Recall from MAP estimation example:

$f(x|\mu) = n(\mu, \sigma^2)$ (σ^2 known) $\qquad D = (x_1, \dots, x_n)$ observed samples

$f(\mu) = n(\mu_0, \sigma_0^2)$ (μ_0, σ_0^2 known)

$$\underbrace{f(\mu|D)}_{\text{posterior}} = \frac{f(D|\mu)f(\mu)}{f(D)} \propto f(D|\mu)f(\mu)$$

$$= \prod_{k=1}^{n} f(x_k|\mu) f(\mu)$$

$$= \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}$$

$$\Rightarrow f(\mu|D) \propto e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}$$

$$\propto e^{-\frac{1}{2\sigma^2}\sum_k (x_k^2 - 2x_k\mu + \mu^2) - \frac{1}{2\sigma_0^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2)}$$

$$\propto e^{-\frac{1}{2}\left\{\underbrace{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}_{\frac{1}{\sigma_n^2}}\mu^2 - 2\underbrace{\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)}_{\frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2}}\mu\right\}}$$

$$\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k = \hat{\mu}_{ML}$$

sample mean

$$\begin{cases} \dfrac{1}{\sigma_n^2} = \dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^2} & \text{or} \quad \sigma_n^2 = \dfrac{\sigma_0^2\sigma^2}{\sigma^2 + \sigma_0^2 n} \\[2ex] \dfrac{\mu_n}{\sigma_n^2} = \dfrac{n}{\sigma^2}\hat{\mu}_n + \dfrac{\mu_0}{\sigma_0^2} & \text{or} \quad \mu_n = \left(\dfrac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\hat{\mu}_n + \dfrac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 \end{cases}$$

Note $f(\mu|D)$ is exponential function of a <u>quadratic</u> function of μ : <u>normal</u> density

$f(\mu|D)$ is called <u>reproducing</u> <u>density</u>

$f(\mu)$ is called <u>conjugate prior</u>.

write $f(\mu|D) = \mathcal{N}(\mu_n, \sigma_n^2)$

$$= \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_n}\right)^2\right]$$

i.e. $f(\mu|D) = \mathcal{N}(\hat{\mu}_{MAP}, \sigma_n^2)$

- $\mu_n = \hat{\mu}_{MAP}$ represents best guess for μ after observing n samples
- σ_n^2 measures uncertainty about this guess
- σ_n^2 decreases monotonically with n, $\to 0$ as σ^2/n
  each additional observation decreases uncertainty about true value of μ
- As n increases, $f(\mu|D)$ becomes more and more sharply peaked (see figure #40).

<u>Bayesian parameter estimation</u>

- want to estimate density $f(\underline{x}|D)$
- Assume $f(\underline{x}|\underline{\theta})$ known
  prior distribution on $\underline{\theta}$: $f(\underline{\theta})$
  a posterior dist. on $\underline{\theta}$: $f(\theta|D)$
- $$f(\underline{x}|D) = \int f(\underline{x}, \underline{\theta}|D)\, d\underline{\theta}$$
  $$= \int f(\underline{x}|\underline{\theta}, D)\, f(\underline{\theta}|D)\, d\underline{\theta}$$

- since $\underline{x}$ and D selected independently,
  $f(\underline{x}|\underline{\theta}, D) = f(\underline{x}|\underline{\theta})$

- $f(\underline{x}|D) = \int f(\underline{x}|\underline{\theta})\, f(\underline{\theta}|D)\, d\underline{\theta}$

- If $f(\underline{\theta}|D)$ peaks sharply about $\hat{\underline{\theta}}$, then $f(\underline{x}|D) \sim f(\underline{x}|\hat{\underline{\theta}})$
- But there is usual uncertainty about exact value of $\underline{\theta}$, so average $f(\underline{x}|\underline{\theta})$ over $\underline{\theta}$.

Back to Gaussian 1-D example:

Assume $f(x|\mu) = n(\mu, \underbrace{\sigma^2}_{\text{known}})$

Assume μ is a r.v. $\sim f(\mu) = n(\mu_0, \sigma_0^2)$

μ_0: best prior guess for μ

σ_0^2: uncertainty about prior guess.

Observed samples $D = \{x_1, \ldots, x_n\}$

$$f(x|D) = \int f(x|\mu)\, \underbrace{f(\mu|D)}_{\text{posterior}}\, d\mu$$

$$= \int n(x|\mu, \sigma^2)\, n(\mu|\mu_n, \sigma_n^2)\, d\mu$$

$$= n(x|\mu_n, \sigma^2 + \sigma_n^2)$$

i.e. $$f(x|D) = \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \cdot \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d\mu$$

$$\propto \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right]$$

$$\Rightarrow f(x|D) = n(\mu_n, \sigma^2+\sigma_n^2)$$

- $f(x|\mu) = n(\mu, \sigma^2) \longrightarrow f(x|D) = n(\mu_n, \sigma^2+\sigma_n^2)$
- Conditional mean μ_n treated as true mean
  - Variance increased to account for additional uncertainty in x from lack of exact knowledge about μ. $(\sigma^2 \to \sigma^2+\sigma_n^2)$
  - Contrast Bayesian approach with ML methods which only estimate $\hat{\mu}$ and $\hat{\sigma}^2$, rather than estimate a distribution $p(x|D)$.

<u>Multivariate case</u>

$$f(\underline{x}|\underline{\mu}) = n(\underset{\text{unknown}}{\underline{\mu}}, \underset{\text{known}}{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu})}$$

$$f(\underline{\mu}) = n(\underbrace{\underline{\mu}_0, \Sigma_0}_{\text{known}}) = \frac{1}{(2\pi)^{d/2}|\Sigma_0|^{1/2}} e^{-\frac{1}{2}(\underline{\mu}-\underline{\mu}_0)^T\Sigma_0^{-1}(\underline{\mu}-\underline{\mu}_0)}$$

$$D = \{\underline{x}_1, \ldots, \underline{x}_n\}$$

$$f(\underline{\mu}|D) \propto \prod_{k=1}^{n} f(\underline{x}_k|\underline{\mu})\, f(\underline{\mu})$$

$$\propto \exp\left[-\frac{1}{2}\left(\underline{\mu}^T(n\Sigma^{-1}+\Sigma_0^{-1})\underline{\mu} - 2\underline{\mu}^T(\Sigma^{-1}\sum_{k=1}^{n} x_k + \Sigma_0^{-1}\mu_0)\right)\right]$$

$$\propto \exp\left[-\tfrac{1}{2}(\underline{\mu}-\underline{\mu}_n)^T \Sigma_n^{-1} (\underline{\mu}-\underline{\mu}_n)\right]$$

$$= n(\underline{\mu}_n, \Sigma_n)$$

equating coefficients,

$$\Sigma_n^{-1} = n\Sigma^{-1} + \Sigma_0^{-1}$$

$$\Sigma_n^{-1}\underline{\mu}_n = n\,\Sigma^{-1}\hat{\underline{\mu}}_n - \Sigma_0^{-1}\underline{\mu}_0$$

where $\hat{\underline{\mu}}_n = \frac{1}{n}\sum_{k=1}^{n} \underline{x}_k$ is <u>sample mean</u> $= \hat{\underline{\mu}}_{ML}$

Now $(A^{-1}+B^{-1})^{-1} = A(A+B)^{-1}B = B(A+B)^{-1}A$

for any nonsingular $d \times d$ matrices A, B.

$$\Sigma_n = \Sigma_0\left(\Sigma_0 + \tfrac{1}{n}\Sigma\right)^{-1}\tfrac{1}{n}\Sigma$$

$$\underline{\mu}_n = \Sigma_0\left(\Sigma_0 + \tfrac{1}{n}\Sigma\right)^{-1}\hat{\underline{\mu}}_n + \tfrac{1}{n}\Sigma\left(\Sigma_0 + \tfrac{1}{n}\Sigma\right)^{-1}\underline{\mu}_0$$

linear combination of $\hat{\underline{\mu}}_n$ and $\underline{\mu}_0$

Now $f(\underline{x}|D) = \int f(\underline{x}|\underline{\mu})\, f(\underline{\mu}|D)\, d\underline{\mu}$

observe (given D) $\underline{X} = \underline{\mu} + \underline{z}$

where $\underline{\mu} \sim n(\underline{\mu}_n, \Sigma_n)$

$\underline{z} \sim n(\underline{0}, \Sigma)$ indep of $\underline{\mu}$

Thus $f(\underline{x}|D) = n(\underline{\mu}_n, \Sigma_n + \Sigma)$

Bernoulli Distribution

$X \in \{0,1\}$ $\quad X = 1$ heads, $\quad X = 0$ tails

$p(X=1|\mu) = \mu \qquad 0 \le \mu \le 1$

$p(X=0|\mu) = 1-\mu$

$E[X] = \mu, \quad \mathrm{var}(X) = \mu(1-\mu) \qquad p(x_1,\dots,x_n|\mu) = \mu^{\sum_i x_i}(1-\mu)^{n-\sum x_i}$

Given $D = \{x_1,\dots,x_n\}$ $\qquad x_k$ iid $\sim \mathrm{Bern}(\mu)$

we showed $\qquad \hat{\mu}_{ML} = \frac{1}{n}\sum_{k=1}^{n} x_k \equiv m$

Beta distribution

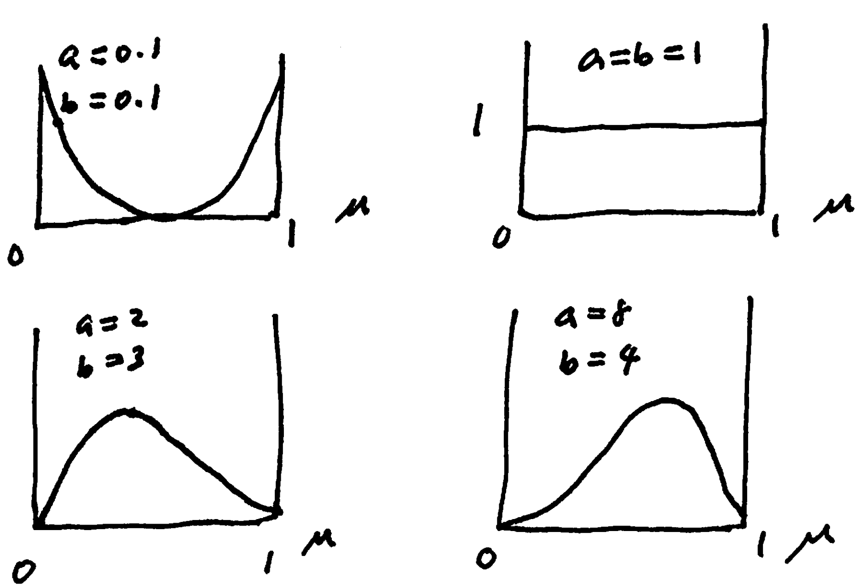
normalization

$$\mathrm{Beta}(\mu|\underbrace{a,b}_{\text{hyperparameters}}) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \qquad \mu \text{ a r.v.} \quad \mu \in [0,1]$$

where $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}\,du$ gamma function

$$E[\mu] = \frac{a}{a+b}$$

$$\mathrm{var}(\mu) = \frac{ab}{(a+b)^2(a+b+1)}$$



$$f(\mu|D) \propto p(D|\mu) f(\mu)$$

$$\propto \left(\prod_{k=1}^{n} \mu^{x_k} (1-\mu)^{1-x_k}\right) \text{Beta}(\mu|a,b)$$

$$\propto \mu^{\Sigma x_k} (1-\mu)^{n-\Sigma x_k} \mu^{a-1} (1-\mu)^{b-1}$$

$$\propto \mu^{m} (1-\mu)^{n-m} \mu^{a-1} (1-\mu)^{b-1} \qquad m = \sum_{k=1}^{n} x_k$$

$$\propto \mu^{m+a-1} (1-\mu)^{n-m+b-1}$$

$$\Rightarrow f(\mu|D) = \text{Beta}(\mu | m+a, n-m+b)$$

a: pseudo count for $x=1$

b: " for $x=0$

Now $$p(X=1|D) = \int_0^1 p(X=1|\mu) f(\mu|D)\, d\mu$$

$$= \int_0^1 \mu f(\mu|D)\, d\mu$$

$$= E[\mu|D] \quad \text{where } f(\mu|D) = \text{Beta}(\mu|m+a, n-m+b)$$

$$= \frac{m+a}{\cancel{m}+a+n-\cancel{m}+b}$$

$$= \frac{m+a}{n+a+b}$$

$$p(X=0|D) = \frac{n-m+b}{n+a+b}$$

Given D, $X \sim \text{Bernoulli}\left(\frac{m+a}{n+a+b}\right)$

~~of a priori~~,

Recap: $D = \{\underline{x}_1, \ldots, \underline{x}_n\}$

ML learning: $\underline{\theta}$ fixed. No prior over $\underline{\theta}$

Find parameter which maximizes likelihood of data:

$$\hat{\underline{\theta}}_{ML} = \underset{\underline{\theta}}{\operatorname{argmax}}\ f(D|\underline{\theta}) \qquad \Big(f(\underline{D}|\underline{\theta}) = \prod_{k=1}^{n} f(\underline{x}_k|\underline{\theta}) \Big)$$

by independence

MAP Learning:

$\underline{\theta}$ is a r.v. $\sim f(\underline{\theta})$.

Find parameter to maximize posterior.

$$\hat{\theta}_{MAP} = \underset{\underline{\theta}}{\operatorname{argmax}}\ f(\underline{\theta}|D)$$

$$= \underset{\underline{\theta}}{\operatorname{argmax}}\ f(D|\underline{\theta})\, f(\underline{\theta})$$

Bayesian learning

$\underline{\theta}$ is a r.v. $\sim f(\underline{\theta})$

compute posterior distr. of $\underline{\theta}$: $f(\underline{\theta}|D)$

Then estimate $f(\underline{x}|D) = \int f(\underline{x}, \underline{\theta}|D)\, d\underline{\theta}$

$$= \int f(\underline{x}|\underline{\theta}, D)\, f(\underline{\theta}|D)\, d\underline{\theta}$$

$$= \int f(\underline{x}|\underline{\theta})\, f(\underline{\theta}|D)\, d\underline{\theta}$$

Recursive Bayes incremental learning

Let $D^n = \{\underline{x}_1, \ldots, \underline{x}_n\}$ if $n > 1$,

$$\underbrace{f(D^n|\underline{\theta})}_{\text{likelihood}} = \prod_{k=1}^{n} f(\underline{x}_k|\underline{\theta}) = f(\underline{x}_n|\underline{\theta}) \prod_{k=1}^{n-1} f(\underline{x}_k|\underline{\theta})$$

$$= f(\underline{x}_n|\underline{\theta})\, f(D^{n-1}|\underline{\theta})$$

where $D^{n-1} = \{\underline{x}_1, \ldots, \underline{x}_{n-1})$

$$f(\underline{\theta}|D^n) = \frac{f(D^n|\underline{\theta})f(\underline{\theta})}{\int f(D^n|\underline{\theta})f(\underline{\theta})d\underline{\theta}} = \frac{f(\underline{x}_n|\underline{\theta})f(D^{n-1}|\underline{\theta})f(\underline{\theta})}{\int f(\underline{x}_n|\underline{\theta})f(D^{n-1}|\underline{\theta})f(\underline{\theta})d\underline{\theta}}$$

$$= \frac{f(\underline{x}_n|\underline{\theta})f(\underline{\theta}|D^{n-1})\cancel{f(D^{n-1})}}{\int f(\underline{x}_n|\underline{\theta})f(\underline{\theta}|D^{n-1})\cancel{f(D^{n-1})}d\underline{\theta}} \quad (1)$$

- Repeated use gives sequence:

$$f(\underline{\theta}|D^0) \equiv f(\underline{\theta}),\ f(\underline{\theta}|\underline{x}_1),\ f(\underline{\theta}|\underline{x}_1,\underline{x}_2), \ldots$$

- Given $f(\underline{\theta}|D^{n-1})$, obtain $f(\underline{\theta}|D^n)$ with $f(\underline{x}_n|\underline{\theta})$ and (1)

- Recursive approach.

- on-line learning: incremental, learning goes on as data is collected

- Batch: non-incremental, all training data must be present before learning takes place. Uses all training data for learning.

Example: Assume

$$f(x|\theta) = u(0,\theta) = \begin{cases} 1/\theta, & 0 \le x \le \theta \\ 0, & o/w. \end{cases}$$

and assume $f(\theta|D^0) = f(\theta) = u(0,10)$

$D = \{4, 7, 2, 8\}$

$$f(\theta|D^1) \propto f(x_1|\theta)\, f(\theta|D^0)$$

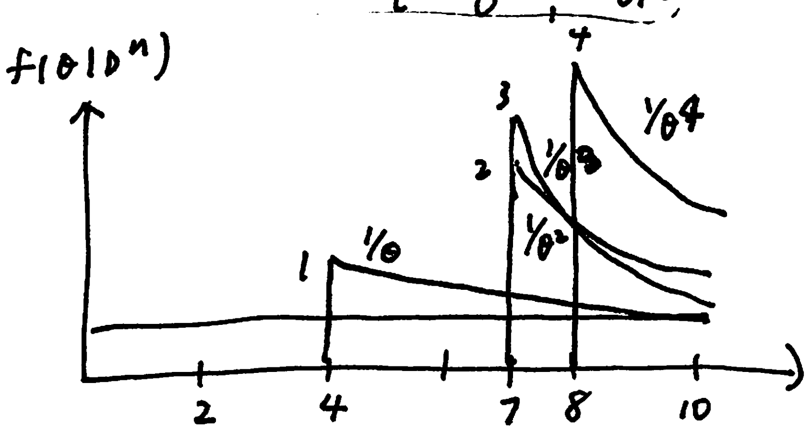
$$\propto \begin{cases} \frac{1}{\theta}, & 4 \le \theta \le 10 \\ 0, & o/w. \end{cases}$$

$$f(\theta|D^2) \propto \underbrace{f(x_2|\theta)}_{\begin{cases} \frac{1}{\theta}, & 7 \le \theta \le 10 \\ 0 & o/w \end{cases}} \underbrace{f(\theta|D^1)}_{\begin{cases} \frac{1}{\theta}, & 4 \le \theta \le 10 \\ 0, & o/w \end{cases}}$$

$$\propto \begin{cases} \frac{1}{\theta^2}, & 7 \le \theta \le 10 \\ 0, & o/w. \end{cases}$$

$$\vdots$$

$$f(\theta|D^n) \propto \begin{cases} \frac{1}{\theta^n}, & \max\{D^n\} \le \theta \le 10 \\ 0, & o/w. \end{cases}$$

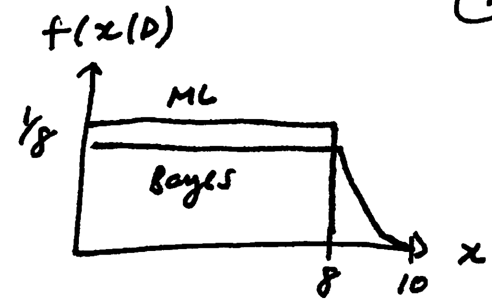


$$f(\theta|D^4) \propto \begin{cases} \frac{1}{\theta^4}, & 8 \le \theta \le 10 \\ 0, & o/w. \end{cases}$$

Given $D = \{4, 7, 2, 8\}$

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\ f(D|\theta) = 8$$

$$\Rightarrow f(x|D) \sim U(0, 8)$$



Bayesian methodology:

$$f(x|D) = \int f(x|\theta)\, f(\theta|D)\, d\theta$$

has tail above $x=8$: prior influence remains