6.437 Notes

Anugrah Chemparathy

L2 - Bayesian Hypothesis Testing

Suppose we are trying to decide which hypothesis from a set $\mathcal{H} = \{H_0, \cdots, H_m\}$ was responsible for some empirical observations y.

Suppose you have some a prioiri probability of each hypothesis class being responsible, and can write the conditional probability of observing the empirical data for each H:

$$p_H(H_m) \qquad\qquad p_{y|H}(\cdot, H_m)$$

Using Bayes' rule we can rewrite the a posteriori probabilities of each hypothesis as:

$$p_{H|y}(H_m|y) = \frac{p(y|H)p(H)}{p(y)} = \frac{p(y|H)p(H)}{\sum_{H_i} p(y|H_i)p(H_i)}$$

Now we consider the simpler binary hypothesis testing problem. We will construct a decision rule in the form of a likelihood ratio which we can show has the lowest expected cost.

- Define $C(H_j, H_i) \triangleq C_{ij}$ as the cost of deciding the hypothesis is H_i when it is actually H_j.
- A **valid** set of costs has $C_{jj} < C_{ij}$ for $i \neq j$

The optimal decision rule is:

$$\hat{H}(\cdot) = \arg\min_{f(\cdot)} \varphi(f) \qquad \text{where } \varphi(f) \triangleq E[C(H, f(y)]$$

- The expectation is taken over both y and H and $f(\cdot)$ is our decision rule.

We call the expected cost $\varphi(f)$ the **Bayes risk**.

Theorem 2.1: Given P_i, valid costs C_{ij}, and data y, the Bayesian decision rule is:

$$L(y) \triangleq \frac{p_{y|H}(y|H_1)}{p_{y|H}(y|H_0)} \underset{\hat{H}=H_0}{\overset{\hat{H}=H_1}{\gtrless}} \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} \triangleq \eta$$

It happens that $L(y)$ is a sufficient statistic. As it turns out any invertible function g of $L(y)$ is also a sufficient statistic (i.e. $g(L(y)) \gtrless g(\eta)$ is an equally good decision rule).

Corollary 2.1 (Minimum Probability of Error): Using the 1-0 error loss function $C_{01} = C_{10} = 1$

$$\varphi(\hat{H}) = P(\hat{H}(y) = H_0, H = H_1) + P(\hat{H}(y) = H_1, H = H_0)$$

has the decision rule $\hat{H}(y) = \arg\max_H p_{H|y}(H|y)$ and (corollary 2.2) when both hypotheses are equally likely is equivalent to $\hat{H}(y) = \arg\max_H p_{y|H}(y|H)$

L3 - NonBayesian Hypothesis Testing

For many formulations of binary hypothesis testing, the optimal deterministic decision rule turns out to be an LRT of the same form as that from L2.

$$P_R \triangleq P_D \triangleq P(\hat{H}(y) = H_1 | H = H_1) \qquad \textbf{Recall}$$
$$P_F \triangleq P(\hat{H}(y) = H_1 | H = H_0) \qquad \text{Size}$$

Alternatively

$$P_E^1 \triangleq P(\hat{H}(y) = H_1 | H = H_0) \qquad \text{Type 1}$$
$$P_E^2 \triangleq P(\hat{H}(y) = H_0 | H = H_1) \qquad \text{Type 2}$$

Also we define the **Precision**

$$P_P \triangleq P(H = H_1 | \hat{H}(y) = H_1) = \frac{p_D p_H(H_1)}{p_{\hat{H}}(\hat{H}(y) = H_1)} = \frac{P_D p_H(H_1)}{P_F P_H(H_0) + P_D P_H(H_1)}$$

As we sweep the LRT threshold η, we trace out a curve of points in the (P_f, P_D) plane where:

$$P_D(\eta) = P(L(y) \geq \eta | H = H_1)$$
$$P_F(\eta) = P(L(y) \geq \eta | H = H_0)$$

this is called the operating characteristic of the LRT (the OC-LRT).

The OC-LRT (figure 3.1) has the general shape of the top left of a quarter circle - i.e. there is a tradeoff between having a large P_D (you will have a high P_F as well). We can actually rearrange the Bayes risk from L2 as $\varphi(f) = \alpha P_F - \beta P_D + \gamma$.

Property 3.1: The OC-LRT is monotonically nondecreasing.

Neyman-Pearson Hypothesis Testing: maximize P_D under a P_F constraint.

$$\arg\max_{\hat{H}(\cdot)} P_D \quad \text{subject to} \quad P_F \leq \alpha$$

Theorem 3.1 (Neyman Pearson lemma) We can show that one deterministic decision rule solution to the above is

$$\frac{p_{y|H}(y|H_1)}{p_{y|H}(y|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda$$

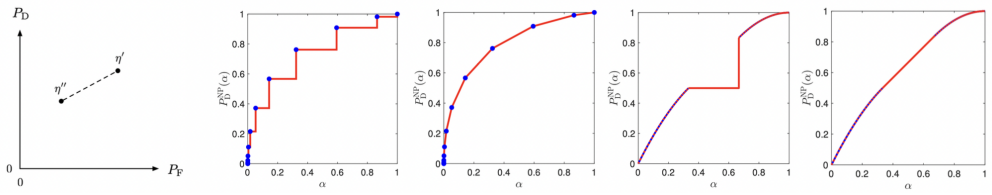
where λ is chosen such that $P_F = \alpha$ exactly using Lagrange Multipliers. This needs no prior!

We also define **p-values** by mapping the $LRT \to [0, 1]$ using an invertible g (as in L2).

L4 - Performance Limits of Hypothesis Testing

The OC-LRT graph traced out by P_D, P_F at varying η is not necessarily continuous. Note that this can happen if the underlying data distribution is discrete, or even continuous but with some piecewise components.

This can be suboptimal, particularly in the Neyman-Pearson approach. If there is a discontinuity along the OC-LRT curve with points corresponding to η' and η'' on each side and $P_F(\eta') < \alpha < P_F(\eta'')$ then we will be forced to use η' as our threshold in the deterministic approach.



We can use a randomized LRT decision rule to interpolate between the two points on the OC-LRT

$$\hat{H}(y) = \begin{cases} \hat{H}'_{\eta'}(y) & u = \eta' \\ \hat{H}''_{\eta''}(y) & u = \eta'' \end{cases}$$

Simply choosing one of two standard LRT decision rules with each threshhold with a tunable Bernoulli p achieves $P_D = pP_D(\eta') + (1-p)P_D(\eta'')$ and $P_F = pP_F(\eta') + (1-p)P_F(\eta'')$.

Equivalently, with $u \sim B(p)$, we can write

$$\hat{H}(y) = \begin{cases} H_0 & L(y) < \eta' \\ H_{1(u=\eta')} & \eta' \le L(y) < \eta'' \\ H_1 & L(y) \ge \eta'' \end{cases} \qquad p_{\hat{H}|y}(H_1|y) = \begin{cases} 0 & L(y) < \eta' \\ p & \eta' \le L(y) < \eta'' \\ 1 & L(y) \ge \eta'' \end{cases}$$

We treat the randomized decision rule $\hat{H}(\cdot)$ as solely a function of the detail through a "Markov chain" formalism: $H \leftrightarrow y \leftrightarrow \hat{H}$. Ultimately we can write stuff like:

$$p_{\hat{H}|y,H}(\cdot|y,H) = P_{\hat{H}|y}(\cdot|y) \quad \text{and} \quad p_{H|y,\hat{H}}(\cdot|y,\hat{H}) = P_{H|y}(\cdot|y)$$

We then define the randomized decision rule **as a conditonal distribution**: $P_{\hat{H}|y}(H_1|y) \triangleq q(y)$

Claim 4.1 A randomized test cannot achieve a lower Bayes' risk than the optimum LRT $(1_{L(y)\ge\eta})$ in binary Bayesian Hypothesis testing.

$$P_F(q) = P(\hat{H}(y) = H_1|H = H_0) = \int q(y)P_{y|H}(y|H_0)dy = E[q(y)|H = H_0]$$

$$p_D(q) = P(\hat{H}(y) = H_1|H = H_1) = \int q(y)P_{y|H}(y|H_1)dy = E[q(y)|H = H_1]$$

Theorem 4.1 Given hypotheses H_0, H_1 and some $\alpha \in [0,1]$, and a Bernoulli $u \sim B(p)$ the decision rule

$$q(y) = \begin{cases} 0 & L(y) < \eta' \\ p & \eta' \leq L(y) < \eta'' \\ 1 & L(y) \geq \eta'' \end{cases}$$

satisfies $P_F(q_*) = \alpha$ and $P_D(q_*) \geq P_D(q)$ for any decision rule q with $P_F(q) \leq \alpha$.

Neyman Pearson Function: We define ζ_{NP} as a function mapping from $P_F \to P_D$ along the Pareto optimal frontier which we define as $\mathcal{F}_{p_Y|H}$.

- The given proof requires the observation that the smallest η_0 with $P(L(y) > \eta_0 | H = H_0) < \alpha$ implies a point mass at $L(y) = \eta_0$, and no point mass when the inequality is no longer strict (i.e. $\leq \alpha$).

Property 4.1: The Neyman-Pearson function satisfies $\zeta_{NP}(1) = 1$ i.e. $(1,1) \in \mathcal{F}_{p_y|H}$

Property 4.2: The Neyman-Pearson function satisfies $\zeta_{NP}(P_F) = P_D \geq P_F$.

Property 4.3: The Neyman-Pearson function is concave

Property 4.4: Let η_0 be any LRT threshold such that there is no point mass at η_0 for under either hypothesis (i.e. $P(L(y) = \eta_0 | H = H_0) = P(L(y) = \eta_0 | H = H_1) = 0$).

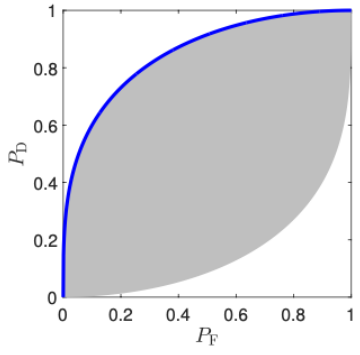
$$\dot{\zeta_{NP}}(P_F(\eta_0)) = \eta_0$$

Then the slope of the Neyman-Pearson function is equal to the threshold at each point η_0.

4.6 Summary - Region of Possible Operating Points

If we have a decision rule corresponding to (P_F^*, P_D^*), then the reversed decision rule (corresponding to reflection over $(1/2, 1/2)$) would correspond to $(1 - P_F^*, 1 - P_D^*)$.

As a result, you cannot do worse than the inverse of the Pareto-optimal frontier (worst = the lower right edge of the figure below). Additionally you can create a decision rule mapping to any point in the interior by simply interpolating between two points on each side.



L5 - Minimax Hypothesis Testing

Under the minimax framework, we assume that nature will choose the most detrimental prior for whatever decision rule we choose, and we must choose our decision rule with this in mind. We allow nature to randomly choose one of $\{H_0, H_1\}$ randomly with probability $p = P(H = H_1)$, making the Bayes risk for r (irrespective of if r is a good decision rule):

$$\varphi(p, r) = (1-p)\underbrace{E[C(H,\hat{H})|H = H_0]}_{\triangleq \varphi_0(r)} + p\underbrace{E[C(H,\hat{H})|H = H_1]}_{\triangleq \varphi_1(r)}$$

The optimum randomized decision rule $\hat{H}_m$ is

$$r_M(\cdot) = \arg\min_r \varphi_M(r) \qquad \varphi_M(r) \triangleq \max_{p\in[0,1]} \varphi(p, r)$$

In Bayesian hypothesis testing we used an LRT decision rule, with prior parameter q and bernoulli decision parameter λ (for use when $L(y) = \eta$ exactly). We define the **mismatch Bayes risk** as

$$\varphi_B(p, q, \lambda) \triangleq \varphi(p, r_B(\cdot; q, \lambda))$$

where p is the true prior for $p(H)$. The **matched Bayes risk** is:

$$\varphi_B^*(p) \triangleq \varphi_B(p, p, \lambda)$$

Claim 5.1: We can show

1. $\varphi_B(\cdot, q, \lambda)$ is a linear function
$$\varphi_B(p, q, \lambda) = \varphi_B^0(q, \lambda) + p[\varphi_B^1(q, \lambda) - \varphi_B^0(q, \lambda)]$$
where φ_B^i is the Bayes risk conditioned on $H = H_i$
2. $\varphi_B(p, q, \lambda)$ is lower bounded by $\varphi_B^*(p)$
3. $\varphi_B^*(\cdot)$ is concave and continous on $[0, 1]$
4. $\varphi_B^*(0) = C_{00}$ and $\varphi_B^*(1) = C_{11}$

Note that in general, $\varphi_B^*(p)$ will be differentiable at p iff the optimum Bayes decision rule is "achieved at a unique point on the efficient frontier". When two optimal decision rules on the efficient frontier corresponding to the same p, there is a non-differentiable point (I guess).

Fact 5.1 For any real valued function g we can show

$$\min_a \max_b \geq \max_b \min_a g(a, b)$$

- Essentially it is optimal to choose last in such games.

Theorem 5.1: Given data models $P_{y|H}(\cdot|H_i)$ and valid costs, a minimax decion rule of $r_*(\cdot) \triangleq r_B(\cdot; p_*, \lambda_*)$ where p_* is the minimax prior:

$$\min_r \max_p \varphi(p, r) = \varphi(p_*, r_*) = \varphi_B^*(p_*)$$

Additionally, when there exists a valid P_F^* such that the line g_M intersects the efficient frontier (i.e. along $\zeta(P_F)$) where we define g_M as:

$$g_M(P_F) \triangleq \left(\frac{C_{01} - C_{00}}{C_{01} - C_{11}}\right) - \left(\frac{C_{10} - C_{00}}{C_{01} - C_{11}}\right) P_F$$

then it corresponds to an optimizing pair (p_*, λ_*) of parameters for a minimax bayesian decision rule. Otherwise, we set λ_* arbitrarily and xchoose p_* according to:

$$p_* = \begin{cases} 0 & \zeta_{NP}(P_F) > g_M(P_F) \quad \forall\ P_F \in [0,1] \\ 1 & \zeta_{NP}(P_F) > g_M(P_F) \end{cases}$$

- The proof of this theorem uses the observation that nature can just choose either $0, 1$ for p depending on which is worst for our choice of decision rule, making our Bayes risk $\max[\varphi_B^0, \varphi_B^1(q)]$. So for an optimal p ends up having $\varphi_B^0 = \varphi_B^1$. The way you actually prove it is with inequalities though as there are some holes in this intuition.

- Additionally for certain experiments, it is possible that the best decision rule to minimize cost would be to always guess one hypothesis (set your $q = 1, 0$) and let nature choose the opposite true prior ($p = 1, 0$) because always getting the decision rule wrong, but having a choice of whether to pay C_{01} or C_{10}, is optimal.

Corollary 5.1: If $r_B(\cdot; p_*, \lambda_*)$ is a minimax decision rule, then:

$$p_* \in \arg_p \max \varphi_B^*(p)$$

Additionally we define $r_B(\cdot; p_*, \lambda_*)$ as an **equalizer rule** when $\varphi_B^0(p_*, \lambda_*) = \varphi_B^1(p_*, \lambda_*)$. If there exist p_*, λ_* belonging to an equalizer rule, then $r_B(\cdot; p_*, \lambda_*)$ is a minimax decision rule.

L6 - Bayesian Parameter Estimation

Suppose we have data y generated under some parameters x which we want to estimate.

We define the **Bayes risk criteria**:

$$\hat{x} = \arg_{f(\cdot)} \min E[C(x, f(y)]$$

We can write the objective function as:

$$E[C(x, f(y))] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} C(x, f(y)) p_{x,y}(x, y)\, dx\, dy = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} C(x, f(y)) p_{x|y}(x|y)\, dx \right) p_y(y)\, dy$$

minimizing this is equivalent to minimizing the interior term for each particular y:

$$\hat{x}(y) = \arg\min \int_{-\infty}^{\infty} C(x, a) p_{x|y} p(x|y)\, dx$$

- The MAE estimate (i.e. $C(a, \hat{a}) = |a - \hat{a}|$) is the median of posterior belief: $p_{x|y}(\cdot|y)$. In other words the MAE estimator is the point at which the CDF equals $1/2$.
- The MUC loss:
$$C(a, \hat{a}) = \begin{cases} a & |a - \hat{a}| > \epsilon \\ 0 & \text{otherwise} \end{cases}$$
is the mode of posterior belief (i.e. the MAP estimate) as $\epsilon \to 0$.
- The MSE loss (Bayes Least Squares) is the mean of posterior belief: $\boxed{\hat{X}_{BLS} = E[x|y = y]}$

We define the **error** of an estimator $e(x, y) = \hat{x}(y) - x$ and the **global bias** b as constant vector:

$$b = E[e(x, y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\hat{x}(y) - x) p_{x,y}(x, y)\, dx\, dy$$

It will be useful to write $e(x, y) = b + (e(x, y) - b)$

We can write the error correlation matrix as $E[ee^T] = \Lambda_e + bb^T$ where $\Lambda_e = E[(e - b)(e - b)^T]$

Claim 6.4: The BLS estimate is unbiased (i.e. $b_{BLS} = 0$).

Claim 6.5 The error covariance matrix of the BLS is the expected covariance of posterior belief:

$$\Lambda_{e,BLS} \triangleq \Lambda_{BLS} = E[\Lambda_{x|y}(y)] = E[E[(x - E[x|y])(x - E[x|y])^T|y]]$$

Theorem 6.1: An estimator $\hat{x}(\cdot)$ is the BLS estimator iff estimation error e is orthogonal to any vector-valued function g of the data (i.e. $E[(\hat{x}(y) - x)g(y)^T] = 0$)

Theorem 6.2: The error covariance Λ_e of an arbitrary estimator $\hat{x}$ satisfies $\Lambda_{BLS} \preceq \Lambda_e$ with equality iff $\hat{x}(y) - E[\hat{x}(y) - x] = \hat{x}_{BLS}(y)$

Corollary 6.1: We can write $E[\Lambda_{x|y}(y)] \preceq \Lambda_x$ with equality iff $E[x|y] = E[x]$

L8 - NonBayesian Parameter Estimation

Define an estimator $f(\cdot)$ as **valid** if $\hat{x} = f(y)$ is in estimate of x based solely on y (and does not depend on x) - i.e. $p(\hat{x}|y, x) = p(\hat{x}|y)$.

Define the error and bias as:

$$e(y) = \hat{x}(y) - x \qquad b_{\hat{x}}(x) = E[e(y)] = E[\hat{x}(y) - x]$$

and the error covariance is $\Lambda_e = E_y[(e(y) - b_{\hat{x}}(x))(e(y) - b_{\hat{x}}(x))^T]$.

Definition 8.2: An estimator $\hat{x}$ for a *nonrandom* parameter x is **unbiased** if $b_{\hat{x}}(x) = 0 \ \forall \ x$

We sometimes want the MVUE (minimum variance unbiased estimator) - the **admissible** (valid and unbiased) estimator with the smallest variance: $\hat{x}_{MVUE} = \arg_x \min \lambda_{\hat{x}}(x)$.

It is not guaranteed that the MVUE exists (there may be several estimators which are each minimum variance for only some values of x, or there may be no admissible estimators)

Theorem 8.1 *(Cramer Rao Bound for scalars)*. Provided $p_y(y; x)$ satisfies the regularity condition

$$E\left[\frac{\partial}{\partial x} \ln p_y(y; x)\right] = 0 \qquad \forall x$$

then for any admissible $\hat{x}$ we have $\lambda_{\hat{x}}(x) \geq \frac{1}{J_y(x)}$ with

$$J_y(x) \triangleq E[S(y; x)^2] \qquad S(y; x) = \frac{\partial}{\partial x} \ln p_y(y; x)$$

where $J_y(x)$ is the **Fisher information** in y about x and $S(y; x)$ is the score function for x based on y.

- The regularity condition enforces that the orders of integration and differentiation can be exchanged: $\int \frac{\partial}{\partial x} p_y(y; x) \ dy = \frac{\partial}{\partial x} \int p_y(y; x) \ dy$
- The proof bounds the correlation of $e(\hat{x}, y)$ with artificial variable $f(y) = S(y; x)$ to achieve the final result, using the limit that $\rho \leq 1$.

Corollary 8.1: The Fisher information can be equivalently expressed as

$$J_y(x) = -E\left[\frac{\partial^2}{\partial x^2} \ln p_y(y; x)\right]$$

Definition 8.4 An unbiased estimator is **efficient** if it satisfies the Cramer-Rao bound with equality.

Corollary 8.2 An estimator $\hat{x}$ is efficient iff it can be expressed as:

$$\hat{x}(y) = x + \frac{1}{J_y(x)} \frac{\partial}{\partial x} \ln p_y(y; x)$$

where the RHS must be independent of x for the estimator to be valid.

Claim 8.1 When an efficient estimator $\hat{x}_{eff}$ exists, it is the MLE (i.e. $\hat{x}_{eff} = \hat{x}_{MLE}$).

Theorem 8.2 *(Cramer Rao bound for vectors).* Provided $p_y(y;x)$ satisfies the regularity condition

$$E\left[\frac{\partial}{\partial x}\ln p_y(y;x)\right] = 0 \qquad \forall x$$

then the covariance matrix $\Lambda_{\hat{x}}(x)$ of any unbiased estimator satisfies:

$$\Lambda_{\hat{x}}(x) \geq J_y^{-1}(x)$$

in the positive semidefiniteness sense, where $J_y(x)$ is the Fisher Information matrix:

$$J_y(x) \triangleq E[S(y;x)^T S(y;x)] \qquad S(y;x) \triangleq \frac{\partial}{\partial x}\ln p_y(y;x)$$

- Additionally, we can write the Fisher matrix as the expectation of the Hessian:

$$J_y(x) = -E\left[\frac{\partial^2}{\partial x^2}\ln p_y(y;x)\right]$$

Corollary 8.3 An unbiased efficient estimate $\hat{x}(y)$ exists iff

$$\hat{x}(y) = x + J_y^{-1}(x)\left[\frac{\partial}{\partial x}\ln p_Y(y;x)\right]$$

is a valid estimator (i.e. the RHS can be rewritten such that it does not depend on x, and thus is equal for any choice of x).

Corollary 8.4 If an efficient unbiased estimate exists, it is the maximum likelihood estimate.

Therew as some discussion of regularization being equivalent to priors.

Theorem 8.3 *(Gauss Markov Theorem).* Suppose data y depends on parameters x through the model:

$$y = Hx + w \qquad w \sim N(0, \Lambda_w)$$

where Λ_w and H are full rank. Then the maximum likelihood estimator is

$$\hat{x}_{ML}(y) = (H^T\Lambda_w^{-1}H)^{-1}H^T\Lambda_w^{-1}y$$

is the solution to the weighted least squares formulation, and it is the MVUE.

- When $\Lambda_w \propto I$ this becomes the ordinary least squares problem.

L9 - Exponential Families

Definition 9.1 A parameterized family of distributions over alphabet $\mathcal{Y}$ is a one-parameter exponential family if it can be expressed as:

$$p_y(y;x) = \exp\{\lambda(x)t(y) - \alpha(x) + \beta(y)\} \qquad \forall\, x \in X\ y \in Y$$

Note that $\alpha(x)$ is a normalizing constant, and the PDF can be rewritten:

$$p_y(y;x) = \frac{1}{Z(x)}\exp\{\lambda(x)t(y) + \beta(y)\} \qquad Z(x) = \int_y \exp\{\lambda(x)t(y) + \beta(y)\}$$

$Z(x)$ and $\alpha(x)$ are sometimes called the **partition** and **log-partition** function in statistical physics.

The notation $y \sim \mathcal{E}(\mathcal{X}, \mathcal{Y}, \lambda(\cdot), t(\cdot), \beta(\cdot))$ indicates that y is exponentially distributed.

Note: adding or subtracting constants from $t(\cdot)$ and $\beta(\cdot)$ does not change the distribution - it simply allows some terms to be moved around into the normalization constant:

$$\ln p_y = \lambda(x)(t(y) - c_1) - \alpha(x) + (\beta(y) - c_2) = \lambda(x)t(y) - \tilde{\alpha}(x) + \beta(y)$$

Additionally, if we choose $c_1 = 0$ and c_2 such that $q(y) \triangleq e^{\beta(y)-c_2}$ is normalized, then we can rewrite the exponential family as

$$p_y(y;x) \propto q(y)\exp\{\lambda(x)t(y)\}$$

Where q is referred to as the **base-distribution** of the family.

We can try to restrict analysis to exponential families where their support (output space) does not depend on x. These are called **regular**.

We have two properties

$$\frac{d}{dx}\alpha(x) = \left(\frac{d}{dx}\lambda(x)\right)E[t(y)] \qquad \frac{d^2}{dx^2}\alpha(x) = \left(\frac{d^2}{dx^2}\lambda(x)\right)E[t(y)] + \left(\frac{d}{dx}\lambda(x)\right)^2 \mathrm{Var}[t(y)]$$

The Fisher information in y about x is $J_y(x) = \ddot{\lambda}(x)\frac{d}{dx}E[t(y)]$

A **canonical** exponential family, is one with $\lambda(x) = x$.

- For canonical families $\dot{\alpha}(x) = E[t(y)]$ and $\ddot{\alpha}(x) = \mathrm{Var}[t(y)] = J_y(x)$

Firstly, any weighted geometric mean of two probability distributions can be written as a (within the equivalence class) unique canonical exponential family (and vice-versa: $p_2(y) = c_1 e^{\beta(y)}$, $p_1(y) = c_2 p_2(y) e^{t(y)}$):

$$p_y(y;x) = \frac{p_1(y)^x p_2(y)^{(1-x)}}{Z(x)} \qquad \ln p_y(y;x) = x \ln \frac{p_1(y)}{p_2(y)} + \ln p_2(y) - \ln Z(x)$$

Theorem 9.1: P is a one-dimensional exponential family iff for any $p_1, p_2, p_3 \in P$, there exists some λ for which p_3 is a weighted geometric mean of p_1, p_2.

There are also tilting distributions, which are scaled versions of any base distribution q expressed as an canonical(?) exponential family. These weren't very interesting, so I haven't included anything.

The ideas of exponential families generalize easily to multi-parameter families.

$$p_y(y; x) = \exp\{\lambda(x)^T t(y) - \alpha(x) + \beta(y)\}$$

And they can be conveniently constructed as such to handle finite alphabets (output space sets/range/image etc.), and under suitable conditions can "replicate" well behaved distributions.

Lastly, note that when analyzing given data, the only thing we really need to maintain for inference about the distribution are $t(y), \beta(y)$ which we can just compute, and then throw away the data. For many applications such as ML estimation we don't even need $\beta(y)$, making $t(y)$ kind of a step in the direction of *sufficient statistics*, the topic of the next lecture.

L10 - Sufficient Statistics

A **statistic** is a deterministic function of the data y, i.e. $t(y)$ - this is itself a random variable due to its dependence on y. Many values of y may map to the same value of t.

Def 10.1: a statistic $t(\cdot)$ is **sufficient** wrt $p_y(\cdot; x)$ if $p_{y|t}(\cdot|\cdot; x)$ does not depend on $x \in X$: $p_{y|t}(; x_1) = p_{y|t}(; x_2)$.

- Essentially the uncertainty in y is not a function of x.
- An equivalent notion of sufficiency - the likelihoods are scalings (which may depend on y): $p_y(y; x) \propto p_t(t(y); x)$

Thm 10.1: A statistic $t(\cdot)$ is sufficient with respect to a model family iff $\frac{L_y(x)}{L_t(x)} = \frac{p_y(y;x)}{p_t(t(y);x)}$ is not a function of x for every y.

Thm 10.2 (Neyman Factorization Theorem): A statistic $t(\cdot)$ is sufficient wrt $p_y(\cdot; x)$ iff there exist functions a, b such that $p_y(y; x) = a(t(y), x)b(y)$.

Def 10.2: A sufficient statistic s is **minimal** if for any other sufficient statistic t, there exists a function $g(\cdot)$ such that $s = g(t)$.

- The minimal sufficient statistic is not unique - any 1-1 function applied will also work.

Def 10.3: A sufficient statistic t is complete if for any function $\phi(\cdot)$ with $E[\phi(t(y))] = 0 \quad \forall x$ must satisfy $P(\phi(t(y)) = 0) = 1$.

- In the notes (L10 and L9.11c!) we show that for exponential families, their natural statistic $t(y)$ is complete.

Thm 10.3: A sufficient statistic t is minimal if it is complete

- Some minimal statistics may not be complete.

We can also choose to model the unknown parameter in a Bayesian framework

Def 10.4: A statistic $t(\cdot)$ is sufficient wrt $p_{x,y}$ iff $p_{y|t,x}(y|t(y), x) = p(y|t(y))$

Thm 10.4: A statistic $t(\cdot)$ is sufficient wrt $p_{x,y}$ iff $p_{x|y}(\cdot|y) = p_{x|t}(\cdot|t(y))$

Thm 10.5: A statistic $t = t(\cdot)$ is sufficient iff $p_{y|x}(y|x) = p_{t|x}(t(y)|x)p_{y|t}(y|t(y)) \quad \forall x, y$

Def 10.5: A variable t is a statistic if it satisfies the Markov chain $x \leftrightarrow y \leftrightarrow t$

Note: Hereafter we use $\propto$ defined as a multiplicative factor that may be non-constant

- Additionally, we define likelihoods as: $L_{y_1}(x) = p_y(y_1; x)$

Thm 10.6 (partition characterization): A statistic t is sufficient iff for all y_1, y_2 such that $t(y_1) = t(y_2)$ we have $L_{y_1}(x) \propto L_{y_2}(x)$ (i.e. there exists g s.t. $L_{y_2}(x) = g(y_1, y_2)L_{y_2}(x)$)

Thm 10.7 (minimal, partition characterization): A sufficient statistic t is minimal iff for all y_1, y_2 s.t. $L_{y_1}(x) \propto L_{y_2}(x)$ we have $t(y_1) = t(y_2)$.

- I don't really understand this and should probably revisit it.

L11 - Inequalities

A function f is **convex** if $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$.

- **strict convexity** requires the inequality to be strict
- **concave** functions are the opposite (obviously)

Thm (Cauchy Schwarz): For any two lists of reals a_i, b_i we have:

$$\left(\sum_{i=1}^{N} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{N} a_i^2\right)\left(\sum_{i=1}^{N} b_i^2\right)$$

Thm 11.1 (Jensen's Inequality): If $\phi(\cdot)$ is a concave function and v is a random variable defined over V then

$$E[\phi(v)] \leq \phi(E[v])$$

If $\phi(\cdot)$ is strictly concave, then the above holds with equality iff v is a deterministic constant.

Thm 11.2 (Csiszar's Inequality): Given positive finite length sequences a_i, b_i and a strictly convex function $f(\cdot)$ we have

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

with equality iff a_i/b_i is a constant

Corollary 11.1 (Log-Sum Inequality) Given positive finite length sequences a_i, b_i:

$$\sum_{i=1}^{N} a_i \log\frac{a_i}{b_i} \geq \left(\sum_{i=1}^{N} a_i\right) \log\frac{\sum_{i=1}^{N} a_i}{\sum_{i=1}^{N} b_i}$$

with equality iff a_i/b_i is a constant

- Simply choose $f(x) = x\log(x)$ in Czisar's inequality.

Thm 11.3 (Gibb's Inequality): if v is a random variable distributed according to $p(\cdot)$, then for any distribution $q(\cdot)$:

$$E_p[\log q(v)] \leq E_p[\log p(v)]$$

with equality iff $q \cong p$.

L12 - The EM algorithm

Given data y generated from $p_y(\cdot; x)$ we wish to compute the ML estimate for $\hat{x}(y) = \arg\max_a \ell_y(a; y)$ where $\ell_y(a; y) \triangleq \log p_y(y; a)$.

We also define a variable z for which $y = g(z)$ for some deterministic g. Essentially z is an arbitrary (possibly fictitious) quantity which can use for other things [what other things?]

We define $\ell_z(a; z) \triangleq \log p_z(z; a)$. We can then write the expectation of the likelihood given a choice of parameter x:

$$\hat{\ell'}_z(x; y) \triangleq E_{p_{z|y}(\cdot|y;x')}[\ell_z(x; z)]$$

Additionally, since $p_z(z; x) = p_{z,y}(z, y; x) = p_{z|y}(z|y; x)p_y(y; x)$, if we take the log and expectation (wrt $p_{z|y}(\cdot|y; x')$ parameterized by arbitrary x') of each side we get:

$$\begin{aligned}
\underbrace{\log p_y(y; x)}_{\ell_Y(x;y)} &= \underbrace{E_{p_{z|y}(\cdot|y;x'))}[\log p_z(z; x)]}_{\triangleq U(x;x')} \underbrace{- E_{p_{z|y}(\cdot|y;x'))}[\log p_{z|y}(z|y; x)]}_{\triangleq V(x;x')} \\
&= U(x; x') + V(x; x') \\
&\geq U(x; x') + V(x'; x') \\
&= [U(x; x') - U(x'; x')] + U(x'; x') + V(x'; x') \\
&= \underbrace{[U(x, x') - U(x', x']}_{\triangleq \Delta(x', x)} + \log p_y(y; x')
\end{aligned}$$

Where in the inequality step we have used Gibbs' inequality. Essentially this means the log-likelihood will increase at every step if at every step we find a new x such that $\Delta(x', x) > 0$.

We have the following iterate EM procedure for approximating ML estimates $\hat{x}(y)$:

1. Initialize/guess a starting estimate for $\hat{x}^{(0)}$
2. **E-step**: Compute $U(x; \hat{x}^{(l)} = E_{p_{z|y}(\cdot|y;\hat{x}^{(l)})}[\log p_z(z; x)]$
3. **M-step**: Solve the maximization $\hat{x}^{(l+1)} = \arg\max_{x \in X} U(x; \hat{x}^{(l)})$
    - We can use the stationary point equations: $\frac{\partial}{\partial x_k} U(x, x') = 0$ for each k.
4. Increment l and go back to step 2. Repeat until convergence

The notes then cover a general application of EM algorithms to an exponential family, and analysis/proof of its convergence to stationary points of $\ell_y(\cdot; y)$.

Then there was an example of the EM algorithm used in a logistic modeling problem. It's worth reviewing to see the actual equations being used.

L13 - Inference as Decision Making

Again, we use the observation model $p_{y|x}(\cdot|\cdot)$ with a bayesian prior $p_x(\cdot)$.

As a new perspective on inference, consider a decision device which returns a distribution $q(\cdot)$ describing the relative likelihood of different x based on observed y.

We introduce a cost criterion - a function like $C(x,q) = A\sum_{a\in X}(q(a) - 1_{a=x})^2 + B(x)$ or $C(x,q) = -A\log q(x) + B(x)$

Def 13.1 A cost function $C(\cdot,\cdot)$ is **proper** if

$$p_{x|y}(\cdot|y) = \arg\min_q E_{p(x|y)}[C(x,q)|y=y] \quad \forall y \in Y$$

- The optimal solution to a proper cost function returns the true $p_{x|y}$

Claim 13.1: The log-loss cost criterion is proper

Def 13.2: A cost function C is local if there exists a function ϕ such that $C(x,q) = \phi(x, q(x))$

- The cost function only considers the estimated belief of x and not $x \pm \epsilon$!

Claim 13.2: The log-loss cost criterion is local

Thm 13.1: When the alphabet X costs of at least 3 values, then the log-loss is the only smooth, local, proper cost function.

We proceed by studying the **minimum** expected cost of the log-loss criterion before we have any observations. Obviously, this happens when $q(\cdot) = p_x(\cdot)$. Thus we define **entropy** $H(x)$:

$$H(x) = -E_{p_x}[\log p_x(x)] = -\sum_a p_x(a)\log p_x(a)$$

- We adopt the convention $0\log 0 = 0$

Claim 13.3: For discrete rv $x \in X$ we have $0 \le H(x) \le \log|X|$ where the lower bound requires x to be constant and the upper bound requires x uniformly distributed.

We can define

- **conditional entropy given** $y=y$: $H(x|y=y) \triangleq -E[\log p_{x|y}|y=y] = -\sum_a p_{x|y}(a|y)\log p(a|y)$
- **conditional entropy**: $H(x|y) \triangleq E_y[H(x|y=y)] = -\sum_{a,b} p_{x,y}(a,b)\log p_{x|y}(a|b)$
- We can show $0 \le H(x|y) \le H(x)$
- **chain rule of entropy**: we can show: $H(x,y) = H(x|y) + H(y)$

The **mutual information** is the cost reduction from making an average observation:

$$I(x;y) \triangleq H(x) - H(x|y) = \sum_{a,b} p_{x,y}(a,b)\log\frac{p_{x,y}(a,b)}{p_x(a)p_y(b)}$$

- **conditional MI**: $I(x;y|z) \triangleq H(x|z) - H(x|y,z) \ge 0$ with equality iff $x \leftrightarrow y \leftrightarrow z$
- nonnegativity: $I(x;y) \ge 0$

- $I(xy) = 0$ iff $x \perp y$
- symmetry: $I(x;y) = I(y;x)$
- chain rule: $I(x;y,z) = I(x;z) + I(x;y|z)$ where $I(x;y|z) \triangleq H(x|z) - H(x|y,z)$

Thm 13.2 (Data Processing Inequality): If $x \leftrightarrow y \leftrightarrow t$ is a Markov chain (i.e. if t is any statistic), then

$$I(x;y) \geq I(x;t)$$

with equality iff t is a sufficient statistic (i.e. $x \leftrightarrow t \leftrightarrow y$)

Corollary 13.1 A statistic t is sufficient iff $I(x;t) = I(x;y)$

Corollary 13.2 For any deterministic $g(\cdot)$, $I(x;y) \geq I(x;g(y))$.

We define **KL-divergence** $D(p||q) \triangleq E_p[\log \frac{p(x)}{q(x)}]$

- By Gibbs' Inequality: $D(p||q) \geq 0$ with equality iff $p = q$
- identity: $D(p||U(X)) = \log|X| - H(p)$
- identity: $I(x;y) = D(p_{x,y}||p_x p_y)$ and $I(x;y|z=z) = D(P_{x,y|z}(|z)||p_{x|z}(|z)p_{y|z}(|z)$
  $\rightarrow$ we can further write $I(x;y|z) = \sum_z p_z(z) I(x;y|z=z)$

This setup can be extended to posterior beliefs upon having observed y, however there are some intricacies, explained in the notes.

Claim 13.4 (connection to Fisher Info): Suppose the family of distributions $p_y(\cdot;x)$ is positive, thrice differentiable for each y and satisfies the regularity conditions:

$$E\left[\frac{\partial}{\partial x}\ln p_y(y;x)\right] = 0 \qquad \left|E\left[\frac{\partial^3}{\partial x^3}\ln p_y(y;x)\right]\right| < \infty$$

Then for all $x \in X$ we can write $D(p_y(;x)||p_y(;x+\delta)) = \frac{\log(e)}{2} J_y(x)\delta^2 + O(\delta^2)$ as $\delta \to 0$

L14 - Information Geometry

We will develop a notion of geometry between probability distributions starting from the KL-divergence $D(p||q) = E_p[\frac{\log p(x)}{q(x)}]$ as a distance metric.

For a finite alphabet $Y = \{1, \cdots, M\}$, we have $\sum p_y(i) = 1$, so each allotment of probability density (each possible distribution) is a point on the M-dimensional simplex.

- a distribution is on the boundary of the simplex if $q(y) = 0$ for some y
- if p, q are both on the interior of the simplex, $D(p||q) < \infty$ while if q is on the boundary, then $D(p||q) = \infty$
  - $\rightarrow$ However if p is on the boundary and q is not, then $D(p||q) < \infty$ again.

Def: Given a nonconstant function $t(y) \rightarrow \mathbb{R}$, the set of distributions in P with $E_p[t(y)] = \sum_y p(y)t(y) = c$ is known as a **linear family** with parameter c.

- This is a convex set

Def 14.1 (I-projection): The information projection of q onto a (nonempty + closed) set of distributions P is $p^* = \arg\min_p D(p||q)$

Thm 14.1 (Pythagoras' Theorem): Let p^* be the I-projection of q onto a (nonempty + closed + convex) set P. Then

$$D(p||q) \geq D(p||p^*) + D(p^*||q) \quad \forall p \in P$$

Cor 14.1 (Pythagoras' Corollary): The I-projection p^* of any q not on the boundary of P^Y onto a linear family P cannot lie on a boundary component of P^Y unless all of P is confined to that particular boundary component

Def 14.2 (Linear Family): A set of M-dimensional distributions $P \subset P^Y$ is a **linear family** if for some $K < M$ there exists functions $\mathbf{t} = [t_1(\cdot), \cdots, t_K(\cdot)]^T$ and constants $\bar{\mathbf{t}} = [\bar{t}_1, \cdots, \bar{t}_K]$ such that $\mathbf{t}[p(1) \cdots p(M)]^T = \bar{\mathbf{t}}$.

Claim 14.1: For every p_1, p_2 from a linear family $\mathcal{L} \subset P^y$, then for every $\lambda \in \mathbb{R}$ (including $\lambda > 1$) $\lambda p_1 + (1 - \lambda)p_2 \in P^y$.

Cor 14.2 (Pythagorean Identity): For any q in the simplex of P^Y and some $\mathcal{L}$ defined over Y, the I-projection p^* of q onto $\mathcal{L}$ satisfies

$$D(p||q) = D(p||p^*) + D(p^*||q) \quad \forall p \in \mathcal{L}$$

- This is stronger than Thm 14.1 - it gives us equality!

Define $\mathcal{L}_{\mathbf{t}}(p^*)$ to be the linear family with given $\mathbf{t}$ which contains p^*. What is the set of all distributions in the simplex whose I-projection onto $\mathcal{L}_{\mathbf{t}}(p^*)$ is p^*?

Thm 14.2: p^* is the I-projection of q onto $\mathcal{L}_{\mathbf{t}}(p^*)$ iff q is in the exponential family $\mathcal{E}_t(p^*) = \{q(y) = p^*(y) \exp\{\mathbf{x}^T\mathbf{t}(y) - \alpha(\mathbf{x})\} \ \forall y \in Y\}$

- We can use this to find the I-projection by finding the member satisfying $\mathbf{t}, \bar{\mathbf{t}}$ where $\alpha(x)$ is the log-partition function normalizing over $\sum_y p(y; x) = 1$

L15 - Modeling as Inference

Usually we try to estimate a parameter x of the distribution (estimation), but we can also try to model the distribution itself by finding a sufficient approximation $q(\cdot)$ (modeling)

Thm 15.1 Let $\{p_y(;,x)\}$ be a class of models, and let $q \in P^Y$ be an **admissible** distribution (does not depend on the parameter x). Then exist weights w for a mixture model $q_w(\cdot) = \sum_x w(x)p_y(\cdot;x)$ such that

$$D(p_y(x)||q_w) \leq D(p_y(x)||q) \quad \forall x \in X$$

We use the minimax framework: $R^+ \triangleq \min_q\{\max_x D(p_y(x)||q)\}$

Lemma 15.1: For any $q \in P^Y$ we have $\max_x D(p_y(x)||q) = \max_w \sum_x w(x)D(p_y(x)||q)$

- This can be used to rewrite $R^+ = \min_q\{\max_w \sum w(x)D(p_y(x)||q)\}$

Thm 15.2 (Redundancy-Capacity Theorem): Optimizing q and w on both sides of the the minmax / maxmin are the same:

$$R^+ = \min_q \max_w \sum w(x)D(p_y(x)||q) = \max_w \min_q \sum w(x)D(p_y(x)||q) = R^-$$

If we visualize R^- above from a Bayesian framework where $w(x)$ is a prior p_x, then through some simple algebra and tricks we can show:

$$R^- = \max_w \sum_x \sum_y w(x)p_y(y;x)\log\frac{p_y(y;x)}{q_w(y)}$$

which can eventually be rearranged to $\max I(x;y)$.

Def 15.1 Let $p_{y|x}$ be a model. The least informative prior p_x^* for the model is given by $p_x^* = \arg\max_{p_x} I(x;y)$

Def 15.2 The **model capacity** C of the model $p_{y|x}$ is the average cost reduction associated with the least informative prior: $C = \max_{p_x} I(x;y)$.

We also have $0 \leq C \leq \log|X|$

Thm 15.3 (Equidistance property): The optimum mixture model q^* with optimum weights w^* is such that

$$D(p_y(x)||q^*) \leq C \quad \forall x \in X$$

with equality for all x s.t. $w^*(x) > 0$.

- This is a very neat property with a kind of complicated proof.

L16 - Information Measures for Continuous Variables

We define **differential entropy**:

$$h(x) \triangleq -\int_{-\infty}^{\infty} p_x(x) \log p_x(x)\ dx$$

- not invariant to coordinate transformations, unlike in the continuous case

We can also write

$$h(x|y=y) = -\int_{-\infty}^{\infty} p_{x|y}(x|y) \log p_{x|y}(x|y)\ dx$$
$$h(x|y) \triangleq \int_{-\infty}^{\infty} p_y(y) h(x|y=y)\ dy$$

Both differential mutual information and information divergence are coordinate-transformation invariant.

Differential mutual information is similar to the discrete case: $I(x;y) = h(x) - h(x|y)$.

Similarly, **information divergence** is $D(p||q) \triangleq \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\ dx$.

- We can write $I(x;y) = D(p_{x,y}||p_x p_y) \geq 0$. Because of this we also know $h(x|y) \leq h(x)$ - conditioning can never increase differential entropy!.

There was a very brief section on modeling with continuous alphabets. The mixture model is now $q_w(y) = \int_x w(x) p_y(y;x)\ dx$, and the model capacity is still $C = \max_{p_x} I(x;y)$

There were also a bunch of results for differential/conditional differential entropy for Gaussian distributions. This was pretty long and straightforward, so just look at the notes if necessary.

L17 - Max Entropy Distributions

The least informative prior we previously discussed in L15 is optimal but challenging in practice, but we can get decent performance using a maximally-ignorant prior. Essentially given some linear constraints $E_p[t_k(y)] = \bar{t}_k$ for K constraints we want to find the distribution with maximum entropy.

In the finite alphabet case, we can find this by minimizing $D(p||U) = \log|Y| - H(p)$ where U is the uniform distribution (and by construction the maximum entropy prior possible over finite alphabets).

Note that $U = q(y) = e^{\beta(y)}$ where $\beta(y) = 0$. By the results of the previous section, we simply find the I-projection of U onto the corresponding linear family:

$$p^*(y) = \exp\left\{\sum_K x_i t_i(y) - \alpha(\mathbf{x})\right\}$$

where x_i are chosen to satisfy the constraints.

For infinite alphabets with both discrete and continuous distributions, the result is similar

Claim 17.1: Among all distributiosn over $Y \subset \mathbb{R}$ in the linear family with finite differential entropy, the following distribution p^*, when it exists is the unique distribution having maximum differential entropy.

$$p^*(y) = \exp\left\{\sum_K x_i t_i(y) - \alpha(\mathbf{x})\right\}$$

- This is the same form as before!
- **Claim 17.2**: This result (and the form of p^*) is still applicable for discrete distributions over infinite alphabets.

We can essentially use this to show that the normal distribution is the maximum entropy distribution of given mean and variance!

L18 - Conjugate Priors

Definition 18.1: For observations of the form $\mathbf{y} = [y_1, \cdots, y_n]^T$ over Y^n and X, then $p_{\mathbf{y}|x}$ is **conditionally i.i.d.** if for every $x \in X, \mathbf{y} \in Y^N$ we have:

$$p_{\mathbf{y}|x}(y_1, \cdots, y_n|x) = \prod_{n=1}^{N} p_{y|x}(y_n|x)$$

- conditional i.i.d. models have permutation invariance, and their sufficient statistics are especially compact

Definition 18.2: A sequence $y_1, \cdots, y_n$ is exchangeable if for every permutation $n_1, \cdots, n_N$ and all $y_1, \cdots, y_n \in Y^N$ we have:

$$p_{y_1,\cdots,y_n}(y_1, \cdots, y_n) = p_{y_{n_1},\cdots,y_{n_N}}(y_1, \cdots, y_n)$$

- **infinitely exchangeable**: exchangeable for every finite N.

Theorem 18.1 (de Finetti): The sequence $y_1, y_2, \cdots$ is infinitely exchangeable iff for every N there exists an alphabet X, distribution p_x, and model $p_{y_1,\cdots,y_N|x} = \prod p_{y|x}(y_i|x)$ such that:

$$p_{y_1,\cdots,y_N}(y_1, \cdots, y_N) = \int p_{y_{n_1},\cdots,y_{n_N}}(y_1, \cdots, y_N|x) p_x(x)\ dx$$

- Think of every infinitely exchangeable sequence as a mixture of i.i.d. distributions

---

We want to keep updating $p_{x|y}$ as we add in more data points y_i to our sequence. Note that:

$$p_{x|\mathbf{y}}(\cdot|y) = T_{\mathbf{y}}[p_x(\cdot)] \triangleq \frac{p_{x,y}(\cdot|\mathbf{y})}{p_{\mathbf{y}}(\mathbf{y})} = \frac{p_{\mathbf{y}|x}(\mathbf{y}|\cdot)p_x(\cdot)}{\int p_{\mathbf{y}|x}(\mathbf{y}|a)p_x(a)\ da}$$

so if we update $p_{x|y_1} \to p_{x|y_2,y_1}$

$$p_{x|y_2,y_1} = T_{y_2|y_1}[p_{x|y_1}] \triangleq \frac{p_{y_2|y_1,x}p_{x|y_1}(x|y_1)}{\int p_{y_2|y_1,x}(y_2|y_1, a)p_{x|y_1}(a|y_1)\ da}$$

We are interested in finding families of priors for which our updated belief after each successive observation y_i is in the same family. We restrict our attention to the case of conditionally i.i.d. models: $p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|x) = \prod p_{y|x}(y_i|x)$.

Def 18.3 A family $Q = \{q(\cdot;\theta) : \theta : \Theta \subset)\}$, where the mapping from Θ to Q is continuous and continuously invertible, is a **conjugate prior family** for the conditional i.i.d. model if for every $y \in Y$ we have $p_{x|y}(\cdot|y) \in Q$ whenever $p_x(\cdot) \in Q$.

- If θ_0 is the parameter specifying p_x, then $\theta(y;\theta_0)$ is the parameter for $p_{x|y}(\cdot|y)$

Claim 18.1: A collection Q of the above form is a conjugate prior family iff $p_{x|y_1,\cdots,y_n}(\cdot|y_1, \cdots, y_n) \in Q$ whenever $p_x(\cdot) \in Q$ for every $N \geq 1$.

Def 18.4: A conjugate prior family is **natural** if for every $y \in Y$ there exists a paramenter value $\theta(y) \in \theta$ s.t. $q(\cdot;\theta(y)) \propto p_{y|x}(y|\cdot)$ with a constant of proportionality that is typically a function of y.

Theorem 18.2: For a conditional i.i.d. model such that $Y \subset \mathbb{R}$ is a region and $p_{y|x}$ is continuous, if a conjugate prior family exists, then for every $N \geq 1$ there exists a continuous function $t_N(\cdot, \ldots, \cdot)$ with finite dimension which is a sufficient statistic for inferences about x.

In one of the optional sections, it is shown that conjugate priors only exist for $p_{y|x}$ which are from exponential families. In the last section, they show how to construct a conjugate prior family from a given exponential family, and essentially that all exponential families have a conjugate prior family.

If $p_{y|x}(y|x) = \exp\{\lambda(x)^T t(y) - \alpha(x) + \beta(y)\}$, then the corresponding conjugate prior family is:

$$Q = \{q(\cdot; \mathbf{t}, N) : q(x; \mathbf{t}, N) = \exp\{[t^T \lambda(x) - N\alpha(x)] - \gamma(t, N)\}\}$$

where $\mathbf{t}$ is an updating function of the $t(y)$ we see, starting from our prior's value of t_0. The corresponding prior is $p_x(x) = q(x; t_0, N_0)$.

Correspondingly, Q is itself a canonical exponential family with natural statistic $[\lambda(x), -\alpha(x)]$ and parameter (known as x in the definition earlier in the notes) $[\mathbf{t}, N]$

L19 - Information Geometry of MLE and EM

We begin by showing that maximum likelihood estimation is simply finding a parametrized distribution with minimum KL divergence to the "empirical distribution".

Fact 19.1: For any deterministic sequence $v = \{v_1, \cdots, v_n\}$ the empirical distribution is $\hat{p}(v; \mathbf{v}) = \frac{1}{N}\sum_N 1_{v=v_n}$ and we can apply functions such as $\frac{1}{N}\sum_N f(v_n) = \sum_v f(v)\hat{p}(v; \mathbf{v})$.

Fact 19.2: The ML estimate of x over parameterized models $p_y(\cdot; x)$ is $\hat{x}_{ML}(\mathbf{y}) = \arg\min_a D(\hat{p}_y(\cdot; y)||p_y(\cdot; a))$

- This is known as the **reverse I-projection**, or also the **M-projection**
- The proof follows from $\tilde{I}(x; y) = \sum_{b\in Y} \hat{p}_y(b; y) \log p_y(bx)$

We can show similar results on the EM algorithm on i.i.d. observations. In EM we want to find:

$$U(x, x') = \sum_{\mathbf{z}\in Z^N} p_{\mathbf{z}|\mathbf{y}}(\mathbf{z}|\mathbf{y}; x') \log p_{\mathbf{z}}(\mathbf{z}; x) = \sum_n \sum_{c\in Z} p_{z|y}(c|y_n; x') \log p_z(c; x)$$

where we have used the independence of $p_{z|y}(z_i|y_i; x')$, $p_z(z_j; x)$ for $i \neq j$ to simplify terms.

Now we construct a hypothetical empirical distribution for p_z drawn from:

$$\hat{P}^Z(\mathbf{y}) \triangleq \{ \sum_{c\in\{g(c)=b\}} \hat{p}_z(c) = \hat{p}_y(b; \mathbf{y}) \ \forall b \in Y\}$$

For $\hat{p}_z^* = \arg\min_{\hat{p}_z\in\hat{P}^Z(\mathbf{y})} D(\hat{p}_z(\cdot)||p_z(\cdot; x))$ we have the following property:

Claim 19.1: The empirical distribution $\hat{p}_z^*(\cdot; x)$ can be expressed as

$$\hat{p}_z^*(z; x) = \frac{p_z(z; x)\hat{p}_y(g(z))}{p_y(g(z); x)}$$

- This follows from a variant of the data processing inequality.
- **Lemma 19.1** (DPI - Decision Form): Let $g : Z \to Y$ be an arbitrary mapping. Then for arbitrary p_z, q_z and corresponding p_y, q_y produced by g, we have $D(p_z||q_z) \geq D(p_y||q_y)$ with equality iff $\frac{p_z(z)}{q_z(z)} = \frac{p_y(g(z))}{q_y(g(z))}$ (presumably for every z).

and by some algebra we can write:

$$\frac{1}{N}U(x, x') = -D(\hat{p}_z^*(z; x')||p_z(\cdot; x)) - H(\hat{p}_z^*(z; x'))$$

so $\arg\max_x U(x, x') = \arg\min_x D(\hat{p}_z^*(z; x')||p_z(\cdot; x))$

So the EM algorithm is equivalent to first finding $\hat{p}_z^*(z; x^{(l-1)}) = \arg\min_{\hat{p}_z\in\hat{P}^Z(\mathbf{y})} D(\hat{p}_z(\cdot)||p_z(\cdot; x))$ (the E-step), and then computing $\hat{x}^l = \arg\min_x D(\hat{p}_z^*(z; x')||p_z(\cdot; x))$ (the M-step).

- Essentially computing a reverse I-projection followed by an I-projection.

Lastly, they show that if we are computing the MLE over a linear family we must have that $\frac{1}{N}\sum_N t_k(y_i) = E_{p_y(\cdot|\hat{x}_{ML})}[t_k(y)]$ for each y.

L20 - Stochastic Approximation

Solving problems in high dimensional systems is pretty hard. Things like marginalizing out a variable $\sum_x p_{x,y}(x,y) = p_y(y)$ can be infeasible. In this lecture we develop sampling methods from complex distributions.

Consider the challenge of approximating sampling from an infeasible distribution p using a simpler distribution q. In fact, we only need access to samples from q and a non-normalized version $p_\circ(x)$ which satisfies $p(x) = \frac{p_\circ(x)}{Z_p}$ (where $Z_p = \sum_x p_\circ(x)$) for both p, q for the following approaches. (In fact we don't even need the exact form of q, beyond samples from it).

Importance Sampling: In order to estimate the mean of p_x we can use the following:

$$\hat{\mu}_f(x) = \sum_{n=1}^{N} \frac{w(x_n)}{\sum_{n'=1}^{N} w(x_{n'})} f(x_n)$$

where $w(x) = \frac{p_\circ(x)}{q_\circ(x)}$ is the **importance weights**

- The proof follows from $E_q[w(x)] = \frac{Z_p}{Z_q}$ and $E_q[w(x)f(x)] = \frac{Z_p}{Z_q}\mu_f$.
- Qualitatively, we want our q to be "close" to p, so we don't have too few samples in high density parts of p.

Rejection Sampling: Suppose there is a constant c such that $cq_\circ(x) > p_\circ(x)$ for all x. Then we can generate samples from q, and then generate $u \sim U([0, cq_\circ(x)])$ and discard if $u \leq p_\circ(x)$.

Def 20.1 (FSHMC): A sequence of variables x_i forms a **finite-state homogeneous Markov chain** if 1) x is finite, 2)for every n we have the Markov chain $x_1 \leftrightarrow \cdots \leftrightarrow x_n$, 3) $p_{x_{n+1}|x_n}$ does not depend on n.

Def 20.2: A distribution p_x is a **stationary distribution** of a FSHMC if $p_x(x) = \sum_{x'} w(x|x')p_x(x')$.

- These are called the **global balance** equations

Def 20.3: A FSHMC is **fully communicating** if it is possible to go from any state to any other state (including itself) within some finite number of time steps.

Thm 20.1: If a FSHMC is fully communicating it has a unique stationary distribution.

Def 20.4: A fully communicating FSHMC is aperiodic if for some x it can return to x at any desired time after some n_0: $p_{x_n|x_1}(x|x) > 0$ for all $n > n_0$.

Thm 20.2: If a fully communicating FSHMC with stationary distribution p_x is aperiodic then for any p_{x_1} we have $\lim_{n\to\infty} p_{x_n}(x) = p_x(x)\ \forall x \in X$.

Def 20.5: A fully communicating FSHMC with distribution p_x is reversible if when $p_{x_1} = p_x$ we have

$$p_{x_1,\cdots,x_n}(x_1,\cdots,x_n) = p_{x_1,\cdots,x_N}(x_n,\cdots,x_1)$$

Prop 20.1: A fully communicating FSHMC is reversible iff its stationary distribution satisfies:

$$p_x(x')w(x|x') = p_x(x)w(x'|x)$$

- These are called the **detailed balance** equations

Now we have the tools for MCMC approximations.

From Thm 20.2, we know for a fully communicating aperiodic FSHMC with stationary distribution p_x we have $\lim_{n\to\infty} D(p_x||p_{x_n}) = 0$.

Prop 20.2: For a finite-state homogenous Markov chain $x_1 \leftrightarrow x_2 \cdots$ with stationary distribution $p_x \in P^X$, for any $p_{x_1} \in P^X$ we have

$$D(p_x||p_{x_{n+1}} \leq D(p_x||p_n)$$

- This doesn't require the stationary distribution to be unique, or for p_{x_n} to converge to one.

Metropolis Hastings Algorithm: We begin with unnormalized $p_\circ$ and transition model $v(\cdot|\cdot)$

1. At time n suppose we have $x_n = x$.
2. Generate x' from $v(\cdot|x)$
3. Compute acceptance factor

$$a(x \to x') \triangleq \min\left\{1, \frac{p_\circ(x')v(x|x')}{p_\circ(x)v(x'|x)}\right\}$$

4. Sample u from bernoulli dist. with parameter $a(x \to x')$.
5. If $u = 1$, accept the transition to $x_{n+1} = x'$. Otherwise $x_{n+1} = x$. Then repeat.

Thm 20.3: Given finite X and our choice of $v(\cdot|\cdot)$ the sequence of samples generated by the MH algorithm corresponds to a sequence of variables that form a reversible Markov chain with transition distribution $w(x'|x) = v(x'|x)a(x \to x')$ for which the target p is the stationary distribution

- The proof simply shows that detailed balance is satisfied.

There are some requirements for the "proposal distribution" $v(\cdot|\cdot)$:

Corr 20.1: If the choice of $v(\cdot|\cdot)$ is such that the induced Markov chain $w(\cdot|\cdot)$ is fully communicating and aperiodic, then x_n will have a distribution approaching the target p as $n \to \infty$.

Corr 20.2: If the chain with transition distribution $v(\cdot|\cdot)$ is fully communicating and $a(x \to x') > 0$ for all distinct pairs x, x', the the induced MH chain is fully communicating.

- Essentially, $v(x'|x) = 0$ iff $v(x|x') = 0$

Corr 20.3: If we have the conditions of Corr 20.2, and $a(x \to x') < 1$ for at least one distinct pair x, x', then the induced MH chain is fully communicating and aperiodic.

L21 - Typical Sequences and Large Deviation

Thm 21.1 (WLLN): Let w_i be a set of i.i.d. random variables with mean μ and $E[|w_i|] < \infty$. Then for any $\epsilon > 0$:

$$\lim_{N\to\infty} P\left(\left|\frac{1}{N}\sum_N w_i - \mu\right| > \epsilon\right) = 0$$

Define the normalized log-likelihood: $L_p(y) = \frac{1}{N}\log p_y(y) = \frac{1}{N}\sum_N \log p(y_i)$.

- From the WLLN we have $\lim_{N\to\infty} P(|L_p(y) + H(p)| > \epsilon) = 0$

Def 21.1 (Typical Set): Let y_i be a sequence of N elements from $\mathcal{Y}$, with a small constant ϵ. The sequence is called ϵ-**typical** wrt p if

$$|L_p(y) + H(p)| \leq \epsilon$$

- $\mathcal{T}_\epsilon(p; N)$ is called the ϵ-typical set wrt p.
- The probability that a generated sequence is in the typical set is approximately 1.
- Since $L_p(y) \approx -H(p)$, we have $p_y(y) \approx 2^{-NH(p)}$, and so $|\mathcal{T}_\epsilon(p; N)| \approx 2^{NH(p)}$
  - $\rightarrow$ Whenever $H(p) < \log M$, only a (vanishingly) exponentially small fraction of possible sequences actually occur.

Define the **total probability of a set**(of sequences) A: $P\{A\} = \sum_{\mathbf{y}\in A} p^N(\mathbf{y})$

Thm 21.2 (Asymptotic Equipartition Property): Given $\mathcal{T}_\epsilon(p; N)$ with $\epsilon > 0$, we have

- $\lim_{N\to\infty} P\{\mathcal{T}_\epsilon(p; N)\} = 1$
- $2^{-N(H(p)+\epsilon)} \leq p^N(y) \leq 2^{-N(H(p)-\epsilon)}$
- $(1-\epsilon)2^{N(H(p)-\epsilon)} \leq |\mathcal{T}_\epsilon(p; N)| \leq 2^{N(H(p)+\epsilon)}$

We can extend these results to continuous distributions, replacing cardinality with "volume".

If we are generating data from p and have some reference q, we can construct a new typical set definition.

Using $L_{p|q}(y) = \frac{1}{N}\log\frac{p^N(y)}{q^N(y)} = \frac{1}{N}\sum_N \log\frac{p(y_i)}{q(y_i)}$ with WLLN we have:

$$\lim_{N\to\infty} P(|L_{p|q}(y) - D(p||q)| > \epsilon) = 0$$

Def 21.2 (Typical set, another): The sequence y is called **divergence** ϵ-**typical** wrt p relative to reference q if:

$$|L_{p|q}(y) - D(p||q)| \leq \epsilon$$

- The set of all such sequences of length N: $\mathcal{T}_\epsilon(p|q; N)$

The two variations of the typical set are not strictly equivalent, but sequences of samples from p will fall in both sets with high probability. The sequences that are unique to one of the variations occur with negligible probability.

We can compute the probability of sampling a sequence from p's typical set using q:

$$q^N(y) \approx p^N(y)2^{-ND(p||q)}$$
$$Q\{\mathcal{T}_\epsilon(p|q;N))\} \approx 2^{-ND(p||q)}$$

The probability of sampling from q and producing any sequence in the typical set of p is exponentially small!

Thm 21.3: Given, p, q, N, then:

$$(1-\epsilon)2^{-N(D(p||q)+\epsilon)} \leq Q\{\mathcal{T}_\epsilon(p|q;N)\} \leq 2^{-N(D(p||q)+\epsilon)}$$

It is useful to characterize the "large deviation probability": $P\left(\frac{1}{N}\sum_N t(y_i) \geq \gamma\right)$

Thm 21.4 (Cramer's): Given y_i generated from q, and statistic $t : Y \to R$ with $\mu = E_q[t(y)] < \infty$. For any $\gamma > \mu$:

$$\lim_{n\to\infty} -\frac{1}{N}\log P\left(\frac{1}{N}\sum_N t(y_i) \geq \gamma\right) = E_C(\gamma)$$

where $E_C(\gamma) \triangleq D(p(\cdot;x)||q)$ is the **chernoff exponent** with $p(y;x) = q(y)e^{xt(y)-\alpha(x)}$ and $x > 0$ satisfying $E_{p(\cdot;x)}[t(y)] = \gamma$.

L22 - Method of Types/Sanov's Thoerem

Def 22.1: The **type**, or empirical distribution, of a sequence y is the probability distribution: $\hat{p}(b; y) = \frac{N_b(y)}{N}$ (where $N_b(y)$ is the count of b in the sequence).

Def 22.2: The **set of types** P_N^Y is the set of all possible types for sequences of length N generated from alphabet Y.

Def 22.3: For some $p \in P_N^Y$, the **type class** of p ($\mathcal{T}_N^Y(p)$) is the set of all sequences whose type is equal to p: $\mathcal{T}_N^Y(p) = \{y \;\; \hat{p}(\cdot; y) \cong p(\cdot)\}$

Identity 22.1: For arbitrary $g(\cdot)$ we can write: $(\prod_N g(y_i))^{1/N} = \prod_M g(b)^{\hat{p}(b;y)}$

We also define **exponential rate notation**: $f(N) \doteq 2^{N\alpha}$ denotes $\lim_{N\to\infty} \frac{\log f(N)}{N} = \alpha$

- Note the use of $\doteq$! This will be used frequently in the notes.
- This notation is a bit imprecise. Here are some cases:
  - $\rightarrow$ $f(N) \doteq \infty \doteq 2^{N\infty}$ - grows superexponentially
  - $\rightarrow$ $f(N) \doteq 0 \doteq 2^{-N\infty}$ - decays superexponentially
  - $\rightarrow$ $f(N) \doteq 1 \doteq 2^{N\cdot 0}$ - grows or decays subexponentially. This could mean many things including convergence to a constant, or N^2 or $1/N$, but the approximation is very coarse for the range of possible cases and doesn't directly separate them.
  - $\rightarrow$ given $f(N) \doteq 2^{N\alpha}, g(N) \doteq 2^{N\beta}$, addition and multiplication behave like $f(N) + g(N) \doteq 2^{N\max(\alpha,\beta)}$ and $f(N) \cdot g(N) \doteq 2^{N(\alpha+\beta)}$

Lemma 22.1 For any finite alphabet Y, $|P_N^Y| \leq (N+1)^{|Y|}$ - the number of total types is polynomial in sequence length.

Lemma 22.2: For a sequence y and distribution q we have:

$$q^N(y) = 2^{-N(D(\hat{p}(\cdot;y)||q)+H(\hat{p}(\cdot;y)))}$$

Lemma 22.3: $|\mathcal{T}_N^Y(p)| \approx 2^{NH(p)}$. Specifically:

$$cN^{-|Y|}2^{NH(p)} \leq |\mathcal{T}_N^Y(p)| \leq 2^{NH(p)}$$

- Every nondegenerate type class contains exponentially many sequences
- The proof uses: **Fact 22.1**: (coarse Stirling's) For $n \geq 1$: $e\left(\frac{n}{e}\right)^n \leq n! \leq ne\left(\frac{n}{e}\right)^n$

Thm 22.1: For $p, q \in P^Y$:

$$cN^{-|Y|}2^{-ND(p||q)} \leq Q\{\mathcal{T}_N^Y(p)\} \leq 2^{-ND(p||q)} \text{ and } Q\{\mathcal{T}_N^Y(p)\} \doteq 2^{-ND(p||q)}$$

- There is an exponentially small probability that a sequence obtained by sampling from q will have a type $p \neq q$
- However, a sequence generated from q will have a type other than q with high probability, but it will be within ϵ of q with high probability, and q will be the most likely candidate.

Given a set $\mathcal{S} \subset P^Y$ we can structure problems as computing the total probability over the set of sequences whose type matches $\mathcal{S}$: $\mathcal{R} = \{y \in Y^N : \hat{p}(\cdot; y) \in \mathcal{S} \cap P_N^Y\}$

Note: We adopt the following abuse of notation: $Q\{\mathcal{S}\} \triangleq P_q(\hat{p}(\cdot; y) \in \mathcal{S}) = Q\{\mathcal{R}\}$

- Note that $Q\{\mathcal{S}\} \triangleq P_q(\hat{p}(\cdot; y) \in \mathcal{S}) = Q\{\mathcal{S} \cap P_N^Y\}$!

Thm 22.2 (Sanov's Theorem): for arbitrary $S \subset P^Y$ and $q \in P^Y$:

$$Q\{S \cap P_N^Y\} \leq (N+1)^{|Y|} 2^{-ND(p_*||q)} \text{ and } Q\{S \cap P_N^Y\} \dot{\leq} 2^{-ND(p_*||q)}$$

where $p_* = \arg\min_{p \in \text{cl}(S)} D(p||q)$ is the I-projection of q onto $\text{cl}(S)$.

- Moreover, if $\text{cl}(S) = \text{cl}(\text{int}(S))$ then $Q\{S \cap P_N^Y\} \approx (N+1)^{|Y|} 2^{-ND(p_*||q)}$
- Note that $\text{cl}(S)$ refers to the *closure* of S - all points in S, and its boundary. $\text{int}(S)$ refers to the interior of S which disqualifies the theorem from having to process extremely weird sets which have random floating isolated points (since such points would not be in $\text{int}(S)$).
- The notes comment probability of any type p generated under q decays exponentially, but its maximized at the type closest to q which dominates. These bounds feel pretty loose to me so the claim seems a bit strong but I guess I should mention it.

We can use these developments to resolve the "large deviations" scenario from the previous lecture: $\frac{1}{N}\sum_N t(y_i) \geq \gamma$.

Define $\mathcal{R} = \{y; \hat{p}(\cdot; y) \in S \cap P_N^Y\}$ where $S = \{p : E_p[t(y)] \geq \gamma\}$. Then S is a linear family within P^Y, and the I-projection of q onto it lies along the one-parameter linear exponential family: $p_y(y; x) = q(y)e^{xt(y)-\alpha(x)}$.

Using these shortcuts, we can handle special large deviation events, such as:

$$P\left(\frac{1}{N}\sum_N y_i \geq \gamma_1 \quad \text{and} \quad \frac{1}{N}\sum_N s(y_i) \geq \gamma_2\right)$$

which the scalar version of Cramer's Theorem cannot. But the vector version of Cramer's theorem can handle this, and has 2 advantages: (1) It is not limited to finite distributions, and specifically works on continuous distributions (2) it has better bounds on the large deviation probability because it doesn't depend on the cardinality $|Y|$ of the alphabet.

Thm 22.3 (Conditional Limit Theorem): Given i.i.d. y generated from $q \in P^Y$, and nonempty closed and convex $S \subset P^Y$, then for any $\epsilon > 0$:

$$\lim_{N\to\infty} P(|\hat{p}(b; y) - p_*(b)| > \epsilon | \hat{p}(\cdot; y) \in S) = 0$$

where p_* is the I-projection from Sanov's theorem.

- Essentially, given that the sequence lies in S, $\hat{p}(\cdot; y)$ converges in probability to $p_*(\cdot)$.
- The proof first shows that as $N \to \infty$, sequences corresponding to types p near p_* (according to KL) will occur with probability 1 when generated from q (conditioning on the sequence begin in $\mathcal{S}$ to begin with). Then they bound $D(p||p_*) \leq \epsilon$, and use Pinsker's Inequality (below) to show convergence to p_*.

The proof of the above uses **Lemma 22.4** (Pinsker's Inequality): For any $q, p \in P^Y$ we have:

$$||p - q||_1 \triangleq \sum_{b \in Y} |p(b) - q(b)| \leq \sqrt{2 \ln 2 D(p||q)}$$

- Two distributions that are close in divergence are close in absolute difference as well!

L23 - Asymptotics of Hypothesis Testing

We extend previous results for the log-likelihood ratio test is formulated as:

$$L'(y) = \frac{1}{N}\log\frac{p_1^N(y)}{p_0^N(y)} = \frac{1}{N}\sum_N \log\frac{p_1(y_i)}{p_0(y_i)} \underset{\hat{H}=H_0}{\overset{\hat{H}=H_1}{\gtrless}} \gamma$$

while the corresponding LRT (i.e. Lec 2) would be:

$$L(y) \triangleq \frac{p_{y|H}(y|H_1)}{p_{y|H}(y|H_0)} = \frac{p_1^N(y)}{p_0^N(y)} \underset{\hat{H}=H_0}{\overset{\hat{H}=H_1}{\gtrless}} 2^{N\gamma}$$

We can define decision regions $\mathcal{R}_i = \{y \in Y^N : L'(y)(\geq / \leq)\gamma\}$

- where we include the equality case in both regions for convenience. This is imprecise, but it doesn't make much difference since these events are very small.

we further define $S_0 = \{p \in P^Y : E_p[\log\frac{p_1(y)}{p_0(y)}] \leq \gamma\}$ (and S_1 resp.)

- So $\mathcal{R}_i = \{y \in Y^N : \hat{p}(\cdot; y) \in S_i \cap P_N^Y\}$

We can write $E_{p_0}[\log\frac{p_1(y)}{p_0(y)}] = -D(p_0||p_1), E_{p_1}[\log\frac{p_1(y)}{p_0(y)}] = D(p_1||p_0)$. So for a "logical" experiment, we would want $-D(p_0||p_1) \leq \gamma \leq D(p_1||p_0)$.

- otherwise in expectation, data generated from p_0 would be classified as H_1 or vice versa, implying a very aggressive/passive detector respectively.

By Sanov's Theorem, the false-alarm and miss probabilities decay exponentially:

$$P_F = P_0\{R_1\} = 2^{-ND(p_0^*||p_0)}$$

where p_0^* is the I-projection of p_0 onto S_1. Analogously $P_1 = 2^P - ND(p_1^*||p_1)$.

The boundary between the distributions is the linear family:

$$\mathcal{L} = \left\{p \in P^Y : E_p\left[\log\frac{p_1(y)}{p_0(y)}\right] = \gamma\right\}$$

and the exponential family of weighted geometric means between p_0, p_1 is:

$$\mathcal{E}_{p_0,p_1} = \left\{p \in P^Y : p(y;x) = \frac{p_0(y)^{1-x}p_1(y)^x}{Z(x)}\right\} = \left\{p \in P^Y : \exp\left\{x\log\frac{p_1(y)}{p_0(y)} - \alpha(x) + \log p_0(y)\right\}\right\}$$

- This uses p_0 as its base distribution, and $t(y) = \log\frac{p_1(y)}{p_0(y)}$. So by thm 14.2 (orthogonal projection), $\mathcal{E}_{p_0,p_1}$ intersects the boundary $\mathcal{L}$ at p_0^*.
- x varies from $0 \leq x \leq 1$ with $p_0(\cdot) = p(\cdot;0), p_1(\cdot) = p(\cdot;1)$

If we reparametrize $\tilde{x} = 1 - x$, we can write $\mathcal{E}_{p_1,p_0}$ using $t(y) = \log\frac{p_0(y)}{p_1(y)}$ and rewrite $\mathcal{L}$ to use this form. This allows us to show that $\mathcal{E}_{p_1,p_0}$ intersects $\mathcal{L}$ at p_1^*, which **implies** $p_0^* = p_1^*$!

Using results from L14, for exp. families of the form $\exp\{xt(y) - \alpha(x) + \ln q(y)\}$ we have $\frac{d}{dx}D(p(\cdot;x)||q) = x\delta_{p(\cdot;x)}[t(y)]$, which applied to the above parametrization yields $D(p(\cdot;x)||p_0)$

is a monotonically increasing function of x. The alternate $\tilde{x}$ parametrization suggests $D(p(\cdot;x)||p_1)$ is a monotonically increasing function of $\tilde{x}$ (and thus monotonically decreasing in x).

We can then show that $E_p[\log \frac{p_1(y)}{p_0(y)}] = D(p(\cdot;x)||p_0) - D(p(\cdot;x)||p_1)$ is a monotonically increasing function of x, and $p_* = p_0^* = p_1^*$ is the unique choice of x satisfying $E_p[\log \frac{p_1(y)}{p_0(y)}] = \gamma$.

Key Equation: $E_{p_*}[\log \frac{p_1(y)}{p_0(y)}] = D(p_*||p_0) - D(p_*||p_1) = \gamma$

- This corresponds to $P_F \approx 2^{-ND(p_*||p_0)}$, $P_M \approx 2^{-ND(p_*||p_1)}$
- We choose x_* so that $D(p_*||p_0) - D(p_*||p_1) = \gamma$

Neyman-Pearson: We constrain P_F and minimize P_M

- Set $\gamma = D(p_0||p_1)$, which implies $p_* = p_0$
- Corresponds to $P_M \doteq 2^{-ND(p_0||p_1)}$ and $P_F \doteq 1$.
  - $\rightarrow$ Keep in mind exponential rate notation - this means $\lim \frac{\log P_F}{N} = 0$ so P_F decays subexponentially (to a constant), while P_M optimally decays with rate $D(p_0||p_1)$.
- **Stein's Lemma**: when $P_F \leq \alpha$, the fastest rate of decay for P_M is $D(p_0||p_1)$.

Bayesian: Given $C_{00} = C_{11} = 0$, the expected cost is:

$$\begin{aligned} E[C] &= C_{01}P_0P_F + C_{10}P_1P_M \\ &\doteq C_{01}P - 02^{-ND(p_*||p_0)} + C_{10}P_1 2^{-ND(p_*||p_1)} \\ &\doteq 2^{-N\min\{D(p_*||p_0), D(p_*||p_1)\}} \end{aligned}$$

Maximizing the rate of cost decay is equivalent to maximizing the minimum of $D(p_*||p_0)$, $D(p_*||p_1)$. Obviously, we maximize this by setting $D(p_*||p_0) = D(p_*||p_1)$.

This corresponds to $\gamma = 0$, and LRT threshold $2^{N\gamma} = 1$.

The notes also define the **Chernoff information** E_C, which didn't seem relevant.

They also commented that the expected cost decays exponentially with E_C, and any subexponential LRT threshold in general.

In the special case that both p_0/p_1 belong to the same set (either of) $\mathcal{S}_i$, we see one of P_M/P_F decay to 0, while the other approaches 1 as $N \rightarrow \infty$

L24 - Convergence of Random Sequences

Def 24.1 (Almost-Sure): A sequence of random variables z_i converges almost surely (or with probability 1) to a random variable z if $P(\lim_{N\to\infty} z_N = z) = 1$.

- Notation: $z_N \xrightarrow{a.s.} z$

Def 24.2 (In Probability): A sequence of random variables z_i converges in probability (or with "high" probability) to a random variable z if $\lim_{N\to\infty} P(|z_N - z| < \epsilon) = 1$.

- Notation: $z_N \xrightarrow{p} z$
- We can use this formulation to alternatively define almost sure convergence :
  - → **Fact 24.1**: for a r.v. collection z_i and z, we have $z_N \xrightarrow{a.s.} z$ iff for every $\epsilon > 0$, $\lim_{N\to\infty} P(|z_N - z| < \epsilon \text{ for all } N > N_0) = 1$
    Essentially, in almost sure convergence the probability that a sample path eventually gets to and remains within ϵ of z approaches one. Meanwhile, converge in probability only suggests the probability that any individual sample in the sequence approaches one as $N \to \infty$.
  - → **Fact 24.2** For a r.v. collection z_i and z we have $z_N \xrightarrow{a.s.} z$ if for every $\epsilon > 0$ we have $\sum_{n=1}^{\infty} P(|z_n - z| > \epsilon) < \infty$
    Essentially, the almost sure convergence is equivalent to convergence in probability decaying sufficiently fast.

Def 24.3 (In Distribution): A sequence of random variables z_i converges in distribution (or in "law") to a random variable z if CDF $\lim_{N\to\infty} P_{z_N}(z) = P_z(z)$ for all z with continous $P_z(\cdot)$

- Notation: $z_N \xrightarrow{d} z$
- equivalently: $\lim_{n\to\infty} E[g(z_N)] = E[g(z)]$ for all bounded, continuous $g(\cdot)$.
- (stronger variant): **Def 24.4** (**In Divergence**): A sequence z_i converges in divergence (or strongly in law) to r.v. z if $\lim_{n\to\infty} D(p_{z_N}||p_z) = 0$

Thm 24.1 (Continuous Mapping): If z_i are defined over R^k and $g : R^k \to R$ is continuous then for $x \in d, p, a.s.$ we have $z_N \xrightarrow{x} z$ implies $g(z_N) \xrightarrow{x} g(z)$.

- Generally, this holds whenever the set of points with $g(\cdot)$ is discontinuous has probability 0 under p_z.

Thm 24.2 (Slutsky's): If $x_N \xrightarrow{d} x$ and $y_N \xrightarrow{d} c$ where c is a finite constant, then:

$$x_N + y_N \xrightarrow{d} x + c \text{ and } x_N y_N \xrightarrow{d} cx$$

and if z_N is any sequence with $x_N = z_N y_N$ and $c \neq 0$, then $z_N \xrightarrow{d} x/c$.

Thm 24.3 (SLLN): The exact same conditions and result as the WLLN (Thm 21.1), but with almost sure convergence (instead of in probability)!

For analyzing estimators, we want a more general version of the SLLN. We can use the following variant, created by Wald:

Thm 24.4 (Uniform LLN): For i.i.d. $w_1, \cdots, w_N \in W$, and compact parameter set Θ, and $g(w;\theta)$ which is continuous in θ for each w and which has some $g_+(w) > 0$ with $E[g_+(w)] < \infty$ and $|g(w;\theta)| \leq g_+(w)$ for all w, θ, then:

$$\max_\theta \left| \frac{1}{N} \sum_N g(w_i;\theta) - \mu(\theta) \right| \xrightarrow{a.s.} 0 \text{ as } N \to \infty$$

where $\mu(\theta) = E[g(w;\theta)]$ - the sample average almost surely converges uniformly to its mean.

Thm 24.5 (CLT): Let w_i be a set of i.i.d. r.v.s with mean μ and variance $\sigma^2 < \infty$. Then:

$$\sqrt{N}\left(\frac{1}{N} \sum_N w_i - \mu \right) \xrightarrow{d} N(0, \sigma^2) \text{ as } N \to \infty$$

- A theorem from Polya tells us if $z_N \xrightarrow{d} z$ and CDF $P_z(\cdot)$ is continuous, then the convergence is uniform ($\lim \max_z |P_{z_N}(z) - P_z(z)| = 0$), so the convergence in the CLT is uniform.
- We even have a $1/\sqrt{N}$ bound on the convergence thanks to Berry-Essen theorem apparently.

The following stronger variant also exists, although I don't exactly understand it. **Thm 24.6** (SCLT): Let w_i be a set of i.i.d. r.v.s with mean μ and variance $\sigma^2 < \infty$. Then:

$$e_N \triangleq \sqrt{N}\left(\frac{1}{N} \sum_N w_i - \mu \right) \xrightarrow{D} N(0, \sigma^2) \text{ as } N \to \infty$$

iff $D(p_{e_N} || N(0, \sigma^2)) < \infty$ for some N.

L25 - Asymptotics of Parameter Estimation

We begin with **nonrandom** parameter estimation

Def 25.1: A parameter x is **identifiable** if $x' \in X$ satisfies $D(p_y(\cdot;x)||p_y(\cdot;x')) = 0$ only when $x' = x$.

Def 25.2: A sequence $\hat{x}_N(y^N)$ of estimates is **weakly consistent** if $\hat{x}_N(y^N)) \xrightarrow{p} x$ as $N \to \infty$

- Strong consistency is the same, but with almost sure convergence

Def 25.3 A sequence of estimates is **asymptotically normal** if $\sqrt{N}(\hat{x}_N - x) \xrightarrow{d} N(0, \frac{1}{\sigma^2(x)})$ for some $\sigma^2(x) > 0$

Def 25.4: A sequence of estimates is **asymptotically efficient** if it is asymptotically normal as in 25.3 with $\sigma^2(x) = J_y(x)$, the fisher information of $p_y(y;x)$.

We also define empirical divergence: $\hat{D}_y^N(a) \triangleq \frac{1}{N}\sum_N \log \frac{p_y(y_n;x)}{p_y(y_n;a)}$

For finite X alphabets, we have the following result using WLLN:

Thm 25.1: For $p_y(\cdot;\cdot)$ with $|Y| < \infty$ and $|X| < \infty$, and y_i generated i.i.d. according to some identifiable $p_y(\cdot;x)$ with $p_y(y;a) > 0$ for all y, a, the MLE $\hat{x}_N(y^N)$ is weakly consistent.

We can also extend to continuous X: $p_y(y;a) = \exp\{at(y) - \alpha(a) + \beta(y)\}$, for which we derived properties of the ML estimate in L19. Here we show the resulting estimate is strongly consistent and asymptotically efficient.

Thm 25.2: For an exponential family as above and y_i generated according to some representative identifiable $p_y(\cdot;x)$ with $\ddot{\alpha}(x) = J_y(x) > 0$, the ML estimate $\hat{x}_N(y^N)$ is strongly consistent.

Thm 25.3: For an exponential family as above and y_i generated according to some representative identifiable $p_y(\cdot;x)$ with $\alpha(x) = J_y(x) > 0$, the ML estimate is asymptotically efficient.

TODO: I skipped some examples which are probably helpful to read

---

We can also prove some results when the parametrization for p_y is of the wrong form.

Thm 25.6: For P with finite $|Y|, |X|$ and samples y_i from $q_y \notin P$ with $q_y(y), p_y(y;x) > 0$ for all y, x and with unique $x_* = \arg\min_x D(q_y||p_{y|x}(\cdot|x))$, then $\hat{x}_N(y_i) \xrightarrow{p} x_*$.

We then evaluate **bayesian** parameter estimation.

Generally we want to be able to integrate $p_{x|y^N}(x|y^N) = \int p_x(a)p_{y^N|x}(y^N|a)\,da$. We can approximate this with Laplace's Method

Laplace's Method: We approximate integrals of the form $\int_a^b e^{Ng(z)}\,dz$ where g is smooth and N is large using a taylor series for $g(z)$ about $\arg\max_{z\in[a,b]} g(z)$.

Thm 25.7 (Laplace's method): for twice-differentiable $g : [a,b] \to R$ with unique maximum $z_* \in (a,b)$ s.t. $\ddot{g}(z_*) < 0$ we have:

$$\frac{\int_a^b e^{Ng(z)}\,dz}{e^{Ng(z_*)}\sqrt{\frac{2\pi}{|N\ddot{g}(z_*)|}}} \to 1 \text{ as } N \to \infty$$

The notes then specialize this to show that when X is continuous, under some suitable conditions the posterior tends towards a gaussian centered at $\hat{x}_{N,MLE}(y)$ $p_{x|y^N} \approx N(\hat{x}_N(y^N), \frac{1}{NJ_y(x)})$

They also prove a result when $p_{y|x}$ is from an exponential family:

Thm 25.8: With $p_{y|x}$ from an exponentntial family and y_i generated i.i.d according to $p_{y|x}$ where x is generated from some prior distribution p_x and all x are identifiable and $\ddot{\alpha}(x) = J_y(x) > 0$. Then:

$$\ln \frac{p_{\tilde{x}|y}(\tilde{a}|y^N)}{p_{\tilde{x}|y}(\tilde{0}|y^N)} \xrightarrow{a.s.} \frac{1}{2} J_y(x)\tilde{a}^2 \text{ as } N \to \infty,\ \tilde{a} \in R$$

where $\tilde{x} \triangleq \sqrt{N}(x - \hat{x}_N(y^N))$