B. Lecture #9. Inference & Learning with Belief Propagation

This lecture will continue the thread of the previous lecture. We will mainly focus on the so-called Belief Propagation, related theory and techniques. In addition to discussing inference with Belief Propagation we will also have a brief discussions (pointers) to respective inverse problem – learning with Graphical Models.

1. Bethe Free Energy & Belief Propagation

Bethe-Peierls or Belief Propagation (we will use the same abbreviation BP for both) usually provides a good approximation, which is provably exact in some important cases (when graph in the underlying GM is a tree). It also provides an empirically good approximation for a very broad family of problems stated on loopy graphs. See the original paper [9], a comprehensive review[10], and lecture notes `http://www.eecs.berkeley.edu/~wainwrig/Talks/A_GraphModel_Tutorial`, for an advanced/additional reading.

Instead of Eq. (III.14) one uses the following BP substitution

$$b(\sigma) \to b_{bp}(\sigma) = \frac{\prod_a b_a(\sigma_a)}{\prod_i (b_i(\sigma_i))^{q_i-1}} \tag{III.19}$$

$$\forall a \in \mathcal{V}_f, \;\; \forall \sigma_a : \quad b_a(\sigma_a) \geq 0 \tag{III.20}$$

$$\forall i \in \mathcal{V}_n, \;\; \forall a \sim i : \quad b_i(\sigma_i) = \sum_{\sigma_a \setminus \sigma_i} b_a(\sigma_a) \tag{III.21}$$

$$\forall i \in \mathcal{V}_n : \quad \sum_{\sigma_i} b_i(\sigma_i) = 1. \tag{III.22}$$

where q_i stands for degree of node i. The physical meaning of the factor $q_i - 1$ on the rhs of Eq. (III.35) is straightforward: by placing beliefs associated with the factor-nodes connected by an edge with a node, i, we over-count contributions of an individual variable q_i times and thus the denominator term in Eq. (III.35) comes as a correction for this over-counting.

Substitution of Eqs. (III.35) into Eq. (III.13) results in what is called Bethe (of BP) Free Energy

$$\mathcal{F}_{bp} \doteq E_{bp} - S_{bp}, \tag{III.23}$$

$$E_{bp} \doteq \sum_a \sum_{\sigma_a} b_a(\sigma_a) \log f_a(\sigma_a) \tag{III.24}$$

$$\mathcal{S}_{bp} = \sum_a \sum_{\sigma_a} b_a(\sigma_a) \log b_a(\sigma_a) - \sum_i \sum_{\sigma_i} (q_i - 1) b_i(\sigma_i) \log b_i(\sigma_i), \tag{III.25}$$

where E_{bp} is the so-called self-energy (physics jargon) and S_{bp} is the BP-entropy (this name should be clear in view of what we have discussed about entropy so far). Thus the BP version of the KL-divergence minimization becomes

$$\arg\min_{b_a, b_i} \mathcal{F}_{bp}\Big|_{\text{Eqs. (III.36,III.21,III.22)}}, \tag{III.26}$$

$$\min_{b_a, b_i} \mathcal{F}_{bp}\Big|_{\text{Eqs. (III.36,III.21,III.22)}} \tag{III.27}$$

Question: Is $\mathcal{F}_{bp}$ a convex function (of its arguments)? [Not always, however for some graphs and/or some factor functions the convexity holds.]

The ML (zero temperature) version of Eq. (III.26) results from the following optimization

$$\min_{b_a, b_i} E_{bp}\Big|_{\text{Eqs. (III.36,III.21,III.22)}} \tag{III.28}$$

Note the optimization is a Linear Programming (LP) — minimizing linear objective over set of linear constraints.

Let us restate Eq. (III.26) as an unconditional optimization. We use the standard method of Lagrangian multipliers to achieve it. The resulting Lagrangian is

$$\mathcal{L}_{bp}(b,\eta,\lambda) \doteq \sum_a \sum_{\sigma_a} b_a(\sigma_a)\log f_a(\sigma_a) - \sum_a \sum_{\sigma_a} b_a(\sigma_a)\log b_a(\sigma_a) + \sum_i \sum_{\sigma_i} (q_i-1) b_i(\sigma_i)\log b_i(\sigma_i)$$
$$- \sum_i \sum_{a\sim i} \sum_{\sigma_i} \eta_{ia}(\sigma_i)\left(b_i(\sigma_i) - \sum_{\sigma_a\setminus\sigma_i} b_a(\sigma_a)\right) + \sum_i \lambda_i\left(\sum_{\sigma_i} b_i(\sigma_i) - 1\right), \tag{III.29}$$

where η and λ are the dual (Lagrangian) variables associated with the conditions Eqs. (III.21,III.22) respectively. Then Eq. (III.26) become the following min-max problem

$$\min_b \max_{\eta,\lambda} \mathcal{L}_{bp}(b,\eta,\lambda). \tag{III.30}$$

Changing the order of optimizations in Eq. (III.30) and then minimizing over η one arrives at the following expressions for the beliefs via messages (check the derivation details)

$$\forall a,\ \forall \sigma_a: \quad b_a(\sigma_a) \sim f_a(\sigma_a)\exp\left(\sum_{i\sim a}\eta_{ia}(\sigma_i)\right) \doteq f_a(\sigma_a)\prod_{i\sim a} n_{i\to a}(\sigma_i) \doteq f_a(\sigma_a)\prod_{i\sim a}\prod_{b\sim i}^{b\neq a} m_{b\to i}(\sigma_i) \tag{III.31}$$

$$\forall i,\ \forall \sigma_i: \quad b_i(\sigma_i) \sim \exp\left(\frac{\sum_{a\sim i}\eta_{ia}(\sigma_i)}{q_i-1}\right) \doteq \prod_{a\sim i} m_{a\to i}(\sigma_i), \tag{III.32}$$

where, as usual, $\sim$ for beliefs means equality up to a constant which guarantees that the sum of respective beliefs is unity, and we have also introduce the auxiliary variables , m and n, called messages, related to the Lagrangian multipliers η as follows

$$\forall i,\ \forall a\sim i: \quad n_{i\to a}(\sigma_i) \doteq \exp\left(\eta_{ia}(\sigma_i)\right) \tag{III.33}$$

$$\forall a,\ \forall i\sim a: \quad m_{a\to i}(\sigma_i) \doteq \exp\left(\frac{\eta_{ia}(\sigma_i)}{q_i-1}\right). \tag{III.34}$$

Combining Eqs. (III.31,III.32,III.33,III.34) with Eq. (III.21) results in the following BP-equations stated in terms of the message variables

$$\forall i,\ \forall a\sim i,\ \forall \sigma_i: \quad n_{i\to a}(\sigma_i) = \prod_{b\sim i}^{b\neq a} m_{a\to i}(\sigma_i) \tag{III.35}$$

$$\forall a,\ \forall i\sim a,\ \forall \sigma_i: \quad m_{a\to i}(\sigma_i) = \sum_{\sigma_a\setminus\sigma_i} f_a(\sigma_a)\prod_{j\sim a}^{j\neq i} n_{j\to a}(\sigma_j). \tag{III.36}$$

Note that if the Bethe Free Energy (III.23) is non-convex there may be multiple fixed points of the Eqs. (III.35,III.36). The following iterative, so called Message Passing (MP), algorithm (5) is used to find a fixed point solution of the BP Eqs. (III.35,III.36)

Algorithm 5 Message Passing, Sum-Product Algorithm [factor graph representation]

Input: The graph. The factors.

1: $\forall i,\ \forall a\sim i, \forall \sigma_i: \quad m_{a\to i} = 1$ [initialize variable-to-factor messages]
2: $\forall a,\ \forall i\sim a, \forall \sigma_i: \quad n_{i\to 1} = 1$ [initialize factor-to-variable messages]
3: **loop**Till convergence within an error [or proceed with a fixed number of iterations]
4: $\quad \forall i,\ \forall a\sim i,\ \forall \sigma_i: \quad n_{i\to a}(\sigma_i) \leftarrow \prod_{b\sim i}^{b\neq a} m_{a\to i}(\sigma_i)$
5: $\quad \forall a,\ \forall i\sim a,\ \forall \sigma_i: \quad m_{a\to i}(\sigma_i) \leftarrow \sum_{\sigma_a\setminus\sigma_i} f_a(\sigma_a)\prod_{j\sim a}^{j\neq i} n_{j\to a}(\sigma_j)$
6: **end loop**

Exercise: Derive the $T=0$ version of the aforementioned (see previous exercise) message-passing equations. A hint: the iterative equations should contain alternating min- and sum- steps — thus the name min-sum algorithm.

Exercise (bonus): Study performance of the message-passing algorithm on example of a small code decoding, for example check (this is a student midterm paper !) `http://www.people.fas.harvard.edu/~rpoddar/Papers/ldpc.pdf` for discussion of decoding of a binary $(3,6)$ code over the Binary-Erasure Channel (BEC). Show how BP decodes and contrast the BP decoding against the MAP decoding. What is the (best) complexity of the MAP decoder for a code over the BEC channel. [Hint: Use Gaussian Elimination over GL(2).]

3. Sufficient Statistics

So far we have been discussing direct (inference) GM problem. In the remainder of this lecture we will briefly talk about inverse problems. This subject will also be discussed (on example of the tree) in the following recitation.

Stated casually - the inverse problem is about 'learning' GM from data/samples. Think about the two room setting. In one room a GM is known and many samples are generated. The samples, but not GM (!!!), are passed to the second room. The task becomes to reconstruct GM from samples.

The first question we should ask is if this is possible in principle, even if we have an infinite number of samples. A very powerful notion of *sufficient statics* helps to answer this question.

Consider the Ising model (not the first time in this course) using a little bit different notations then before

$$P(\sigma) = \frac{1}{Z(\theta)} \exp\left\{\sum_{i\in V} \theta_i\sigma_i + \sum_{\{i,j\}\in E} \theta_{ij}\sigma_i\sigma_j\right\} = \exp\{\theta^T\phi(\sigma) - \log Z(\theta)\}, \tag{III.37}$$

where $\sigma_i \in \{-1, 1\}$ and the *partition function* $Z(\theta)$ serves to normalize the probability distribution. In fact, Eq. (III.37) describes what is called an *exponential family* - emphasizing 'exponential' dependence on the factors θ.

Exercise: Show that any pairwise GM over binary variables can be represented as an Ising model.

Consider collection of all first and second moments (but only these two) of the spin variables, $\mu^{(1)} \doteq (\mu_i = \mathbb{E}[\sigma_i], i \in V)$ and $\mu^{(2)} \doteq (\mu_{ij} = \mathbb{E}[\sigma_i\sigma_j], \{i,j\} \in E)$. The *sufficient statistics* statement is that to reconstruct θ, fully defining the GM, it is *sufficient* to know $\mu^{(1)}$ and $\mu^{(2)}$.

4. Maximum-Likelihood Estimation/Learning of GM

Let us turn the *sufficiency* into a constructive statement – the *Maximum-likelihood estimation* over an exponential family.

First, notice that (according to the definition of μ)

$$\forall i: \quad \partial_{\theta_i} \log Z(\theta) = -\mu_i, \quad \forall i,j: \quad \partial_{\theta_{ij}} \log Z(\theta) = -\mu_{ij}. \tag{III.38}$$

This leads to the following statement: if we know how to compute log-partition function for any values of θ - reconstructing 'correct' θ is a convex optimization problem (over θ):

$$\theta^* = \arg\max_\theta \{\mu^T\theta - \log Z(\theta)\} \tag{III.39}$$

If P represents the empirical distribution of a set of independent identically-distributed (i.i.d.) samples $\{\sigma^{(s)}, s = 1, \dots, S\}$ then μ are the corresponding empirical moments, e.g. $\mu_{ij} = \frac{1}{S}\sum_s \sigma_i^{(s)}\sigma_j^{(s)}$.

General (concluding) Remarks about GM Learning. The ML parameter Estimation (III.39) is the best we can do. It is fundamental for the task of Machine Learning, and in fact it generalizes beyond the case of the Ising model.

Unfortunately, there are only very few nontrivial cases when the partition function can be calculated efficiently for any values of θ (or parametrization parameters if we work with more general class of GM than described by the Ising models).

Therefore, to make the task of parameter estimation practical one needs to rely on one of the following approaches:

- Limit consideration to the class of functions for which computation of the partition function can be done efficiently for any values of the parameters. We will discuss such case in the recitation – this will be the so-called tree (Chow-Lou) learning. (In fact, the partition function can also be computed efficiently in the case of the planar Ising - one of the suggested projects covers this advance subject.)
- Relay on approximations, e.g. such as MF and BP (but there are also other).
- There exists a very innovative new approach - which allows to learn GM efficiently however using more information than suggested by the notion of the *sufficient statistics*. How one of the scientists contributing to this line of research put it – 'the sufficient statistics is not sufficient'. This is a fascinating novel subjects, which is however beyond the scope of this course.

5. Recitation: Direct (Inference) and Inverse (Learning) Problems over Trees.

IV. THEME #4: STOCHASTIC MODELING & OPTIMIZATION

A. Lecture #10. Space-time Continuous Stochastic Processes

In this lecture we discuss stochastic dynamics of continuous variables governed by the Langevin equation. We discuss how to derive the so-called Fokker-Planck equations, describing temporal evolution of the probability of a state. We then go into some additional details for a basic example of stochastic dynamics in a free space (no potential) describing the Brownian motion where Fokker-Planck equations becomes the diffusion equation.

1. Langevin equation in continuous time and discrete time

Stochastic process in 1d is described in the continuous-time and discrete-time forms as follows

$$\dot{x} = -F(x) + \sqrt{D}\xi(t), \quad \langle \xi(t)\rangle = 0, \quad \langle \xi(t_1)\xi(t_2)\rangle = \delta(t_1 - t_2) \tag{IV.1}$$

$$x_{n+1} - x_n = -\Delta F(x_n) + \sqrt{D\Delta}\zeta(t_n), \quad \langle \zeta(t_n)\rangle = 0, \quad \langle \zeta(t_n)\zeta(t_k)\rangle = \delta_{kn} \tag{IV.2}$$

The first and second terms on the rhs of Eq. (IV.1) stand for the force and the "noise" respectively. The noise is considered independent at each time step. These equations, also called Langevin equations, describe evolution of a "particle" positioned at $x \in \mathbb{R}$. The two terms on the rhs of Eq. (IV.1)) correspond to deterministic advancement of the particle (also dependent on its position at the previous time step) and, respectively, on a random correction/increment. The random correction models uncertainty of the environment the particles moves through (we can also think of it as representing random kicks by other "invisible" particles). The uncertainty is represented in a probabilistic way – therefore we will be talking about the probability distribution function of paths, i.e. trajectories of the particle.

The square root on the rhs of Eq. (IV.2) may seem mysterious, let us clarify its origin on basic (no force/potential) example of $F(x) = 0$. (This will be the running example through out this lecture.) In this case the Langevin equation describes Brownian motion. Direct integration of the linear equation with the inhomogeneous source results in this case in

$$\forall t \geq 0: \quad x(t) = \int_0^t dt' \xi(t'), \tag{IV.3}$$

$$\forall t \geq 0 \quad \langle x^2(t)\rangle = \int_0^t dt_1 \int_0^t dt_2 D\delta(t_1 - t_2) = D\int_0^t dt_1 = Dt, \tag{IV.4}$$

where we have also accounted that $x(0) = 0$. Infinitesimal version of Eq. (IV.5) is

$$\delta x = \sqrt{D\Delta}, \tag{IV.5}$$

which is thus the Brownian (no force) version of Eq. (IV.2).

2. From the Langevin Equation to the Path Integral

The Langevin equation can also be viewed as relating the change in $x(t)$, i.e. dynamic of interest, to stochastic dynamic of the δ-correlated source $\zeta(t_n) = \zeta_n$ characterized by the Probability Density Function (PDF)

$$P(\zeta_1, \cdots, \zeta_N) = (2\pi)^{-N/2} \exp\left(-\sum_{n=1}^{N} \frac{\zeta_n^2}{2}\right) \tag{IV.6}$$

Eqs. (IV.1,IV.2,IV.7) are starting points for our further derivations, but they should also be viewed as a way to simulate the Langevin equation on computer by generating many paths at once, i.e. simultaneously. Notice, for completeness, that there are also other ways to simulate the Langevin equation, e.g. through the telegraph process.

Let us express ζ_n via x_n from Eq. (IV.2) and substitute it into Eq. (IV.7)

$$P(\zeta_1, \cdots, \zeta_{N-1}) \to P(x_1, \cdots, x_N) = (2\pi D)^{-(N-1)/2} \exp\left(-\frac{1}{2D\Delta}\sum_{n=1}^{N-1} (x_{n+1} - x_n + \Delta F(x))^2\right) \tag{IV.7}$$

one gets an explicit expression for the measure over a path written in the discretized way. And here is a typical way of how we state it in the continuous form (e.g. as a notational shortcut)

$$P\{x(t)\} \propto \exp\left(-\frac{1}{2D}\int_0^T dt\,(\dot{x}+F(x))^2\right) \tag{IV.8}$$

This object is called (in physics and math) "path integral" and/or Feynmann/Kac integral.

3. From the Path Integral to the Fokker-Planch (through sequential Gaussian integrations)

Probability Density Function of a path is a useful general object. However we may also want to marginalize it thus extracting the marginal PDF for being at the position x_N at the (temporal) step N from the joint probability distribution (of the path), $P(x_1,\cdots,x_N)$, and also from the prior/initial (distribution) $P_1(x_1)$ – both assumed known:

$$P_N(x_N) = \int dx_1\cdots dx_N P(x_1,\cdots,x_N)P_1(x_1). \tag{IV.9}$$

It is convenient to derive relation between $P_N(\cdot)$ and $P_1(\cdot)$ in steps, i.e. through a recurrence, integrating over $dx_1,\cdots,dx_N$ sequentially. Let us proceed analyzing the case of the Brownian motion where, $F=0$. Then the first step of the induction becomes

$$P_2(x_2) = (2\pi D)^{-1/2}\int dx_1 \exp\left(-\frac{1}{2D\Delta}(x_2-x_1)^2\right)P_1(x_1) \tag{IV.10}$$

$$= (2\pi D)^{-1/2}\int d\epsilon \exp\left(-\frac{\epsilon^2}{2D\Delta}\right)P_1(x_2-\epsilon) \tag{IV.11}$$

$$\approx (2\pi D)^{-1/2}\int d\epsilon \exp\left(-\frac{\epsilon^2}{2D\Delta}\right)\left(P_1(x_2)-\epsilon\partial_x P_1(x_2)+\frac{\epsilon^2}{2}\partial_x^2 P_1(x_2)\right) \tag{IV.12}$$

$$= P_1(x_2)+\Delta\frac{D}{2}\partial_x^2 P_1(x_2), \tag{IV.13}$$

where transitioning from Eq. () to Eq. () one makes Taylor expansion in ϵ, also assuming that $\epsilon\sim\sqrt{\Delta}$ and keeping only the leading terms in Δ. The resulting Gaussian integrations are straightforward. We arrive at the discretized (in time) version of the diffusion equation

$$\partial_t P_t(x) = \frac{D}{2}\partial_x^2 P_t(x). \tag{IV.14}$$

Of course it is not surprising that the case of the Brownian motion has resulted in the diffusion equation for the marginal PDF. Restoring the $U(x)$ term (derivation is straightforward) one arrives at the Fokker-Planck equation, generalizing the zero-force diffusion equation

$$\partial_t P_t(x) - \partial_x(U(x)P_t(x)) = \frac{D}{2}\partial_x^2 P_t(x). \tag{IV.15}$$

4. Analysis of the Fokker-Planck Equation: General Features and Examples

Here we only give a very brief and incomplete description on the properties of the distribution which analysis is of a fundamental importance for Statistical Physics. See e.g. [1]. Some selected subjects will also be discussed in the recitations.

The Fokker-Planck equation (IV.15) is linear and deterministic Partial Differential Equation (PDE). It describes continuous in phase space, x, and time, t, evolution/flow of the probability distribution.

Derivation was for a particle moving in 1d, $\mathbb{R}$, but the same ideology and logic extends to higher dimensions, $\mathbb{R}^d$, $d=1,2,\cdots$. There are also extension of this consideration to compact continuous spaces. Thus one can analyze dynamics on a circle, sphere or torus.

Analogs of the Fokker-Planck can be derived and analyzed for more complicated probabilities than just the marginal probability of the state (path integral marginalized to given time). An example – of the "first passage" probability will be given in the recitation.

The temporal evolution is driven by two terms - "diffusion" and "advecton" - the terminology is from fluid mechanics - indeed not only fluids but also probabilities can flow. The flow of probability is in the phase space. The diffusion originates from the stochastic source while advection from the deterministic (but possibly nonlinear) deterministic force.

Linearity of the Fokker-Planck does not imply that it is simpler than the original nonlinear problem. Deriving the Fokker-Planck we made a transition from nonlinear, stochastic but ODE to linear PDE. This type of transition from nonlinear representation of many trajectories to linear probabilistic representation is typical in math/statistics/physics. The linear Fokker-Planck equation can be viewed as the continuous-time, continuous-space version of the discrete-time/discrete space Master equation describing evolution of a (finite dimensional) probability vector in the case of a Markov Chain.

The Fokker-Planck Eq. (IV.15) can be represented in the 'flux' form:

$$\partial_t P_t + \partial_x J_t(x) = 0 \tag{IV.16}$$

where $J_t(x)$ is the flux of probability through the space-state point x at the moment of time t. The fact that the second (flux) term in Eq. (IV.16) has a gradient form, corresponds to the global conservation of probability. Indeed, integrating Eq. (IV.16) over the whole continuous domain of achievable x, and assuming that if the domain is bounded there is no injection (or dissipation) of probability on the boundary, one finds that the integral of the second term is zero (according to the standard Gauss theorem of calculus) and thus, $\partial_t \int dx P_t(x) = 0$. In the steady state, when $\partial_t P_t = 0$ for all x (and not only in the result of integration over the entire domain) the flux is constant - does not depend on x. The case of zero-flux is the special case of the so-called 'equilibrium' statistical mechanics. (See some further comments below on the latter.)

If the initial probability distribution, $P_{t=0}(x)$ is known, $P_t(x)$ for any consecutive t is well defined, in the sense that the Fokker-Planck is the Cauchi (initial value) problem with unique solution.

Remarks about simulations. One can solve PDE but can also analyze stochastic ODE approaching the problem in two complementary ways - correspondent to Eulerian and Lagrangian analysis in Fluid Mechanics describing "incompressible" flows in the probability space.

Main and the simplest (already mentioned) example of the Langevin dynamic is the Brownian motion, i.e. the case of $F = 0$. Another example, principal for the so-called 'equilibrium statistical physics', is of the potential force $F = \partial_x U(x)$, where $U(x)$ is a potential. Think, for example about x representing a particle connected to the origin by a spring. $U(x)$ is the potential/energy stored within the spring. In this case of the gradient force the stationary (i.e. time-independent) solution of the Fokker Planck Eq. (IV.15) can be found explicitly,

$$P_{st}(x) = Z^{-1} \exp\left(-\frac{U(x)}{D}\right). \tag{IV.17}$$

This solution is called Gibbs distribution, or equilibrium distribution.

Exercise: Show that the dynamic in the gradient force case obeys the Detailed Balance.

5. Recitation. Homogeneous and Forced Brownian Motion.

6. Recitation. First Passage Problem. Effects of Boundaries. Kramers Escape Problem.