CS 6347

Lecture 20-21

Exponential Families &
Expectation Propagation

Discrete State Spaces

- We have been focusing on the case of MRFs over discrete state spaces
- Probability distributions over discrete spaces correspond to vectors of probabilities for each element in the space such that the vector sums to one
  - The partition function is simply a sum over all of the possible values for each variable
  - Entropy of the distribution is nonnegative and is also computed by summing over the state space

Continuous State Spaces

$$p(x) = \frac{1}{Z} \prod_C \psi_{C(x_C)}$$

- For continuous state spaces, the partition function is now an integral

$$Z = \int \prod_C \psi_{C(x_C)}\, dx$$

- The entropy becomes

$$H(x) = -\int p(x) \log p(x)\, dx$$

Differential Entropy

$$H(x) = -\int p(x) \log p(x)\, dx$$

- This is called the differential entropy
  - It is not always greater than or equal to zero
    - Easy to construct such distributions:
      - Let $q(x)$ be the uniform distribution over the interval $[a, b]$, what is the entropy of $q(x)$?

Differential Entropy

$$H(x) = -\int p(x) \log p(x)\, dx$$

- This is called the differential entropy
  - It is not always greater than or equal to zero
    - Easy to construct such distributions:
      - Let $q(x)$ be the uniform distribution over the interval $[a, b]$, what is the entropy of $q(x)$?

$$H(q) = -\int_a^b \frac{1}{b-a} \log \frac{1}{b-a}\, dx = \log(b-a)$$

KL Divergence

$$d(q||p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

- The KL-divergence is still nonnegative, even though it contains the differential entropy
  - This means that all of the observations that we made for finite state spaces will carry over to the continuous case
    - The EM algorithm, mean-field methods, etc.
  - Most importantly

$$\log Z \geq H(q) + \sum_C \int q_C(x_C) \log \psi_C(x_C)\, dx_C$$

Continuous State Spaces

- Examples of probability distributions over continuous state spaces
  - The uniform distribution over the interval $[a, b]$

$$q(x) = \frac{1_{x \in [a,b]}}{b - a}$$

  - The multivariate normal distribution with mean μ and covariance matrix Σ

$$q(x) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Continuous State Spaces

- What makes continuous distributions so difficult to deal with?
  - They may not be compactly representable
  - Families of continuous distributions need not be closed under marginalization
    - The marginal distributions of multivariate normal distributions are again (multivariate) normal distributions
  - Integration problems of interest (e.g., the partition function or marginal distributions) may not have closed form solutions
    - Integrals may also not exist!

The Exponential Family

$$p(x|\theta) = \mathrm{h(x)} \cdot \exp(\langle\theta, \phi(x)\rangle - \log Z(\theta))$$

- A distribution is a member of the exponential family if its probability density function can be expressed as above for some choice of parameters θ and potential functions $\phi(x)$
- We are only interested in models for which $Z(\theta)$ is finite
- The family of log-linear models that we have been focusing on in the discrete case belong to the exponential family

The Exponential Family

$$p(x|\theta) = \mathrm{h(x)} \cdot \exp(\langle \theta, \phi(x) \rangle - \log Z(\theta))$$

- As in the discrete case, there is not necessarily a unique way to express a distribution in this form

- We say that the representation is minimal if there does not exist a vector a such that

$$\langle a, \phi(x) \rangle = const$$

  – In this case, there is a unique parameter vector associated with each member of the family

  – The ϕ are called sufficient statistics for the distribution

The Multivariate Normal

$$q(x|\mu, \Sigma) = \frac{1}{Z}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

- The multivariate normal distribution is a member of the exponential family

$$q(x|\theta) = \frac{1}{Z(\theta)}\exp\left(\sum_i \theta_i x_i + \sum_{i \geq j} \theta_{ij} x_i x_j\right)$$

- The mean and the covariance matrix (must be positive semidefinite) are sufficient statistics of the multivariate normal distribution

The Exponential Family

- Many of the discrete distributions that you have seen before are members of the exponential family
  - Binomial, Poisson, Bernoulli, Gamma, Beta, Laplace, Categorical, etc.
- The exponential family, while not the most general parametric family, is one of the easiest to work with and captures a variety of different distributions

Continuous Bethe Approximation

- Recall that, from the nonnegativity of the KL-divergence

$$\log Z \geq H(q) + \sum_C \int q_C(x_C) \log \psi_C(x_C)\, dx_C$$

for any probability distribution q

- We can make the same approximations that we did in the discrete case to approximate $Z(\theta)$ in the continuous case

Continuous Bethe Approximation

$$\max_{\tau \in \mathbf{T}} H_B(\tau) + \sum_C \int \tau_C(x_C) \log \psi_c(x_c)\, dx_C$$

where

$$H_B(\tau) = -\sum_{i \in V} \int \tau_i(x_i) \log \tau_i(x_i)\, dx_i - \sum_C \int \tau_C(x_C) \log \frac{\tau_C(x_C)}{\prod_{i \in C} \tau_i(x_i)} dx_C$$

and T is a vector of locally consistent marginals

- This approximation is exact on trees

Continuous Belief Propagation

$$p(x) = \frac{1}{Z} \prod_{i \in V} \phi_i(x_i) \prod_{(i,j) \in E} \psi_{ij}(x_i, x_j)$$

- The messages passed by belief propagation are

$$m_{ij}(x_j) = \int \phi_i(x_i) \psi_{ij}(x_i, x_j) \prod_{k \in N(i) \setminus k} m_{ki}(x_i)\, dx_i$$

- Depending on the functional form of the potential functions, the message update may not have a closed form solution
  – We can't necessarily compute the correct marginal distributions/partition function even in the case of a tree!

Gaussian Belief Propagation

- When $p(x)$ is a multivariate normal distribution, the message updates can be computed in closed form
  - In this case, max-product and sum-product are equivalent
  - Note that computing the mode of a multivariate normal is equivalent to solving a linear system of equations
  - Called Gaussian belief propagation or GaBP
  - Does not converge for all multivariate normal
    - The messages can have a non-positive definite inverse covariance matrix

Properties of Exponential Families

- Exponential families are
  - Closed under multiplication
  - Not closed under marginalization
    - Easy to get mixtures of Gaussians when a model has both discrete and continuous variables
      - Let $p(x, y)$ be such that $x \in \mathbb{R}^n$ and $y \in \{1, \dots, k\}$ such that $p(x|y)$ is normally distributed and $p(y)$ is multinomially distributed
      - $p(x)$ is then a Gaussian mixture (mixtures of exponential family distributions are not generally in the exponential family)

Properties of Exponential Families

- Derivatives of the log-partition function correspond to expectations of the sufficient statistics

$$\nabla_\theta \log Z(\theta) = \int p(x|\theta) T(x) dx$$

- So do second derivatives

$$\frac{\partial^2}{\partial \theta_k \partial \theta_l} \log Z(\theta) =$$

$$\int p(x|\theta) T(x)_k T(x)_l dx - \left(\int p(x|\theta) T(x)_k dx \right) \left(\int p(x|\theta) T(x)_l dx \right)$$

Mean Parameters

- Exponential family distributions can be equivalently characterized in terms of their mean parameters

- Consider the set of all vectors μ such that

$$\mu_k = \int q(x)\phi(x)_k dx$$

for some probability distribution $q(x)$

- If the representation is minimal, then every collection of mean parameters can be realized (perhaps as a limit) by some exponential family
  - This characterization is unique

KL-Divergence and Exponential Families

- Minimizing KL divergence is equivalent to "moment matching"
- Let $q(x|\theta) = \mathrm{h(x)} \cdot \exp(\langle\theta, \phi(x)\rangle - \log Z(\theta))$ and let $p(x)$ be an arbitrary distribution

$$d(p||q) = \int p(x) \log \frac{p(x)}{q(x|\theta)} dx$$

- This KL divergence is minimized when

$$\int p(x)\phi(x)_k dx = \int q(x|\theta)\phi(x)_k dx$$

Expectation Propagation

- Key idea: given $p(x) = \frac{1}{Z}\prod_C \psi_C(x_C)$ approximate it by a simpler distribution $p(x) \approx \tilde{p}(x) = \frac{1}{\tilde{Z}}\prod_C \tilde{\psi}_C(x_C)$

- We could just replace each factor with a member of some exponential family that best describes it, but this can result in a poor approximation unless each ψ_C is essentially a member of the exponential family already

- Instead, we construct the approximating distribution by performing a series of optimizations

Expectation Propagation

- Input $p(x) = \frac{1}{Z} \prod_C \psi_C(x_C)$
- Initialize the approximate distribution $\tilde{p}(x) = \frac{1}{\tilde{Z}} \prod_C \tilde{\psi}_C(x_C)$ so that each $\tilde{\psi}_C(x_C)$ is a member of some exponential family
- Repeat until convergence
  - For each C
    - Let $q(x) = \frac{\tilde{p}(x)}{\tilde{\psi}_C(x_C)} \psi_C(x_C)$
    - Set $\tilde{p}(x) = \operatorname{argmin}_{q'} d(q||q')$ where the minimization is over all exponential families q' of the chosen form

Expectation Propagation

- EP over fully factorized exponential family distributions maximizes the Bethe free energy subject to the following moment matching conditions (instead of the marginalization conditions)

$$\int \tau_i(x_i)\phi_i(x_i)dx_i = \int \tau_C(x_C)\phi_i(x_i)\, dx_C$$

where ϕ_i is a vector of sufficient statistics

Expectation Propagation

- Maximizing the Bethe free energy subject to these moment matching constraints is equivalent to a form of belief propagation where the beliefs are projected onto a set of allowable marginal distributions (e.g., those in a specific exponential family)
- This is the approach that is often used to handle continuous distributions in practice
- Other methods include discretization/sampling methods that make use of BP in a discrete setting