A Brief Introduction to Variational Bayesian Inference

Mark Johnson

CG168 notes

Bayes rule

- Bayes theorem:

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)}$$

- Bayes inversion: swap direction of arcs in Bayes net
- Interpreted as a recipe for "belief updating":

$$\underbrace{P(\theta|D)}_{\text{Posterior}} \propto \underbrace{P(D|\theta)}_{\text{Likelihood}}\ \underbrace{P(\theta)}_{\text{Prior}}$$

- The normalizing constant (which you have to divide Likelihood times Prior by) is:

$$P(D) = \sum_{\theta'} P(D|\theta')\,P(\theta')$$

which is the probability of generating the data under *any* model

Categorical distributions

- A *categorical distribution* has a finite set of outcomes $1, \ldots, m$
- A categorical distribution is parameterized by a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, where $\mathrm{P}(X = j|\boldsymbol{\theta}) = \theta_j$ (so $\sum_{j=1}^{m} \theta_j = 1$)
  - Example: An m-sided die, where $\theta_j =$ prob. of face j
- Suppose $\boldsymbol{X} = (X_1, \ldots, X_n)$ and each $X_i|\boldsymbol{\theta} \sim \mathrm{Categorical}(\boldsymbol{\theta})$. Then:

$$\mathrm{P}(\boldsymbol{X}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathrm{Categorical}(X_i; \boldsymbol{\theta}) = \prod_{j=1}^{m} \theta_j^{N_j}$$

  where N_j is the number of times j occurs in $\boldsymbol{X}$.
- Goal of next few slides: compute $\mathrm{P}(\boldsymbol{\theta}|\boldsymbol{X})$

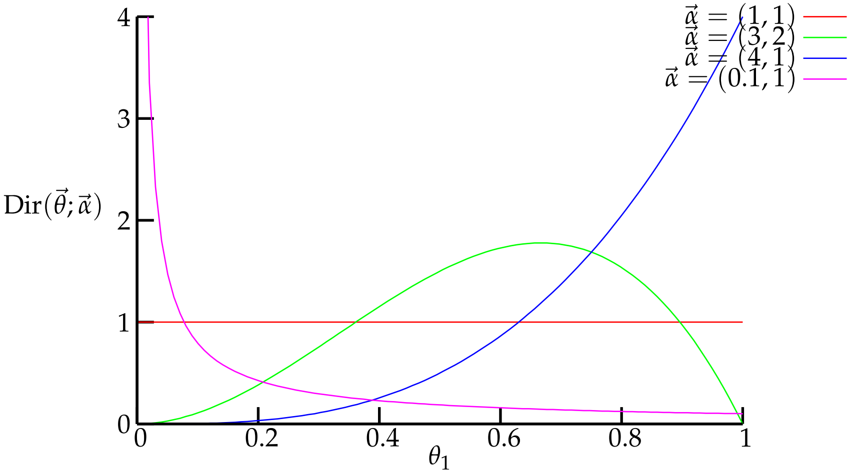
Dirichlet distributions

- *Dirichlet distributions* are probability distributions over multinomial parameter vectors
  - called *Beta distributions* when $m = 2$
- Parameterized by a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ where $\alpha_j > 0$ that determines the shape of the distribution

$$
\begin{aligned}
\mathrm{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) &= \frac{1}{C(\boldsymbol{\alpha})} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} \\
C(\boldsymbol{\alpha}) &= \int \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} \, d\boldsymbol{\theta} = \frac{\prod_{j=1}^{m} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{m} \alpha_j)}
\end{aligned}
$$

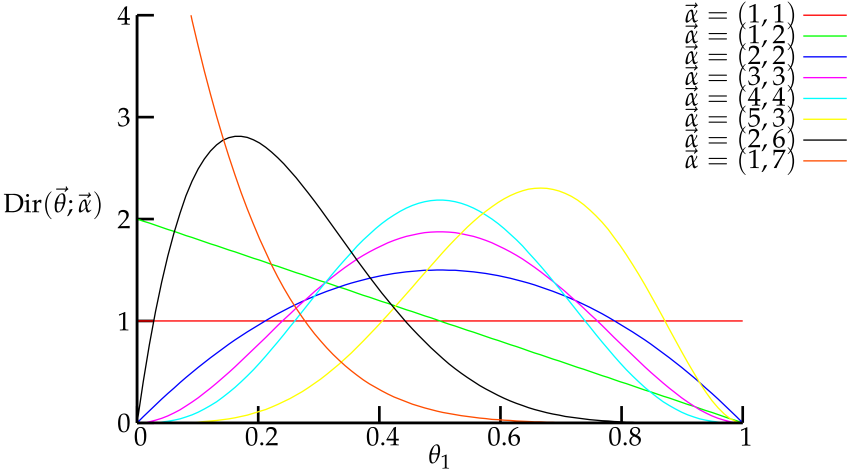
- Γ is a generalization of the factorial function
- $\Gamma(k) = (k-1)!$ for positive integer k
- $\Gamma(x) = (x-1)\Gamma(x-1)$ for all x

Plots of the Dirichlet distribution



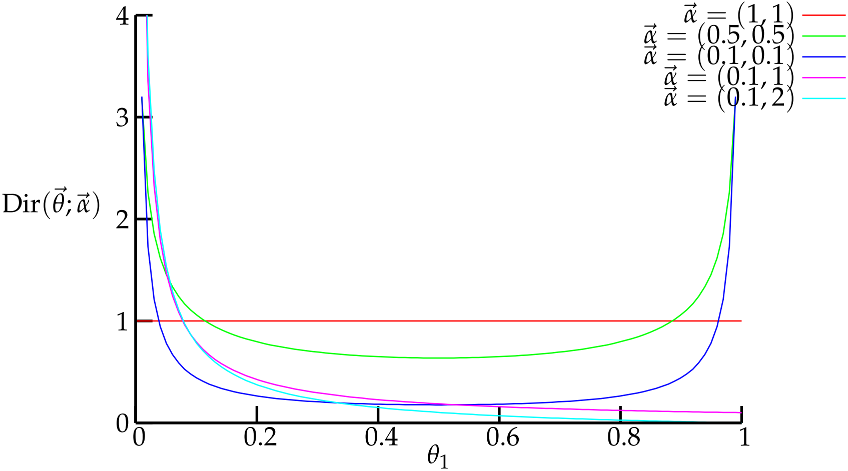
$$\mathrm{Dir}(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{m} \alpha_j)}{\prod_{j=1}^{m} \Gamma(\alpha_j)} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1}$$

Plots of the Dirichlet distribution (2)



$$\mathrm{Dir}(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{m} \alpha_j)}{\prod_{j=1}^{m} \Gamma(\alpha_j)} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1}$$

Sparse priors when $\boldsymbol{\alpha} < 1$



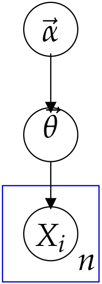
$$\mathrm{Dir}(\boldsymbol{\theta};\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{m} \alpha_j)}{\prod_{j=1}^{m} \Gamma(\alpha_j)} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1}$$

Dirichlet distributions as priors for $\boldsymbol{\theta}$

- Generative model:

$$\begin{array}{rcccl} \boldsymbol{\theta} & | & \boldsymbol{\alpha} & \sim & \mathrm{Dir}(\boldsymbol{\alpha}) \\ X_i & | & \boldsymbol{\theta} & \sim & \mathrm{Categorical}(\boldsymbol{\theta}), \quad i = 1, \ldots, n \end{array}$$

- We can depict this as a Bayes net using *plates*, which indicate *replication*



Inference for $\boldsymbol{\theta}$ with Dirichlet priors

- Data $\boldsymbol{X} = (X_1, \ldots, X_n)$ generated i.i.d. from Categorical$(\boldsymbol{\theta})$
- Prior is Dir$(\boldsymbol{\alpha})$. By Bayes Rule, posterior is:

$$\begin{aligned}
\mathrm{P}(\boldsymbol{\theta}|\boldsymbol{X}) &\propto \mathrm{P}(\boldsymbol{X}|\boldsymbol{\theta})\ \mathrm{P}(\boldsymbol{\theta}) \\
&\propto \left(\prod_{j=1}^{m} \theta_j^{N_j}\right) \left(\prod_{j=1}^{m} \theta_j^{\alpha_j - 1}\right) \\
&= \prod_{j=1}^{m} \theta_j^{N_j + \alpha_j - 1}, \text{ so} \\
\mathrm{P}(\boldsymbol{\theta}|\boldsymbol{X}) &= \mathrm{Dir}(\boldsymbol{N} + \boldsymbol{\alpha})
\end{aligned}$$

- So if prior is Dirichlet with parameters $\boldsymbol{\alpha}$,
  posterior is Dirichlet with parameters $\boldsymbol{N} + \boldsymbol{\alpha}$

$\Rightarrow$ can regard Dirichlet parameters $\boldsymbol{\alpha}$ as "pseudo-counts" from "pseudo-data"

Bayesian inference with hidden data

- Data consists of *visible* or *observed* variable x
- Model also involves a *hidden* or *latent* variable y
- Goal: estimate *joint* distribution over y and parameters θ

$$\mathrm{P}(y, \theta \mid x) \quad = \quad \frac{\mathrm{P}(y, x \mid \theta)\,\mathrm{P}(\theta)}{\mathrm{P}(x)}$$

- For most models this is intractable
  - Variational Bayes (assumes $\mathrm{P}(y, x \mid \theta) \approx Q(y)Q(\theta)$)
  - Markov Chain Monte Carlo sampling methods

Variational Bayes

- For any distribution $Q(y,\theta)$:

$$
\begin{aligned}
\log \mathrm{P}(x) &= F(Q) + \mathrm{KL}(Q \,\|\, \mathrm{P}(y,\theta \mid x)), \text{ where:} \\
\log \mathrm{P}(x) &= \log \sum_y \int \mathrm{P}(y,x,\theta)\, d\theta, \\
F(Q) &= \sum_y \int Q(y,\theta) \log \frac{\mathrm{P}(y,x,\theta)}{Q(y,\theta)}\, d\theta, \text{ and} \\
\mathrm{KL}(Q \,\|\, \mathrm{P}(y,\theta \mid x)) &= -\sum_y \int Q(y,\theta) \log \frac{\mathrm{P}(y,\theta \mid x)}{Q(y,\theta)}\, d\theta
\end{aligned}
$$

- Maximize $F \Leftrightarrow$ minimize KL-divergence

$\Rightarrow$ F is optimized when $Q(y,\theta) = \mathrm{P}(y,\theta \mid x)$

- Variational inference: optimize over a restricted class of Q functions

Mean field approximation in Variational Inference

- Mean field approximation: require that Q factorizes

$$Q(y, \theta) \;\; = \;\; Q(y)Q(\theta)$$

- In general $\mathrm{P}(y, \theta \mid x)$ does *not* factor
  - Cluster parameters θ vary depending on cluster assignment y
- But this may be approximately true
  - as data size grows, posterior becomes increasingly peaked

Mean field Variational Bayes

- Maximize F wrt $Q(y)$ and $Q(\theta)$

$$F(Q(y), Q(\theta)) = \sum_y \int Q(y)Q(\theta) \log \frac{\mathrm{P}(y, x \mid \theta)\mathrm{P}(\theta)}{Q(y)Q(\theta)} \, d\theta$$

- Add Lagrangians for constraints $\sum_y Q(y) = 1$ and $\int Q(\theta) \, d\theta = 1$, differentiate and set to zero
- Leads to an EM-like *alternating maximization procedure* for F
  - optimize $Q(y)$ while holding $Q(\theta)$ fixed
  - optimize $Q(\theta)$ while holding $Q(y)$ fixed

$$\begin{aligned} \log Q(y) &= \mathrm{E}_{Q(\theta)}[\log \mathrm{P}(y, x \mid \theta)] - \log Z \\ \log Q(\theta) &= \log \mathrm{P}(\theta) + \mathrm{E}_{Q(y)}[\log \mathrm{P}(y, x \mid \theta)] - \log Z' \end{aligned}$$

Variational Bayes for Dirichlet-Multinomials

$$\log \mathrm{P}(y, x \mid \boldsymbol{\theta}) = \log \prod_{j=1}^{\ell} \prod_{k=1}^{m_j} \theta_{j,k}^{n_{j,k}(x,y)} = \sum_{j=1}^{\ell} \sum_{k=1}^{m_j} n_{j,k}(x,y) \log \theta_{j,k}$$

$$\log \mathrm{P}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \log \prod_{j=1}^{\ell} \mathrm{Dir}(\boldsymbol{\theta}_j \mid \boldsymbol{\alpha}_j) = \sum_{j=1}^{\ell} \sum_{k=1}^{m_j} (\alpha_{j,k} - 1) \log \theta_{j,k} - c$$

Plugging these back into the VB mean-field formulae:

$$Q(\boldsymbol{\theta}) = \prod_{j=1}^{\ell} \mathrm{Dir}(\boldsymbol{\theta}_j \mid \boldsymbol{\alpha}'_j), \text{ where } \boldsymbol{\alpha}'_j = \boldsymbol{\alpha}_j + \mathrm{E}_{Q(y)}[\boldsymbol{n}_j]$$

$$Q(y) \propto \mathrm{P}(y, x \mid \boldsymbol{\theta}'), \text{ where } \log \theta'_{j,k} = \mathrm{E}_{Q(\boldsymbol{\theta})}[\log \theta_{j,k}]$$

$$\theta'_{j,k} = \exp\left(\Psi(\alpha'_{j,k}) - \Psi(\sum_{k'=1}^{m_j} \alpha'_{j,k'}) \right)$$

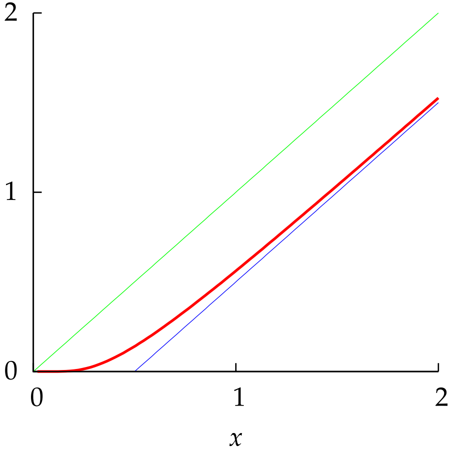
Mean-field Variational Bayes EM for Dirchlet-Multinomials

$$
\begin{aligned}
n_{j,k}^{(t)} &= \mathrm{E}_{\boldsymbol{\theta}^{(t)}}[n_{j,k}] \\
\theta_{j,k}^{(t+1)} &= \exp\left(\Psi(\alpha_{j,k} + n_{j,k}^{(t)}) - \Psi(\sum_{k'=1}^{m_j} \alpha_{j,k'} + n_{j,k'}^{(t)})\right) \\
&= \frac{\exp\left(\Psi(\alpha_{j,k} + n_{j,k}^{(t)})\right)}{\exp\left(\Psi(\sum_{k'=1}^{m_j} \alpha_{j,k'} + n_{j,k'}^{(t)})\right)}
\end{aligned}
$$

- E-step computes *expected counts*, just as in ordinary EM
- M-step is now more complicated
  - Add Dirichlet pseudo-count $\alpha_{j,k}$ to expected count $n_{j,k}^{(t)}$
  - Pass these through $\exp(\Psi(\cdot))$

Exponential of Digamma function

- The *digamma function* $\Psi(x) = d \log \Gamma(x)/dx$



- Plot shows $\exp(\Psi(x))$ function (in red), bounded by functions x and $x - 0.5$ (in green and blue respectively).

Things to be aware of

- $\boldsymbol{\theta}$ is *subnormalized*, i.e., $\sum_{k=1}^{m_j} \theta_{j,k}^{(t+1)} \leq 1$

$$\theta_{j,k}^{(t+1)} = \exp\left(\Psi(\alpha_{j,k} + n_{j,k}^{(t)}) - \Psi(\sum_{k'=1}^{m_j} \alpha_{j,k'} + n_{j,k'}^{(t)})\right)$$

- Each iteration updates

$$F(Q(y), Q(\theta)) = \log \mathrm{P}(x) - \mathrm{KL}(Q(y)Q(\theta) \parallel \mathrm{P}(y, \theta \mid x))$$

  so it does *not* always increase log likelihood ($\log \mathrm{P}(x \mid \theta)$)
- The notes describe how to compute F
- Code for computing the digamma function $\Psi(\cdot)$ is on course web page
- VB often takes longer to converge than EM