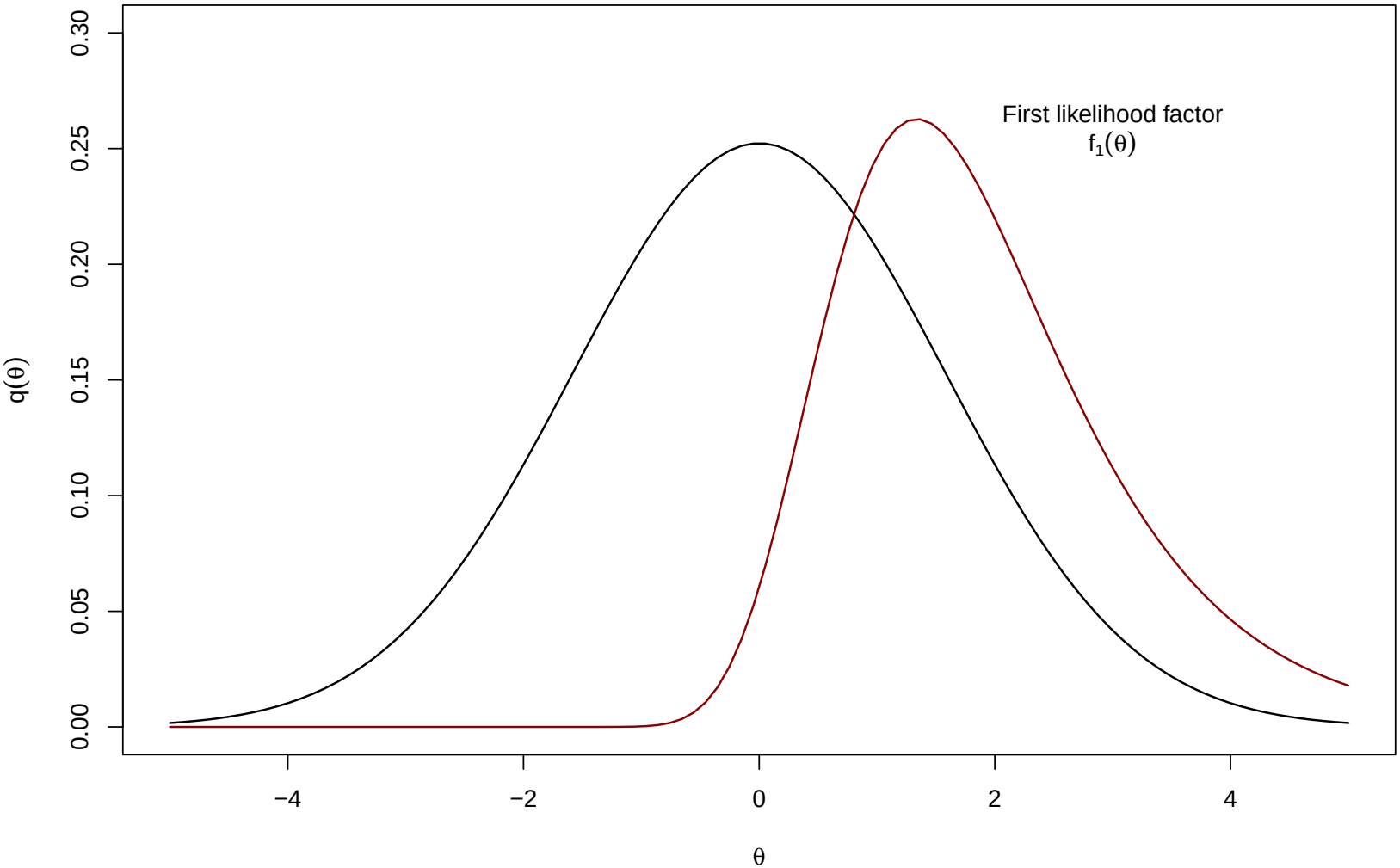
An illustration



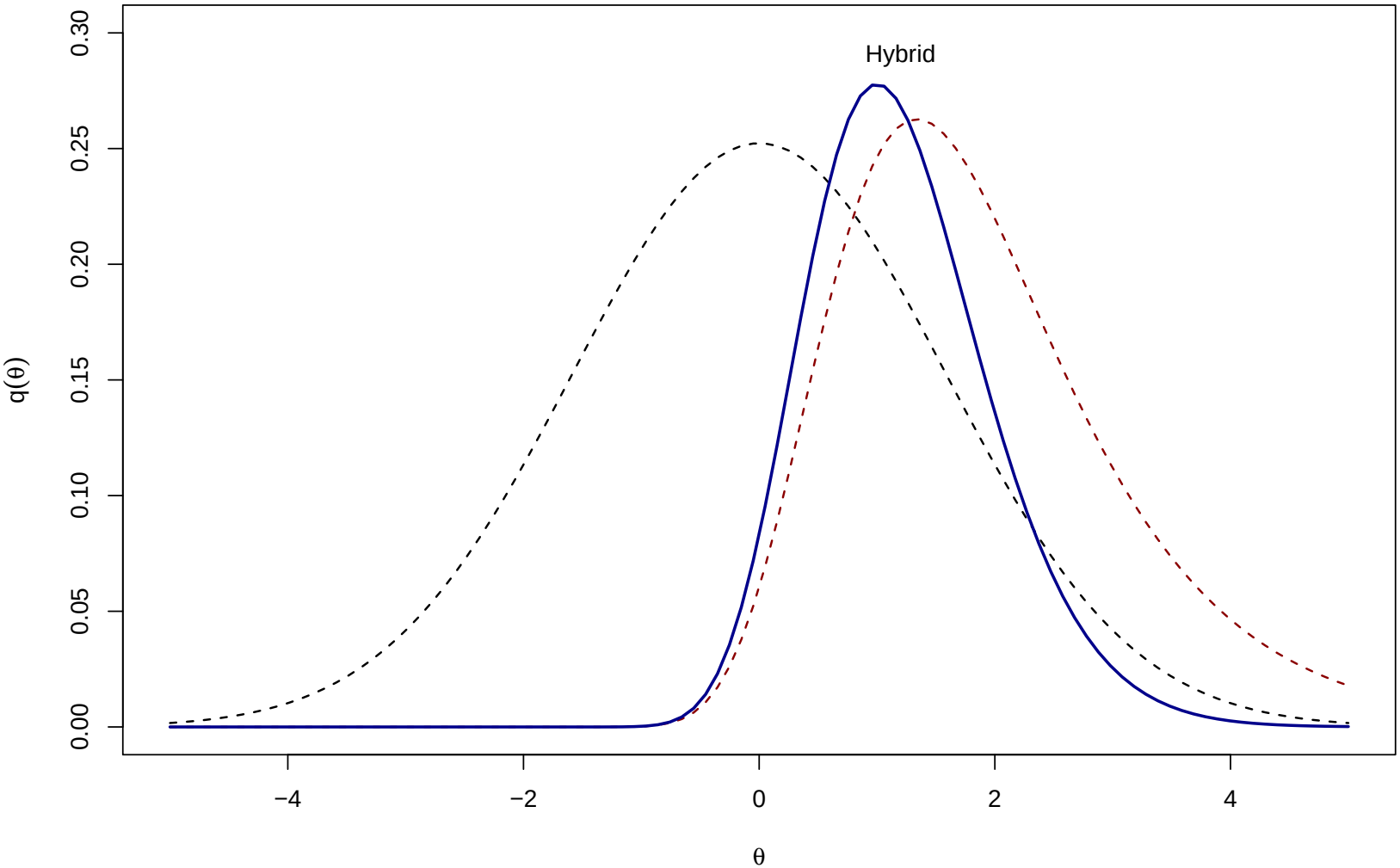
An illustration



An illustration



An illustration



Implementing EP

The basic algorithm is very simple, but good implementations are tricky (and bad ones don't work at all).

Three things will help:

1. The linear subspace property
2. Parallelisation
3. Stabilisation

The linear subspace property

In regression models:

$$y_i = \mathbf{x}_i^t \theta + \epsilon$$

GLMs are similar: each likelihood site only depends on a one-dimensional linear combination of the parameters.

This leads to enormous computational savings and is often what makes EP possible

The linear subspace property

If the likelihood site can be expressed as

$$l_i(\theta) = g_i(\mathbf{A}_i\theta)$$

such that $\mathbf{A}_i\theta$ has dimension $k < p$, then the hybrid moments can be computed in a marginal hybrid distribution of dimension k.

In GLMs $k = 1$, meaning that all the moments can be computed using simple quadrature methods!

Parallelisation

EP is extremely easy to parallelise: you can compute all site updates in parallel instead of one-by-one (Cseke & Heskes, 2011).

In practice mini-batches work best, see Barthelme, Chopin, Cottet (2015): update m sites in parallel, then move on to the next batch.

Stabilisation

- Sometimes EP diverges, especially when faced with relatively nasty likelihood sites.
- Fixes: slow-down iterations, use Power-EP (Minka, 2004), parallel EP.
- I'll come back to the issue when we look at large-n properties.

EP in practice: logistic regression

EP is really good at GLMs, including logistic regression (Ridgway & Chopin, 2015).

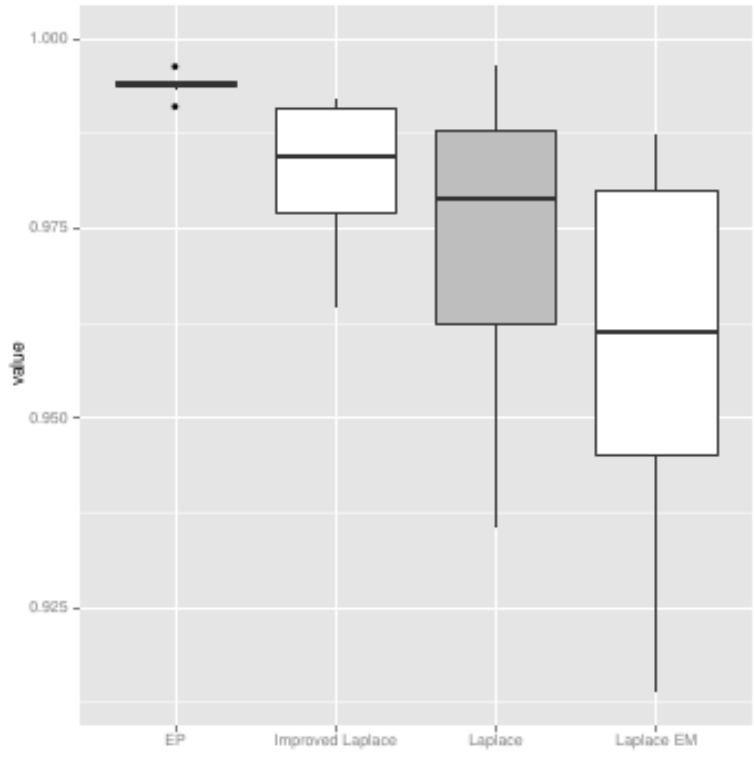
$$p(y_i = 1) = \Phi\left(\mathbf{x}_i^t \theta + \epsilon\right)$$

Ridgway & Chopin evaluated EP's performance on real data (from the UCI datasets), imposing Cauchy priors on the regression coefficients.

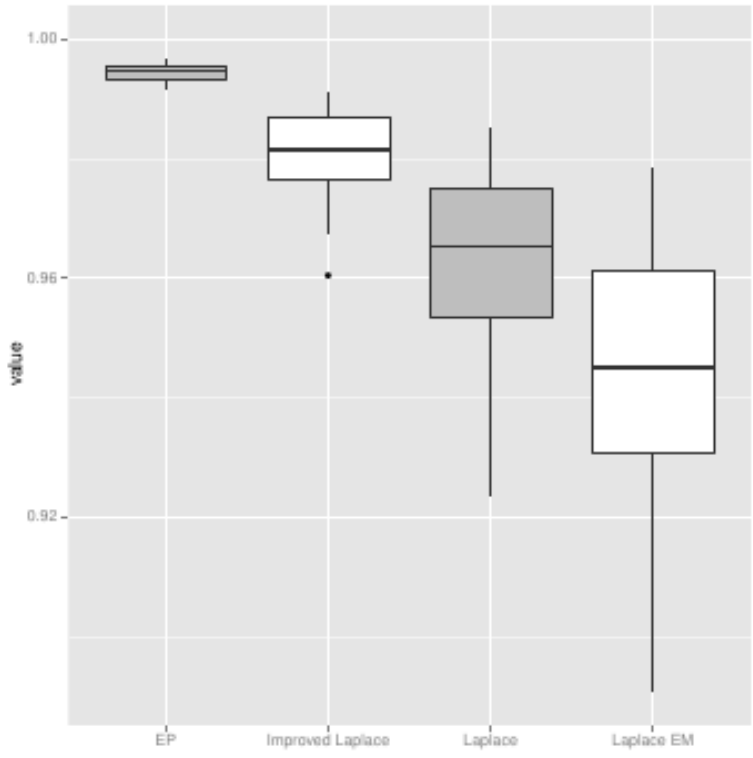
Note that EP must then approximate both likelihood and prior!

The measured error is on the marginals of $p(\theta|\mathbf{y})$.

EP in practice: logistic regression



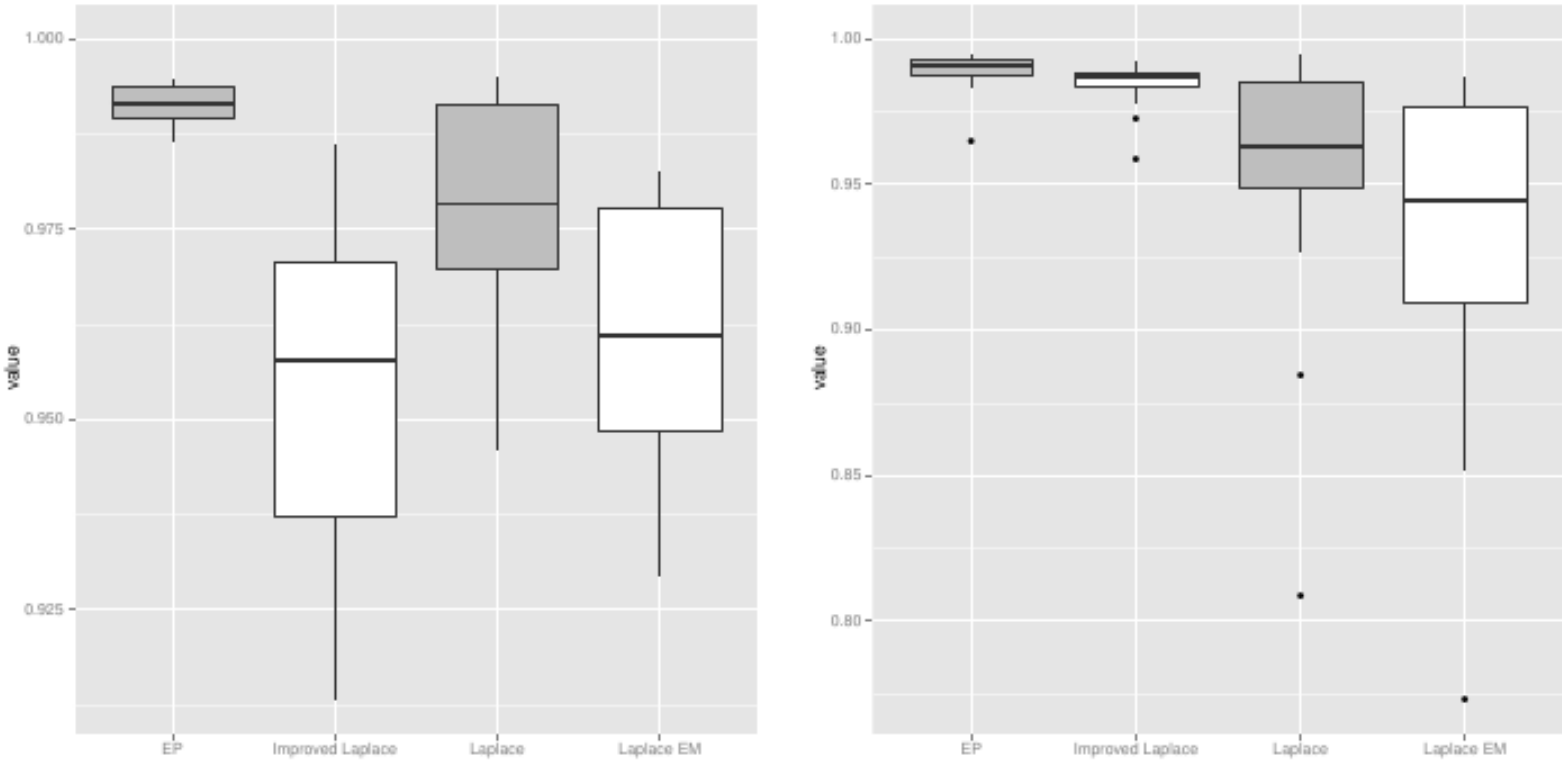
(A) Pima



(B) Heart

Ridgway & Chopin (2015)

EP in practice: logistic regression



(c) Breast (d) German

Ridgway & Chopin (2015)

EP in practice: Gaussian process classification

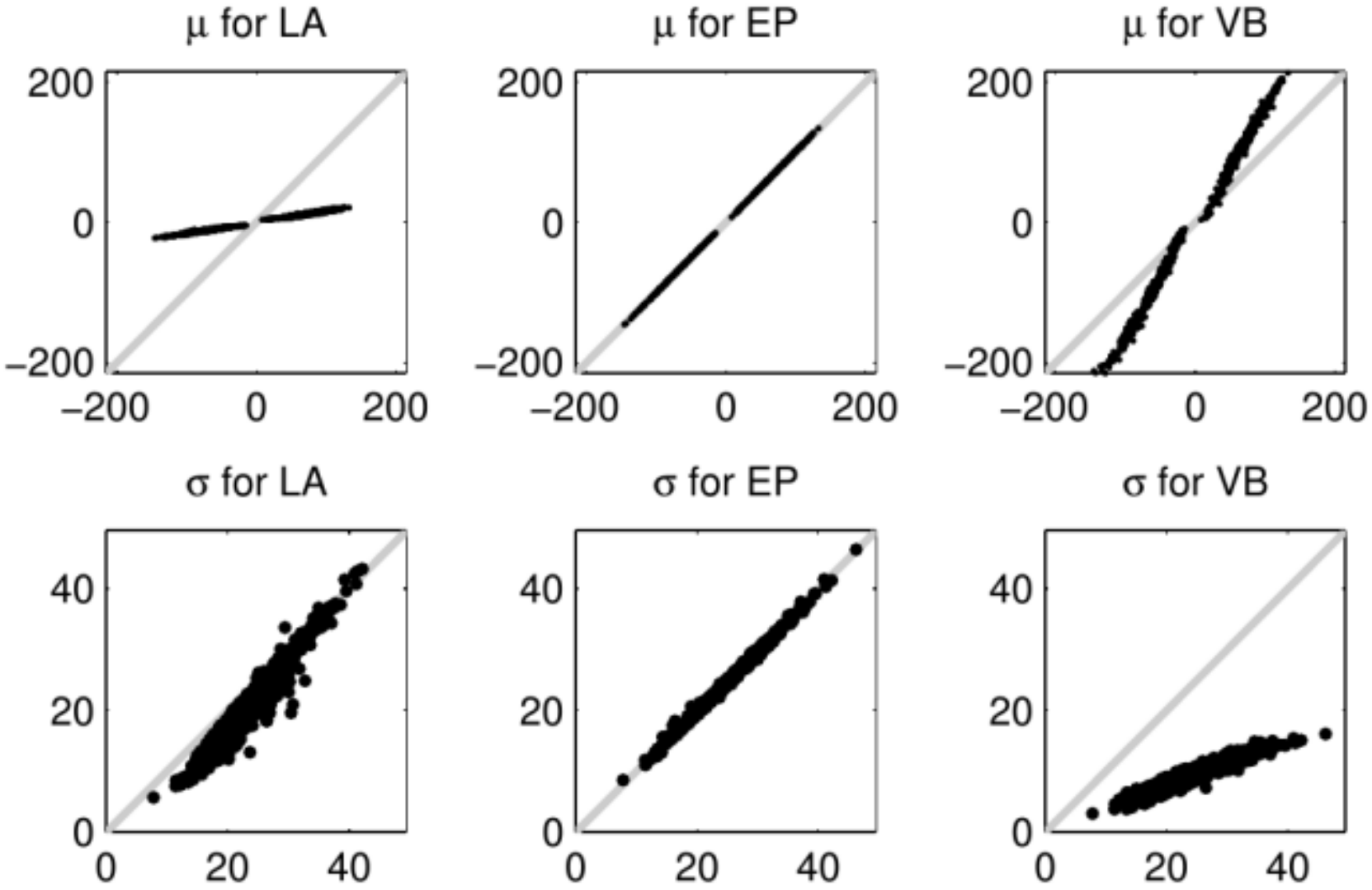
Gaussian Process classification is very similar to probit regression (it's simply a non-parametric variant).

$$p(y_i = 1) = \Phi\left(f(x_i) + \epsilon\right)$$

where $f(x)$ is drawn from a Gaussian process.

The posterior can be quite non-Gaussian because there are many effective parameters compared to the size of the data.

EP in practice: Gaussian process classification



Nickish & Rasmussen (2008)

EP in practice: Approximate Bayesian Computation

In ABC settings the likelihood function is intractable: we cannot compute $p(y|\theta)$, we can only sample it.

It makes inference much harder and much more expensive. EP can speed things up considerably.

EP-ABC

Basic idea: factorise the posterior over datapoints and use the ABC approximation in each factor

$$p_\epsilon(\theta|\mathbf{y}^\star) \propto p(\theta) \prod_{i=1}^{n} \left\{ \int f_i(y_i|\theta) \mathbb{I}_{\left\{ \|y_i - y_i^\star\| \leq \epsilon \right\}} \, dy_i \right\}$$

You can't do that using normal ABC, but you can using EP!

Barthelmé, Chopin (2014). Barthelmé, Chopin, Cottet (2015).

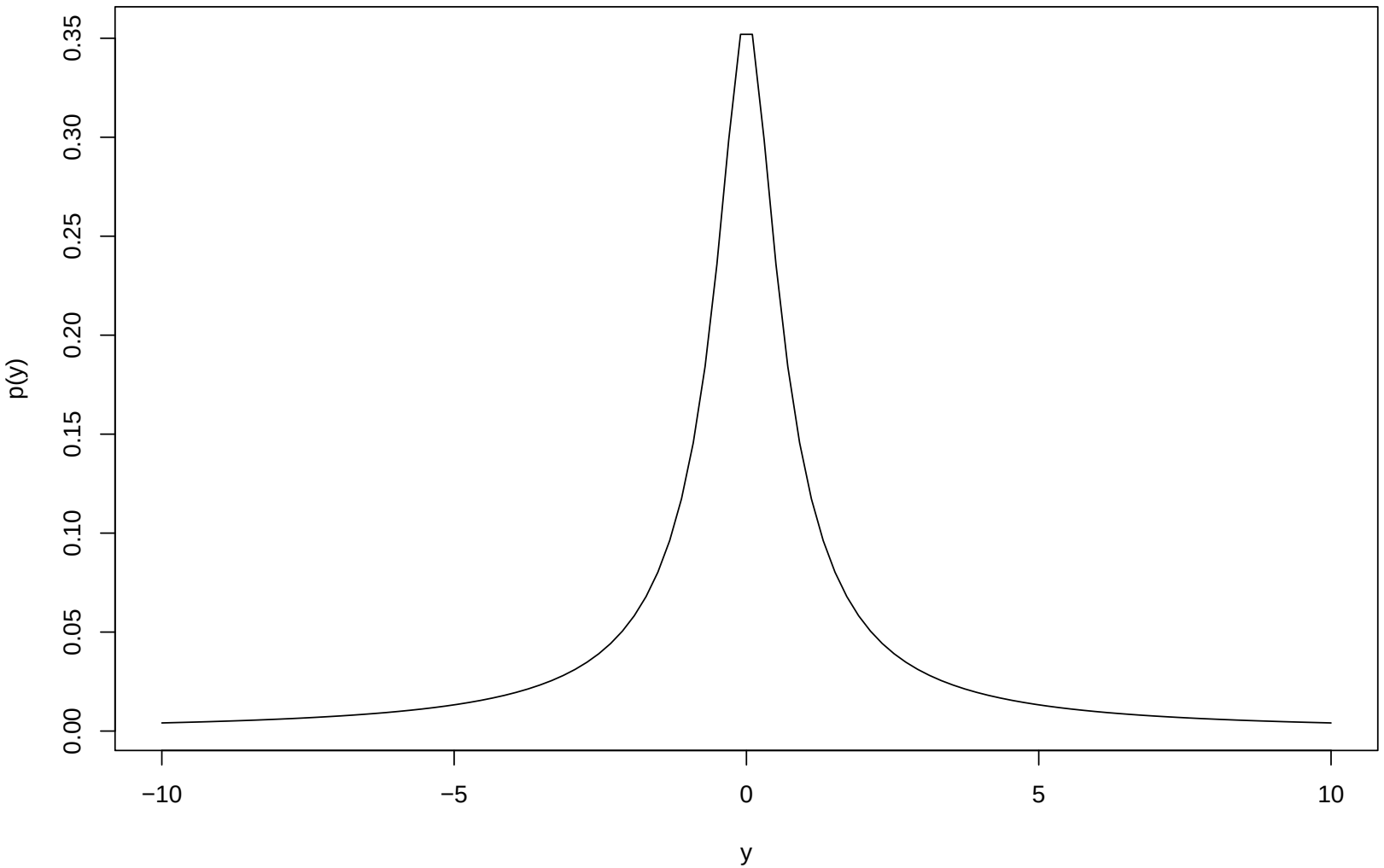
EP-ABC

- Since the posterior distribution is just a product of factors, we can go over the sites one-by-one.
- This means doing just a little ABC step, integrating one datapoint at a time using a very simple rejection mechanism
- Very fast (if properly implemented, which can take a while)

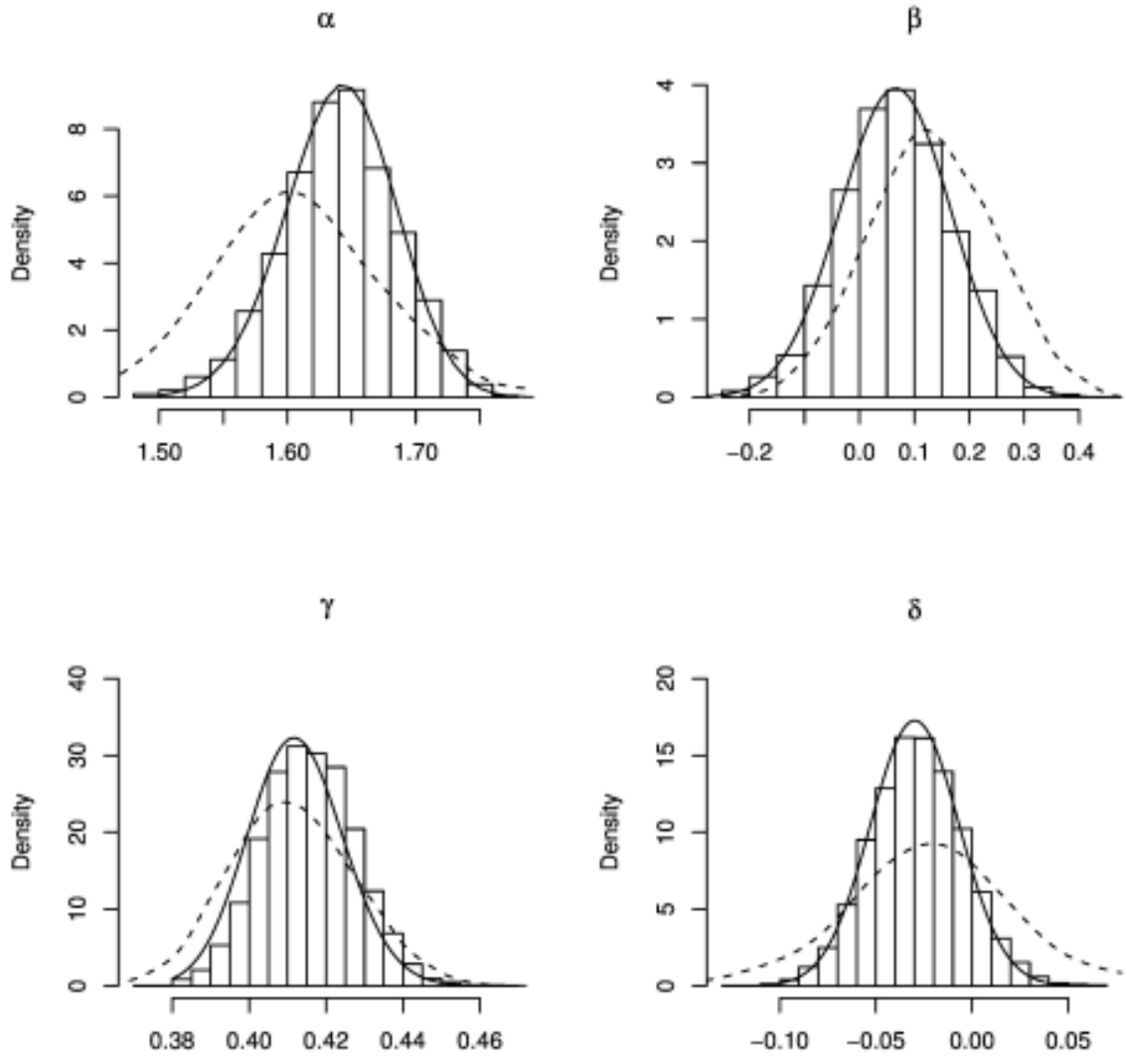
EP-ABC: results in an alpha-stable model

- Alpha-stable distributions are univariate distributions with heavy tails
- Likelihood is very expensive to compute

EP-ABC: results in an alpha-stable model



EP-ABC: results in an alpha-stable model



Barthelmé & Chopin (2014)

Part II: Some Theory

Why does EP work so well?

- Until recently (2014), theory on EP was essentially non-existent.
- Unknown:
  - Do EP iterations converge?
  - Is EP asymptotically exact?
  - Can we give guarantees on how good the approximations are going to be?
- Some progress in Dehaene & Barthelmé (2015a), Dehaene & Barthelmé (2015b)

EP in the large-n regime

- We gave EP the traditional statistics treatment: increase size of the data set, see if algorithm behaves well.
- The math is difficult and we had to use strong regularity conditions (likelihood functions are assumed strongly log-concave).
- We were able to show that *EP is asymptotically exact* and that *EP iterations tend to those of Newton's method*

Large-n assumptions

- We assume that
  - you have n independent datapoints
  - you use a matching factorisation with n sites
  - the likelihood obeys some (strong-ish) regularity conditions
- We study the $n \leftrightarrow \infty$ limit

The Gaussian limit of posterior distributions

- Bernstein-von Mises theorem: posterior distributions become Gaussian around the mode
- Specifically:

$$\lim p(\theta|y_1 \dots y_n) = N(\theta^\star, \mathbf{H}^{-1})$$

  where $\mathbf{H}$ is the Hessian matrix at the mode
- See Panov & Spokoiny (2015) for a proof in a very general framework

Result 1: (Gaussian) EP is asymptotically exact

- We show that in the large-n limit, EP recovers the limiting Gaussian posterior
- In particular, this implies good frequentist properties:
- EP estimates converge to the MLE
- EP estimates are asymptotically efficient and consistent

Result 2: EP is asymptotically *very* accurate

- Main result of Dehaene & Barthelmé, NIPS, 2015:

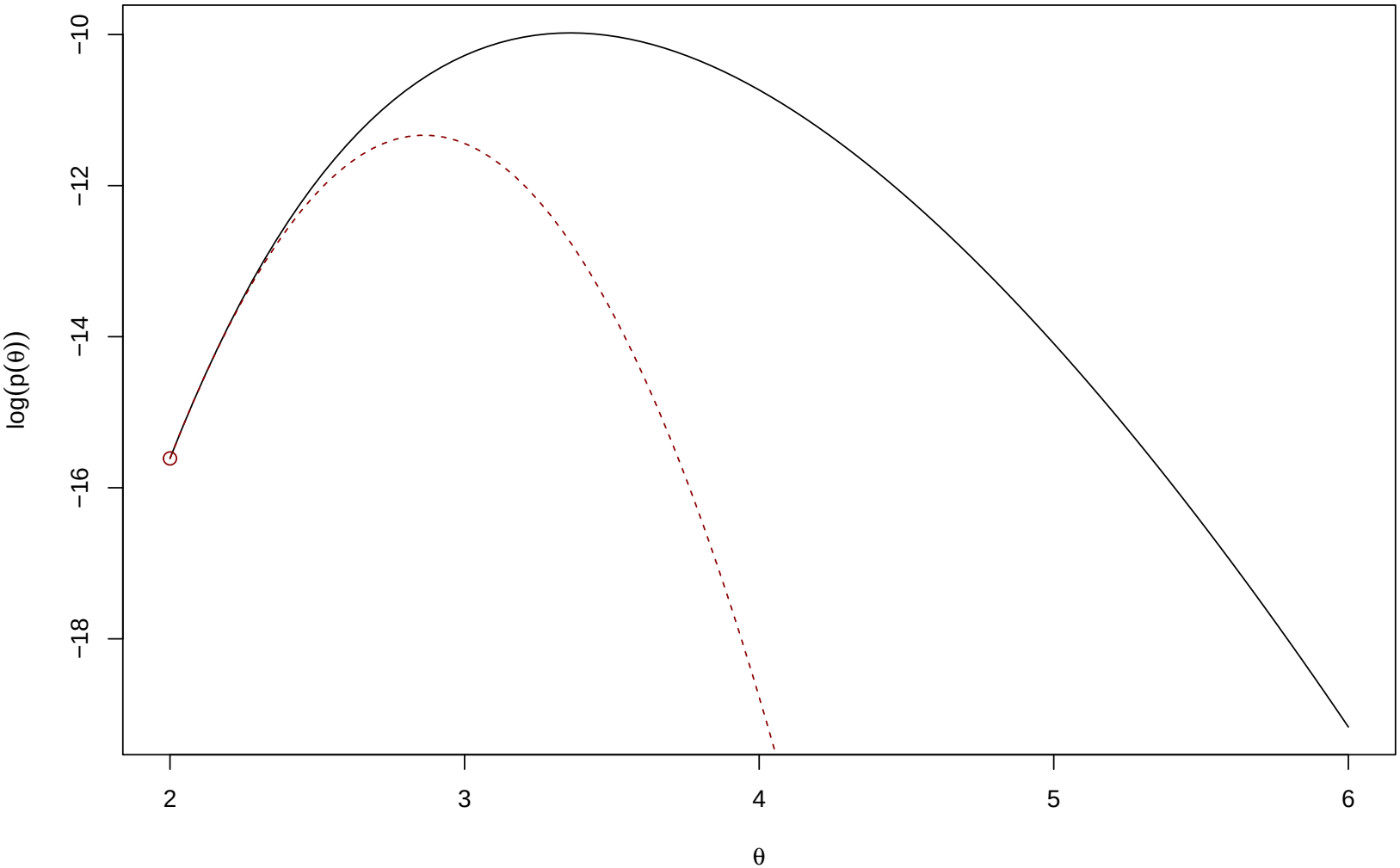
$$|\mu_{ep} - \mu| \leq \mathcal{O}(n^{-2})$$

- In other words, EP's approximation of the mean converges *very fast*
- Laplace: $\mathcal{O}(n^{-1})$ rate
- Warning: result derived under very strong (unrealistic) assumptions

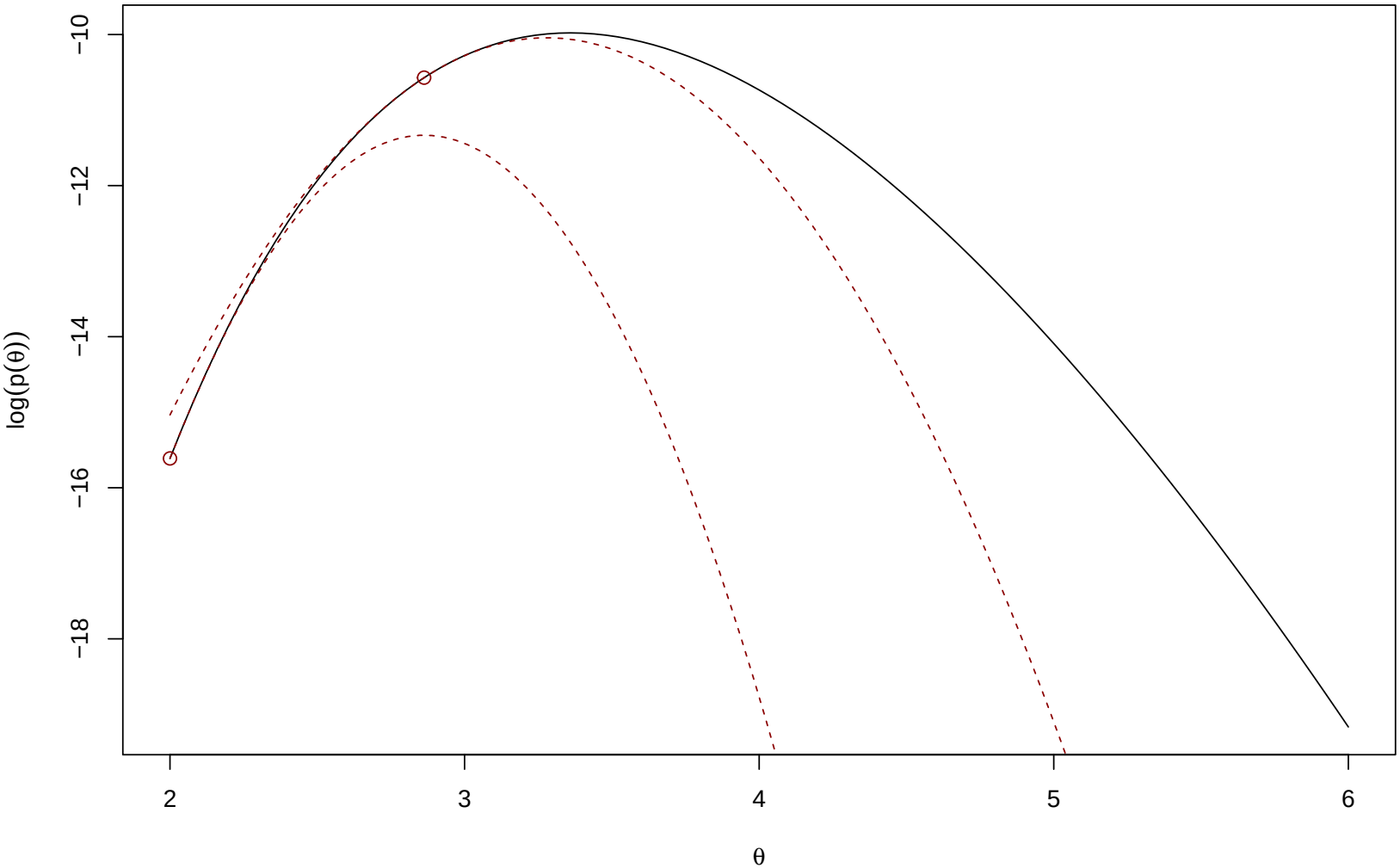
Result 3: EP resembles Newton's method

- The previous result concerns *fixed points* of EP
- The next result is interesting for the dynamics of EP (how the iterations behave)
- In large n, (parallel) EP resembles Newton's method, meaning they produce a similar sequence of Gaussian approximations

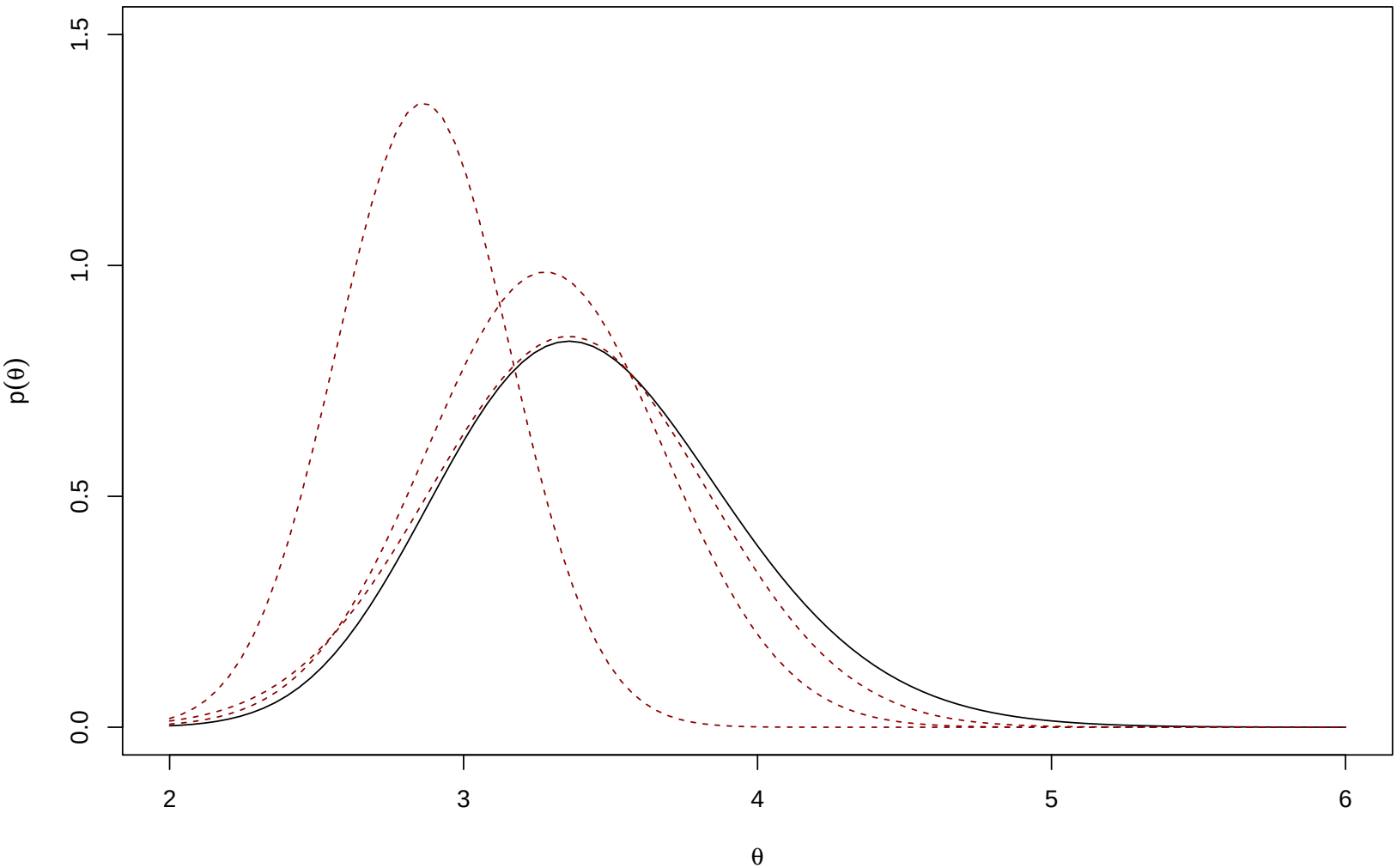
Newton's method



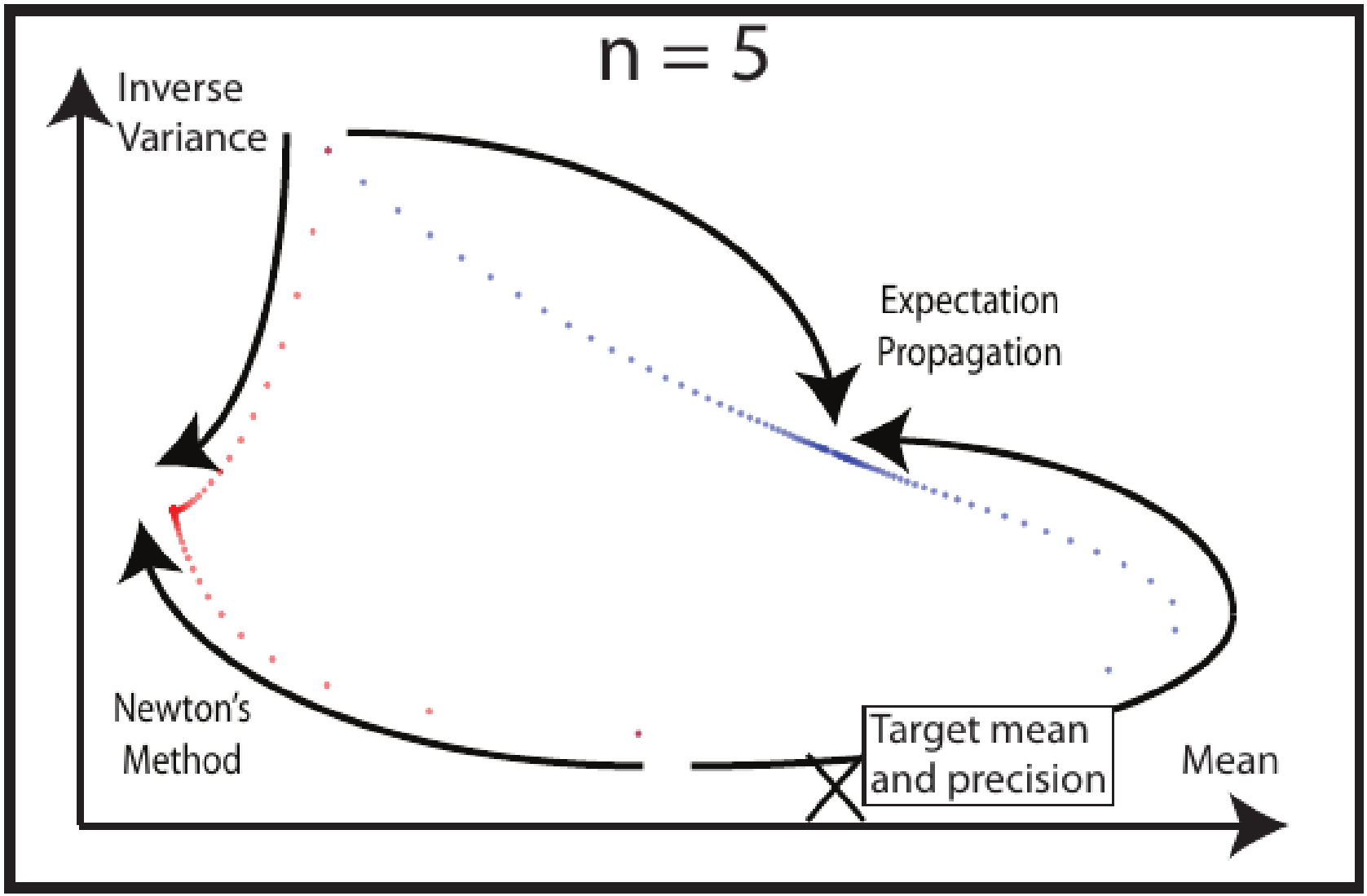
Newton's method



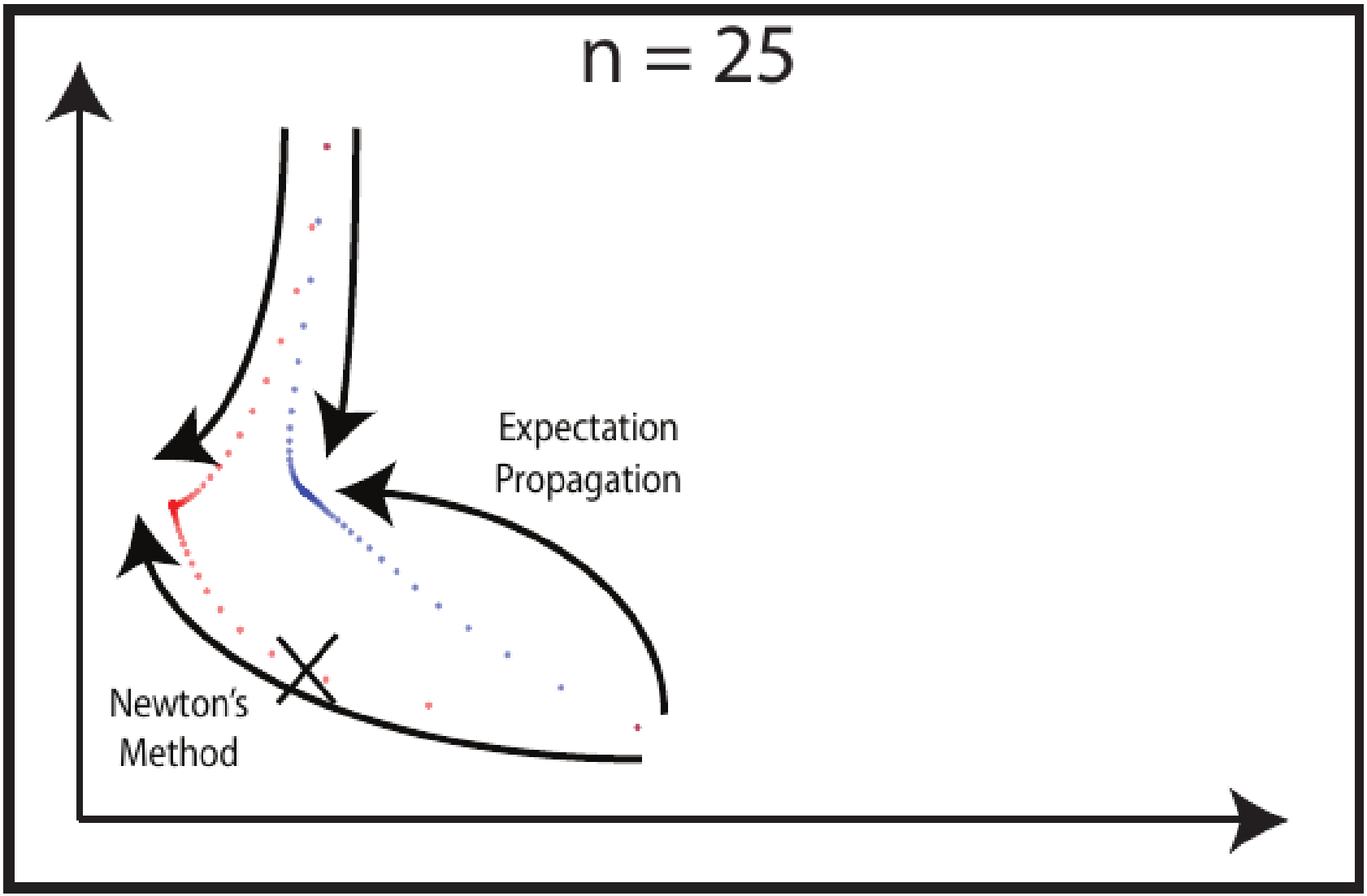
Newton's method as a sequence of Gaussian approx.



EP tends to Newton's



EP tends to Newton's



Consequences

- Newton's method can diverge when initialised far from the mode and so can EP!
- Gives you an easy way to check for potential problems
- Either:
  - slow down iterations
  - initialise EP near the mode

Further consequence

- EP will get captured by local modes, meaning that if the distribution is multimodal it will center on a single mode
- This is contrary to what you'd expect from the usual motivation from KL divergence
- It does however make EP usable on mixture models

Conclusion (theory part)

- If your likelihood is well-behaved (log-concave) then EP is guaranteed to do no worse than a Laplace approximation and will ordinarily do much better.
- If your posterior distribution is multi-modal with well-separated modes then everything will depend on the starting point. You should get good approximations of the local modes.
- If your posterior distribution has heavy-tails or a high-dimensional banana shape, we can't say much.

Conclusion (overall)

- EP is a powerful algorithm than can replace MCMC in a variety of problems
- In ABC applications it can provide substantial speedups.
- It now comes with theoretical guarantees

Caveats:

- The cost of implementation is relatively high
- Scaling in large p (number of parameters) is a problem

Collaborators

ABC-EP: Nicolas Chopin (ENSAE) and Vincent Cottet (ENSAE)

Asymptotics of EP: Guillaume Dehaene (Université de Genève)

Other current directions

Large-n problems: combine EP and MCMC. Invented independently by Xu et al. (2014), Gelman et al. (2014).

Using EP to speed up MCMC: EP for pseudo-marginals (Fillipone & Girolami, 2014)

Corrections to EP: you can improve the results of EP using the hybrids (Paquet, Winter, Opper, 2013).

References

Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence (pp. 362-369). Morgan Kaufmann Publishers Inc.

Panov, M., & Spokoiny, V. (2015). Finite Sample Bernstein–von Mises Theorem for Semiparametric Problems. Bayesian Analysis, 10(3), 665-710.

Filippone, M., & Girolami, M. (2014). Pseudo-marginal Bayesian inference for Gaussian processes. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 36(11), 2214-2226.

Xu, M.,Lakshminarayanan, B., Teh, Y. W., Zhu, J., & Zhang, B. (2014). Distributed Bayesian posterior sampling via moment sharing. In Advances in Neural Information Processing Systems (pp. 3356-3364).

Gelman, A.,Vehtari, A., Jylänki, P., Robert, C., Chopin, N., & Cunningham, J. P. (2014). Expectation propagation as a way of life. arXiv preprint arXiv:1412.4869.

Opper, M., Paquet, U., & Winther, O. (2013). Perturbative corrections for approximate inference in gaussian

latent variable models. The Journal of Machine Learning Research, 14(1), 2857-2898.