A Variational Principle for Graphical Models

[Martin J. Wainwright and Michael I. Jordan (2005)]

Introduction

- Overview of graphical models
- Exponential families in more generality
- A general variational representation for inference
- Exact inference in variational form

Introduction

- Variational methods provides an alternative approach to computing approximate marginal probabilities and expectations in graphical models
- Examples:
    - sum-product algorithm [Yedidia et al., 2001; McEliece et al., 1998]
    - mean-field algorithm [Jordan et al., 1999; Zhang, 1996]
- Goal of my project:
    - to give a mathematically precise and computationally-oriented meaning to the term "variational" in the setting of graphical models
    - to formulate the optimization problem over a finite-dimensional set M of *realizable mean parameters*

Overview of graphical models

- A graph G = (V,E) is formed by a collection of vertices V, and a collection of edges E
- Associated with each vertex s ∈ V is a random variable x_s taking values in some set X_s
- For any subset A of the vertex set V, we define $x_A := \{x_s \mid s \in A\}$

- Directed graphical model

$$p(\mathbf{x}) = \prod_{s \in V} p(x_s \mid x_{\pi(s)}). \tag{11.1}$$

- Undirected graphical model (MRF)

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{C} \psi_C(x_C), \tag{11.2}$$

Overview of graphical models

- Inference problems and exact algorithms

  (a) computing the likelihood.

  (b) computing the marginal distribution $p(x_A)$ over a particular subset $A \subset V$ of nodes.

  (c) computing the conditional distribution $p(x_A \mid x_B)$, for disjoint subsets A and B, where $A \cup B$ is in general a proper subset of V.

  (d) computing a mode of the density (i.e., an element $\widehat{\mathbf{x}}$ in the set $\arg\max_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x})$).

- Sum-product algorithm and Max-product algorithm

Graphical models in exponential form

- Maximum entropy
  - Given a collection of functions $\phi_\alpha\colon \mathcal{X}^n \rightarrow \mathbb{R}$, suppose that we have observed their expected values

$$\mathbb{E}[\phi_\alpha(\boldsymbol{x})] = \mu_\alpha \ \textit{for all } \alpha \in I \tag{11.11}$$

    - Goal: infer a full probability distribution
    - Let P denote the set of all probability distributions p. Since there are (in general) many distributions p ∈ P that are consistent with the observations (11.11), we need a principled method for choosing among them
  - *The principle of maximum entropy*
    - Choose the distribution p_{ME} such that its *entropy*

$$H(p) := -\sum_{\mathbf{x}\in\mathcal{X}^n} p(\mathbf{x}) \log p(\mathbf{x})$$

    is maximized

Graphical models in exponential form

- Maximum entropy
  - *The principle of maximum entropy*
    - More formally,

$$p_{ME} := \arg\max_{p \in \mathcal{P}} H(p) \qquad \text{subject to constraints (11.11).} \qquad (11.12)$$

    - Intuition: choose the distribution with maximal uncertainty while remaining faithful to the data
    - Presuming that problem (11.12) is feasible, we can show using a Lagrangian formulation that its optimal solution takes the form

$$p(\mathbf{x};\theta) \;\propto\; \exp\Big\{\sum_{\alpha \in \mathcal{I}} \theta_\theta \phi_\alpha(\mathbf{x})\Big\}, \qquad (11.13)$$

      - $\theta \in \mathbb{R}_d$ is known as the *canonical parameter*
      - $\phi = \{\phi_\alpha \mid \alpha \in I\}$ are known as *sufficient statistics*

Graphical models in exponential form

- Exponential families in more generality
  - The *exponential family* associated with ϕ consists of the following parameterized collection of density functions:

$$p(\mathbf{x};\theta) = \exp\left\{\langle\theta, \boldsymbol{\phi}(\mathbf{x})\rangle - A(\theta)\right\}. \qquad (11.14)$$

    - The quantity A, known as the *log partition function* or *cumulant generating function*, is defined by the integral:

$$A(\theta) = \log \int_{\mathcal{X}^n} \exp\langle\theta, \boldsymbol{\phi}(\mathbf{x})\rangle\, \boldsymbol{\nu}(d\mathbf{x}). \qquad (11.15)$$

      - *v*: a fixed based measure, typically counting measure (discrete), or Lebesgue measure (e.g. Gaussian families)
      - ⟨a, b⟩: the ordinary Euclidean inner product
  - *Lemma 11.1* The cumulant generating function A is convex in terms of θ. It is infinitely differentiable on Θ, and its derivatives correspond to cumulants.
    - As an important special case, the first derivatives of A take the form

$$\frac{\partial A}{\partial \theta_\alpha} = \int_{\mathcal{X}^n} \phi_\alpha(\mathbf{x}) p(\mathbf{x};\theta)\boldsymbol{\nu}(d\mathbf{x}) = \mathbb{E}_\theta[\phi_\alpha(\mathbf{x})], \qquad (11.17)$$

Exact Variational Principle

- Re-phrase inference problems in the language of exponential families

(a) computing the cumulant generating function $A(\theta)$

- The problem of computing the cumulant generating function arises in a variety of signal processing problems, including likelihood ratio tests and parameter estimation.

(b) computing the vector of mean parameters $\mu := \mathbb{E}_\theta[\boldsymbol{\phi}(\mathbf{x})]$

- The computation of mean parameters is also fundamental, and takes different forms depending on the underlying graphical model.
- It corresponds to computing means and covariances in the Gaussian case, whereas for a multinomial MRF it corresponds to computing marginal distributions

Exact Variational Principle

- Conjugate duality [Rockafellar (1970); Hiriart-Urruty and Lemar´echal (1993)]
  - Associated with any convex function $f\colon \mathbb{R}^d \to \mathbb{R}_*$ a conjugate dual function $f_*\colon \mathbb{R}^d \to \mathbb{R}_*$ is defined as:

$$f^*(y) := \sup_{x\in\mathbb{R}^d} \{\langle y, x\rangle - f(x)\}. \tag{11.24}$$

  - This definition illustrates the concept of a variational definition: the function value f^* is specified as the solution of an optimization problem parameterized by the vector $y \in \mathbb{R}^d$

  - Meeting certain technical conditions, taking the dual twice recovers the original function

$$f(x) = \sup_{y\in\mathbb{R}^d} \{\langle x, y\rangle - f^*(y)\}. \tag{11.25}$$

  - Goal: apply conjugacy to the cumulant generating function A associated with an exponential family, as defined in equation (11.15)

$$A^*(\mu) := \sup_{\theta\in\Theta} \{\langle \theta, \mu\rangle - A(\theta)\}, \tag{11.26}$$

Exact Variational Principle

- Conjugate duality [Rockafellar (1970); Hiriart-Urruty and Lemar´echal (1993)]
  - *Example 11.7* To illustrate the computation of a dual function, consider a scalar Bernoulli random variable x ∈ {0,1}

$$p(x;\theta) = \exp\{\theta x - A(\theta)\}$$
$$A(\theta) = \log[1 + \exp(\theta)]$$

  - Thus, the variational problem (11.26) defining A^* takes the form:

$$A^*(\mu) = \sup_{\theta \in \mathbb{R}} \{\theta\mu - \log[1 + \exp(\theta)]\}$$
$$\mu = \mathbb{E}_\theta[x]$$

  - Taking derivatives shows that the supremum is attained at the unique θ satisfying logistic relation

$$\theta = \log[\mu/(1-\mu)]$$

  - Substituting back we have

$$A^*(\mu) = \mu \log \mu + (1-\mu)\log(1-\mu)$$

  - Variation form

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\}.$$

Exact Variational Principle

- Sets of realizable mean parameters

  - For a given $\mu \in \mathbb{R}^d$, consider the optimization problem of equation (11.26):

$$A^*(\mu) := \sup_{\theta \in \Theta} \{\langle \theta, \mu \rangle - A(\theta)\}, \tag{11.26}$$

  - Take the derivative with respect to θ and set it equal to zero. Doing so yields the zero-gradient condition:

$$\mu \;=\; \nabla A(\theta) \;=\; \mathbb{E}_\theta[\boldsymbol{\phi}(\mathbf{x})], \tag{11.28}$$

  - We now need to determine the set of $\mu \in \mathbb{R}^d$ for which (11.28) has a solution. Observe that any $\mu \in \mathbb{R}^d$ satisfying this equation has a natural interpretation as a *globally realizable mean parameter*

$$\mathcal{M} \;:=\; \{\; \mu \in \mathbb{R}^d \mid \exists \;\, p(\cdot) \;\text{ such that }\; \int \boldsymbol{\phi}(\mathbf{x}) p(\mathbf{x}) \boldsymbol{\nu}(d\mathbf{x}) = \mu\}$$

Exact Variational Principle

- Entropy in terms of mean parameters
  - As expected, the form of the dual function turns out to be closely related to entropy
  - Given a density function p taken with respect to base measure ν, its entropy is given by

$$H(p) = -\int_{\mathcal{X}^n} p(\mathbf{x}) \log [p(\mathbf{x})]\, \boldsymbol{\nu}(d\mathbf{x}) \;=\; -\mathbb{E}_p[\log p(\mathbf{x})]. \tag{11.33}$$

  - Suppose that μ belongs to the interior of the set of realizable mean parameters

$$\mathbb{E}_{\theta(\mu)}[\boldsymbol{\phi}(\mathbf{x})] = \mu$$

  - Substituting this relation into the definition (11.26) of the dual function yields

$$A^*(\mu) \;=\; \langle \mu,\, \theta(\mu)\rangle - A(\theta(\mu)) \;=\; \mathbb{E}_{\theta(\mu)}\left[\log p(\mathbf{x}; \theta(\mu))\right]$$

    which we recognize as the negative entropy
  - Summarizing our development:

$$A^*(\mu) \;=\; \max_{p \in \mathcal{P}} H(p) \quad \text{such that } \mathbb{E}_p[\phi_\alpha(\mathbf{x})] = \mu_\alpha \text{ for all } \alpha \in \mathcal{I}. \tag{11.36}$$

Exact Variational Principle

- Entropy in terms of mean parameters
  - Given the form (11.35) of the dual function, we can now use the conjugate dual relation (11.25) to express A in terms of an optimization problem involving its dual function and the mean parameters:

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\}. \tag{11.37}$$

  - Note that the optimization is restricted to the set M of globally realizable mean parameters
  - In addition to representing the value A(θ) of the cumulant generating function, the nature of our dual construction ensures that the optimum is always attained at the vector of mean parameters

$$\mu = \mathbb{E}_\theta[\boldsymbol{\phi}(\mathbf{x})].$$

  - Consequently, solving this optimization problem yields both the value of the cumulant generating function as well as the full set of mean parameters.

Discussion & Limitation

- Exact variational principle is intractable to solve:
  The constraint set $\mathcal{M}$ is hard to characterize;
  A^* typically lacks an explicit form
  - a broad class of methods for approximate inference are based on the use of approximations to $\mathcal{M}$ and A^*

- Deal exclusively with exponential family models
  - one approach to exploiting variational ideas for nonparametric models is through exponential family approximations of nonparametric distributions [Blei and Jordan (2004)]

- Variational methods in general turn inference into an optimization problem via exponential families and convex duality

- A broad class of message-passing algorithms (mean field updates, sum-product, max-product), can all be understood as solving either exact or approximate versions of a variational principle for graphical models

Next step

- Provide derivation of approximate inference
- Analysis of sum-product and study of Bethe entropy approximation:
  - want to show that sum-product can be derived from a variational perspective
- How some of the examples discussed in class could be solved as optimization problem