today: Bayesian approach
- model selection

**Variational methods vs. sampling**

Recall for mean field target dist.

$$\min_{q \in \mathcal{Q}} \text{KL}(q \parallel p)$$

$q \in \mathcal{Q}$ fully specifies distribution (where $q(x) = \mathcal{T}(q(x))$

in case where $p(x)$ was from Ising model

- can coordinate descent on $q_i$’s

approximate marginal

$$q_i(x_i=1) = \hat{p}_i$$

say we have converged to stationary $\pi_i^*$

usually $\hat{p}_i \neq p(x_i=1)$

"biased" / "unbiased"

in contrast: sampling methods are usually

asymptotically unbiased

examples: if use Gibbs sampling to get $x(t)$ (samples)

then

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} x(t) = p(x=1)$$

from Ergodic Thm

"asymptotically unbiased"

Problem: "mixing time" = how long it takes to forget initial conditions

sometimes sample very long "sticky chain"

slow convergence of Monte Carlo estimate

in practice, you can not use the first few samples to reduce the bias of estimate
Summary:

- Variational approach:
  - Often faster than sampling
  - But it is inexact

  e.g., QMR-DT network vs. variational vs. sampling

- Easier to debug the variational

Bayesian statistics

\[ p(x | \theta) \]

Model

\[ \theta \rightarrow \text{data} \]

Random

Statistics

"Subjective Bayesian"

- Use probability
  - Bayes rule

"Frequentist" [traditional statistics]

- Log of likelihoods
  - ML
  - Regularized ML
  - Max entropy
  - Moment matching
  - Etc...

Culculine: Bayesian is "optimistic" (think that can come up with good models)

- Obtain a method by pulling the Bayesian crank

Frequentist is more pessimistic: Use analysis tools

Bayesian:

\[ p(x | \theta) \text{ "likelihood model"} \]

\[ p(\theta) \text{ "prior"} \]

Posterior:

\[ p(\theta | x) = \frac{p(x | \theta) p(\theta)}{p(x)} \text{ Bayes rule} \]

\[ p(x) \text{ "marginal likelihood" normalization} \]
**Teaching Page 3**

**Example 6: biased coin**

**Bayesian model**

\[ x_i \in \{0, 1\} \]

\[ x_0 \text{ iid Bernoulli}(\Theta) \quad p(x_0 = 0) = \Theta \]

\[ \Theta \sim \text{Uniform}[0, 1] \quad \text{(uniform prior)} \]

**Graph model:**

\[
\begin{align*}
\Theta & \rightarrow x_0 \\
\Theta & \rightarrow x_1 \\
\Theta & \rightarrow x_n \\
\end{align*}
\]

\[
\begin{align*}
\Theta \in [0, 1] \\
\end{align*}
\]

**Posterior:**

\[ p(\Theta | x_1 : n) \propto p(x_1 : n | \Theta) p(\Theta) \]

\[ = \frac{\prod_{i=1}^{n} p(x_i | \Theta)}{p(\Theta)} \]

\[ = \prod_{i=1}^{n} \frac{\Theta^{x_i} (1-\Theta)^{1-x_i}}{\Theta^0 (1-\Theta)^0} \]

\[ = \frac{\Gamma(n+1)}{\Gamma(x_1 + 0)(\Gamma(n-x_1 + 1))} \]

\[ \sim \text{Beta}(x_1 + 1, n-x_1 + 1) \text{ distribution when } \alpha = n_1 + 1, \beta = n - n_1 + 1 \]

**Question:** What is probability of next flip = 1?

(\text{Frequentist}) \quad \hat{\Theta}_{\text{ML}} = \frac{n_1}{n}

(\text{Bayesian}) \quad \Theta \sim \text{Beta}(\alpha, \beta)

**Bayesian integrates out the uncertainty**

\[ p(x_{n+1} | x_1 : n) = \int \theta p(x_{n+1} | \theta) p(\theta | x_1 : n) \, d\theta \]

**Predictive distribution**

\[ p(x_{n+1} = 1 | x_1 : n) = \int \theta p(\theta | x_1 : n) \, d\theta = \text{posterior mean} \]
mean of Beta(\(\alpha, \beta\)) is \(\frac{\alpha}{\alpha + \beta} = \frac{n+1}{n+2}\)  \text{New}

notice that for \(n=0\) \(\to \) get \(\frac{1}{2}\)  \text{graphical version of ML}

\[
\hat{\theta}_{\text{MLE mean}} = \frac{n_x}{n} \left[ \frac{\alpha}{n+2} \right] + \frac{1}{2} \left[ \frac{\beta}{n+2} \right] = \frac{n+1}{n+2}
\]

\(= \hat{\theta}_{\text{MLE prior (1-F)}}\)

convex combination of \(\hat{\theta}_{\text{MLE}} \& \hat{\theta}_{\text{prior}}\)

\[\text{An} \overset{\text{as}}{\longrightarrow} \text{1}\]

variance of a beta is \(\frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}\)

\(= (n_x/n)(1-n_x/n) \text{O(n)}\)

\(\overset{\text{as}}{\longrightarrow} 0\)

posterior "contracts" around \(\hat{\theta}_{\text{PM}} \overset{\text{as}}{\longrightarrow} \hat{\theta}_{\text{MLE}} \& \text{true param}\)

\(\hat{\theta}_{\text{MLE}} \overset{\text{as}}{\longrightarrow} \theta^*\)

"Bernstein von- Mises thm"

\(\Rightarrow\) "Bayesian central limit thm." which basically says that if prior puts non-zero mass around true model \(\theta^*\) then posterior concentrates around \(\theta^*\) as a Gaussian asymptotically

for a Gaussian mean of mode are the same

so can approximate \(E[\theta|\text{data}]\)

with \(\hat{\theta}_{\text{MAP}} = \arg\max \rho(\theta|\text{data})\)

[true for large \(n\)]

revisiting the earlier example is:

\(\theta\) coins picked randomly

each flipped \(n\) times...
as a frequentist, empirical distribution on \(X_1: n\)
will converge as \(T \to \infty\)
to \(p(x_1, \ldots, x_n) = \int_{\Theta} \left(\prod_{i=1}^{n} p(x_i | \theta_i) d\theta_i\right) d\theta_i\)

(cotinuous) mixture distribution

Here, \(X_1, \ldots, X_n\) are not independent
on the other hand, \(p(x_1, \ldots, x_n) = p(X_{\pi(1)}, \ldots, X_{\pi(n)})\)
for any permutation \(\pi: 1:n \to 1:n\)
\(X_1, \ldots, X_n\) are "exchangeable" weaker than independence

De Finetti's representation theorem:
\(X_1, X_2, \ldots\) is infinitely exchangeable

\(\iff\) \(\exists\) unique \(p(\theta)\) on some space \(\Theta\)
\[p(x_1, \ldots, x_n) = \int_{\Theta} \left(\prod_{i=1}^{n} p(x_i | \theta_i) d\theta_i\right) d\theta_i\]

Multinomial model

e.g. modeling words...
\(X | \Theta \sim \text{Mult}(\Theta, 1)\) where \(\Theta \in \Delta^K\)
\(i.e. \sum_{k=1}^{K} \Theta_k = 1\)
\(\Theta_k \geq 0\)

\[\hat{\Theta}_k^{ML} = \frac{N_k}{n}\] 
if \(K > n\)
than some \(\hat{\Theta}_k^{ML} = 0\)
for some \(k \Rightarrow\) overfitting

as a Bayesian, put prior on \(\Delta_K = \Theta\)
a convenient property of prior family is "conjugacy"

consider family of distributions \(F = \{ p(\theta | a) : a \geq 0 \}\)
say that \(F\) is "conjugate family" to observation model \(p(x | \theta)\)
say that $F$ is a "conjugate family" to observation model $p(x | s)$ if posterior $p(s | x, p) = \frac{p(x | s) p(s | p)}{p(x | p)}$ for hyperparameter $\theta$

i.e. $E_a^\theta \text{ s.t. } p(s | x, \theta) = p(s | \theta)$

for multinomial $x$

likelihood $p(x | \alpha) = \frac{n!}{\prod_{i=1}^{k} \alpha_i!} \prod_{i=1}^{k} x_i^{\alpha_i}$

$\alpha = (\alpha_1, \cdots, \alpha_k)$

if use prior $\propto \prod_{i=1}^{k} \alpha_i^{\theta_{0i}}$

Dirichlet distribution $\rightarrow$ dist. over $\Delta_k$

$\text{Dir}(\theta | \alpha_1, \cdots, \alpha_k) = \frac{1}{\text{Beta}^k(\alpha)} \prod_{i=1}^{k} \alpha_i^{\theta_{0i}}$

valid $\alpha > 0$

$E[\theta_1 | \alpha_1, \cdots, \alpha_k] = \frac{\alpha_1}{\sum_1} \theta_{01}$

variance $(\theta_1) = O(\frac{\theta_{01}}{\sum_1})$

- $\alpha_1 = 1 \rightarrow \text{get uniform distribution}$
- $k = 2 \rightarrow \text{get Beta distribution}$
- $\alpha < 1 \cup \text{ shape distribution (no mode)}$
- $\alpha > 0 \cup \text{ unimodal bump}$

for multinomial model; if prior $p(s | \alpha) = \text{Dir}(\theta | \alpha)$

posterior $p(s | x_1, \cdots, x_n) \propto \prod_{i=1}^{k} \alpha_i^{x_i + \alpha_i - 1}$

$= \text{Dir}(\theta | (\alpha_0 + x_1, \cdots, \alpha_0 + x_k))$

posterior mean: $E[\theta_1 | \text{data}] = \frac{\alpha_1}{\sum_{1}^{k}}$
Bayesian linear regression

\[ y = \mathbf{x} \mathbf{w} + \varepsilon \]

observation model: \( p(y | \mathbf{x}, \mathbf{w}) = N(y | \mathbf{w} \mathbf{x}, \sigma^2) \)

prior: \( p(\mathbf{w}) = N(\mathbf{w} | \mathbf{0}, \frac{1}{N}) \)

(conjugate)

posterior: \( p(\mathbf{w} | y_0, \mathbf{x}_0, \mathbf{x}_{1:n}) \) is also Gaussian

with covariance \( \Sigma_n = N \mathbf{I} + \mathbf{X} \mathbf{X}^T \sigma^2 \)

posterior mean \( \hat{\mathbf{w}}_n = N \mathbf{X} \mathbf{X}^T \sigma^2 \)

La same as in ridge regression with \( \delta = \sigma^2 A \)

as a Bayesian:

compute predictive dist. \( p(y_{\text{new}} | y_{0:n}, x_{0:n}) \)

\[
\int_{\mathbf{w}} p(y_{\text{new}} | x_{\text{new}}, \mathbf{w}) p(\mathbf{w} | \text{data}) d\mathbf{w}
\]

\( \propto \) Gaussian prior \( \propto \) Gaussian posterior

\[ = N(y_{\text{new}} | \hat{\mathbf{w}}_n \mathbf{X}^T x_{\text{new}}, \sigma^2 \text{predictive}) \]

\( \sigma^2 \text{predictive} (x_{\text{new}}) = \sigma^2 + \frac{x_{\text{new}}^T \mathbf{X} \Sigma_n \mathbf{X}^T x_{\text{new}}}{N} \)

Model Selection

Say want to choose between \( M_1 \) vs. \( M_2 \)

\( M_1 \)

\( M_2 \)
as a Bayesain

\[
\begin{align*}
\hat{\Theta}^{M_2} &= \operatorname{argmax}_{\Theta_1, \Theta_2} \log p(\text{data} | G_1, \Theta_2, M_2) \\
\hat{\Theta}^{M_1} &= \operatorname{argmax}_{\Theta_1, \Theta_2} \log p(\text{data} | G_1, \tilde{\Theta}_2, M_2)
\end{align*}
\]

compare \( \log p(\text{data} | \hat{\Theta}^{M_1}, M_1) \) vs. \( \log p(\text{data} | \hat{\Theta}^{M_2}, M_2) \)?

\[ \text{we have } M_1 \prec M_2 \Rightarrow P < 1 \]

\[ \text{we have likelihood is useless...} \]

\[ \text{instead, use cross-validation} \]

\[ \text{\# here, Bayesian alternatives...} \]

true Bayesian, sum over models

\[ p(\text{new data}) = \sum_M \left( \int p(\text{new data} | G, M) \frac{p(M, G | D)}{p(G | D, M) p(M | \text{data})} \, d\Theta \right) \]

\[ = \sum_M p(M | \text{data}) \left[ \int p(\text{new data} | G, M) \frac{p(G | D, M) p(M | \text{data})}{p(G | D, M) p(M | \text{data})} \, d\Theta \right] \]

\[ \text{model averaging} \]

\[ \text{\# if force to pick one model...} \]

\[ \text{pick model which maximizes } p(M | \text{data}) p(\text{data} | M) p(M) \]

\[ \text{\# marginal likelihood} \]

\[ p(D | M) = \int p(D | G, M) p(G | M) \, d\Theta \]

\[ \text{here, step by step } \int p(M | \text{data}) p(\text{data} | M) \, d\Theta \]
Bayesian factor: 

\[
\frac{p(M_1 | D)}{p(M_2 | D)} = \frac{p(D | M_1) p(M_1)}{p(D | M_2) p(M_2)}
\]

Pick \( M = \text{arg} \max \ p(D | M) \) as "empirical Bayes"
(\( \text{Type 2 ML} \))

Too many models \( \Rightarrow \) can overfit

Consider why marginal likelihood works vs. ML's:

\( M_1 \subseteq M_2 \subseteq M_3 \)

BIC criterion \( \Rightarrow \) approximation \( p(D | M) \)