14 Bayesian Estimation

– Thomas Bayes (18th-century mathematician and statistician)
– Sir Harold Jeffreys (famous 20th-century mathematician and statistician) wrote that Bayes' theorem "is to the theory of probability what Pythagoras's theorem is to geometry"

14.1 Review: Properties of ML Estimator

– Data: i.i.d. sample of size n drawn from $P(X|\theta)$

– Consistency: the sequence of MLE estimates $\widehat{\theta}$ converges in probability to the true parameter value θ

– Asymptotic Normality: as the sample size increases, the distribution of the MLE tends to the Gaussian distribution with mean θ (and covariance matrix equal to the inverse of the Fisher information matrix)

– Efficiency: No consistent estimator has lower asymptotic mean squared error than the ML estimator (ML estimator achieves the Cramer-Rao lower bound when the sample size tends to infinity)

14.2 Bayes' Rule / Theorem

For events A and B, $P(A|B) = P(B|A)P(A)/P(B)$
– Proof follows from our definition of conditional probability, i.e., $P(X|Y) := P(X \cap Y)/P(Y)$

14.3 Example (Coin Flip)

– Consider that we don't know if a coin is fair / unfair
– We have 2 possibilities in our mind:
(1) Coin fair, i.e., $P(\text{head}) = p = 0.5$
(2) Coin biased towards heads with $P(\text{head}) = q = 0.7$
– We have a belief (**prior** to observing data) that $P(\text{CoinFair}) = 0.8$
– Now we experiment with the coin, collect data, and recompute the probability that the coin is fair

$$P(\text{CoinFair}|\text{Data}) = P(\text{Data}|\text{CoinFair})P(\text{CoinFair})/P(\text{Data})$$

– Given: We have data = n observations with r heads and $(n-r)$ tails. What does the data do to our belief ?

$$\begin{aligned} P(\text{Data}|\text{CoinFair}) &= C_r^n 0.5^r 0.5^{n-r} \\ P(\text{Data}|\text{CoinUnfair}) &= C_r^n 0.7^r 0.3^{n-r} \\ P(\text{Data}) &= P(\text{Data}|\text{CoinFair})P(\text{CoinFair}) + P(\text{Data}|\text{CoinUnfair})P(\text{CoinUnfair}) \\ P(\text{CoinFair}|\text{Data}) &= \frac{0.5^r 0.5^{n-r} \times 0.8}{0.5^r 0.5^{n-r} \times 0.8 + 0.7^r 0.3^{n-r} \times 0.2} \end{aligned}$$

– **Case 1:** If $n = 20, r = 11$, then $P(\text{CoinFair}|\text{Data}) = 0.9074$ which is more than 0.8. So the data has strengthened our belief !!
– Why has this happened ? Because 11 heads out of 20 is more like the fair coin.

– **Case 2:** If $n = 20, r = 13$, then $P(\text{CoinFair}|\text{Data}) = 0.6429$ which is less than 0.8. So the data has weakened our belief !!
– Why has this happened ? Because 13 heads out of 20 is more like the unfair coin.

14.4 Example (Box)

There are two boxes:
(i) one with 4 black balls and 1 white ball
(ii) another with 1 black ball and 3 white balls

You pick one box at random (*prior* probability of picking any box is 0.5).

Then select a ball from the box. It turns out to be white (*data*).

Given that the ball is white, what is the probability that you picked the 1st box ?

Solution: $P(Box1|W) = P(W|Box1)P(Box1)/P(W)$ where,
using total probability, $P(W) = P(W|Box1)P(Box1) + P(W|Box2)P(Box2)$

14.5 Example: Gaussian (Unknown mean, Known variance)

– Given: Data $\{x_i\}_{i=1}^N$ derived from a Gaussian distribution with known variance σ^2, but unknown mean μ

– *Treat mean μ as a random variable*

– Prior belief on μ is that it is derived from a Gaussian with mean μ_0 and variance σ_0^2

– Associated Generative Model here: first draw μ from prior, then draw data given μ

– Goal: Estimate μ, given prior and data

– What if we ignore the prior ? (ML estimation seen before)
– What if we ignore the likelihood / data ? ($\mu = \mu_0$)

– A possible solution: Maximize posterior w.r.t. μ
Posterior: $P(\mu|x_1, \cdots, x_N) = P(x_1, \cdots, x_N|\mu)P(\mu)/P(x_1, \cdots, x_N)$
Assume sample mean = $\bar{x}$.
Then MAP estimate for the mean is :

$$\mu = \frac{\bar{x}\sigma_0^2 + \mu_0\sigma^2/N}{\sigma_0^2 + \sigma^2/N}$$

– What if $N = 1$?
– What if $N \to \infty$? (data dominates the prior)
– What if $\sigma_0 \to \infty$? (weak prior: ignore the prior)
– What if $\sigma_0 \to 0$? (strong prior: ignore the data)

14.6 Posterior Mean Estimate to Minimize MSE

– Given data: $\{x_i\}_{i=1}^n$ drawn from $P(X|\theta)$

– We have a prior $P(\theta)$ on RV θ

– Posterior = conditional density $P(\theta|x_1, \cdots, x_n) = \frac{P(x_1, \cdots, x_n|\theta)P(\theta)}{\int_\theta P(x_1, \cdots, x_n, \theta)d\theta}$

– Question: Given a PDF $P(\theta|x_1, \cdots, x_n)$ on the true parameter θ, what is the best estimate $\widehat{\theta}^*$ to minimize mean squared error $E_{P(\theta|x_1, \cdots, x_n)}[(\widehat{\theta} - \theta)^2]$?

– Answer: The PDF mean $E_{P(\theta|x_1, \cdots, x_n)}[\theta]$. This is also a Bayes estimate.

14.7 Loss functions and Risk functions

– Loss function $L(\widehat{\theta}|\theta)$ = loss incurred for estimating $\widehat{\theta}$, when the true value is θ

– Risk function $R(\widehat{\theta}|\theta)$ = expected loss = expectation of the loss function under the posterior PDF $P(\theta|x_1, \cdots, x_n)$

– Choose $\widehat{\theta}$ to minimize risk

– Example: Squared-error loss function: $L(\widehat{\theta}) = (\widehat{\theta} - \theta)^2$

Risk function $= E_{P(\theta|x_1,\cdots,x_n)}[(\widehat{\theta} - \theta)^2]$ = mean squared error

Let risk minimizer = θ^*

Then, $\frac{\partial}{\partial \widehat{\theta}} E_{P(\theta|x_1,\cdots,x_n)}[(\widehat{\theta} - \theta)^2]\Big|_{\widehat{\theta}=\theta^*} = 0$

Thus, $\theta^* = E_{P(\theta|x_1,\cdots,x_n)}[\theta] =$ Posterior mean

– Example: Zero-one loss function (case of discrete RV θ): $L(\widehat{\theta}) = I(\widehat{\theta} \neq \theta)$

Risk function $= R(\widehat{\theta}) = E_{P(\theta|x_1,\cdots,x_n)}[I(\widehat{\theta} \neq \theta)]$

$= \sum_{\theta \neq \widehat{\theta}} P(\theta|x_1, \cdots, x_n)$
$= 1 - P(\theta = \widehat{\theta}|x_1, \cdots, x_n)$

Thus, the risk function is minimized when $\widehat{\theta} = \arg\max_\theta P(\theta|x_1, \cdots, x_n)$ = MAP estimate

– Example: Zero-one loss function (case of continuous RV θ)

Assume that the loss function is an *inverted* rectangular pulse —_— with height 1 and an infinitesimally small width $\epsilon > 0$ (we do NOT make $\epsilon = 0$), with center of the pulse at the true parameter value θ. i.e.,
$L(\widehat{\theta}|\theta) = 0$; if $\theta \in (\widehat{\theta} - \epsilon/2, \widehat{\theta} + \epsilon/2)$
$L(\widehat{\theta}|\theta) = 1$; otherwise

For such a loss function, the risk function $1 - \int_{\widehat{\theta}-\epsilon/2}^{\widehat{\theta}+\epsilon/2} P(\theta|x_1, \cdots, x_n)$ is minimized when the pulse center is placed at the mode of the PDF.

– Example: Absolute-error loss function $L(\widehat{\theta}) = |\widehat{\theta} - \theta|$

Risk function $= E_{P(\theta|x)}[|\widehat{\theta} - \theta|]$

$= \int_{-\infty}^{\infty} |\widehat{\theta} - \theta| P(\theta|x) d\theta$
$= \int_{-\infty}^{\widehat{\theta}} (\widehat{\theta} - \theta) P(\theta|x) d\theta + \int_{\widehat{\theta}}^{\infty} (\theta - \widehat{\theta}) P(\theta|x) d\theta$

The risk function is minimized when its derivative is zero.

How to take the derivative of an integral where the limits are also a function of the variable of interest ?
Leibniz's Integral Rule (draw picture):
$\frac{\partial}{\partial a} \int_{l(a)}^{u(a)} f(z, a) dz = \int_{l(a)}^{u(a)} \frac{\partial f}{\partial a} dz + f(z = u(a), a) \frac{\partial u}{\partial a} - f(z = l(a), a) \frac{\partial l}{\partial a}$

In our case, $f(z \equiv \theta, a \equiv \widehat{\theta}) \propto (\widehat{\theta} - \theta) P(\theta|x)$
In our case, for the 1st integral: $f(z = u(a), a) = 0$ and the lower-limit term doesn't arise
In our case, for the 2nd integral: $f(z = l(a), a) = 0$ and the upper-limit term doesn't arise
Thus, the derivative of our risk function w.r.t. $\widehat{\theta}$ is:
$= \int_{-\infty}^{\widehat{\theta}} (+1) P(\theta|x) d\theta + \int_{\widehat{\theta}}^{\infty} (-1) P(\theta|x) d\theta$

$= \int_{-\infty}^{\widehat{\theta}} P(\theta|x) d\theta - \int_{\widehat{\theta}}^{\infty} P(\theta|x) d\theta$
This is zero when $\widehat{\theta}$ = median of $P(\theta|x)$

The median will be a minimizer if the 2nd derivative is positive. Is that so ?

In this case, for both integrals, $\frac{\partial f}{\partial a} = 0$
In this case, for 1st integral, the lower-limit term doesn't arise
In this case, for 2nd integral, the upper-limit term doesn't arise
Thus, the 2nd derivative of our risk function w.r.t. $\widehat{\theta}$, evaluated at $\widehat{\theta}$ = median of $P(\theta|x)$, is:
$= P(\widehat{\theta}|x) + P(\widehat{\theta}|x) \geq 0$

Note: the median $\widehat{\theta}$ isn't unique if $P(\widehat{\theta}|x) = 0$

14.8 Example: i.i.d. Bernoulli

– Given: $X_1, \cdots, X_n$ are i.i.d. Bernoulli with parameter θ and PDF $P(x = 1|\theta) = \theta, P(x = 0|\theta) = 1 - \theta$

– Data: $x_1, \cdots, x_n$

– Estimate $\theta \in (0, 1)$

– Prior $P(\theta) = 1, \forall\theta \in (0, 1)$

– Answer:

Rewrite PDF as $P(x|\theta) = \theta^x(1 - \theta)^{1-x}$, where $x \in \{0, 1\}$

$P(\theta|x_1, \cdots, x_n) = P(x_1, \cdots, x_n|\theta)/P(x_1, \cdots, x_n)$
where
Numerator $= \theta^{\sum_i x_i}(1 - \theta)^{n-\sum_i x_i}$
Denominator $= \int_0^1 \theta^{\sum_i x_i}(1 - \theta)^{n-\sum_i x_i} d\theta$

Now we exploit the result / trick: $\int_0^1 \theta^m(1 - \theta)^r d\theta = m!r!/(m + r + 1)!$

Let $x = \sum_i x_i$
Then, $P(\theta|x_1, \cdots, x_n) = \frac{(n+1)!}{x!(n-x)!}\theta^x(1 - \theta)^{n-x}$

Thus, $E_{P(\theta|x_1,\cdots,x_n)}[\theta] = \int_0^1 \theta \frac{(n+1)!}{x!(n-x)!}\theta^x(1 - \theta)^{n-x} d\theta = \frac{x+1}{n+2}$

Thus, Bayes estimator $= \frac{\sum_i X_i+1}{n+2}$

Note: ML estimator $= \max_\theta \log \left(\theta^{\sum_i X_i}(1 - \theta)^{n-\sum_i X_i}\right)$
$= \max_\theta X \log\theta + (n - X) \log(1 - \theta)$, where $X := \sum_i X_i$
$= X/n$
$= \sum_i X_i/n$

Check that the 2nd derivative is negative (Use the facts: $X \geq 0$ and $n - X \geq 0$ and $0 < \theta < 1$)

Note: Asymptotically, i.e., as $n \to \infty$, the Bayes estimator tends to the ML estimator

14.9 Example: i.i.d. Gaussian

– Given: $X_1, \cdots, X_n$ i.i.d. $G(\theta, \sigma_0^2)$. Unknown mean. Known variance.
– Prior: $P(\theta) := G(\theta; \mu; \sigma^2)$

– Bayes estimate = posterior mean = ?

– Answer:

Property 1: Product of 2 Gaussians is another Gaussian: $G(z;\mu_1,\sigma_1^2)G(z;\mu_2,\sigma_2^2) \propto G(z;\mu_3,\sigma_3^2)$

Numerator exponent $= \frac{(z-\mu_1)^2}{2\sigma_1^2} + \frac{(z-\mu_2)^2}{2\sigma_2^2}$
$= \frac{1}{2\sigma_1^2\sigma_2^2}\left(z^2(\sigma_2^2+\sigma_1^2) - (2\mu_1\sigma_2^2 + 2\mu_2\sigma_1^2)z + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_2^2\right)$
$= \frac{1}{2\sigma_1^2\sigma_2^2}\left(z^2(\sigma_2^2+\sigma_1^2) - (2\mu_1\sigma_2^2 + 2\mu_2\sigma_1^2)z\right) + c$, where $c =$ constant independent of z
$= \frac{\sigma_2^2+\sigma_1^2}{2\sigma_1^2\sigma_2^2}\left(z^2 - \frac{2\mu_1\sigma_2^2+2\mu_2\sigma_1^2}{\sigma_2^2+\sigma_1^2}z\right) + c$, where $c =$ constant independent of z
$= \frac{\sigma_2^2+\sigma_1^2}{2\sigma_1^2\sigma_2^2}\left(z^2 - 2\mu_3 z + \mu_3^2\right) + c'$, where $c' =$ constant independent of z and where $\mu_3 = \frac{\mu_1\sigma_2^2+\mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}$
$= \frac{1}{2\sigma_3^2}(z-\mu_3)^2 + c'$, where $c' =$ constant independent of z where $\sigma_3^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$

In our case, we have two PDFs on θ, i.e.,

– Prior $P(\theta) = \frac{1}{\sigma\sqrt{2\pi}}\exp((\theta-\mu)^2/(2\sigma^2)) = G(\theta;\mu,\sigma^2)$

– Likelihood $P(x_1,\cdots,x_n|\theta) = \frac{1}{(2\pi)^{n/2}\sigma_0^n}\exp(-\sum_i(x_i-\theta)^2/(2\sigma_0^2)) = G(\theta;x_1,\sigma_0^2)\cdots G(\theta;x_n,\sigma_0^2)$

The negative exponent here can be written as:
$(n\theta^2 - 2(\sum_i x_i)\theta)/(2\sigma_0^2) + c$, where $c =$ constant independent of θ
$= (\theta^2 - 2(\sum_i x_i/n)\theta)/(2\sigma_0^2/n) + c$
$\propto G(\theta;\sum_i x_i/n, \sigma_0^2/n)$

Let $x = \sum_i x_i/n$

Thus, the (normalized) product of the prior and the likelihood gives a Gaussian $G(\theta;\mu^*,\sigma^{*2})$, where
$\mu^* = \frac{\mu\sigma_0^2/n + x\sigma^2}{\sigma^2+\sigma_0^2/n}, \sigma^{*2} = \frac{\sigma^2\sigma_0^2/n}{\sigma^2+\sigma_0^2/n}$

Bayes estimate = mean of posterior = μ^*, which also happens to be the Gaussian posterior's mode = MAP estimate

Note: As the data sample size $n \to \infty$, the mean $\mu^* \to x$ and variance $\sigma^{*2} \to 0$.
Thus, the posterior becomes a delta function at $\theta = x =$ sample mean
In this case, the Bayes estimate converges to the ML estimate = sample mean

14.10 MAP Estimation and ML Estimation

– Consider the likelihood function $P(x_1,\cdots,x_n|\theta)$

– Consider prior $P(\theta) = 1/(b-a)$ for $\theta \in (a,b)$, i.e., a uniform distribution over (a,b)

– Then, posterior PDF $= \frac{P(x_1,\cdots,x_n|\theta)P(\theta)}{\int_a^b P(x_1,\cdots,x_n|\theta)P(\theta)d\theta}$, for $\theta \in (a,b)$
$= \frac{P(x_1,\cdots,x_n|\theta)}{\int_a^b P(x_1,\cdots,x_n|\theta)d\theta}$, for $\theta \in (a,b)$

– Maximum of the posterior within (a,b) = maximum of $P(x_1,\cdots,x_n|\theta)$ within (a,b)
– If the mode of the likelihood function lied within (a,b), then the mode of the posterior $\equiv$ ML estimate

14.11 Bayes Interval Estimate

– Previous analysis gives a point estimate for the parameter θ
– How do we get an interval estimate for the parameter θ ?
– We can do this by finding a,b such that $\int_a^b P(\theta|x_1,\cdots,x_n)d\theta = 1-\alpha$, where probability α is given.
– We can get such information in some special cases, relatively easily

14.11.1 Example: Gaussian

– Question: Suppose signal of value s is sent from A to B.
Because of the noisy communication channel, signal received at B has a Gaussian PDF with mean s and variance 60.
A priori, it is known that the signal s being sent is selected from a Gaussian PDF with mean 50 and variance 100.
Given, value received at B is 40.
Find an interval (a, b) s.t. the probability of the signal being in that interval is 0.9

– Answer:
Using formulas derived before for the posterior $P(s|x_1 = 40)$ of parameter s given data x_1,
Posterior mean = $\frac{50*60+40*100}{60+100} = 43.75$
Posterior variance = $\frac{60*100}{60+100} = 37.5$

We know that the posterior PDF is Gaussian
Thus, $Z := \frac{S-43.75}{\sqrt{37.5}}$ has a standard Normal PDF
For a standard Normal PDF, we know that the probability mass within $Z \in (-1.645, +1.645)$ is 0.9

Thus, we want to find S s.t. $P(-1.645 < Z < 1.645|\text{data}) = 0.9$
i.e., $P(-1.645 < \frac{S-43.75}{\sqrt{37.5}} < 1.645|\text{data}) = 0.9$
i.e., $P(33.68 < S < 53.83|\text{data}) = 0.9$

Thus, the desired interval is $(a = 33.68, b = 53.83)$

14.12 Conjugate Priors

– If the posterior PDFs $P(\theta|x)$ are in the same family as the prior PDF $P(\theta)$, then:
(i) the prior and posterior are called *conjugate* PDFs, and
(ii) the prior is called the conjugate prior for the likelihood function

– Advantage of conjugate priors: The posterior has a closed-form expression because the denominator / normalization constant has a closed-form expression
$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\int P(x|\theta)P(\theta)d\theta}$
Otherwise, a hard numerical integration may be required to approximate the normalization factor

– Example: Binomial Likelihood and Beta prior

1) Likelihood of s successes in n tries: $P(s, n|\theta) =^n C_s \theta^s (1-\theta)^{n-s}$
2) Prior: $P(\theta) = \text{beta}(\theta; a > 0, b > 0) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}$
3) Posterior $\propto \theta^{s+a-1}(1-\theta)^{n-s+b-1} \equiv \text{beta}(\theta; a+s, b+n-s)$

We know that the mean of the beta PDF beta$(\theta; a, b)$ is $a/(a+b)$

Thus, Bayes estimate = posterior mean = $(a+s)/(a+b+n)$
$= w(a/(a+b)) + (1-w)s/n$, where weight $w = (a+b)/(a+b+n)$

Note: When the sample size $n = 0$, the posterior mean $= a/(a+b) =$ prior mean

Note: As the sample size $n \to \infty$, the weight $w \to 0$ and the posterior mean $\to$ ML estimate

If prior $P(\theta) = 1$ is uniform over $\theta \in (0, 1)$, i.e., beta$(\theta, 1, 1)$
In that case, the likelihood determines the posterior

– Example: Gaussian (known mean μ, unknown variance θ)

1) Likelihood: $P(x_1, \cdots, x_n|\mu, \theta) \propto \prod_{i=1}^n \theta^{-0.5} \exp(-0.5(x_i - \mu)^2/\theta)$

2) Prior = Inverse Gamma PDF: $P(\theta; a, b) \propto \theta^{-a-1} \exp(-b/\theta)$, where $a > 0, b > 0$
3) Posterior = Inverse Gamma PDF: $P(\theta; a + n/2, b + \sum_i (x_i - \mu)^2/2)$

Mean of the inverse Gamma $P(\theta; a, b) = b/(a-1)$, for $a > 1$

Thus, Bayes estimate = posterior mean = $(b + \sum_i (x_i - \mu)^2/2)/(a + n/2 - 1)$
$= (2b + \sum_i (x_i - \mu)^2)/(2a + n - 2)$
$= w(b/(a-1)) + (1-w) \sum_i (x_i - \mu)^2/n$, where weight $w = (2a-2)/(2a+n-2)$

Note: When the sample size $n = 0$, the weight $w = 1$ and the posterior mean $= b/(a-1) =$ prior mean

Note: As the sample size $n \to \infty$, the weight $w \to 0$ and the posterior mean $\to$ ML estimate

• An *"uninformative"* (misnomer) prior for the Gaussian mean is the (improper) uniform PDF.

– Why improper ? Because it doesn't integrate to a finite number

– Why uninformative ? Because:
i) posterior PDF driven by the likelihood function alone
ii) the prior on θ is invariant to translation of the data x_i (Duda-Hart-Stork). Note: translation of data also implies that the MLE estimate of the mean also gets translated.

– Uninformative priors express "objective" (impersonal; unaffected by personal beliefs) information such as "the variable is positive" or "the variable is less than some limit".
– Uninformative priors yield results close to what we would get with ML (non-Bayesian) analysis

• An *"uninformative"* (and improper) prior for the Gaussian standard deciation σ is $P(\sigma) = 1/sigma$

– Why uninformative ? Because of scale invariance, as follows.
Consider the RVs $\log(x)$ and $\log(\sigma)$. If the data x get scaled (which implies that the MLE for the standard deviation σ also gets scaled) in the original domain by factor a, then a term $\log(a)$ gets added in the log domain. A scale-invariant prior implies that the prior leads to a uniform PDF on σ in the $\log(\sigma)$ domain.
Transform the RV $u = \log(\sigma)$ with $P(U) = c$, to get $v = \exp(u)$. Transformation of variables implies that $P(\sigma) = c/\sigma$.

– Example: Poisson PDF and Gamma prior

Use this example to motivate the general result for exponential families later
1) Likelihood: $P(k_1, \cdots, k_n|\lambda) = \prod_i \lambda^{k_i} \exp(-\lambda)/k_i!$, where $\lambda > 0, k_i > 0$
2) Prior: $P(\theta) =$ Gamma$(\lambda|\alpha, \beta) \propto \lambda^{\alpha-1} \exp(-\beta\lambda)$, where $\alpha > 0, \beta > 0, \lambda > 0$
3) Posterior: $\propto \lambda^{\sum_i k_i + \alpha - 1} \exp(-n\lambda - \beta\lambda) \equiv$ Gamma$(\lambda; \sum_i k_i + \alpha, n + \beta)$

For a Gamma distribution Gamma$(\lambda|\alpha, \beta)$, we know that the mean is α/β

Thus, the Bayes estimate = posterior mean = $(\sum_i k_i + \alpha)/(n + \beta)$
$= w(\alpha/\beta) + (1-w) \sum_i k_i/n$, where weight $w = \beta/(\beta + n)$
$= w(\alpha/\beta) + (1-w)\widehat{\lambda}_{\text{MLE}}$

Note: When the sample size $n = 0$, the weight $w = 1$ and the posterior mean $= \alpha/\beta =$ prior mean

Note: As the sample size $n \to \infty$, the weight $w \to 0$ and the posterior mean $\to$ ML estimate

Note: It is good that the prior gets ignored when the sample size becomes infinite; our beliefs shouldn't affect results when we have infinite data

14.13 Exponential Family of PDFs

– Interesting result: All PDFs in the exponential family have conjugate priors.

– Definition: A single-parameter exponential family is a set of PDFs where each PDF can be expressed in the form:
$P(x|\theta) = \exp(\eta(\theta)T(x) - A(\theta) + B(x)) = h(x)g(\theta)\exp(\eta(\theta)T(x))$
where $T(x), B(x), \eta(\theta), A(\theta)$ are known functions.

– Interpretation: The parameters θ and observation variables x must *factorize* either directly or within either part of an exponential operation

– Example: Gaussian

– Counter example: $P(x|\theta) = [f(x)g(\theta)]^{h(x)j(\theta)} = \exp([h(x)\log f(x)]j(\theta) + h(x)[j(\theta)\log g(\theta)])$

• How do we go about guessing what the conjugate prior is ?

– Consider the *canonical form* of the exponential family where $\eta(\theta) := \theta$, i.e., $\eta(\cdot)$ is identity
Note: It is always possible to convert an exponential family to canonical form, by defining a transformed parameter $\theta' = \eta(\theta)$

– Step (1) For the exponential family, the likelihood function for data $\{x_i\}_{i=1}^N$ is:
$L(\theta|x_1, \cdots, x_N) = (\Pi_i \exp(B(x_i)))\exp(\theta(\sum_i T(x_i)) - NA(\theta))$

– Step (2) Consider the prior $P(\theta|\alpha, \beta) = H(\alpha, \beta)\exp(\alpha\theta - \beta A(\theta))$

Diaconis and Ylvisaker 1979 gave conditions on the hyper-parameters α, β under which this PDF is integrable (i.e., proper)

– Step (3) The posterior PDF $\propto \exp(\theta(\alpha + \sum_i T(x_i)) - (\beta + N)A(\theta))$ that belongs to the exponential family w.r.t. variable θ and has the same form as the prior
The conversion from the prior to the posterior simply replaces $\alpha \to \alpha + \sum_i T(x_i)$ and $\beta \to \beta + N$

Because the prior can be normalized, so can the posterior