Binary hypothesis testing (2)

Find $\mathsf{H} \in \{H_0, H_1\}$ prior $P_0 = \mathbb{P}[H = H_0]$ and $P_1 = 1 - P_0$. Model consists of $p_{\mathsf{H}}, p_{\mathsf{y}|\mathsf{H}} \to \hat{H}(y)$. With cost function C_{ij} (guess i, truth j), we seek $\hat{H} = \operatorname{argmin}_{f(\cdot)} \varphi(f)$, for $\varphi(f) = \mathbb{E}[C(\hat{H}, f(y))]$. The **likelihood ratio test** minimizes **Bayes' risk** φ:

$$L(y) := \frac{p_{\mathsf{y}|\mathsf{H}}(y|H_1)}{p_{\mathsf{y}|\mathsf{H}}(y|H_0)} \gtrless_{H_0}^{H_1} \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} := \eta$$

Proof sketch: $f(y)$ can be computed pointwise on y, so compare the expected cost for $\hat{H}(y) = H_0, H_1$.

- **maximum a-posteriori rule**: C is 0-1 loss, rule compare $p_{\mathsf{H}|\mathsf{y}}(H_i|y)$.
- **maximum likelihood rule**: if also $P_1 = P_2 = \frac{1}{2}$, rule compare $p_{\mathsf{y}|\mathsf{H}}(y|H_i)$.

Neyman-Pearson/OCs (3,4)

$P_F = \mathbb{P}[\hat{H} = H_1 | \mathsf{H} = H_0]$, and $P_D = \mathbb{P}[\hat{H} = H_1 | \mathsf{H} = H_1]$. For LRTs $L(\mathsf{y}) \gtreqless \eta$, the **operating characteristic** is set $\{(P_F, P_D) \mid \eta\}$. The OC is $(1,1) \to (0,0)$ as η decreases, and is convex. Bayesian framework selects $\eta : (P_F, P_D)$ on LRT minimizing $\alpha P_F - \beta P_D + \gamma$ (determined by costs).

$$\zeta_{\mathrm{NP}}(\alpha) = \operatorname*{argmax}_{\hat{H}(\cdot)} P_D \text{ s.t. } P_F \le \alpha.$$

$\zeta_{\mathrm{NP}}(\alpha)$ is concave nondecreasing from $(0,0)$ to $(1,1)$.

Continuous case: if $L(\mathsf{y})$ continuous then the argmax $\hat{H}(\cdot)$ is of the form $\mathbf{1}[L(\mathsf{y}) \ge \eta]$.

Discrete case: equality in the LRT is possible; randomization $p_{\hat{H}|\mathsf{y}}(\cdot|y)$ can improve over deterministic decision rules. In particular, for rule $\hat{H}$ that picks $\hat{H}'$ with probability p and $\hat{H}''$ with probability $1-p$, we have $P_D(\hat{H}) = pP_D(\hat{H}') + (1-p)P_D(\hat{H}'')$, and same for P_F.

Neyman-Pearson Lemma: There exists an optimum Neyman-Pearson rule of the form $q_*(y) = 0$ if $L(y) < \eta$, p if $L(y) = \eta$, and 1 if $L(y) > \eta$.

Minimax hypothesis testing (5)

Costs C_{ij} known, but nature picks worst prior. Let $\varphi(p, r)$ be Bayes risk of $r(\cdot) = p_{\hat{H}|\mathsf{y}}(\cdot|y)$ under prior p. Seek $r_M(\cdot) = \operatorname{argmin}_{r(\cdot)} \max_p \varphi(p, r)$.

The **mismatched Bayes risk** $\varphi_B(p, q, \lambda)$ is risk $\varphi(p, r_B(\cdot; q, \lambda))$, where $r_B(\cdot; q, \lambda)$ is the Bayes rule for prior q. Hence **Bayes risk** is $\varphi_B^*(p) = \varphi_B(p, p, \lambda)$, independent of λ. Note φ_B is linear in p, and minimized over q at $q = p$. Also $\varphi_B^*(p)$ is concave, and $\varphi_B^*(0) = C_{00}, \varphi_B^*(1) = C_{11}$.

The **minimax decision rule** is $r_B(\cdot; p_*, \lambda_*)$, where p_*, λ_* corresopnd to the point (P_F^*, P_D^*) at the intersection of $\zeta_{\mathrm{NP}}(P_F^*)$ and

$$g_M(P_F^*) := \frac{C_{01} - C_{00}}{C_{01} - C_{11}} - \frac{C_{10} - C_{00}}{C_{01} - C_{11}} P_F.$$

If there is no intersection, set $p_* = 0$ if $\zeta_{\mathrm{NP}} > g_M(P_F)$ and $p_* = 1$ otherwise, and λ can be set arbitrarily

Minimax Inequality: in general,

$$\min_a \max_b g(a, b) \ge \max_b \min_a g(a, b).$$

In this case,

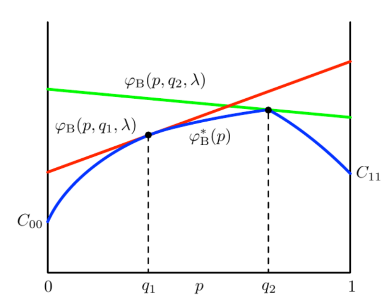
$$\min_{r(\cdot)} \max_p \varphi(p, r) = \max_p \min_{r(\cdot)} \varphi(p, r)$$

by Von Neumann's Theorem; one can show $\varphi(p, r)$ has a saddle point.

A **equalizer rule** is one such that

$$\mathbb{E}[C(H, \hat{H}) | H_0] = \mathbb{E}[C(H, \hat{H}) | H_1],$$

i.e. the Bayes risk of the rule is the same regardless of the prior p. The minimax rule is equalizer rule if $p_* \in \{0, 1\}$.



Bayesian parameter estimation (6)

With the posterior $p_{\mathsf{x}|\mathsf{y}}(x|y) = \frac{p_{\mathsf{y}|\mathsf{x}}(y|x)}{\int p_{\mathsf{y}|\mathsf{x}}(y|x)p_{\mathsf{x}}(x)\,dx}$, choose $\hat{x} = \operatorname{argmin}_{f(\cdot)}[C(\mathsf{x}, f(y))]$ for some cost function C. $\hat{x}(y)$ should be minimized pointwise.

Let error $e(\mathsf{x}, \mathsf{y}) = \hat{x}(y) - \mathsf{x}$, so bias $b = \mathbb{E}[e(\mathsf{x}, \mathsf{y})]$. **Error covariance matrix** $\Lambda_e = \mathbb{E}[(e-b)(e-b)^\intercal]$, **error correlation matrix** $\mathbb{E}[ee^\intercal] = \Lambda_e + bb^\intercal$. MSE is $\operatorname{tr}(\mathbb{E}[ee^\intercal])$.

- **min absolute error**: $C(x, \hat{x}) = |x - \hat{x}|$ gives the median of the posterior.
- **min uniform cost** $C(x, \hat{x}) = \mathbf{1}[|x - \hat{x}| > \varepsilon]$ gives the mode as $\varepsilon \to 0$ (MAP estimator).

Bayesian Least Squares $C(a, \hat{a}) = \|a - \hat{a}\|^2$ gives mean of posterior, $\hat{x}(y) = \mathbb{E}[\mathsf{x}|\mathsf{y} = y]$. The BLS estimator is unbiased, with error covariance & correlation matrix $\mathbb{E}[\Lambda_{\mathsf{x}|\mathsf{y}}(\mathsf{y})]$.

Orthogonality: $\hat{x}$ is $\hat{x}_{\mathrm{BLS}}$ iff unbiased and

$$\mathbb{E}[(\hat{x}(\mathsf{y}) - \mathsf{x})g(\mathsf{y})^\intercal] = 0 \text{ for all } g(\cdot).$$

Linear least-squares* (7)

LLS estimator $\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) = \operatorname{argmin}_{f(\cdot)\in\mathcal{B}} \mathbb{E}[\|x - f(\mathbf{y})\|^2]$, where $\mathcal{B} = \{f(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{d}\}$.

Orthogonality: $\hat{\mathbf{x}}$ is the LLS estimator iff unbiased and $\mathbb{E}[(\hat{\mathbf{x}} - \mathbf{x})\mathbf{y}^\intercal] = 0$. This is equivalent to $\mathbb{E}[(\hat{\mathbf{x}}(y) - \mathbf{x})(\mathbf{F}\mathbf{y} + \mathbf{g})^\intercal] = 0 \quad \forall \mathbf{F}, \mathbf{g}$. Closed-form:

$$\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) = \mu_{\mathsf{x}} + \Lambda_{\mathsf{xy}}\Lambda_{\mathsf{y}}^{-1}(\mathbf{y} - \mu_{\mathsf{y}}).$$

Gaussian random variables $x = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, vectors $p_{\mathsf{x}}(\mathbf{x}) = \frac{1}{\sqrt{|2\pi\Lambda|}}e^{-\frac{1}{2}(\mathbf{x}-\mu)^\intercal\Lambda^{-1}(\mathbf{x}-\mu)}$. Then $a^\intercal\mathbf{x}$ is Gaussian. If $\mathbf{x}, \mathbf{y}$ are jointly Gaussian, $\hat{x}_{\mathrm{BLS}} = \hat{x}_{\mathrm{LLS}}$.

Nonbayesian param. estimation (8)

Only consider *valid* $f(\cdot)$, i.e. independent of x. Let $e(\mathsf{y}) = \hat{\mathsf{x}} - x$, $b_{\hat{\mathsf{x}}}(x) = \mathbb{E}[e(\mathsf{y})]$, and $\Lambda_e(x) = \mathbb{E}[(e-b)(e-b)^\intercal]$. MSE is trace of

$$\mathbb{E}[e(\mathsf{y})e(\mathsf{y})^\intercal] = \Lambda_e(x) + b_{\hat{\mathsf{x}}}(x)b_{\hat{\mathsf{x}}}(x)^\intercal.$$

A **minimum variance unbiased** estimator satisfies $\lambda_{\hat{\mathsf{x}}^*}(x) \leq \lambda_{\hat{\mathsf{x}}}(x)$ for every x. This does not have to exist.

Cramér-Rao: if $\mathbb{E}\left[\frac{\partial}{\partial x}\ln p_{\mathsf{y}}(\mathsf{y}; x)\right] = 0$ for all x, then variance of unbiased estimator $\hat{x}$ is at least

$$\lambda_{\hat{\mathsf{x}}}(x) \geq 1/J_{\mathsf{y}}(x) \quad J_{\mathsf{y}}(x) = \mathbb{E}[S(\mathsf{y}; x)^2]$$
$$S(y; x) = \frac{\partial}{\partial x}\ln p_{\mathsf{y}}(y; x).$$

Vector Cramér-Rao: if $\mathbb{E}\left[\frac{\partial}{\partial \mathbf{x}}\ln p_{\mathsf{y}}(\mathsf{y}; \mathbf{x})\right] = 0$ then for any unbiased $\hat{\mathsf{x}}(\cdot)$, $\Lambda_{\hat{\mathsf{x}}} = \operatorname{Cov}[\hat{\mathsf{x}}|\mathbf{x}]$ satisfies

$$\Lambda_{\hat{\mathsf{x}}} - J_{\mathsf{y}}^{-1}(\mathbf{x}) \succeq 0, \quad J_{\mathsf{y}}(\mathbf{x}) = \mathbb{E}[S(\mathsf{y}; \mathbf{x})S(\mathsf{y}; \mathbf{x})^\intercal]$$
$$S(\mathsf{y}; \mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}\ln p_{\mathsf{y}}(\mathsf{y}; \mathbf{x}).$$

Also, $\mathbb{E}\left[\left(\frac{\partial}{\partial x}\ln p_{\mathsf{y}}(y; x)\right)^2 + \frac{\partial^2}{\partial x^2}\ln p_{\mathsf{y}}(y; x)\right] = 0$, implies $J_{\mathsf{y}}(x) = -\mathbb{E}\left[\frac{\partial^2}{\partial x^2}\ln p_{\mathsf{y}}(\mathsf{y}; x)\right]$.

Efficent estimators match equality of Cramér-Rao bound for all x, i.e. is of the form

$$\hat{x}(y) = x + \frac{1}{J_{\mathsf{y}}(x)}S(y; x),$$

(where the RHS must be independent of x). Efficient estimators are unbiased & unique & MVU if existent.

The **Maximum Likelihood estimator** is

$$\hat{x}_{\mathrm{ML}}(y) = \operatorname*{argmax}_x p_{\mathsf{y}}(y; x).$$

$\hat{x}_{\mathrm{ML}}$ commutes with invertible transformations: if $\theta = g(x)$, $\theta_{\mathrm{ML}}(\cdot) = g(\hat{x}_{\mathrm{ML}}(\cdot))$. Also, if both exist, $\hat{x}_{\mathrm{ML}}(\cdot) = \hat{x}_{\mathrm{eff}}(\cdot)$.

Gauss-Markov Theorem: If $\mathsf{y} = \mathbf{H}\mathsf{x} + \mathsf{w}$ for $\mathsf{w} \sim N(0, \Lambda)$, then MVU estimator $\hat{\mathbf{x}}_{\mathrm{ML}}(\mathbf{y})$ is

$$(\mathbf{H}^\intercal\Lambda^{-1}\mathbf{H})^{-1}\mathbf{H}^\intercal\Lambda^{-1}\mathbf{y} = \operatorname*{argmin}_{\mathbf{x}} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2.$$

Exponential Families (9)

An **single-param. exponential family** is of form

$$p_{\mathsf{y}}(y; x) = \exp(\lambda(x)t(y) - \alpha(x) + \beta(y)),$$

"Natural parameter" $\lambda(\cdot)$, "natural statistic" $t(\cdot)$, "log base function" $\beta(\cdot)$. Note $\alpha(\cdot)$ normalizes, and $t(\cdot)$, $\beta(\cdot)$ can be shifted by constants. Call $q(y) \propto \exp(\beta(y))$ the **base distribution**.

$$\dot\alpha(x) = \dot\lambda(x)\mathbb{E}[t(\mathsf{y})]$$
$$\ddot\alpha(x) = \ddot\lambda(x)\mathbb{E}[t(\mathsf{y})] + \dot\lambda(x)^2\mathbb{V}[t(\mathsf{y})]$$
$$J_{\mathsf{y}}(x) = \dot\lambda(x)\frac{d}{dx}\mathbb{E}[t(\mathsf{y})]$$

A **canonical exponential family** has $\lambda(x) = x$. $e^{\alpha(x)} = \mathbb{E}_q[e^{xy}]$, $\dot\alpha(x) = \mathbb{E}[t(y)]$, $\ddot\alpha(x) = \mathbb{V}[t(y)]$. Examples include the *geometric mean* $p(\cdot;x) \propto q_1(\cdot)^x q_2(\cdot)^y$ and *tilted* $p(\cdot;x) \propto q(y)e^{xt(y)}$.

Multi-parameter exponential family:

$$p_{\mathsf{y}}(\mathbf{y};\mathbf{x}) = \exp(\lambda(x)^\intercal \mathbf{t}(\mathbf{y}) - \alpha(\mathbf{x}) + \beta(\mathbf{y})).$$

Over finite alphabet $\mathcal{Y}$, any distribution q can be generated by one exponential family $p_{\mathsf{y}}(y;\mathbf{x}) = \exp\left(\sum_{i\in\mathcal{Y}} x_i t_i(y) - \alpha(\mathbf{x})\right)$ via $x_i = \ln q(i)$.

Sufficient Statistics (10)

A *sufficient* statistic $t(\cdot)$ w.r.t. family $\{p_{\mathsf{y}}(\cdot;x)\}$ is such that $p_{\mathsf{y}|\mathsf{t}}(\cdot|\cdot;x)$ is independent of x. Equivalently $\frac{L_y(x)}{L_t(x)} := \frac{p_{\mathsf{y}}(y;x)}{p_{\mathsf{t}}(t(y);x)}$ is not a function of x.

Neyman Factorization: $t(\cdot)$ is sufficient iff exist $a(\cdot,\cdot)$ and $b(\cdot)$ such that

$$p_{\mathsf{y}}(y;x) = a(t(y),x)b(y).$$

Sufficient statistic $s(\cdot)$ is **minimal** if for any sufficient t, exists g such that $s = g \circ t$.

For example, for finite $\mathcal{X}$, $t(y) = \langle p_{\mathsf{y}}(y;x_i)\rangle$ is sufficient, and $t(y) = \frac{p_{\mathsf{y}}(y;x_i)}{p_{\mathsf{y}}(y;x_0)}$ is minimal.

A sufficient statistic t is **complete** if the only φ such that $\mathbb{E}[\varphi(t(\mathsf{y}))] = 0$ for all $x \in \mathcal{X}$ is $\varphi(\cdot) \equiv 0$. Completeness implies minimality: if t is complete and s is minimal, let $s = g(t)$; then $\mathbb{E}[\mathsf{t}|\mathsf{s} = s]$ is a function of s and hence t, so $\mathsf{t} - \mathbb{E}[\mathsf{t}|\mathsf{s} = s]$ has mean zero, hence is zero, so t is a function of s.

Bayesian formulation: $t(\cdot)$ is sufficient w.r.t. $p_{\mathsf{x},\mathsf{y}}$ if $p_{\mathsf{x}|\mathsf{y}}(\cdot|y) = p_{\mathsf{x}|\mathsf{t}}(\cdot|t(y))$. Neyman factorization looks like $p_{\mathsf{y}|\mathsf{x}}(y|x) = p_{\mathsf{t}|\mathsf{x}}(t(y)|x)p_{\mathsf{y}|\mathsf{t}}(y|t(y))$.

Statistics are partitions: If $L_{y_1}(x) \propto L_{y_2}(x)$ then y_1, y_2 provide "the same information" about x. Sufficient statistics group all y_i with proportional $L_{y_i}(\cdot)$ together; minimum sufficient statistics in addition group all y_i with non-proportional $L_{y_i}(\cdot)$ apart.

EM algorithm (12)

Observed data y from r.v. $\mathsf{y} \sim p_{\mathsf{y}}(\cdot;x)$. We wish to compute $\hat{x}_{\mathrm{ML}}$. Find some **complete data** z such that $\mathsf{y} = g(\mathsf{z})$, such that ML estimator of z is easy.

1. Set $\ell = 0$ and initialize $\hat{x}$ arbitrarily.
2. Find $U(x;\hat{x}) = \mathbb{E}_{p_{\mathsf{z}|\mathsf{y}}(\cdot|y;\hat{x})}[\log p_{\mathsf{z}}(\mathsf{z}|x)]$.
3. Set new $\hat{x}$ to $\operatorname{argmax}_x U(x;\hat{x})$.

It is possible to show

$$\ell_{\mathsf{y}}(x;y) = U(x,x') - U(x',x') + \ell_{\mathsf{y}}(x';y),$$

So the new $\hat{x}$ increases log-likelihood $\ell_{\mathsf{y}}(\hat{x};y)$ at each step. This converges to a stationary point.

For *Gaussian mixture models*, i.e. of the form $p_{\mathsf{y}}(\mathbf{y};\theta) = \sum_k \pi_k N(\mathbf{y};\mu_k,\Lambda_k)$, and complete data is y along with the weights π_i. More generally, for *exponential families*, i.e. complete data $\mathbf{z}$ is in

$$p_{\mathsf{z}}(\mathbf{z};\mathbf{x}) = \exp(\mathbf{x}^\intercal t(\mathbf{z}) - \alpha(\mathbf{x}) + \beta(\mathbf{z})),$$

we have $U(\mathbf{x},\mathbf{x}') = \mathbf{x}^\intercal \mathbb{E}_{p_{\mathsf{z}|\mathsf{y}}(\cdot|\mathbf{y};\mathbf{x}')}[t(\mathbf{z})] - \alpha(\mathbf{x}) + \mathbb{E}_{p_{\mathsf{z}|\mathsf{y}}(\cdot|\mathbf{y};\mathbf{x}')}[\beta(\mathsf{z})]$; solve partials and substitute $\frac{\partial}{\partial x_i}\alpha(\mathbf{x}) = \mathbb{E}_{p_{\mathsf{z}}(\cdot;\mathbf{x})}[t_i(\mathsf{z})]$ to get new $\hat{\mathbf{x}}'$ from $\hat{\mathbf{x}}$:

$$\mathbb{E}_{p_{\mathsf{z}}(\cdot;\hat{\mathbf{x}}')}[t_k(\mathsf{z})] = \mathbb{E}_{p_{\mathsf{z}|\mathsf{y}}(\cdot|\mathbf{y};\hat{\mathbf{x}})}[t_k(\mathsf{z})].$$

Decision theory (13)

Inference predicts a probability distribution $q(\cdot)$ for x. The only smooth cost function $C(x,q)$ s.t. **proper**: $p_{\mathsf{x}|\mathsf{y}}(\cdot|y) = \operatorname{argmin}_q \mathbb{E}[C(x,q)|\mathsf{y} = y]$ and **local**: $C(x,q)$ is a function of x and $q(x)$, is **log-loss**: $-A\log q(x) + B(x)$, as long as $|\mathcal{X}| \geq 3$.

Entropy: $H = -\sum_a p(a)\log p(a) = \mathbb{E}[C(\mathsf{x},p_{\mathsf{x}})]$. This is concave in p. **Conditional entropy**:

$H(\mathsf{x}|\mathsf{y}) = \mathbb{E}[C(\mathsf{x}, p_{\mathsf{x}|\mathsf{y}})]$; expands

$$H(\mathsf{x}|\mathsf{y}) = \sum_y p_{\mathsf{y}}(y) H(\mathsf{x}|\mathsf{y} = y).$$

Note $H(\mathsf{x}|\mathsf{y}) \le H(\mathsf{x}) \le \log|\mathcal{X}|$, and $H(\mathsf{x},\mathsf{y}) = H(\mathsf{x}|\mathsf{y}) + H(\mathsf{y})$.

Mutual information $I(\mathsf{x};\mathsf{y}) = H(\mathsf{x}) - H(\mathsf{x}|\mathsf{y})$. This is *symmetric*, since it can be computed as $\sum_{x,y} p_{\mathsf{x},\mathsf{y}}(x,y) \log \frac{p_{\mathsf{x},\mathsf{y}}(x,y)}{p_{\mathsf{x}}(x)p_{\mathsf{y}}(y)}$. Some identities:

- $I(\mathsf{x};\mathsf{y}) = 0$ iff x,y independent
- $I(\mathsf{x};\mathsf{y},\mathsf{z}) = I(\mathsf{x};\mathsf{z}) + I(\mathsf{x};\mathsf{y}|\mathsf{z})$ where $I(\mathsf{x};\mathsf{y}|\mathsf{z}) = H(\mathsf{x}|\mathsf{z}) - H(\mathsf{x}|\mathsf{y},\mathsf{z})$
- $H(\mathsf{x},\mathsf{y}) = H(\mathsf{x}) + H(\mathsf{y}) - I(\mathsf{x};\mathsf{y})$.

Data Processing Inequality: $I(\mathsf{x};\mathsf{y}) \ge I(\mathsf{x};\mathsf{t})$, with equality if and only if $t(\mathsf{y})$ is sufficient.

Information (KL) Divergence: given $\mathsf{x} \sim p(\cdot)$, $D(p\|q) = \mathbb{E}[C(\mathsf{x},q) - C(\mathsf{x},p)] = \sum_a p(a) \log \frac{p(a)}{q(a)}$. $D(p\|U) = \log|\mathcal{X}| - H(p)$, $I(\mathsf{x};\mathsf{y}) = D(p_{\mathsf{x},\mathsf{y}}\|p_{\mathsf{x}}p_{\mathsf{y}})$. This is convex in p and q.

If $\mathbb{E}[\frac{\partial}{\partial x} \ln p_{\mathsf{y};x}] = 0$ and $\mathbb{E}[\frac{\partial^3}{\partial x^3} \ln p_{\mathsf{y};x}] \ne \pm\infty$, then $D(p_{\mathsf{y}}(\cdot;x)\|p_{\mathsf{y}}(\cdot;x+\delta)) = \frac{\log(e)}{2} J_{\mathsf{y}}(x)\delta^2 + o(\delta^2)$.

Information geometry (14)

I-projection $p^* = \mathrm{argmin}_{p\in\mathcal{P}} D(p\|q)$. If p^* is projection of q onto closed/convex $\mathcal{P}$, then $D(p\|q) \ge D(p\|p^*) + D(p^*\|q)$ for any $p \in \mathcal{P}$.

Linear family: family s.t. $\{p : \mathbb{E}_p[t_i(\mathsf{y})] = \overline{t_i}\}$. For $p_1, p_2 \in \mathcal{P}$, $\lambda p_1 + (1-\lambda)p_2 \in \mathcal{P}$ for any $\lambda \in \mathbb{R}$. Equality holds in Pythagorean Theorem if $\mathcal{P}$ is linear.

Orthogonal family: given p^* in linear $\mathcal{P}$ with parameters $t(\cdot)$, $\mathcal{E}_t(p^*)$ is all p projecting to p^*.

$$\mathcal{E}_t(p^*) = \{q \mid q(y) = p^*(y)\exp(\mathbf{x}^\intercal t(y) - \alpha(\mathbf{x}))\}.$$

Modeling as inference (15)

Mixture models $q_w(y) = \sum_{x\in\mathcal{X}} w(x) p_{\mathsf{y}}(y;x)$; w is a "prior". *Optimality*: for any distribution $q(\cdot)$, there exist weights w such that q_w is strictly better: $D(p_{\mathsf{y}}(\cdot;x)\|q_w) \le D(p_{\mathsf{y}}(\cdot;x)\|q)$ for all x.

Minimax perspective: $\min_{q(\cdot)} \max_x D(p_{\mathsf{y}}(\cdot;x)\|q)$. **Redundancy-capacity**: $\max\min = \min\max$. Note the inner maximization is

$$\max_x D(p_{\mathsf{y}}(\cdot;x)\|q) = \max_w \sum_x w(x) D(p_{\mathsf{y}}(\cdot;x)\|q).$$

Interpret w as a **least informative prior** $p_{\mathsf{x}}^* = \mathrm{argmax}_{p_{\mathsf{x}}} I(\mathsf{x};\mathsf{y})$. **Model capacity** is $C = \max_{p_{\mathsf{x}}} I(\mathsf{x};\mathsf{y}) \le \log|\mathcal{X}|$. **Equidistance property**: at mixture q^* and weights w^*, $D(p_{\mathsf{y}}(\cdot;x)\|q^*) \le C\ \forall x$, equality when $w^*(x) > 0$.

Continuous information theory (16)

- Entropy $h(\mathsf{x}) = -\int p(x)\log p(x)\,dx$. Conditional $h(\mathsf{x}|\mathsf{y}) = -\int p_{\mathsf{y}}(y) h(\mathsf{x}|\mathsf{y}=y)\,dy$.
- Mutual information $I(\mathsf{x};\mathsf{y}) = h(\mathsf{x}) - h(\mathsf{x}|\mathsf{y})$.
- Divergence $D(p\|q) = \int p(x) \log \frac{p(x)}{q(x)}\,dx$.

For Gaussians, $h(\mathsf{x}) = \frac{\log e}{2}(\ln|2\pi\Lambda| + \mathrm{tr}(I))$, $h(\mathsf{x}|\mathsf{y}) = \frac{1}{2}\log|2\pi e(\Lambda_{\mathsf{x}} - \Lambda_{\mathsf{xy}}\Lambda_{\mathsf{y}}^{-1}\Lambda_{\mathsf{xy}}^\intercal)|$.

Some useful theorems (11 & misc.)

Jensen's inequality: For concave $\varphi(\cdot)$ and random variable v,

$$\mathbb{E}[\varphi(v)] \gtreqless \varphi(\mathbb{E}[v]).$$

Csiszár's inequality: for convex f,

$$\sum_{i=1}^N b_i f\left(\frac{a_i}{b_i}\right) \ge \left(\sum_{i=1}^N b_i\right) f\left(\frac{\sum_{i=1}^N a_i}{\sum_{i=1}^N b_i}\right).$$

Log-sum: $\sum a_i \log \frac{a_i}{b_i} \ge (\sum a_i)\log\frac{\sum a_i}{\sum b_i}$.

Gibbs' inequality: let r.v. $\mathsf{v} \sim p(\cdot)$. For any distribution $q(\cdot)$, $\mathbb{E}_p[\log p(\mathsf{v})] \ge \mathbb{E}_p[\log q(\mathsf{v})]$.

Lagrange multipliers: $\nabla f(\hat{\mathbf{z}}) + \lambda^\intercal \nabla g(\hat{\mathbf{z}}) = 0$.

Taylor's Theorem: for some t' between t and t_0,

$$f(t) = \sum_{j=0}^J (t-t_0)^j \frac{f^{(j)}(t_0)}{j!} + (t-t_0)^{J+1}\frac{f^{(J+1)}(t')}{(J+1)!}$$