Machine Learning Notes

Matthew Scicluna

January 3, 2018

Contents

1 Notation

Notation	Meaning
X	random variable
x	instantiation of random variable
$\int f(x)d\mu(x)$	Lebesgue integral of f w.r.t. measure μ
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^n$	Set of n-tuples of real numbers
$\mathbb{R}^{n\times m}$	Set of n by m matrices of real numbers
$\mathbb{N}$	Set of natural numbers
$\sup(A)$	Supremum of a set A
$\inf(A)$	Infemum of a set A
$\mathbb{1}_A(x)$	Indicator function of set A
$\mathbb{E}_{X\sim p}\{f(X)\}$	The expected value of $f(X)$ where $X \sim p$
$\mathbb{E}\{X\|Y\}$	The expected value of X conditioned on Y
$\mathcal{Y}^{\mathcal{X}}$	the set of functions $f : \mathcal{X} \mapsto \mathcal{Y}$
$\frac{\partial f}{\partial x_i}$	partial derivative of f w.r.t component x_i
$\nabla_x f(x)$	gradient of f evaluated at x
$Hf(x)$	Hessian of f evaluated at x
$Jf(x)$	Jacobian Matrix of f evaluated at x
$[A]_{ij}$	ith row and jth column of matrix A
$B_r(x)$	Ball of radius r centred at x
$\text{int}(A)$	Interior points of set A
$\text{bd}(A)$	Boundary points of set A
$\text{dom}(f)$	Domain of function f

2 Preliminaries

2.1 Vector Calculus

Let $S \subseteq \mathbb{R}^n$ and x_0 an interior point of S with $B_r(x_0) \subseteq S$. Given a function $f : S \to \mathbb{R}^m$ we say that f is **Differentiable** at x_0 if there exists an $A \in \mathbb{R}^{m \times n}$ depending only on x_0 s.t. $\forall ||\Delta|| < r$

$$f(x_0 + \Delta) - f(x_0) = A(x_0)\Delta + r_{x_0}(\Delta)$$
$$\text{where } \frac{r_{x_0}(\Delta)}{||\Delta||} \to 0 \text{ as } \Delta \to 0$$

If f is differentiable at every point of an open subset E of S, then f is **Differentiable** on E. We call $df(x_0, \Delta) = A(x_0)\Delta \in \mathbb{R}^{m \times 1}$ the first **Differential** of f at x_0.

Theorem 2.1. *f is **Differentiable** at x_0 if and only if each component of f denoted as f_i $i = 1, \cdots, m$ is differentiable at x_0. In that case $[df(x_0, \Delta)]_i = df_i(x_0, \Delta)$*

Proof. Magnus and Neudecker chapter 5 [1] □

So f is only differentiable if each of its m components are seperately differentiable. Let f_i be the ith component of f, with f and x_0 defined as before, e_j be the jth unit vector of $\mathbb{R}^n$ we define the **Partial Derivative** of f_i w.r.t x_j as

$$\frac{\partial f_i(x_0)}{\partial x_j} = \lim_{t \to 0} \frac{f_i(x_0 + te_j) - f_i(x_0)}{t}$$

We define the **Jacobian Matrix** of f as:

$$Jf(x_0) = \begin{bmatrix} \frac{\partial f_1(x_0)}{\partial x_1} & \cdots & \frac{\partial f_1(x_0)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(x_0)}{\partial x_1} & \cdots & \frac{\partial f_m(x_0)}{\partial x_n} \end{bmatrix}$$

It can be shown that if f is differentiable, then all its partial derivatives exist, although the converse is not true! This is why Theorem 2.1 only holds in one direction. Note that the Jacobian is defined at any point where all the partial derivatives exist, even if f is not differentiable at that point! When the Jacobian is square, we call its determinant the **Jacobian** of f.

Theorem 2.2. $[A(x_0)]_{ij} = \left.\frac{\partial f_i(x)}{\partial x_j}\right|_{x=x_0}$

Proof. Magnus and Neudecker chapter 5 theorem 5 [1] □

By construction, $Jf(x_0) = A(x_0)$. We call the transpose of the Jacobian Matrix the **Gradient** and denote it as $\nabla_x f(x_0)$. Finally, we give an important result: the chain rule.

Theorem 2.3 (Chain Rule). *Let f,x_0 defined as before and $g : T \to \mathbb{R}^p$. Suppose that $T \subseteq \mathbb{R}^m$, $f(S) \subseteq T$, $f(x_0)$ is an interior point of T, and that g is differentiable at $f(x_0)$. Then $Jg \circ f(x_0) = Jg(f(x_0))Jf(x_0)$*

Proof. Magnus and Neudecker chapter 5 theorem 8 [1] □

As a sanity check, we can see that $Jg \circ f(x_0) \in \mathbb{R}^{p\times n}$ while $Jg(f(x_0)) \in \mathbb{R}^{p\times m}$ and $Jf(x_0) \in \mathbb{R}^{m\times n}$.

2.1.1 Many to One Functions and the Hessian

The majority of functions we work with in machine learning are real valued, meaning they are of form $f : \mathbb{R}^n \to \mathbb{R}$. This is since they are either loss functions or probability densities. Hence we focus on this case. The most common use of vector calculus is to optimize a function by finding its stationary points (where the gradient is 0) and to check whether it is a maxima or minima by checking the Hessian. First we present a simpler version of the chain rule for gradients.

Theorem 2.4 (Simplified Chain Rule). *If we consider only functions g with $p = 1$ then the gradient is:*

$$\nabla_x g \circ f(x) = Jf(x)^T \nabla_x g(f(x))$$

The Jacobian matrix for f takes the following form:

$$Jf(x) = \left(\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n}\right) \in \mathbb{R}^{1\times n}$$

The Gradient for f is then:

$$\nabla_x f(x) = \left(\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n}\right)^T \in \mathbb{R}^{n\times 1}$$

We define the **Second Partial Derivative** for a real valued f as follows. Let f, e_i as before. The second partial derivative of f w.r.t x_i and x_j is:

$$\frac{\partial^2 f(x_0)}{\partial x_i \partial x_j} = \lim_{t\to 0} \frac{\frac{\partial f(x_0+te_i)}{\partial x_j} - \frac{\partial f(x_0)}{\partial x_j}}{t}$$

We define the **Hessian** for real valued functions as:

$$Hf(x_0) = \begin{bmatrix} \frac{\partial^2 f(x_0)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x_0)}{\partial x_n \partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x_0)}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f(x_0)}{\partial x_n^2} \end{bmatrix} \in \mathbb{R}^{n\times n}$$

2.1.2 Some Important Examples

We now get to the payoff, some important results which will appear often in these notes. Let $b \in \mathbb{R}^n$, $B \in \mathbb{R}^{n \times n}$

Theorem 2.5. $\nabla_x x^T B x = x^T(B^T + B)$

Proof. Notice

$$\begin{aligned}(x+\Delta)^T B(x+\Delta) - x^T B x &= \Delta^T B x + x^T B \Delta + \Delta^T B \Delta \\ &= x^T(B + B^T)\Delta + \Delta^T B \Delta \\ &= A(x)\Delta + r_x(\Delta)\end{aligned}$$

Where $r_x = \Delta^T B \Delta$ and $A(x) = x^T(B + B^T)$. We see that $r_x \in O(||\Delta||)$, and so the differential is $x^T(B + B^T)\Delta$. Therefore, $\nabla_x x^T B x = (B + B^T)x$, the transpose. □

Theorem 2.6. $\nabla_\mu (x-b)^T B (x-b) = (x-b)^T (B + B^T)$

Proof. We use the chain rule where $f(x) = x - b$ and $g(y) = y^T B y$.

$$\begin{aligned}&\nabla_y g(y) = (B + B^T)y \\ &Jf(x) = I \\ &\nabla_x g \circ f(x) = (B + B^T)(x - b)\end{aligned}$$

□

For the final example, we must refine our definition of differentiability to include matrices. That's right, you can actually do that. First we define the norm of a matrix A as:

$$||A|| = tr(A^T A)^{\frac{1}{2}}$$

Let $S \subseteq \mathbb{R}^{n \times m}$ and x_0 an interior point of S with $B_r(x_0) \subseteq S$. Given a function $f : S \to \mathbb{R}$ we say that f is **Differentiable** at x_0 if there exists an $A \in \mathbb{R}^{n \times m}$ depending only on x_0 s.t. $\forall ||\Delta|| < r$

$$\begin{aligned}&f(x_0 + \Delta) - f(x_0) = tr(A(x_0)^T \Delta) + r_{x_0}(\Delta) \\ &\text{where } \frac{r_{x_0}(\Delta)}{||\Delta||} \to 0 \text{ as } \Delta \to 0\end{aligned}$$

We define the Jacobian and Gradient in terms of $A(x_0)$ in the same way that we did for vector valued functions.

Theorem 2.7. $\nabla_B \log\det B = B^{-1}$ *for* B *symmetric, invertible*

Proof. First notice the following decomposition

$$\begin{aligned}\log\det(B+\Delta) &= \log\det\left(B^{\frac{1}{2}}\left(I + B^{-\frac{1}{2}}\Delta B^{-\frac{1}{2}}\right)B^{\frac{1}{2}}\right)\\ &= \log\left(\det\left(B^{\frac{1}{2}}\right)\det\left(I + B^{-\frac{1}{2}}\Delta B^{-\frac{1}{2}}\right)\det\left(B^{\frac{1}{2}}\right)\right)\\ &= \log\det\left(I + B^{-\frac{1}{2}}\Delta B^{-\frac{1}{2}}\right) + \log\det B\end{aligned}$$

We try to find our differential in the same way we did in the first example:

$$\begin{aligned}\log\det(B+\Delta) - \log\det B &= \log\det\left(I + B^{-\frac{1}{2}}\Delta B^{-\frac{1}{2}}\right)\\ &\overset{(a)}{=} \sum_i \log\left(1 + \lambda\left(B^{-\frac{1}{2}}\Delta B^{-\frac{1}{2}}\right)\right)\\ &\overset{(b)}{=} \sum_i \lambda\left(B^{-\frac{1}{2}}\Delta B^{-\frac{1}{2}}\right) + O(||\Delta||)\\ &\overset{(c)}{=} \operatorname{tr}\left(B^{-\frac{1}{2}}\Delta B^{-\frac{1}{2}}\right) + O(||\Delta||)\\ &= \operatorname{tr}(B^{-1}\Delta) + O(||\Delta||)\end{aligned}$$

Where (a) follows since $\det A = \prod_i \lambda_i(A)$; (b) using $\log(1+x) = x + O(x^2)$ for $|x| < 1$; and (c) using $\operatorname{tr}(A) = \sum_i \lambda_i(A)$. We use the characterization of the gradient for functions of matrices to get that $\nabla_B \log\det B = B^{-1}$. □

Finally, we mention in passing some other useful identities.

- $\nabla_x b^T x = \nabla_x x^T b = b$
- $\nabla_B x^T B x = x x^T$

2.2 Optimization

We now discuss how to solve optimization problems. We want to minimize our **Objective Function** $f_0 : \mathbb{R}^n \to \mathbb{R}$ w.r.t some **Optimization Variable** $x \in \mathbb{R}^n$ subject to some **Inequality Constraints** $f_i : \mathbb{R}^n \to \mathbb{R}$ and some **Equality Constraints** $h_i : \mathbb{R}^n \to \mathbb{R}$. If there are no constraints (i.e.) $m = p = 0$ then we call the problem **Unconstrained**. We set $f_0(x) = \infty$ for $x \notin \text{dom}(f_0)$. Formally, our optimization problem can be written as:

$$\begin{aligned} &\text{Minimize } f_0(x) \\ &\text{subject to } f_i(x) \leq 0 \ i = 1, \cdots, m \\ &\qquad\qquad h_i(x) = 0 \ i = 1, \cdots, p \end{aligned}$$

We denote the intersection of the domains of f_0, the f_i's and the h_i's as $\mathcal{D}$. We call a point x **Feasible** if $x \in \mathcal{D}$ and x satisfies the constraints. The **Optimal Value** of this problem is $p^* = \inf\{f_0(x) : \text{x feasible}\}$. If x feasible and $f_0(x) = p^*$ then we call it an **Optimal Point**. The set of Optimal Points is called the **Optimal Set**. Often we will find points which minimize f_0 only over points within some radius. Such points are called **Locally Optimal**. We note two degenerate cases to be careful for:

- the problem has no feasible points, so we set $p^* = \infty$
- the problem has arbitrarily small optimal values, $p^* = -\infty$

2.2.1 Convex Sets and Functions

We describe a highly desirable property of functions we are optimizing over: Convexity. If a function is convex then any locally optimal point is globally optimal. A set A is called **Convex** if:

$$\begin{aligned} (1-\alpha)x + \alpha y \in A \\ \forall x, y \in A \text{ and } \forall \alpha \in [0,1] \end{aligned}$$

A function f is **Convex** if:

$$\begin{aligned} f((1-\alpha)x + \alpha y) \leq (1-\alpha)f(x) + \alpha f(y) \\ \forall x, y \in \text{dom(f) and } \forall \alpha \in [0,1] \end{aligned}$$

If the equality is strict (i.e. $<$) then f is called **Strictly Convex**. We call a function **Concave** if $-f$ is convex. There is a relationship between convexity of a set and that of a function. We define the **Epigraph** of a function is as:

$$\Big\{(x,t) : x \in \text{dom(f)}, t \geq f(x)\Big\}$$

Theorem 2.8. *A function f is convex if and only if its epigraph is a convex set*

Proof. Boyd [2] □

We present some examples of Convex functions.

- $\exp ax, -\log x$
- Any norm over $\mathbb{R}^n$
- non-negative weighted sums of convex functions f_i's
- Let f be convex with domain $\mathcal{X} \times \mathcal{Y}$. If for each $y \in \mathcal{Y}$ we have that $f(x, y)$ is convex in x; then

$$g(x) = \sup_{y \in \mathcal{Y}} f(x, y) \text{ is convex in } x$$

To check for convexity, we present first and second order conditions which are both necessary and sufficient:

Theorem 2.9. *A function f is convex if and only if dom(f) is convex and $\forall x, y \in$dom(f)*

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

Proof. Boyd section 3.1.3 [2] □

This is a significant result since this demonstrates how local information from a convex function (i.e. its derivative) can yield global information about it (the inequality holds across its entire domain)!

Theorem 2.10. *A function f is convex if and only if dom(f) is convex and $\forall x \in$dom(f) $Hf(x)$ is positive semi definite.*

Proof. Boyd section 3.1.4 [2] □

We can strengthen the result. If $\text{dom}(f)$ is convex and $Hf(x)$ is positive definite, then f is strictly convex. The converse, however, does not hold. A **Convex Optimization Problem** is an optimization problem where $f_0, f_1, \cdots, f_m$ are convex and h_i are **Affine**: $h(x) = 0$ can be written as $Ax = b$, where $A \in \mathbb{R}^{d \times n}$, $b \in \mathbb{R}^n$. These problems have a very appealing property:

Theorem 2.11. *For any convex optimization problem, any locally optimal point is optimal.*

Proof. Boyd section 4.2.2 [2] □

2.2.2 Unconstrained Optimization

For unconstrained problems it is easy to verify whether a point is a local optima as these points have clear first and second order necessary and sufficent conditions. If x^* is a local minima of f_0 and f_0 is twice continuously differentiable and $x^* \in \text{int}(\mathcal{D})$; then the following are necessary: $\nabla_x f_0(x^*) = 0$ and $Hf_0(x^*)$ positive semi-definite. The following are sufficient: $\nabla_x f_0(x^*) = 0$ and $Hf_0(x^*)$ is positive definite.

If f_0 is convex then the local minima is the optimal point. If f_0 is not convex the procedure becomes more complicated. We must compare the points satisfying the sufficient conditions (or just the necessary ones if none satisfy the sufficient ones) along with any points in $\text{bd}(\mathcal{D})$ to find the minima.

2.2.3 Langrangian Duality

We can solve our optimization problem using the **Lagrangian** $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ with domain $\mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \tag{1}$$

The Langrangian itself may be difficult to minimize. The problem can be simplified by introducing the **Lagrange dual function** $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) \tag{2}$$

We call (1) the **Primal Problem** and (2) the **Dual Problem**. We note that g is concave (regardless of f_0) and can be $-\infty$. We call (λ, ν) **Dual Feasible** if $\lambda \geq 0$ and $(\lambda, \nu) \in \text{dom}(g)$, where $\text{dom}(g)=\{(\lambda, \nu) | g(\lambda, \nu) > -\infty\}$ [1]. We denote d^* as the **Dual Optimum Value**: the infemum of g. We say that (λ, ν) is **Dual Optimum** if it is dual feasible and $g(\lambda, \nu) = d^*$. We now relate the primal and dual problems:

Theorem 2.12 (Lower Bound Property). *Let $\lambda \geq 0$, then $g(\lambda, \nu) \leq p^*$*

Proof. Note that for any feasible $\tilde{x}$ and $\lambda \geq 0$:

$$f_0(\tilde{x}) \geq L(\tilde{x}, \lambda, \nu) \geq \inf_{x \in D} L(x, \lambda, \nu) = g(\lambda, \nu)$$

and since p^* is the infemum of all feasible $\tilde{x}$, it follows that $p^* \geq g(\lambda, \nu)$ □

Instead of minimizing f_0 using the primal we can maximize the lower bound g. This may be an easier problem since g is always concave. It is clear that we always have **Weak Duality** $p^* \geq d^*$, although we want **Strong Duality**: $p^* = d^*$. A sufficient condition for strong duality is **Slater's Condition**.

Theorem 2.13 (Slater's Condition). *Suppose we have a convex primal, and that int($\mathcal{D}$) is non empty. If $\exists x \in$int($\mathcal{D}$) satisfying $f_1(x) < 0, \cdots, f_m(x) < 0$ and $h_1(x) = 0, \cdots, h_p(x) = 0$, we have strong duality. If any f_i are affine, we only need them to satisfy $f_i(x) = 0$.*

Proof. Boyd section 5.3.2 [2] □

[1] We note that we could add extra constraints to prevent $g(\lambda, \nu) = -\infty$ and this would not change anything. See Boyd section 5.2.1 [2]

2.2.4 Saddle Point Interpretation

Given a $w^* \in W$, $z^* \in Z$ we say that (w^*, z^*) is a **Saddle Point** for function f with domain $W \times Z$ if $f(w^*, z) \leq f(w^*, z^*) \leq f(w, z^*)$ for all $(w, z) \in W \times Z$. This means that w^* minimizes $f(w, z^*)$ over W and z^* maximizes $f(w^*, z)$ over Z:

$$f(w^*, z^*) = \inf_{w \in W} f(w, z^*) = \sup_{z \in Z} f(w^*, z)$$

First we notice the following:

$$\begin{aligned}\sup_{\substack{\lambda \geq 0 \\ \nu}} L(x, \lambda, \nu) &= \sup_{\substack{\lambda \geq 0 \\ \nu}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right) \\ &= \begin{cases} f_0(x) & x \text{ is feasible} \\ \infty & \text{o.w.} \end{cases}\end{aligned}$$

We can show this in cases. Suppose x is feasible. $h_i(x) = 0$ so we can ignore them. Since $f_i(x) \leq 0,\ i = 1, \cdots, m;\ \lambda = 0$ would optimize L since $\sum_{i=1}^{m} \lambda_i f_i(x) \leq 0$. Now suppose x was infeasible and WLOG suppose that $\exists j\ s.t.\ f_j(x) > 0$, then we can make L arbitrarily large by taking $\lambda_j \to \infty$. The same reasoning works for any $h_i(x) \neq 0$. Hence the result. From this we can express p^* as:

$$\begin{aligned}p^* &= \inf_x f_0(x),\ \text{x feasible} \\ &= \inf_x \sup_{\substack{\lambda \geq 0 \\ \nu}} L(x, \lambda, \nu)\end{aligned}$$

and our dual is defined as

$$d^* = \sup_{\substack{\lambda \geq 0 \\ \nu}} \inf_x L(x, \lambda, \nu)$$

Hence strong duality can be expressed in the following way, clearly showing that (x, λ, ν) is a saddle point for L.

$$\inf_x \sup_{\substack{\lambda \geq 0 \\ \nu}} L(x, \lambda, \nu) = \sup_{\substack{\lambda \geq 0 \\ \nu}} \inf_x L(x, \lambda, \nu)$$

2.2.5 KKT Conditions

We now state some conditions often used to determine whether a solution x^* of f is optimal. These are the **KKT Conditions**.

Theorem 2.14 (KKT Conditions). *Let f_0, f_i, h_i be differentiable (and so they have open domains). If x^* is optimal and (λ^*, ν^*) are dual optimal and we have strong duality; then the following Conditions must be satisfied:*

- $\nabla_x f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla_x f_i(x^*) + \sum_{i=1}^p \nu_i^* \nabla_x h_i(x^*) = 0 \rightarrow$ *Stationarity*
- $\lambda_i^* f_i(x^*) = 0 \rightarrow$ *Complementary Slackness*
- x^* *is feasible* $\rightarrow$ *Primal Feasibility*
- $\lambda^* \geq 0 \rightarrow$ *Dual Feasibility*

Proof. We have to show Stationarity and Complementary Slackness. Notice that

$$f_0(x^*) = g(\lambda^*, \nu^*) = \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p \nu_i^* h_i(x) \right) \tag{3}$$

$$\leq f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p \nu_i^* h_i(x^*) \tag{4}$$

$$\overset{(a)}{\leq} f_0(x^*) \tag{5}$$

Where (a) comes from the condition that $h_i(x^*) = 0 \Rightarrow \sum_{i=1}^p \nu_i^* h_i(x^*) = 0$ and $f_i(x^*) \leq 0, \lambda_i \geq 0 \Rightarrow \sum_{i=1}^m \lambda_i^* f_i(x^*) \leq 0$. We get that $\inf_x L(x, \lambda^*, \nu^*) = f_0(x^*)$, i.e. x^* minimizes $L(x, \lambda^*, \nu^*)$ which implies Stationarity. Complementary slackness comes from $\sum_{i=1}^m \lambda_i^* f_i(x^*) = 0$ and $\lambda_i^* f_i(x^*) \leq 0 \Rightarrow \lambda_i^* f_i(x^*) = 0$ for every i. □

Theorem 2.15. *If x^*, λ^*, ν^* satisfy the KKT conditions and the primal is convex; then x^* is optimal, λ^*, ν^* are dual optimal and we have strong duality.*

Proof.

$$g(\lambda^*, \nu^*) = \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p \nu_i^* h_i(x) \right) \tag{6}$$

$$\overset{(a)}{=} f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p \nu_i^* h_i(x^*) \tag{7}$$

$$\overset{(b)}{=} f_0(x^*) \tag{8}$$

Since $\lambda^* \geq 0$, $L(x, \lambda^*, \nu^*)$ is convex in x. By the KKT conditions, x^* is primal feasible and $\nabla_x L(x, \lambda^*, \nu^*)$ evaluated at x^* is $0 \Rightarrow x^*$ minimizes $L(x, \lambda^*, \nu^*)$ (a). (b) follows from complementary slackness. Therefore, we have strong duality and so x^* is optimal and λ^*, ν^* are dual optimal. □

In some cases we can find optimal points by using the Dual to solve the Primal. Specifically; if:

1. strong duality holds
2. (λ^*, ν^*) is a dual optimal solution
3. $L(x, \lambda^*, \nu^*)$ has a unique minimum value (e.g. $L(x, \lambda^*, \nu^*)$ is strictly convex in x)

Let x^* be the unique optima, then either:

1. x^* is feasible; and so x^* must be optimal.
2. x^* is not feasible and no optimal can exist.

2.3 Basic Probability

We now discuss the requisite probability theory needed for Machine Learning. What is a probabilistic question? One example of this would be:

"Suppose you flipped 4 coins, what is the probability you see exactly 2 Heads?"

Before we answer this we should list the **Sample Space**: the set of all **Outcomes**. For us, this is:

$$\Omega = \Big\{ \text{HHHH}, \text{THHH}, \cdots, \text{TTTT} \Big\}$$

There are 16 outcomes. An **Experiment** consists of observing an outcome. If we believe that each outcome is equiprobable then we can assign a probability of $\frac{1}{16}$ to each. A subset of our probability space is called an **Event**. For example, the set of outcomes where 2 heads are observed is an event. This event has $\binom{4}{2} = 6$ outcomes in it. It follows that the probability of this event would be $\frac{6}{16}$.

Now lets consider a more complicated example. Suppose the experiment consisted of viewing a number in $(0, 1)$ at random. What is the probability you observe a rational number? To answer this question we need to introduce some notions from Measure Theory. We define the **σ-algebra** $\mathcal{F}$ on Ω to be a set of Events which satisfy the following properties:

- $\Omega \in \mathcal{F}$
- $E \in F$ implies that $E^{\complement} \in \mathcal{F}$ *Closure under complements*
- $E_1, E_2, \cdots$ are each in $\mathcal{F}$ then $\bigcup_{i=1}^{\infty} E_i \in \mathcal{F}$ *Closure under countable unions*

Given a Sample space Ω and a σ-algebra $\mathcal{F}$ on Ω we can define the **Probability** as a mapping $P : \mathcal{F} \mapsto [0, 1]$, which is interpreted as the "probability" of the set. We require that it satisfies the following 3 axioms:

- For any $E \in \mathcal{F}$, $P(E) \geq 0$. *probabilities must be non-negative*
- $P(S) = 1$ *We are certain something in the Sample Space was observed*
- For any countable collection of disjoint sets $E_1, E_2, \cdots$ we have that:

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P\left(E_i\right)$$

A subtle point here: not every σ-algebra can have a valid probability measure associated with it[2]. We can compute Probabilities of Events Conditioned on

[2]For example, if we take Ω to be $(0, 1)$ and F to be $\mathcal{F} = 2^{\Omega}$ (which always is a σ-algebra), we can construct a set for which there is no Probability measure which simulteneously satisfies axioms 2 and 3. This requires invoking the Axiom of Choice and is very non-trivial. You can find the construction in page 3 of this book [3]

other Events. We define the **Conditional Probability** of event A on event B with $P(B) > 0$ as:

$$P(A \mid B) = \frac{P(A \cap B)}{p(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$
$$\text{for } P(B) > 0$$

Notice that $P(\cdot|B)$ is just another probability measure on $\mathcal{F}$. Whats more, $P(\cdot|\Omega) = P$! We can define $P(\cdot|B)$ for $P(B) = 0$, but must be careful when we do so or we will get results like the Borel-Kolmogorov paradox! A set of events $\{A_i\}$ are **Mutually Independent** if, for any subset of $\{A_j\}_{j\in k}$:

$$P\left(\bigcap_{j\in k} A_j\right) = \prod_{j\in k} P(A_j)$$

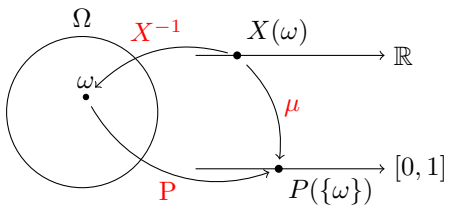
Finally, we define the **Probability Space** as the triple (Ω, F, P) as defined in the previous section. An example of such a space for a finite Ω is as follows: take $\mathcal{F} = 2^\Omega$ and use the **Counting Measure**: $P(E) = \frac{|E|}{|\Omega|}$. For $\Omega = \mathbb{R}$ we use the $\mathbb{B}$: the **Borel σ-algebra**. This consists of all intervals of the form $(-\infty, t],\ t \in \mathbb{R}$ along with their countable unions, countable intersections and complements. We will see in the next section how to define a probability for this.

2.3.1 Random Variables and Vectors

A **Random Variable** (RV) is a $\mathbb{B}$-Measurable function $X : (\Omega, \mathcal{F}) \mapsto (\mathbb{R}, \mathbb{B})$[3]. We can use X to transform questions about arbitrary sample spaces into questions about numbers, which Statisticians like more. For a **Measurable** set A (meaning $A \in \mathbb{B}$) we can compute $P(X \in A)$ as:

$$P(X \in A) := P(X^{-1}(A)) = P(\{\omega \in \Omega : X(\omega) \in A\})$$

Where $P(X^{-1}(\cdot)) := \mu(\cdot)$ is called the **Distribution** of X. By construction, X induces a new Probability Space $(\mathbb{R}, \mathbb{B}, \mu)$ from the original $(\Omega, \mathcal{F}, P)$. This means that all the probability axioms are satisfied for this new triple. We provide a simple example in the picture below for reference. Notice the event is the singleton set $\{\omega\}$:



[3]Note the measurability condition is to ensure that the CDF is well formed

We define the **CDF** for an RV as:

$$F_X(t) := \mu\left((-\infty, t]\right) = P(X \leq t)$$

They always satisfy the folling properties:

1. Right Continuous
2. Non-Negative
3. $\lim_{t\to\infty} F_X(t) = 1$, $\lim_{t\to-\infty} F_X(t) = 0$

The RV X is actually completely specified by F_X! Furthermore, any function satisfying the above axioms is the CDF for some RV. This means that, for any such F_X, $(\mathbb{R}, \mathbb{B}, F_X)$ is a probability triple.

If F_X is **Absolutely Continuous**: it is differentiable a.e. and $\exists f_X(x)$ s.t. $F_X(x) = \int_{-\infty}^{x} f_X(u)du$; then we call X a **Continuous** RV. We have that $\frac{d}{dx}F_X(x) = f_X(x)$ wherever F_X is differentiable, and we call f_X the **PDF**. If $X(\Omega)$ is countable then we call X Discrete. Unlike in the continuous case, $f_X(x) := P(\{X = x\})$ and is called the **PMF**. We often just write $p(x) := f_X(x)$ in this case. The CDF here would be $F_X(t) = \sum_{i=0}^{t} f_X(i)$[4].

We can define a probability density over a vector of random variables too (we consider only cases where every RV is continous or every RV is discrete). We define the **Joint CDF** for $X = (X_1, X_2, ..., X_n)$ as:

$$F_X(t_1, ..., t_n) = P(X_1 \leq t_1, ..., X_n \leq t_n)$$

The **Marginal PDF** of $X_{1:p} = (X_1, \cdots X_p)$ is:

$$f_{X_{1:p}}\left(t_{1:p}\right) = \int_{X_{(p+1):n}} f_X\left(t_{1:p}, X_{(p+1):n}\right) dX_{(p+1):n}$$

The **Conditional PDF** on $X_{1:p}$ given $X_{(p+1):n} = t_{(p+1):n}$ is:

$$f_{X_{1:p}|X_{(p+1):n}}\left(t_{1:p}, t_{(p+1):n}\right) = \frac{f_X\left(t_{1:p}, t_{(p+1):n}\right)}{f_{X_{(p+1):n}}\left(t_{(p+1):n}\right)}$$

Where $f_{X_{(p+1):n}}\left(t_{(p+1):n}\right) > 0$

We can define these analogously for the discrete case by replacing the integral with the appropriate sum.

[4]RVs can be either Continuous, Discrete, Mixed or Singular

Random Variables can be **Independent** too. For continuous or discrete RVs X, Y, Z we say that X is **Independent** of Y: $X \perp Y$ iff

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \ \forall x, y$$

We say that X is **Conditionally Independent** of Y given Z: $X \perp Y \mid Z$ iff

$$f_{X,Y|Z}(x,y|z) = f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z) \ \forall x, y, z$$

For continuous or discrete RVs X, Y, Z, W, we have the following useful properties:

- *Symmetry* $X \perp Y \mid Z \Rightarrow Y \perp X \mid Z$
- *Decomposition* $X \perp Y, W \mid Z \Rightarrow \begin{cases} X \perp Y \mid Z \\ X \perp W \mid Y \end{cases}$
- *Weak Union* $X \perp Y, W \mid Z \Rightarrow \begin{cases} X \perp Y \mid Z, W \\ X \perp W \mid Z, Y \end{cases}$
- *Contraction* $\begin{cases} X \perp Y \mid Z, W \\ X \perp W \mid Z \end{cases} \Rightarrow X \perp Y, W \mid Z$

2.3.2 Moments of a Random Variable

We denote $\mathbb{E}_X(X^r) := \mathbb{E}(X^r)$ as the **rth Moment** of X under the distribution of X. Note that moments need not exist (i.e. $E(|X^r|) = \pm\infty$). We define the first moment as $\mathbb{E}(X) = \int_0^\infty 1 - F_X(t)\, dt - \int_{-\infty}^0 F_X(t)\, dt$. If X Continuous, then this simplifies to $\mathbb{E}(X) = \int_{-\infty}^\infty t \cdot f_X(t)\, dt$. If X discrete, we just replace the integral with a sum. We now state two important and surprising results:

Theorem 2.16 (Law of the Unconscious Statistician).

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(t) \cdot f_X(t)\, dt$$

Theorem 2.17 (Jensens Inequality). *For a convex function g*

$$g(\mathbb{E}\{X\}) \leq \mathbb{E}\{g(X)\}$$

We can generate Moments using the **MGF** of X: $M_X(t) = \mathbb{E}\left(\exp(Xt)\right)$, which exists when $\exists \epsilon > 0$ s.t. $\forall |t| < \epsilon$, $M_X(t) < \infty$. We present some results:

1. $\exists \epsilon > 0$ s.t. $\forall |t| < \epsilon$, $M_X(t) = M_Y(t) \Rightarrow X$ and Y have same distribution
2. $\mathbb{E}(|X^r|) = \frac{\partial^r}{\partial^r t} M_X(t)\big|_{t=0}$, if M_X exists.
3. If $\{X_i\}$ independent RVs, then $M_{\sum X_i}(t) = \prod M_{X_i}(t)$

The two moments most commonly analyzed are:

1. **Mean** of X: $\mathbb{E}(X) := \mu_X$
2. **Variance** of X: $Var(X) = \mathbb{E}((X - \mu_X)^2) = \sigma_X^2$

For random vectors X we have:

1. $\mathbb{E}\{X\} = \begin{bmatrix} \mathbb{E}\{X_1\} \\ \vdots \\ \mathbb{E}\{X_n\} \end{bmatrix} = \mu$
2. $Cov(X) = \mathbb{E}[(X - \mu)(X - \mu)^T] = \Sigma$

2.3.3 Frequentistism vs Bayesianism

We have described Probability theory rigorously but haven't provided any semantics. In the world of Probability and Statistics there are two competing interpretations for what the probability of an Event really *means*:

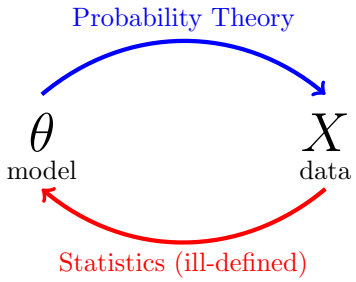
1. **Frequentists**: the *limiting frequency* of the event
2. **Bayesians**: the *reasonable expectation* that the event occurs

The *reasonable expectation* can be further broken down into two views. The **Objective Bayesians** view the *reasonable expectation* as the *state of knowledge*. They view probability as an extension of propositional logic, which is described in [4]. The **Subjective Bayesians** view probability as a quantification of *personal belief*[5]. These competing interpretations lead to very different ways of modelling and analyzing data, which we will see in the next section.

[5]The main difference between the groups is in how they choose their priors: the Subjective Bayesians use knowledge about or prior experience with model parameters, wheras the Objectivists try to introduce as little prior knowledge as possible, using noninformative priors

2.4 Basic Statistics

Before we describe the requisite Statistical ideas needed for Machine Learning, we will provide some definitions. We say **Probability** quantifies uncertainty of data given a model. We have seen that probabilistic questions are fairly well-formed mathematical problems. **Statistics** is the inverse problem: given data, how likely is model? Unlike with Probability theory, these questions tend to be ill-formed since many models can generate the same data!



Statistical Inference is the process of making propositions about the underlying PDF/PMF using data. The problem begins by determining the general form of the PDF/PMF. This is challenging since there are infinitely many! Often we break this problem up by first choosing a **Parametric Family**: $\{p_\theta\}_{\theta \in \Theta}$[6]. This is a set of PDFs/PMFs indexed by some parameter(s) θ. We call Θ the **Parametric Space** and require for it to be finite dimensional[7]. For a concrete example, lets recall the coin tossing scenerio. We could model this using the **Binomial Distribution**. We write $X \sim Bin(N, \theta)$ if:

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

Where dom(X)= $\{0, \cdots, N\}$, $\Theta = [0, 1]$. Semantically, this family is used to model the probability of k successes in N IID trials, where the probability of a success in any single trial is θ. For our example, if we denote X as the number of heads, then $X \sim Bin(4, \frac{1}{2})$. There are many Parametric Families; each of which can be used to capture different aspects of the data. Choosing the best Parametric Family is a matter of choosing the best representation for the data. Typically we don't know θ, and have to infer it from the data. In Statistics this is called **Point Estimation**[8]!

[6] where $p_\theta = f_X(\cdot|\theta)$

[7] If Θ is infinite dimensional, we call the corresponding model **Non Parametric**

[8] We could also provide a **Confidence Interval** or **Credible Interval**; but those options are not typically used in ML

2.4.1 Point Estimation

We want to use data $D = \{X_1, \cdots, X_n\}$ to estimate an unknown parameter θ. We define a **Statistic** T to be any function of D which does not depend on θ. Notice T is a RV and should contain as much information about θ as D does. A Statistic T which contains no information about T is called **Ancillary**. Formally, T is Ancillary if $T(D) \perp \theta$. On the other extreme, T is **Sufficient** if $D \perp \theta \mid T(D)$. In this case T contains all the information about θ. We can identify Sufficiency using the following theorem:

Theorem 2.18 (Neyman Factorization Critereon). *T is sufficient for θ if and only if*

$$f(D|\theta) = g(T(D), \theta)h(D)$$

We next describe a general principle which will allow us to estimate it from some data! The **Likelihood Principle** says that all the evidence in a sample relevent to parameters θ is contained in the likelihood function. The likelihood is defined as follows:

$$\mathcal{L}(\theta|D) = P(D|\theta)$$

If you have trouble accepting the Likelihood Principle, it has been shown to be equivalent to two milder principles:

1. **Sufficiency Principle**: If two different observations x, y are such that $T(x) = T(y)$ for a sufficient statistic T, then inference based on x and y should be the same.

2. **Conditionality Principle**: If an experiment concerning inference about θ is chosen from a collection of possible experiments independently, then any experiment not chosen is irrelevant to the inference.

Of these principles, Sufficiency is accepted by both Frequentists and Bayesians, while the Conditionality principle is debated.

This principle is the guiding idea behind **Maximum Likelihood Estimation** (MLE). The basic idea is as follows: Since the likelihood contains all information relevent to θ, we can find the most probable θ simply by maximizing the likelihood using the optimization techniques described in 2.2.

We can take this one step further: supposing that θ is an RV with its own PDF/PMF. We can use the laws of probability to get that:

$$\begin{aligned} P(\theta|D) &= \frac{P(D|\theta)P(\theta)}{P(X)} \\ &\propto \mathcal{L}(\theta|D)P(\theta) \end{aligned}$$

Where

- $P(\theta|D)$ is the **Posterior Distribution**. The probability of θ *after* observing data
- $P(D|\theta)$ is the Likelihood
- $P(\theta)$ is the **Prior Distribution** of θ, which can be a parametric model itself (whose parameters would be referred to as **Hyperparameters**)
- $P(D)$ is the **Marginal Likelihood**. This is to ensure that the distribution is valid (i.e. sums/integrates to 1)

So if instead of maximizing $\mathcal{L}$ we maximize the posterior, we get the **Maximum A Posteriori Estimate** (MAP). Notice that if the Prior is uniform then the MAP is equivalent to the MLE. We need a theoretical justification for why we assume the existence of a prior distribution on θ in the first place! The justification for this requires the **Infinite Exchangeable** assumption. This is satisfied if; given a random infinite sequence of RVs $\{X_i\}_{i=1}^{\infty}$, any finite subset $\{X_j\}_{j=1}^{n}$, and any permutation of this subset $\pi_{1:n}$, we have that:

$$P(X_1, \cdots, X_n) = P(X_{\pi_1}, \cdots, X_{\pi_n}) \tag{9}$$

It turns out the above is equivalent to assuming the existence of the prior! The following theorem makes this precise.

Theorem 2.19 (De Finetti Theorem). *A sequence is Infinite Exchangeable iff for any n*

$$P(X_1, \cdots X_n) = \int \prod_{i=1}^{n} P(X_i|\theta) d\mu(\theta)$$

for some measure μ on θ. Also, if θ has a density (e.g. is discrete or continuous) then $d\mu(\theta) = p(\theta)d\theta$. Note: θ may be infinite!

This theory says that, if we assume exchangable data (and iid $\Rightarrow$ exchangeable), then there must exist a θ, $p(X|\theta)$ and distribution μ on θ! So the idea of having a prior distribution on the parameters does have theory to back it up!

2.4.2 The Multinomial Distribution

Suppose instead of flipping a coin we were to roll a die instead. We can model the process of observing one of $K \geq 2$ possible outcomes using the **Multinoulli**

Distribution. Here, X is a discrete RV over K choices. We encode X as a **one-hot encoding**: a random vector taking values in the unit bases in $\mathbb{R}^K$

i.e. $\text{dom}(X) = \{e_1, e_2, \ldots, e_K\}$ where $e_j = \Big(0 \ldots \underset{\substack{\uparrow \\ j^{\text{th}} \text{ coordinate}}}{1} \ldots 0\Big)^T \in \mathbb{R}^K$

$\Theta = \Delta_K$ is the **Probability Simplex** on K choices, and is given by:

$$\Delta_K = \left\{ \pi \in \mathbb{R}^K \ ; \ \forall j \ \pi_j \geq 0 \text{ and } \sum_{j=1}^{K} \pi_j = 1 \right\}$$

Our PMF has the following form:

$$p(X|\pi) = \prod_{j=1}^{K} \pi_j^{X_j}$$

where $X_j \in \{0, 1\}$ is the j^{th} component of X. If we were to roll a die multiple times, we could model the result using the **Multinomial Distribution**. This is just $X = \sum_{i=1}^{N} X_i$, where each X_i are IID multinoulli with same parameter $\pi \in \Delta_K$. We then have that:

$$p(X|\pi) = \frac{N!}{\prod_{j=1}^{K} n_j} \prod_{j=1}^{K} \pi_j^{n_j}$$

$$\text{dom}(X) = \left\{ (n_1, \ldots, n_K) \ ; \ \forall j \ n_j \in \mathbb{N} \text{ and } \sum_{j=1}^{K} n_j = N \right\}$$

$$\text{where } n_j = \sum_{i=1}^{N} X_{i,j}, \text{ with } X_{i,j} \text{ being the } j^{\text{th}} \text{ component of } X_i$$

Notice that the n_j's are a sufficient statistic for X. Now suppose we have some data $X_1, \cdots X_n$ from a Multinoulli model and we want to estimate the π_j's using MLE. Instead of maximizing the Likelihood we can maximize the log likelihood since it is monotonic function. We observe that:

$$\log \mathcal{L}(\pi|x_1, \cdots x_n) \propto \sum_{j=1}^{K} n_j \log \pi_j$$

And we ignore the normalizing constant since it wont affect the maximum of $\log \mathcal{L}$. To ensure that $\pi \in \Delta_K$ we must introduce the constraint $\sum_{j=1}^{K} \pi_j = 1$. The Lagrangian is

$$L(\pi, \lambda) = \sum_{j=1}^{K} n_j \log \pi_j + \lambda \left(1 - \sum_{j=1}^{K} \pi_j \right)$$

Notice that $\nabla_\pi L(\pi, \lambda) = 0$ gives us a stationary point where the j^{th} cooridinate is $\frac{n_j}{\lambda}$, and$\nabla_\lambda L(\pi, \lambda) = 0 \Rightarrow \lambda = N$. Since L is concave being a stationary point is a sufficient condition for being a maximum. Therefore:

$$\boxed{\hat{\pi}_j^{\text{ML}} = \frac{n_j}{N}}$$

We can do MAP inference on this as well. We suppose that π comes from a **Dirichlet Distribution**, that is:

$$P(\pi|\alpha) = \frac{\Gamma\left(\sum_{j=1}^k \alpha_j\right)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k \pi_j^{\alpha_j - 1}$$
$$\text{where } \text{dom}(\pi) = \Delta_K$$

The Dirichlet Distribution is **Conjugate** to the Multinomial Distribution. This means that the Prior and Posterior are from the same parametric family. We can verify this:

$$\begin{aligned} P(\theta|X) &\propto P(X|\theta)P(\theta) \\ &= \prod_{j=1}^k \pi_j^{n_j} \prod_{j=1}^k \pi_j^{\alpha_j - 1} = \prod_{j=1}^k \pi_j^{n_j + \alpha_j - 1} \end{aligned}$$

We recognize the above as the unnormalized density of a Dirichlet Random Variable, and so $\pi \mid X \sim Dir\left(\{n_j + \alpha_j\}_{j=1}^k\right)$. We maximize over this to compute the MAP estimate. We assume that $\alpha_j > 1 \ \forall j$. We can look for stationary points of the Lagrangian L and find that

$$\begin{aligned} \nabla_{\pi_j} L(\pi, \lambda) &= \frac{n_j + \alpha_j - 1}{\pi_j} - \lambda = 0 \\ &\Rightarrow \pi_j = \frac{n_j + \alpha_j - 1}{\lambda} \end{aligned}$$

And

$$\begin{aligned} \nabla_\lambda L(\pi, \lambda) &= 1 - \sum_{i=1}^k \pi_i = 0 \\ &\Rightarrow 1 - \sum_{i=1}^k \frac{n_i + \alpha_i - 1}{\lambda} = 0 \\ &\Rightarrow \lambda = \sum_{i=1}^k n_i + \alpha_i - 1 = N + \sum_{i=1}^k (\alpha_i - 1) \end{aligned}$$

Putting this together, and assuming that the determinent of the Hessian at this point is negative, we get that:

$$\boxed{\hat{\pi}_j^{\text{MAP}} = \frac{n_j + \alpha_j - 1}{N + \sum_{i=1}^k (\alpha_i - 1)}}$$

We can see that the prior adds "pseudocounts" to the estimator, which makes it more robust in cases where the true π_j is small. In these cases $\hat{\pi}_j^{\text{MLE}}$ usually is 0, but the pseudocounts will ensure that $\hat{\pi}_j^{\text{MAP}}$ isn't exactly 0.

2.4.3 The Normal Distribution

The most popular model for unbounded continuous data, is the **Normal Distribution**. We say $X \sim \mathcal{N}_p(\mu, \Sigma)$ if:

$$p(x|\mu, \Sigma) = (2\pi)^{-\frac{p}{2}} \det \Sigma^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

For $x, \mu \in \mathbb{R}^p$, and $\Sigma \in \mathbb{R}^{p \times p}$ Symmetric and Positive Definite.

2.4.4 The Empirical Density Function

We now describe a the simplest representation of our data. Suppose we are given some data $x_1, \cdots x_n \sim F$ from an unknown distribution F, which we want to approximate. We define the **Empirical Distribution** $\hat{F}$ of the data as:

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq t\}} \tag{10}$$

It can be shown that $\hat{F}(t) \to F(t)$ *a.s.* $\forall t$, justifying its use as an approximation of F, provided enough data has been observed. As with F we can approximate f. We define the **Empirical Density Function** $\hat{f}$:

$$\hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i, t) \tag{11}$$

Where δ is defined differently in the continuous and discrete case. In the continuous case it is called the **Dirac Delta Function**:

$$\delta(x, y) = \begin{cases} \infty & x = y \\ 0 & \text{o.w.} \end{cases} \tag{12}$$

Additionally, we suppose that:

1. $\int_{-\infty}^{\infty} \delta(t, y) dt = 1$
2. $\int \delta(t, y) f(t) dt = f(y)$, for any f with compact support that is continuous around y

This is not a function, but is called a *Generalized Function*. In the discrete case things are much simpler, as we can use the simpler **Kronecker delta function**:

$$\delta(x, y) = \begin{cases} 1 & x = y \\ 0 & \text{o.w.} \end{cases} \tag{13}$$

Finally, we notice that $\hat{f}$ and $\hat{F}$ satisfy an important relationship that would be expected from the cdf and pdf: $\int_{-\infty}^{t} \hat{f}(y)dy = \hat{F}(t)$.

$$\begin{aligned}
\int_{-\infty}^{t} \hat{f}(y)dy &= \int_{-\infty}^{t} \frac{1}{n}\sum_{i=1}^{n} \delta(x_i, y)dy \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{t} \delta(x_i, y)dy \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \mathbb{1}_{\{x_i \leq y\}} \delta(x_i, y)dy \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq y\}} \\
&= \hat{F}(t)
\end{aligned}$$

2.5 Statistical Decision Theory

In the next chapter we will describe a general framework for how to make data-driven decisions under uncertainty: **Statistical Decision Theory**.

We observe some **Data** $D \in \mathcal{D}$, which comes from some **Data Generating Distribution** $D \sim p$. Let $p \in \mathcal{P}$[9], where $\mathcal{P}$ is the set of possible distributions. Let $\mathcal{A}$ be our set of possible actions. To determine how good an action is, we define the **Loss** (cost) of doing that action as $L : \mathcal{P} \times \mathcal{A} \mapsto \mathbb{R}$. The goal is to determine a **Decision Rule** $\delta : \mathcal{D} \mapsto \mathcal{A}$ which, given data, produces an action.

Typically we consider $\mathcal{P}$ as a Parametric Family of distributions, and we use Θ interchangibly with $\mathcal{P}$, using that $p := p_\theta$. We need a way to assign a value to any δ and a way to compare these values to find which one is "best". One such way of doing this is via the Frequentist Risk.

2.5.1 Frequentist Risk Perspective

The first approach seeks to minimize the **Frequentist Risk**, which is defined as:

$$R(P, \delta) = \mathbb{E}_{D\sim p}\{L(p, \delta(D))\} \tag{14}$$

If we want to compare decision rules δ_1, δ_2 using def 14, we have to take into account p, since R varies with both p and δ. Sometimes one decision rule δ_1 is better then another δ_2 regardless of p, in which case we say δ_1 **Dominates** δ_2. More formally:

$$R(p, \delta_1) \leq R(p, \delta_2)\ \forall p \in \mathcal{P}\ and$$
$$\exists p \in \mathcal{P},\ R(p, \delta_1) < R(p, \delta_2)$$

This basically means that δ_1 is a better decision rule then δ_2. Sometimes, there may be a "best" δ, one which isn't dominated by any other δ_0. We say δ is **Admissible** if $\nexists \delta_0$ s.t. δ_0 dominates δ. Note: we should rule out inadmissible decision rules (except for simplicity or efficiency) but not necessarily accept Admissible ones!

Unfortunately, different p's usually produce different optimal δ's! we must take into account the unknown p when minimizing (14). One way to take this into account is to use the **Minimax Criteria**: the optimal δ minimizes the Frequentist Risk in worst case scenerio.

$$\delta_{minimax} = \min_{\delta} \max_{p\in\mathcal{P}}\ R(p, \delta) \tag{15}$$

If $\mathcal{P}$ is a Parameteric Family, we can handle the dependence of R on p by averaging it out, adding weights π over Θ to put more weight on certain θ's. We can then minimize over δ. This is called the **Bayes Risk**, even though it is Frequentist concept since it averages over D via (14).

$$\delta_{bayes} = \arg\min_{\delta} \int_{\Theta} R(p_\theta, \delta)\pi(\theta)d\theta \tag{16}$$

[9]Often p will describe an IID process, e.g. $D = (X_1, ..., X_n)$ where $X_i \overset{iid}{\sim} P_0$. In this case, the loss is usually written w.r.t p_0 instead of p.

Where δ_{bayes} is called the **Bayes Rule**. Note that the Bayes Rule may not exist, and when they do they may not be unique.

2.5.2 Bayesian Risk Perspective

Note that (14) does not consider that we only observed one D. We can define a Risk function that does. The **Posterior Risk** is

$$R_B(\delta|D) = \int_\Theta L(P_\theta, \delta)p(\theta|D)d\theta \tag{17}$$

Where $p(\theta|D)$ is the posterior for a given prior $\pi(\theta)$. We can choose our decision rule based on this new risk function. This is called the **Bayes Estimator** or **Bayes Action** (not to be confused with the **Bayes Rule** above).

$$\delta_{post} = \arg\min_\delta R_B(\delta|D) \tag{18}$$

Notice that in (17), we do not consider different unobserved values of D, since the Bayesian would say they are irrelevent courtesy of the Conditionality Principle. For them, only the observed D matters for inference. Additionally, θ is integrated out in (17), meaning that (18) gives the undisputed optimal δ!

Note that the Frequentist can still use (18) by interpreting it as (16) with π as the "true" prior for Θ. We would then get that:

$$\begin{aligned}\int_\Theta R(p_\theta, \delta)\pi(\theta)d\theta &= \int_\Theta \int_D L(p_\theta, \delta)p(D|\theta)p(\theta)dDd\theta \\ &\overset{(a)}{=} \int_D \int_\Theta L(p_\theta, \delta)p(\theta|D)p(D)d\theta dD \\ &= \int_D R_B(\delta|D)p(D)dD\end{aligned}$$

Where (a) is due to Fubini's theorem (provided the integral is finite). It turns out that a *Bayes rule* can be obtained by taking the *Bayes action* for each particular D! See [5] for more details.

2.6 Types of Procedures

2.6.1 Parameter Estimation

Given a Parametric Family $\{p_\theta\}_{\theta\in\Theta}$, typically we have data $D = (X^{(1)}, \cdots, X^{(n)})$ where each $X^{(i)} \overset{iid}{\sim} p_\theta$. We want to use this data to estimate the true parameters θ. Hence, $\mathcal{A} = \Theta$ and $\delta(D)$ is some an **Estimator** of θ. The estimator should minimize the loss (more specifically the risk). One popular loss function is the **Squared Loss**: $L(\theta, \delta(D)) = ||\theta - \delta(D)||^2$. Note that since the data are IID we use the marginal density over X instead of the joint over D in the loss function.

If we take the expectation of the loss function above (the frequentist risk), we can decompose it nicely into two pieces:

$$\begin{aligned} R(P,\delta) &= \mathbb{E}_{D\sim p}\{||\theta - \delta(D)||^2\} \\ &= \mathbb{E}_{D\sim p}\{(\theta - \mathbb{E}_{D\sim p}\{\delta(D)\} + \mathbb{E}_{D\sim p}\{\delta(D)\} - \delta(D))^2\} \\ &= \underbrace{(\theta - \mathbb{E}_{D\sim p}\{\delta(D)\})^2}_{Bias^2} + \underbrace{\mathbb{E}_{D\sim p}\{(\mathbb{E}_{D\sim p}\{\delta(D)\} - \delta(D))^2\}}_{Variance} \end{aligned}$$

The above says that when we average this loss over all possible datasets, we can compare how much of the loss is due to the Bias and how much to the Variance of the estimator δ. This idea works for other loss functions, but the decomposition is not nearly as clean. Finally, this is a Frequentist idea since it involves taking an expectation over the data generating distribution, an idea doesn't appeal to Bayesians since it is contrary to the conditionality principle.

We conclude by showing an interesting result: for parameter estimation, the bayes action for the squared loss $\delta_{post}(D) = \mathbb{E}\{\theta|D\}$. This is a simple optimization problem:

$$\begin{aligned} R_B(\delta|D) &= \int_\Theta ||\theta - \delta(D)||^2 p(\theta|D)d\theta \\ &= \delta(D)^2 - 2\delta(D)\int_\Theta \theta p(\theta|D)d\theta + \int_\Theta \theta^2 p(\theta|D)d\theta \end{aligned}$$

and taking the derivative and setting to 0 yields:

$$\begin{aligned} \frac{\partial R_B}{\partial \delta} &= 2\delta(D) - 2\int_\Theta \theta p(\theta|D)d\theta = 0 \\ \Rightarrow \delta(D) &= \int_\Theta \theta p(\theta|D)d\theta = \mathbb{E}\{\theta|D\} \end{aligned}$$

2.6.2 Prediction

Let $D = ((X^{(1)}, Y^{(1)}), \cdots, (X^{(n)}, Y^{(n)}))$ where $X^{(i)} \in \mathcal{X}$ and $Y^{(i)} \in \mathcal{Y}$.

We put a density on X and Y: $(X^{(i)}, Y^{(i)}) \overset{iid}{\sim} P_{XY}$. Our action space $\mathcal{A} = \mathcal{Y}^{\mathcal{X}}$. Hence $\delta(D)$ is a **Learning Algorithm** which learns a function i.e. $\delta(D) = \hat{f}$.

We can evaluate the performance of f using a *prediction loss* $l : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$, a measure of the distance between a given prediction and its associated ground truth. We define the **Generalization Error** from l as:

$$L(P, f) = \mathbb{E}_{(X,Y)\sim P_{XY}}\{l(Y, f(X))\} \tag{19}$$

This is often called the Risk in Machine Learning. Note that we do not know (19), but we can approximate it using the below formula. This is called **Empirical Risk Minimization**

$$L(P, f) = \frac{1}{n}\sum_{i=1}^{n} l\left(Y^{(i)}, f(X^{(i)})\right) \tag{20}$$

Notice that (20) is just (19) with the empirical density substituted in place of the true, unknown one.

Prediction problems are called different things depending on whether $\mathcal{Y}$ is discrete or continuous. If $\mathcal{Y}$ discrete then the problem is called **Classification**, and if $\mathcal{Y}$ is not discrete e.g. $\mathcal{Y} = \mathbb{R}$, then it is referred to as **Regression**.

We previously described Prediction problems and noted that if $\mathcal{Y}$ is discrete then we refer to prediction as **Classification** and otherwise we refer to it as **Regression**.

2.6.3 Regression

We look at Regression. Let $D = ((X^{(1)}, T^{(1)}), \cdots, (X^{(n)}, T^{(n)}))$ with $X^{(i)} \in \mathcal{X}$ and $T^{(i)} \in \mathcal{T}$. Let $(X^{(i)}, T^{(i)}) \overset{iid}{\sim} P_{XT}$ and $l(t, y(x)) = |t - y(x)|^2$ (the squared loss). We need to find a function $y : \mathcal{X} \to \mathcal{T}$ which minimizes the Generalization Error:

$$\begin{aligned} L(P, y) &= \mathbb{E}_{(X,T)\sim P_{XT}}\{|t - y(x)|^2\} \\ &= \int\int |t - y(x)|^2 p(x, t) dx dt \end{aligned}$$

In this case we can find the optimal y by using calculus of variations. That is to say, the problem is reduced to an optimization problem. We denote $G(y, y', x) = \int |t - y(x)|^2 p(x, t) dt$ and use the Euler Lagrange equations to get that our stationary point must occur at

$$\begin{aligned} &\frac{\partial G(y, y', x)}{\partial y} - \frac{d}{dx}\frac{\partial G(y, y', x)}{\partial y'} = 0 \\ &\Rightarrow \frac{\partial G(y, y', x)}{\partial y} = 0 \end{aligned}$$

Since $\frac{\partial G(y,y',x)}{\partial y'} = 0$ since y' is not in G. We then solve:

$$\begin{aligned}\frac{\partial L(P,y)}{\partial y(x)} &= \frac{\partial}{\partial y(x)} \int |t - y(x)|^2 p(x,t) dt \\ &= 2 \int (t - y(x)) p(x,t) dt = 0\end{aligned}$$

Solving for the above we have that

$$\begin{aligned}\int (t - y(x)) p(x,t) dt = \int t p(x,t) dt - \int y(x) p(x,t) dt = 0 \\ \Rightarrow \int t p(x,t) dt = y(x) p(x) \\ \Rightarrow y(x) = \int t \frac{p(x,t)}{p(x)} dt = \mathbb{E}_{t \sim p(t|x)}\{t|x\}\end{aligned}$$

And so our learning algorithm $\delta(D)$ simply returns $y(x) = \mathbb{E}_{t \sim p(t|x)}\{t|x\}$. Keep in mind that we don't know $p(t|x)$ yet! We now look at a generalization of squared loss function – a family of loss functions called the **Minkowski Loss**. This family has the following form:

$$L_q(P,y) = \int \int |t - y(x)|^q p(x,t) dx dt \tag{21}$$

We solve for the optimal $y(x)$ and set this to 0:

$$\frac{\partial L_q(P,y)}{\partial y(x)} = \int q |t - y(x)|^{q-1} sgn(t - y(x)) p(x,t) dt \tag{22}$$

$$= \int_{y(x)}^{\infty} q |t - y(x)|^{q-1} p(x,t) dt - \int_{-\infty}^{y(x)} q |t - y(x)|^{q-1} p(x,t) dt \tag{23}$$

$$\Rightarrow \int_{-\infty}^{y(x)} |t - y(x)|^{q-1} p(x,t) dt = \int_{y(x)}^{\infty} |t - y(x)|^{q-1} p(x,t) dt \tag{24}$$

For $q = 1$, we see that $y(x)$ is the conditional median of t.

$$\int_{-\infty}^{y(x)} p(x,t) dt = \int_{y(x)}^{\infty} p(x,t) dt \tag{25}$$

Finally, as $q \to 0$, the $y(x)$ given by the Minkowski loss is the conditional mode of t. Notice again we need to know the underlying data generating pdf i.e. $p(x,t)$. Determining this is called **Inference** and will be dealt with later.

3 Representation

3.1 Exponential Family

The **(Canonical) Exponential Family** is a parametric family of distributions which have the following form:

$$p(x|\eta) = \exp\{\eta^T T(x) - A(\eta)\}h(x) \tag{26}$$

Where:

1. $h(x)d\mu(x)$ is the **Reference Measure** on X
   (a) $h(x)$ is the **Reference Density** $\rightarrow$ defines the support and must not depend on η!
   (b) $d\mu(x)$ is the **Base Measure**
      - the Counting measure for discrete $\mathcal{X}$
      - the Lebesgue measure for continuous $\mathcal{X}$
2. $T : \mathcal{X} \rightarrow \mathbb{R}^p \rightarrow$ the **Sufficient Statistics** $\rightarrow$ functions of x that fully summarizes x within the density function
3. η is called the **Canonical Parameter**
4. $A(\eta)$ is the **Cumulant Function** $\rightarrow$ ensures that the density sums/integrates to one

Note that any member of the exponential family is fully specified by 1 and 2. $A(\eta)$ is dependent on the choice of 1 and 2, and so is not chosen. We can see this by the following calculation:

$$\begin{aligned}
1 = \int_x p(x|\nu)d\mu(x) &= \int_x \exp\{\eta^T T(x)\}e^{-A(\eta)}h(x)d\mu(x) \\
&= e^{-A(\eta)} \int_x \exp\{\eta^T T(x)\}h(x)d\mu(x) \\
&\Rightarrow A(\eta) = \log \int_x \exp\{\eta^T T(x)\}h(x)d\mu(x) \\
&\Rightarrow A(\eta) = \log Z(\eta)
\end{aligned}$$

Where $Z(\eta)$ is called the **Partition Function**. Since A is a function of η, we must restrict η to ensure that $p(x|\eta)$ is well defined. We let $\Omega = \{\eta \in \mathbb{R}^p | A(\eta) < \infty\}$ and call this the **Natural Parameter Space**. Members of the Exponential family (sets of $h(x)d\mu(x)$ and $T(x)$) with non-empty, open Ω are called **Regular**. We are interested in these members since they have valid pdfs.

We are also interested in **Minimal** exponential families. These are families which contain non-redundant η's and $T(x)$'s. What we mean by this is that neither have any affine equality constraints:

1. $\nexists$ non-zero a, b s.t. $a^T T(x) + b = 0 \ \forall x \ s.t. \ h(x) = 0$
2. $\nexists$ non-zero c, d s.t. $c^T \eta + d = 0 \ \forall x \ s.t. \ h(x) = 0$

More generally, given an open connected subset $\Theta \in \mathbb{R}^p$ and mapping $\eta : \Theta \to \Omega$, we can write this as:

$$p(x|\theta) = p(x|\eta(\theta)) = \exp\{\eta(\theta)^T T(x) - A(\eta(\theta))\}h(x) \tag{27}$$

If the Jacobian of η is not full rank, then we call this a **Curved Exponential Family**.

3.1.1 Properties of Exponential Families

For canonical exponential families we have the following results:

Theorem 3.1. $\nabla_\eta A(\eta) = \mathbb{E}\{T(x)\}$

Proof.

$$\begin{aligned}
\nabla_\eta A(\eta) &= \nabla_\eta \log \int_x \exp\{\eta^T T(x)\}h(x)d\mu(x) \\
&= \frac{1}{Z(\eta)} \nabla_\eta \int_x \exp\{\eta^T T(x)\}h(x)d\mu(x) \\
&\overset{(a)}{=} \frac{1}{Z(\eta)} \int_x \nabla_\eta \exp\{\eta^T T(x)\}h(x)d\mu(x) \\
&= \frac{1}{Z(\eta)} \int_x T(x) \exp\{\eta^T T(x)\}h(x)d\mu(x) \\
&\overset{(b)}{=} \int_x T(x) \exp\{\eta^T T(x) - A(\eta)\}h(x)d\mu(x) \\
&= \mathbb{E}\{T(x)\}
\end{aligned}$$

Where (a) follows from the Dominated Convergence Theorem and (b) follows from the defintion of $A(\eta)$ □

Theorem 3.2. $\frac{\partial^2}{\partial \eta_i \partial \eta_j} A(\eta) = Cov\{T_i(x), T_j(x)\}$ *and so* $HA(\eta) = Cov\{T(x)\}$

Proof.

$$\begin{aligned}
\frac{\partial^2}{\partial\eta_i\partial\eta_j}A(\eta) &= \frac{\partial}{\partial\eta_i}\int_x T_j(x)\exp\{\eta^T T(x) - A(\eta)\}h(x)d\mu(x) \\
&= \int_x T_j(x)\frac{\partial}{\partial\eta_i}\exp\{\eta^T T(x) - A(\eta)\}h(x)d\mu(x) \\
&= \int_x T_j(x)\exp\{\eta^T T(x) - A(\eta)\}\left(T_i(x) - \frac{\partial}{\partial\eta_i}A(\eta)\right)h(x)d\mu(x) \\
&= \int_x T_j(x)p(x|\eta)\left(T_i(x) - \mathbb{E}\{T_i(x)\}\right)d\mu(x) \\
&= \int_x T_j(x)T_i(x)p(x|\eta) - T_j(x)p(x|\eta)\mathbb{E}\{T_i(x)\}d\mu(x) \\
&= \mathbb{E}\{T_j(x)T_i(x)\} - \mathbb{E}\{T_j(x)\}\mathbb{E}\{T_i(x)\} \\
&= Cov\{T_i(x), T_j(x)\}
\end{aligned}$$

□

Theorem 3.3. *Ω is a convex set and $A(\eta)$ is a convex function. If the family is minimal then $A(\eta)$ is strictly convex.*

Proof. Since $HA(\eta) = Cov\{T(x)\}$ is always positive semi-definite, we have that $A(\eta)$ is convex. Since Ω is the epigraph of A, it follows that Ω is a convex set. Lastly, we show strict convexity whenever an exponential family is minimal. This follows since, for any $a \neq 0$ $a^T T(x)$ is not constant, and so $Cov\{a^T T(x)\} \neq 0$. Since $Cov\{a^T T(x)\} \geq 0$ from positive semi definiteness, we have that $Cov\{a^T T(x)\} > 0$. Then:

$$Cov\{a^T T(x)\} = a^T Cov\{T(x)\}a = a^T HA(\eta)a > 0$$

And so $HA(\eta)$ positive definite and therefore is strictly convex. □

3.1.2 Estimation in the Exponential Family

Given an IID sample $X_1, \cdots, X_n \sim p(X|\eta)$ from a Canonical Exponential Family, we have that

$$p(x_1, \cdots, x_n|\eta) = \left(\prod_{i=1}^n h(x_i)\right)\exp\left\{\eta^T\left(\sum_{i=1}^n T(x_i)\right) - nA(\eta)\right\}$$

We see that this is also in the exponential family. Specifically:

1. the new sufficient statistic is $\sum_{i=1}^n T(x_i)$
2. the new reference density is $\prod_{i=1}^n h(x_i)$
3. the new cumulant function is $nA(\eta)$
4. η and Ω remain the same

Notice that the sufficient statistics of this new density are simply sums of the sufficient statistics from $p(x_i|\eta)$ (i.e. $T \in \mathbb{R}^p$ regardless of n). We can compute the log likelihood of $x_1, \cdots, x_n$ as

$$l(\eta|x_1, \cdots, x_n) = \sum_{i=1}^{n} \log h(x_i) + \eta^T \left(\sum_{i=1}^{n} T(x_i) \right) - nA(\eta)$$

Notice that this is a concave funtion, and so it has a global maximum. We show that the MLE estimate for an exponential family is equivalent to Moment Matching. We take the gradient and set it to zero:

$$\begin{aligned} \nabla_\eta l(\eta|x_1, \cdots, x_n) &= \sum_{i=1}^{n} T(x_i) - n\nabla_\eta A(\eta) = 0 \\ &\Rightarrow \frac{1}{n} \sum_{i=1}^{n} T(x_i) = \nabla_\eta A(\eta) \\ &\Rightarrow \frac{1}{n} \sum_{i=1}^{n} T(x_i) = \mathbb{E}\{T(x)\} \end{aligned}$$

3.1.3 Conjugate Priors of the Exponential Family

We note that the exponential family is closed under multiplication but not closed under marginalization. Because of closure under multiplication we can deduce; for every member of the exponential family, a conjugate prior. The prior has the form:

$$p(\eta|\tau, n_0) = \exp\{\tau^T \eta - n_0 A(\eta)\} g(\tau, n_0)$$

and the corresponding posterior is:

$$\begin{aligned} p(\eta|x_1, \cdots, x_n) &= p(x_1, \cdots, x_n|\eta) p(\eta|\tau, n_0) \\ &\propto \exp\left\{ \eta^T \left(\tau + \sum_{i=1}^{n} T(x_i) \right) - (n + n_0) A(\eta) \right\} \end{aligned}$$

Theorem 3.4. $\mathbb{E}_{\eta \sim p(\cdot|\tau, n_0)}\{\mathbb{E}_\eta\{T(x)\}\} = \kappa \frac{\tau}{n_0} + (1 - \kappa) \frac{\sum_{i=1}^{n} T(x_i)}{n}$
Where $\kappa = \frac{n_0}{n_0 + n}$

Proof. Jordan course notes, lecture 4 page 4 [6]. □

4 Information Theory

We want a function I which measures how much information you learn from observing some event E. We want it to satisfy some properties, mainly:

1. Highly probable E have low $I(E)$ and conversely $\rightarrow$ *rare events give more information.*
2. $I(E) \geq 0 \rightarrow$ *Information is non-negative.*
3. if $p(E) = 1$ then $I(E) = 0 \rightarrow$ *Events that always occur provide no information.*
4. If E_1, E_2 are independent events then $I(E_1 \cap E_2) = I(E_1) + I(E_2) \rightarrow$ *information due to independent events are additive.*

From 1. and 3. we see that I should be a function of the probability of an events occurence, i.e. $I(E) = f(p(E))$ for some f. From 4., given independent events E_1, E_2, we have that:

$$f(p(E_1)p(E_2)) = f(p(E_1 \cap E_2)) = f(p(E_1)) + f(p(E_2)) \tag{28}$$
$$f(x \cdot y) = f(x) + f(y) \tag{29}$$

If we assume that I is continuous, then only $I(E) = Klogp(E)$ satisfies (28) [7]. Finally, using 2., we see that $K < 0$. We can then define I as:

$$I(E) = -logp(E) \tag{30}$$

Where the choice of K decides the base of the logarithm. In this case we set it to 1 for clarity.

4.1 Entropy

We can extend this notion to a discrete Random Variable $X \sim p$ with finite domain $\mathcal{X}$. By defining the **Shannon Entropy** $H(X)$ as the average amount of information i.e.

$$H(X) = \mathbb{E}_{X\sim p}\{I(X)\} = \mathbb{E}_{X\sim p}\{-\log p(X)\} = -\sum_{x\in\mathcal{X}} p(x)\log p(x) \tag{31}$$

We can also denote this as $H(p)$ where $p \sim X$ depending on what we want to emphasize. Note that WLOG we can assume that $p(x) > 0\ \forall x \in \mathcal{X}$. This is because we can use the convention that $0 \cdot log0 = 0$ (based on continuity arguments). Hence zero probability outcomes do not contribute to $H(X)$ anyways. We can further extend this for two Random Variables X, Y with finite domain $\mathcal{X} \times \mathcal{Y}$ by defining the **Joint Entropy** as:

$$H(X,Y) = -\sum_{x,y} p_{XY}(x,y)\log p_{XY}(x,y) \tag{32}$$

The **Conditional Entropy** is defined as:

$$H(X|Y) = \mathbb{E}_{X|Y}\{-\log p(X|Y)\} = -\sum_{x,y} p_{XY}(x,y)\log p_{X|Y}(x|y) \tag{33}$$

These quantities have nice properties:

1. *Non-negativity*: $H(X) \geq 0$, with equality only when X is a constant.

   PROOF: WLOG we assume that $p(x) > 0\ \forall x \in \mathcal{X}$. We have that $H(X) = -\sum_x p(x) log p(x) = \sum_x p(x) log p(x)^{-1} \geq 0$, since $p(x) > 0$ and $p(x)^{-1} \geq 1$. If $H(X) = 0$ then $\exists \alpha$ such that $p(\alpha)^{-1} = 1 \Rightarrow p(\alpha) = 1$. Hence X must be a constant, as needed.

2. *Chain Rule*: $H(X,Y) = H(X \mid Y) + H(Y) = H(Y \mid X) + H(X)$

3. *Monotonicity*: $H(X \mid Y) \leq H(X)$

4.2 KL Divergence

We can now look at the **KL Divergence** or **Relative Entropy**. This quantity measures the "distance" between two probability mass functions p and q.

$$KL(p||q) = \mathbb{E}_{X\sim p}\left\{\log \frac{p(X)}{q(X)}\right\} = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \tag{34}$$

The KL divergence has some nice properties.

1. $KL(p||q) \geq 0$ with equality iff $p = q$

   PROOF: If there exists $x \in \mathcal{X}$ such that $p(x) = 0$ and $q(x) > 0$, then $KL(p||q) = \infty$. Otherwise:

$$\begin{aligned} -KL(p||q) &= \mathbb{E}_{X\sim p}\left\{\log\frac{q(X)}{p(X)}\right\} \\ &\overset{(a)}{\leq} \log \mathbb{E}_{X\sim p}\left\{\frac{q(X)}{p(X)}\right\} \\ &= \log\sum_x p(x)\frac{q(x)}{p(x)} = \log\sum_x q(x) = 0 \end{aligned}$$

   Where (a) follows from Jensen's inequality. $KL(p||q) = 0$ only occurs when there is equality in Jensen's inequality, which only occurs when $p(x) = cq(x)$ for some c. Since $\sum_x cq(x) = c\sum_x q(x) = c \Rightarrow c = 1$, so $p = q$ as needed.

2. $KL(p||q)$ is strictly convex in each argument

3. $KL(p||q) \neq KL(q||p)$ so it is not a metric

4. We can decompose the KL divergence into two seperate terms:

$$KL(p||q) = \sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}} p(x)\log q(x) \tag{35}$$
$$= -H(p) + \mathbb{E}_{X\sim p}\{-\log q(x)\} \tag{36}$$
$$= -H(p) + CE(p,q) \tag{37}$$

Where the $H(p)$ is the Entropy and $CE(p,q)$ is called the **Cross Entropy**.

4.3 Mutual Information

We can quantify the amount of information obtained about one discrete random variable X, through another Y by defining the **Mutual Information** as:

$$I(X,Y) = \sum_{x,y} p_{X,Y}(x,y)\log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} \tag{38}$$

We again assume WLOG that $p(x,y) > 0\ \forall(x,y) \in \mathcal{X}\times\mathcal{Y}$. We note the following properties of I:

1. $I(X,X) = H(X) \rightarrow$ Sometimes the Entropy is called the **Self Information**
2. $I(X,Y) = KL(p_{XY}||p_Xp_Y)$
3. $I(X,Y) \geq 0$

   Proof. Notice that $I(X,Y) = KL(p_{XY}||p_Xp_Y) \geq 0$ by the positiveness of $KL(\cdot||\cdot)$ □

4. $I(X,Y) = H(p_X) + H(p_Y) - H(p_{XY})$

   Proof. We use property 2. of I and property 4. of $KL(\cdot||\cdot)$

$$\begin{aligned}
I(X,Y) &= KL(p_{XY}||p_Xp_Y)\\
&= -H(p_{XY}) + CE(p_{XY}, p_Xp_Y)\\
&= -H(p_{XY}) - \sum_{x,y} p_{X,Y}(x,y)\log p_X(x)p_Y(y)\\
&= -H(p_{XY}) - \left(\sum_{x,y} p_{X,Y}(x,y)\log p_X(x) + \sum_{x,y} p_{X,Y}(x,y)\log p_Y(y)\right)\\
&= -H(p_{XY}) - \left(\sum_{x} p_X(x)\log p_X(x) + \sum_{y} p_Y(y)\log p_Y(y)\right)\\
&= -H(p_{XY}) + H(p_X) + H(p_Y)
\end{aligned}$$

□

4.4 Differential Entropy

We can define the Entropy, KL divergence and Mutual Information for continuous random variables.

$$H(p) = -\int_{x\in\mathcal{X}} p(x)\log p(x)d\mu(x) \tag{39}$$

$$KL(p,q) = \mathbb{E}_{X\sim p}\left\{\log\frac{p(X)}{q(X)}\right\} = \int p(x)\log\frac{p(x)}{q(x)}d\mu(x) \tag{40}$$

$$I(X,Y) = \int\int p_{X,Y}(x,y)\log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}d\mu(x)d\mu(y) \tag{41}$$

In the continuous case, the properties previously described hold except that the entropy is no longer necessarily non negative. For an example of this let $p = Uniform\left(\frac{1}{2},1\right)$. Then $H(p) = log(\frac{1}{2}) < 0$.

4.5 Entropy and Estimation

The KL divergence can be used within the Decision Theoretic framework. Semantically, $KL(p||q)$ represents how well some distribution q approximates the "true" p. Suppose we wanted to estimate a distribution p which we knew belonged to a Parametric Family $p \in \{p_\theta\}_{\theta\in\Theta}$. Let $\mathcal{A} = \{p_\theta\}_{\theta\in\Theta}$ and $\delta(D) = p_\theta$. Recall that this is similar to the Estimation problem described in 2.6.1 [10]. We can define the **Negative Log Loss**:

$$L(p,p_\theta) = -\log p_\theta(X) \tag{42}$$

This loss makes sense as $p_\theta(X)$ small means that the model has not taken into account X, and the corresponding loss will be large. The Cross Entropy is the corresponding risk function for this:

$$R(p,p_\theta) = \mathbb{E}_{X\sim p}\{-\log p_\theta(X)\} \tag{43}$$

This risk also makes sense. $KL(p,p_\theta) = -H(p) + L(p,p_\theta)$, and since $H(p)$ is constant, minimizing the KL is equivalent to minimizing the cross entropy. Since $KL(p,p_\theta) \geq 0$ we see that the minimum is attained at $L(p,p_\theta) = H(p)$, which occurs when $p_\theta = p$ i.e. when our prediction matches the "true" density.

4.5.1 Maximum Likelihood Estimation

We don't know p, so we cannot compute (43). Instead, we can use in its place the empirical density function $\hat{p}$, as defined in 10. Given X discrete, it turns out that the MLE for θ is the same as $\arg\min_{\theta\in\Theta} KL(\hat{p}||p_\theta)$. This is because:

[10] We modify the problem to make explicit the intention of estimating the density rather then the parameter. These goals are the same provided the parametric family is **identifiable**

$$
\begin{aligned}
KL(\hat{p}||p_\theta) &= -H(\hat{p}) + CE(\hat{p}, p_\theta) \\
&= -H(\hat{p}) - \sum_{x \in \mathcal{X}} \hat{p}(x) \log p_\theta(x) \\
&= -H(\hat{p}) - \frac{1}{n} \sum_{x \in \mathcal{X}} \sum_{i=1}^{n} \delta(x, x^{(i)}) \log p_\theta(x) \\
&= -H(\hat{p}) - \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x^{(i)}) \\
&= -H(\hat{p}) - \frac{1}{n} l(\theta \mid x^{(1)}, \cdots, x^{(n)})
\end{aligned}
$$

This provides a nice interpretation for the MLE - it is finding the $p \in \{p_\theta\}_{\theta \in \Theta}$ which minimizes the dissimilarity between the empirical distribution of the training set and itself as measured by the KL divergence. Conversly we can justify the use of the Cross Entropy loss through its equivalence to Maximum Likelihood. Note that this holds for X continuous, we just have to change the sums for integrals.

On a final note, one may think that the quantity $KL(p_\theta||\hat{p})$ could be interesting. They would be wrong. This is since $p_\theta(x) = 0 \Rightarrow \hat{p}_\theta(x) = 0$ but $\hat{p}_\theta(x) = 0 \not\Rightarrow p_\theta(x) = 0$ since $\hat{p}_\theta(x) = 0$ only means that the particular value of x wasn't observed in the sample.

4.5.2 Maximum Entropy Principle

The **Principle of Maximum Entropy** (MaxENT) states that the probability distribution which best represents the "current state of knowledge" is the one with the largest entropy. More specifically, given some subset of distributions on $\mathcal{X}$ denoted as $\mathcal{M}$, we want to chose as our estimated distribution:

$$\arg\max_{q \in \mathcal{M}} H(q)$$

We may impose constraints to this in the form of **Testible Information**– statements about q with well-defined truth or falsity. The most basic of these is that $\int_{\mathcal{X}} q(x)dx = 1$. We now show a few maximum entropy distributions.

Theorem 4.1. *Let $X \sim p$ be a RV with finite support $\mathcal{X}$, $|\mathcal{X}| = k$, and $\mathcal{M} = \Delta_k$. The uniform density is the MaxENT density.*

Proof. We derive the following upper bound for $H(p)$

$$H(p) \leq \log k \tag{44}$$

To derive this inequality, let $q \sim Uniform$ on $\mathcal{X}$. We have that:

$$\begin{aligned} D(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x) \\ &= -H(p) + \sum_x p(x) \log k \\ &= -H(p) + \log k \end{aligned}$$

and so $H(p) = \log k - D(p||q) \Rightarrow H(p) \leq \log k$ as needed. Since $H(q) = \log k$ we can see that equality holds iff $p \sim Uniform$. □

So we have that, for densities with finite support and no testible information (apart from being a valid pmf), the MaxENT solution is uniform.

Theorem 4.2. *The MaxENT density for Random Variables $X_1 \in \mathcal{X}_1$ and $X_2 \in \mathcal{X}_2$ with $X_1 \sim p_1$ and $X_2 \sim p_2$ is $(X_1, X_2) \sim p_1 p_2$. i.e. higher entropy assumes independence.*

Proof. Properties 3. and 4. of I gives us that $I(X_1, X_2) \geq 0 \Rightarrow H(X_1) + H(X_2) \geq H(X_1, X_2)$, and so the maximal entropy of (X_1, X_2) is $H(X_1) + H(X_2)$. By definition this only occurs when $I(X_1, X_2) = 0$, which only occurs if $p_{1,2}(x_1, x_2) = p_1(x_1)p_2(x_2)\ \forall x_1, x_2 \in \mathcal{X}_1 \times \mathcal{X}_2$. □

Theorem 4.3. *The MaxENT of X with $\mathcal{X} = \mathbb{N}$ and with testible information $E(X) = \alpha$ is the Geometric Distribution $p(k) = \left(\frac{\alpha}{1+\alpha}\right)^k \frac{1}{1+\alpha}$*

Proof. We want to find the distribution which maximizes the entropy $H(p)$ satisfying the constraints $\mathbb{E}(X) = \alpha$ and $\sum_{i=0}^{\infty} p(i) = 1$. We form the Lagrangian:

$$L(p, \nu, C) = -H(p) + \nu \left(\sum_{i=0}^{\infty} ip(i) - \alpha \right) + C \left(\sum_{i=0}^{\infty} p(i) - 1 \right)$$

Taking the derivative w.r.t. $p(k)$ we get:

$$\frac{\partial}{\partial p(k)} L(p, \nu, C) = -\log p(k) - 1 + k\nu + C \tag{45}$$

$$\Rightarrow p(k) = \exp\{k\nu\} \exp\{C - 1\} \tag{46}$$

And using that $\sum_{i=0}^{\infty} p(i) = 1$ we have that

$$\sum_{i=0}^{\infty} \exp\{i\nu\} \exp\{C-1\} = 1 \Rightarrow \exp\{-C+1\} = \sum_{i=0}^{\infty} \exp\{i\nu\} \tag{47}$$

we substitute (47) into (45) to eliminate C

$$p(k) = \frac{\exp\{k\nu\}}{\sum_{i=0}^{\infty} \exp\{i\nu\}} \tag{48}$$

We then solve for α

$$\begin{aligned}
\mathbb{E}(X) = \sum_{k=0}^{\infty} \frac{k \exp\{k\nu\}}{\sum_{i=0}^{\infty} \exp\{i\nu\}} &= \alpha \\
\Rightarrow \sum_{k=0}^{\infty} k \exp\{k\nu\} &= \alpha \sum_{i=0}^{\infty} \exp\{i\nu\} \\
\overset{(a)}{\Rightarrow} \frac{\exp\{\nu\}}{(1-\exp\{\nu\})^2} &= \frac{\alpha}{(1-\exp\{\nu\})} \\
\Rightarrow \exp\{\nu\} &= \frac{\alpha}{1+\alpha}
\end{aligned}$$

Where (a) comes from the geometric series. Finally, we sub this value into (48) to get the familiar formula:

$$p(k) = \left(\frac{\alpha}{1+\alpha}\right)^k \frac{1}{1+\alpha} \tag{49}$$

□

4.5.3 MaxENT and the Exponential Family

It turns out that if the only testible information we have about our pdf are moment constraints, then the MaxENT solution always belongs to the Exponential Family from 3.1.

Theorem 4.4. *If $X_1, \cdots, X_n$ are an IID sample and $T_1(X), \cdots, T_d(X)$ are statistics, then the MaxENT estimator satisfying $\mathbb{E}_q\{T_j(X)\} = \mathbb{E}_{\hat{p}}\{T_j(X)\}$ $j = 1, \cdots, d$ is the MLE distribution in the exponential family with sufficient statistics $T(X)$*

Proof. For simplicity let X be finite with $\mathcal{X}$ and k as defined before. Suppose we have statistics $T_1(X), \cdots, T_d(X)$ and we define $\mathcal{M}$ as

$$\mathcal{M} = \Big\{ q : \underset{\substack{\text{model expected} \\ \text{feature count}}}{\mathbb{E}_q\{T_j(X)\}} = \underset{\substack{\text{empirical} \\ \text{feature count}}}{\mathbb{E}_{\hat{p}}\{T_j(X)\}} \ j = 1, \cdots, d \Big\} \tag{50}$$

Our testible information are the *d moment constraints*. Using the relation $H(p) = \log k - D(p||q)$ derived from theorem 4.1 we have the following alternative characterization of MaxENT:

$$\arg\max_{q \in \mathcal{M}} H(q) = \arg\min_{q \in \mathcal{M}} KL(q, Uniform) \tag{51}$$

We then pose the MaxENT problem as an optimization problem from 2.2

$$\begin{aligned} \text{Minimize } & \sum_x q(x) \log \frac{q(x)}{u(x)} \\ \text{subject to } & q(x) \geq 0 \\ & \sum_x q(x) = 1 \\ & \sum_x q(x) T_j(x) = \alpha_j \ j = 1, \cdots, d \end{aligned}$$

Where $u(x) = \frac{1}{k} \ \forall x \in \mathcal{X}$. Our Lagrangian is:

$$\begin{aligned} L(q, \lambda, \nu) &= \sum_x q(x) \log \frac{q(x)}{u(x)} + \sum_{j=1}^{d} \lambda_j \left(\alpha_j - \sum_x q(x) T_j(x) \right) + \nu \left(1 - \sum_x q(x) \right) \\ &= \mathbb{E}_q \left\{ \log \frac{q(x)}{u(x)} \right\} + \alpha^T \lambda - \mathbb{E}_q \left\{ \lambda^T T(x) \right\} + \nu - \mathbb{E}_q\{\nu\} \end{aligned}$$

We find the dual function (2). First we find the q which minimizes L:

$$\begin{aligned} \frac{\partial L(q|\lambda, \nu)}{\partial q(x)} &= 1 + \log \frac{q(x)}{u(x)} - \lambda^T T(x) - \nu = 0 \\ &\Rightarrow q^*(x|\nu, \lambda) = u(x) \exp\{\lambda^T T(x) + \nu - 1\} \end{aligned}$$

We then compute the dual:

$$\begin{aligned}
g(\lambda,\nu) &= \min_{q\in\mathcal{M}} L(q^*(x|\nu,\lambda),\lambda,\nu) \\
&= L(q^*(x|\nu,\lambda),\lambda,\nu) \\
&= \mathbb{E}_{q^*}\left\{\lambda^T T(x)+\nu-1\right\}+\alpha^T\lambda-\mathbb{E}_{q*}\left\{\lambda^T T(x)\right\}+\nu-\mathbb{E}_{q^*}\{\nu\} \\
&= \alpha^T\lambda+\nu-\mathbb{E}_{q^*}\{1\} \\
&= \alpha^T\lambda+\nu-\underbrace{\sum_x u(x)\exp\{\lambda^T T(x)\}}_{=Z(\lambda)}e^{\nu-1}
\end{aligned}$$

We claim that Slaters condition is satisfied. We see that L is convex since f_0 is the KL divergence, which is convex in q, and the constraints are all linear. We claim that $\exists q \in \text{int}(\mathcal{M})$ s.t. $q(x) > 0\ \forall x$. We claim that $\text{int}(\mathcal{M})$ is nonempty and assume WLOG that such a q exists since, if it didn't, we could just restrict our domain to $\mathcal{X}\setminus\{x|q(x)=0\}$. Satisfying Slaters condition gives us strong duality, and so if we find a dual optimal (λ^*,ν^*), it will be enough to minimize the strictly convex $L(q,\lambda^*,\nu^*)$ over q.

We first maximize w.r.t ν

$$\begin{aligned}
\frac{\partial g(\lambda,\nu)}{\partial\nu} &= 1-Z(\lambda)e^{\nu-1}=0 \\
&\Rightarrow e^{\nu^*-1}=\frac{1}{Z(\lambda)}
\end{aligned}$$

And we substitute our optimum ν^*:

$$\begin{aligned}
\max_{\nu\in\mathbb{R}} L(q^*(x|\nu,\lambda),\lambda,\nu) &= L(q^*(x|\nu^*,\lambda),\lambda,\nu^*) \\
&= \alpha^T\lambda+\nu-\underbrace{Z(\lambda)e^{\nu^*-1}}_{=1} \\
&= \alpha^T\lambda+\underbrace{\nu-1}_{-\log Z(\lambda)}
\end{aligned}$$

Finally, we use that $\alpha_j=\mathbb{E}_{\hat{p}}\{T_j(X)\}$

$$\begin{aligned}
L(q^*(x|\nu^*,\lambda),\lambda,\nu^*) &= \alpha^T\lambda-\log Z(\lambda) \\
&= \mathbb{E}_{\hat{p}}\{T(X)\}^T\lambda-\log Z(\lambda) \\
&= \frac{1}{n}\sum_{i=1}^n\left(T(X_i)^T\lambda-\log Z(\lambda)\right)
\end{aligned}$$

Let $p(X_i|\lambda)=q^*(X_i\mid\nu^*,\lambda)=u(x)\exp\{\lambda^T T(x)-\log Z(\lambda)\}$. This is a pdf belonging to the Exponential family. We see the correspondence between

maximizing the dual and maximum likelihood on the Exponential family since:

$$\begin{aligned} L(q^*, \lambda, \nu^*) &\propto \frac{1}{n} \sum_{i=1}^{n} \Big(\log p(X_i \mid \lambda) \Big) \\ &= \frac{1}{n} l(X_1, \cdots, X_n | \lambda) \end{aligned}$$

□

Supposing we solved the MLE problem above, we then have that our optima is:

$$q^*(x) = u(x) \exp\{(\lambda^*)^T T(x) - \log Z(\lambda^*)\}$$

What this means is that, given only Moment constraints and no other restrictions, the distribution with the most "Randomness" is precisely the distribution from the exponential family with matching moments.

References

[1] J. Magnus and H. Neudecker, *Matrix differential calculus with applications in statistics and econometrics*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section, Wiley, 1988.

[2] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[3] J. Seth Rosenthal, *A First Look at Rigorous Probability Theory*. 01 2006.

[4] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[5] P. Hoff, "Bayes estimators." Notes, 2013.

[6] M. I. Jordan, "260 course notes." Slides.

[7] T. Carter, "An introduction to information theory and entropy." Slides, 2004.