CPSC 486/586: Probabilistic Machine Learning	January 30, 2023
Lecture 5	
Lecturer: Andre Wibisono	*Scribe: Daniel Kim*

1 Review

From last class, recall $\theta \sim p_0(\theta)$ as the prior distribution and $x \mid \theta \sim p(x \mid \theta)$ as the data likelihood.

Then the posterior distribution is

$$\theta \mid x \sim p_1(\theta \mid x) = \frac{p_0(\theta) \cdot p(x \mid \theta)}{p(x)}$$

where $p(x)$ is constant in θ.

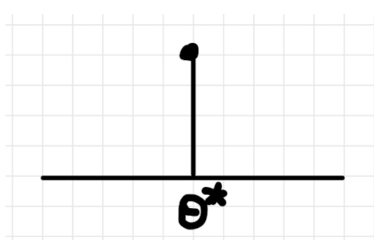
Definition 1. *A conjugate family consists of a prior $p_0 \in Q$, a "nice" class of distributions and a likelihood $p(x \mid \theta)$ such that the posterior $p(\theta \mid x) \in Q$ is also "nice."*

Prior $p_0(\theta)$, $\theta \in \Theta$	**Likelihood $p(x \mid \theta)$, $x \in \mathcal{X}$**	**Posterior $p(\theta \mid x)$**
Gaussian, $\Theta = \mathbb{R}$	Gaussian ($x \in \mathbb{R}^d$), $\mathcal{X} = \mathbb{R}$	Gaussian
Beta, $\Theta = [0, 1]$	Bernoulli ($x \in \{0, 1\}$), $\mathcal{X} = \{0, 1\}$	Beta
Dirichlet	Categorical ($x \in \{1, \ldots, k\}$), $\mathcal{X} = \{1, \ldots, k\}$	Dirichlet
Exp. family	Exp. family	Exp. family
Beta	Geometric	Beta

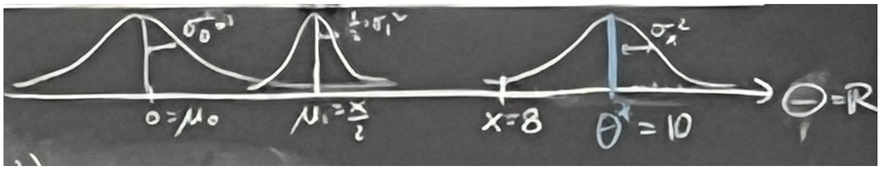
2 Inference

Suppose $\theta^* \in \Theta$ unknown, observe $X_1, \ldots, X_n \mid \theta^* \sim p(x \mid \theta^*)$ iid. Take a prior $p_0(\theta)$. We want to compute the posterior $p_n(\theta) = p(\theta \mid x_1, \ldots, x_n)$.

We will see that $\lim_{n \to \infty} p_n = \delta_{\theta^*}$, where δ_{θ^*} is the infinite point mass at θ^* and zero everywhere else.



Example 1. *Take* $\Theta = \mathcal{X} = \mathbb{R}$*, and* $p_0 = \mathcal{N}(\mu_0, \sigma_0^2)$ *and* $p(x \mid \theta) = \mathcal{N}(\theta, \sigma_x^2)$ *e.g. where* $\mu_0 = 0$ *and* $\sigma_0^2 = \sigma_x^2 = 1$.



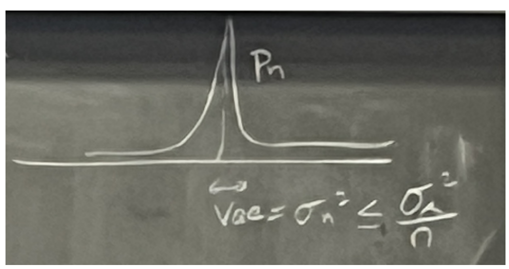
Then $p(\theta \mid x) = \mathcal{N}(\mu_1, \sigma_1^2)$ *where we can compute*

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma_x^2} \implies \sigma_1^2 = \frac{\sigma_0^2 \cdot \sigma_x^2}{\sigma_0^2 + \sigma_x^2} \leq \min\{\sigma_0^2, \sigma_x^2\},$$
$$\frac{1}{\sigma_1^2}\mu_1 = \frac{1}{\sigma_0^2}\mu_0 + \frac{1}{\sigma_x^2}x \implies \mu_1 = \frac{\sigma_x^2}{\sigma_0^2 + \sigma_x^2}\mu_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma_x^2}x$$

using same techniques as last class, where $\mu_0 = 0$*,* $\sigma_0^2 = \sigma_x^2 = 1$*, which implies* $\mu_1 = \frac{1}{2}x$*. This suggests* $p_1(\theta \mid x) = \mathcal{N}(\frac{1}{2}x, \frac{1}{2})$.

Example 2. *Observe* $X_1, \ldots, X_n \sim p(x \mid \theta^*)$ *iid where* $p_0 = \mathcal{N}(\mu_0, \sigma_0^2)$*,* $p_n(\theta) = p_n(\theta \mid x_1, \ldots, x_n)$*, and* $p_n = \mathcal{N}(\mu_n, \sigma_n^2)$ *and we can similarly compute*

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma_x^2} \implies \sigma_n^2 = \frac{\sigma_0^2 \cdot \sigma_x^2}{\sigma_x^2 + n\sigma_0^2} \leq \min\left\{\sigma_0^2, \frac{\sigma_x^2}{n}\right\},$$
$$\mu_n = \frac{\sigma_x^2}{\sigma_x^2 + n\sigma_0^2}\mu_0 + \frac{n\sigma_0^2}{\sigma_x^2 + n\sigma_0^2}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\sigma_x^2}{\sigma_x^2 + n\sigma_0^2}\mu_0 + \frac{n\sigma_0^2}{\sigma_x^2 + n\sigma_0^2}\overline{X}_n$$



Notice that

- $\frac{\sigma_x^2}{n} \to 0$ *as* $n \to \infty$
- $\overline{X}_n \to \mathbb{E}[X_1] = \theta^*$ *as* $n \to \infty$
- $\frac{\sigma_x^2}{\sigma_x^2 + n\sigma_0^2} \to 0$ *as* $n \to \infty$

- $\frac{n\sigma_0^2}{\sigma_x^2+n\sigma_0^2} \to 1$ *as* $n \to \infty$

which implies $\lim_{n\to\infty} \mu_n = \theta^*$ *and* $\lim_{n\to\infty} \sigma_n^2 = 0$. *This means* $\lim_{n\to\infty} p_n = \mathcal{N}(\theta^*, 0) = \delta_{\theta^*}$.

Definition 2. *The Bernoulli distribution is denoted as Ber(p) on* $\mathcal{X} = \{0,1\}$ *for* $0 \le p \le 1$. *We say* $X \sim$ *Ber(p)* $\iff \mathbb{P}(X=1) = p$ *and* $\mathbb{P}(X=0) = 1-p$.

Definition 3. *The Bernoulli density is* $\rho : \{0,1\} \to \mathbb{R}$ *with*

$$\rho(x) = p^x(1-p)^{1-x}$$

with the consequential properties

- $\rho(0) \ge 0$
- $\rho(1) \ge 0$
- $\rho(0) + \rho(1) = 1$
- $\rho(x) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \end{cases}$

Example 3. *The Bernoulli distribution is in an exponential family. We can write*

$$\begin{aligned}\rho(x) &= p^x(1-p)^{1-x} \cdot \mathbf{1}\{x \in \{0,1\}\} \\ &= \exp\left(x \log p + (1-x)\log(1-p)\right) \mathbf{1}\{x \in \{0,1\}\} \\ &= \exp\left(x \log\left(\frac{p}{1-p}\right) + \log(1-p)\right) \mathbf{1}\{x \in \{0,1\}\},\end{aligned}$$

implying

$$\rho_\theta(x) = \exp(\langle T(x), \theta\rangle - A(\theta)) \cdot h(x)$$

so Ber(p) is in an exponential family with $T(x) = x$ *and* $\theta = \log\left(\frac{p}{1-p}\right)$. *The normalizing constant* $A(\theta)$ *will be* $-\log(1-p) = \log(1+e^\theta)$. *The base measure* $h(x)$ *is equal to* $\mathbf{1}\{x \in \{0,1\}\}$.

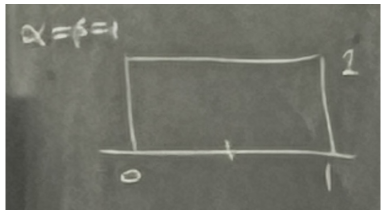
Definition 4. *The Beta distribution Beta*(α, β) *on* $p \in [0,1]$ *for some parameters* $\alpha, \beta > 0$ *has the density*

$$\rho(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}$$

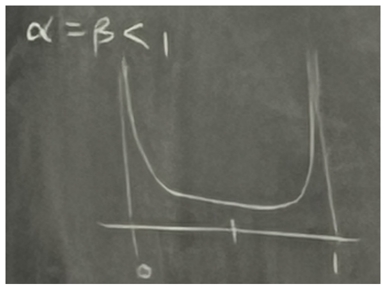
where we need to include the normalizing constant $B(\alpha,\beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}\, dp = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+b)}$.

Recall that the Gamma function is defined $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ and $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$.

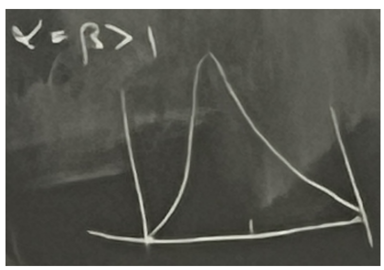
Consider $\alpha = \beta = 1$. Then the density is shaped as



Consider $\alpha = \beta < 1$. Then the density is shaped as



Consider $\alpha = \beta > 1$. Then the density is shaped as



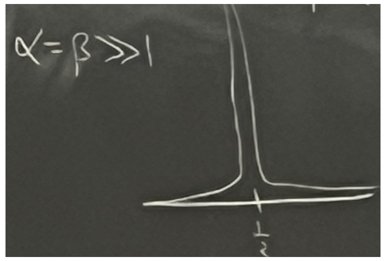
If $p \sim \text{Beta}(\alpha, \beta)$, then

$$\mathbb{E}[p] = \int_0^1 p\left(\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\right)dp = \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} = \frac{\alpha}{\alpha+\beta}$$

$$\text{Var}(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Using the AM-GM inequality applied to $\alpha, \beta \geq 0$, we know $\alpha\beta \leq \frac{(\alpha+\beta)^2}{4}$, so it can be shown that $\text{Var}(p) \leq \frac{1}{4(\alpha+\beta+1)}$.

Then if $\alpha, \beta \gg 1$, then the density would be shaped as



Example 4 (Beta-Bernoulli)**.** *Consider a prior $p \sim \rho_0 = Beta(\alpha_0, \beta_0)$ and likelihood $x \mid p \sim \rho(x \mid p) = Ber(p)$. We observe $x_1, \ldots, x_n \sim \rho(x \mid p^*)$ iid where p^* is unknown.*

By Bayes rule, the posterior can be computed as

$$\begin{aligned} p \mid x \sim \rho(p \mid x) &\propto \rho_0(p)\rho(x \mid p) \\ &\propto p^{\alpha_0 - 1}(1-p)^{\beta_0 - 1} p^x (1-p)^{1-x} \\ &\propto p^{\alpha_0 + x - 1}(1-p)^{\beta_0 + (1-x) - 1} \end{aligned}$$

implying

$$\rho(p \mid x) = Beta(\alpha_0 + x, \beta_0 + 1 - x).$$

After observing $x_1, \ldots, x_n$, we have

$$\rho(p \mid x_1, \ldots, x_n) = Beta\left(\alpha_0 + \sum_{i=1}^n x_i, \beta_0 + \sum_{i=1}^n (1 - x_i)\right)$$

so we can compute

$$\mathbb{E}\left[p \mid x_1, \ldots, x_n\right] = \frac{\alpha_0 + \sum_{i=1}^n x_i}{\alpha_0 + \beta_0 + n} = \frac{\alpha_0 + n\overline{X}_n}{\alpha_0 + \beta_0 + n} \to \mathbb{E}[x_i] = p^*,$$

$$Var(p \mid x_1, \ldots, x_n) \leq \frac{1}{4(\alpha_0 + \beta_0 + n)} \to 0$$

as $n \to \infty$.

Definition 5. *The sigmoid function is*

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

mapping $\mathbb{R} \to [0, 1]$.

Example 5 (Bayesian Logistic Regression)**.** *Consider $\theta \in \mathbb{R}^d$, $\theta \sim p_0 = \mathcal{N}(0, I)$ and $x \in \mathbb{R}^d$ covariates, with $y \in \{0, 1\}$ labels. Suppose*

$$y \mid \theta, x \sim Ber\left(\frac{1}{1 + e^{-\theta^T x}}\right) = Ber(\sigma(\theta^T x)).$$

By Bayes rule,

$$p(\theta \mid y, x) \propto p_0(\theta) \cdot \underbrace{p(y \mid \theta, x)}_{Ber(\sigma(\theta^T x))(y)}$$

$$\propto \exp\left(-\frac{1}{2}\|\theta\|^2 + y \log \sigma(\theta^T x) + (1-y)\log(1-\sigma(x^T\theta))\right)$$

and left as an exercise, it turns out that

$$p(\theta \mid y, x) \propto_\theta \exp\left(-\frac{1}{2}\|\theta\|^2 + yx^T\theta - \log(1+e^{\theta^T x})\right)$$

so

$$p(\theta \mid x, y) \propto \exp(-f(\theta))$$

but

$$f(\theta) = \frac{1}{2}\|\theta\|^2 + \log(1+e^{\theta^T x}) - yx^T\theta$$

is not quadratic, so $p(\theta \mid x, y)$ is not a Gaussian. This is an example where the posterior is not in the same family of distributions as the prior.