Expectation Propagation: Factorization and Entropy Approximation

John P. Cunningham

Department of Statistics
Columbia University

with Alp Kuckelbir, Philipp Hennig, Simon Lacoste-Julien

22 May, 2015

Outline

Outline

Preliminaries

EP for normal integrals

EP from an entropy perspective

Ideas for entropy improvements

Summary

Numerical Integration

Approximate inference is a problem of numerical integration.

Notation:

- prior: $p_0(x) = \frac{1}{Z_0} \exp\left\{\theta^\top \phi(x)\right\}$, for $x \in \mathbb{R}^d$
- likelihood terms: $t_i(x) = p(y_i|x) = \exp\left\{\bar{\theta}_i^\top \bar{\phi}_i(x)\right\}$, for $i = 1, ..., n$.
- posterior: $p(x|y_1, ..., y_n) = \frac{1}{Z} p_0(x) \prod_i t_i(x)$
- key object of interest:

$$Z = \int p_0(x) \prod_{i=1}^{n} t_i(x) dx = \int \exp\left\{\theta^\top \phi(x) + \sum_{i=1}^{n} \bar{\theta}_i^\top \bar{\phi}_i(x)\right\} dx$$

- (note: an intractable exponential family)

Expectation Propagation (loosely)

- Approximates 0th, 1st, and 2nd moments of an intractable family.
- Typical presentation: minimize $KL(p||q)$ for some tractable family q:

$$q(x) = \frac{1}{Z_{EP}} \exp\left\{ \left(\theta + \sum_{j=1}^{n} \lambda_j \right)^{\top} \phi(x) \right\} = \arg\min_{q \in \mathcal{N}} KL(p||q)$$

- But this is again intractable. Instead, iterate one term at a time:

$$\tilde{t}_i^{new} = \arg\min_{\tilde{t}_i \in \mathcal{N}} KL\left(q^{\setminus i} t_i || q^{\setminus i} \tilde{t}_i \right)$$

where $q^{\setminus i} t_i \propto \exp\left\{ \left(\theta + \sum_{j \neq i}^{n} \lambda_i \right)^{\top} \phi(x) + \bar{\theta}_i^{\top} \bar{\phi}(x)_i \right\}$.

- ... and hope it all works out.

Typical EP Example: Bayesian Probit Regression

- $p_0(x) = \mathcal{N}(x; m, K)$ is the prior on *weights* $x \in \mathbb{R}^d$.
- $t_i(x) = p(y_i|x) = \Phi\left(x^\top(c_i y_i)\right)$, $i = 1, ..., n$. (normal cdf)

- clarification:
  - x are the shape parameters of the hyperplane
  - $c_i \in \mathbb{R}^d$ are the inputs to the probit regression, the query points along the latent hyperplane defined by x.

- given data $\mathcal{D} = \{c_i, y_i\}_{i=1,\dots,n}$:
  - inference: $p(x|\mathcal{D})$
  - model selection: $\arg\max_{m,K} p(\mathcal{D})$

- Critical points for GP approx workshop:
  - GP classification is a special case
  - there are n rank-one factors, and $n \neq d$
  - EP updates are still typical unidmensional normal-probit integrals

Some relevant literature

- EP works very well for GP Classification

[Kuss and Rasmussen (2005) JMLR]

[Rasmussen and Williams (2006) MIT Press]

- EP can be improved with perturbative corrections

[Opper, Paquet, Winther (2009) NIPS]

[Paquet, Opper, Winther (2009) JMLR]

[Opper, Paquet, Winther (2013) JMLR]

- EP with step functions can be accurate...

[Cunningham, Hennig, Lacoste-Julien (2011) arXiv]

[Opper, Paquet, Winther (2013) JMLR]

- EP with $n \gg d$ is worth considering (and can be correct?)

[Dehaene and Barthelme (2015) arXiv]

[Gelman, Vehtari, et al (2014) arXiv]

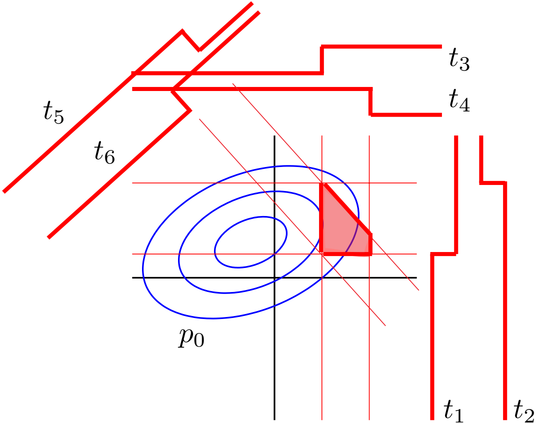
[Xu et al (2014) NIPS]

[Hernandez-Lobato and Hernandez-Lobato (today)]

Outline

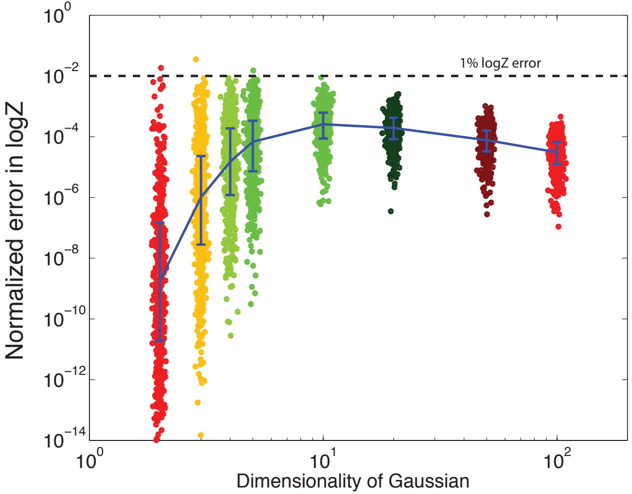
Limit of Bayesian Probit Regression

- $t_i(x) = p(y_i|x) = \Phi\left(x^\top(c_i y_i)\right)$, $i = 1, ..., n$. (standard normal cdf)
- Now let $||c_i|| \to \infty$
  - t_i are Heaviside step functions oriented in the the direction of $y_i c_i$
  - $p(\mathcal{D}) = \int p_0(x) \prod_i t_i(x) dx = \int_{\mathcal{A}} p_0(x) dx$
  - where $\mathcal{A}$ is the polyhedron defined by the factors t_i.



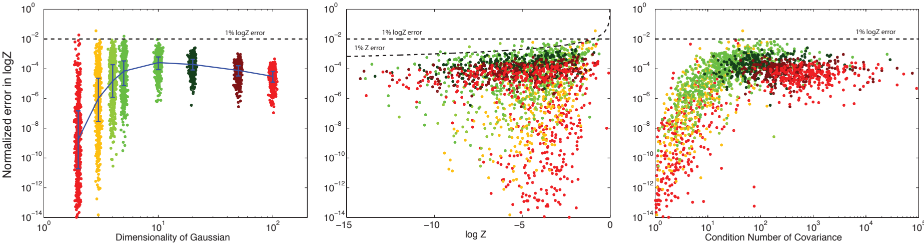
Empirical Performance

- EP works generally quite well for these integrals $Z = \int_{\mathcal{A}} p_0(x)dx$.
- Note: hyperrectangular regions, $n = d$.



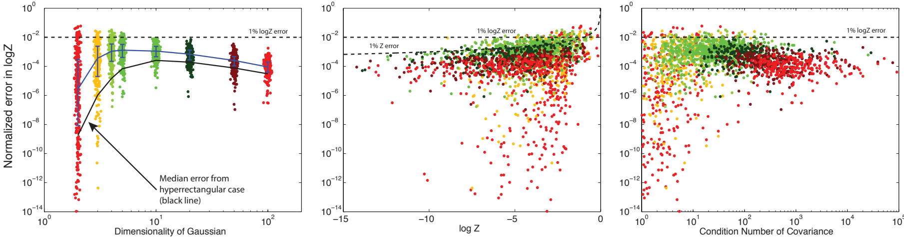
Empirical Performance

- EP works generally quite well for these integrals $Z = \int_{\mathcal{A}} p_0(x)dx$.
- Note: hyperrectangular regions, $n = d$.



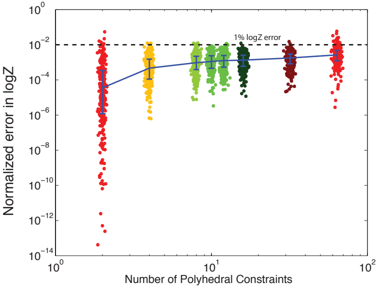
Empirical Performance

- ...EP still works well for these integrals $Z = \int_{\mathcal{A}} p_0(x)dx$.
- Note: polyhedral regions, $n \neq d$.



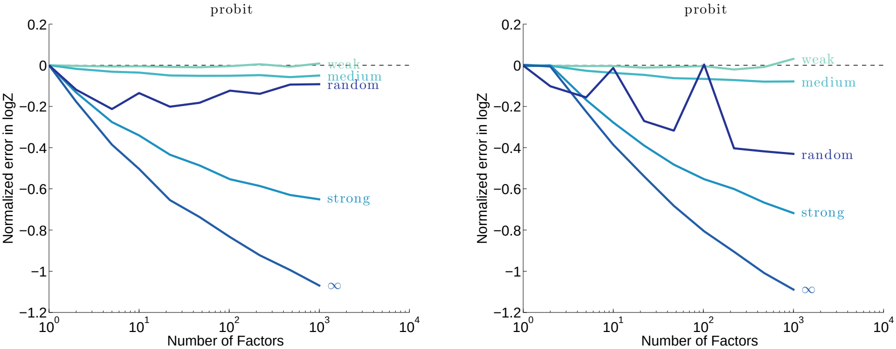
Empirical Performance

- Errors are increasing in the number of constraints c_i...



EP redundant factorization in BPR/GPC

- Redundant factorization $n \gg d$ can erode quality of EP estimates.
- True of probit regression generally, not just the limiting case.



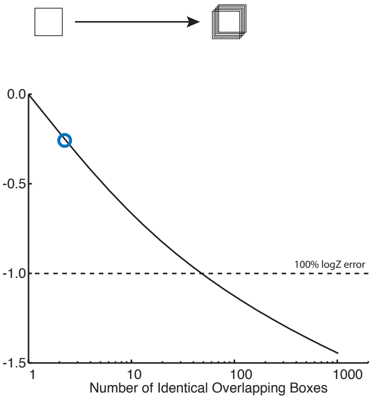
- (very unusual property of EP, unlike mean field, Laplace, etc.)

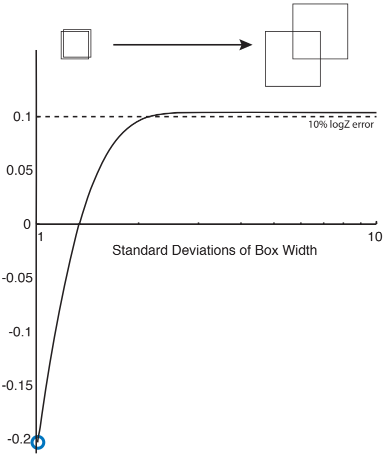
(but see [Hensman, Zwießle, Lawrence (2014) AISTATS])

- (generally agrees with Ole's comment $\log R = \log \frac{Z}{Z_{ep}} > 0$)

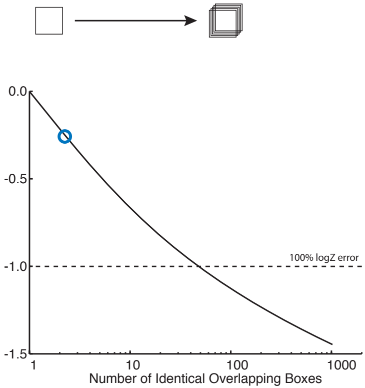
Not quite an upper bound

- A few contrived examples:





Not quite an upper bound



- tractable: spherical prior with $t_1(x) = \mathbb{1}\{x \in [-1,1]^2\}$.
- No redundant factors:

$$\log Z = \int p_0(x) t_1(x) dx$$

EP becomes: $\min KL(p_0 t_1 || p_0 \tilde{t}_1)$

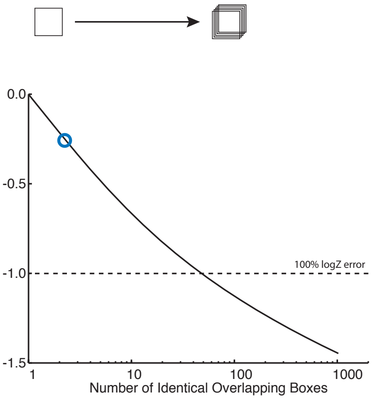
- α redundant factors:

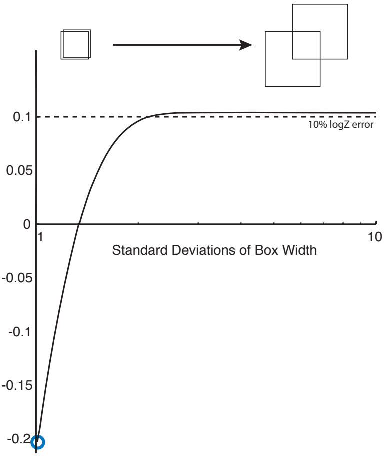
$$\log Z = \int p_0(x) t_1(x)^\alpha dx$$

EP becomes: $\min D_{\frac{1}{\alpha}}(p_0 t_1 || p_0 \tilde{t}_1)$

Not quite an upper bound

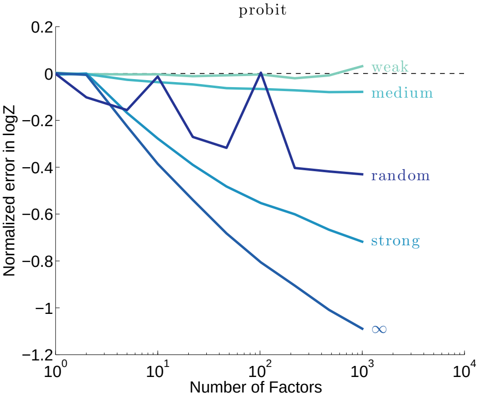
- Why this happens:





Pause... is this contradictory?

- Key differences vs. many GP classification examples:
  - GP Classification considers the $n = d$ regime
  - Axis aligned factors reduce redundancy ($t_i(x) = t(x_i)$)
  - Sample points are typically close to origin $\rightarrow$ "weak" non-gaussianity



Outline

Expectation Propagation (precisely)

- Recall for exponential families $\frac{1}{Z(\theta)} \exp\left\{\theta^\top \phi(x)\right\}$:
  - mean parameters $\mu := E_\theta(\phi(x)) = \nabla_\theta \log Z(\theta)$
  - natural parameters $\theta(\mu) = \nabla_\mu A^*(\mu)$
  - conjugate dual $A^*(\mu) = -H(p_{\theta(\mu)})$ (conditions apply)
- Our intractable exponential family distribution:

$$p(x) = \frac{1}{Z} p_0(x) \prod_{i=1}^{n} t_i(x) = \frac{1}{Z} \exp\left\{\theta^\top \phi(x) + \sum_{i=1}^{n} \bar{\theta}_i^\top \bar{\phi}_i(x)\right\}.$$

- A variational representation of our object of interest:

$$\log Z = \max_{\mu, \bar{\mu} \in \mathcal{M}} \left\{\theta^\top \mu + \sum_{i=1}^{n} \bar{\theta}_i^\top \bar{\mu}_i + H\left(p_{\theta(\mu, \bar{\mu})}\right)\right\}.$$

- (side note: this is a minimization of $KL(q||p)$, not $KL(p||q)$).

[Wainwright and Jordan (2008) FTML]

The EP entropy approximation

$$\log Z = \max_{\mu,\bar{\mu}\in\mathcal{M}} \left\{ \theta^\top \mu + \sum_{i=1}^{n} \bar{\theta}_i^\top \bar{\mu}_i + H\left(p_{\theta(\mu,\bar{\mu})}\right) \right\}$$

- Also intractable, so EP solves a relaxation to this variational problem:

$$\log Z \approx \max_{\mu,\bar{\mu}\in\mathcal{M}'} \left\{ \theta^\top \mu + \sum_{i=1}^{n} \bar{\theta}_i^\top \bar{\mu}_i + H_{ep}\left(\mu,\bar{\mu}\right) \right\},$$

- where $H_{ep}\left(\mu,\bar{\mu}\right) = H(q_{\theta(\mu)}) + \sum_{i=1}^{n} \left(H(q_{\theta(\mu,\bar{\mu}_i)}) - H(q_{\theta(\mu)}) \right)$.
- Results in the moment matching of $q^{\setminus i} t_i$ and $q^{\setminus i} \tilde{t}_i$, yielding

$$H(p) \approx H_{ep} = H(q) + \sum_{i=1}^{n} \left(H(q^{\setminus i} t_i) - H(q^{\setminus i} \tilde{t}_i) \right)$$

Why the EP entropy approximation is interesting

- The entropy form offers a key hint as to why EP can go wrong...

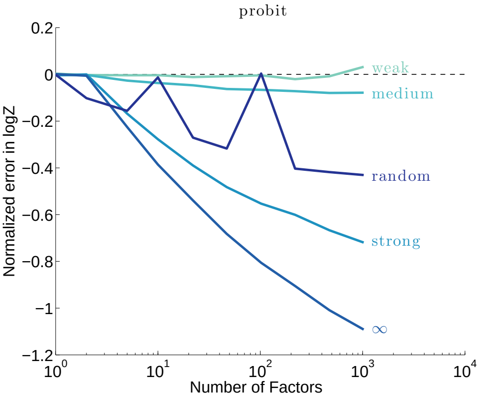
$$H(p) \approx H_{ep} = H(q) + \sum_{i=1}^{n} \Big(H(q^{\setminus i} t_i) - H(q^{\setminus i} \tilde{t}_i) \Big)$$
$$= H(q) - \sum_{i=1}^{n} KL(q^{\setminus i} t_i || q^{\setminus i} \tilde{t}_i)$$

($q^{\setminus i} t_i$ and $q^{\setminus i} \tilde{t}_i$ are moment matched, not entropy matched!)

- Of course, errors in our problem could come from:
  - the entropy approximation itself $H(p) \approx H_{ep}$
  - the relaxed constraint set $\mathcal{M} \in \mathcal{M}'$
  - the optimization routine (e.g. no fixed point, etc.)

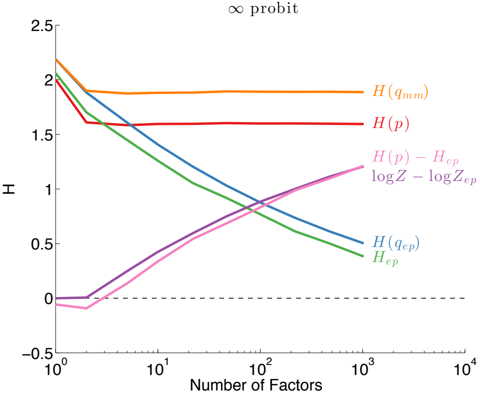
$\log Z$ errors and $H(p) \approx H_{ep}$

- Return to the errors in $\log Z$...



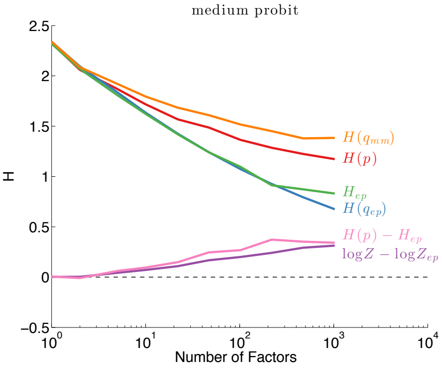
$\log Z$ errors and $H(p) \approx H_{ep}$

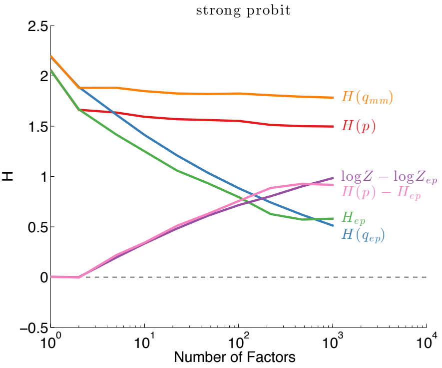
- Consider the step function case:



$\log Z$ errors and $H(p) \approx H_{ep}$

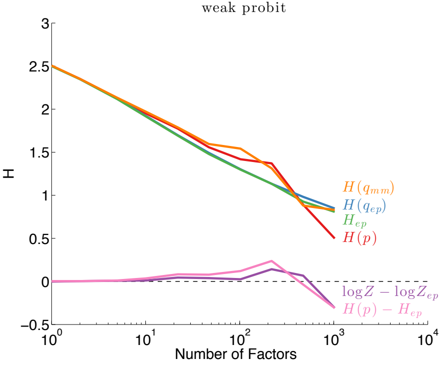
- Consider other cases:





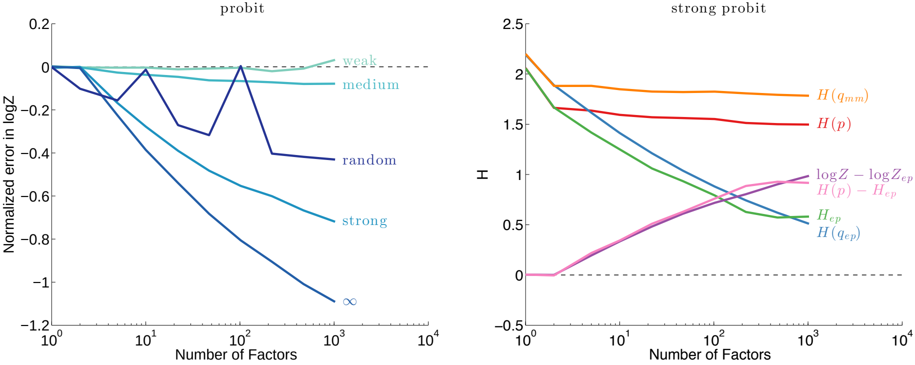
$\log Z$ errors and $H(p) \approx H_{ep}$

- Consider other cases:



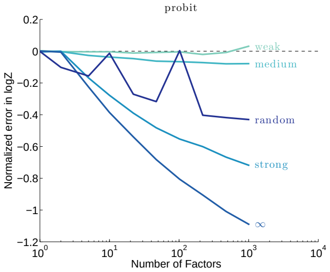
$\log Z$ errors and $H(p) \approx H_{ep}$

- At least in the Bayesian Probit Regression case, much of the EP error from redundant factors is due to entropy approximation.



Outline

Correcting EP errors via the entropy connection

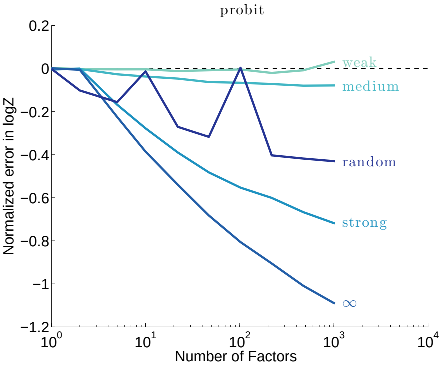


- Remove bias via sampling-type entropy approximation
  - feels like cheating...
- Change approximation to remove negative bias
  - connecting back to α-EP

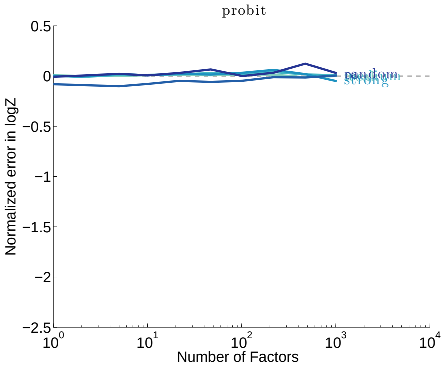
EP redundancy with entropy corrections

- For completeness...

Before correction...



After entropy correction...



Correcting bias in the EP entropy approximation

- EP entropy approximation:

$$
\begin{aligned}
H(p) \quad \approx \quad H_{ep} \quad &= \quad H(q) + \sum_{i=1}^{n} \Big(H(q^{\setminus i} t_i) - H(q^{\setminus i} \tilde{t}_i) \Big) \\
&= \quad H(q) - \sum_{i=1}^{n} KL(q^{\setminus i} t_i || q^{\setminus i} \tilde{t}_i).
\end{aligned}
$$

- α-EP

$$
\begin{aligned}
H(p) \quad \approx \quad H_{ep} \quad &= \quad H(q) + \sum_{i=1}^{n} \frac{1}{\alpha_i} \Big(H(q^{\setminus i} t_i) - H(q^{\setminus i} \tilde{t}_i) \Big) \\
&\neq \quad H(q) - \sum_{i=1}^{n} \frac{1}{\alpha_i} D_{\alpha_i}(q^{\setminus i} t_i || q^{\setminus i} \tilde{t}_i).
\end{aligned}
$$

- so one can heuristically choose α_i to remove the bias...

Outline

Takeaways for Discussion

- EP is unusual in its dependence on the *number* of factors.
- EP is very useful for normal probabilities, GPC, BPR.
- Factorization can lead to significant problems, especially $n \gg d$.
- More work is needed to understand the conditions of these errors.
- The entropy approximation gives a handle on the inherent bias...
- ...and potential for improvements.

Thanks...

- People: Alp Kuckelbir, Philipp Hennig, Simon Lacoste-Julien
- Funding: Sloan Foundation, Simons Foundation, Gatsby Trust