

Probabilistic & Unsupervised Learning

Exponential families: convexity, duality and free energies

Maneesh Sahani

maneesh@gatsby.ucl.ac.uk

**Gatsby Computational Neuroscience Unit, and
MSc ML/CSML, Dept Computer Science
University College London**

Term 1, Autumn 2019

Exponential families: the log partition function

Consider an exponential family distribution with sufficient statistic $s(X)$ and natural parameter θ (and no base factor in X alone). We can write its probability or density function as

$$p(X|\theta) = \exp\left(\theta^T s(X) - \Phi(\theta)\right)$$

where $\Phi(\theta)$ is the [log partition function](#)

$$\Phi(\theta) = \log \sum_x \exp\left(\theta^T s(x)\right)$$

$\Phi(\theta)$ plays an important role in the theory of the exponential family. For example, it maps natural parameters to the moments of the sufficient statistics:

$$\frac{\partial}{\partial \theta} \Phi(\theta) = e^{-\Phi(\theta)} \sum_x s(x) e^{\theta^T s(x)} = \mathbb{E}_\theta [s(X)] = \mu(\theta) = \mu$$

$$\frac{\partial^2}{\partial \theta^2} \Phi(\theta) = e^{-\Phi(\theta)} \sum_x s(x)^2 e^{\theta^T s(x)} - e^{-2\Phi(\theta)} \left[\sum_x s(x) e^{\theta^T s(x)} \right]^2 = \mathbb{V}_\theta [s(X)]$$

The second derivative is thus positive semi-definite, and so $\Phi(\theta)$ is [convex in \$\theta\$](#) .

Exponential families: mean parameters and negative entropy

A (minimal) exponential family distribution can also be parameterised by the **means of the sufficient statistics**.

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} [s(X)]$$

Consider the **negative entropy** of the distribution as a function of the mean parameter:

$$\Psi(\boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{\theta}} [\log p(X|\boldsymbol{\theta}(\boldsymbol{\mu}))] = \boldsymbol{\theta}^T \boldsymbol{\mu} - \Phi(\boldsymbol{\theta})$$

so

$$\boldsymbol{\theta}^T \boldsymbol{\mu} = \Phi(\boldsymbol{\theta}) + \Psi(\boldsymbol{\mu})$$

The negative entropy is **dual** to the log-partition function. For example,

$$\begin{aligned} \frac{d}{d\boldsymbol{\mu}} \Psi(\boldsymbol{\mu}) &= \frac{\partial}{\partial \boldsymbol{\mu}} (\boldsymbol{\theta}^T \boldsymbol{\mu} - \Phi(\boldsymbol{\theta})) + \frac{d\boldsymbol{\theta}}{d\boldsymbol{\mu}} \frac{\partial}{\partial \boldsymbol{\theta}} (\boldsymbol{\theta}^T \boldsymbol{\mu} - \Phi(\boldsymbol{\theta})) \\ &= \boldsymbol{\theta} + \frac{d\boldsymbol{\theta}}{d\boldsymbol{\mu}} (\boldsymbol{\mu} - \boldsymbol{\mu}) = \boldsymbol{\theta} \end{aligned}$$

Exponential families: duality

In fact, the log partition function and negative entropy are **Legendre dual** or **convex conjugate** functions.

Consider the KL divergence between distributions with natural parameters θ and θ' :

$$\begin{aligned}\mathbf{KL}[\theta||\theta'] &= \mathbf{KL}[p(X|\theta)||p(X|\theta')] = \mathbb{E}_{\theta} [-\log p(X|\theta') + \log p(X|\theta)] \\ &= -\theta'^T \mu + \Phi(\theta') + \Psi(\mu) \geq 0 \\ \Rightarrow \Psi(\mu) &\geq \theta'^T \mu - \Phi(\theta')\end{aligned}$$

where μ are the mean parameters corresponding to θ .

Now, the minimum KL divergence of zero is reached iff $\theta = \theta'$, so

$$\Psi(\mu) = \sup_{\theta'} [\theta'^T \mu - \Phi(\theta')] \quad \text{and, if finite} \quad \theta(\mu) = \operatorname{argmax}_{\theta'} [\theta'^T \mu - \Phi(\theta')]$$

The left-hand equation is the definition of the conjugate dual of a convex function.

Continuous functions are reciprocally dual, so we also have:

$$\Phi(\theta) = \sup_{\mu'} [\theta^T \mu' - \Psi(\mu')] \quad \text{and, if finite} \quad \mu(\theta) = \operatorname{argmax}_{\mu'} [\theta^T \mu' - \Psi(\mu')]$$

Thus, duality gives us another relation between θ and μ .

Duality, inference and the free energy

Consider a joint exponential family distribution on observed \mathbf{x} and latent \mathbf{z} .

$$p(\mathbf{x}, \mathbf{z}) = \exp \left[\boldsymbol{\theta}^\top s(\mathbf{x}, \mathbf{z}) - \Phi_{XZ}(\boldsymbol{\theta}) \right]$$

The posterior on \mathbf{z} is also in the exponential family, with the **clamped** sufficient statistic $s_Z(\mathbf{z}; \mathbf{x}) = s_{XZ}(\mathbf{x}^{\text{obs}}, \mathbf{z})$; the **same** (now possibly redundant) natural parameter $\boldsymbol{\theta}$; and partition function $\Phi_Z(\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} \exp \boldsymbol{\theta}^\top s_Z(\mathbf{z})$.

The likelihood is

$$\mathcal{L}(\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{\mathbf{z}} e^{\boldsymbol{\theta}^\top s(\mathbf{x}, \mathbf{z}) - \Phi_{XZ}(\boldsymbol{\theta})} = \sum_{\mathbf{z}} e^{\boldsymbol{\theta}^\top s_Z(\mathbf{z}; \mathbf{x})} e^{-\Phi_{XZ}(\boldsymbol{\theta})} = \exp[\Phi_Z(\boldsymbol{\theta}) - \Phi_{XZ}(\boldsymbol{\theta})]$$

So we can write the log-likelihood as

$$\ell(\boldsymbol{\theta}) = \sup_{\boldsymbol{\mu}_Z} \left[\underbrace{\boldsymbol{\theta}^\top \boldsymbol{\mu}_Z - \Phi_{XZ}(\boldsymbol{\theta})}_{\langle \log p(\mathbf{x}, \mathbf{z}) \rangle_q} - \underbrace{\Psi(\boldsymbol{\mu}_Z)}_{-H[q]} \right] = \sup_{\boldsymbol{\mu}_Z} \mathcal{F}(\boldsymbol{\theta}, \boldsymbol{\mu}_Z)$$

This is the familiar free energy with $q(\mathbf{z})$ represented by its mean parameters $\boldsymbol{\mu}_Z$!

Inference with mean parameters

We have described inference in terms of the distribution q , approximating as needed, then computing expected suff stats. Can we describe it instead as an optimisation over μ directly?

$$\mu_z^* = \operatorname{argmax}_{\mu_z} [\theta^\top \mu_z - \Psi(\mu_z)]$$

Inference with mean parameters

We have described inference in terms of the distribution q , approximating as needed, then computing expected suff stats. Can we describe it instead as an optimisation over μ directly?

$$\mu_z^* = \operatorname{argmax}_{\mu_z} [\theta^\top \mu_z - \Psi(\mu_z)]$$

Concave maximisation(!), but two complications:

- ▶ The optimum must be found over **feasible** means. Interdependence of the sufficient statistics may prevent arbitrary sets of mean sufficient statistics being achieved

Inference with mean parameters

We have described inference in terms of the distribution q , approximating as needed, then computing expected suff stats. Can we describe it instead as an optimisation over μ directly?

$$\mu_Z^* = \operatorname{argmax}_{\mu_Z} [\theta^\top \mu_Z - \Psi(\mu_Z)]$$

Concave maximisation(!), but two complications:

- ▶ The optimum must be found over **feasible** means. Interdependence of the sufficient statistics may prevent arbitrary sets of mean sufficient statistics being achieved
 - ▶ Feasible means are convex combinations of all the single-configuration sufficient statistics.

$$\mu = \sum_{\mathbf{x}} \nu(\mathbf{x}) s(\mathbf{x}) \quad \sum_{\mathbf{x}} \nu(\mathbf{x}) = 1$$

Inference with mean parameters

We have described inference in terms of the distribution q , approximating as needed, then computing expected suff stats. Can we describe it instead as an optimisation over μ directly?

$$\mu_Z^* = \operatorname{argmax}_{\mu_Z} [\theta^\top \mu_Z - \Psi(\mu_Z)]$$

Concave maximisation(!), but two complications:

- ▶ The optimum must be found over **feasible** means. Interdependence of the sufficient statistics may prevent arbitrary sets of mean sufficient statistics being achieved
 - ▶ Feasible means are convex combinations of all the single-configuration sufficient statistics.

$$\mu = \sum_{\mathbf{x}} \nu(\mathbf{x}) s(\mathbf{x}) \quad \sum_{\mathbf{x}} \nu(\mathbf{x}) = 1$$

- ▶ Take a Boltzmann machine on two variables, x_1, x_2 .
- ▶ The sufficient stats are $s(\mathbf{x}) = [x_1, x_2, x_1 x_2]$.
- ▶ Clearly only the stats $\mathcal{S} = \{[0, 0, 0], [0, 1, 0], [1, 0, 0], [1, 1, 1]\}$ are possible.
- ▶ Thus $\mu \in \operatorname{convex hull}(\mathcal{S})$.

Inference with mean parameters

We have described inference in terms of the distribution q , approximating as needed, then computing expected suff stats. Can we describe it instead as an optimisation over μ directly?

$$\mu_Z^* = \operatorname{argmax}_{\mu_Z} [\theta^\top \mu_Z - \Psi(\mu_Z)]$$

Concave maximisation(!), but two complications:

- ▶ The optimum must be found over **feasible** means. Interdependence of the sufficient statistics may prevent arbitrary sets of mean sufficient statistics being achieved
 - ▶ Feasible means are convex combinations of all the single-configuration sufficient statistics.

$$\mu = \sum_{\mathbf{x}} \nu(\mathbf{x}) s(\mathbf{x}) \quad \sum_{\mathbf{x}} \nu(\mathbf{x}) = 1$$

- ▶ Take a Boltzmann machine on two variables, x_1, x_2 .
 - ▶ The sufficient stats are $s(\mathbf{x}) = [x_1, x_2, x_1 x_2]$.
 - ▶ Clearly only the stats $\mathcal{S} = \{[0, 0, 0], [0, 1, 0], [1, 0, 0], [1, 1, 1]\}$ are possible.
 - ▶ Thus $\mu \in \operatorname{convex hull}(\mathcal{S})$.
- ▶ For a discrete distribution, this space of possible means is bounded by exponentially many hyperplanes connecting the discrete configuration stats: called the **marginal polytope**.

Inference with mean parameters

We have described inference in terms of the distribution q , approximating as needed, then computing expected suff stats. Can we describe it instead as an optimisation over μ directly?

$$\mu_z^* = \operatorname{argmax}_{\mu_z} [\theta^\top \mu_z - \Psi(\mu_z)]$$

Concave maximisation(!), but two complications:

- ▶ The optimum must be found over **feasible** means. Interdependence of the sufficient statistics may prevent arbitrary sets of mean sufficient statistics being achieved
 - ▶ Feasible means are convex combinations of all the single-configuration sufficient statistics.

$$\mu = \sum_{\mathbf{x}} \nu(\mathbf{x}) s(\mathbf{x}) \quad \sum_{\mathbf{x}} \nu(\mathbf{x}) = 1$$

- ▶ Take a Boltzmann machine on two variables, x_1, x_2 .
 - ▶ The sufficient stats are $s(\mathbf{x}) = [x_1, x_2, x_1 x_2]$.
 - ▶ Clearly only the stats $\mathcal{S} = \{[0, 0, 0], [0, 1, 0], [1, 0, 0], [1, 1, 1]\}$ are possible.
 - ▶ Thus $\mu \in \operatorname{convex hull}(\mathcal{S})$.
- ▶ For a discrete distribution, this space of possible means is bounded by exponentially many hyperplanes connecting the discrete configuration stats: called the **marginal polytope**.
- ▶ Even when restricted to the marginal polytope, evaluating $\Psi(\mu)$ can be challenging.

Convexity and undirected trees

- ▶ We can parametrise a discrete pairwise MRF as follows:

$$\begin{aligned} p(\mathbf{X}) &= \frac{1}{Z} \prod_i f_i(X) \prod_{(ij)} f_{ij}(X_i, X_j) \\ &= \exp \left(\sum_i \sum_k \theta_i(k) \delta(X_i = k) + \sum_{(ij)} \sum_{k,l} \theta_{ij}(k, l) \delta(X_i = k) \delta(X_j = l) - \Phi(\theta) \right) \end{aligned}$$

Convexity and undirected trees

- ▶ We can parametrise a discrete pairwise MRF as follows:

$$\begin{aligned} p(\mathbf{X}) &= \frac{1}{Z} \prod_i f_i(X) \prod_{(ij)} f_{ij}(X_i, X_j) \\ &= \exp \left(\sum_i \sum_k \theta_i(k) \delta(X_i = k) + \sum_{(ij)} \sum_{k,l} \theta_{ij}(k, l) \delta(X_i = k) \delta(X_j = l) - \Phi(\theta) \right) \end{aligned}$$

- ▶ So discrete MRFs are always exponential family, with natural and mean parameters:

$$\theta = [\theta_i(k), \theta_{ij}(k, l) \quad \forall i, j, k, l]$$

$$\mu = [p(X_i = k), p(X_i = k, X_j = l) \quad \forall i, j, k, l]$$

In particular, the mean parameters are just the singleton and pairwise probability tables.

Convexity and undirected trees

- ▶ We can parametrise a discrete pairwise MRF as follows:

$$\begin{aligned} p(\mathbf{X}) &= \frac{1}{Z} \prod_i f_i(X_i) \prod_{(ij)} f_{ij}(X_i, X_j) \\ &= \exp \left(\sum_i \sum_k \theta_i(k) \delta(X_i = k) + \sum_{(ij)} \sum_{k,l} \theta_{ij}(k, l) \delta(X_i = k) \delta(X_j = l) - \Phi(\theta) \right) \end{aligned}$$

- ▶ So discrete MRFs are always exponential family, with natural and mean parameters:

$$\theta = [\theta_i(k), \theta_{ij}(k, l) \quad \forall i, j, k, l]$$

$$\mu = [p(X_i = k), p(X_i = k, X_j = l) \quad \forall i, j, k, l]$$

In particular, the mean parameters are just the singleton and pairwise probability tables.

- ▶ If the MRF has tree structure T , the negative entropy can be written in terms of the single-site entropies and mutual informations on edges:

$$\begin{aligned} \Psi(\mu_T) &= \mathbb{E}_{\theta_T} \left[\log \prod_i p(X_i) \prod_{(ij) \in T} \frac{p(X_i, X_j)}{p(X_i)p(X_j)} \right] \\ &= - \sum_i H(X_i) + \sum_{(ij) \in T} I(X_i, X_j) \end{aligned}$$

The Bethe free energy again

We can see the Bethe free energy problem as a relaxation of the true free-energy optimisation:

$$\mu_Z^* = \operatorname{argmax}_{\mu_Z \in \mathcal{M}} [\theta^\top \mu_Z - \Psi(\mu_Z)]$$

where \mathcal{M} is the set of feasible means.

The Bethe free energy again

We can see the Bethe free energy problem as a relaxation of the true free-energy optimisation:

$$\mu_Z^* = \operatorname{argmax}_{\mu_Z \in \mathcal{M}} [\theta^\top \mu_Z - \Psi(\mu_Z)]$$

where \mathcal{M} is the set of feasible means.

1. Relax $\mathcal{M} \rightarrow \mathcal{L}$, where \mathcal{L} is the set of **locally consistent** means (i.e. all nested means marginalise correctly).

The Bethe free energy again

We can see the Bethe free energy problem as a relaxation of the true free-energy optimisation:

$$\mu_Z^* = \operatorname{argmax}_{\mu_Z \in \mathcal{M}} [\theta^\top \mu_Z - \Psi(\mu_Z)]$$

where \mathcal{M} is the set of feasible means.

1. Relax $\mathcal{M} \rightarrow \mathcal{L}$, where \mathcal{L} is the set of **locally consistent** means (i.e. all nested means marginalise correctly).
2. **Approximate** $\Psi(\mu_Z)$ by the tree-structured form

$$\Psi_{\text{Bethe}}(\mu_Z) = - \sum_i H(X_i) + \sum_{(ij) \in G} I(X_i, X_j)$$

The Bethe free energy again

We can see the Bethe free energy problem as a relaxation of the true free-energy optimisation:

$$\mu_Z^* = \operatorname{argmax}_{\mu_Z \in \mathcal{M}} [\theta^\top \mu_Z - \Psi(\mu_Z)]$$

where \mathcal{M} is the set of feasible means.

1. Relax $\mathcal{M} \rightarrow \mathcal{L}$, where \mathcal{L} is the set of **locally consistent** means (i.e. all nested means marginalise correctly).
2. **Approximate** $\Psi(\mu_Z)$ by the tree-structured form

$$\Psi_{\text{Bethe}}(\mu_Z) = - \sum_i H(X_i) + \sum_{(ij) \in \mathcal{G}} I(X_i, X_j)$$

\mathcal{L} is still a convex set (polytope for discrete problems). However Ψ_{Bethe} is not convex.

Convexifying BP

Consider instead an **upper bound** on $\Phi(\theta)$:

Imagine a set of spanning trees T for the MRF, each with its own parameters θ_T, μ_T . By padding entries corresponding to off-tree edges with zero, we can assume that θ_T has the same dimensionality as θ .

Suppose also that we have a distribution β over the spanning trees so that $\mathbb{E}_\beta [\theta_T] = \theta$.

Then by the convexity of $\Phi(\theta)$,

$$\Phi(\theta) = \Phi(\mathbb{E}_\beta [\theta_T]) \leq \mathbb{E}_\beta [\Phi(\theta_T)]$$

If we were to **tighten** the upper bound we might obtain a good approximation to Φ :

$$\Phi(\theta) \leq \inf_{\beta, \theta_T: \mathbb{E}_\beta [\theta_T] = \theta} \mathbb{E}_\beta [\Phi(\theta_T)]$$

Convex Upper Bounds on the Log Partition Function

$$\Phi(\boldsymbol{\theta}) \leq \inf_{\boldsymbol{\theta}_T: \mathbb{E}_\beta[\boldsymbol{\theta}_T] = \boldsymbol{\theta}} \mathbb{E}_\beta [\Phi(\boldsymbol{\theta}_T)]$$

Solve this constrained optimisation problem using Lagrange multipliers:

$$\mathcal{L} = \mathbb{E}_\beta [\Phi(\boldsymbol{\theta}_T)] - \boldsymbol{\lambda}^\top (\mathbb{E}_\beta [\boldsymbol{\theta}_T] - \boldsymbol{\theta})$$

Setting the derivatives wrt $\boldsymbol{\theta}_T$ to zero, we get:

$$\begin{aligned} \beta(T)\boldsymbol{\lambda}_T - \beta(T)\boldsymbol{\Pi}_T(\boldsymbol{\lambda}) &= 0 \\ \boldsymbol{\lambda}_T &= \boldsymbol{\Pi}_T(\boldsymbol{\lambda}) \end{aligned}$$

where $\boldsymbol{\Pi}_T(\boldsymbol{\lambda})$ are the Lagrange multipliers corresponding to vertices and edges on the tree T .

Although there can be many $\boldsymbol{\theta}_T$ parameters, at optimum they are all constrained: their corresponding mean parameters are all consistent with each other and with $\boldsymbol{\lambda}$.

Convex Upper Bounds on the Log Partition Function

$$\begin{aligned}\Phi(\boldsymbol{\theta}) &\leq \sup_{\boldsymbol{\lambda}} \inf_{\boldsymbol{\theta}_T} \mathbb{E}_{\beta} [\Phi(\boldsymbol{\theta}_T)] - \boldsymbol{\lambda}^T (\mathbb{E}_{\beta} [\boldsymbol{\theta}_T] - \boldsymbol{\theta}) \\ &= \sup_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \boldsymbol{\theta} + \mathbb{E}_{\beta} \left[\inf_{\boldsymbol{\theta}_T} \Phi(\boldsymbol{\theta}_T) - \boldsymbol{\theta}_T^T \boldsymbol{\Pi}_T(\boldsymbol{\lambda}) \right] \\ &= \sup_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \boldsymbol{\theta} + \mathbb{E}_{\beta} [-\Psi(\boldsymbol{\Pi}_T(\boldsymbol{\lambda}))] \\ &= \sup_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \boldsymbol{\theta} + \mathbb{E}_{\beta} \left[\sum_i H_{\boldsymbol{\lambda}}(X_i) - \sum_{(ij) \in \mathcal{T}} I_{\boldsymbol{\lambda}}(X_i, X_j) \right] \\ &= \sup_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \boldsymbol{\theta} + \sum_i H_{\boldsymbol{\lambda}}(X_i) - \sum_{(ij)} \beta_{ij} I_{\boldsymbol{\lambda}}(X_i, X_j)\end{aligned}$$

This is a **convexified** Bethe free energy.

EP free energy

A Bethe-like approach also casts EP as a variational energy fixed point method.

Consider finding marginals of a (posterior) distribution defined by clique potentials:

$$P(\mathcal{Z}) \propto f_0(\mathcal{Z}) \prod_i f_i(\mathcal{Z}_i)$$

where all factors have exponential form, f_0 is in a tractable exponential family (possibly uniform) but the f_i are **jointly intractable** – i.e. product cannot be marginalised, although individual terms may be (numerically) tractable.

Augment by including tractable ExpFam terms with zero natural parameters

$$P(\mathcal{Z}) \propto e^{\theta_0^T \mathbf{s}_0(\mathcal{Z})} \prod_i e^{\theta_i^T \mathbf{s}_i(\mathcal{Z}_i)} e^{\mathbf{0}^T \tilde{\mathbf{s}}_i(\mathcal{Z}_i)} = e^{\theta_0^T \mathbf{s}_0(\mathcal{Z}) + \sum_i (\theta_i^T \mathbf{s}_i(\mathcal{Z}_i) + \tilde{\theta}^T \tilde{\mathbf{s}}_i(\mathcal{Z}_i))}$$

Now, the variational dual principle tells us that the expected sufficient statistics:

$$\boldsymbol{\mu}_0^* = \langle \mathbf{s}_0 \rangle_P; \quad \boldsymbol{\mu}_i^* = \langle \mathbf{s}_i(\mathcal{Z}_i) \rangle_P; \quad \tilde{\boldsymbol{\mu}}_i^* = \langle \tilde{\mathbf{s}}_i \rangle_P$$

are given by

$$\{\boldsymbol{\mu}_0^*, \boldsymbol{\mu}_i^*, \tilde{\boldsymbol{\mu}}_i^*\} = \underset{\{\boldsymbol{\mu}_0, \boldsymbol{\mu}_i, \tilde{\boldsymbol{\mu}}_i\} \in \mathcal{M}}{\text{argmax}} \left[\boldsymbol{\theta}_0^T \boldsymbol{\mu}_0 + \sum_i \left(\boldsymbol{\theta}_i^T \boldsymbol{\mu}_i + \mathbf{0}^T \tilde{\boldsymbol{\mu}}_i \right) - \Psi(\boldsymbol{\mu}_0, \boldsymbol{\mu}_i, \tilde{\boldsymbol{\mu}}_i) \right]$$

EP relaxation

The EP algorithm relaxes this optimisation:

- ▶ Relax \mathcal{M} to **locally consistent** marginals, retaining consistency across each edge connecting $\{\mu_0, \tilde{\mu}_i\}$ (as in BP on a junction graph); and between pairs $(\mu_i, \tilde{\mu}_i)$.
- ▶ Replace negative entropy by $\Psi_{\text{Bethe}}(\{\mu_0, \tilde{\mu}_i\}) - \sum_i (\mathbf{H}[\mu_i, \tilde{\mu}_i] - \mathbf{H}[\tilde{\mu}_i])$.
- ▶ In effect, drop links between different μ_i and run reparameterisation on a junction graph.

EP relaxation

The EP algorithm relaxes this optimisation:

- ▶ Relax \mathcal{M} to **locally consistent** marginals, retaining consistency across each edge connecting $\{\mu_0, \tilde{\mu}_i\}$ (as in BP on a junction graph); and between pairs $(\mu_i, \tilde{\mu}_i)$.
- ▶ Replace negative entropy by $\Psi_{\text{Bethe}}(\{\mu_0, \tilde{\mu}_i\}) - \sum_i (\mathbf{H}[\mu_i, \tilde{\mu}_i] - \mathbf{H}[\tilde{\mu}_i])$.
- ▶ In effect, drop links between different μ_i and run reparameterisation on a junction graph.

The free-energy-based approximate marginals include μ_i which are refined during updates.

- ▶ Direct learning on the EP free-energy would use these marginals rather than the approximate ones (and a local normaliser formed by integrating over $f_i(\mathcal{Z}_i)q_{-i}(\mathcal{Z}_i)$).
- ▶ These estimates may yield more accurate results than optimising θ according to expectations under the tractable marginals $\tilde{\mu}_i$.

References

- ▶ **Graphical Models, Exponential Families, and Variational Inference.** Wainwright and Jordan. **Foundations and Trends in Machine Learning, 2008 1:1-305.**
- ▶ Exact Maximum A Posteriori Estimation for Binary Images. Greig, Porteous and Seheult, *Journal of the Royal Statistical Society B*, 51(2):271-279, 1989.
- ▶ Fast Approximate Energy Minimization via Graph Cuts. Boykov, Veksler and Zabih, *International Conference on Computer Vision* 1999.
- ▶ MAP estimation via agreement on (hyper)trees: Message-passing and linear-programming approaches. Wainwright, Jaakkola and Willsky, *IEEE Transactions on Information Theory*, 2005, 51(11):3697-3717.
- ▶ Learning Associative Markov Networks. Taskar, Chatalbashev and Koller, *International Conference on Machine Learning*, 2004.
- ▶ A New Class of Upper Bounds on the Log Partition Function. Wainwright, Jaakkola and Willsky. *IEEE Transactions on Information Theory*, 2005, 51(7):2313-2335.
- ▶ MAP Estimation, Linear Programming and Belief Propagation with Convex Free Energies. Weiss, Yanover and Meltzer, *Uncertainty in Artificial Intelligence*, 2007.