Expectation Propagation in the large data limit

Guillaume Dehaene, Simon Barthelmé

February 29, 2016

Bayesian inference is hard

- Bayesian inference is a powerful statistical methodology
- But for most generative models, it's also computationally intractable

$$p(x|O_1 \dots O_n) = p(x) \prod_i p(O_i|x)$$

Bayesian inference is hard

- Bayesian inference is a powerful statistical methodology
- But for most generative models, it's also computationally intractable

$$p(x|O_1 \dots O_n) = p(x) \prod_i p(O_i|x)$$

- What can we do ?
  - Point estimates ! (maximum likelihood, MAP)

Bayesian inference is hard

- Bayesian inference is a powerful statistical methodology
- But for most generative models, it's also computationally intractable

$$p(x|O_1 \dots O_n) = p(x) \prod_i p(O_i|x)$$

- What can we do ?
  - Point estimates ! (maximum likelihood, MAP)
  - Sampling methods ! Generate $X \sim p(x|O_1 \dots O_n)$

Bayesian inference is hard

- Bayesian inference is a powerful statistical methodology
- But for most generative models, it's also computationally intractable

$$p(x|O_1 \dots O_n) = p(x) \prod_i p(O_i|x)$$

- What can we do ?
  - Point estimates ! (maximum likelihood, MAP)
  - Sampling methods ! Generate $X \sim p(x|O_1 \dots O_n)$
  - Approximate inference ! Find $q \approx p$

Expectation Propagation

- It's used to match players in skill level in Halo (Microsoft True Skill, XBox)

EP is powerful ...

- EP has great potential:
  - It's powerful (Kuss et al, 05; Nickish et al, 08):
    - Empirically, it gives high-quality approximations at minimal cost
  - It's universal:
    - it can be applied to any $p(x)$ with a simple factor structure
  - Can perform the computation in parallel

but poorly understood !!

- But EP is also very poorly known !!
- Open questions:
  - How good are the approximations ?
  - Does it always terminate ?

but poorly understood !!

- But EP is also very poorly known !!
- Open questions:
    - How good are the approximations ?
    - Does it always terminate ?
- We've been able to tackle those questions in the large-data limit:
    - We prove it gives good approximations
    - We prove that it has a simple limit behavior

Contents

Notations

We will approximate a 1D probability distribution $p(x)$ **that has a simple factor structure**

$$\begin{aligned} p(x|O_1 \dots O_n) &= p(x) \prod_{i=1}^{n} p(O_i|x) \\ p(x) &= \prod_{i=1}^{n} f_i(x) \end{aligned}$$

Notations

We will approximate a 1D probability distribution $p(x)$ **that has a simple factor structure**

$$p(x|O_1 \dots O_n) = p(x) \prod_{i=1}^{n} p(O_i|x)$$
$$p(x) = \prod_{i=1}^{n} f_i(x)$$

We will approximate p with a Gaussian g that also factorizes:

$$g(x) = \prod_{i=1}^{n} g_i(x) \approx p(x)$$
$$\forall i \quad g_i(x) \approx f_i(x)$$

Notations

We will approximate a 1D probability distribution $p(x)$ **that has a simple factor structure**

$$\begin{aligned} p(x|O_1 \ldots O_n) &= p(x) \prod_{i=1}^{n} p(O_i|x) \\ p(x) &= \prod_{i=1}^{n} f_i(x) \end{aligned}$$

We will approximate p with a Gaussian g that also factorizes:

$$\begin{aligned} g(x) = \prod_{i=1}^{n} g_i(x) &\approx p(x) \\ \forall i \quad g_i(x) &\approx f_i(x) \end{aligned}$$

We will often work with negative logs:

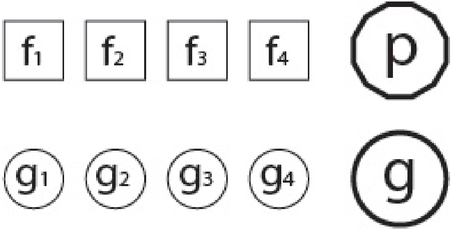
$$\begin{aligned} \psi(x) &= -\log[p(x)] \\ \phi_i(x) &= -\log[f_i(x)] \end{aligned}$$

The EP loop

- EP aims to find a factorized approximation of $p(x)$:

$$g(x) = \prod_{i=1}^{n} g_i(x) \approx \prod_{i=1}^{n} f_i(x) = p(x)$$

- EP proceeds **iteratively**
- The basic idea: How do we improve a current approximation $[g_i^t(x)]$??

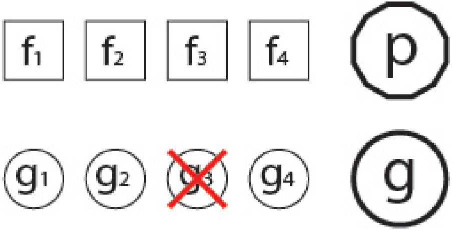


The EP loop

- EP aims to find a factorized approximation of $p(x)$:

$$g(x) = \prod_{i=1}^{n} g_i(x) \approx \prod_{i=1}^{n} f_i(x) = p(x)$$

- EP proceeds **iteratively**
- The basic idea: How do we improve a current approximation $[g_i^t(x)]$??

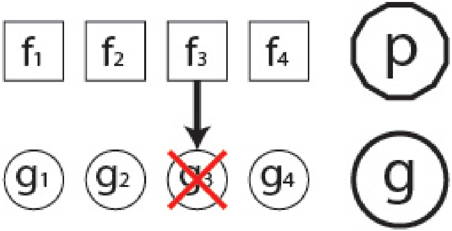


The EP loop

- EP aims to find a factorized approximation of $p(x)$:

$$g(x) = \prod_{i=1}^{n} g_i(x) \approx \prod_{i=1}^{n} f_i(x) = p(x)$$

- EP proceeds **iteratively**
- The basic idea: How do we improve a current approximation $[g_i^t(x)]$??

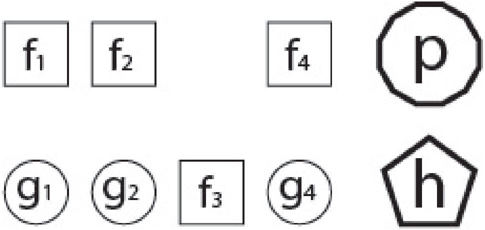


The EP loop

- EP aims to find a factorized approximation of $p(x)$:

$$g(x) = \prod_{i=1}^{n} g_i(x) \approx \prod_{i=1}^{n} f_i(x) = p(x)$$

- EP proceeds **iteratively**
- The basic idea: How do we improve a current approximation $[g_i^t(x)]$??

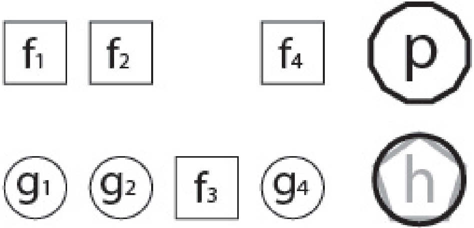


The EP loop

- EP aims to find a factorized approximation of $p(x)$:

$$g(x) = \prod_{i=1}^{n} g_i(x) \approx \prod_{i=1}^{n} f_i(x) = p(x)$$

- EP proceeds **iteratively**
- The basic idea: How do we improve a current approximation $[g_i^t(x)]$??

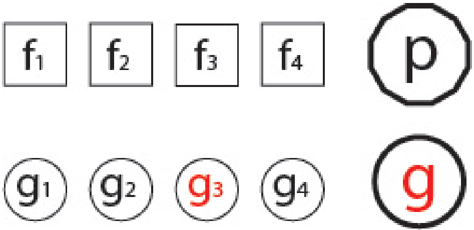


The EP loop

- EP aims to find a factorized approximation of $p(x)$:

$$g(x) = \prod_{i=1}^{n} g_i(x) \approx \prod_{i=1}^{n} f_i(x) = p(x)$$

- EP proceeds **iteratively**
- The basic idea: How do we improve a current approximation $[g_i^t(x)]$??



The EP loop

- Select i for updating
- Compute

$$h_i(x) = f_i(x) \prod_{j \neq i} g_j(x)$$

- Compute a Gaussian approximation:

$$g^{t+1}(x) \approx h_i(x)$$

- update the approximation of f_i:

$$g_i^{t+1} = \frac{g^{t+1}}{\prod_{j \neq i} g_j(x)}$$

The EP loop

- Select i for updating
- Compute

$$h_i(x) = f_i(x) \prod_{j \neq i} g_j(x)$$

- Compute a Gaussian approximation:

$$g^{t+1}(x) \approx h_i(x)$$

- update the approximation of f_i:

$$g_i^{t+1} = \frac{g^{t+1}}{\prod_{j \neq i} g_j(x)}$$

- Terminology:
  - $g_{-i} = \prod g_j$ is the **cavity distribution**
  - h_i is the **hybrid**

Approximating the hybrid

- How do we compute $g^{t+1} \approx h_i$?

Approximating the hybrid

- How do we compute $g^{t+1} \approx h_i$?
- Minimize the **Kullback-Leibler** divergence

$$g^{t+1} = \mathrm{argmin}_g KL\left(h_i, g\right)$$

- Gives a good approximation
- Is simple to compute

Minimizing KL

- Inside exponential families, minimizing KL is easy

Minimizing KL

- Inside exponential families, minimizing KL is easy
- Gaussians are an exponential family:

$$g(x|r,\beta) \propto \exp\left(rx - \beta\frac{x^2}{2}\right)$$

Minimizing KL

- Inside exponential families, minimizing KL is easy
- Gaussians are an exponential family:

$$g(x|r,\beta) \propto \exp\left(rx - \beta\frac{x^2}{2}\right)$$

- Relation between r, β and the moments:

$$\begin{aligned} \mu &= \frac{r}{\beta} \\ \text{var} &= \beta^{-1} \end{aligned}$$

Minimizing KL

To find $\text{argmin}_g KL(h_i, g)$

- Compute the mean and variance of h_i
- Compute the Gaussian with that mean and variance:

$$\begin{aligned} r &= \text{var}^{-1}\mu \\ \beta &= \text{var}^{-1} \end{aligned}$$

Working in natural parameters

- Working in the space of Gaussians: $g_i \in \mathcal{G}$: impossible to visualize

Working in natural parameters

- Working in the space of Gaussians: $g_i \in \mathcal{G}$: impossible to visualize
- Working in moments:

$$\begin{aligned} \mu_i &= E_{g_i}(x) \\ v_i &= \mathrm{var}_{g_i}(x) \end{aligned}$$

Better, but hard to multiply and divide Gaussians

Working in natural parameters

- Working in the space of Gaussians: $g_i \in \mathcal{G}$: impossible to visualize
- Working in moments:

$$\begin{aligned} \mu_i &= E_{g_i}(x) \\ v_i &= \mathrm{var}_{g_i}(x) \end{aligned}$$

Better, but hard to multiply and divide Gaussians

- Working in natural parameters:

$$g_i(x) \propto \exp\left(r_i x - \beta_i \frac{x^2}{2}\right)$$

Multiplication and division of Gaussians = sums and differences of natural parameters !

EP in natural parameters

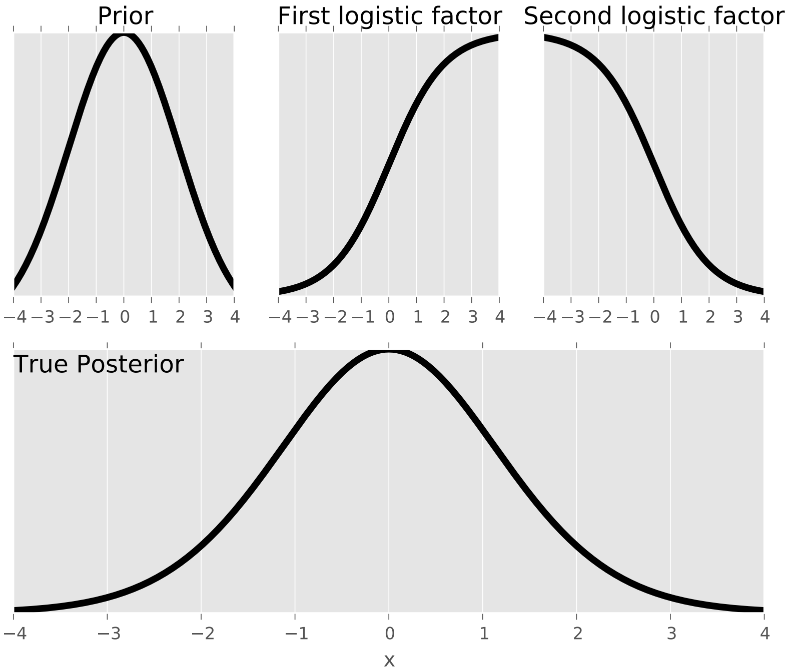
Sequential algorithm, operating on (2n) dimensional space $[r_i, \beta_i]$

- For i in $1 \dots n$
  1. Compute "cavity" parameters: $r_{-i} = \sum_{j \neq i} r_j$, $\beta_{-i} \sum_{j \neq i} \beta_j$
     1. Compute hybrid distribution $h_i(x) = f_i(x) \mathcal{N}(x | r_{-i}, \beta_{-i})$
     2. Compute $E_{h_i}(x)$ and $var_{h_i}(x)$
     3. Update r_i and β_i from the moments of the hybrid

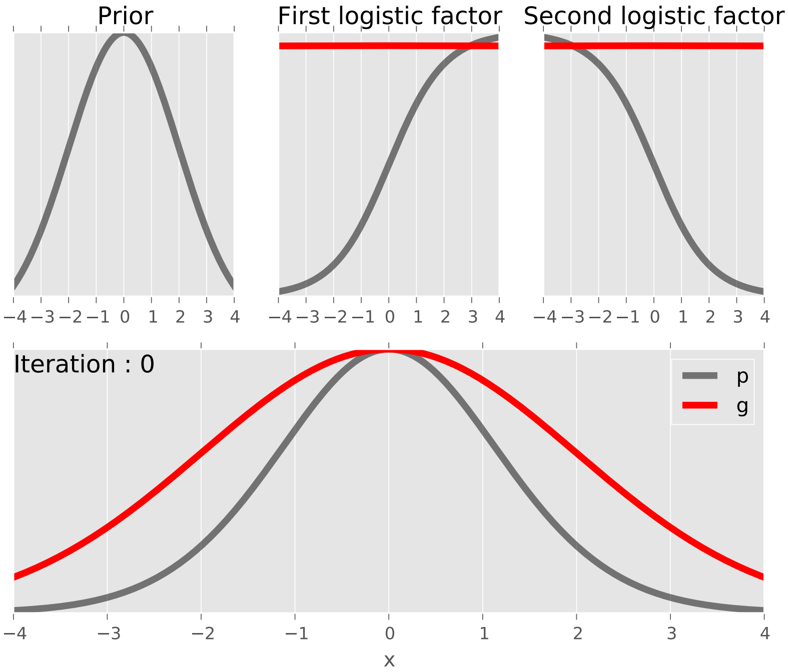
$$
\begin{aligned}
r_i &= \frac{E_{h_i}(x)}{var_{h_i}} - r_{-i} \\
\beta_i &= \frac{1}{var_{h_i}} - \beta_{-i}
\end{aligned}
$$

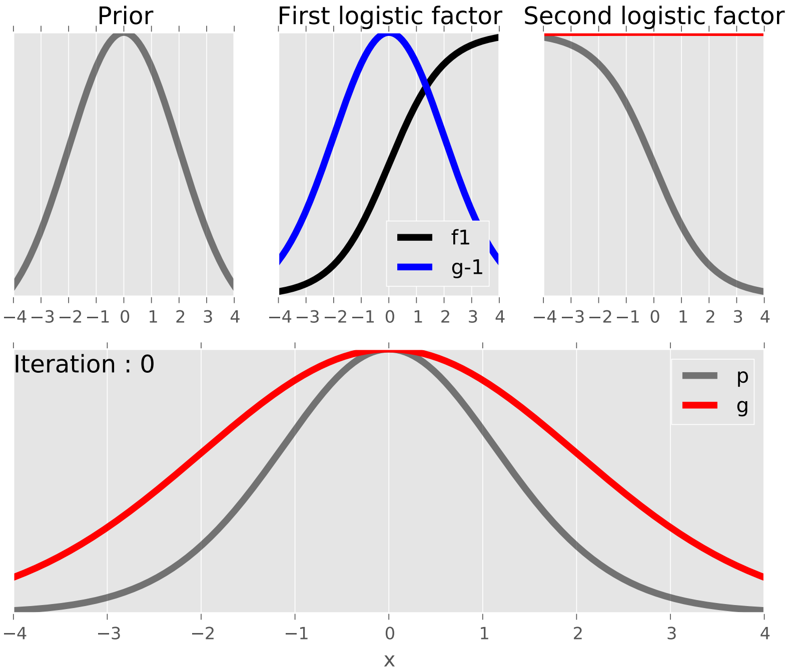
An example !

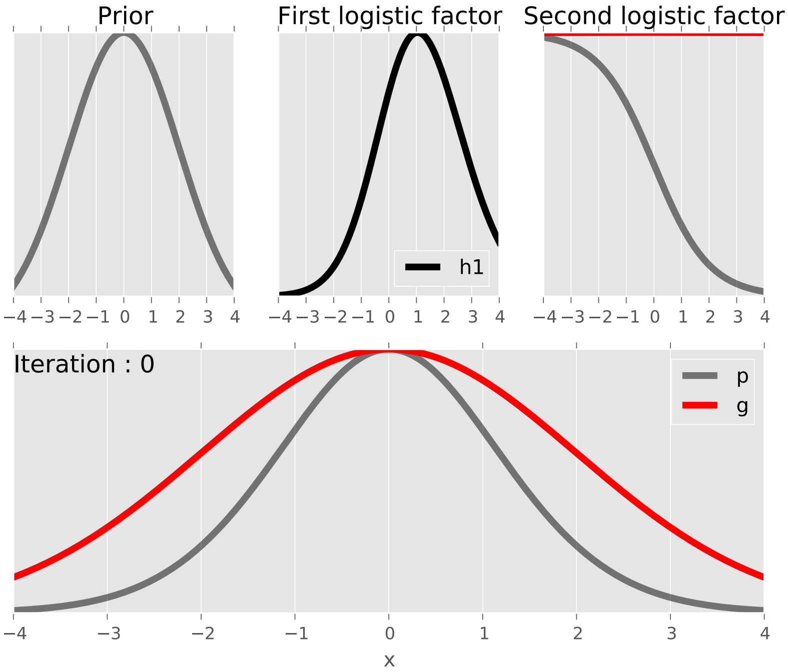
- 3 factors $f_i(x)$:
    - 2 logistic (likelihoods)
    - 1 Gaussian (prior)

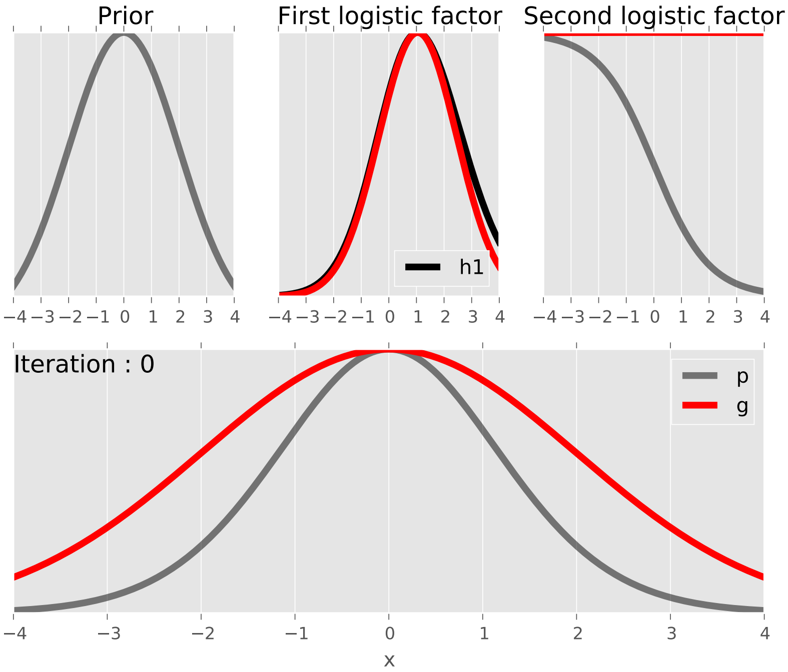


Prior
First logistic factor
Second logistic factor
−4 −3 −2 −1 0 1 2 3 4
True Posterior
−4 −3 −2 −1 0 1 2 3 4
x

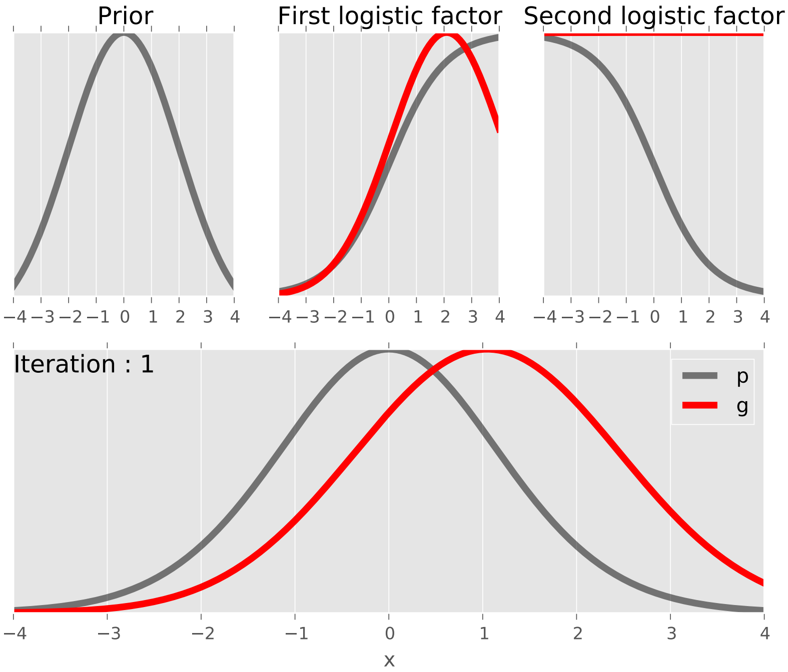

Prior
First logistic factor
Second logistic factor
−4 −3 −2 −1 0 1 2 3 4
Iteration : 0
p
g
x


Prior
First logistic factor
Second logistic factor
f1
g-1
Iteration : 0
p
g
−4
−3
−2
−1
0
1
2
3
4
x

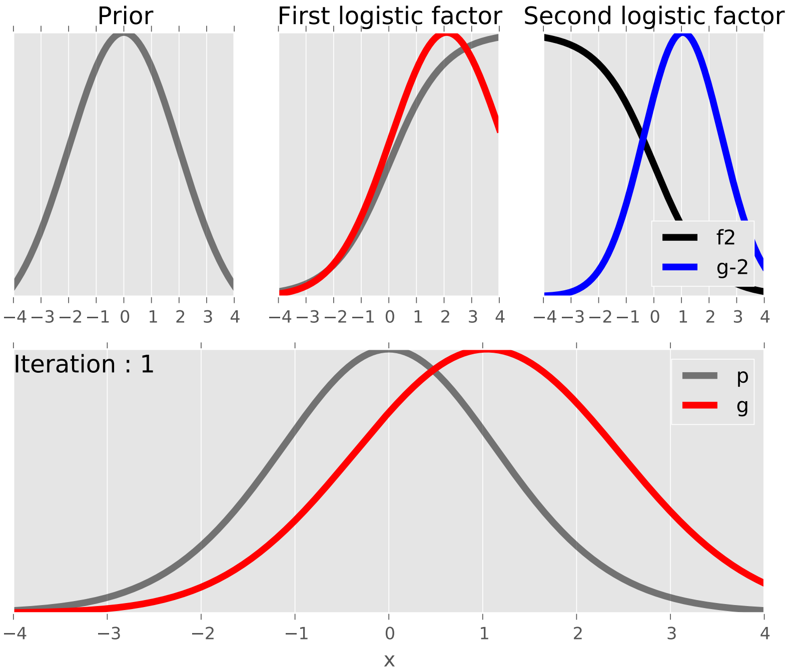

Prior
First logistic factor
Second logistic factor
h1
Iteration : 0
p
g
−4 −3 −2 −1 0 1 2 3 4
x



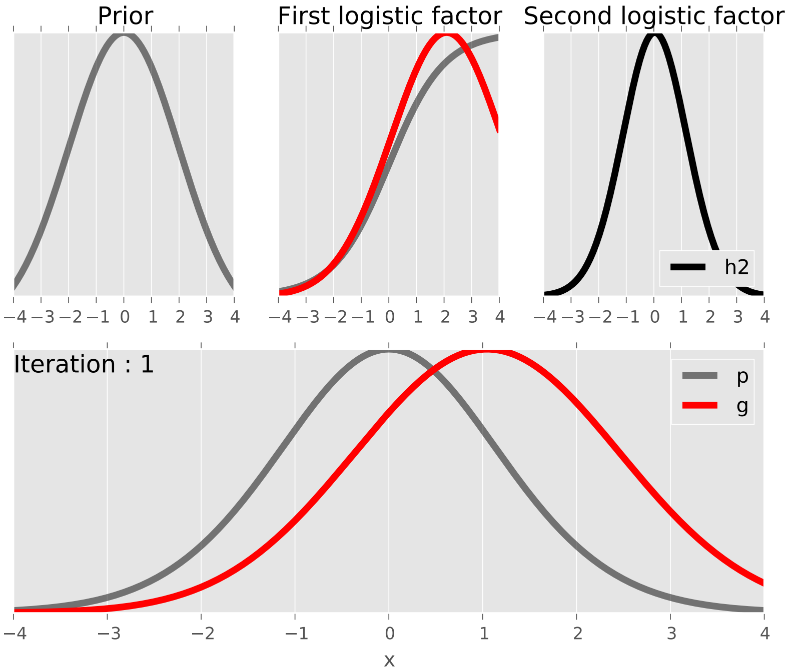
Prior
First logistic factor
Second logistic factor
h1
Iteration : 0
p
g
x



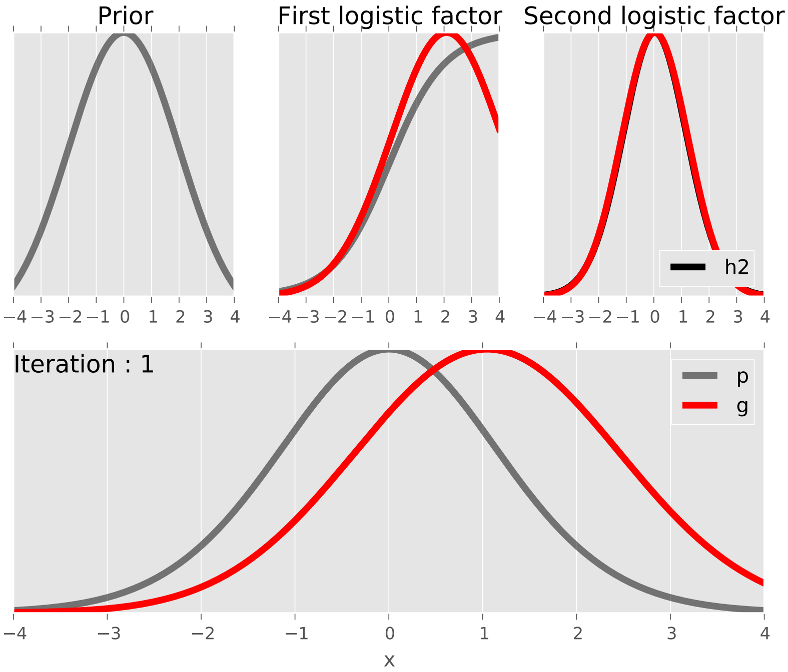
Prior
First logistic factor
Second logistic factor
−4 −3 −2 −1 0 1 2 3 4
Iteration : 1
p
g
−4 −3 −2 −1 0 1 2 3 4
x



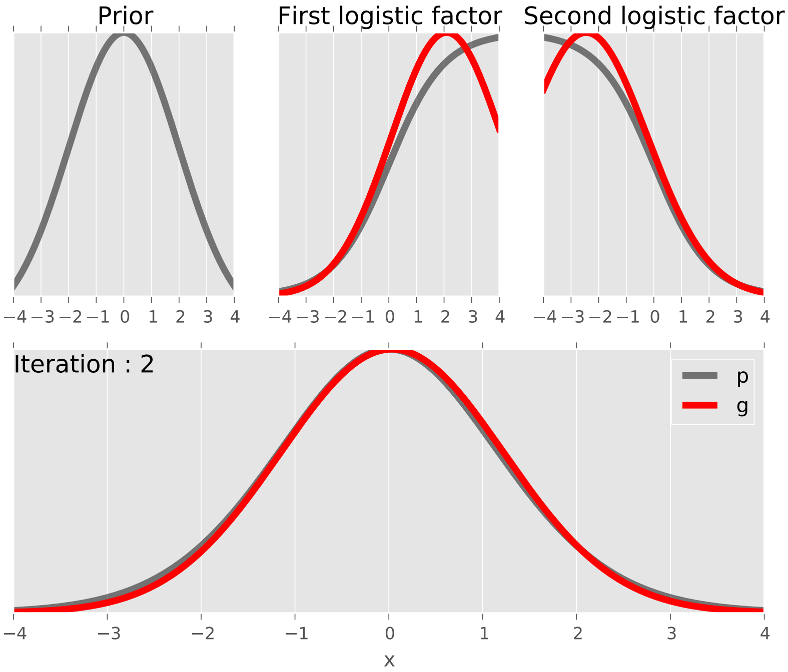
Prior
First logistic factor
Second logistic factor
f2
g-2
Iteration : 1
p
g
x



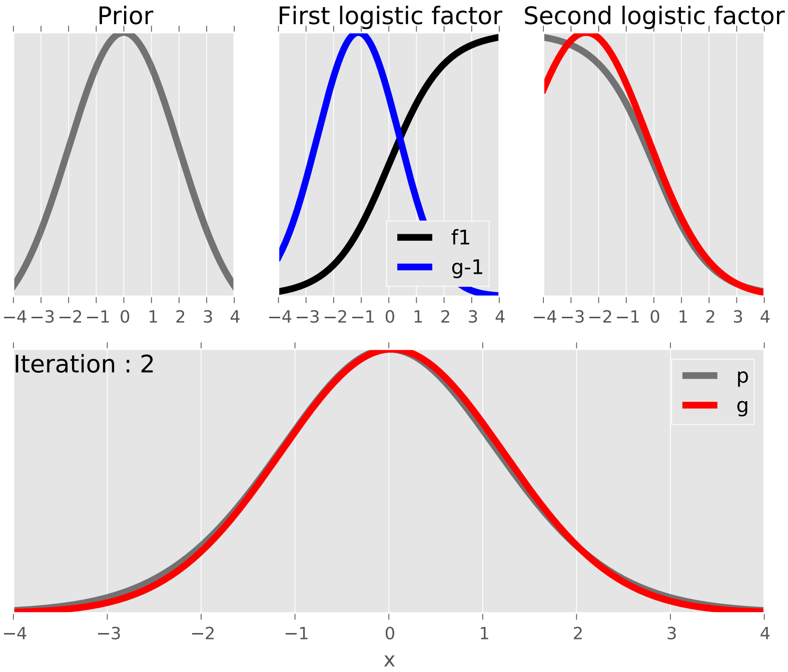
Prior
First logistic factor
Second logistic factor
h2
Iteration : 1
p
g
x



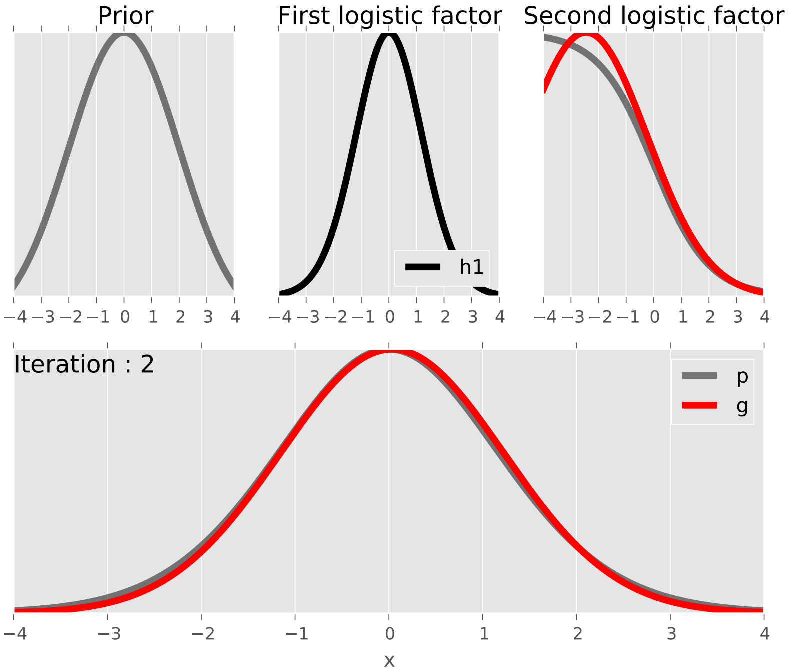
Prior
First logistic factor
Second logistic factor
h2
Iteration : 1
p
g
−4
−3
−2
−1
0
1
2
3
4
x



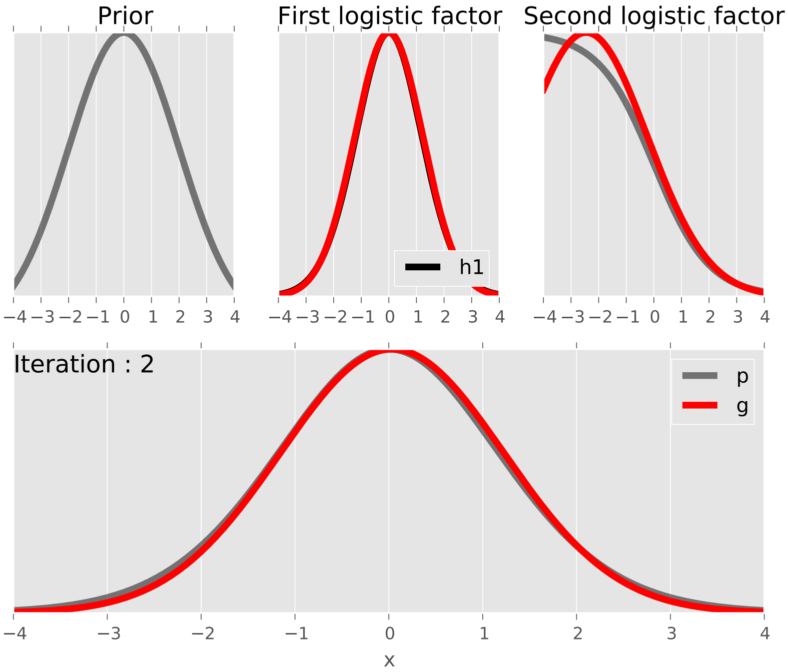
Prior
First logistic factor
Second logistic factor
−4 −3 −2 −1 0 1 2 3 4
Iteration : 2
p
g
−4 −3 −2 −1 0 1 2 3 4
x



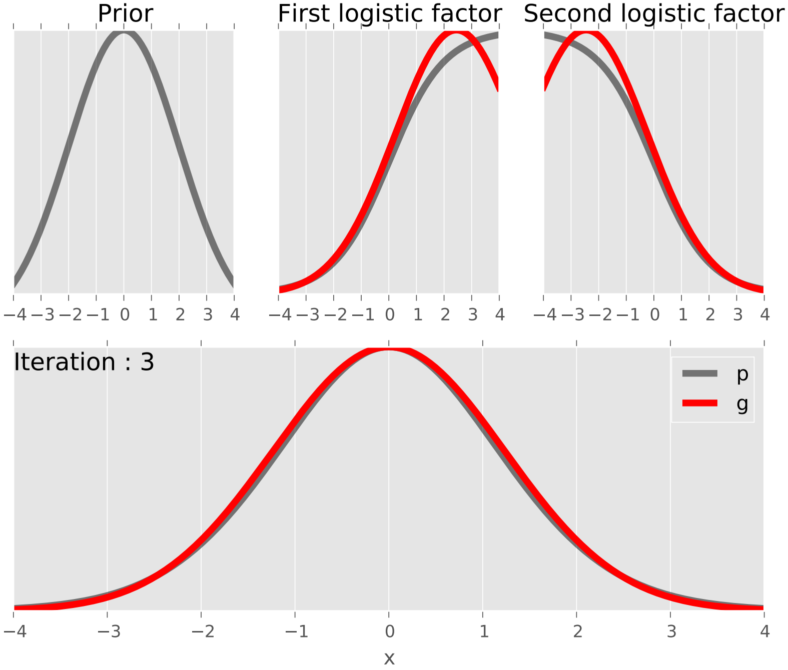
Prior
First logistic factor
Second logistic factor
f1
g-1
−4 −3 −2 −1 0 1 2 3 4
Iteration : 2
p
g
−4 −3 −2 −1 0 1 2 3 4
x

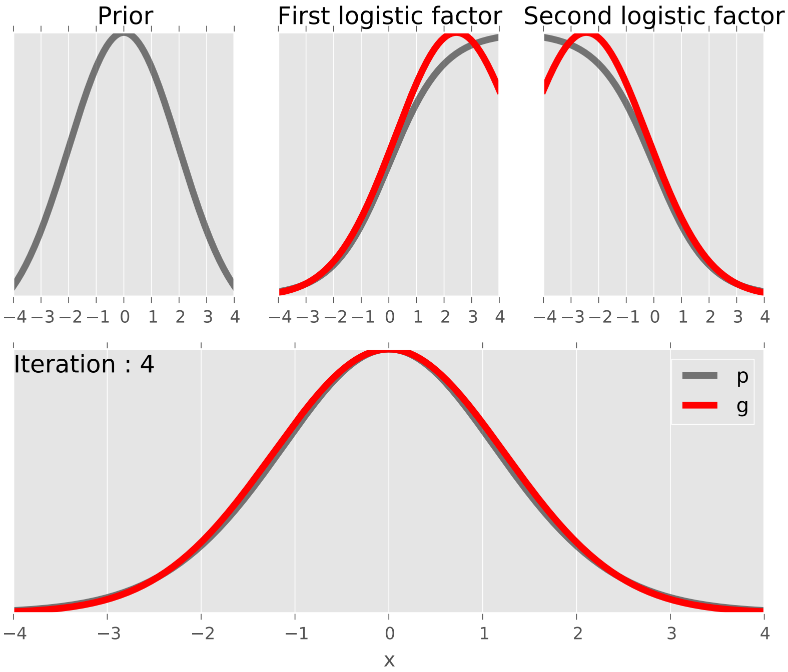


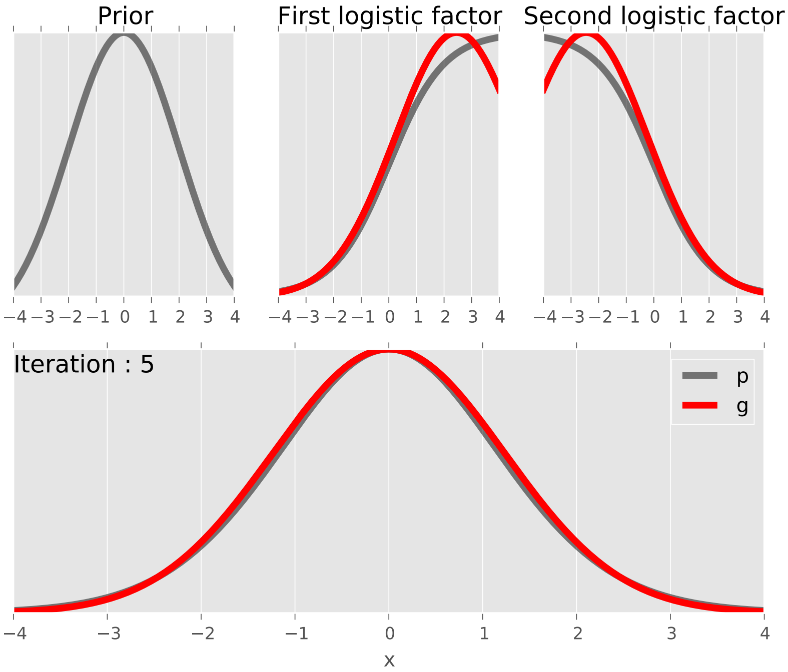
Prior
First logistic factor
Second logistic factor
h1
Iteration : 2
p
g
x

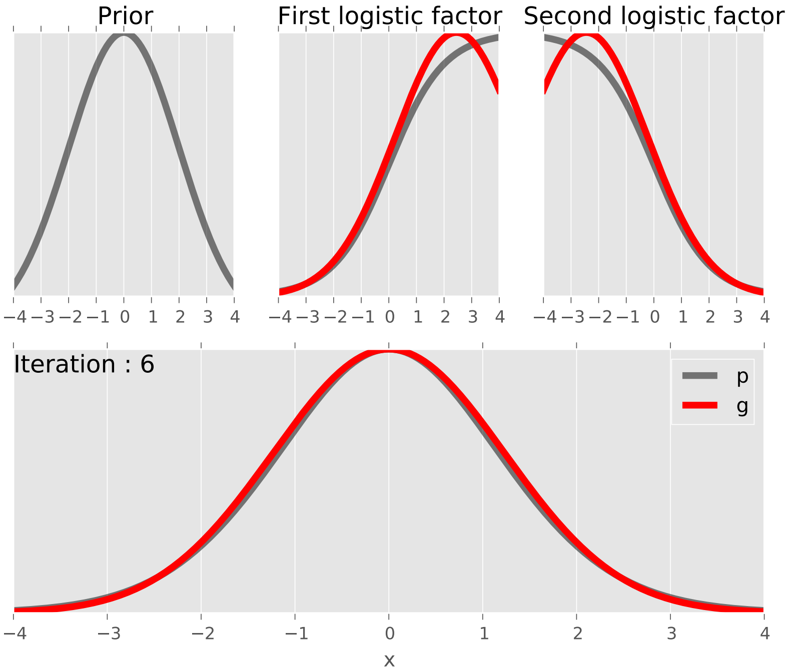


Prior
First logistic factor
Second logistic factor
h1
Iteration : 2
p
g
−4 −3 −2 −1 0 1 2 3 4
x

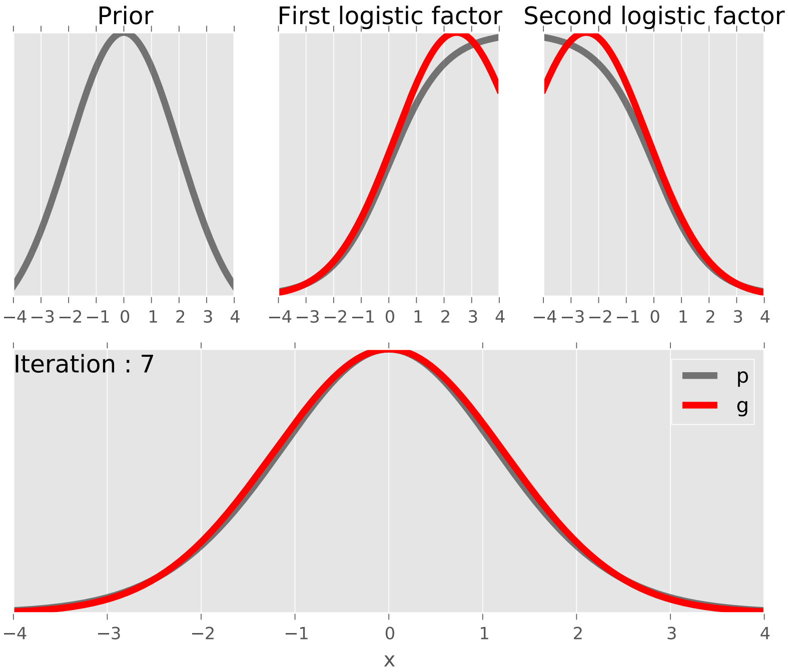

Prior
First logistic factor
Second logistic factor
−4 −3 −2 −1 0 1 2 3 4
−4 −3 −2 −1 0 1 2 3 4
−4 −3 −2 −1 0 1 2 3 4
Iteration : 3
p
g
−4 −3 −2 −1 0 1 2 3 4
x


Prior
First logistic factor
Second logistic factor
Iteration : 4
p
g
x
−4 −3 −2 −1 0 1 2 3 4


Prior
First logistic factor
Second logistic factor
Iteration : 5
p
g
x



Prior
First logistic factor
Second logistic factor
Iteration : 6
p
g
x


Prior
First logistic factor
Second logistic factor
Iteration : 7
p
g
x
−4
−3
−2
−1
0
1
2
3
4

EP variants

- We presented the base algorithm which is **sequential**:
  - Pick i, then update g_i, ...

EP variants

- We presented the base algorithm which is **sequential**:
    - Pick i, then update g_i, ...
- We can also:
    - Update all approximations at once (**parallel EP**)
    - Update 10% (batch EP)
    - Update asynchronously

EP variants

- We presented the base algorithm which is **sequential**:
  - Pick i, then update g_i, ...
- We can also:
  - Update all approximations at once (**parallel EP**)
  - Update 10% (batch EP)
  - Update asynchronously
- We can also "slow-down" the updates

EP variants

- We presented the base algorithm which is **sequential**:
  - Pick i, then update g_i, ...
- We can also:
  - Update all approximations at once (**parallel EP**)
  - Update 10% (batch EP)
  - Update asynchronously
- We can also "slow-down" the updates
- All of those don't modify the fixed-points !

EP summary

- EP approximates each factor f_i as a Gaussian g_i, and refines these approximations iteratively
- $h_i = f_i g_{-i}$ is a better approximation than g

EP summary

- EP approximates each factor f_i as a Gaussian g_i, and refines these approximations iteratively
- $h_i = f_i g_{-i}$ is a better approximation than g
- The parameter space is the natural parameters: (r_i, β_i)
- Variants of EP modify the updating schedule, or change the updating rule

When to use EP

- EP **apparently** works well if all h_i are almost Gaussian
- EP is dangerous to use on multimodal distributions. Like VB, EP sometimes fits a single mode of $p(x)$, missing most of the probability mass
- The EP iteration can be frustrating:
    - **slow it down** or **do it sequentially**

The large-data limit

- Large-data limit: number of observations tends to ∞
- Frequentist result: **Central Limit Theorem**: the distribution of empirical means become Gaussian

The large-data limit

- Large-data limit: number of observations tends to ∞
- Frequentist result: **Central Limit Theorem**: the distribution of empirical means become Gaussian
- Bayesian result: **Bernstein-von Mises**: posteriors converge to Gaussian distributions
- And the variance quickly goes to 0:

$$\mathrm{var}_p(x) \propto n^{-1}$$

The large-data limit and EP

- If approximate inference methods aren't exact in the large-data limit, they shouldn't be used

The large-data limit and EP

- If approximate inference methods aren't exact in the large-data limit, they shouldn't be used
- The large-data limit makes theoretical analysis simple
    - The influence of a single factor f_i becomes negligible

The large-data limit and EP

- If approximate inference methods aren't exact in the large-data limit, they shouldn't be used
- The large-data limit makes theoretical analysis simple
  - The influence of a single factor f_i becomes negligible
  - In the hybrid distribution, $h_i = f_i g_{-i}$, the cavity dominates

Contents

EP gives very good approximations

- We know empirically that fixed-points of EP give very good approximations of $p(x)$:
    - can we prove it ?

Assumptions

- We will constrain the factors $f_i(x)$

Assumptions

- We will constrain the factors $f_i(x)$
- We will assume that all $f_i \propto \exp(-\phi_i)$ are **strongly log-concave**:

$$\phi_i''(x) \geq \beta_m$$

- This is an unrealistic assumption

Assumptions

- We will constrain the factors $f_i(x)$
- We will assume that all $f_i \propto \exp(-\phi_i)$ are **strongly log-concave**:

$$\phi_i''(x) \geq \beta_m$$

- This is an unrealistic assumption
- We will assume that the higher-derivatives are bounded:

$$\left|\phi_i^{(d)}(x)\right| \leq K_d$$

Assumptions

- We will constrain the factors $f_i(x)$
- We will assume that all $f_i \propto \exp(-\phi_i)$ are **strongly log-concave**:

$$\phi_i''(x) \geq \beta_m$$

- This is an unrealistic assumption
- We will assume that the higher-derivatives are bounded:

$$\left|\phi_i^{(d)}(x)\right| \leq K_d$$

- These assumptions transfer from the f_i to $p = \prod f_i$

The "Laplace" approximation

- We will compare EP fixed-points to the "Laplace" approximation (LA):

The "Laplace" approximation

- We will compare EP fixed-points to the "Laplace" approximation (LA):
- Find the mode $x^{\star}$ of $p(x)$
- At $x^{\star}$, compute $\psi''(x^{\star})$

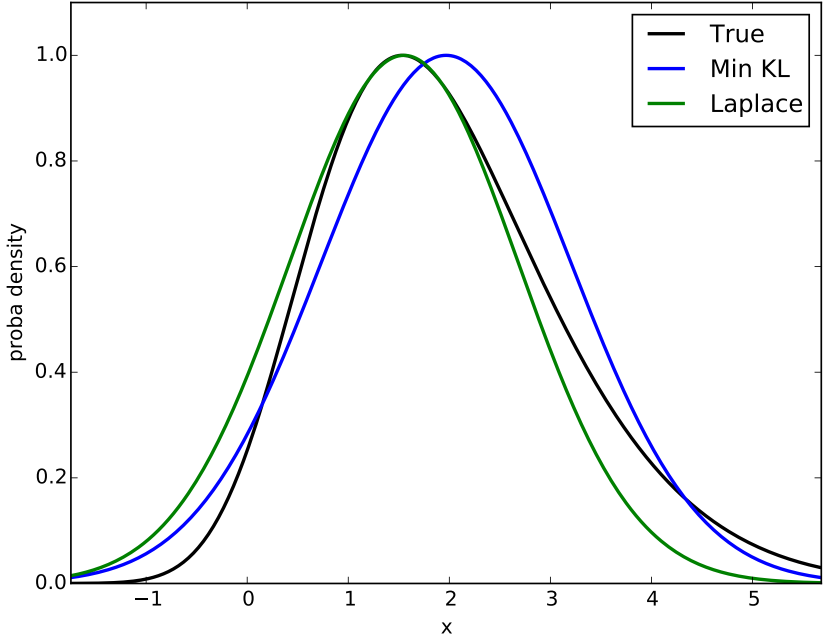
$$p(x) \approx \exp\left(-\psi''(x^{\star})\frac{(x-x^{\star})^2}{2}\right)$$

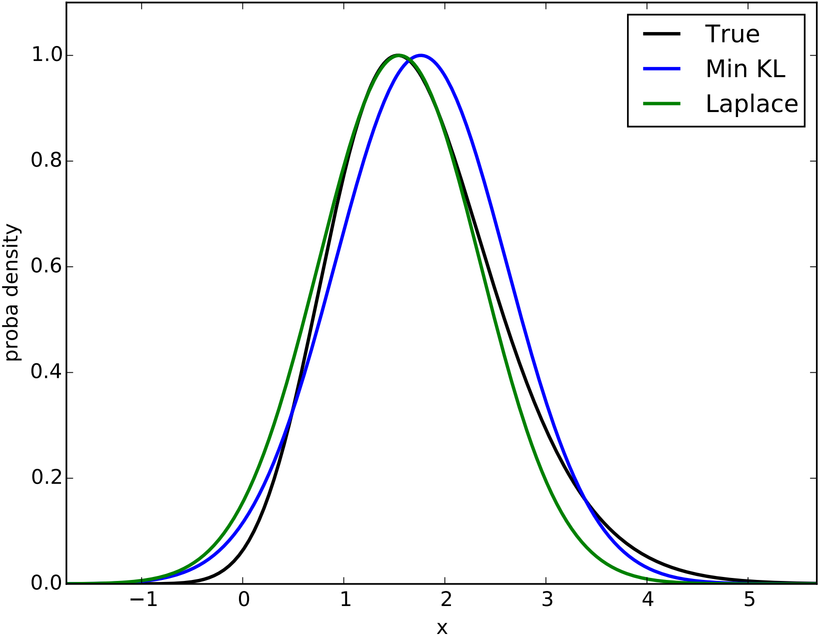
Why LA is good

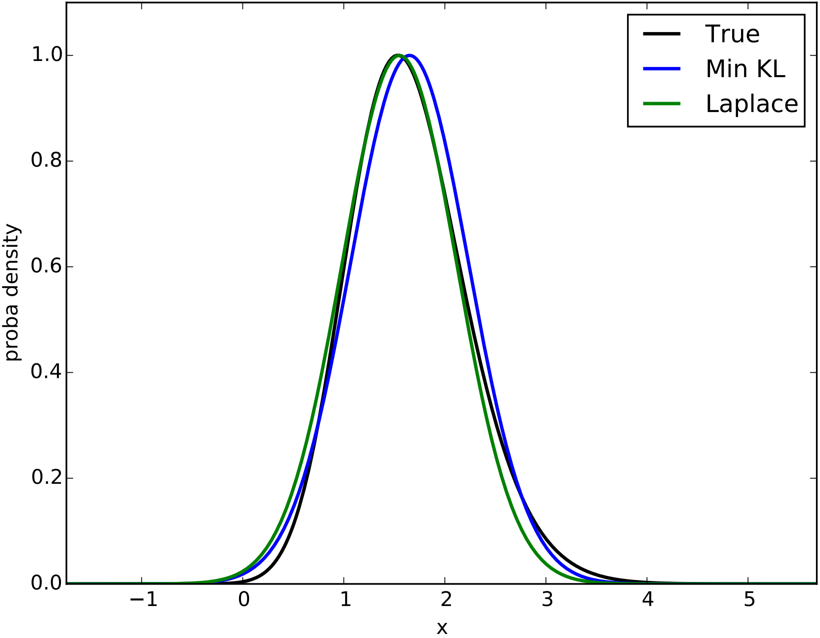
- The Bernstein-von Mises theorem justifies the LA:
  - In the large-data limit, $p_n(x) \rightarrow g_{LA}(x)$

Why LA is good

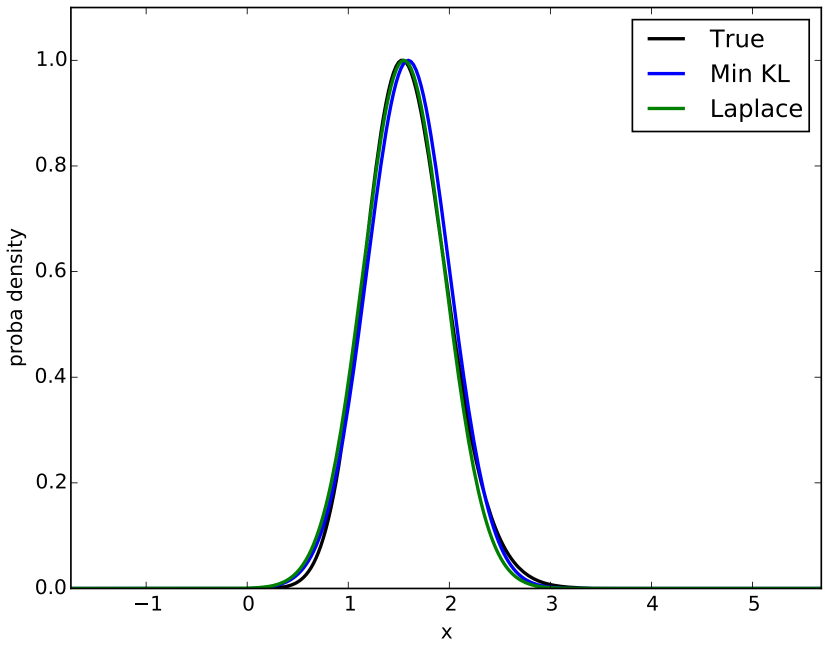
- The Bernstein-von Mises theorem justifies the LA:
  - In the large-data limit, $p_n(x) \to g_{LA}(x)$
- But that doesn't mean it's perfect:
  - it looks at point estimates
  - it ignores higher derivatives


True
Min KL
Laplace
proba density
x
1.0
0.8
0.6
0.4
0.2
0.0
−1
0
1
2
3
4
5


True
Min KL
Laplace
proba density
1.0
0.8
0.6
0.4
0.2
0.0
−1
0
1
2
3
4
5
x



True
Min KL
Laplace
proba density
x
1.0
0.8
0.6
0.4
0.2
0.0
−1
0
1
2
3
4
5


True
Min KL
Laplace
proba density
x
0.0
0.2
0.4
0.6
0.8
1.0
−1
0
1
2
3
4
5

Why LA is good

- We can derive the expression of the bias:

$$
\begin{aligned}
x^{\star} - \mu &= -\frac{\psi^{(3)}\left(x^{\star}\right)}{\psi''\left(x^{\star}\right)^{2}} + O\left(n^{-2}\right) \\
\psi''\left(x^{\star}\right) - v &= O\left(n^{-2}\right)
\end{aligned}
$$

Why LA is good

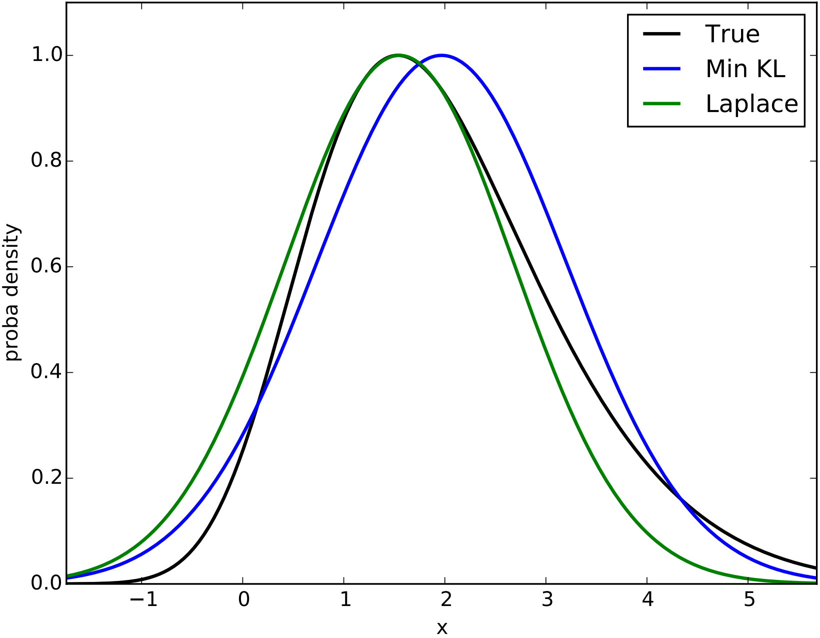
- We can derive the expression of the bias:

$$
\begin{aligned}
x^\star - \mu &= -\frac{\psi^{(3)}\left(x^\star\right)}{\psi''\left(x^\star\right)^2} + O\left(n^{-2}\right) \\
\psi''\left(x^\star\right) - v &= O\left(n^{-2}\right)
\end{aligned}
$$

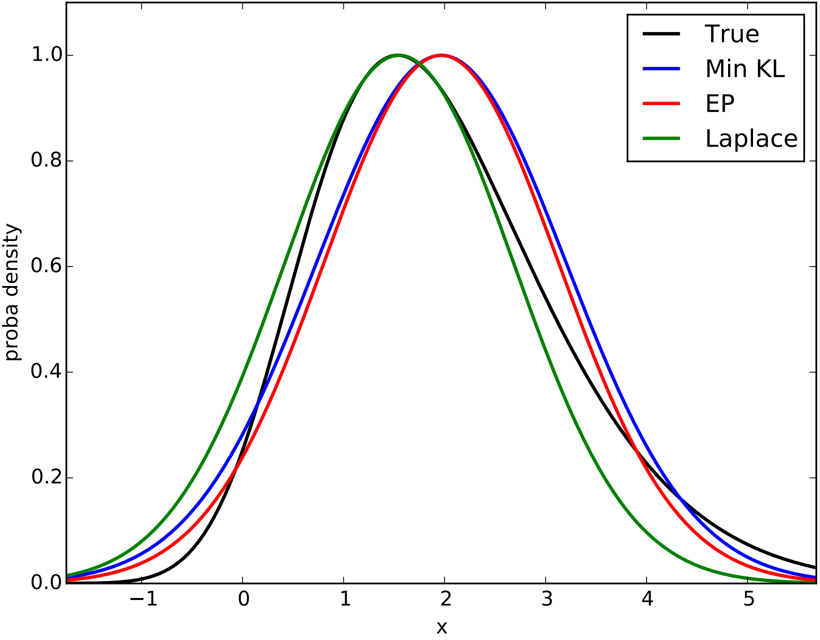
- Since LA misses the mean consistently, there is **room for improvement**
- If EP is able to always correct this miss, it will improve on LA

Why EP is better !!

- Consider an EP fixed-point $g(x) = \prod g_i(x) \approx p(x)$
- EP captures the $\psi^{(3)}(x^\star)$ deviation



True
Min KL
Laplace
proba density
x
0.0
0.2
0.4
0.6
0.8
1.0
−1
0
1
2
3
4
5


True
Min KL
EP
Laplace
proba density
x
0.0
0.2
0.4
0.6
0.8
1.0
−1
0
1
2
3
4
5

Why EP is better !!

- With a similar proof as for the LA result, we prove:

$$
\begin{aligned}
\mu_{EP} - \mu &= O\left(n^{-2}\right) \\
v_{EP} - v &= O\left(n^{-2}\right)
\end{aligned}
$$

Comparing LA and EP

Theorem (Quality of the LA and EP approximations)

- *LA:*

$$\begin{aligned} \mu - x^\star &= O\left(n^{-1}\right) \\ v - \left[\psi''\left(x^\star\right)\right]^{-1} &= O\left(n^{-2}\right) \end{aligned}$$

- *EP:*

$$\begin{aligned} \mu - \mu_{EP} &= O\left(n^{-2}\right) \\ v - v_{EP} &= O\left(n^{-2}\right) \end{aligned}$$

- *The first term of the error for the variance is slightly smaller for EP than for the LA*

The log-concavity assumption

- The strongly log-concave sites assumption is unrealistic
- However, simple log-concavity should be enough
- proof ?

Summary

- Both EP and LA give asymptotically correct approximations of $p(x)$
- But LA fails slightly on asymmetric distributions whereas EP doesn't

Summary

- Both EP and LA give asymptotically correct approximations of $p(x)$
- But LA fails slightly on asymmetric distributions whereas EP doesn't
- Important result for credible intervals from EP approximations
- **But** problematic assumptions

Contents

Understanding the EP iteration

- The EP iteration has **one complicated step**: the site-approximation update
  1. Compute hybrid distribution $h_i = f_i g_{-i}$
  2. Compute $E_{h_i}(x)$ and var_{h_i}
  3. Compute the Gaussian with same mean and variance: $\mathcal{N}\left(x|E_{h_i}(x); var_{h_i}\right)$
  4. update g_i

Understanding the EP iteration

- The EP iteration has **one complicated step**: the site-approximation update
    1. Compute hybrid distribution $h_i = f_i g_{-i}$
    2. Compute $E_{h_i}(x)$ and var_{h_i}
    3. Compute the Gaussian with same mean and variance: $\mathcal{N}\left(x | E_{h_i}(x); var_{h_i}\right)$
    4. update g_i
- This is the step we need to understand

Assumptions

- Much looser assumptions

Assumptions

- Much looser assumptions
- We bound the range of the second derivatives

$$\forall i,\ \max\left(\phi_i''\right) - \min\left(\phi_i''\right) \leq B$$

- Still uniform bound on the higher derivatives

$$\left|\phi_i^{(d)}(x)\right| \leq K_d$$

Assumptions

- Much looser assumptions
- We bound the range of the second derivatives

$$\forall i,\ \max\left(\phi_i^{''}\right) - \min\left(\phi_i^{''}\right) \leq B$$

- Still uniform bound on the higher derivatives

$$\left|\phi_i^{(d)}(x)\right| \leq K_d$$

- Applies to any GLM, can be extended so that B and K_d depend on n

Intuitive understanding

- Rewrite KL minization:

$$
\begin{aligned}
g_i &= [g_{-i}]^{-1} \operatorname{argmin}_g KL(h_i, g) \\
&\approx \operatorname{argmin}_g \int h_i \left[\log(f_i) - \log(g_i)\right]
\end{aligned}
$$

- h_i tells us where g_i needs to fit f_i

Intuitive understanding

- Rewrite KL minization:

$$
\begin{aligned}
g_i &= [g_{-i}]^{-1} \operatorname{argmin}_g KL(h_i, g) \\
&\approx \operatorname{argmin}_g \int h_i [\log(f_i) - \log(g_i)]
\end{aligned}
$$

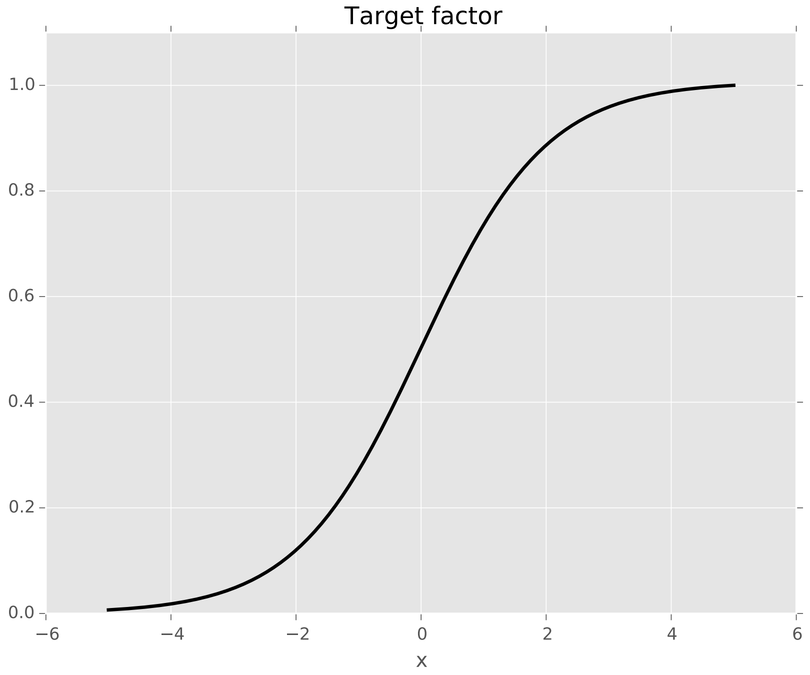
- h_i tells us where g_i needs to fit f_i
- If g_{-i} has very small variance (ie: β_{-i} is big):
  - g_{-i} is almost Dirac
  - $h_i \approx g_{-i}$ and is also Dirac

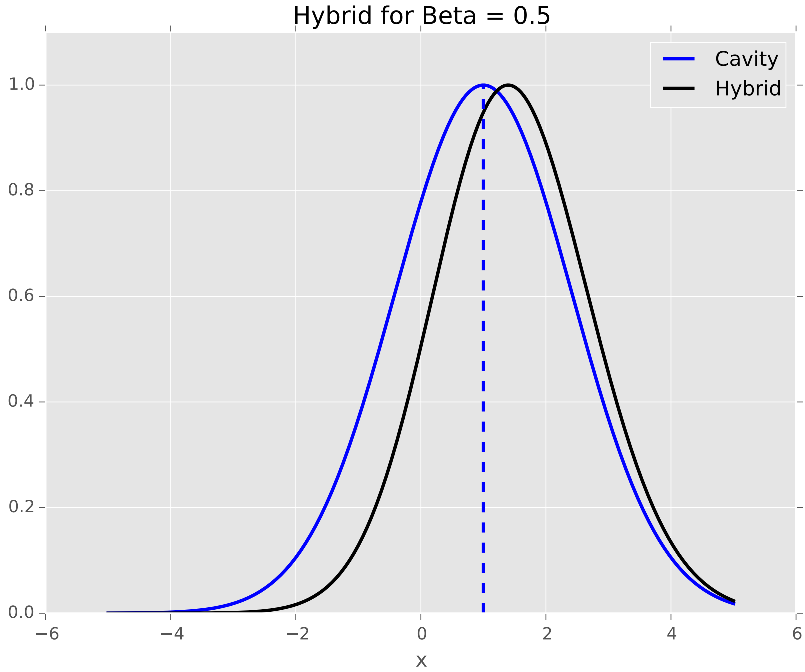
Intuitive understanding

- Rewrite KL minization:

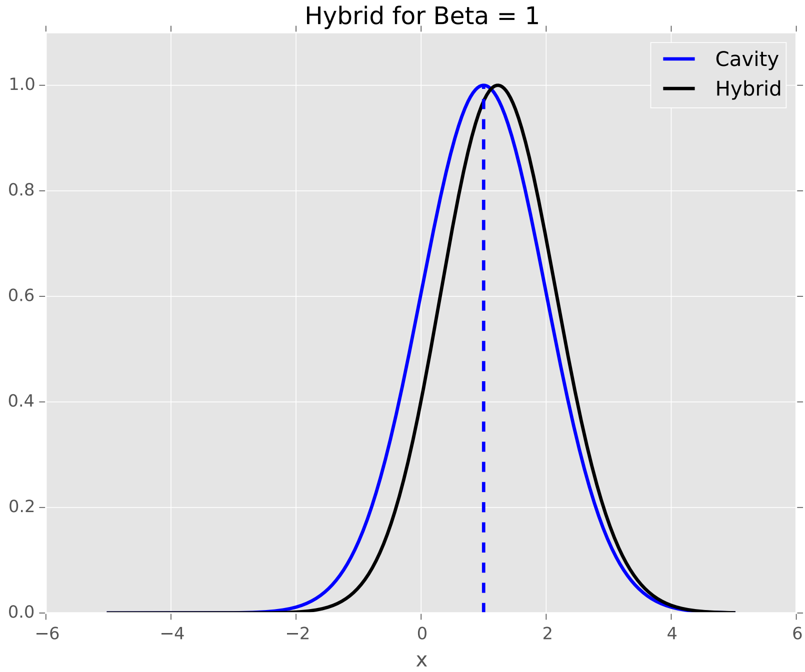
$$\begin{aligned} g_i &= [g_{-i}]^{-1} \operatorname{argmin}_g KL(h_i, g) \\ &\approx \operatorname{argmin}_g \int h_i [\log(f_i) - \log(g_i)] \end{aligned}$$

- h_i tells us where g_i needs to fit f_i
- If g_{-i} has very small variance (ie: β_{-i} is big):
    - g_{-i} is almost Dirac
    - $h_i \approx g_{-i}$ and is also Dirac
- The best approximation is the Taylor expansion of $\log(f_i)$.

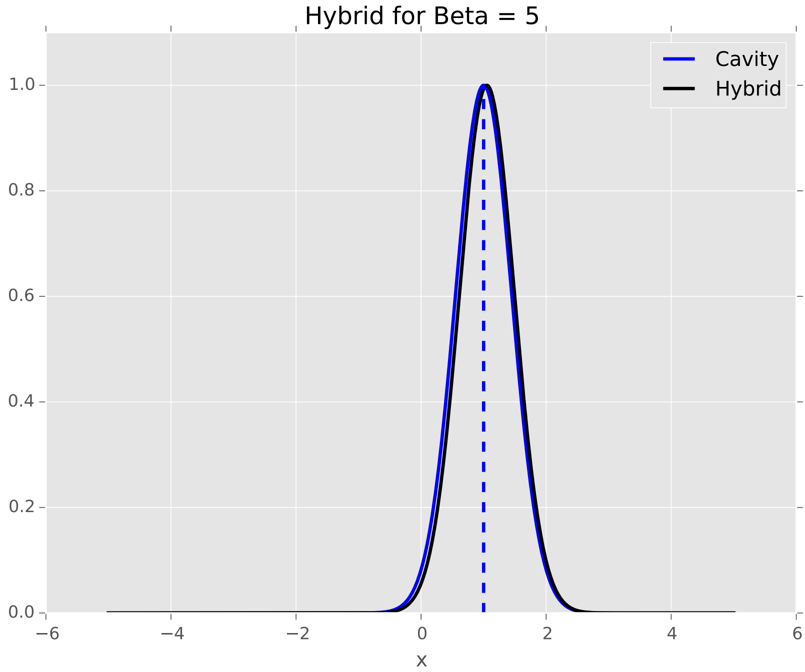

Target factor
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
x



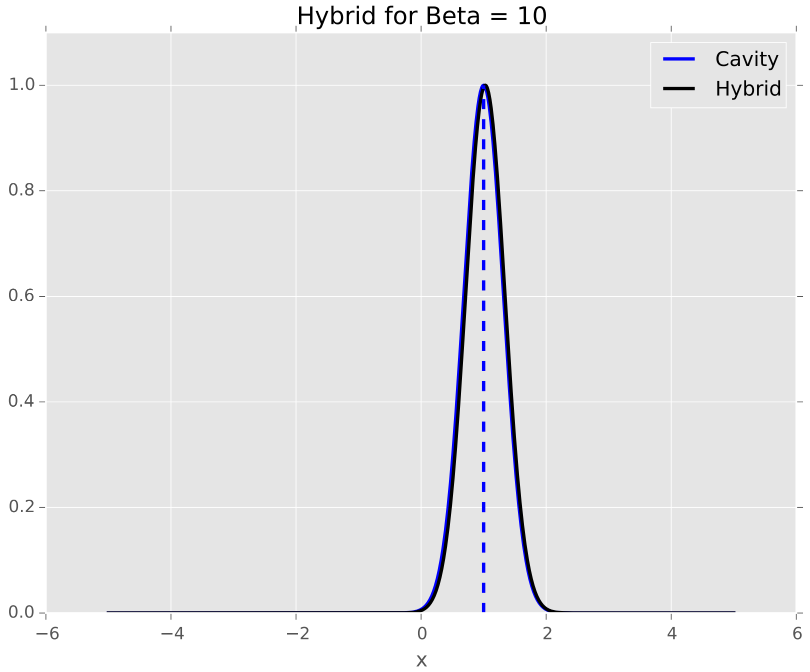
Hybrid for Beta = 0.5
Cavity
Hybrid
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
x



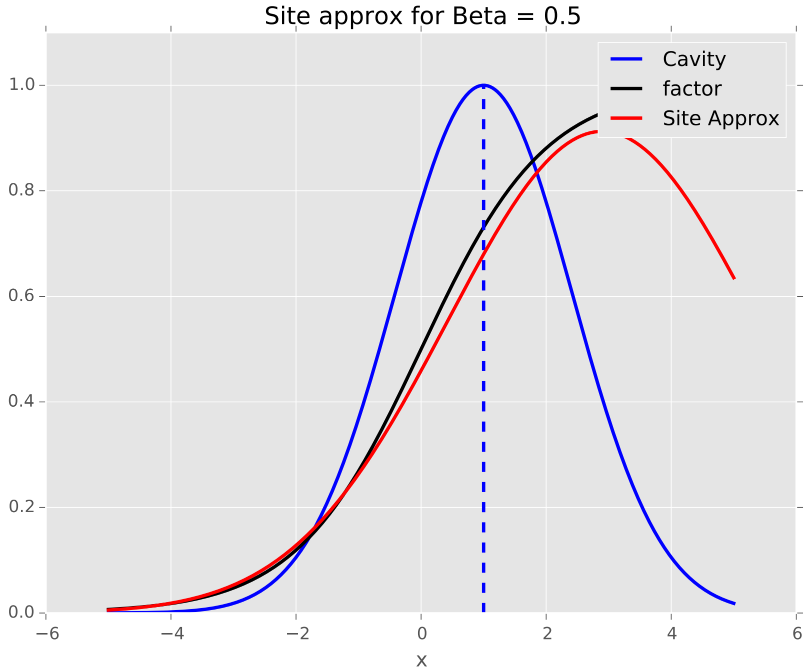
Hybrid for Beta = 1
Cavity
Hybrid
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
x

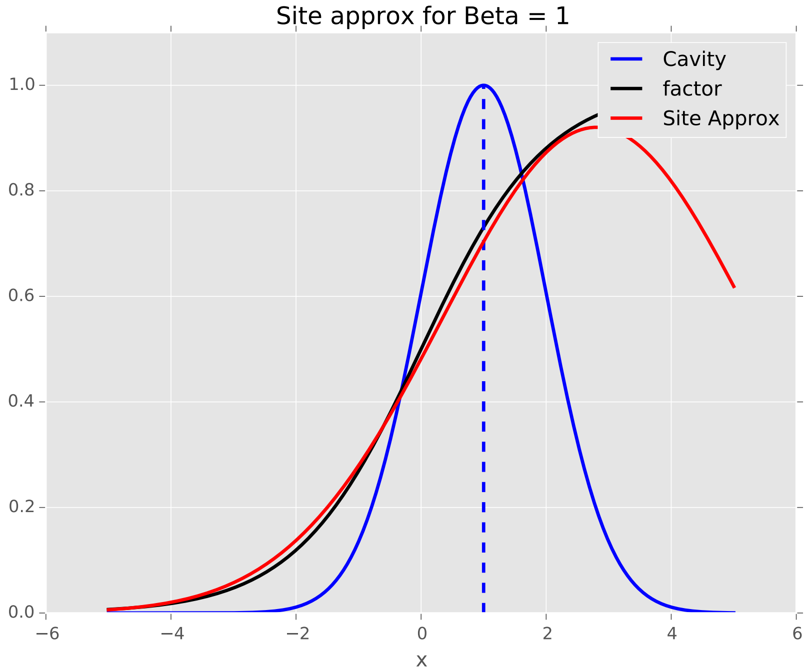


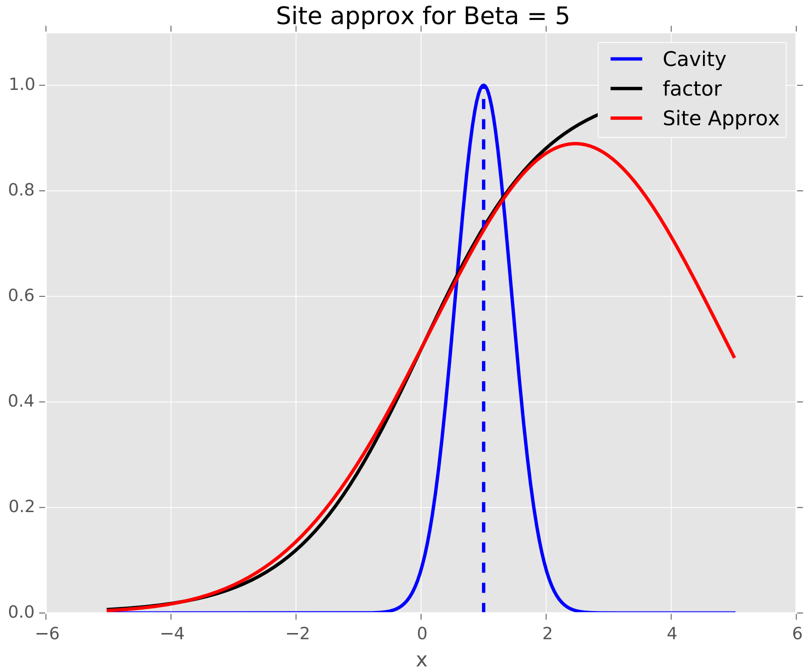
Hybrid for Beta = 5
Cavity
Hybrid
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
x

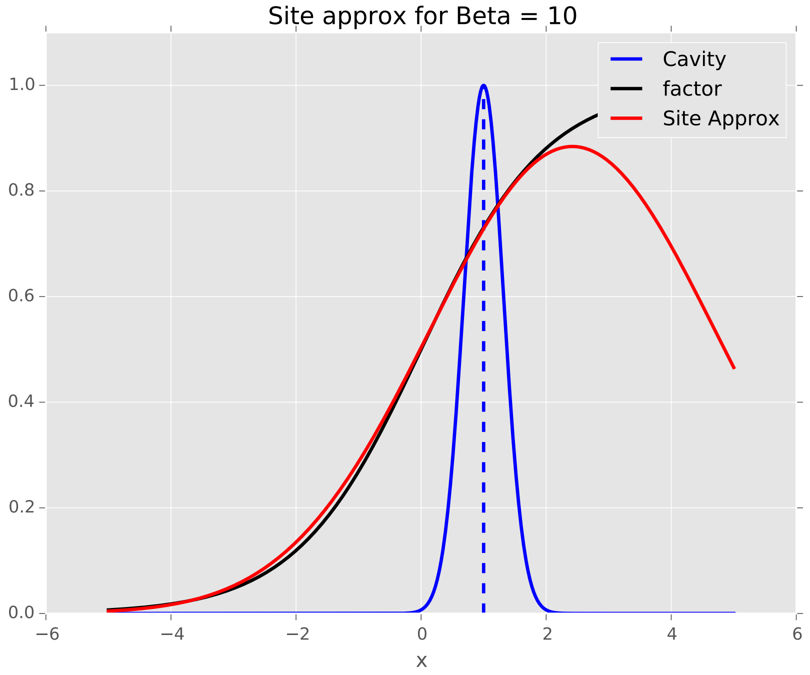


Hybrid for Beta = 10
Cavity
Hybrid
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
x


Site approx for Beta = 0.5
Cavity
factor
Site Approx
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
x


Site approx for Beta = 1
Cavity
factor
Site Approx
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
x


Site approx for Beta = 5
Cavity
factor
Site Approx
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
x



Site approx for Beta = 10
Cavity
factor
Site Approx
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
x

Limit behavior of the approximation

Theorem (Limit behavior of the factor approximation)

When $\beta_{-i} \rightarrow \infty$, the limit of the EP approximation of $f_i \propto \exp\left(-\phi_i\right)$ is:

$$g_i^{\infty} \propto \exp\left(-\phi_i'\left(\mu_{-i}\right)\left(x-\mu_{-i}\right)-\frac{\phi''\left(\mu_{-i}\right)}{2}\left(x-\mu_{-i}\right)^2\right)$$

- **Many important details in the error term:** non-uniform convergence in μ_{-i}, ...

Limit behavior of EP iterations

- The limit behavior of parallel EP = the sum of the limit behaviors

Theorem (Limit behavior of EP)

When $\beta_{-i} \to \infty$ for all i, the limit of the next EP approximation of $p(x) \propto \exp(-\psi(x))$ is:

$$q_{t+1}^{\infty} \propto \exp\left(-\psi'(\mu_t)(x-\mu_t) - \frac{\psi''(\mu_t)}{2}(x-\mu_t)^2\right)$$

Limit behavior of EP iterations

- The limit behavior of parallel EP = the sum of the limit behaviors

Theorem (Limit behavior of EP)

When $\beta_{-i} \to \infty$ for all i, the limit of the next EP approximation of $p(x) \propto \exp(-\psi(x))$ is:

$$q_{t+1}^{\infty} \propto \exp\left(-\psi'(\mu_t)(x-\mu_t) - \frac{\psi''(\mu_t)}{2}(x-\mu_t)^2\right)$$

- Did you recognize Newton's algorithm ?

Newton's algorithm

- Objective of Newton's algorithm:
  - find the mode $x^\star$ (in order to compute the LA)

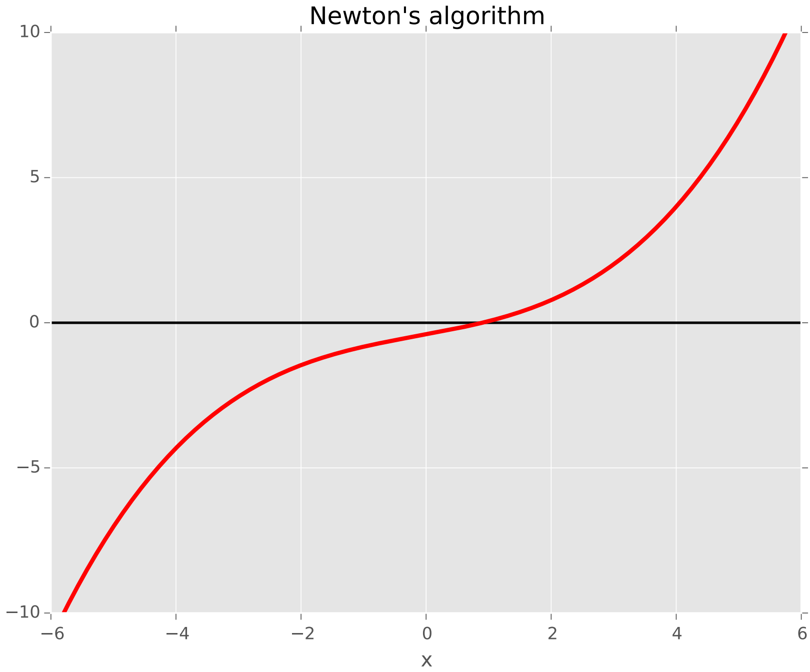
$$\psi^{'}(x^\star) = 0$$

Newton's algorithm

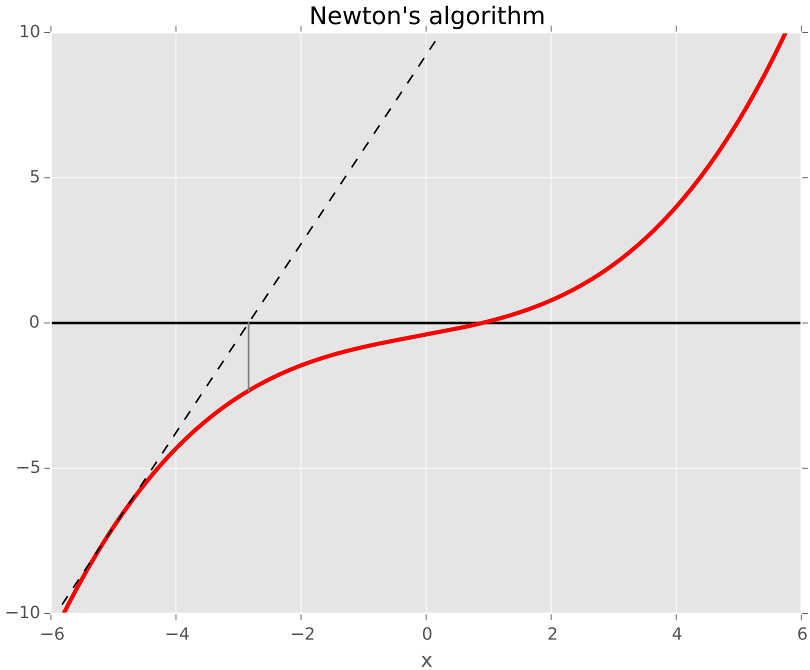
- Objective of Newton's algorithm:
  - find the mode $x^\star$ (in order to compute the LA)

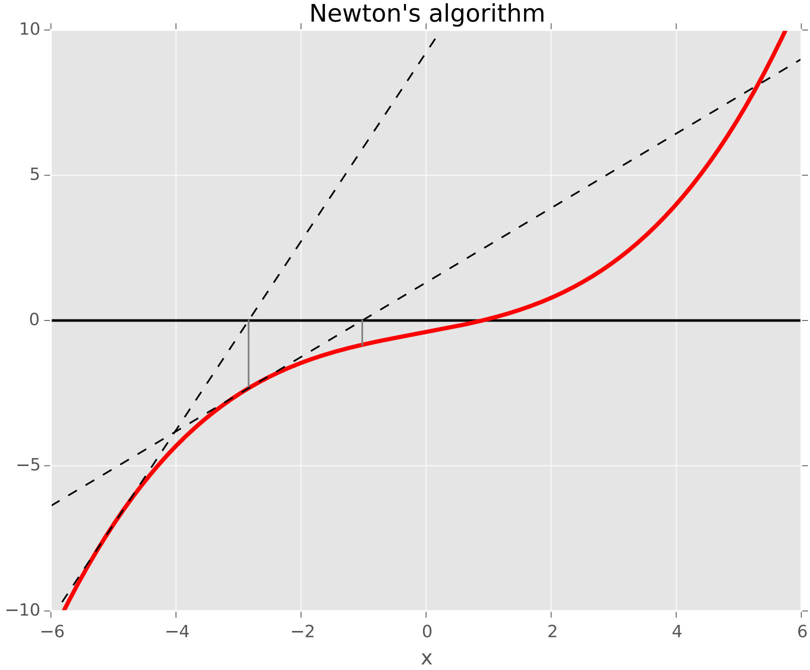
$$\psi^{'}(x^\star) = 0$$

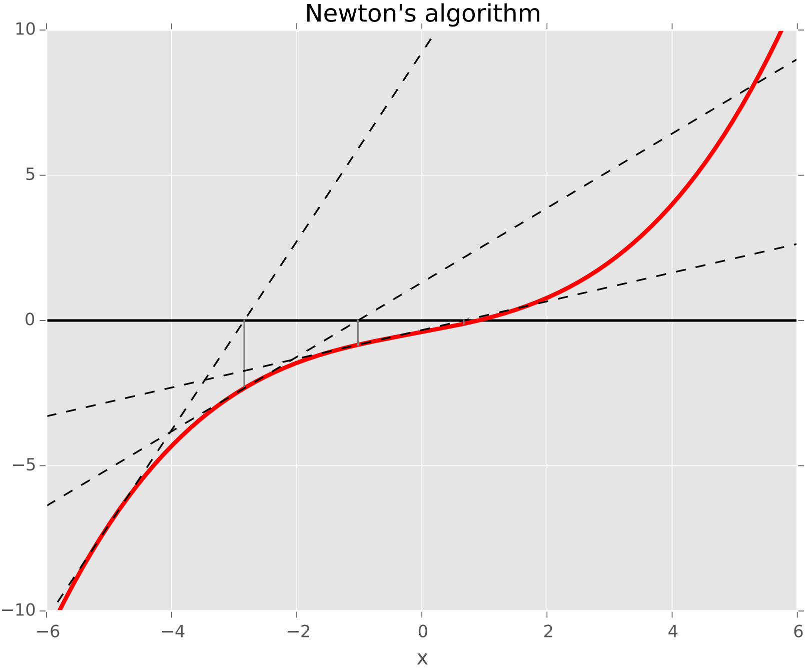
- Iterative algorithm:
  - At μ_t, compute the tangent to $\psi^{'}$
  - Solve tangent = 0: this is μ_{t+1}

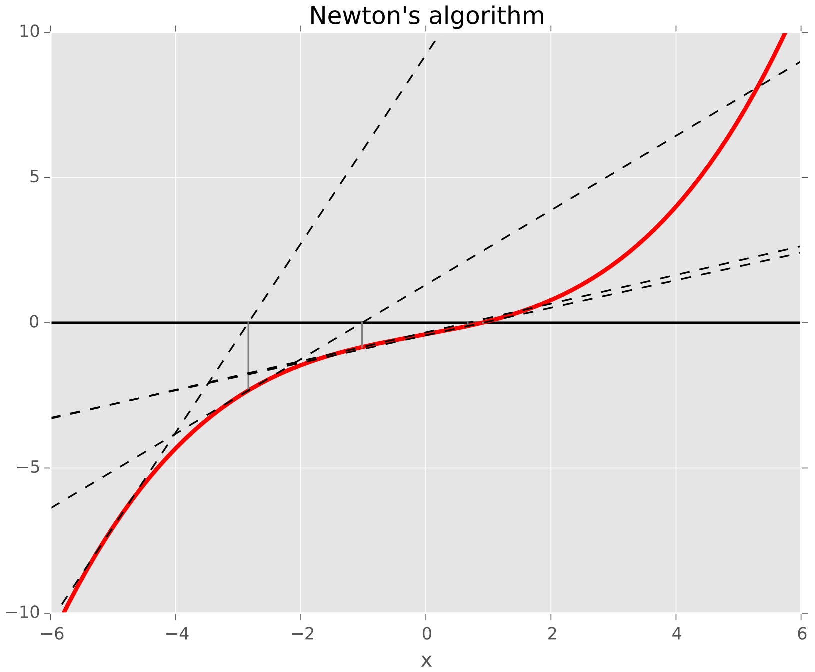


Newton's algorithm
10
5
0
−5
−10
−6
−4
−2
0
2
4
6
x


Newton's algorithm
10
5
0
−5
−10
−6
−4
−2
0
2
4
6
x


Newton's algorithm
10
5
0
−5
−10
−6
−4
−2
0
2
4
6
x


Newton's algorithm
10
5
0
−5
−10
−6
−4
−2
0
2
4
6
x


Newton's algorithm
10
5
0
−5
−10
−6
−4
−2
0
2
4
6
x

Newton and EP

- At each step, we approximate

$$\begin{aligned}\psi' &\approx ax + b\\ \psi &\approx a\frac{x^2}{2} + b\\ p(x) &\approx \exp\left(-a\frac{x^2}{2} - b\right)\end{aligned}$$

- Newton is iterating over Gaussian approximations of $p(x)$!!!

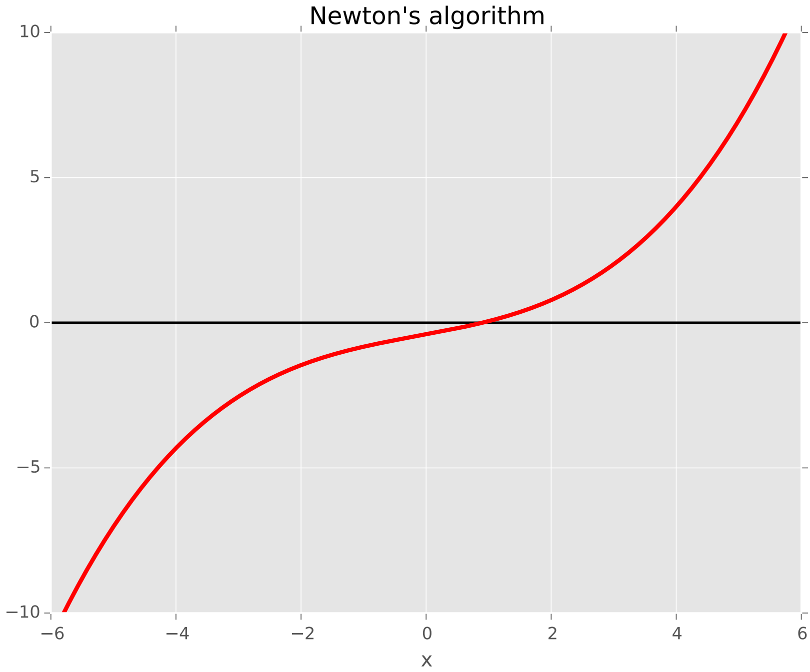
Newton and EP

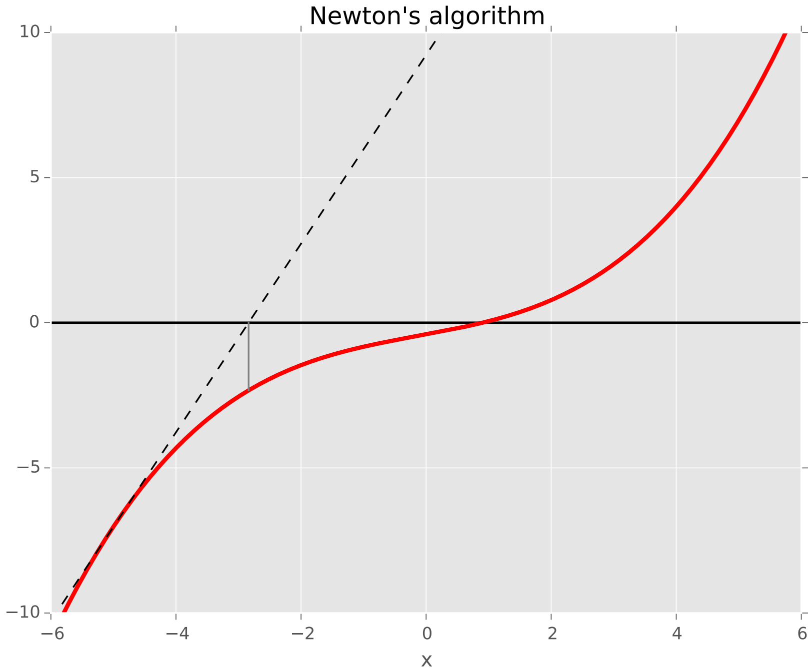
- At each step, we approximate

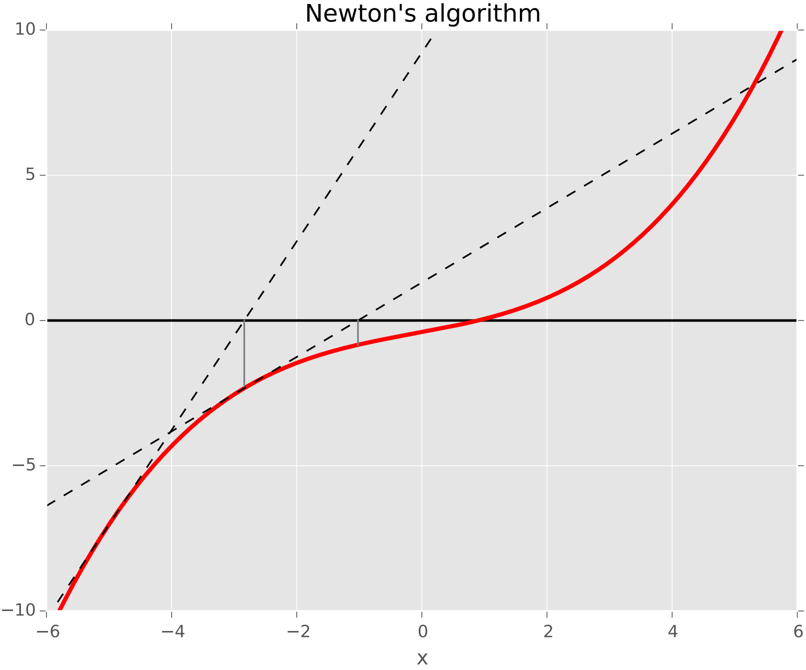
$$\begin{aligned}\psi' &\approx ax + b\\ \psi &\approx a\frac{x^2}{2} + b\\ p(x) &\approx \exp\left(-a\frac{x^2}{2} - b\right)\end{aligned}$$

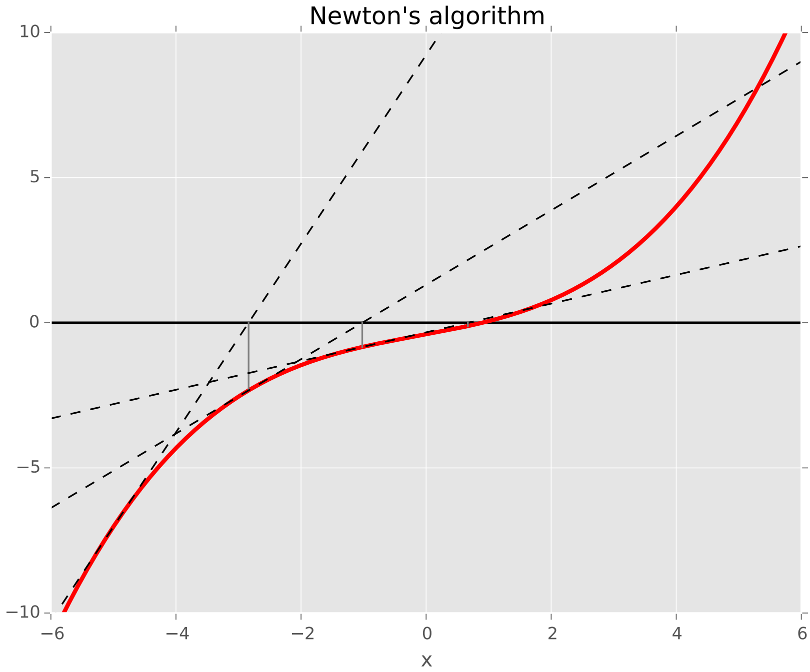
- Newton is iterating over Gaussian approximations of $p(x)$!!!
- The EP limit approximation is:

$$p(x) \approx \exp\left(-\psi'(\mu_t)(x-\mu_t) - \frac{\psi''(\mu_t)}{2}(x-\mu_t)^2\right)$$

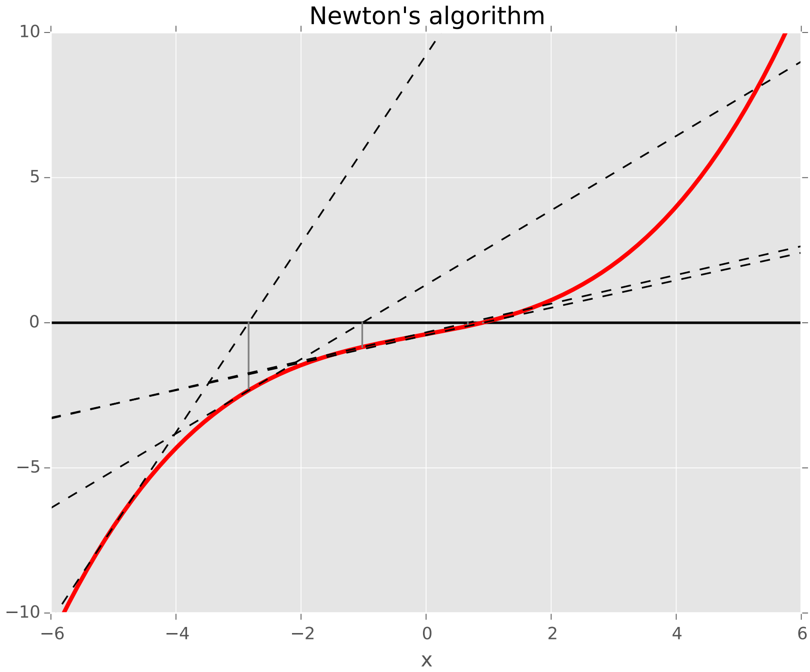

Newton's algorithm
10
5
0
−5
−10
−6
−4
−2
0
2
4
6
x

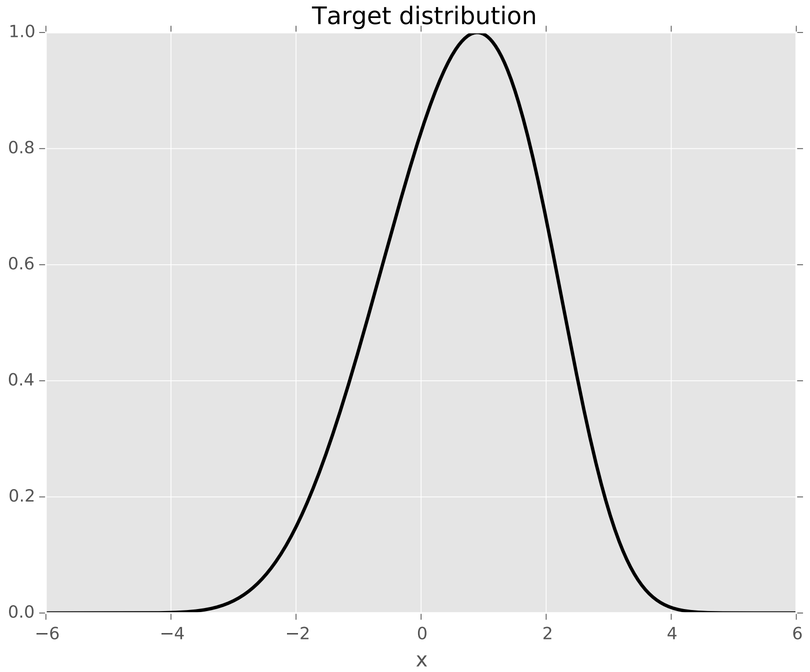

Newton's algorithm
10
5
0
−5
−10
−6
−4
−2
0
2
4
6
x

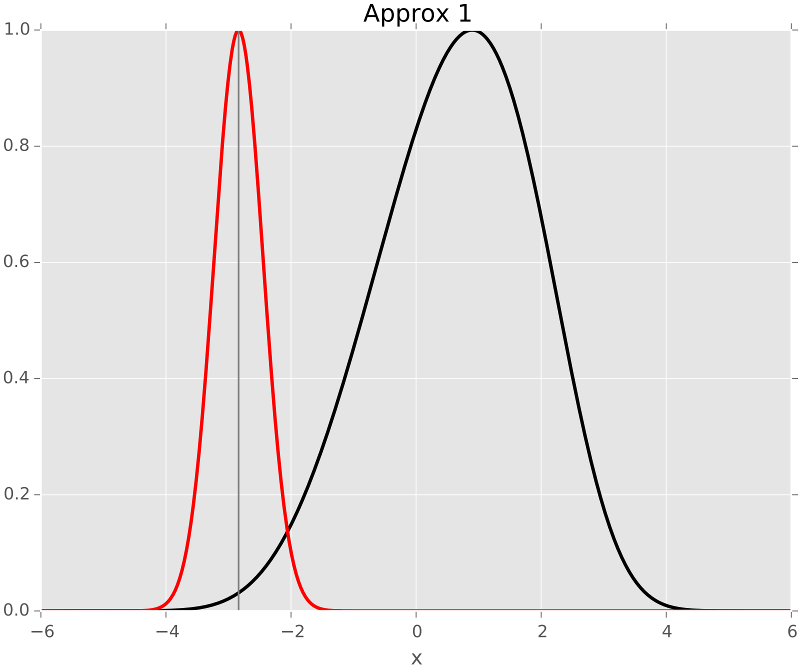

Newton's algorithm
10
5
0
−5
−10
−6
−4
−2
0
2
4
6
x

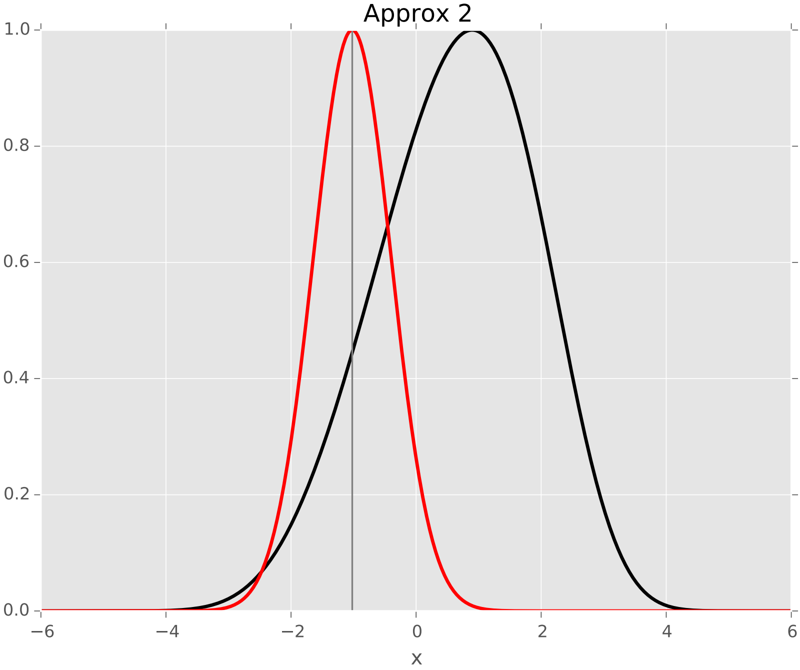


Newton's algorithm
10
5
0
−5
−10
−6
−4
−2
0
2
4
6
x

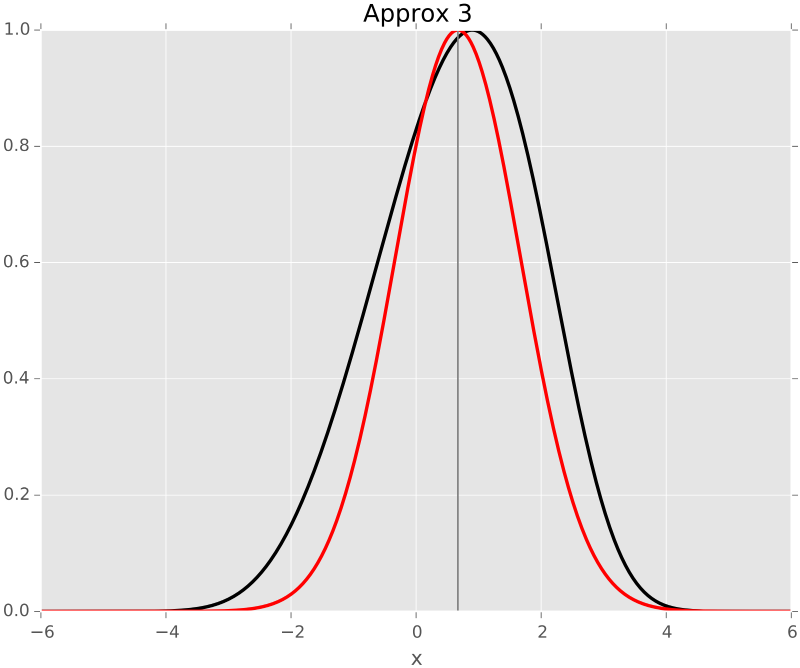

Newton's algorithm
10
5
0
−5
−10
−6
−4
−2
0
2
4
6
x


Target distribution
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
x

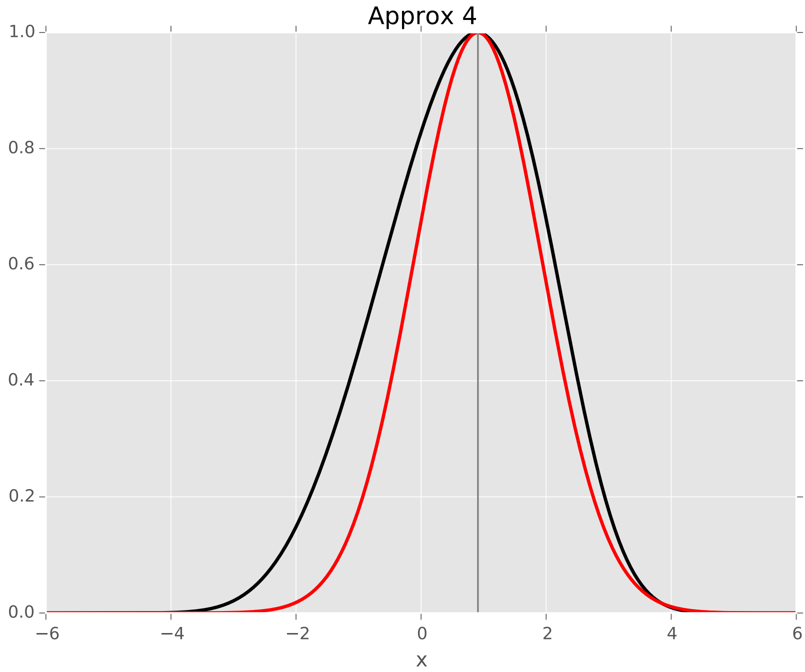

Approx 1
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
x



Approx 2
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
x



Approx 3
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
x



Approx 4
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
x

Reaching the high-precision limit

- High-precision limit $\neq$ large-data limit

Reaching the high-precision limit

- High-precision limit $\neq$ large-data limit
- Approximation quality result:

$$\begin{aligned} \mathrm{var}_{p_n} &\propto n^{-1} \\ v_{EP} &\approx \mathrm{var}_{p_n} \end{aligned}$$

- Around fixed-points (where it matters), EP is close to Newton

Reaching the high-precision limit

- High-precision limit $\neq$ large-data limit
- Approximation quality result:

$$\begin{aligned} \mathrm{var}_{p_n} &\propto n^{-1} \\ v_{EP} &\approx \mathrm{var}_{p_n} \end{aligned}$$

- Around fixed-points (where it matters), EP is close to Newton
- We can derive other links
- We can always check

Intuitions from *EP* $\approx$ *Newton*

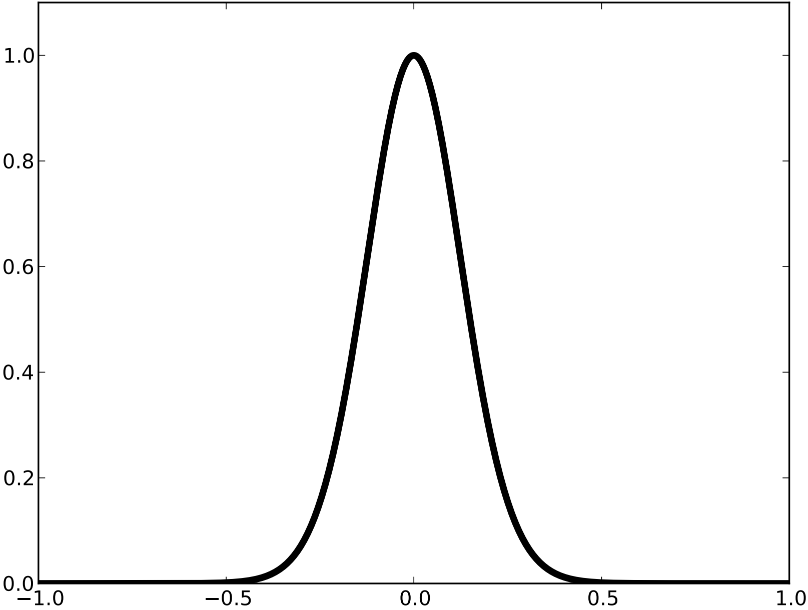
- Intuitively, if EP behaves like Newton in some limit, even away from that limit, it should have similar properties

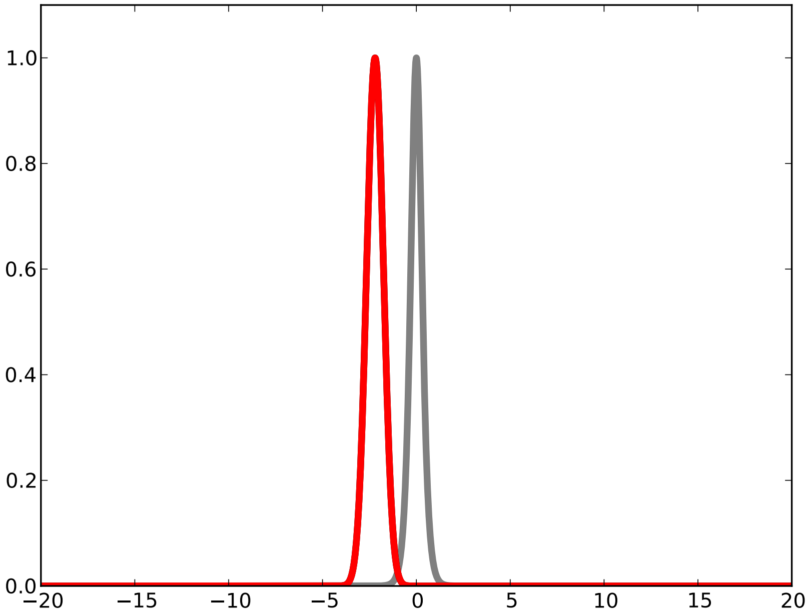
Intuitions from *EP* $\approx$ *Newton*

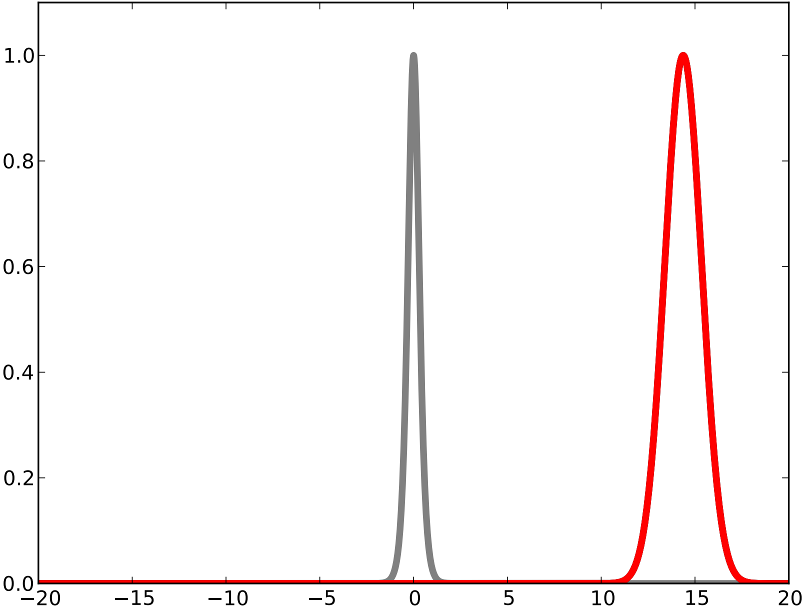
- Intuitively, if EP behaves like Newton in some limit, even away from that limit, it should have similar properties
- Newton is very well-known
    - It converges extremely fast once it gets close to its fixed-point
    - **But** it can fail to converge
    - We have to supplement it with line-search algorithms
    - If we don't, it can "bounce" around its fixed-point

Intuitions from $EP \approx Newton$

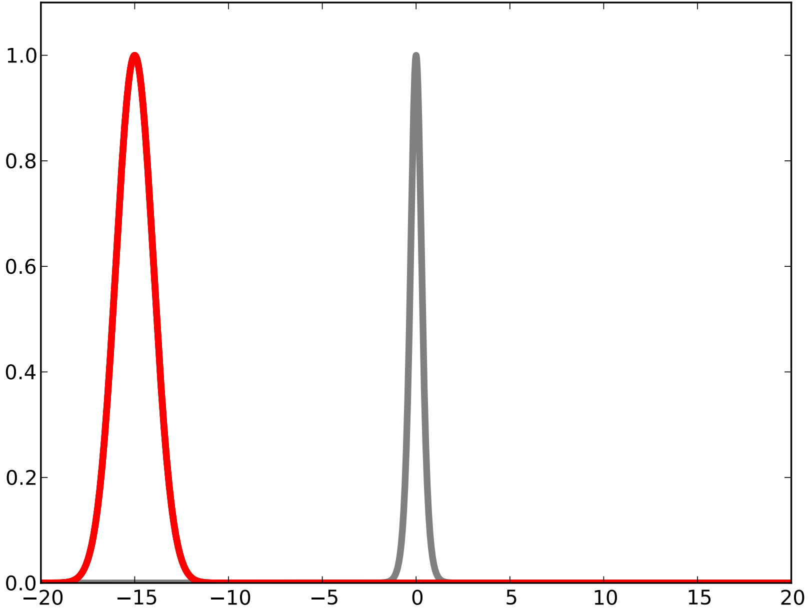
- Intuitively, if EP behaves like Newton in some limit, even away from that limit, it should have similar properties
- Newton is very well-known
    - It converges extremely fast once it gets close to its fixed-point
    - **But** it can fail to converge
    - We have to supplement it with line-search algorithms
    - If we don't, it can "bounce" around its fixed-point
- EP probably has similar properties !

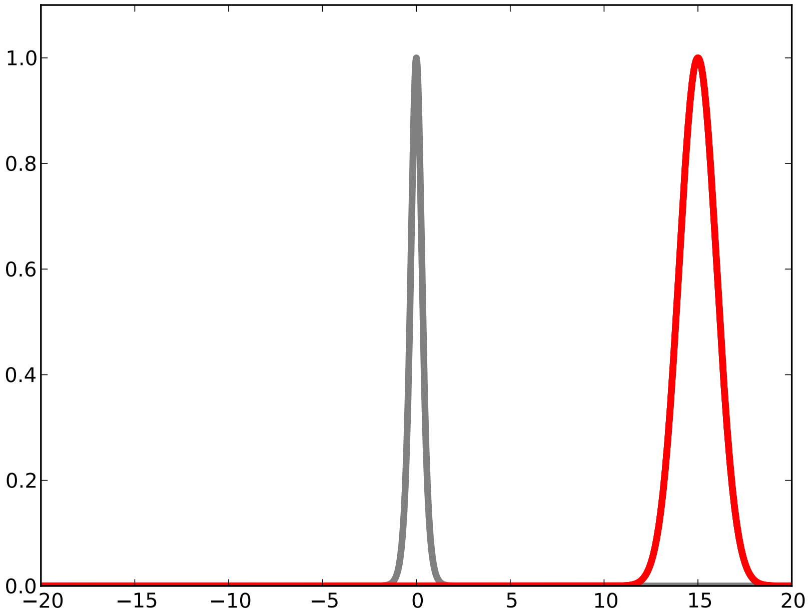

1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0


1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20

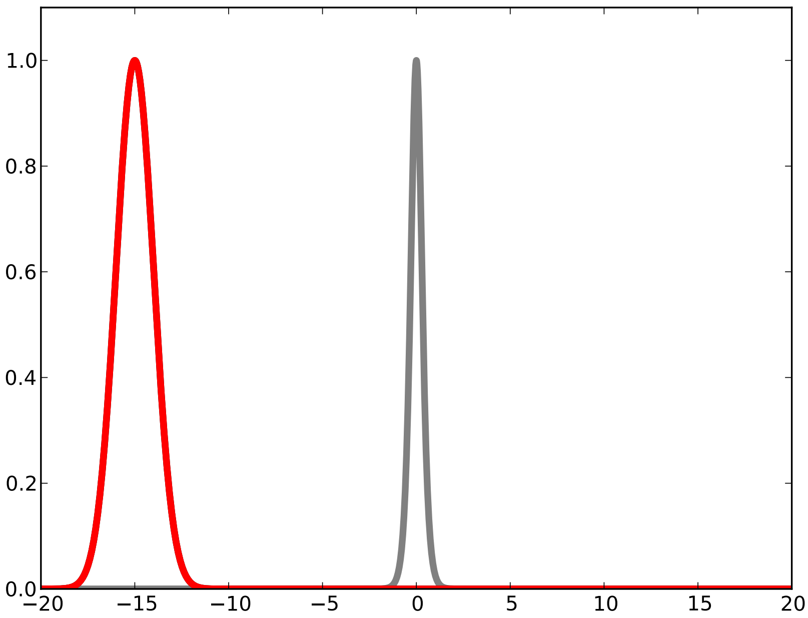


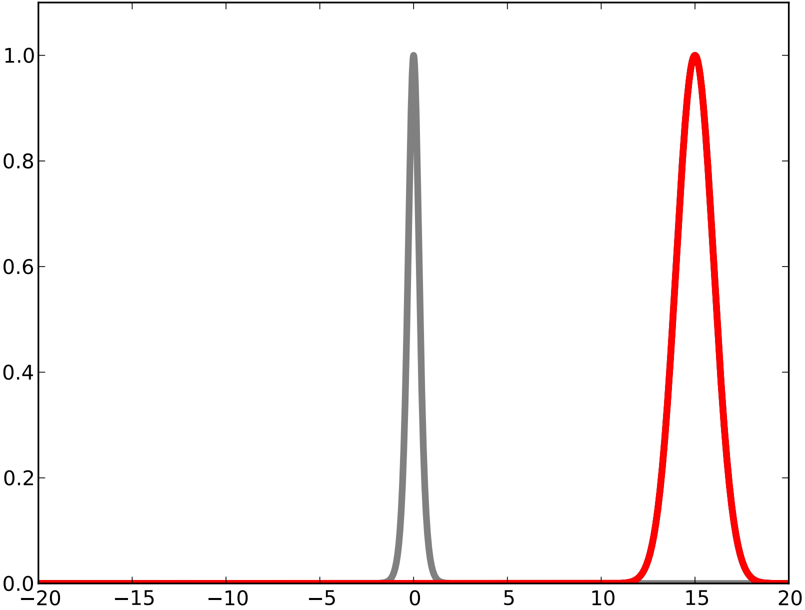
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20

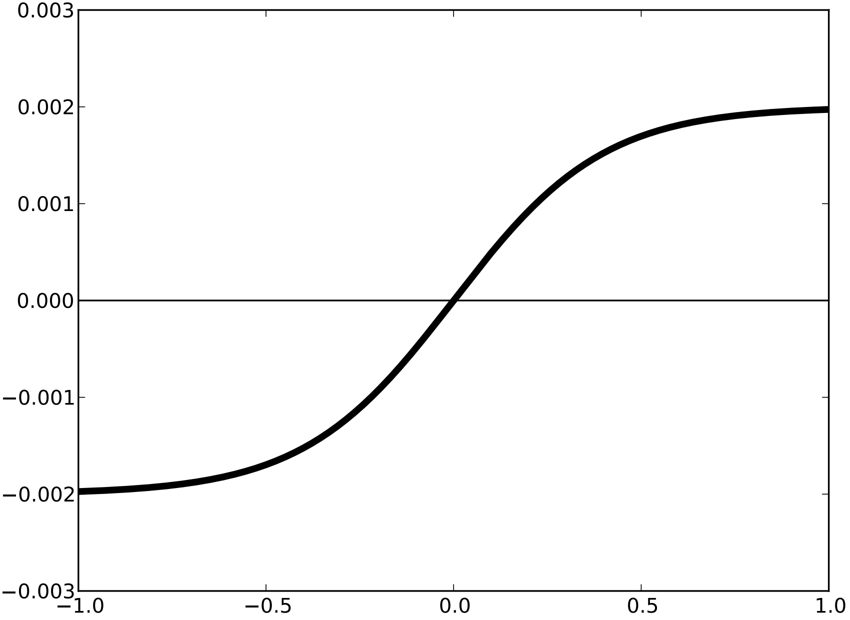

1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20



1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20


1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20


1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20


0.003
0.002
0.001
0.000
−0.001
−0.002
−0.003
−1.0
−0.5
0.0
0.5
1.0

Intuitions from *EP* $\approx$ *Newton*

- On a multi-modal $p(x)$, Newton has multiple fixed-points

Intuitions from *EP* $\approx$ *Newton*

- On a multi-modal $p(x)$, Newton has multiple fixed-points
- We can prove that sufficiently peaked modes have an EP fixed-point
  - EP can be "captured" by a mode and miss most of the probability mass
  - Avoid multi-modal distributions like the plague

Summary of our results

- EP is a better approximation than LA (with some caveats)

Summary of our results

- EP is a better approximation than LA (with some caveats)
- EP behaves like Newton's algorithm in the high-precision limit
- The high-precision limit is reached in the large-data regime

Summary of our results

- EP is a better approximation than LA (with some caveats)
- EP behaves like Newton's algorithm in the high-precision limit
- The high-precision limit is reached in the large-data regime
- This sheds new light on the dynamical behavior of EP:
  - It can bounce around it's fixed-point
  - We might need to supplement EP with line-search algorithms
  - EP can be captured by modes

References

- Our work:
  - "Bounding errors of Expectation-Propagation", Dehaene, Barthelmé, 2015, NIPS
  - "Expectation Propagation is Newton-like in the large-data limit", Dehaene, Barthelmé, 2015, In review
- Further references:
  - "Birth" of EP: Minka, 2001, UAI
  - Best explanation: Seeger, 2008, Berkely course notes
  - *EP as a way of life*: Gelman, Vehtari, Jylanki, Robert, Chopin, Cunningham