Inference and Representation

David Sontag

New York University

Lecture 7, Oct. 28, 2014

Approximate marginal inference

- Given the joint $p(x_1, \ldots, x_n)$ represented as a graphical model, how do we perform **marginal inference**, e.g. to compute $p(x_1 \mid e)$?
- We showed in Lecture 4 that doing this exactly is NP-hard
- Nearly all *approximate inference* algorithms are either:
  1. Monte-carlo methods (e.g., Gibbs sampling, likelihood reweighting, MCMC)
  2. Variational algorithms (e.g., mean-field, loopy belief propagation)

Latent Dirichlet allocation (LDA)

- **Topic models** are powerful tools for exploring large data sets and for making inferences about the content of documents

Documents

Topics

politics	religion	sports
president	hindu	baseball
obama	judiasm	soccer
washington	ethics	basketball
religion	buddhism	football
...	...	...

- Many applications in information retrieval, document summarization, and classification

New document

Words $w_1, \ldots, w_N$

What is this document about?

weather	.50
finance	.49
sports	.01

Distribution of topics θ

- LDA is one of the simplest and most widely used topic models

Generative model for a document in LDA

1. Sample the document's **topic distribution** θ (aka topic vector)

$$\theta \sim \text{Dirichlet}(\alpha_{1:T})$$

   where the $\{\alpha_t\}_{t=1}^{T}$ are fixed hyperparameters. Thus θ is a distribution over T topics with mean $\theta_t = \alpha_t / \sum_{t'} \alpha_{t'}$

2. For $i = 1$ to N, sample the **topic** z_i of the i'th word

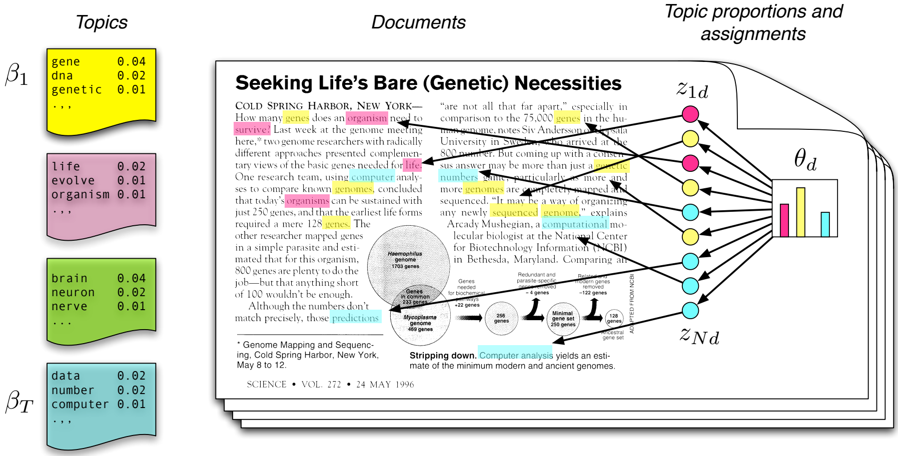
$$z_i | \theta \sim \theta$$

3. ... and then sample the actual **word** w_i from the z_i'th topic

$$w_i | z_i \sim \beta_{z_i}$$

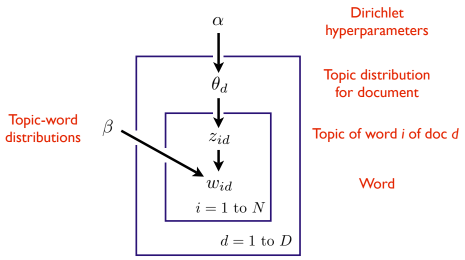
   where $\{\beta_t\}_{t=1}^{T}$ are the *topics* (a fixed collection of distributions on words)

Example of using LDA



(Blei, *Introduction to Probabilistic Topic Models*, 2011)

Approximate inference for latent Dirichlet Allocation



- Parameters are α and β
- Both θ_d and $\mathbf{z}_d$ are unobserved
- The difficulty here is that **inference** is intractable – because of the Dirichlet prior on $\vec{\theta}_d$, which encourages sparsity among the T topics

Variational methods

- **Goal**: Approximate difficult distribution $p(\mathbf{x} \mid \mathbf{e})$ with a new distribution $q(\mathbf{x})$ such that:
  1. $p(\mathbf{x} \mid \mathbf{e})$ and $q(\mathbf{x})$ are "close"
  2. Computation on $q(\mathbf{x})$ is easy
- How should we measure distance between distributions?
- The **Kullback-Leibler divergence** (KL-divergence) between two distributions p and q is defined as

$$D(p\|q) = \sum_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}.$$

  (measures the expected number of extra bits required to describe *samples from* $p(\mathbf{x})$ using a code based on q instead of p)
- $D(p \,\|\, q) \geq 0$ for all p, q, with equality if and only if $p = q$
- Notice that KL-divergence is **asymmetric**

KL-divergence *(see Section 2.8.2 of Murphy)*

$$D(p\|q) = \sum_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}.$$

- Suppose p is the true distribution we wish to do inference with
- What is the difference between the solution to

$$\arg\min_q D(p\|q)$$

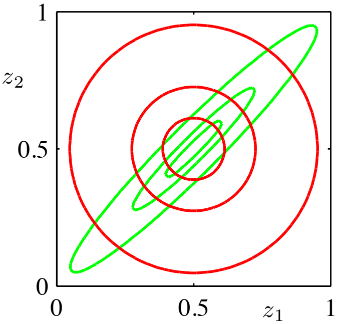
  (called the *M-projection* of q onto p) and

$$\arg\min_q D(q\|p)$$

  (called the *I-projection*)?
- These two will differ only when q is minimized over a restricted set of probability distributions $Q = \{q_1, \ldots\}$, and in particular when $p \notin Q$

$$q^* = \arg\min_{q \in Q} D(p\|q) = \sum_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}.$$

For example, suppose that $p(\mathbf{z})$ is a 2D Gaussian and Q is the set of all Gaussian distributions with diagonal covariance matrices:

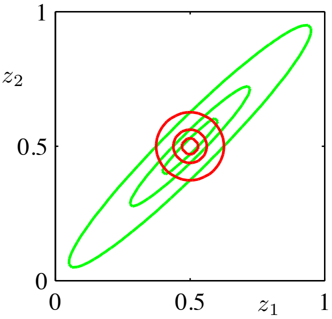


p=Green, q^*=Red

KL-divergence – I-projection

$$q^* = \arg\min_{q \in Q} D(q \| p) = \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})}.$$

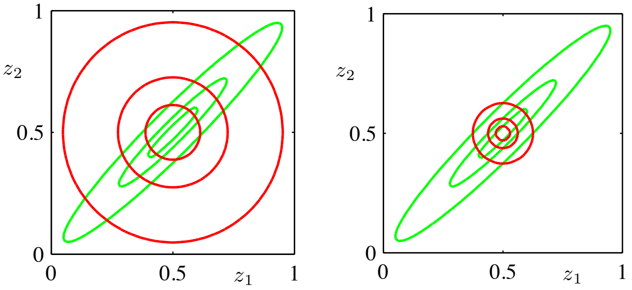
For example, suppose that $p(\mathbf{z})$ is a 2D Gaussian and Q is the set of all Gaussian distributions with diagonal covariance matrices:



p=Green, q^*=Red

KL-divergence (single Gaussian)

In this simple example, both the M-projection and I-projection find an approximate $q(\mathbf{x})$ that has the correct mean (i.e. $E_p[\mathbf{z}] = E_q[\mathbf{z}]$):

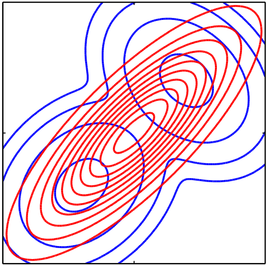


What if $p(\mathbf{x})$ is multi-modal?

KL-divergence – M-projection (mixture of Gaussians)

$$q^* = \arg\min_{q \in Q} D(p\|q) = \sum_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}.$$

Now suppose that $p(\mathbf{x})$ is mixture of two 2D Gaussians and Q is the set of all 2D Gaussian distributions (with arbitrary covariance matrices):

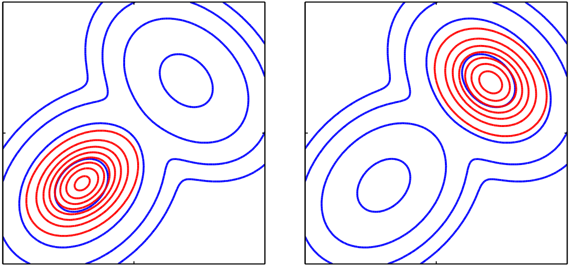


p=Blue, q^*=Red

M-projection yields distribution $q(\mathbf{x})$ with the correct mean and covariance.

KL-divergence – I-projection (mixture of Gaussians)

$$q^* = \arg\min_{q \in Q} D(q\|p) = \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})}.$$



p=Blue, q^*=Red (two local minima!)

Unlike the M-projection, the I-projection does not necessarily yield the correct moments.

Mapping of distributions to/from moments

- Recall the definition of probability distributions in the exponential family:
$$q(\mathbf{x};\eta) = h(\mathbf{x}) \exp\{\eta \cdot \mathbf{f}(\mathbf{x}) - \ln Z(\eta)\}$$

  $\mathbf{f}(\mathbf{x})$ are called the *sufficient statistics*

- In the exponential family, there is a one-to-one correspondance between distributions $q(\mathbf{x};\eta)$ and marginal vectors $E_q[\mathbf{f}(\mathbf{x})]$
- For example, when q is a Gaussian distribution,
$$q(\mathbf{x};\mu,\Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

  then $\mathbf{f}(\mathbf{x}) = [x_1, x_2, \ldots, x_k, x_1^2, x_1x_2, x_1x_3, \ldots, x_2^2, x_2x_3, \ldots]$

- The expectation of $\mathbf{f}(\mathbf{x})$ gives the first and second-order (non-central) moments, from which one can solve for μ and Σ

Properties of exponential families

The derivative of the log-partition function is equal to the distribution's marginals:

$$
\begin{aligned}
\partial_{\eta_i} \ln Z(\eta) &= \partial_{\eta_i} \ln \sum_{\mathbf{x}} \exp\{\eta \cdot \mathbf{f}(\mathbf{x})\} \\
&= \frac{1}{\sum_{\mathbf{x}} \exp\{\eta \cdot \mathbf{f}(\mathbf{x})\}} \partial_{\eta_i} \sum_{\mathbf{x}} \exp\{\eta \cdot \mathbf{f}(\mathbf{x})\} \\
&= \frac{1}{\sum_{\mathbf{x}} \exp\{\eta \cdot \mathbf{f}(\mathbf{x})\}} \sum_{\mathbf{x}} \partial_{\eta_i} \exp\{\eta \cdot \mathbf{f}(\mathbf{x})\} \\
&= \frac{1}{\sum_{\mathbf{x}} \exp\{\eta \cdot \mathbf{f}(\mathbf{x})\}} \sum_{\mathbf{x}} \exp\{\eta \cdot \mathbf{f}(\mathbf{x})\} \partial_{\eta_i} \eta \cdot \mathbf{f}(\mathbf{x}) \\
&= \frac{1}{\sum_{\mathbf{x}} \exp\{\eta \cdot \mathbf{f}(\mathbf{x})\}} \sum_{\mathbf{x}} \exp\{\eta \cdot \mathbf{f}(\mathbf{x})\} f_i(\mathbf{x}) \\
&= \sum_{\mathbf{x}} \frac{\exp\{\eta \cdot \mathbf{f}(\mathbf{x})\}}{\sum_{\hat{\mathbf{x}}} \exp\{\eta \cdot \mathbf{f}(\hat{\mathbf{x}})\}} f_i(\mathbf{x}) = \sum_{\mathbf{x}} q(\mathbf{x}) f_i(\mathbf{x}) = E_q[f_i(\mathbf{x})].
\end{aligned}
$$

M-projection does moment matching

- Recall that the M-projection is:
$$q^* = \arg\min_{q \in Q} D(p \| q) = \sum_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}.$$
- Suppose that Q is an exponential family ($p(\mathbf{x})$ can be arbitrary) and that we perform the M-projection, finding q^*
- **Theorem:** The expected sufficient statistics, with respect to $q^*(\mathbf{x})$, are *exactly* the marginals of $p(\mathbf{x})$:
$$E_{q^*}[\mathbf{f}(\mathbf{x})] = E_p[\mathbf{f}(\mathbf{x})]$$
- Thus, solving for the M-projection (exactly) is just as hard as the original inference problem

M-projection does moment matching

- Recall that the M-projection is:

$$q^* = \arg\min_{q(\mathbf{x};\eta)\in Q} D(p\|q) = \sum_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}.$$

- **Theorem:** $E_{q^*}[\mathbf{f}(\mathbf{x})] = E_p[\mathbf{f}(\mathbf{x})]$.
- **Proof:** Look at the first-order optimality conditions.

$$\begin{aligned}
\partial_{\eta_i} D(p\|q) &= -\partial_{\eta_i} \sum_{\mathbf{x}} p(\mathbf{x}) \log q(\mathbf{x}) \\
&= -\partial_{\eta_i} \sum_{\mathbf{x}} p(\mathbf{x}) \log \left\{ h(\mathbf{x}) \exp\{\eta \cdot \mathbf{f}(\mathbf{x}) - \ln Z(\eta)\} \right\} \\
&= -\partial_{\eta_i} \sum_{\mathbf{x}} p(\mathbf{x}) \left\{ \eta \cdot \mathbf{f}(\mathbf{x}) - \ln Z(\eta) \right\} \\
&= -\sum_{\mathbf{x}} p(\mathbf{x}) f_i(\mathbf{x}) + E_{q(\mathbf{x};\eta)}[f_i(\mathbf{x})] \\
&= -E_p[f_i(\mathbf{x})] + E_{q(\mathbf{x};\eta)}[f_i(\mathbf{x})] = 0.
\end{aligned}$$

- **Corollary**: Even computing the gradients is hard (can't do gradient descent)

Most variational inference algorithms make use of the I-projection

Variational methods

- Suppose that we have an arbitrary graphical model:

$$p(\mathbf{x};\theta) = \frac{1}{Z(\theta)} \prod_{\mathbf{c}\in C} \phi_c(\mathbf{x_c}) = \exp\Big(\sum_{\mathbf{c}\in C} \theta_c(\mathbf{x_c}) - \ln Z(\theta)\Big)$$

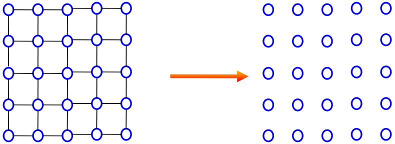
- All of the approaches begin as follows:

$$\begin{aligned}
D(q\|p) &= \sum_{\mathbf{x}} q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} \\
&= -\sum_{\mathbf{x}} q(\mathbf{x}) \ln p(\mathbf{x}) - \sum_{\mathbf{x}} q(\mathbf{x}) \ln \frac{1}{q(\mathbf{x})} \\
&= -\sum_{\mathbf{x}} q(\mathbf{x}) \big(\sum_{\mathbf{c}\in C} \theta_c(\mathbf{x_c}) - \ln Z(\theta)\big) - H(q(\mathbf{x})) \\
&= -\sum_{\mathbf{c}\in C} \sum_{\mathbf{x}} q(\mathbf{x}) \theta_c(\mathbf{x_c}) + \sum_{\mathbf{x}} q(\mathbf{x}) \ln Z(\theta) - H(q(\mathbf{x})) \\
&= -\sum_{\mathbf{c}\in C} E_q[\theta_c(\mathbf{x_c})] + \ln Z(\theta) - H(q(\mathbf{x})).
\end{aligned}$$

Mean field algorithms for variational inference

$$\max_{q \in Q} \sum_{\mathbf{c} \in C} \sum_{\mathbf{x_c}} q(\mathbf{x_c}) \theta_c(\mathbf{x_c}) + H(q(\mathbf{x})).$$

- Although this function is concave and thus in theory should be easy to optimize, we need some compact way of representing $q(\mathbf{x})$
- *Mean field* algorithms assume a factored representation of the joint distribution, e.g.



$$q(\mathbf{x}) = \prod_{i \in V} q_i(x_i) \qquad \text{(called } \textit{naive} \text{ mean field)}$$

Naive mean-field

- Suppose that Q consists of all fully factored distributions, of the form $q(\mathbf{x}) = \prod_{i \in V} q_i(x_i)$
- We can use this to simplify

$$\max_{q \in Q} \sum_{\mathbf{c} \in C} \sum_{\mathbf{x_c}} q(\mathbf{x_c}) \theta_c(\mathbf{x_c}) + H(q)$$

- First, note that $q(\mathbf{x}_c) = \prod_{i \in c} q_i(x_i)$
- Next, notice that the joint entropy decomposes as a sum of local entropies:

$$\begin{aligned} H(q) &= -\sum_{\mathbf{x}} q(\mathbf{x}) \ln q(\mathbf{x}) \\ &= -\sum_{\mathbf{x}} q(\mathbf{x}) \ln \prod_{i \in V} q_i(x_i) = -\sum_{\mathbf{x}} q(\mathbf{x}) \sum_{i \in V} \ln q_i(x_i) \\ &= -\sum_{i \in V} \sum_{\mathbf{x}} q(\mathbf{x}) \ln q_i(x_i) \\ &= -\sum_{i \in V} \sum_{x_i} q_i(x_i) \ln q_i(x_i) \sum_{\mathbf{x}_{V \setminus i}} q(\mathbf{x}_{V \setminus i} \mid x_i) = \sum_{i \in V} H(q_i). \end{aligned}$$

Naive mean-field

- Suppose that Q consists of all fully factored distributions, of the form $q(\mathbf{x}) = \prod_{i \in V} q_i(x_i)$
- We can use this to simplify

$$\max_{q \in Q} \sum_{\mathbf{c} \in C} \sum_{\mathbf{x_c}} q(\mathbf{x_c}) \theta_c(\mathbf{x_c}) + H(q)$$

- First, note that $q(\mathbf{x}_c) = \prod_{i \in c} q_i(x_i)$
- Next, notice that the joint entropy decomposes as $H(q) = \sum_{i \in V} H(q_i)$.
- Putting these together, we obtain the following variational objective:

$$(*) \max_{q} \sum_{\mathbf{c} \in C} \sum_{\mathbf{x_c}} \theta_c(\mathbf{x_c}) \prod_{i \in c} q_i(x_i) + \sum_{i \in V} H(q_i)$$

subject to the constraints

$$q_i(x_i) \geq 0 \quad \forall i \in V, x_i \in \mathrm{Val}(X_i)$$

$$\sum_{x_i \in \mathrm{Val}(X_i)} q_i(x_i) = 1 \quad \forall i \in V$$

Naive mean-field for pairwise MRFs

- How do we maximize the variational objective?

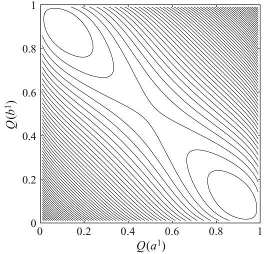
$$(*) \max_{q} \sum_{ij \in E} \sum_{x_i, x_j} \theta_{ij}(x_i, x_j) q_i(x_i) q_j(x_j) - \sum_{i \in V} \sum_{x_i} q_i(x_i) \ln q_i(x_i)$$

- This is a non-concave optimization problem, with many local maxima!
- Nonetheless, we can greedily maximize it using **block coordinate ascent**:
  1. Iterate over each of the variables $i \in V$. For variable i,
  2. Fully maximize (*) with respect to $\{q_i(x_i), \forall x_i \in \mathrm{Val}(X_i)\}$.
  3. Repeat until convergence.
- Constructing the Lagrangian, taking the derivative, setting to zero, and solving yields the update: (*shown on blackboard*)

$$q_i(x_i) \leftarrow \frac{1}{Z_i} \exp\left\{\theta_i(x_i) + \sum_{j \in N(i)} \sum_{x_j} q_j(x_j) \theta_{ij}(x_i, x_j)\right\}$$

How accurate will the approximation be?

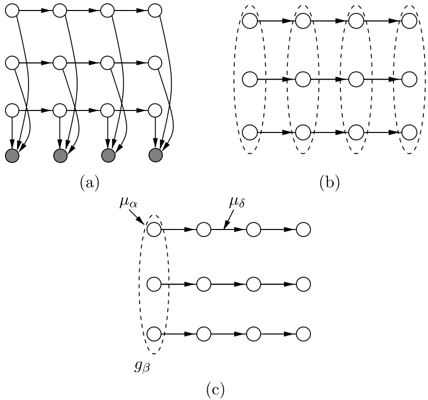
- Consider a distribution which is an XOR of two binary variables A and B: $p(a,b) = 0.5 - \epsilon$ if $a \neq b$ and $p(a,b) = \epsilon$ if $a = b$
- The contour plot of the variational objective is:



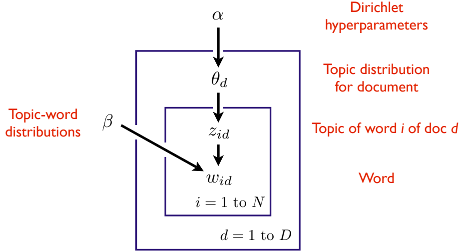
- Even for a single edge, mean field can give very wrong answers!
- Interestingly, once $\epsilon > 0.1$, mean field has a single maximum point at the uniform distribution (thus, exact)

Structured mean-field approximations

- Rather than assuming a fully-factored distribution for q, we can use a *structured* approximation, such as a spanning tree
- For example, for a factorial HMM, a good approximation may be a product of chain-structured models:



Approximate inference for latent Dirichlet Allocation



- Parameters are α and β
- Both θ_d and $\mathbf{z}_d$ are unobserved
- Use the mean field approximation:

$$q(\theta_d, \mathbf{z}_d | \gamma_d, \phi_d) = q(\theta_d \mid \gamma_d) \prod_{n=1}^{N} q(z_n \mid \phi_{dn})$$