Approximate Inference

IPAM Summer School

Ruslan Salakhutdinov
BCS, MIT
Deprtment of Statistics, University of Toronto

Plan

1. Introduction/Notation.

2. Illustrative Examples.

3. Laplace Approximation.

4. Variational Inference / Mean-Field.

References/Acknowledgements

- Chris Bishop's book: **Pattern Recognition and Machine Learning**, chapter 11 (many figures are borrowed from this book).

- David MacKay's book: **Information Theory, Inference, and Learning Algorithms**, chapters 29-32.

- Radford Neals's technical report on **Probabilistic Inference Using Markov Chain Monte Carlo Methods**.

- Zoubin Ghahramani's ICML tutorial on Bayesian Machine Learning: http://www.gatsby.ucl.ac.uk/~zoubin/ICML04-tutorial.html

Inference Problem

Given a dataset $\mathcal{D} = \{x_1, ..., x_n\}$:

Bayes Rule:

$$P(\theta|\mathcal{D}) = \frac{P(D|\theta)P(\theta)}{P(\mathcal{D})}$$

$P(\mathcal{D}\|\theta)$	Likelihood function of θ
$P(\theta)$	Prior probability of θ
$P(\theta\|\mathcal{D})$	Posterior distribution over θ

Computing posterior distribution is known as the **inference** problem.
But:

$$P(\mathcal{D}) = \int P(\mathcal{D}, \theta) d\theta$$

This integral can be very high-dimensional and difficult to compute.

Prediction

$$P(\theta|\mathcal{D}) = \frac{P(D|\theta)P(\theta)}{P(\mathcal{D})}$$

$P(\mathcal{D}|\theta)$	Likelihood function of θ
$P(\theta)$	Prior probability of θ
$P(\theta|\mathcal{D})$	Posterior distribution over θ

Prediction: Given $\mathcal{D}$, computing conditional probability of x^* requires computing the following integral:

$$\begin{aligned} P(x^*|\mathcal{D}) &= \int P(x^*|\theta, \mathcal{D}) P(\theta|\mathcal{D}) d\theta \\ &= \mathbb{E}_{P(\theta|\mathcal{D})}[P(x^*|\theta, \mathcal{D})] \end{aligned}$$

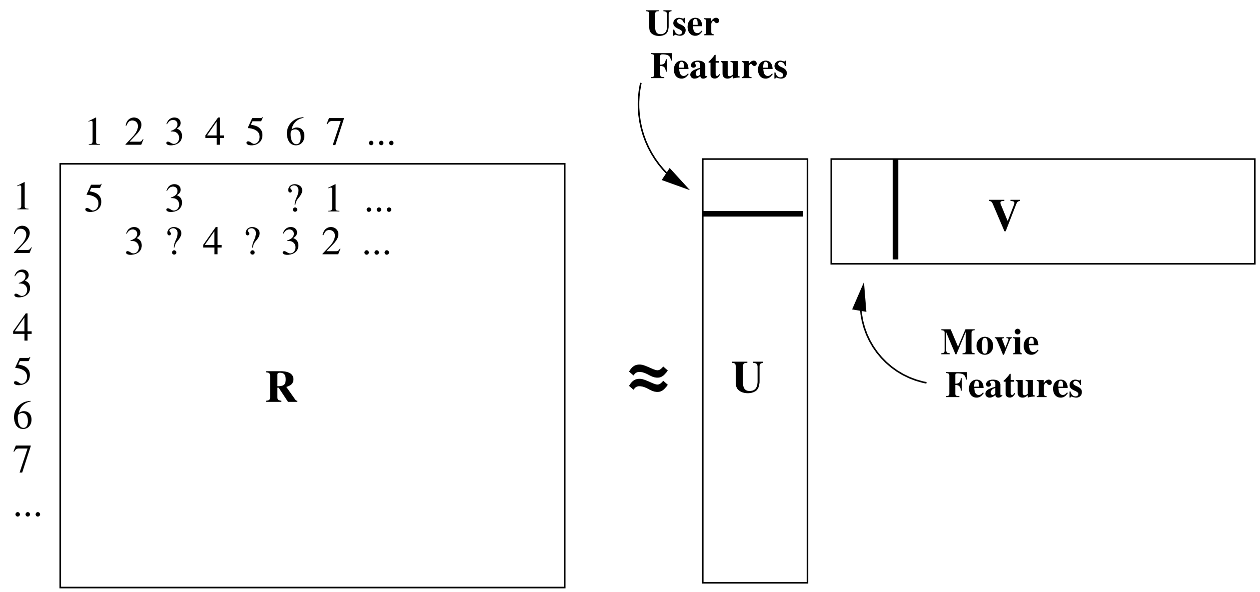
which is sometimes called **predictive distribution**.

Computing predictive distribution requires posterior $P(\theta|\mathcal{D})$.

Computational Challenges

- Computing marginal likelihoods often requires computing very high-dimensional integrals.
- Computing posterior distributions (and hence predictive distributions) is often analytically intractable.
- First, let us look at some examples.

Bayesian PMF



We have N users, M movies, and integer rating values from 1 to K.

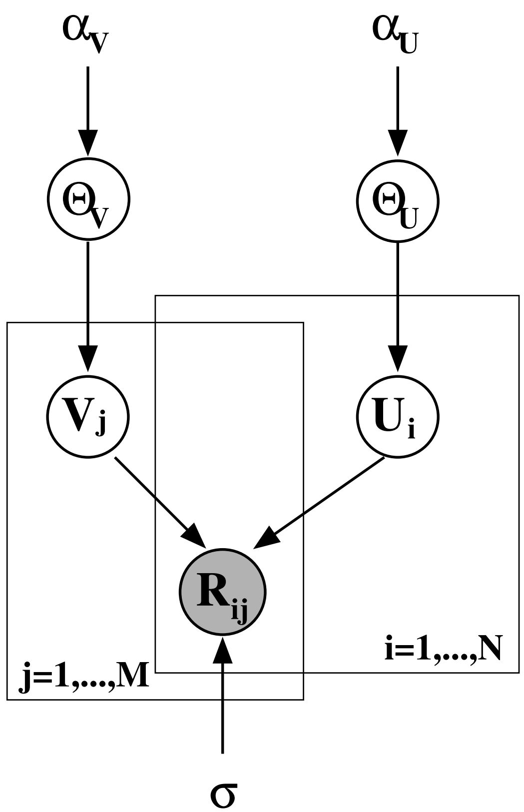
Let r_{ij} be the rating of user i for movie j, and $U \in R^{D \times N}$, $V \in R^{D \times M}$ be latent user and movie feature matrices:

$$R \approx U^\top V$$

Goal: Predict missing ratings.

Salakhutdinov and Mnih, NIPS 2008.

Bayesian PMF



Probabilistic linear model with Gaussian observation noise. Likelihood:

$$p(r_{ij}|u_i, v_j, \sigma^2) = \mathcal{N}(r_{ij}|u_i^\top v_j, \sigma^2)$$

Gaussian Priors over parameters:

$$p(U|\mu_U, \Lambda_U) = \prod_{i=1}^{N} \mathcal{N}(u_i|\mu_u, \Sigma_u),$$

$$p(V|\mu_V, \Lambda_V) = \prod_{i=1}^{M} \mathcal{N}(v_i|\mu_v, \Sigma_v).$$

Conjugate Gaussian-inverse-Wishart priors on the user and movie hyperparameters $\Theta_U = \{\mu_u, \Sigma_u\}$ and $\Theta_V = \{\mu_v, \Sigma_v\}$.

Hierarchical Prior.

Bayesian PMF

Predictive distribution: Consider predicting a rating r_{ij}^* for user i and query movie j:

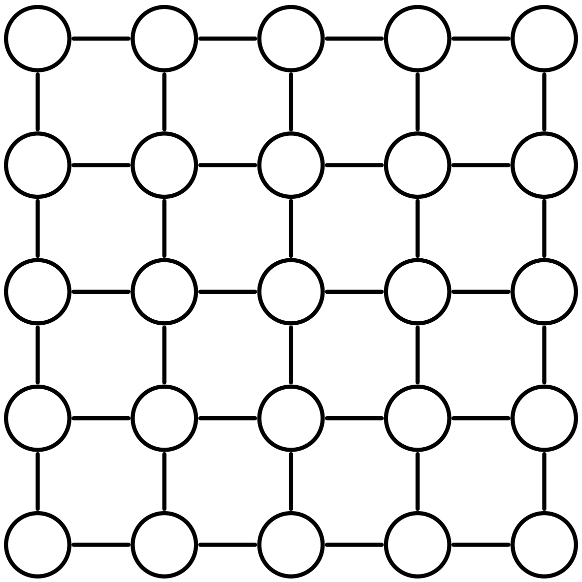
$$p(r_{ij}^*|R) = \iint p(r_{ij}^*|u_i, v_j)\underbrace{p(U, V, \Theta_U, \Theta_V|R)}_{\text{Posterior over parameters and hyperparameters}} d\{U, V\} d\{\Theta_U, \Theta_V\}$$

Exact evaluation of this predictive distribution is analytically intractable.

Posterior distribution $p(U, V, \Theta_U, \Theta_V|R)$ is complicated and does not have a closed form expression.

Need to approximate.

Undirected Models



$\mathbf{x}$ is a binary random vector with $x_i \in \{+1, -1\}$:

$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}} \exp\big(\sum_{(i,j) \in E} \theta_{ij} x_i x_j + \sum_{i \in V} \theta_i x_i \big).$$

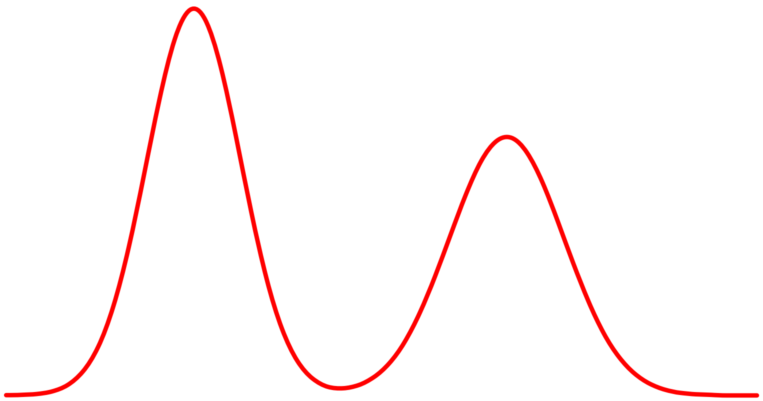
where $\mathcal{Z}$ is known as partition function:

$$\mathcal{Z} = \sum_{\mathbf{x}} \exp\big(\sum_{(i,j) \in E} \theta_{ij} x_i x_j + \sum_{i \in V} \theta_i x_i \big).$$

If $\mathbf{x}$ is 100-dimensional, need to sum over 2^{100} terms.
The sum might decompose (e.g. junction tree). Otherwise we need to approximate.

Remark: Compare to marginal likelihood.

Inference



For most situations we will be interested in evaluating the expectation:

$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})dz$$

We will use the following notation: $p(\mathbf{z}) = \frac{\tilde{p}(\mathbf{z})}{\mathcal{Z}}$.

We can evaluate $\tilde{p}(\mathbf{z})$ pointwise, but cannot evaluate $\mathcal{Z}$.

- Posterior distribution: $P(\theta|\mathcal{D}) = \frac{1}{P(\mathcal{D})}P(\mathcal{D}|\theta)P(\theta)$
- Markov random fields: $P(z) = \frac{1}{\mathcal{Z}}\exp(-E(z))$

Plan

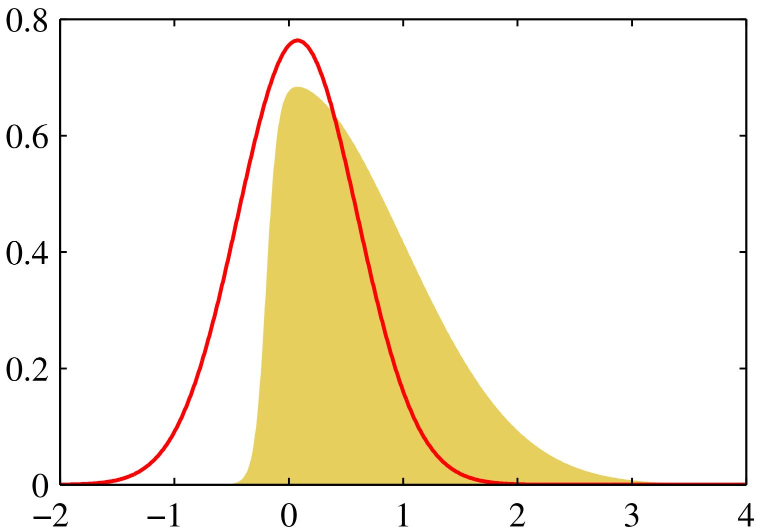
1. Introduction/Notation.

2. Illustrative Examples.

3. Laplace Approximation.

4. Variational Inference / Mean-Field.

Laplace Approximation



Consider:

$$p(\mathbf{z}) = \frac{\tilde{p}(\mathbf{z})}{\mathcal{Z}}$$

Goal: Find a Gaussian approximation $q(\mathbf{z})$ which is centered on a mode of the distribution $p(\mathbf{z})$.

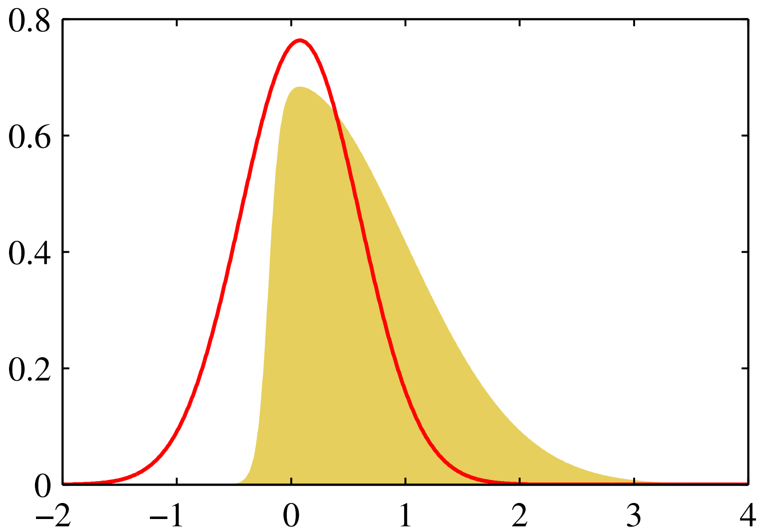
At a stationary point $\mathbf{z}_0$ the gradient $\bigtriangledown\tilde{p}(\mathbf{z})$ vanishes. Consider a Taylor expansion of $\ln\tilde{p}(\mathbf{z})$:

$$\ln\tilde{p}(\mathbf{z}) \approx \ln\tilde{p}(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T A(\mathbf{z}-\mathbf{z}_0)$$

where A is a Hessian matrix:

$$A = -\bigtriangledown\bigtriangledown\ln\tilde{p}(\mathbf{z})|_{z=z_0}$$

Laplace Approximation



Consider:

$$p(\mathbf{z}) = \frac{\tilde{p}(\mathbf{z})}{\mathcal{Z}}$$

Goal: Find a Gaussian approximation $q(\mathbf{z})$ which is centered on a mode of the distribution $p(\mathbf{z})$.

Exponentiating both sides:

$$\tilde{p}(\mathbf{z}) \approx \tilde{p}(\mathbf{z}_0) \exp\left(-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T A(\mathbf{z} - \mathbf{z}_0) \right)$$

We get a multivariate Gaussian approximation:

$$q(\mathbf{z}) = \frac{|A|^{1/2}}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T A(\mathbf{z} - \mathbf{z}_0) \right)$$

Laplace Approximation

Remember $p(\mathbf{z}) = \frac{\tilde{p}(\mathbf{z})}{\mathcal{Z}}$, where we approximate:

$$\mathcal{Z} = \int \tilde{p}(\mathbf{z}) d\mathbf{z} \approx \tilde{p}(\mathbf{z}_0) \int \exp\Big(-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T A(\mathbf{z} - \mathbf{z}_0)\Big) = \tilde{p}(\mathbf{z}_0) \frac{(2\pi)^{D/2}}{|A|^{1/2}}$$

Bayesian Inference: $P(\theta|\mathcal{D}) = \frac{1}{P(\mathcal{D})} P(\mathcal{D}|\theta)P(\theta)$.

Identify: $\tilde{p}(\theta|\mathcal{D}) = P(\mathcal{D}|\theta)P(\theta)$ and $\mathcal{Z} = P(\mathcal{D})$:

- The posterior is approximately Gaussian around the MAP estimate θ_{MAP}

$$p(\theta|\mathcal{D}) \approx \frac{|A|^{1/2}}{(2\pi)^{D/2}} \exp\Big(-\frac{1}{2}(\theta - \theta_{MAP})^T A(\theta - \theta_{MAP})\Big)$$

Laplace Approximation

Remember $p(\mathbf{z}) = \frac{\tilde{p}(\mathbf{z})}{\mathcal{Z}}$, where we approximate:

$$\mathcal{Z} = \int \tilde{p}(\mathbf{z}) d\mathbf{z} \approx \tilde{p}(\mathbf{z}_0) \int \exp\Big(-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T A (\mathbf{z} - \mathbf{z}_0)\Big) = \tilde{p}(\mathbf{z}_0) \frac{(2\pi)^{D/2}}{|A|^{1/2}}$$

Bayesian Inference: $P(\theta|\mathcal{D}) = \frac{1}{P(\mathcal{D})} P(\mathcal{D}|\theta) P(\theta)$.

Identify: $\tilde{p}(\theta|\mathcal{D}) = P(\mathcal{D}|\theta)P(\theta)$ and $\mathcal{Z} = P(\mathcal{D})$:

- Can approximate Model Evidence:

$$P(\mathcal{D}) = \int P(\mathcal{D}|\theta) P(\theta) d\theta$$

- Using Laplace approximation

$$\ln P(\mathcal{D}) \approx \ln P(D|\theta_{MAP}) + \underbrace{\ln P(\theta_{MAP}) + \frac{D}{2} \ln 2\pi - \frac{1}{2} \ln |A|}_{\text{Occam factor: penalize model complexity}}$$

Bayesian Information Criterion

BIC can be obtained from the Laplace approximation:

$$\ln P(\mathcal{D}) \approx \ln P(D|\theta_{MAP}) + \ln P(\theta_{MAP}) + \frac{D}{2}\ln 2\pi - \frac{1}{2}\ln|A|$$

by taking the large sample limit ($N \to \infty$) where N is the number of data points:

$$\ln P(\mathcal{D}) \approx P(D|\theta_{MAP}) - \frac{1}{2}D\ln N$$

- Quick, easy, does not depend on the prior.
- Can use maximum likelihood estimate of θ instead of the MAP estimate
- D denotes the number of "well-determined parameters"
- **Danger:** Counting parameters can be tricky (e.g. infinite models)

Plan

1. Introduction/Notation.

2. Illustrative Examples.

3. Laplace Approximation.

4. Variational Inference / Mean-Field.

Variational Inference

Key Idea: Approximate intractable distribution $p(\theta|D)$ with simpler, tractable distribution $q(\theta)$.

We can lower bound the marginal likelihood using Jensen's inequality:

$$\begin{aligned}
\ln p(\mathcal{D}) &= \ln \int p(\mathcal{D}, \theta) d\theta = \ln \int q(\theta) \frac{P(\mathcal{D}, \theta)}{q(\theta)} d\theta \\
&\geq \int q(\theta) \ln \frac{p(\mathcal{D}, \theta)}{q(\theta)} d\theta = \underbrace{\int q(\theta) \ln p(\mathcal{D}, \theta) d\theta + \underbrace{\int q(\theta) \ln \frac{1}{q(\theta)} d\theta}_{\text{Entropy functional}}}_{\text{Variational Lower-Bound}} \\
&= \ln p(\mathcal{D}) - \mathrm{KL}(q(\theta)||p(\theta|D)) = \mathcal{L}(q)
\end{aligned}$$

where $\mathrm{KL}(q||p)$ is a Kullback–Leibler divergence – a non-symmetric measure of the difference between two distributions q and p: $\mathrm{KL}(q||p) = \int q(\theta) \ln \frac{q(\theta)}{p(\theta)} dx$.

The goal of variational inference is to maximize the variational lower-bound w.r.t. approximate q distribution, or minimize $\mathrm{KL}(q||p)$.

Mean-Field Approximation

Key Idea: Approximate intractable distribution $p(\theta|D)$ with simpler, tractable distribution $q(\theta)$ by minimizing $\mathrm{KL}(q(\theta)||p(\theta|D))$.

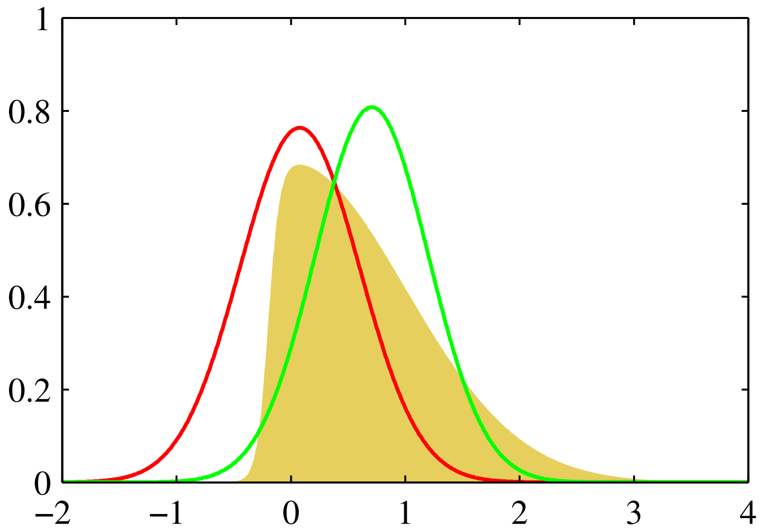
We can choose a fully factorized distribution: $q(\theta) = \prod_{i=1}^{D} q_i(\theta_i)$, also known as a mean-field approximation.

The variational lower-bound takes form:

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\theta) \ln p(\mathcal{D}, \theta) d\theta + \int q(\theta) \ln \frac{1}{q(\theta)} d\theta \\
&= \int q_j(\theta_j) \underbrace{\left[\ln p(\mathcal{D}, \theta) \prod_{i \neq j} q_i(\theta_i) d\theta_i \right]}_{\mathbb{E}_{i \neq j}[\ln p(\mathcal{D}, \theta)]} d\theta_j + \sum_i \int q_i(\theta_i) \ln \frac{1}{q(\theta_i)} d\theta_i
\end{aligned}$$

Suppose we keep $\{q_{i \neq j}\}$ fixed and maximize $\mathcal{L}(q)$ w.r.t. all possible forms for the distribution $q_j(\theta_j)$.

Mean-Field Approximation



The plot shows the original distribution (yellow), along with the Laplace (red) and variational (green) approximations.

By maximizing $\mathcal{L}(q)$ w.r.t. all possible forms for the distribution $q_j(\theta_j)$ we obtain a general expression:

$$q_j^*(\theta_j) = \frac{\exp(\mathbb{E}_{i \neq j}[\ln p(\mathcal{D}, \theta)])}{\int \exp(\mathbb{E}_{i \neq j}[\ln p(\mathcal{D}, \theta)]) d\theta_j}$$

Iterative Procedure: Initialize all q_j and then iterate through the factors replacing each in turn with a revised estimate.

Convergence is guaranteed as the bound is convex w.r.t. each of the factors q_j (see Bishop, chapter 10).

Other Variational Methods

Many other existing techniques:

- Loopy Belief Propagation.
- Expectation Propagation.
- Various other Message Passing algorithms.

We will see more of variational inference in tomorrow's lecture on Deep Networks.