## ① Review of exponential family

- $p(x|\eta) = h(x) \exp\left(\eta^T t(x) - a(\eta)\right)$

  $\eta \in \mathbb{R}^d$ "natural parameter"

  $t(x) \in \mathbb{R}^d$ "sufficient statistics"

  $h(x) \in \mathbb{R}$ "base measure"

  $a(\eta) \in \mathbb{R}$ "log normalizer"

- $h(x) \exp(\eta^T t(x))$ is the "kernel"

- $a(\eta) = \log \int \underbrace{h(x) \exp(\eta^T t(x))}_{\text{kernel}} dx$

---

typically there is a "standard" parameterization

- $p(x|\theta) = h(x) \exp(g(\theta)^T t(x) - a(\theta))$
- $g(\theta) = \eta$ is the "link function"
- $g^{-1}(\eta) = \theta$ is the "inverse link"

---

then is also a "mean parameterization"

$$\nabla_\eta a(\eta) = \nabla_\eta \log \int h(x) \exp(\eta^T t(x)) dx$$

$$= \frac{\nabla_\eta \int \cdots dx}{\int \cdots dx} = \frac{\int h(x) \nabla_\eta \exp(\eta^T t(x)) dx}{\int \cdots dx}$$

$$= \int t(x) \frac{h(x) \exp(\eta^T t(x)) dx}{\int \cdots dx}$$

$$= \mathbb{E}_\eta[t(x)] \triangleq \mu$$

Example: Binomial (for fixed $n$)

$$X \sim \text{Binom}(N, \theta)$$

$$p(x \mid \theta, N) = \binom{N}{x} \theta^x (1-\theta)^{N-x}$$

$$= \binom{N}{x} \exp\left( x \log \theta + (N-x) \log(1-\theta) \right)$$

$$= \binom{N}{x} \exp\left( x \log \frac{\theta}{1-\theta} + N \log(1-\theta) \right)$$

$\binom{N}{x}$: $\triangleq h(x)$; $x$: $\triangleq t(x)$; $\log \frac{\theta}{1-\theta}$: $g(\theta)$; $N \log(1-\theta)$: $\triangleq -a(\theta)$

$$= h(x) \exp\left( t(x)^T g(\theta) - a(\theta) \right)$$

$$g(\theta) = \text{logit}(\theta) = \log \frac{\theta}{1-\theta} = \eta$$

$$g^{-1}(\eta) = \text{logistic}(\eta) = \frac{e^\eta}{1+e^\eta} = \theta$$

$$= h(x) \exp\left( t(x)^T \eta - N \log(1+e^\eta) \right)$$

$N \log(1+e^\eta) \equiv a(\eta)$

confirm that $\mathbb{E}[x] = N\theta$:

$$\mu = \nabla_\eta a(\eta) = \nabla_\eta N \log(1+e^\eta)$$

$$= \frac{N \nabla_\eta (1+e^\eta)}{(1+e^\eta)} = \frac{N e^\eta}{1+e^\eta} = N\theta$$

## (2) Conjugacy

Recall beta-binomial conjugacy:

$$\vartheta \sim \text{Beta}(a, b)$$

$$x_i \overset{\text{ind.}}{\sim} \text{Binomial}(N_i, \vartheta), \quad i = 1 \dots n$$

$$p(\vartheta \mid x_1 \dots x_n, a, b) \propto_\vartheta \vartheta^{a-1} (1-\vartheta)^{b-1} \prod_{i=1}^n \vartheta^{x_i} (1-\vartheta)^{N_i - x_i}$$

$$\propto_\vartheta \text{Beta}\left(\vartheta;\ \underbrace{a + \sum_{i=1}^n x_i}_{\triangleq a_n},\ \underbrace{b + \sum_{i=1}^n (N_i - x_i)}_{\triangleq b_n}\right)$$

This is an instance of a general pattern w/ expfam dists.

---

Generally:

$$\eta \sim F(\cdot \mid \lambda)$$

$$x_i \overset{\text{ind.}}{\sim} G(\cdot \mid \eta) \quad i = 1 \dots n$$

conjugate if

$$p(\eta \mid x_{1:n}, \lambda) \propto F(\eta \mid \lambda) \prod_{i=1}^n G(x_i \mid \eta)$$

$$\propto F(\eta \mid \lambda_n)$$

(same family as prior)

---

All expfams $G(\cdot \mid \eta)$ (likelihood) have a conjugate prior.

$$p(x_i \mid \eta) = h_\ell(x_i) \exp\left(\eta^\top t_\ell(x_i) - a_\ell(\eta)\right)$$

$$p(\eta \mid \lambda) = h_c(\eta) \exp\left(\lambda_1^\top \eta + \lambda_2 \left(-a_\ell(\eta)\right) - a_c(\lambda)\right)$$

- $\lambda \equiv [\lambda_1, \lambda_2] \in \mathbb{R}^{\dim(\eta)+1}$ "natural parameter"
- $t_c(\eta) \equiv [\eta, -a_\ell(\eta)]$ "suff stats"

**Example:** Binomial's conjugate prior

$$\eta = g(\theta) = \text{logit}\,\theta, \quad \dim(\eta) = 1$$

$$a_\ell(\eta) = N \log(1 + e^\eta)$$

Set $N = 1$ (for now):

$$p(\eta \mid \lambda) \propto_\eta h_c(\eta) \exp\left(\lambda_1 \eta - \lambda_2 \log(1 + e^\eta)\right)$$

notice that $\eta = g_\ell(\theta)$ is the likelihood's natural parametrization of $\theta$. When deriving the prior, we will often want to think about $\theta = g_\ell^{-1}(\eta)$.

Change of variables to $\theta \in (0,1)$:

$$p_\theta(\theta \mid \lambda) = \left|\frac{\partial}{\partial \theta} g_\ell(\theta)\right| p_\eta\left(g_\ell^{-1}(\theta) \mid \lambda\right)$$

$$\propto_\theta \left|\frac{\partial}{\partial \theta} \log \frac{\theta}{1-\theta}\right| \mathbb{1}(\theta \in (0,1)) h_c(g_\ell(\theta)) \exp\left(\lambda_1 \log \frac{\theta}{1-\theta} + \lambda_2 \log(1-\theta)\right)$$

$$\propto_\theta \frac{1}{\theta(1-\theta)} \mathbb{1}(\theta \in (0,1)) h_c(g_\ell(\theta)) \exp\left(\lambda_1 \log \theta + [-\lambda_1 + \lambda_2] \log(1-\theta)\right)$$

$$\propto_\theta \frac{1}{\theta(1-\theta)} \mathbb{1}(\theta \in (0,1)) h_c(g_\ell(\theta)) \theta^{\lambda_1} (1-\theta)^{-\lambda_1 + \lambda_2}$$

We recognize the kernel of a Beta:

Let $h_c(g_\ell(\theta)) = 1$, $\alpha = \lambda_1$, $\beta = -\lambda_1 + \lambda_2$:

$$\propto_\theta \theta^{\alpha - 1} (1-\theta)^{\beta - 1} \quad \Longrightarrow \text{conjugate prior: Beta.}$$

(3) Conjugate posterior under indep. sampling

$$\eta \sim F(\cdot \mid \lambda)$$

$$x_i \overset{\text{ind.}}{\sim} G(\cdot \mid \eta) \quad i = 1 \dots n$$

$$P(\eta \mid x_1 \dots x_n, \lambda) \propto_\eta F(\eta \mid \lambda) \prod_{i=1}^{n} G(x_i \mid \eta)$$

$$\propto_\eta h_c(\eta) \exp\left(\lambda_1^\top \eta + \lambda_2 \left(-a_\ell(\eta)\right) - a_c(\lambda)\right) \left[\prod_{i=1}^{n} h_\ell(x_i) \exp\left(\eta^\top t_\ell(x_i) - a_\ell(\eta)\right)\right]$$

$$\propto_\eta h_c(\eta) \exp\left(\lambda_1^\top \eta + \lambda_2 \left(-a_\ell(\eta)\right)\right) \prod_i \exp\left(\eta^\top t_\ell(x_i) - a_\ell(\eta)\right)$$

$$\propto_\eta h_c(\eta) \exp\left(\eta^\top \left[\lambda_1 + \sum_{i=1}^{n} t_\ell(x_i)\right] - a_\ell(\eta)\left[\lambda_2 + n\right]\right)$$

Same family as $F(\cdot \mid \lambda)$ prior but with updated natural parameters:

$$\lambda_n \equiv \begin{bmatrix} \lambda_{n1} \\ \lambda_{n2} \end{bmatrix} = \begin{bmatrix} \lambda_1 + \sum_{i=1}^{n} t_\ell(x_i) \\ \lambda_2 + n \end{bmatrix}$$

↳ posterior updating = adding / incrementing!

Example:

$$\theta \sim \text{Beta}(\alpha, \beta)$$

$$x_i \overset{\text{ind.}}{\sim} \text{Binom}(N_i, \theta) \quad i = 1 \dots n$$

$$p(\theta \mid -) \propto \theta^{\alpha + \sum_i x_i - 1} (1-\theta)^{\beta + \sum_i (N_i - x_i) - 1}$$

$$\alpha_n = \alpha + \sum_{i=1}^{n} x_i \quad , \quad \beta_n = \beta + \sum_{i=1}^{n} (N_i - x_i)$$

From above:

$$\lambda_{n1} = \alpha_n \quad , \quad -\lambda_{n1} + \lambda_{n2} = \beta_n$$

So: $\leftarrow (x_i \equiv t_e(x_i))$

$$\lambda_{n1} = \lambda_1 + \sum_{i=1}^{n} x_i$$

$$\lambda_{n2} = \beta_n + \alpha_n = \beta + \sum_{i=1}^{n} (N_i - x_i) + \alpha + \sum_{i=1}^{n} x_i$$

$$= \alpha + \beta + \sum_{i=1}^{n} N_i$$

$$= \lambda_1 + (\lambda_2 - \lambda_1) + \sum_{i=1}^{n} N_i$$

$$= \lambda_2 + \sum_{i=1}^{n} N_i$$

---

(4) Likelihood principle

Notice the "sample size" above is $\sum_{i=1}^{n} N_i$.

This is because $x_i \sim \text{Binom}(N_i, \theta)$ is the same as...

$$b_{ij} \overset{iid}{\sim} \text{Binom}(1, \theta), \quad \begin{matrix} i = 1 \dots n \\ j = 1 \dots N_i \end{matrix}$$

... from the perspective of $\theta$.

$$\prod_{i=1}^{n} \text{Binom}(x_i; N_i, \theta) \propto_\theta \prod_{i=1}^{n} \prod_{j=1}^{N_i} \text{Binom}(b_{ij}; 1, \theta)$$

---

We can also understand this as satisfying the likelihood principle: all information in the data about $\theta$ should be contained in the likelihood function:

$$P(X \mid \theta) \propto_\theta L(\theta; X)$$

(likelihood dist.) (likelihood func.)

If two likelihood functions are proportional in $\theta$, then they contain the same information.

↳ posterior (Bayesian) inference satisfies the likelihood principle generically.

---

Another example: $x_i \overset{ind.}{\sim} NB(r_i, \theta)$ $i = 1 \dots n$

$$\prod_{i=1}^{n} NB(x_i; r_i, \theta) \propto_\theta \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{\sum_{i=1}^{n} r_i}$$

(negative binomial)

↳ this "experiment" contains the same info about $\theta$ as observing $\sum_{i=1}^{n} x_i$ "successes" and $\sum_{i=1}^{n} r_i$ "failures" in $\sum_{i=1}^{n}(x_i + r_i)$ indep. Bernoulli trials!

---

## (5) Posterior predictive in conjugate models

$$p(x_{n+1} \mid x_1 \dots x_n)$$

$$= \int p(x_{n+1} \mid \eta)\, p(\eta \mid x_{1:n})\, d\eta$$

$$= \int G(x_{n+1} \mid \eta)\, F(\eta \mid \lambda_n)\, d\eta$$

$$= \int h_\ell(x_{n+1}) \exp\left( t(x_{n+1})^\top \eta - a_\ell(\eta) \right) h_c(\eta) \exp\left( \eta^\top \lambda_{n1} + \lambda_{n2}(-a_\ell(\eta)) - a_c(\lambda_n) \right) d\eta$$

$$= h_\ell(x_{n+1}) \exp(-a_c(\lambda_n)) \underbrace{\int \exp\left( \eta^\top [t(x_{n+1}) + \lambda_{n1}] + (-a_\ell(\eta))[1 + \lambda_{n2}] \right) d\eta}$$

$$= \exp\left( a_c\left( [\lambda_{n1} + t(x_{n+1}),\ \lambda_{n2} + 1] \right) \right)$$

$$= h_\ell(x_{n+1}) \frac{\exp\left( a_c\left( [\lambda_{n1} + t(x_{n+1}),\ \lambda_{n2} + 1] \right) \right)}{\exp\left( a_c\left( [\lambda_{n1}\ ,\ \lambda_{n2}\ ] \right) \right)}$$

ratio of normalizers!

## Example: Beta-binomial

- $a_c(\lambda_1, \lambda_2) = B(\lambda_1, \lambda_2 - \lambda_1)$ beta function
- $h_\ell(x_{n+1}) = \binom{N_{n+1}}{x_{n+1}}$ binomial coefficient

$$p(x_{n+1} \mid N_{n+1}, x_{1:n}, N_{1:n})$$

$$= \binom{N_{n+1}}{x_{n+1}} \frac{B(\lambda_{n1} + x_{n+1}, \lambda_{n2} - \lambda_{n1} + 1)}{B(\lambda_{n1}, \lambda_{n2} - \lambda_{n1})}$$

$$= \binom{N_{n+1}}{x_{n+1}} \frac{B(\alpha + \sum_{i=1}^n x_i + x_{n+1}, \beta + \sum_{i=1}^n (N_i - x_i) + N_{n+1} - x_{n+1})}{B(\alpha + \sum_{i=1}^n x_i, \beta + \sum_{i=1}^n (N_i - x_i))}$$

$$= \text{BetaBinom}(x_{n+1}; N_{n+1}, \alpha_n, \beta_n)$$

---

(6) "Prior predictive"

$$n = 0: \quad p(x_{n+1} \mid x_{1:n}) \equiv p(x_1)$$

$$= h_\ell(x_1) \frac{\exp\left(a_c([\lambda_1 + t(x_1), \lambda_2 + 1])\right)}{\exp\left(a_c([\lambda_1, \lambda_{n2}])\right)}$$

(this is the "marginal likelihood", or "model evidence", or "normalizing constant")

## (7) Alternative parameterization of conjugate prior

Let $\lambda_1 = t_0 \kappa_0$, $\lambda_2 = \kappa_0$, so that the prior:

$$F(\eta \mid t_0, \kappa_0) \propto_\eta h_c(\eta) \exp\left(t_0 \kappa_0^\top \eta - \kappa_0 a(\eta)\right)$$

Then the posterior is:

$$p(\eta \mid t_0, \kappa_0, X_{1:n}) = f(\eta \mid t_n, \kappa_n)$$

where

$$\kappa_n = \kappa_0 + n$$

$$t_n = \frac{\kappa_0 t_0 + \sum_{i=1}^n t(x_i)}{\kappa_0 + n} = \left(\frac{\kappa_0}{\kappa_0 + n}\right) t_0 + \left(\frac{n}{\kappa_0 + n}\right) \frac{1}{n} \sum_{i=1}^n t(x_i)$$

Recall that the MLE is:

$$\operatorname*{argmax}_\eta \sum_{i=1}^n \log G(x_i \mid \eta)$$

$$= \operatorname*{argmax}_\eta \underbrace{\eta^\top \left(\sum_{i=1}^n t(x_i)\right) - n\, a_\ell(\eta)}_{\triangleq LL(\eta; x)}$$

$$\nabla_\eta LL(\eta; x) = \sum_{i=1}^n t(x_i) - n \underbrace{\nabla_\eta a_\ell(\eta)}_{= \mathbb{E}_\eta[t(x)]} = 0$$

↳ $\hat{\eta}^{MLE}$: $\frac{1}{n} \sum_{i=1}^n t(x_i) = \mathbb{E}_{\hat{\eta}^{MLE}}[t(x)]$ "moment matching"

⑧ Shannon entropy and information

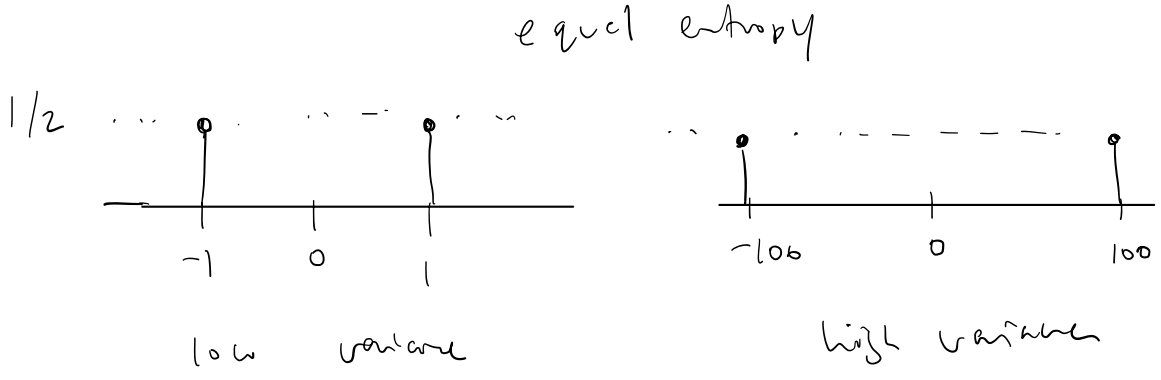
"information content" of event $X = x$

$$h(x) = \log\left[\frac{1}{p(x)}\right]$$

"entropy" of random variable $X$

$$H(X) = \mathop{\mathbb{E}}_{X}\left[\log\left[\frac{1}{p(x)}\right]\right]$$

measure of uncertainty



(9) Expfam dists are uniquely maximum entropy

- consider an expfam dist for $x \in S$ (S: support):

$$p_\eta(x) = \underline{1}(x \in S)\, h(x) \exp\left(\eta^T t(x) - a(\eta)\right)$$

- define the set of distributions

$$\mathcal{P}_\alpha \overset{D}{=} \left\{ p : \mathop{\mathbb{E}}_{x \sim p(x)} [t(x)] = \alpha, \quad x \in S \right\}$$

all distributions, with support S, that satisfy the moment constraint $\mathbb{E}[t(x)] = \alpha$.

- THEOREM: if $p_\eta(x) \in \mathcal{P}_\alpha$, then:

$$H(p_\eta) > H(p') \quad \forall\, p' \in \mathcal{P}_\alpha$$

"$p_\eta$ is the unique maximum entropy dist in the set $\mathcal{P}_\alpha$"

- Proof: ① show $H(p_\eta) \geq H(p') \;\forall\; p' \in \mathcal{P}_\alpha$

② show uniqueness

# Show $H(p_\eta) \geq H(p') \;\forall\; p' \in \mathcal{P}_\alpha$

- for simplicity, discrete $X \in S$:

$$p(X = k) \equiv p(k), \quad \sum_{k \in S} p(k) = 1$$

- introduce Lagrangian, with Lagrange multipliers

$$\eta \equiv (\eta_1 \dots \eta_d), \quad \eta_0$$

$$\Lambda(p, \eta, \eta_0) = -H(p) + \eta^T (\alpha - \mathbb{E}_{x \sim p}[t(x)]) + \eta_0 \left( \sum_{k \in S} p(k) - 1 \right)$$

← moment constraints

↑ normalization constraint

- argmin $\Lambda(p, \eta, \eta_0)$ will maximize $H(p)$
  $p$ subject to constraints

- Derivate wrt a single component of $p(j)$, $j \in S$:

$$\frac{\partial}{\partial p(j)} \Lambda(p, \eta, \eta_0) = \frac{\partial}{\partial p(j)} \sum_{k \in S} p(k) \log p(k) = \log p(j) + 1$$

$$+ \frac{\partial}{\partial p(j)} \eta^T \left( \alpha - \sum_k p(k) t(k) \right) = -\eta^T t(j)$$

$$+ \frac{\partial}{\partial p(j)} \eta_0 \left( \sum_k p(k) - 1 \right) = \eta_0$$

- set to zero

$$\log p(j) + 1 - \eta^T t(j) + \eta_0 = 0$$

$$p(j) = \exp\left(\eta^T t(j) - \eta_0 - 1\right)$$

- $p(j) \equiv p_{\eta_0, \eta}(j)$ is parametrized by $\eta_0, \eta$
- We can drop $\eta_0$ by setting it to a value that satisfies the constraint it enforces

$$\eta_0 \triangleq a(\eta) - 1$$

- $p(j) \equiv p_\eta(j) = \exp\left(\eta^T t(j) - a(\eta)\right)$

  ↳ $\sum_{k \in S} p_\eta(k) = 1$ (by defn of $a(\eta)$)

- now if $\exists \eta$ s.t. $\mathbb{E}_{x \sim p_\eta}[t(x)] = \alpha$, then:

$$p_\eta \in \mathcal{P}_\alpha \text{ and } H(p_\eta) \geq H(p'), \; p' \in \mathcal{P}_\alpha$$

↳ $p_\eta$ is a max ent dist in $\mathcal{P}_\alpha$.

---

Show uniqueness: $H(p_\eta) > H(p'), \; p' \in \mathcal{P}_\alpha$

- consider any $p \in \mathcal{P}_\alpha$

- $H(p) = -\sum_k p(k) \log p(k)$

$$= -\sum_k p(k) \log \left[ p(k) \frac{p_\eta(k)}{p_\eta(k)} \right]$$

$$= -\sum_k p(k) \log \frac{p(k)}{p_\eta(k)} - \sum_k p(k) \underbrace{\log p_\eta(k)}_{= \eta^T t(k) - a(\eta)}$$

$$= -KL(p \,||\, p_\eta) - \sum_k p(k) \left[ \eta^T t(x) - a(\eta) \right]$$

$$= -KL(p \,||\, p_\eta) - \left[ \eta^T \mathop{\mathbb{E}}_{x \sim p} [t(x)] - a(\eta) \right]$$

since $p \in \mathcal{S}_\mathcal{E}$, $\mathbb{E}_p[t(x)] = \alpha = \mathbb{E}_{p_\eta}[t(x)]$:

$$= -KL(p \,||\, p_\eta) - \underbrace{\sum_k p_\eta(k) \log p_\eta(k)}_{= H(p_\eta)}$$

So:

$$H(p) = H(p_\eta) - \underbrace{KL(p \,||\, p_\eta)}_{\geq 0 \text{ IFF } p = p_\eta}.$$

(10) Back to Shannon information and entropy

"Game of submarine" (Mackay Chap 4.1)

- submarine is somewhere in 64-cell search grid
- $P(z=k) = \frac{1}{64}$ (no prior info)
- $p(\text{hit on } \underline{\text{first}} \text{ guess}) \equiv p(x_1 = 1) = \frac{1}{64}$
- info content of <u>event</u> $x_1 = 1$?

$$h(x_1 = 1) = \log_2 64 = 6 \text{ bits}$$

- info content of event $x_1 = 1$?

$$h(x_1 = 0) = \log_2 64/63 \approx 0.0227 \text{ bits}$$

- prob. guessing correctly on <u>second</u> guess:

$$p(x_2 = 1 \mid x_1 = 0) = \frac{1}{63} \leftarrow \text{one cell eliminated by } x_1$$

$\hookrightarrow h(x_2 = 0 \mid x_1 = 0) = \log_2 63/62 \approx 0.023$ bits

- <u>total</u> info content in $x_1 = 0, x_2 = 0$:

$$h(x_2 = 0, x_1 = 0) = \log_2 \frac{1}{p(x_1 = 0, x_2 = 0)}$$

$$= \log_2 \frac{64}{63} + \log_2 \frac{63}{62}$$

- total info content in 32 misses: $x_1 = 0 \ldots x_{32} = 0$

$$h(x_1 = 0 \ldots x_{32} = 0) = \log_2 \frac{64}{63} + \log_2 \frac{63}{62} \cdots + \log_2 \frac{33}{32}$$

$$= \log_2 \frac{64}{32} = \log_2 2 = \underline{1 \text{ bit}}$$

- this makes sense since:

$$p(z \in \{1, \dots, 32\}) = \frac{32}{64} = \frac{1}{2}$$

- similarly:

$$h(X_{(1)}=0 \dots X_{(48)}=0, X_{(49)}=1) =$$

$$= \log_2 \frac{64}{16} + \log_2 16 = \underline{6 \text{ bits}}$$

No matter <u>when</u> you find the sub, the total info gained is <u>6 bits</u>.

↳ $H(p) = \sum_{k=1}^{64} p(z=k) \log_2 \frac{1}{p(z=k)} = \sum \frac{1}{64} \log_2 64 = 6.$