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 λ 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,

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

Kullback-Leibler Divergence:

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

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

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

- f convex iff $f''(x) \geq 0$
- f concave iff $f''(x) \leq 0$ 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, $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$

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. $\frac{\partial^2}{\partial \lambda_k \partial \lambda_j} \log Z_k = \mathrm{Cov}_{\,T_k(X), T_j(X)}$
3. $\log Z_k$ is convex in λ 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 λ 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 $x \neq y$

Kraft-McMillan Inequality: For all prefix codes C,

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

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

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

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

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

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

Block coding: If instead we let $X_{1:n} \overset{\text{iid}}{\sim} p$ and $C_n^* = \arg\min_{C_n \text{ on } X_{1:n}} \mathbb{E}_p |C_n(X_{1:n})|$, then

$$H(p) \le \frac{1}{n}\mathbb{E}_p |C_n^*(X_{1:n})| \le H(p) + \frac{1}{n}$$

So by coding large enough blocks, we can get arbitrarily close the $H(p)$ bits/symbol.

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:

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

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

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

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

Kernel Density Estimation: for a kernel k satisfying

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

we define the kernel density estimator

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

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:

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

Mean Integrated Square Error:

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

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

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

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

Cross-Validation Estimator:

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

Stone's Theorem: For

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

we have

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

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

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