CS 2243 Fall 2024 Scribes: <names>
11/4 Based on notes by: Aayush Karan

Lecture 17: Intro to Graphical Models and Variational Inference

1 Physics, inference, and sampling

Many problems in machine learning and statistics can be framed as the following inference setting. Suppose we have a prior distribution $\mathcal{D}$ where we draw a signal $X \sim \mathcal{D}$. The signal undergoes some kind of noisy channel, and the observed "noisy measurement" is Y. The goal is to understand how the knowledge of Y induces a posterior belief on the original X. For example,

- **Learning neural networks/teacher-student setting:** If X denotes the weights of a neural network F sampled from some complicated prior, Y is the set $\{(x_i, y_i) : x_i \sim \mathcal{N}(0, \mathrm{Id}_d), y_i = F(x_i)\}$ of input and output pairs of F, where the input is sampled from a multivariate Gaussian.
- **Denoising:** If X is an image from a distribution $\mathcal{D}$, Y is the image with added noise terms to the pixels. Can we recover the original X? This setting turns out to be the core primitive of diffusion models for generative modeling.

From Bayes Rule, we can immediately express the posterior on X given Y:

$$P(X = x|Y = y) = \frac{P(Y = y|X = x)P(X = x)}{P(Y = y)}. \tag{1}$$

$P(X = x)$ is our prior on X, $P(Y = y|X = x)$ is the likelihood given a model of the noisy channel, and $P(Y = y)$ is the normalization constant, also known as the *partition function/evidence*. We can write this in a suggestive manner:

$$P(X = x|Y = y) = \frac{1}{Z} \cdot \exp\left(-\mathcal{E}(x)\right), \tag{2}$$

where the *energy function* $\mathcal{E}(x) = -\log\left(P(Y = y|X = x)P(X = x)\right)$. This is highly reminiscent of the *Boltzmann energy distribution* from statistical physics. Finding the minimal energy corresponds to finding the signal X at which the posterior is maximized.
Further details on the background used in this and subsequent sections can be found in [MM09], [KZ22], [Mon11], [WJ08], and [Mon14].

2 Undirected graphical models with pairwise interactions

Consider the energy model given by

$$\mathcal{E}(x) = -\sum_{(i,j)\in\mathcal{F}} \log \psi_{ij}(x_i, x_j), \tag{3}$$

with $\mathcal{F}$ some family of edges/subset of $[n] \times [n]$, inputs $x \in \{\pm 1\}^n$, and *compatibility functions* $\psi_{ij} : \{\pm 1\}^2 \to \mathbb{R}$. The posterior distribution is also known as the *Gibbs measure* μ over $\{\pm 1\}^n$, where $\mu = \frac{1}{Z} \cdot \exp\left(-\mathcal{E}(x)\right)$ and the partition function $Z = \sum_{x\in\{\pm 1\}^n} \exp\left(-\mathcal{E}(x)\right)$.

2.1 Ising Model

Suppose we have the *compatibility functions*

$$\psi_{ij} = \exp\left(-\beta x_i x_j A_{ij}\right) \tag{4}$$

for some matrix A such that $A_i i = 0$ (WLOG, since constants can be absorbed in the normalization) and A is symmetric. In this setting β can be interpreted as the *inverse temperature* and A as the *Hamiltonian* or *interaction matrix* of the system. It follows from the definitions above that the Gibbs measure is given by

$$\mu \propto \exp\left(-\frac{\beta}{2} x^\top A x\right). \tag{5}$$

For some limiting cases with the inverse temperature β, notice the high temperature limit $\beta \to 0$ corresponds to a uniform distribution over $\{\pm 1\}^n$, while the low temperature limit $\beta \to \infty$ corresponds to the uniform distribution over $x \in \{\pm 1\}^n$ such that $x^\top A x$ is maximal.

2.2 Markov Property

We can also interpret A as the (negative) adjacency matrix of a weighted graph G. Given $i \in [n]$, define $\partial_i = \{j \in [n] : (i,j) \in \mathcal{F}\}$ for some edgeset $\mathcal{F}$ corresponding to A; i.e., the neighborhood of i where $A_{ij} \neq 0$.

The *Markov property* states that for $S \subseteq [n]$ such that $[n]/S$ consists of two disjoint pieces in G, then conditioning on some assignment of spins/x values on S, the resulting marginals on the two disjoint pieces are independent. For example, if $S = \partial_i$, then G is broken into i and everything not connected to i. Then the marginal on i conditioned on some assignment to ∂_i is independent of the rest of the graph, i.e.,

$$P(x_i = \sigma | x_{\partial_i} = s) \propto \exp\left(-\beta \sum_{j \in \partial_i} A_{ij} s_j \sigma\right). \tag{6}$$

3 Variational Inference

In the types of inference problems detailed above, there are two fundamental tasks: computing the normalization constant/partition function Z, and sampling from the Gibbs measure μ. For discrete but exponentially sized domains, computing Z is incredibly difficult. Take the example where $\psi_{ij}(x_i, x_j) = 1[x_i x_j = 0]$. Then no two adjacent x_i, x_j can both be 1. It follows that Z counts the number of independent sets in G, which is #P complete (very hard).

Since we cannot compute Z or sample from μ exactly, we must use approximative methods, which include MCMC (Markov Chain Monte Carlo), variational inference, or diffusion models. For now, we focus on variational inference, which aims to approximate μ by metric of distance from a family $\mathcal{P}$ of simpler distributions (e.g. product distributions). Then the goal is the following:

$$\min_{\nu \in \mathcal{P}} KL(\nu || \mu). \tag{7}$$

Of course, if $\mathcal{P}$ is the family of all distributions, then the minimizer is simply $\nu = \mu$ by Gibbs' inequality on the KL distance. While SOS pseudodistributions expand the space of distributions, variational inference diminishes it. The KL optimizer is hard to evaluate, but we can get around it. Notice

$$\mathbb{E}_{x \in \nu}\left[\log \frac{\nu(x)}{\mu(x)}\right] = \mathbb{E}_{x \in \nu}\left[\log \frac{\nu(x)}{\frac{1}{Z} \cdot \exp(-\mathcal{E}(x))}\right] = \mathbb{E}_\nu[\log \nu] + \mathbb{E}_\nu[\mathcal{E}(x)] - \log \frac{1}{Z} \tag{8}$$

so

$$KL(\nu || \mu) = \mathbb{E}_\nu[\log \nu] + \mathbb{E}_\nu[\mathcal{E}(x)] - \log \frac{1}{Z}. \tag{9}$$

For the purposes of minimization over ν, the last term is independent, so only the first two terms – the negative entropy and average energy respectively – are important. The sum can be written as the Gibbs free energy functional

$$G[\nu] = \mathbb{E}_\nu[\log \nu] + \mathbb{E}_\nu[\mathcal{E}(x)]. \tag{10}$$

For the Ising model,

$$G[\nu] = \mathbb{E}_\nu[\frac{\beta}{2} x^\top A x] - H(\nu). \tag{11}$$

When β is small, maximizing entropy minimizes G. When β is large, minimizing average energy minimizes G. We will see how belief propagation is a heuristic for minimizing $G[\nu]$ in the next lecture. Belief propagation is a natural dynamic programming algorithm that gives an exact answer when the graph is a tree and finds stationary points of the Lagrangian dual of $G[\nu]$ called the Bethe free energy.

References

[KZ22] Florent Krzakala and Lenka Zdeborová. *Statistical Physics Methods in Optimization and Machine Learning*. 2022. An Introduction to Replica, Cavity & Message-Passing techniques.

[MM09] Marc Mézard and Andrea Montanari. *Information, Physics and Computation*. 2009.

[Mon11] Andrea Montanari. *Inference in Graphical Models*. 2011. Lecture Notes for Stat 375.

[Mon14] Andrea Montanari. *Statistical Mechanics and Algorithms on Sparse and Random Graphs*. 2014.

[WJ08] Martin J. Wainwright and Michael I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Foundations and Trends in Machine Learning, 2008.