Probabilistic & Unsupervised Learning

Convex Algorithms in Approximate Inference

Maneesh Sahani
maneesh@gatsby.ucl.ac.uk

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

Term 1, Autumn 2014

Convexity and Approximate Inference

The theory of convex functions and convex spaces has long been central to optimisation. It has recently also found application in the theory of free energy and approximation:

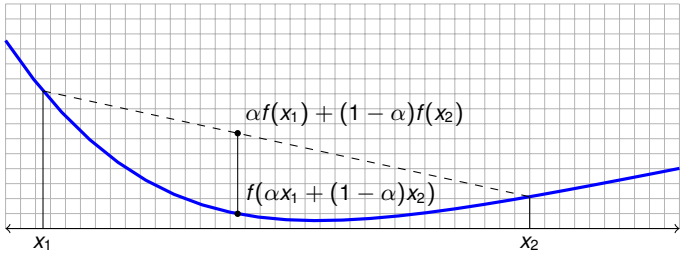
- Linear programming relaxation as an approximate method to find the MAP assignment in Markov random fields.
- Attractive Markov random fields: binary case exact and related to a maximum flow-minimum cut problem in graph theory (a linear program). Approximate otherwise.
- Unified view of approximate inference as optimization on the marginal polytope.
- Tree-structured convex upper bounds on the log partition function (convexified belief propagation).
- Learning graphical models using maximum margin principles and convex approximate inference.

Convexity

A convex function $f : X \to \mathbb{R}$ is one where

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2)$$

for any $x_1, x_2 \in X$ and $0 \leq \alpha \leq 1$.



Convex functions have a global infimum (unless not bounded below) and there are efficient algorithms to find a minimum subject to convex constraints.

Examples: linear programs (LP), quadratic programs (QP), second-order cone programs (SOCP), semi-definite programs (SDP), geometric programs.

LP Relaxation for Markov Random Fields

Consider a discrete Markov random field (MRF) with pairwise interactions:

$$p(\mathbf{X}) = \frac{1}{Z} \prod_{(ij)} f_{ij}(X_i, X_j) \prod_i f_i(X_i) = \frac{1}{Z} \exp\left(\sum_{(ij)} E_{ij}(X_i, X_j) + \sum_i E_i(X_i) \right)$$

The problem is to find the most likely configuration $\mathbf{X}^{\text{MAP}}$:

$$\mathbf{X}^{\text{MAP}} = \operatorname*{argmax}_{\mathbf{X}} \sum_{(ij)} E_{ij}(X_i, X_j) + \sum_i E_i(X_i)$$

Reformulate in terms of indicator variables:

$$b_i(x_i) = \delta(X_i = x_i)$$
$$b_{ij}(x_i, x_j) = \delta(X_i = x_i)\delta(X_j = x_j)$$

where $\delta(\cdot) = 1$ if argument is true, 0 otherwise. Each $b_i(x_i)$ is an indicator for whether variable X_i takes on value x_i. The indicator variables need to satisfy certain constraints:

$b_i(x_i), b_{ij}(x_i, x_j) \in \{0, 1\}$ Indicator variables are binary variables.

$\sum_{x_i} b_i(x_i) = 1$ X_i takes on exactly one value.

$\sum_{x_j} b_{ij}(x_i, x_j) = b_i(x_i)$ Pairwise indicators are consistent with single-site indicators.

LP Relaxation for Markov Random Fields

MAP assignment problem is equivalent to:

$$\underset{\{b_i, b_{ij}\}}{\text{argmax}} \sum_{(ij)} \sum_{x_i, x_j} b_{ij}(x_i, x_j) E_{ij}(x_i, x_j) + \sum_i \sum_{x_i} b_i(x_i) E_i(x_i)$$

with constraints:

$$\forall i, j, x_i, x_j : \quad b_i(x_i), b_{ij}(x_i, x_j) \in \{0, 1\} \quad \sum_{x_i} b_i(x_i) = 1 \quad \sum_{x_j} b_{ij}(x_i, x_j) = b_i(x_i)$$

The linear programming relaxation for MRFs is:

$$\underset{\{b_i, b_{ij}\}}{\text{argmax}} \sum_{(ij)} \sum_{x_i, x_j} b_{ij}(x_i, x_j) E_{ij}(x_i, x_j) + \sum_i \sum_{x_i} b_i(x_i) E_i(x_i)$$

with constraints:

$$\forall i, j, x_i, x_j : \quad b_i(x_i), b_{ij}(x_i, x_j) \in [0, 1] \quad \sum_{x_i} b_i(x_i) = 1 \quad \sum_{x_j} b_{ij}(x_i, x_j) = b_i(x_i)$$

LP Relaxation for Markov Random Fields

- The LP relaxation is a linear program which can be solved efficiently.
- If the solution is integral, i.e. each $b_i(x_i), b_{ij}(x_i, x_j) \in \{0, 1\}$, then the solution corresponds to **the MAP solution $\mathbf{X}^{\text{MAP}}$**.
- LP relaxation is a zero-temperature version of the Bethe free energy formulation of loopy BP, where the Bethe entropy term can be ignored.
- If the MRF is binary and attractive, then (a slightly different reformulation of LP relaxation) will **always give the MAP solution**.
- Next: we show how to find the MAP solution directly for binary attractive MRFs using network flow.

Attractive Binary MRFs and Max Flow-Min Cut

Binary MRFs:

$$p(\mathbf{X}) = \frac{1}{Z} \exp\left(\sum_{(ij)} W_{ij} \delta(X_i = X_j) + \sum_i c_i X_i \right)$$

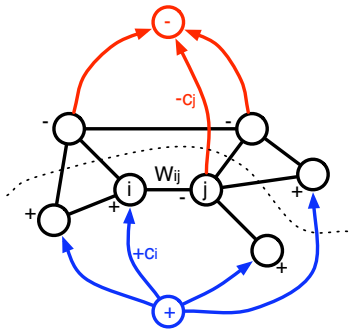
The binary MRF is attractive if $W_{ij} \geq 0$ for all i, j.

- Neighbouring variables 'prefer' to be in the same state.
- No loss of generality; any Boltzmann machines with positive interactions can be reparametrised to this form.
- Many practical MRFs are attractive, e.g. image segmentation, webpage classification.
- MAP **X** can be found efficiently by converting problem into a maximum flow-minimum cut program.

Attractive Binary MRFs and Max Flow-Min Cut

The MAP problem:

$$\underset{\mathbf{x}}{\text{argmax}} \sum_{(ij)} W_{ij} \delta(x_i = x_j) + \sum_i c_i x_i$$



Construct a network as follows:

1. Edges (ij) are undirected with weight $\lambda_{ij} = W_{ij}$;
2. Add a source s and a sink t node;
3. $c_i > 0$: Connect the source node to variable i with weight $\lambda_{si} = c_i$;
4. $c_j < 0$: Connect variable j to the sink node with weight $\lambda_{jt} = -c_j$.

A **cut** is a partition of the nodes into S and T with $s \in S$ and $t \in T$. The weight of the cut is

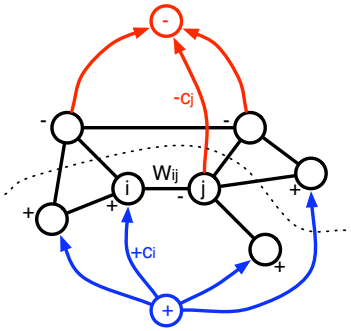
$$\Lambda(S, T) = \sum_{i \in S, j \in T} \lambda_{ij}$$

The **minimum cut** problem is to find the cut with minimum weight.

Attractive Binary MRFs and Max Flow-Min Cut

Identify an assignment $\mathbf{X} = \mathbf{x}$ with a cut:

$$S = \{s\} \cup \{i : x_i = 1\}$$
$$T = \{t\} \cup \{j : x_j = 0\}$$



The weight of the cut is:

$$\begin{aligned}\Lambda(S,T) &= \sum_{(ij)} W_{ij}\delta(x_i \neq x_j) \\ &\quad + \sum_i (1 - x_i)\max(0, c_i) \\ &\quad + \sum_j x_j \max(0, -c_j) \\ &= -\sum_{(ij)} W_{ij}\delta(x_i = x_j) - \sum_i x_i c_i + \text{constant}\end{aligned}$$

So finding the minimum cut corresponds to finding the MAP assignment.

How do we find the minimum cut? The minimum cut problem is dual to the **maximum flow problem**, i.e. find the maximum flow allowable from the source to the sink through the network. This can be solved extremely efficiently (see wikipedia entry).

The framework can be generalized to general attractive MRFs, but will not be exact anymore.

Exponential families: the log partition function

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

$$p(X|\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^\top s(X) - \Phi(\boldsymbol{\theta})\right)$$

where $\Phi(\boldsymbol{\theta})$ is the log partition function

$$\Phi(\boldsymbol{\theta}) = \log \sum_x \exp\left(\boldsymbol{\theta}^\top s(x)\right)$$

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

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

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

The second derivative is thus positive semi-definite, and so $\Phi(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$.

Exponential families: mean parameters and negative entropy

An exponential family distribution can also (almost always) be parameterised by the means of the sufficient statistics.

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

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

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

so

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

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

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

Exponential families: duality

In fact, the log partition function and negative entropy are conjugate dual functions.

Consider the KL divergence between distributions with natural parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$:

$$\begin{aligned}\mathbf{KL}\left[\boldsymbol{\theta}\|\boldsymbol{\theta}'\right] &= \mathbf{KL}\left[p(X|\boldsymbol{\theta})\|p(X|\boldsymbol{\theta}')\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[-\log p(X|\boldsymbol{\theta}') + \log p(X|\boldsymbol{\theta})\right] \\ &= -\boldsymbol{\theta}'^\top\boldsymbol{\mu} + \Phi(\boldsymbol{\theta}') + \Psi(\boldsymbol{\mu}) \geq 0 \\ \Rightarrow \Psi(\boldsymbol{\mu}) &\geq \boldsymbol{\theta}'^\top\boldsymbol{\mu} - \Phi(\boldsymbol{\theta}')\end{aligned}$$

where $\boldsymbol{\mu}$ are the mean parameters corresponding to $\boldsymbol{\theta}$.

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

$$\Psi(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta}'}\left[\boldsymbol{\theta}'^\top\boldsymbol{\mu} - \Phi(\boldsymbol{\theta}')\right] \qquad \text{and, if finite} \qquad \boldsymbol{\mu}(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}'}\left[\boldsymbol{\theta}'^\top\boldsymbol{\mu} - \Phi(\boldsymbol{\theta}')\right]$$

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

Continuous functions are reciprocally dual, so we also have:

$$\Phi(\boldsymbol{\theta}) = \sup_{\boldsymbol{\mu}'}\left[\boldsymbol{\theta}^\top\boldsymbol{\mu}' - \Psi(\boldsymbol{\mu}')\right] \qquad \text{and, if finite} \qquad \boldsymbol{\mu}(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\mu}'}\left[\boldsymbol{\theta}^\top\boldsymbol{\mu}' - \Psi(\boldsymbol{\mu}')\right]$$

Thus, duality gives us another relation between $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$.

Duality, inference and the free energy

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

$$p(\mathbf{x}, \mathbf{y}) = \exp\left[\boldsymbol{\theta}^\mathsf{T} s(\mathbf{x}, \mathbf{y}) - \Phi_{XY}(\boldsymbol{\theta})\right]$$

The posterior on $\mathbf{y}$ is also in the exponential family, with the clamped sufficient statistic $s_Y(\mathbf{y}; \mathbf{x}) = s_{XY}(\mathbf{x}^{\text{obs}}, \mathbf{y})$; the same (now possibly redundant) natural parameter $\boldsymbol{\theta}$; and partition function $\Phi_Y(\boldsymbol{\theta}) = \log \sum_{\mathbf{y}} \exp \boldsymbol{\theta}^\mathsf{T} s_Y(\mathbf{y})$.

The likelihood is

$$\mathcal{L}(\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{\mathbf{y}} e^{\boldsymbol{\theta}^\mathsf{T} s(\mathbf{x},\mathbf{y}) - \Phi_{XY}(\boldsymbol{\theta})} = \sum_{\mathbf{y}} e^{\boldsymbol{\theta}^\mathsf{T} s_Y(\mathbf{y};\mathbf{x})} e^{-\Phi_{XY}(\boldsymbol{\theta})} = \exp[\Phi_Y(\boldsymbol{\theta}) - \Phi_{XY}(\boldsymbol{\theta})]$$

So we can write the log-likelihood as

$$\ell(\boldsymbol{\theta}) = \sup_{\boldsymbol{\mu}_Y} [\underbrace{\boldsymbol{\theta}^\mathsf{T} \boldsymbol{\mu}_Y - \Phi_{XY}(\boldsymbol{\theta})}_{\langle \log p(\mathbf{x},\mathbf{y}) \rangle_q} - \underbrace{\Psi(\boldsymbol{\mu}_Y)}_{\mathbf{H}[q]}] = \sup_{\boldsymbol{\mu}_Y} \mathcal{F}(\boldsymbol{\theta}, \boldsymbol{\mu}_Y)$$

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

Inference with mean parameters

We have described inference in terms of the distribution q, approximating as needed. Can we describe it instead as an optimisation over $\boldsymbol{\mu}$ directly?

$$\boldsymbol{\mu}_Y^* = \operatorname*{argmax}_{\boldsymbol{\mu}_Y} [\boldsymbol{\theta}^\mathsf{T} \boldsymbol{\mu}_Y - \Psi(\boldsymbol{\mu}_Y)]$$

Concave maximisation(!), but two complications:

- The optimum must be found over feasible means. Interdependance 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.
    $$\boldsymbol{\mu} = \sum_{\mathbf{x}} \nu(\mathbf{x}) s(\mathbf{x}) \qquad \sum_{\mathbf{x}} \nu(\mathbf{x}) = 1$$
    - 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 $\boldsymbol{\mu} \in$ 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 $\Psi(\boldsymbol{\mu})$ can be challenging.

Convexity and undirected trees

- We can parametrise a discrete pairwise MRF as follows:
  $$p(\mathbf{X}) = \frac{1}{Z} \prod_i f_i(X) \prod_{(ij)} f_{ij}(X_i, X_j)$$
  $$= \exp\left(\sum_i \sum_{x_i} \boldsymbol{\theta}_i(x_i) \delta(X_i = x_i) + \sum_{(ij)} \sum_{x_i, x_j} \boldsymbol{\theta}_{ij}(x_i, x_j) \delta(X_i = x_i) \delta(X_j = x_j) - \Phi(\boldsymbol{\theta}) \right)$$
- So discrete MRFs are always exponential family, with natural and mean parameters:
  $$\boldsymbol{\theta} = \left[\boldsymbol{\theta}_i(x_i), \boldsymbol{\theta}_{ij}(x_i, x_j) \quad \forall i, j, x_i, x_j\right]$$
  $$\boldsymbol{\mu} = \left[p(X_i = x_i), p(X_i = x_i, X_j = x_j) \quad \forall i, j, x_i, x_j\right]$$
  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:
  $$\Psi(\boldsymbol{\mu}_T) = \mathbb{E}_{\boldsymbol{\theta}_T} \left[\log \prod_i p(X_i) \prod_{(ij) \in T} \frac{p(X_i, X_j)}{p(X_i) p(X_j)} \right]$$
  $$= -\sum_i H(X_i) + \sum_{(ij) \in T} I(X_i, X_j)$$

The Bethe free energy again

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

$$\boldsymbol{\mu}_Y^* = \operatorname*{argmax}_{\boldsymbol{\mu}_Y \in \mathcal{M}} [\boldsymbol{\theta}^\mathsf{T} \boldsymbol{\mu}_Y - \Psi(\boldsymbol{\mu}_Y)]$$

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

1. Relax $\mathcal{M} \to \mathcal{L}$, where $\mathcal{L}$ is the set of locally consistent means (i.e. all nested means marginalise correctly).
2. Approximate $\Psi(\boldsymbol{\mu}_Y)$ by the tree-structured form
   $$\Psi_{\text{Bethe}}(\boldsymbol{\mu}_Y) = -\sum_i H(X_i) + \sum_{(ij) \in T} I(X_i, X_j)$$

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

Convexifying BP

Consider instead an upper bound on $\Phi(\boldsymbol{\theta})$:

Imagine a set of spanning trees T for the MRF, each with its own parameters $\boldsymbol{\theta}_T, \boldsymbol{\mu}_T$. By padding entries corresponding to off-tree edges with zero, we can assume that $\boldsymbol{\theta}_T$ has the same dimensionality as $\boldsymbol{\theta}$.

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

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

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

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

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

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:

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

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

$$\beta(T)\boldsymbol{\lambda}_T - \beta(T)\Pi_T(\boldsymbol{\lambda}) = 0$$
$$\boldsymbol{\lambda}_T = \Pi_T(\boldsymbol{\lambda})$$

where $\Pi_T(\boldsymbol{\lambda})$ 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

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

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

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.