# Entropy

The number of arrangements of $n$ states $\{1,\ldots,s\}$ that yield a distribution $\widehat{p}$:

$$C(\widehat{p}) = \binom{n}{\widehat{p}_1 n, \ldots, \widehat{p}_s n} = \frac{n!}{(\widehat{p}_1 n)! \cdots (\widehat{p}_s)!}$$

**Stirling's Approximation:**

$$k! \approx k^k e^{-k} \sqrt{2\pi k}$$

**Shannon Entropy:**

- For $p$ a pmf, $H(p) = -\sum_{x=1}^s p(x) \log p(x)$
- For $p$ a pdf, $H(p) = -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx$

**Entropy Approximation:** $C(\widehat{p}) \approx e^{nH(p)}$

**Maximum Entropy Principle:** Let $p$ satisfy

1. $\sum_{x=1}^s p(x) = 1$
2. $\sum_{x=1}^s p(x)\mathcal{E}(x) \approx \theta$
3. $p$ maximizes $H(p)$

then $p$ has the form

$$p(x) = \frac{1}{Z_\lambda} e^{\lambda \mathcal{E}(x)} = \frac{1}{\sum_{x=1}^s e^{\lambda \mathcal{E}(x)}} e^{\lambda \mathcal{E}(x)}$$

where $\lambda$ is found via the constrain $\sum p(x)\mathcal{E}(x) = \theta$.

---

In the case of seeking $\arg\max_p H(p)$ subject to constraints $\sum p_x = 1$ and $\sum p_x \mathcal{E}_i(x) = \theta_i$ for $i = 1 : k$, $p$ will have form

$$p(x) = \frac{1}{Z_\lambda} \exp[\sum_{i=1}^k \lambda_i \mathcal{E}_i(x)]$$

**Large Deviation Principle:** For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p$ a pmf on $\{1,\ldots,s\}$, if $\mathcal{E} : \{1,\ldots,s\} \to \mathbb{R}$ satisfies $\frac{1}{n}\sum_{i=1}^n \mathcal{E}(X_k) = \theta$, then $\mathbb{E}_{\widehat{p}}[\mathcal{E}(X)] = \theta$

**Observations:** For $q$ a distribution on $\{1,\ldots,s\}$,

$$\mathbb{P}(\widehat{p} = q) = \binom{n}{n_1, \ldots, n_s} \prod_{x=1}^s p_x^{q_x \cdot n}$$

Further,

$$\frac{e^{-nD(q\|p)}}{(n+1)^s} \le \mathbb{P}(\widehat{p} = q) \le e^{-nD(q\|p)}$$

**Kullback-Leibler Divergence:**

$$D(q \parallel p) = -\sum_{x=1}^s q_x \log \frac{p_x}{q_x}$$

- $D(q \parallel p) \ge 0$
- $D(q \parallel p) = 0 \iff p = q$

**Convexity:** $f$ convex if $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$ for $\lambda \in [0,1]$.

- $f$ convex iff $f''(x) \ge 0$
- $f$ concave iff $f''(x) \le 0$ iff $-f$ is convex
- For $x \in \mathbb{R}^s$, $f$ convex iff $h(\lambda) = f(\lambda x + (1-\lambda)y)$ convex

**Jensen's Inequality:** For $g : \mathbb{R} \to \mathbb{R}$ convex, $\mathbb{E}[g(X)] \ge g(\mathbb{E}[X])$

**Sanov's Theorem:** For $B$ an open subset of the space of distributions on $\{1,\ldots,s\}$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\widehat{p} \in B) = -\inf_{q \in B} D(q \parallel p)$$

Further, if $p^* = \arg\min_{q\in B} D(q \parallel p)$ and $\widehat{p} \in B$, $\widehat{p} \xrightarrow{\mathbb{P}} p^*$

**Exponential Families:**

$$p(x) = \frac{1}{Z(\lambda)} h(x) e^{\lambda \cdot T(x)}$$

where $Z(\lambda)$ satisfies:

1. $\frac{\partial}{\partial \lambda_k} \log Z_k \mathbb{E}_{T_k(X)}$
2. $\frac{\partial^2}{\partial \lambda_k \partial \lambda_j} \log Z_k = \text{Cov}_{T_k(X), T_j(X)}$
3. $\log Z_k$ is convex in $\lambda$ and strictly convex unless $\exists a \in \mathbb{R}^k$ such that $\sum a_k T_k(x) = b$ for $b$ constant.
4. $\log Z(\lambda) - \sum \lambda_k \theta_k$ is convex in $\lambda$ and minimized when $\mathbb{E}[T(X)] = \theta_k$.

# Source Coding

**Source Code:** $C : \{1,\ldots,t\} \to \{0,1\}^*$

**Prefix Code:** a code $C$ for which $C(x)$ is not a prefix of $C(y)$ for $x \ne y$

**Kraft-McMillan Inequality:** For all prefix codes $C$,

$$\sum_{x=1}^t 2^{-|C(x)|} \le 1$$

and for any code lengths $\ell_1, \ldots, \ell_t$ satisfying

$$\sum_{x=1}^t 2^{-\ell_x} \le 1$$

there exists a prefix code $C$ with $|C(x)| = \ell_x$

**Optimal Coding:** Let $\vec{X} \sim p$. For the optimal code $C^* = \arg\min_{C \text{ prefix} \mathbb{E}_p |C(X)|}$,

$$H(p) \le \mathbb{E}_p |C(X)| \le H(p) + 1$$

# Statistical Inference

**Unbiased Estimator:** $\mathbb{E}[\widehat{\theta}] = \theta$

**Consistent Estimator:**

- Almost sure consistency: $\mathbb{P}(\lim_{n\to\infty} \widehat{\theta}_n = \theta) = 1$
- Consistent in probability: $\forall \varepsilon > 0, \mathbb{P}(\left|\widehat{\theta}_n - \theta\right| > \varepsilon) \to 0$
- Consistent in mean square: $\mathbb{E}[(\widehat{\theta}_n - \theta)^2] \to 0$.

**Mean Square Error:**

$$\text{MSE}(\widehat{\theta}_n) = \mathbb{E}[(\widehat{\theta}_n - \theta)^2] = \text{Var}\,[\widehat{\theta}] + \text{Bias}(\widehat{\theta})^2$$

**Theorem:** $\text{MSE}[\widehat{\theta}_n] \to 0 \implies \mathbb{P}(\left|\widehat{\theta}_n - \theta\right| > \varepsilon) \to 0$

**Bias:** $\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta}] - \theta$

**Variance:** $\text{Var}\,[\widehat{\theta}] = \mathbb{E}[\widehat{\theta}^2] - \mathbb{E}[\widehat{\theta}]^2$

**Kernel Density Estimation:** for a kernel $k$ satisfying

1. $k(x) \geq 0$
2. $\int xk(x)\, dx = 0$
3. $\int x^2 k(x)\, dx = 1$

we define the kernel density estimator

$$\widehat{f}_{n,w}(x; X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} k_w(x - X_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{w} k\left(\frac{x - X_i}{w}\right)$$

**Convolution:** Let $Z \sim f$ and $Y \sim g$ be independent. Then

$$Z + Y \sim (f \star g)(x) = \int_{\mathbb{R}} f(t)g(x-t)\, dt$$

**Integrated Square Error:**

$$\text{ISE}(\widehat{f}) = \int_{\mathbb{R}} \left|\widehat{f}_n(x; X_{1:n}) - f(x)\right|^2\, dx$$

**Mean Integrated Square Error:**

$$\text{MISE}(\widehat{f}) = \mathbb{E}[\text{ISE}(\widehat{f})] = \int_{\mathbb{R}} \mathbb{E}\left|\widehat{f}(x; X_{1:n}) - f(x)\right|^2\, dx$$

**Asymptotics:** For $f, k$ smooth, as $w \to 0$,

$$\text{MISE}(\widehat{f}_{n,w}) = \alpha w^4 + \frac{\beta}{nw} + \text{ error}$$

**Sylverman's Rule of Thumb:** For parameters $\alpha, \beta$ unknown, choose the kernel bandwidth $w \propto n^{-1/5}$

**Cross-Validation Estimator:**

$$\widehat{f}^{(i)}_{n,w}(x; X_{1:n}) = \frac{1}{n}\sum_{j \neq i} \widehat{f}_{n-1,w}(X_i)$$

**Stone's Theorem:** For

$$\widehat{w}_n = \arg\min_w \int \widehat{f}^2_{n,w}(x) - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}^{(i)}_{n-1,w}(X_i)\, dx$$

we have

$$\text{ISE}(\widehat{f}_{\widehat{w}_n}) \xrightarrow{a.s.} \inf_w \text{ISE}(\widehat{f}_{w,n}, f)$$

**Maximum Likelihood Estimation:** Sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_\theta$ for $\theta$ unknown. Then

$$\widehat{\theta} = \arg\max_\theta p_\theta(X_1 = x_1, \ldots, X_n = x_n) = \arg\min_\theta D(\widehat{p} \,\|\, p_\theta)$$

# Classification

Let $X \in \mathbb{R}^d$ be a RV with classes $Y \in \{1, \ldots, c\}$. Define $\pi_i = \mathbb{P}(Y = i)$ and $f_i(x)$ the class conditioned density.

Then for any set $A$,

$$\mathbb{P}_X(A) = \sum_{i=1}^{c} \pi_i \mathbb{P}(A)$$

$$f_x(A) = \sum_{i=1}^{c} \pi_i f_i(A)$$

**Bayes' Classification Rule:**

$$h^*(x) = \arg\max_{i=1:c} \mathbb{P}(Y = i \mid X = x)$$

**Neyman-Pearson Classification:** Fix $t \in (0, \infty)$. Then

$$h_t(x) = \begin{cases} 1 & \text{if } \frac{f_2(x)}{f_1(x)} < t \\ 2 & \text{if } \frac{f_2(x)}{f_1(x)} > t \end{cases}$$

**Note:** In the case $t = \frac{\pi_1}{\pi_2}$, then NP is equivalent to Bayes'.

**Detection Rate:** $\mathbb{P}(h(X) = 2 \mid Y = 2)$

**False Alarm Rate:** $\mathbb{P}(h(X) = 2 \mid Y = 1)$

**Theorem:** Fix $t \in (0, \infty)$ and choose any classifier $h$. If $\text{FAR}(h) \leq \text{FAR}(h_t)$, then $\text{DR}(h) \leq \text{DR}(h_t)$

## Generative Classifiers

*Motivation:* From Bayes,

$$h*(x) = \arg\max_{x \in \{1,\ldots,s\}} \mathbb{P}(Y = c \mid X = x) = \arg\max_c \frac{\pi_c f_c(x)}{\mathbb{P}(X = x)}$$

so we can estimate $\widehat{\pi}_c \approx \pi_c$ and then it suffices to estimate $f_c(x)$.

**Naive Bayes:** Assume $f_c(x_1, \ldots, x_d) = \prod_{j=1}^{d} f_{c,j}(x_j)$. Then instead of needing to find $(f_c)^s$ with $f_c : \mathbb{R}^d \to \mathbb{R}$, it suffices to find $f_{c,j} : \mathbb{R} \to \mathbb{R}$.

**Quadratic Discriminant Analysis:** Assume $f_c(x) \sim \mathcal{N}(\mu_c, \Sigma_c)$. Then

$$f_c(x, \mu_c, \Sigma_c) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{(x-\mu_c)\Sigma_c^{-1}(x-\mu_c)}{2}\right)$$

for $x, \mu_c in \mathbb{R}^d$ and $\Sigma_c \in \mathbb{R}^{d\times d}$.

**Linear Discriminant Analysis:** Assume $f_c(x) \sim \mathcal{N}(\mu_c, \Sigma)$.

## Discriminative Classifiers:

*Motivation:*

$$h*(x) = \arg\max_{x\in\{1,\dots,s\}} \mathbb{P}(Y = c \mid X = x) = \arg\max_{c} r_c(x)$$

**Linear Regression:** Assume $s = 2$ so $r_1(x) + r_2(x) = 1$. Then

$$\log\frac{r_2(x)}{r_1(x)} = \alpha + \beta x \implies r_2 = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

and it suffices to use MLE to estimate $\alpha, \beta$.

**Softmax:** Assume $s > 2$. Then model

$$\log\frac{r_k(x)}{r_1(x)} = \alpha_k + \beta_k x \implies r_k(x) = \frac{e^{\alpha_k+\beta_k x}}{1+\sum_{k=n}^{s} e^{\alpha_k+\beta_k x}}$$