

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: 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: 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

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

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

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

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

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

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

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

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 $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$:

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

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

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

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

Continuous functions are reciprocally dual, so we also have:

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

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{z} .

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

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

The likelihood is

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

So we can write the log-likelihood as

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

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

Convexity and undirected trees

► We can parametrise a discrete pairwise MRF as follows:

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

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

$$\begin{aligned} \boldsymbol{\theta} &= [\boldsymbol{\theta}_i(k), \boldsymbol{\theta}_{ij}(k, l) \quad \forall i, j, k, l] \\ \boldsymbol{\mu} &= [p(X_i = k), p(X_i = k, X_j = l) \quad \forall i, j, k, l] \end{aligned}$$

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

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

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

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 $\boldsymbol{\mu}$ directly?

$$\boldsymbol{\mu}_{\mathbf{z}}^* = \operatorname{argmax}_{\boldsymbol{\mu}_{\mathbf{z}}} [\boldsymbol{\theta}^\top \boldsymbol{\mu}_{\mathbf{z}} - \Psi(\boldsymbol{\mu}_{\mathbf{z}})]$$

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.

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

- 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 $\boldsymbol{\mu} \in \text{convex hull}(\mathcal{S})$.

- 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.

The Bethe free energy again

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

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

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** $\Psi(\boldsymbol{\mu}_{\mathbf{z}})$ by the tree-structured form

$$\Psi_{\text{Bethe}}(\boldsymbol{\mu}_{\mathbf{z}}) = - \sum_i H(X_i) + \sum_{(ij) \in \mathcal{G}} I(X_i, X_j)$$

\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 \mathcal{T} for the MRF, each with its own parameters $\theta_{\mathcal{T}}, \mu_{\mathcal{T}}$. By padding entries corresponding to off-tree edges with zero, we can assume that $\theta_{\mathcal{T}}$ has the same dimensionality as θ .

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

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

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

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

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

Convex Upper Bounds on the Log Partition Function

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

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

Convex Upper Bounds on the Log Partition Function

$$\Phi(\theta) \leq \inf_{\theta_{\mathcal{T}}: \mathbb{E}_{\beta} [\theta_{\mathcal{T}}] = \theta} \mathbb{E}_{\beta} [\Phi(\theta_{\mathcal{T}})]$$

Solve this constrained optimisation problem using Lagrange multipliers:

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

Setting the derivatives wrt $\theta_{\mathcal{T}}$ to zero, we get:

$$\begin{aligned} \beta(\mathcal{T}) \lambda_{\mathcal{T}} - \beta(\mathcal{T}) \Pi_{\mathcal{T}}(\lambda) &= 0 \\ \lambda_{\mathcal{T}} &= \Pi_{\mathcal{T}}(\lambda) \end{aligned}$$

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

Although there can be many $\theta_{\mathcal{T}}$ parameters, at optimum they are all constrained: their corresponding mean parameters are all consistent with each other and with λ .

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 factor 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

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

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

$$\mu_0^* = \langle \mathbf{s}_0 \rangle_p; \quad \mu_i^* = \langle \mathbf{s}_i(\mathcal{Z}_i) \rangle_p; \quad \tilde{\mu}_i^* = \langle \tilde{\mathbf{s}}_i \rangle_p$$

are given by

$$\{\mu_0^*, \mu_i^*, \tilde{\mu}_i^*\} = \underset{\{\mu_0, \mu_i, \tilde{\mu}_i\} \in \mathcal{M}}{\operatorname{argmax}} \left[\theta_0^{\top} \mu_0 + \sum_i (\theta_i^{\top} \mu_i + \tilde{\theta}^{\top} \tilde{\mu}_i) - \Psi(\mu_0, \mu_i, \tilde{\mu}_i) \right]$$

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.

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.