Expectation Propagation and Intro to Sampling Methods

Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

March 24, 2018

Expectation Propagation

Minimizing KL by Moment Matching

- VB minimizes $\text{KL}(q||p)$ w.r.t. q. Consider minimizing the "reverse", i.e., $\text{KL}(p||q)$
- Assume $p(\mathbf{Z})$ fixed (but unknown) and $q(\mathbf{Z})$ to be an exponential family dist.

$$q(\mathbf{Z}) = h(\mathbf{Z})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^\top T(\mathbf{Z}))$$

- Then $KL(p||q) = \int p(\mathbf{Z}) \log \frac{p(\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z}$ can then be written as

$$KL(p||q) = -\log g(\boldsymbol{\eta}) - \boldsymbol{\eta}^\top \mathbb{E}_{p(\mathbf{Z})}[T(\mathbf{Z})] + \text{const}$$

- Minimizing w.r.t. q means minimizing w.r.t. $\boldsymbol{\eta}$, and is attained when

$$-\nabla \log g(\boldsymbol{\eta}) = \mathbb{E}_{p(\boldsymbol{z})}[T(\mathbf{Z})]$$

- Note that $-\nabla \log g(\boldsymbol{\eta})$ is equal to $\mathbb{E}_{q(\mathbf{Z})}[T(\mathbf{Z})]$, we will have

$$\mathbb{E}_{q(\mathbf{Z})}[T(\mathbf{Z})] = \mathbb{E}_{p(\mathbf{Z})}[T(\mathbf{Z})]$$

  which is a simple moment matching problem. How about using this idea for approximate inference?

Expectation Propagation

- Denote the unknowns by θ. Assume the true posterior distribution over θ given data $\mathcal{D}$

$$p(\theta|\mathcal{D}) = \frac{1}{p(\mathcal{D})}\prod_i f_i(\theta)$$

- For many problems $p(\theta|\mathcal{D})$ has this form: one factor $f_n(\boldsymbol{x}|\theta)$ per data point $\boldsymbol{x}_n$, plus one factor $f_0(\theta) = p(\theta)$ for the prior θ (total $N+1$ factors)
- Assume we approximate it with another product of total $N+1$ factors

$$q(\theta) = \frac{1}{Z}\prod_i \tilde{f}_i(\theta)$$

- Expectation Propagation (EP) is based on minimizing the following KL divergence

$$\mathrm{KL}\,(p||q) = \mathrm{KL}\left(\frac{1}{p(\mathcal{D})}\prod_i f_i(\boldsymbol{\theta}) \middle\| \frac{1}{Z}\prod_i \widetilde{f}_i(\boldsymbol{\theta})\right)$$

- But the above KL minimization is a hard problem in general
  - Reason: Since $KL(p||q) = \int p(\theta|\mathcal{D}) \log \frac{p(\theta|\mathcal{D})}{q(\theta)} d\theta$, this requires averaging over the true posterior

Expectation Propagation

- EP is an iterative scheme of solving the above problem by matching each $\tilde{f}_j(\theta)$ with $f_j(\theta)$
  - But instead of doing it independently for each $\tilde{f}_j(\theta)$, we do it in the context of all other $\tilde{f}_i(\theta)$, $i \neq j$
- The idea is to refine $\tilde{f}_j(\theta)$ s.t. the new approx. posterior

$$q^{\text{new}}(\boldsymbol{\theta}) \propto \widetilde{f}_j(\boldsymbol{\theta}) \prod_{i \neq j} \widetilde{f}_i(\boldsymbol{\theta})$$

  is as close as possible to the distribution $\propto f_j(\theta) \prod_{i \neq j} \tilde{f}_i(\theta)$
- Define $q^{\setminus j} = \frac{q(\theta)}{\tilde{f}_j(\theta)}$; can get it easily by subtracting off the natural parameters of $\tilde{f}_j$ from q
- Now solve the following *simpler* KL minimization problem w.r.t. $q^{new}(\theta)$

$$\text{KL}\left(\frac{f_j(\boldsymbol{\theta}) q^{\setminus j}(\boldsymbol{\theta})}{Z_j} \middle\| q^{\text{new}}(\boldsymbol{\theta}) \right)$$

  where $Z_j = \int f_j(\theta) q^{\setminus j}(\theta) d\theta$
- The above can be solved by matching the moments of q^{new} with $\frac{f_j(\theta) q^{\setminus j}(\theta)}{Z_j}$

Expectation Propagation

- From $q^{new}(\theta)$, we can get the required factor $\tilde{f}_j(\theta)$

$$\tilde{f}_j(\theta) = K \frac{q^{new}(\theta)}{q^{\setminus j}(\theta)}$$

where $K = \int \tilde{f}_j(\theta) q^{\setminus j}(\theta) d\theta$

- Finally, using zeroth order moment matching

$$\int \tilde{f}_j(\theta) q^{\setminus j}(\theta) d\theta = \int f_j(\theta) q^{\setminus j}(\theta) d\theta$$

we get $K = Z_j$

- This is repeated over each factor $\tilde{f}_j(\theta)$, for several passes
- Look at the clutter problem in PRML (sec 10.7.1) for a concrete example

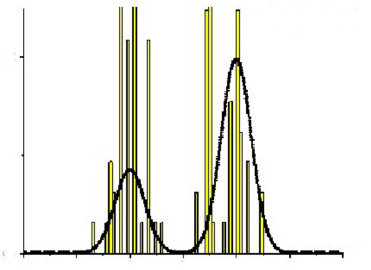
Sampling Methods for Approximate Inference

Sampling for Approximate Inference

- Some typical inference tasks
  - Compute a (possibly intractable) posterior distribution: $p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$
  - Compute a difficult expectation of a random quantity w.r.t. a distribution (an integral), e.g.,
    - The posterior predictive (an expectation w.r.t the posterior over θ)
$$p(\mathcal{D}^{new}|\mathcal{D}) = \int p(\mathcal{D}^{new}|\theta)p(\theta|\mathcal{D})d\theta = \mathbb{E}_{p(\theta|\mathcal{D})}[p(\mathcal{D}^{new}|\theta)]$$
    - The marginal likelihood or "evidence" (an expectation over the prior)
$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\theta)p(\theta|m)d\theta = \mathbb{E}_{p(\theta|m)}[p(\mathcal{D}|\theta)]$$
    - The expected complete data log-likelihood needed for doing MLE/MAP in LVMs (recall EM)
$$\text{Exp-CLL} = \int p(\boldsymbol{z}|\theta, \boldsymbol{x})p(\boldsymbol{x}, \boldsymbol{z}|\theta)d\boldsymbol{z} = \mathbb{E}_{p(\boldsymbol{z}|\theta,\boldsymbol{x})}[p(\boldsymbol{x}, \boldsymbol{z}|\theta)]$$
    - The ELBO in variational inference
$$\mathcal{L}(q) = \mathbb{E}_q[\log p(\boldsymbol{x}, \boldsymbol{z})] - \mathbb{E}_q[\log p(\boldsymbol{z})[$$
- Sampling methods provide a general way to (approximately) solve these problems

The Basic Idea

- Can approximate any distribution using a set of randomly drawn samples from it

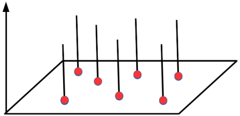


- Usually straightforward to generate samples if it is a simple/standard distribution
- Good News: Even if the distribution is "difficult" (e.g., an intractable posterior), it is often possible to generate random samples from such a distribution
    - Therefore these random samples can be used to approximate such difficult distributions

Empirical Distribution

- Sampling based approximation of a distribution can be represented using an empirical distribution
- Given L "points" $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$, the empirical distribution of these points is defined as

$$p_L(A) = \sum_{\ell=1}^{L} w_\ell \delta_{\boldsymbol{z}^{(\ell)}}(A)$$



- Here $w_1, \ldots, w_L$ are weights that sum to 1, i.e., $\sum_{\ell=1}^{L} w_\ell = 1$ (for uniform weights, $w_\ell = 1/L$)
- Here $\delta_{\boldsymbol{z}}(A)$ denotes the Dirac distribution defined as

$$\delta_{\boldsymbol{z}}(A) = \begin{cases} 0 & \text{if} \quad \boldsymbol{z} \notin A \\ 1 & \text{if} \quad \boldsymbol{z} \in A \end{cases}$$

- $p_L(A)$ is a discrete distribution with finite support $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$ (can think of it as a histogram)

Sampling: Some Basic Methods

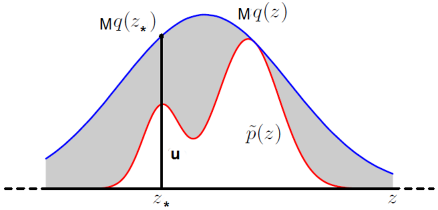
- Most of these basic methods are based on the idea of transformation
- Given a sample x from an "easy" distribution $p(x)$, transform it into a random sample z from a "less easy" distribution $p(z)$
- Some popular examples of transformation methods
  - Inverse CDF method
$$x \sim \text{Unif}(0, 1) \Rightarrow z = \text{Inv-CDF}_{p(z)}(x) \sim p(z)$$
  - Reparametrization method
$$x \sim \mathcal{N}(0, 1) \Rightarrow z = \mu + \sigma x \sim \mathcal{N}(\mu, \sigma^2)$$
  - Box-Muller method: Given (x_1, x_2) from $\text{Unif}(-1, +1)$, generate (z_1, z_2) from 2D Gaussian $\mathcal{N}(0, \mathbf{I})$
- Transformation Methods are simple but have limitations
  - Mostly limited to standard distributions and/or distributions with very few variables

Rejection Sampling

- Want to sample from $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$. Suppose we can only evaluate the numerator $\tilde{p}(\boldsymbol{z})$ at any $\boldsymbol{z}$
- Suppose we have a proposal distribution $q(\boldsymbol{z})$ that we can generate samples from, and

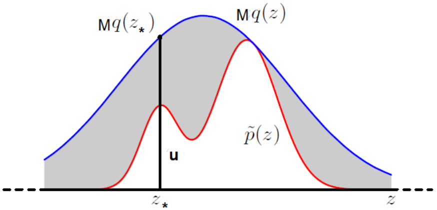
$$Mq(\boldsymbol{z}) \geq \tilde{p}(\boldsymbol{z}) \qquad \forall \boldsymbol{z} \quad \text{(where } M > 0 \text{ is some const.)}$$

- Basic idea: Generate samples from the proposal $q(\boldsymbol{z})$ and accept/reject based on some condition
  1. Sample an r.v. $\boldsymbol{z}_*$ from $q(\boldsymbol{z})$
  2. Sampling a uniform r.v. $u \sim \text{Unif}[0, Mq(\boldsymbol{z}_*)]$



  3. If $u \leq \tilde{p}(\boldsymbol{z}_*)$ then accept $\boldsymbol{z}_*$ else reject

Rejection Sampling



- Why $\boldsymbol{z} \sim q(\boldsymbol{z})$ + accept/reject rule is equivalent to $\boldsymbol{z} \sim p(\boldsymbol{z})$?
- Let's look at the pdf of $\boldsymbol{z}$'s that were accepted, i.e., $p(\boldsymbol{z}|\text{accept})$

$$
\begin{aligned}
p(\text{accept}|\boldsymbol{z}) &= \int_0^{\tilde{p}(\boldsymbol{z})} \frac{1}{Mq(\boldsymbol{z})} du = \frac{\tilde{p}(\boldsymbol{z})}{Mq(\boldsymbol{z})} \\
p(\boldsymbol{z}, \text{accept}) &= q(\boldsymbol{z}) p(\text{accept}|\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{M} \\
p(\text{accept}) &= \int \frac{\tilde{p}(\boldsymbol{z})}{M} d\boldsymbol{z} = \frac{Z_p}{M} \\
p(\boldsymbol{z}|\text{accept}) &= \frac{p(\boldsymbol{z}, \text{accept})}{p(\text{accept})} = \frac{\tilde{p}(\boldsymbol{z})}{Z_p} = p(\boldsymbol{z})
\end{aligned}
$$

Sampling Methods for Approximating Expectations

- Suppose $f(\boldsymbol{z})$ is function of a random variable $\boldsymbol{z} \sim p(\boldsymbol{z})$
- Wish to compute $\mathbb{E}[f] = \mathbb{E}_{p(\boldsymbol{z})}[f(\boldsymbol{z})] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$
- Given L independent samples $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L}$ from $p(\boldsymbol{z})$, we can approximate the above as

$$\mathbb{E}[f] \approx \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)}) \quad \text{(Monte Carlo sampling)}$$

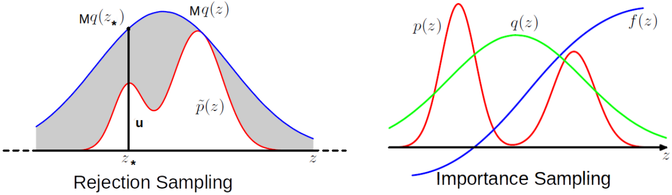
- What if we can't generate samples from $p(\boldsymbol{z})$? Answer: Use Importance Sampling
  - If we can generate L indep. samples $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L}$ from a different "proposal" distribution $q(\boldsymbol{z})$ then

$$\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \int f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}q(\boldsymbol{z})d\boldsymbol{z} \approx \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)})\frac{p(\boldsymbol{z}^{(\ell)})}{q(\boldsymbol{z}^{(\ell)})}$$

  - IS only requires that we can evaluate $p(\boldsymbol{z})$ at any $\boldsymbol{z}$ (in fact, with a small modification to the above, IS works even when we can evaluate $p(\boldsymbol{z})$ only up to a proportionality constant)

Limitations of Basic Sampling Methods

- Transformation based methods: Usually limited to drawing from standard distributions
- Rejection Sampling and Importance Sampling: Require good proposal distributions



- Difficult to find good prop. distr. especially when $\boldsymbol{z}$ is high-dim. (e.g., models with many params)
  - In high dimensions, most of the mass of $p(\boldsymbol{z})$ is concentrated in a tiny region of the $\boldsymbol{z}$ space
  - Difficult to *a priori* know what those regions are, thus difficult to come up with good proposal dist.
- Next Class: MCMC methods