# Machine Learning Notes

Matthew Scicluna

January 16, 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 $\mu$ |
| $\mathbb{R}$ | 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 |
| $:=$ | equals by definition |
| $\sup(A)$ | Supremum of a set $A$ |
| $\inf(A)$ | Infimum 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 $x$ |
| $[A]_{ij}$ | $i$th row and $j$th 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 $f$ |
| $\text{im}(f)$ | Image of function $f$ |
| $o(f)$ | "Little Oh" of $f$ |

# 2 Preliminaries

## 2.1 Linear Algebra and Topology

We define some basic notions from Linear Algebra which will appear throughout these notes.

### 2.1.1 Inner Products, Norms

We define the **Inner Product** on $\mathbb{R}^n$ as:

$$\langle x, y \rangle := \sum_{i=1}^{n} x_i y_i$$

Throughout these notes we denote this as $x^T y$. We define the **Euclidean Norm** as:

$$||x|| := (x^T x)^{\frac{1}{2}} = \left( \sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}}$$

We can define an Inner Product on the space of matrices $\mathbb{R}^{m \times n}$ as:

$$\langle X, Y \rangle := \sum_{i=1}^{m} \sum_{j=1}^{n} [X]_{ij} [Y]_{ij}$$

Which can be written as $tr(X^T Y)$. Notice that this inner product is just the usual inner product on $\mathbb{R}^n$ treating $X$ as a vector of size $mn$. We define the **Frobenius Norm** as:

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

From these notions we can define the **Ball** $B_r(x)$, the set of all points $y$ in $\mathbb{R}^n$ (or $\mathbb{R}^{m \times n}$) s.t. $||x - y|| < r$.

Suppose we have norms $|| \cdot ||_a$ and $|| \cdot ||_b$ on spaces $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. The **Operator Norm** is defined on $\mathbb{R}^{m \times n}$ as:

$$||X||_{a,b} = \sup\{||Xu||_a \mid ||u||_b \leq 1\}$$

When $|| \cdot ||_a$ and $|| \cdot ||_b$ are both Eucliedean Norms, then the Operator Norm of $X$ is its **Maximum Singular Value**, and is denoted as $||X||_2$.

### 2.1.2 Open and Closed Sets

An element $x \in C \subseteq \mathbb{R}^n$ is called an **Interior Point** if $\exists r > 0$ s.t. $B_r(x) \subseteq C$. The set of all interior points of $C$ is called the **Interior** and is denoted int($C$). $C$ is **Open** if $C$=int($C$). A set is **Closed** if its complement $\mathbb{R}^n \setminus C$ is open. Alternatively, a set is called closed if it contains the limit point of every convergent sequence in it. The **Closure** of a set $C$, is the set of limit points of convergent sequences in $C$, denoted cl($C$). A point is said to be in the boundary $x \in$bd($C$) if $\forall \epsilon > 0 \; \exists y \in C, z \notin C$ s.t. $||y-x|| \leq \epsilon$ and $||z-x|| \leq \epsilon$. Semantically, this means that points in the boundary are aribtrarily close to points both inside of $C$ and outside of $C$. $C$ is closed if it contains all its boundary points, and open if it contains none of them. If the boundary is empty, the set is both closed and open, and it is called **Clopen**.

### 2.1.3 Range and Nullspace

Let $A \in \mathbb{R}^{m \times n}$. We say the **Range** of $A$ is the set of vectors that can be written as linear combinations of $A$, i.e. $range(A) = \{Ax | x \in \mathbb{R}^n\}$. This is a subspace of $\mathbb{R}^m$, with dimension equal to $rank(A)$. The **Nullspace** of $A$ is the set of vectors which can be mapped to 0 $null(A) = \{x | Ax = 0\}$, which is a subset of $\mathbb{R}^n$.

### 2.1.4 Eigen-Decomposition and Singlular Value Decomposition

We call $\lambda \in \mathbb{R}$ an **Eigenvalue** and $q \in \mathbb{R}^n$ an **Eigenvector** of a square matrix $A \in \mathbb{R}^{n \times n}$ iff $Aq = \lambda q$. We have that $\det(A) = \prod_{i=1}^{n} \lambda_i$ and $tr(A) = \sum_{i=1}^{n} \lambda_i$. We call $A$ **Positive Semi Definite** if $x^T Ax \geq 0 \; \forall x$. If the inequality is strict (i.e. $>$ instead of $\geq$) we say that $A$ is **Positive Definite**. We order the eigenvalues as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and refer to $\lambda_i(A)$ as the $i^{th}$ largest eigenvalue of $A$.

If $A$ is symmetric, we can define $Q$ as a matrix whose $i^{th}$ column $q_i$ is an eigenvector of $A$, whose corresponding eigenvalue is $\lambda_i = [\Lambda]_{ii}$ for $\Lambda$ diagonal. We define the **(Symmetric) Eigen-Decomposition** as:

$$A = Q \Lambda Q^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T$$

Where $Q \in \mathbb{R}^{n \times n}$ is **Orthogonal** i.e. $Q^T Q = I$. Semantically, this says that any transformation $A$ can be decomposed into a rotation, a scaling, and then a reverse rotation (since $Q^T = Q^{-1}$).

We define the **Singular Value Decomposition** of a matrix $A$ with $rank(A) = r$ as:

$$A = UDV^T$$
$$\text{where } U \in \mathbb{R}^{m \times r},\ V \in \mathbb{R}^{n \times r} \text{ with } U^TU = I,\ V^TV = I$$
$$\text{and } D \text{ diagonal with } [D]_{ii} = \sigma_i$$

$\sigma_i$ is called the $i^{th}$ **Singular Value** of $A$ and $\sigma_1 \geq \cdots \geq \sigma_r > 0$. We can write this as:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

Where $u_i, v_i$ are the $i^{th}$ columns of $U$ and $V$ respectively. Unlike with the Eigen-Decomposition, this is defined for any matrix, even if it is non-square. There is a relationship between the singular values and the Eigen-Decomposition. Notice that:

$$A^TA = VD^TU^TUDV^T = VD^TDV^T \Rightarrow A^TAV = VD^TD$$
$$AA^T = UDV^TVD^TU^T = UDD^TU^T \Rightarrow AA^TU = UDD^T$$

And so each $v_i$ is an eigenvector of $A^TA$ and each $u_i$ is an eigenvector of $AA^T$. Furthermore, the singular values of a symmetric positive semi definite matrix are the same as its nonzero eigenvalues. We define the **Spectral Radius** of $A$ as $\max_i |\lambda_i(A)|$. Contrast this with $||A||_2$, which is the maximum singular value $\sigma_1$.

## 2.2 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. $\forall ||\Delta|| < r$

$$f(x_0 + \Delta) - f(x_0) = A(x_0)\Delta + o(||\Delta||)$$

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 $i$th component of $f$, with $f$ and $x_0$ defined as before, $e_j$ be the $j$th 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 $Jf(x_0) \in \mathbb{R}^{m \times n}$.

### 2.2.1 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:*

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

For a many to one function $f$, the Jacobian matrix 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:

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

### 2.2.2 Vector Derivatives

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

**Theorem 2.5.** $\nabla_x b^T x = \nabla_x x^T b = b$

*Proof.* Notice

$$(x + \Delta)^T b - x^T b = \Delta^T b$$

And the result immediately follows □

**Theorem 2.6.** $\nabla_x x^T B x = x^T (B^T + B)$

*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 $r_x$ is $o(||\Delta||)$, 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 2.7.** $\nabla_\mu (x - b)^T B (x - b) = (x - b)^T (B + B^T)$

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

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

□

### 2.2.3 Matrix Derivatives

We can refine our definition of differentiability to include matrices. This amounts to replacing the Euclidean norm with the Frobenius norm. 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. $\forall ||\Delta|| < r$

$$f(x_0 + \Delta) - f(x_0) = tr(A(x_0)^T \Delta) + o(||\Delta||)$$

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

**Theorem 2.8.** $\nabla_X tr(BX) = B^T$

*Proof.* The result is immediate since

$$tr(B(X+\Delta)) - tr(BX) = tr(B\Delta)$$

□

**Theorem 2.9.** $\nabla_X \log\det X = X^{-1}$ *for* $X$ *symmetric, invertible*

*Proof.* First notice the following decomposition

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

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

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

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

The result is not as surprising when we compare it with the 1D result $\frac{d}{dx}\log x = \frac{1}{x}$. We now use a previous result to compute a useful identity.

**Theorem 2.10.** $\nabla_X b^T X b = bb^T$

*Proof.* Notice that

$$\begin{aligned}
\nabla_X b^T X b &= \nabla_X tr(b^T X b) \\
&= \nabla_X tr(bb^T X) \\
&= bb^T
\end{aligned}$$

□

## 2.3 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, $p^* = -\infty$

### 2.3.1 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:

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

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 2.11.** *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.

- $\exp ax,\ -\log x,\ x\log x,\ \log\det X$
- Any norm over $\mathbb{R}^n$
- non-negative weighted sums of convex functions $f_i$'s
- Affine compositions: let $f$ be convex, $A$ a matrix, and $b$ a vector; then $g(x) = f(Ax+b)$ with $\text{dom}(g) = \{x \mid Ax+b \in \text{dom}(f)\}$ is convex
- 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 $g(x) = \sup_{y\in\mathcal{Y}} f(x,y)$ is convex in $x$
- Let $f$ be a function, then the **Conjugate** of $f$ denoted $f^*$ is defined as $f^*(y) := \sup_{x\in\text{dom}(f)} (y^T x - f(x))$, where $\text{dom}(f^*) = \mathbb{R}$ is always convex

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

**Theorem 2.12.** *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 (since the inequality holds across its entire domain)!

**Theorem 2.13.** *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 2.14.** *For any convex optimization problem, any locally optimal point is optimal.*

*Proof.* Boyd section 4.2.2 [2] □

### 2.3.2 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.

### 2.3.3 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 $(\lambda, \nu)$ **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 $(\lambda, \nu)$ is **Dual Optimum** if it is dual feasible and $g(\lambda, \nu) = d^*$. We now relate the primal and dual problems:

**Theorem 2.15 (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**.

[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]

**Theorem 2.16 (Slater's Condition).** *Suppose we have a convex primal, and that int($\mathcal{D}$) is non empty. If $\exists x \in int(\mathcal{D})$[2] 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] □

### 2.3.4 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 $(x, \lambda, \nu)$ 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]The conditions that int($\mathcal{D}$) is non empty and $\exists x \in$int($\mathcal{D}$) can be generalized to $\exists x \in$relint($\mathcal{D}$)

### 2.3.5 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 2.17 (KKT Conditions).** *Let $f_0, f_i, h_i$ be differentiable (and so they have open domains). If $x^*$ is optimal and $(\lambda^*, \nu^*)$ 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 2.18.** *If $x^*, \lambda^*, \nu^*$ satisfy the KKT conditions and the primal is convex; then $x^*$ is optimal, $\lambda^*, \nu^*$ 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 $\lambda^*, \nu^*$ 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. $(\lambda^*, \nu^*)$ 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.

## 2.4 Optimization Techniques

We will encounter scenerios where we cannot find an optimal solution using the techniques described above. In these cases we must use an Optimization Algorithm. Our basic set up is as follows: suppose we have an objective function $f(x)$ which we want to minimize over $x \in \mathbb{R}^d$, i.e. unconstrained optimization. We describe two methods: a first order method – **Gradient Descent**, and a second order method – **Newtons Method**. In both cases, our algorithm produces a sequence of iterates $\{x_t\}$ s.t. (ideally) $x_t \to x^*$ for $x^*$ an optimal point.

### 2.4.1 Newtons Method

The motivation for Newtons Method is that we are minimizing a *quadratic approximation* of $f$ – the second order Taylor Expansion:

$$f(x) = Q(x) + O(||x - x_t||^3)$$
$$\text{where } Q(x) = f(x_t) + \nabla_x f(x_t)^T (x - x_t) + \frac{1}{2}(x - x_t)^T Hf(x_t)(x - x_t)$$

The algorithm works as follows: given $x_t$, we look for $x_{t+1}$ which minimizes $Q(x)$. Taking the gradient and setting it to 0 gives us:

$$\nabla_x Q(x) = \nabla_x f(x_t) + Hf(x_t)(x - x_t) = 0$$
$$\Rightarrow x = x_t - Hf(x_t)^{-1} \nabla_x f(x_t)$$

And we set $x_{t+1} = x$, and repeat iterating until convergence. Note that for each step, $Hf(x_t)^{-1}$. This operation has time complexity $O(d^3)$, which becomes prohibitive if $d$ is large.

### 2.4.2 Gradient Descent

We first describe a property of functions that will be needed. We say a function $f$ is $L$ **Lipschitz Continuous** if

$$|f(x) - f(y)| = L||x - y||^2 \tag{9}$$

Such a function has the following property: $||\nabla_x f(x)|| \leq L$. The Gradient Descent algorithm iterates in the following way, given $x_t$ set

$$x_{t+1} = x_t - \gamma \nabla_x f(x_t) \tag{10}$$

and iterate until convergence (or boredom). The motivation for this is to minimize a *linear approximation* of $f$ – the first order Taylor Expansion (contrast this with Newtons method). We call $\gamma$ the **Learning Rate**[3]. The following theorem provides a way to choose the step size.

**Theorem 2.19.** *Let $f$ be convex and $L$-Lipschitz then if we choose $\gamma = \frac{||x_1 - x^*||}{L\sqrt{T}}$, we have that*

$$f\left(\frac{1}{T}\sum_{i=1}^{T} x_t\right) - f(x^*) \leq \frac{||x_1 - x^*||L}{\sqrt{T}}$$

*Proof.* We use the convexity of $f$ and (2.12) to get that:

$$\begin{aligned} f(x_t) - f(x^*) &\leq \nabla_x f(x_t)^T (x_t - x^*) \\ &\overset{(a)}{=} -\frac{1}{\gamma}(x_{t+1} - x_t)^T (x_t - x^*) \\ &= \frac{1}{2\gamma}\Big(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2 + \gamma^2 ||\nabla_x f(x_t)||^2\Big) \\ &\overset{(b)}{=} \frac{1}{2\gamma}\Big(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2\Big) + \frac{\gamma}{2}L^2 \end{aligned}$$

where (a) follows by (10) and (b) follows from the Lipschitz continuity of $f$. □

[3]This is also called the **Step Size** in Optimization

# 3 Probability and Statistics

## 3.1 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:

$\Omega =$ 

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 **$\sigma$-algebra** $\mathcal{F}$ on $\Omega$ 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 $\Omega$ and a $\sigma$-algebra $\mathcal{F}$ on $\Omega$ 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 $\sigma$-algebra can have a valid probability measure associated with it[4]. We can compute Probabilities of Events Conditioned on

[4]For example, if we take $\Omega$ to be $(0, 1)$ and $F$ to be $\mathcal{F} = 2^{\Omega}$ (which always is a $\sigma$-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 $\{A_j\}_{j\in k}$:

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

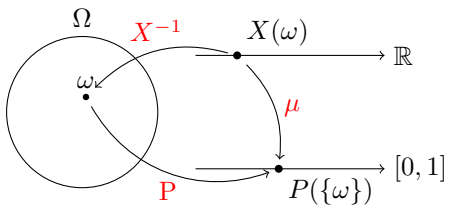
Finally, we define the **Probability Space** as the triple $(\Omega, F, P)$ as defined in the previous section. An example of such a space for a finite $\Omega$ is as follows: take $\mathcal{F} = 2^\Omega$ and use the **Counting Measure**: $P(E) = \frac{|E|}{|\Omega|}$. For $\Omega = \mathbb{R}$ we use the $\mathbb{B}$: the **Borel $\sigma$-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.

### 3.1.1 Random Variables and Vectors

A **Random Variable** (RV) is a $\mathbb{B}$-Measurable function $X : (\Omega, \mathcal{F}) \mapsto (\mathbb{R}, \mathbb{B})$[5]. 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 $\{\omega\}$:




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

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

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

They always satisfy the folling properties:

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

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.

We say that $F_X$ is **Absolutely Continuous**: it is differentiable a.e. and $\exists f_X(x)$ s.t. $F_X(x) = \int_{-\infty}^{x} f_X(u)du$. If $F_X$ is absolutely continuous 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 $\text{im}(X)$ is countable then we call $X$ **Discrete**[6]. Unlike in the continuous case, $f_X(x) = P(X = x)$ and is called the **PMF**. Note that in either case, we usually just write $p(x) := f_X(x)$. We will use this convention throughout these notes.

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:

$$F_X(x_1, ..., x_n) := P(X_1 \leq x_1, ..., X_n \leq x_n)$$

Let $A \in \{1, \cdots, n\}$ and let $X_A := \{X_i : i \in A\}$ and $X_{\neg A} := \{X_i : i \notin A\}$. The **Marginal PDF** of $X_A$ is:

$$p(x_A) := \int_{X_{\neg A}} p\left(x_A, X_{\neg A}\right) dX_{\neg A}$$

The **Conditional PDF** on $X_A$ given $x_{\neg A}$ is:

$$p\left(x_A | x_{\neg A}\right) := \frac{p\left(x_A, x_{\neg A}\right)}{p(x_{\neg A})}$$
$$\text{Where } p(x_{\neg A}) > 0$$

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

[6] 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

$$p(x, y) = p(x)p(y) \ \forall x, y$$

We note a useful property of RVs, the **Chain Rule**:

$$p(x_1, \cdots, x_n) = \prod_{i=1}^{n} p(x_i | x_1, \cdots, x_{i-1})$$

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

$$p(x, y|z) = p(x \mid z)p(y \mid z) \ \forall x, y, z$$

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* $\begin{cases} X \perp Y \mid Z, W \\ X \perp W \mid Z \end{cases} \Rightarrow X \perp Y, W \mid Z$

### 3.1.2 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 $\mathbb{E}(X) = \int_0^\infty 1 - F_X(x)\,dx - \int_{-\infty}^0 F_X(x)\,dx$. If $X$ Continuous, then this simplifies to $\mathbb{E}(X) = \int_{-\infty}^\infty x \cdot p(x)\,dx$. If $X$ discrete, we just replace the integral with a sum. We now state two important and surprising results:

**Theorem 3.1 (Law of the Unconscious Statistician).**

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot p(x)\,dx$$

**Theorem 3.2 (Jensens Inequality).** *For a convex function g*

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

We can generate Moments using the **MGF** of $X$:

$$M_X(t) := \mathbb{E}\{(\exp(Xt)\}$$
$$\text{when } \exists \epsilon > 0 \text{ s.t. } \forall |t| < \epsilon,\ M_X(t) < \infty$$

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 $\{X_i\}$ are 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. $\mathbb{E}\{X\} := \begin{bmatrix} \mathbb{E}\{X_1\} \\ \vdots \\ \mathbb{E}\{X_n\} \end{bmatrix} := \mu_X$
2. $Cov(X) := \mathbb{E}[(X - \mu)(X - \mu)^T] := \Sigma_X$

### 3.1.3 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*:

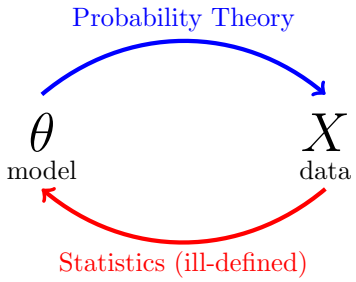
1. **Frequentists**: the *limiting frequency* of the event
2. **Bayesians**: the *reasonable expectation* that the event occurs

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*[7]. These competing interpretations lead to very different ways of modelling and analyzing data, which we will see in the next section.

[7]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

## 3.2 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, what is the most likely model? Unlike with Probability theory, these questions tend to be ill-formed since many models can generate the same data!




In addition, Statistical techniques can be seperated based on

### 3.2.1 Statistical Inference

**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 possible ones! Often we break this problem up by first choosing a **Parametric Family**: $\{p_\theta\}_{\theta\in\Theta}$[8]. This is a set of PDFs/PMFs indexed by some parameter(s) $\theta$. We call $\Theta$ the **Parametric Space** and require for it to be finite dimensional[9]. 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:

$$p(x|\theta) := \binom{N}{x}\theta^x(1-\theta)^{N-x}$$
$$\text{Where dom}(X)= \{0, \cdots, N\},\ \Theta = [0, 1]$$

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 $\theta$. 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 $\theta$, and have to infer it from the data. In Statistics this is called **Point Estimation**[10]!

[8] where $p_\theta = f_X(\cdot|\theta)$ and we write as $p(\cdot|\theta)$ for clarity

[9] If $\Theta$ is infinite dimensional, we call the corresponding model **Non Parametric**

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

### 3.2.2 Maximum Likelihood Estimation

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

**Theorem 3.3 (Neyman Factorization Critereon).** *$T$ is sufficient for $\theta$ if and only if*

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

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

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

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 $\theta$ 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.

Th likelihood principle is the guiding idea behind **Maximum Likelihood Estimation** (MLE). The basic idea is as follows: Since the likelihood contains all information relevent to $\theta$, we can find the most probable $\theta$ simply by maximizing the likelihood using the optimization techniques described in 2.3.

### 3.2.3 MAP Estimation and Method of Moments

In addition to MLE, we describe two other estimators. Suppose that $\theta$ is an RV with its own PDF/PMF. We can use the laws of probability to get that:

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

Where

- $P(\theta|X)$ is the **Posterior Distribution**. The probability of $\theta$ *after* observing data
- $P(X|\theta)$ is the Likelihood
- $P(\theta)$ is the **Prior Distribution** of $\theta$, which can be a parametric model itself (whose parameters would be referred to as **Hyperparameters**)
- $P(X)$ 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 now discuss the **Method of Moments** (MoM). The idea is to equate $K$ moments to each corresponding **Sample Moment**. This amounts to solving a system of $K$ system of equations:

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^{N} X_i^k = \mathbb{E}\{X^k\} \\ &\text{For } k = \{1, \cdots, K\} \end{aligned}$$

### 3.2.4 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 $e_j = \begin{pmatrix} 0 & \ldots & \underset{\substack{\uparrow \\ j^{\text{th}} \text{ coordinate}}}{1} & \ldots & 0 \end{pmatrix}^T \in \mathbb{R}^K$

$\Theta = \Delta_K$ 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 $j^{\text{th}}$ 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 the 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 $\pi_j$'s using MLE. Instead of maximizing the Likelihood we can maximize the log likelihood $\ell$ since it is monotonic function. We observe that:

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

And we ignore the normalizing constant since it wont affect the maximum of $\ell$. 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 $j^{\text{th}}$ 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 $\pi$ 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 the **Conjugate Prior** of 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 $\pi_j$ 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. Finally, notice that if all $\alpha_j = 1$ then the MLE and MAP are the same. This is because a Dirichlet with all $\alpha_j = 1$ is a Uniform distribution.

### 3.2.5 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:

$$\hat{F}(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq t\}} \tag{11}$$

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** $\hat{f}$:

$$\hat{f}(t) := \frac{1}{n} \sum_{i=1}^{n} \delta(x_i, t) \tag{12}$$

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

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

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**:

$$\delta(x, y) := \begin{cases} 1 & x = y \\ 0 & \text{o.w.} \end{cases} \tag{14}$$

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

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

## 3.3 Statistical Decision Theory

We now 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 $p \in \mathcal{P}$[11], 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 a Parametric Family of distributions ($\mathcal{P} = \{p_\theta\}_{\theta \in \Theta}$) and we use $\Theta$ interchangibly with $\mathcal{P}$. We need a way to assign a value to any $\delta$ and a way to compare these values to find which one is "best". One such way of doing this is via the Frequentist Risk.

### 3.3.1 Frequentist Risk Perspective

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

$$R(p, \delta) := \mathbb{E}_{D \sim p}\{L(p, \delta(D))\} \tag{15}$$

If we want to compare decision rules $\delta_1, \delta_2$ using (15), we have to take into account $p$, since $R$ varies with both $p$ and $\delta$. Sometimes one decision rule $\delta_1$ is better then another $\delta_2$ regardless of $p$, in which case we say $\delta_1$ **Dominates** $\delta_2$. 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 $\delta_1$ is a better decision rule then $\delta_2$. Sometimes, there may be a "best" $\delta$, one which isn't dominated by any other $\delta_0$. We say $\delta$ is **Admissible** if $\nexists \delta_0$ s.t. $\delta_0$ dominates $\delta$. 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 $\delta$'s! we must take into account the unknown $p$ when minimizing (15). One way to take this into account is to use the **Minimax Criteria**: the optimal $\delta$ minimizes the Frequentist Risk in worst case scenerio.

$$\delta_{minimax} := \min_{\delta} \max_{p \in \mathcal{P}} R(p, \delta) \tag{16}$$

If $\mathcal{P}$ is a Parameteric Family we can handle the dependence of $R$ on $p$ by averaging out $\Theta$. We can add weights $\pi$ over $\Theta$ to put more weight on certain $\theta$'s. This results in a new weighted risk called the **Bayes Risk**. If we minimize this over $\delta$ we get $\delta_{bayes}$ the **Bayes Rule**.

$$\delta_{bayes} := \arg\min_{\delta} \int_{\Theta} R(p_\theta, \delta)\pi(\theta)d\theta \tag{17}$$

[11] 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 $p$.

Note that the Bayes Rule may not exist, and when they do they may not be unique.

### 3.3.2 Bayesian Risk Perspective

Note that (15) does not consider that we only observed one $D$. We can define a Risk function that does. The **Posterior Risk** is

$$R_B(\delta|D) = \int_\Theta L(p_\theta, \delta)p(\theta|D)d\theta \tag{18}$$

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).

$$\delta_{post} = \arg\min_\delta R_B(\delta|D) \tag{19}$$

Notice that in (18), 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, $\theta$ is integrated out in (18), meaning that (19) gives the undisputed optimal $\delta$!

Note that the Frequentist can still use (19) by interpreting it as (17) with $\pi$ as the "true" prior for $\Theta$. 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.

### 3.3.3 Frequentist Decisition Theory and Estimation

We now describe how we can use Statistical Decision theory as a Frequentist to solve recurring problems throughout these notes. Lets return to the problem of estimating the parameters from a model in a chosen 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 $\theta$. Hence, $\mathcal{A} = \Theta$ and $\delta(D)$ is some an **Estimator** of $\theta$. The estimator should minimize the frequentist 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 $\delta$. This idea works for other loss functions, but the decomposition is not nearly as clean. Notice that Bayesians do not accept this idea since it involves taking an expectation over the data generating distribution, which is contrary to the conditionality principle.

The Bias and Variance of an Estimator are examples of Properties of Estimators. We have already seen a number of estimators: the MLE, MoM and MAP. How can we decide which is best? For frequentists, deciding upon an estimator is often a matter of choosing which estimator has the most desirable properties. One such property for an estimator is consistency. We say an estimator $\hat{\theta}$ is **Consistent** if $\hat{\theta} \xrightarrow{p} \theta$.

The MLE has many desirable properties, which is why it is a favorite for Frequentists. Under regularity conditions on $\Theta$, we have that:

- the MLE is Consistent
- the MLE invariant under reparameterization
- Cramer-Rao lower bound

### 3.3.4 Bayesian Decisition Theory and Estimation

Notice the situation is very different for a Bayesian. A Bayesian would approach Parameter estimation by choosing the Bayes action. We derive the Bayes action for the squared loss:

$$\delta_{post}(D) = \mathbb{E}\{\theta|D\}$$

We distribute the terms in the square:

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

and taking the derivative and setting to 0 yields:

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

We can compute the Bayes rule for the Multinomial Distrubution example from before. We use that $\pi \mid X \sim Dir\left(\{n_j + \alpha_j\}_{j=1}^k\right)$ and get that

$$\mathbb{E}_{\theta|D}\{\theta|D\} = \frac{n_j + \alpha_j}{\sum_{j=1}^k n_j + \alpha_j}$$

And so

$$\boxed{\delta_{post}(D) = \frac{n_j + \alpha_j}{\sum_{j=1}^k n_j + \alpha_j}}$$

Notice this estimate is different then $\hat{\pi}_j^{\text{MLE}}$ and $\hat{\pi}_j^{\text{MAP}}$, which we computed earlier.

## 3.4 The Exponential Family

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

$$p(x|\eta) = \exp\{\eta^T T(x) - A(\eta)\}h(x) \tag{20}$$

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 $\eta$!
   (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. $\eta$ 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 $\eta$, we must restrict $\eta$ 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 $\Omega$ 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 $\eta$'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:

$$p(x|\theta) = p(x|\eta(\theta)) = \exp\{\eta(\theta)^T T(x) - A(\eta(\theta))\}h(x) \tag{21}$$

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

### 3.4.1 Properties of Exponential Families

For canonical exponential families we have the following results:

**Theorem 3.4.** $\nabla_\eta A(\eta) = \mathbb{E}\{T(x)\}$

*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 $A(\eta)$ $\square$

**Theorem 3.5.** $\frac{\partial^2}{\partial\eta_i\partial\eta_j}A(\eta) = Cov\{T_i(x), T_j(x)\}$ *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 3.6.** $\Omega$ *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 $\Omega$ is the epigraph of $A$, it follows that $\Omega$ 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. □

### 3.4.2 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. $\eta$ and $\Omega$ 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

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

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:

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

### 3.4.3 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 3.7.** $\mathbb{E}_{\eta\sim p(\cdot|\tau,n_0)}\{\mathbb{E}_\eta\{T(x)\}\} = \kappa\frac{\tau}{n_0} + (1-\kappa)\frac{\sum_{i=1}^n T(x_i)}{n}$
*Where* $\kappa = \frac{n_0}{n_0+n}$

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

### 3.4.4 The Gaussian Distribution

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

$$p(x|\mu, \Sigma) = (2\pi)^{-\frac{d}{2}} \det\Sigma^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\} \tag{22}$$

For $x, \mu \in \mathbb{R}^p$, $\Sigma \in \mathbb{R}^{p\times p}$ Symmetric and Positive Definite. This distribution has a very surprising role in Statistics through the following theorem:

**Theorem 3.8 (Central Limit Theorem).** *Let $X_1, \cdots X_N$ be IID RVs with finite variances, then*

$$\sqrt{N}\left(\frac{\sum_{i=1}^N X_i}{N} - \mu\right) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$
$$\textit{Where } \mu = \mathbb{E}\{X_i\} \textit{ and } \Sigma = Cov\{X_i\}$$

For the Gaussian distribution, $\mathbb{E}\{X\} = \mu$, $Cov\{X\} = \Sigma$. We can compute the Maximum likelihood estimates for these. We find any stationary points:

$$\begin{aligned} \nabla_\mu \ell(\mu, \Sigma|x_1, \cdots, x_n) &= \nabla_\mu \sum_{i=1}^{n}\left(-\frac{1}{2}\log\det\Sigma - \frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right) \\ &= \sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1} = 0 \\ &\Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

$$\boxed{\hat{\mu}^{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

For the covariance, we take the gradient w.r.t. $\Sigma^{-1}$ to get that:

$$\begin{aligned}
\nabla_{\Sigma^{-1}}\ell(\mu,\Sigma|x_1,\cdots,x_n) &\propto \nabla_{\Sigma^{-1}}\sum_{i=1}^{n}\left(-\frac{1}{2}\log\det\Sigma - \frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right) \\
&= \nabla_{\Sigma^{-1}}\left(\frac{n}{2}\log\det\Sigma^{-1} - \frac{1}{2}\sum_{i=1}^{n} tr((x_i-\mu)^T\Sigma^{-1}(x_i-\mu))\right) \\
&= \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T = 0 \\
&\Rightarrow \Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T
\end{aligned}$$

$$\boxed{\hat{\Sigma}^{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu}^{\text{ML}})(x_i-\hat{\mu}^{\text{ML}})^T}$$

Where we use that the MLE is invariant to reparameterization i.e. $(\hat{\Sigma}^{\text{ML}})^{-1} = (\hat{\Sigma}^{-1})^{\text{ML}}$. We can verify that $(\hat{\mu}^{\text{ML}}, \hat{\Sigma}^{\text{ML}})$ is an optimal point by checking the Hessian and boundary cases. Finally, notice that $\hat{\Sigma}^{-1})^{\text{ML}}$ is biased. We can compare this to the unbiased **Sample Covariance**

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{\mu}^{\text{ML}})(x_i-\hat{\mu}^{\text{ML}})^T$$

We note some surprising properties of the Gaussian:

- The conditional and marginal density of a Gaussian is Gaussian
- Uncorrelated Gaussians are always independant

# 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:

$$f(p(E_1)p(E_2)) = f(p(E_1 \cap E_2)) = f(p(E_1)) + f(p(E_2)) \tag{23}$$

$$f(x \cdot y) = f(x) + f(y) \tag{24}$$

If we assume that $I$ is continuous, then only $I(E) = Klogp(E)$ satisfies (23) [7]. Finally, using 2., we see that $K < 0$. We can then define $I$ as:

$$I(E) = -logp(E) \tag{25}$$

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

## 4.1 Basic Concepts

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.

$$H(X) = \mathbb{E}_{X\sim p}\{I(X)\} = \mathbb{E}_{X\sim p}\{-\log p(X)\} = -\sum_{x\in\mathcal{X}} p(x)\log p(x) \tag{26}$$

We can also denote this as $H(p)$ 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:

$$H(X,Y) = -\sum_{x,y} p(x,y)\log p(x,y) \tag{27}$$

The **Conditional Entropy** is defined as:

$$H(X|Y) = \mathbb{E}_{X|Y}\{-\log p(X|Y)\} = -\sum_{x,y} p(x,y)\log p(x|y) \tag{28}$$

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*: $H(X \mid Y) \leq H(X)$

### 4.1.1 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 $q$.

$$KL(p||q) = \mathbb{E}_{X\sim p}\left\{\log \frac{p(X)}{q(X)}\right\} = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \tag{29}$$

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:

$$KL(p||q) = \sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}} p(x)\log q(x) \tag{30}$$
$$= -H(p) + \mathbb{E}_{X\sim p}\{-\log q(x)\} \tag{31}$$
$$= -H(p) + CE(p,q) \tag{32}$$

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

### 4.1.2 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:

$$I(X,Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{33}$$

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. $I(X,Y) = KL(p(x,y)||p(x)p(y))$
3. $I(X,Y) \geq 0$

   *Proof.* Notice that $I(X,Y) = KL(p(x,y)||p(x)p(y)) \geq 0$ by the positiveness of $KL(\cdot||\cdot)$ □

4. $I(X,Y) = H(X) + H(Y) - H(X,Y)$

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

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

□

### 4.1.3 Differential Entropy

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

$$H(p) = -\int_{x\in\mathcal{X}} p(x)\log p(x)d\mu(x) \tag{34}$$

$$KL(p,q) = \mathbb{E}_{X\sim p}\left\{\log\frac{p(X)}{q(X)}\right\} = \int p(x)\log\frac{p(x)}{q(x)}d\mu(x) \tag{35}$$

$$I(X,Y) = \int\int p_{X,Y}(x,y)\log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}d\mu(x)d\mu(y) \tag{36}$$

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 $H(p) = log(\frac{1}{2}) < 0$.

## 4.2 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 (3.2.2) [12]. We can define the **Negative Log Loss**:

$$L(p,p_\theta) = -\log p_\theta(X) \tag{37}$$

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:

$$R(p,p_\theta) = \mathbb{E}_{X\sim p}\{-\log p_\theta(X)\} \tag{38}$$

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.

### 4.2.1 Maximum Likelihood Estimation

We don't know $p$, so we cannot compute (38). Instead, we can use in its place the empirical density function $\hat{p}$, as defined in (11). Given $X$ discrete, it turns out that the MLE for $\theta$ is the same as $\arg\min_{\theta\in\Theta} KL(\hat{p}||p_\theta)$. This is because:

[12] 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**

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

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.

### 4.2.2 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 $H(p)$

$$H(p) \leq \log k \tag{39}$$

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:

$$\frac{\partial}{\partial p(k)} L(p, \nu, C) = -\log p(k) - 1 + k\nu + C \tag{40}$$

$$\Rightarrow p(k) = \exp\{k\nu\} \exp\{C - 1\} \tag{41}$$

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

$$\sum_{i=0}^{\infty} \exp\{i\nu\} \exp\{C-1\} = 1 \Rightarrow \exp\{-C+1\} = \sum_{i=0}^{\infty} \exp\{i\nu\} \tag{42}$$

we substitute (42) into (40) to eliminate $C$

$$p(k) = \frac{\exp\{k\nu\}}{\sum_{i=0}^{\infty} \exp\{i\nu\}} \tag{43}$$

We then solve for $\alpha$

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

Where (a) comes from the geometric series. Finally, we sub this value into (43) to get the familiar formula:

$$p(k) = \left(\frac{\alpha}{1+\alpha}\right)^k \frac{1}{1+\alpha} \tag{44}$$

□

### 4.2.3 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 3.4.

**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

$$\mathcal{M} = \left\{ q : \underset{\substack{\text{model expected} \\ \text{feature count}}}{\mathbb{E}_q\{T_j(X)\}} = \underset{\substack{\text{empirical} \\ \text{feature count}}}{\mathbb{E}_{\hat{p}}\{T_j(X)\}} \;\; j = 1, \cdots, d \right\} \tag{45}$$

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:

$$\arg\max_{q \in \mathcal{M}} H(q) = \arg\min_{q \in \mathcal{M}} KL(q, Uniform) \tag{46}$$

We then pose the MaxENT problem as an optimization problem from 2.3

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

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 $(\lambda^*,\nu^*)$, it will be enough to minimize the strictly convex $L(q,\lambda^*,\nu^*)$ over $q$.

We first maximize w.r.t $\nu$

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

And we substitute our optimum $\nu^*$:

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

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:

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

□

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.

# 5 Supervised Learning

We now describe a procedure that is crucial for Machine Learning – **Supervised Learning**. We define it using the Statistical Theoretic Framework we developed in (3.3). The idea is that we want to find a function $f$ which uses an RV $X \in \mathcal{X}$ to predict another RV $Y \in \mathcal{Y}$. We suppose that there is a density on $X \times Y$: $(X, Y) \sim p$. Our action space is $\mathcal{A} = \mathcal{Y}^{\mathcal{X}}$. We can evaluate the performance of $f$ using a *prediction loss* $V : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$. We define the **Generalization Error** of $f$ as:

$$L(f) = \mathbb{E}_{(X,Y)\sim p}\{V(Y, f(X))\} \tag{47}$$

This is often called the **Risk** in Machine Learning. Notice that we do not know $p$, and so cannot compute (47), and so we would have to infer it using a technique like Maximum Likelihood.

Supervised learning problems are called different things depending on whether $\mathcal{Y}$ is discrete or continuous (3.1.1). 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**. Usually $\mathcal{X} = \mathbb{R}^d$, and we assume this for the remainder of this section.

## 5.1 Optimum Actions for Regression and Classification

Given a Regression and Classification problem, we want to find optimal actions $f$ using ideas from decision theory described before. We analyze some common loss functions, computing their corresponding risk and optimal $f$ (i.e. the $f$ which minimizes the Generalization error). We will notice that in each case we will need to infer the true distribution $p$. We will tackle this problem in the subsequent subsection by assuming $p$ from a Parametric family and estimating its parameters using Maximum Likelihood. We will then derive a more general solution for estimating densities using an estimate of (47) called **Empirical Risk Minimization**. We begin with regression.

### 5.1.1 Regression

A common choice for $V$ is the squared loss:

$$V(Y, f(X)) = |Y - f(X)|^2$$

We want to find a function $f : \mathcal{X} \to \mathcal{Y}$ which minimizes the Generalization Error:

$$\begin{aligned} L(f) &= \mathbb{E}_{(X,Y)\sim p}\{|Y - f(X)|^2\} \\ &= \int\int |y - f(x)|^2 p(x, y)dxdy \end{aligned}$$

For simplicity we will assume the existance of the PDF $p(x, y)$. In this case we can find the optimal $f$ by using calculus of variations. That is to say, the problem is reduced to an optimization problem. We denote $G(f, f', x) =$

$\int |y - f(x)|^2 p(x,y)dy$ and use the Euler Lagrange equations to get that our stationary point must occur at

$$\frac{\partial G(f, f', x)}{\partial f} - \frac{d}{dx}\frac{\partial G(f, f', x)}{\partial f'} = 0$$
$$\Rightarrow \frac{\partial G(f, f', x)}{\partial f} = 0$$

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

$$\frac{\partial L(f)}{\partial f(x)} = \frac{\partial}{\partial f(x)} \int |y - f(x)|^2 p(x,y)dy$$
$$= 2\int (y - f(x))p(x,y)dy = 0$$

Solving for the above we have that

$$\int (y - f(x))p(x,y)dy = \int tp(x,y)dy - \int f(x)p(x,y)dy = 0$$
$$\Rightarrow \int yp(x,y)dy = f(x)p(x)$$
$$\Rightarrow f(x) = \int y\frac{p(x,y)}{p(x)}dy = \mathbb{E}_{y\sim p(y|x)}\{y|x\}$$

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

$$L_q(y) = \int\int |y - f(x)|^q p(x,y)dxdy \tag{48}$$

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

$$\frac{\partial L_q(y)}{\partial f(x)} = \int q|y - f(x)|^{q-1} sgn(y - f(x))p(x,y)dy \tag{49}$$
$$= \int_{f(x)}^{\infty} q|y - f(x)|^{q-1}p(x,y)dy - \int_{-\infty}^{f(x)} q|y - f(x)|^{q-1}p(x,y)dy \tag{50}$$
$$\Rightarrow \int_{-\infty}^{f(x)} |y - f(x)|^{q-1}p(x,y)dy = \int_{f(x)}^{\infty} |y - f(x)|^{q-1}p(x,y)dy \tag{51}$$

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

$$\int_{-\infty}^{f(x)} p(x,y)dy = \int_{f(x)}^{\infty} p(x,y)dy \tag{52}$$

Finally, as $q \to 0$, the $f(x)$ given by the Minkowski loss is the conditional mode of $y$. Notice again we need to know the underlying data generating pdf i.e. $p(x,y)$.

### 5.1.2 Classification

For Regression problems we saw that we could use the Minkowski error to yield a learning algorithm for Regression problems. Despite this we need a model for our data, which we will discuss in the next section. In this section we will explore a loss function for classification problems. Our decision problem can be expressed as finding the **Decision Region** $\mathcal{R}_j$ for each class, i.e. if $x \in \mathcal{R}_j$ then $f(x) = j$ ($f$ classifies $x$ as being in class $j$). Our loss function $V$ can be represented as a matrix, where $[V]_{ij}$ is the loss obtained when we predict $j$ but $y = i$. We have that our risk is:

$$\begin{aligned} L(f) &= \mathbb{E}_{(X,Y)\sim P_{XY}}\{l\} \\ &= \sum_i \sum_j \int_{\mathcal{R}_j} [V]_{ij} p(x,i) dx \end{aligned}$$

Minimizing the Loss consists of finding the regions $\mathcal{R}_j$ which minimize the loss. We consider the Binary case with the **Zero-One** loss. Our loss function would be

$$V = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

The corresponding risk is:

$$\begin{aligned} L(f) &= \int_{\mathcal{R}_0} p(x,1) dx + \int_{\mathcal{R}_1} p(x,0) dx \\ &\propto \int_{\mathcal{R}_0} p(1|x) dx + \int_{\mathcal{R}_1} p(0|x) dx \end{aligned}$$

Given some $x$, we want to minimize this risk. If $p(1|x) > p(0|x)$, then $f$ should assign $x$ to class 1 (i.e. $x \in \mathcal{R}_1$). Otherwise, $f$ should assign $x$ to class 0. Formally, we can write $f$ as:

$$f(x) = \begin{cases} 1 & \text{if } p(1|x) > p(0|x) \\ 0 & \text{if } p(0|x) > p(1|x) \end{cases}$$

We see that, as with Regression, we need $p(x,y)$ (or $p(y|x)$) to proceed.

## 5.2 Parametric Models

We typically choose $p(x, y)$ from a parametric family and use one of two approaches to model it. The **Generative Approach** seeks to model $p(x, y)$ directly, and the **Discriminative Approach** seeks only to model $p(y|x)$. Note that since $p(x, y) = p(y|x)p(x)$, the only difference in the approaches is that in the generative framework we also model $p(x)$. The Generative approach makes more modelling assumptions and so is less robust for predictions; but one can generate samples from it. We now describe some discriminative models for supervised learning used in practice – Linear Regression and Logistic Regression.

### 5.2.1 Linear Regression

We make the following assumptions about the data. Let $\mathcal{Y} = \mathbb{R}$ and

$$p(y|x, \theta) = \mathcal{N}(\theta^T x, \sigma^2) \text{ where } x, \theta \in \mathbb{R}^d, \sigma^2 > 0$$

We append a 1 to the beginning of $x$ as a bias term, i.e. $\theta^T x = \sum_{i=2}^{d} x_i \theta_i + \theta_1$. Given data $\{X_1, \cdots, X_n\}$, we define the **Design Matrix** as

$$X := \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix}$$

We can estimate $\theta$ using MLE as before. Letting $Y = [y_1, \cdots, y_n]^T$ We get that:

$$\ell(\theta|Y, X) \propto (Y - X\theta)^T (Y - X\theta)$$

Ingoring all the terms which don't depend on $\theta$. This is a convex optimization problem in $\theta$, so it is enough to find a stationary point. We solve for this:

$$\begin{aligned} \nabla_\theta \ell(\theta|Y, X) &= \nabla_\theta \left( Y^T Y - 2X\theta^T Y - \theta^T X^T X\theta \right) \\ &= 0 - 2X^T Y + 2X^T X\theta = 0 \\ &\Rightarrow X^T X\theta = X^T Y \end{aligned}$$

This is sometimes called the **Normal Equations**. If $X^T X$ is invertible then we have a unique solution for $\theta$: $(X^T X)^{-1} X^T Y$. If $X^T X$ is not invertible, then any $\theta$ satisfying the normal equations is an MLE solution.

Supposing that we solved the normal equations to obtain $\hat{\theta}^{\text{MLE}}$ (and by extension $p(y|x)$) the prediction which minimizes the squared loss is:

$$f(x) = \mathbb{E}_{p(y|x)}\{y|x\} = (\hat{\theta}^{\text{ML}})^T x$$

Where we append a 1 to the beginning of $x$ as described earlier.

### 5.2.2 Logistic Regression

We now suppose that $\mathcal{Y} = \{0, 1\}$ and $x, \theta \in \mathbb{R}^d$. Suppose that $Y|X \sim Bern(\sigma(\theta^T x))$, then:

$$\begin{aligned} p(y = 1|\theta) &= \sigma(\theta^T x) \\ p(y = 0|\theta) &= \sigma(-\theta^T x) \\ p(y|\theta) &= \sigma(\theta^T x)^y \sigma(-\theta^T x)^{1-y} \end{aligned}$$

Where $\sigma$ is the **Sigmoid** function.

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

This function has some useful properties:

- $\sigma(-z) = 1 - \sigma(z)$
- $\frac{\partial}{\partial z}\sigma(z) = \sigma(z)\sigma(-z)$

We now look for stationary points:

$$\begin{aligned} \nabla_\theta \ell(\theta|y_1, \cdots, y_n, x_1, \cdots, x_n) &= \nabla_\theta \sum_{i=1}^{n} \left(y_i \log \sigma(\theta^T x_i) + (1 - y_i) \log \sigma(-\theta^T x_i)\right) \\ &= \sum_{i=1}^{n} x_i \left(y_i \frac{x_i \sigma(\theta^T x_i)\sigma(-\theta^T x_i)}{\sigma(\theta^T x_i)} - (1 - y_i)\frac{x_i \sigma(-\theta^T x_i)\sigma(\theta^T x_i)}{\sigma(-\theta^T x_i)}\right) \\ &= \sum_{i=1}^{n} x_i \left(y_i \left(\sigma(-\theta^T x_i) + \sigma(\theta^T x_i)\right) - \sigma(\theta^T x_i)\right) \\ &= \sum_{i=1}^{n} \left(x_i (y_i - \sigma(\theta^T x_i))\right) = 0 \end{aligned}$$

The function is not linear in $\theta$. Contrast this with linear regression, which is linear $\theta$. This can be seen by observing the scalar form of the normal equations:

$$\sum_{i=1}^{n} \left(x_i (y_i - \theta^T x_i)\right) = 0$$

Unlike with Linear Regression, we must use optimization techniques to find the stationary points for this model. We first show that this is a convex optimization problem. It is enough to show that $\ell$ is concave, since $-\ell$ would be convex. We have that:

$$\begin{aligned} \nabla_\theta(\ell(\theta)) &= \sum_{i=1}^{n} \left(x_i (y_i - \sigma(\theta^T x_i))\right) \\ \Rightarrow H\ell(\theta) &= -\sum_{i=1}^{n} x_i x_i^T \left(\sigma(\theta^T x_i)\sigma(-\theta^T x_i)\right) \end{aligned}$$

It is enough to show that $H\ell(\theta)$ is negative semi definite. Given $v \in \mathbb{R}^n$

$$v^T H\ell(\theta)v = -\sum_{i=1}^{n} \underbrace{v^T x_i x_i^T v}_{=(v^T x_i)^2 \geq 0} \underbrace{\left(\sigma(\theta^T x_i)\sigma(-\theta^T x_i)\right)}_{\geq 0}$$

Hence $-\ell(\theta)$ is positive semi definite, as needed. Since this is a convex optimization problem, the MLE and stationary points coincide once again. We can use Newtons Method to find the MLE. We compute Newtons update. Let $\mu_i = \sigma(\theta^T x_i)$ we have:

$$\nabla_\theta - \ell(\theta) = \sum_{i=1}^{n} x_i(\mu_i - y_i) = X^T(\mu - Y)$$
$$H - \ell(\theta) = \sum_{i=1}^{n} x_i x_i^T \left(\mu_i(1-\mu_i)\right) = X^T D X$$
$$\text{Where } D \text{ diagonal and } [D]_{ii} = \mu_i(1-\mu_i)$$

Newtons update can be written as:

$$\begin{aligned}
\theta_{t+1} &= \theta_t - (X^T D_t X)^{-1} X^T(\mu_t - Y_t) \\
&= (X^T D_t X)^{-1}\left((X^T D_t X)\theta_t + X^T(Y_t - \mu_t)\right) \\
&= (X^T D_t X)^{-1}\left(X^T D_t Z_t\right) \\
&\text{Where } Z_t \text{ is } X\theta_t + D_t^{-1}(Y - \mu_t)
\end{aligned}$$

Finally, to make a prediction, we simply use the class with the highest posterior probability, which means that:

$$f(x) = 1 \text{ iff } \sigma((\hat{\theta}^{\mathrm{ML}})^T x)) > 0.5$$

We look at the decision boundary, i.e. the boundary between $\mathcal{R}_0$ and $\mathcal{R}_1$. Notice that $\sigma((\hat{\theta}^{\mathrm{ML}})^T x)) = 0.5 \iff (\hat{\theta}^{\mathrm{ML}})^T x = 0$, and so our boundary is a linear function of $x$!

### 5.2.3 Generative Parametric Models

We now describe the **Fisher Linear Discriminant Analysis** model, a Generative Parametric Model used for classification. We highlight the differences between it and Logistic regression to demonstrate the distinction between Discriminative and Generative approaches. We assume that $Y \in \{0, 1\}$ and

$$\begin{aligned}
&p(x, y = i|\theta) = p(x|y = i, \theta)p(y = i|\theta) \\
&\text{where } X|Y = i) \sim \mathcal{N}(\mu_i, \Sigma) \text{ and } Y \sim Bern(\pi) \\
&\text{and } \theta = (\mu_0, \mu_1, \Sigma, \pi)
\end{aligned}$$

It can be easily shown that

$$\begin{aligned}
\hat{\mu}_0^{\text{ML}} &= \frac{1}{n_0} \sum_{i=1}^{n} \mathbb{1}_{\{y_i=0\}} x_i \\
\hat{\mu}_1^{\text{ML}} &= \frac{1}{n_1} \sum_{i=1}^{n} \mathbb{1}_{\{y_i=1\}} x_i \\
\hat{\Sigma}^{\text{ML}} &= \frac{n_0}{n} \hat{\Sigma}_0 + \frac{n_1}{n} \hat{\Sigma}_1 \\
\hat{\pi}_0^{\text{ML}} &= \frac{n_1}{n}
\end{aligned}$$

Like Logistic Regression, this model also produces a linear decision boundary. This can be seen by considering the conditional distribution:

$$\begin{aligned}
p(y = 1|x) &= \frac{p(x|y = 1)p(y = 1)}{p(x|y = 0)p(y = 0) + p(x|y = 1)p(y = 1)} \\
&= \frac{1}{1 + \exp\left\{-\log \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)}\right\}}
\end{aligned}$$

And

$$\begin{aligned}
&\log \frac{p(x|y = 1)p(y = 1)}{p(x|y = 0)p(y = 0)} \\
&= \log \frac{\pi}{1 - \pi} - \frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1) + \frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0) \\
&= \log \frac{\pi}{1 - \pi} + \mu_1^T \Sigma^{-1} x + \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} x - \frac{1}{2}\mu_0^T \Sigma^{-1} \mu_0 \\
&= \log \frac{\pi}{1 - \pi} + \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 - \frac{1}{2}\mu_0^T \Sigma^{-1} \mu_0 + (\mu_1 - \mu_0)^T \Sigma^{-1} x
\end{aligned}$$

Which is clearly linear in $x$. Notice the decision boundary for Logistic Regression and of LDA will not coincide unless the assumptions on the class conditional distributions are satisfied (i.e. that they are Gaussian with shared covariance). We can see from this example that the Generative LDA has stronger assumptions then the discriminative Logistic Regression and is less robust. However, if the assumptions are satisfied then Generative models are more efficient then discriminative ones.

## 5.3 Empirical Risk Minimization and Regularization

We now expand on the decision theoretic framework for Supervised learning that we described in the beginning of this chapter. Let $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$ (we call $D$ the **Training Data**). We assume that $(X_i, Y_i) \overset{iid}{\sim} p$. We call $\delta$ a **Learning Algorithm** which learns a **Model** $f$ using the training data i.e. $\delta(D) = f$. As before, we can evaluate the performance of $f$ using $V$. In this context $V$ is a measure of the distance between a given **Prediction** $f(x_i)$ and its associated **Ground Truth** $y_i$.

We do not know the Generalization Error since we do not know $p$, but we can approximate it using the **Empirical Risk Minimization** (ERM):

$$L(f) = \frac{1}{n} \sum_{i=1}^{n} V\left(y_i, f(x_i)\right) \tag{53}$$

Notice that (53) is just (47) with the empirical density (11) substituted in place of $p$. We restrict $\mathcal{A} = \mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ and call this the **Hypothesis Space**.

We now have 2 approaches to finding $f$:

- We could choose $V$ and finding $f$ via ERM on a chosen $\mathcal{H}$
- We could choose $V$ and find the optimal $f$ given $p$, and infer $p$ using MLE

We show that both of these approaches can be equivalent. Consider the Linear Regression model from before. Let $V = (Y - f(X))^2$. If we let $\mathcal{H}$ be the set of linear functions, doing ERM is equivalent to doing MLE and assuming that $p(y|x, \theta) = \mathcal{N}(\theta^T x, \sigma^2)$. For Logistic Regression, let

$$\mathcal{H} = \{f : f(x) = \sigma(\theta^T x) \text{ for } \theta \in \mathbb{R}^{d+1}\}$$

where we append a 1 to the end of $x$ like before. If we let

$$V(y_i, \sigma(x_i)) = y_i \log f(x_i) + (1 - y_i) \log(1 - f(x_i)) \tag{54}$$

then ERM over this is equivalent to assuming that $p(y = 1|x, \theta) = \sigma(\theta^T x)$ and doing the optimal decision on the zero-one loss. Notice that the relationship between the first and second approach can be seen in terms through (4.2), noticing that (54) is the Negative Log Loss.

### 5.3.1 Capacity and Generalization

We can split our data into 3 sets:

- **Training Set** $\rightarrow$ This becomes the **Train Data** used to learn $f$
- **Validation Set** $\rightarrow$ used to select **Hyperparameters** (which we will describe later)
- **Test Set** $\rightarrow$ used to estimate the Generalization Error (since we want to assess the performance of $f$ on unseen data)

We define the **Capacity** of $f$ as its "flexibility". A **Hyper-parameter** is a parameter that is not trained (it is specified prior to training). The process of choosing the best hyper-parameters is called **Model Selection**.

A model may not generalize well because of its capacity. If a trained model has high generalization error because it doesn't have sufficient capacity, we say that the model is **Underfitting**. Likewise, if the model has too much capacity, we say the model is **Overfitting**. The generalization error can be decomposed into a bias and variance term, like in (3.3.3). In this context, the bias term represents how close $f$ is to the "true" $f$, and the variance term represents how much $f$ changes for different values of the training set.

We can avoid overfitting by including a **Regularization** term to the ERM, which acts to penalize model complexity.

# 6 Bayesian Inference

We need a theoretical justification for why we assume the existence of a prior distribution on $\theta$ 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:

$$P(X_1, \cdots, X_n) = P(X_{\pi_1}, \cdots, X_{\pi_n}) \tag{55}$$

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

**Theorem 6.1 (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 $\mu$ on $\theta$. Also, if $\theta$ has a density (e.g. is discrete or continuous) then $d\mu(\theta) = p(\theta)d\theta$. Note: $\theta$ may be infinite!*

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

# 7 Probabilistic Graphical Models

An ordering $I : v \mapsto \{1, \cdots, n\}$ is **Topological** iff $j \in \pi_i \Rightarrow I(j) < I(i)$ *i.e. parents always come before their children* If $G$ is a DAG, then $\exists$ a Topological Ordering on $G$ A **Forest** is a Graph s.t. each node has at most one parent A **Tree** is a Forest if it is connected. The Factorization Property of DAGs is as follows: Given DAG $G = (V, E)$

$$\mathcal{L}(G) = \left\{ p \text{ is a dist over } x_v : \exists \text{ factors } f_i \text{ s.t. } p(x_v) = \prod_{i=1}^{n} f_i(x_i; x_{\pi_i}) \right\}$$

$$\text{where } f_i \text{ satisfies: } \begin{array}{l} f_i > 0 \\ \sum_{x_i} f_i(x_i; x_{\pi_i}) = 1 \\ f_i : Dom(x_i)^2 \mapsto [0, 1] \end{array}$$

**Theorem 7.1 (Leaf Plucking Property).** *if $n$ is a leaf of $G$, $p(x_v) \in \mathcal{L}(G - \{n\})$*

*Proof.* $P(x_v) = p(x_{1:n-1}, x_n) = f_n(x_n; x_{\pi_n}) \prod_{i \neq n} f_i(x_i, x_{\pi_i})$. Next marginalizing out $x_n$ and using that it is a leaf, we have:

$$p(x_{1:n-1}) = \underbrace{\sum_{x_n} f_n(x_n; x_{\pi_n})}_{\text{sums to 1}} \underbrace{\prod_{i=1}^{n-1} f_i(x_i; x_{\pi_i})}_{\text{Does not contain } x_n} = \prod_{i=1}^{n-1} f_i(x_i; x_{\pi_i})$$

So $p(x_v) \in \mathcal{L}(G - \{n\})$ as needed. □

**Theorem 7.2 (Factors are Conditional PMFs).** *Let $p \in \mathcal{L}(G)$ and $\{f_j\}$ be a factorization, then $\forall i, P(x_i | x_{\pi_i}) = f_i(x_i; x_{\pi_i})$*

*Proof.* WLOG let $\{1, \cdots, n\}$ be a Topological Ordering and use Theorem 1 to get that $p(x_{1:i}) \in \mathcal{L}(G - \{i+1, \cdots, n\})$, and so $p(x_{1:i}) = \underbrace{\prod_{j=1}^{i-1} f_j(x_j; x_{\pi_j})}_{\text{Call this } g(x_{1:i-1})} f_i(x_i; x_{\pi_i})$.

We partition $\{1 : i\}$ as $\{i\} \cup \pi_i \cup A$ and get that:

$$p(x_i \mid x_{\pi_i}) = \frac{\sum_{x_A} f_i(x_i; x_{\pi_i}) g(x_{1:i-1})}{\sum_{x_A} \sum_{x'_i} f_i(x'_i; x'_{\pi_i}) g(x_{1:i-1})} = \frac{f_i(x_i; x_{\pi_i}) \sum_{x_A} g(x_{1:i-1})}{\sum_{x'_i} f_i(x'_i; x'_{\pi_i}) \sum_{x_A} g(x_{1:i-1})} = f_i(x_i; x_{\pi_i})$$

□

Note: adding edges adds more distributions i.e. $E \subseteq E'$ and $G' = (V, E')$ then $\mathcal{L}(G) \subseteq \mathcal{L}(G')$

**Theorem 7.3.** $p \in \mathcal{L}(G) \iff x_i \perp x_{nd(i)} \mid \pi_i$

*Proof.* $(\Rightarrow)$ $(\Leftarrow)$ □

# 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.