1 Hypothesis Testing

The material on hypothesis testing follows the book of [JN20].

Recall our setup of hypothesis testing: there is a space of observations $\mathcal{X}$ and some sets of probability distributions $\mathcal{P}_i$. Each observation X is assumed to come from a distribution in one of these sets, and we wish to know which set $\mathcal{P}_i$ that is. For simplicity, we will work with 2 sets of probability distributions $\mathcal{P}_1, \mathcal{P}_2 \subset \mathcal{P}(\mathcal{X})$. We define a **hypothesis** as deciding whether an observation comes from one set or the other:

$$H_1 : X \sim P \in \mathcal{P}_1$$
$$H_2 : X \sim P \in \mathcal{P}_2$$

We call our testing procedure *simple* if $\mathcal{P}_1$ and $\mathcal{P}_2$ are singleton sets. Otherwise, they are *composite*. We can also define **tests** as measurable functions $T : X \to \{1, 2\}$ which assign observations to guesses, and the **risks** involved:

$$R_1\{T|H_1, H_2\} = \sup_{P \in \mathcal{P}_1} P\{T(X) = 2\}$$
$$R_2\{T|H_1, H_2\} = \sup_{P \in \mathcal{P}_2} P\{T(X) = 1\}$$

These are directly analagous to the notions of Type-1 and Type-2 error seen in statistics. The overall risk for a test can be defined as

$$R(T|H_1, H_2) := R_1(T) + R_2(T)$$

though there are alternate formulations, such as taking the maximum or average. We are interested in finding an optimal test T^* that minimizes the above total risk.

2 Simple Hypothesis Tests

Consider the following simple setup: $\mathcal{P}_1 = \{P_1\}, \mathcal{P}_2 = \{P_2\}$. Without loss of generality, let $\mathcal{X} = \mathbb{R}^d$, so each P_i has an associated probability density function (pdf) p_i:

$$P_1(dx) = p_1(x)dx$$
$$P_2(dx) = p_2(x)dx$$

In general, this assumption is not necessary; we can consider instead the densities of P_1 and P_2 w.r.t. to the measure $P_1 + P_2$.

Theorem 2.1 *The optimal hypothesis test T^* is*

$$T^*(x) = \begin{cases} 1 & p_1(x) \geq p_2(x) \\ 2 & p_2(x) > p_1(x) \end{cases} \tag{1}$$

Further, the minimal risks are $R_1(T^*) = R_2(T^*) = \int \min(p_1(x), p_2(x))dx$.

Let us consider the case when the two distributions are multivariate Gaussians: $p_1 = \mathcal{N}(x_1, I_d), p_2 = \mathcal{N}(x_2, I_d)$, where I_d is the identity covariance matrix. A similar proof could be followed for other distributions.

We'll begin by noting that our test looks for when $\frac{p_1(x)}{p_2(x)} \geq 1$. Expanding this, we see

$$\frac{p_1(x)}{p_2(x)} = \frac{e^{-\frac{1}{2}\|x-x_1\|^2}}{e^{-\frac{1}{2}\|x-x_2\|^2}} = e^{\frac{1}{2}(\|x-x_2\|^2 - \|x-x_1\|^2)}$$

So the fraction being greater than 1 occurs when the exponent is ≥ 0, or equivalently, $\|x-x_1\|^2 \leq \|x-x_2\|^2$. We can massage this into a familiar form:

$$\begin{aligned}\|x-x_1\|^2 \leq \|x-x_2\|^2 &\implies \|x\|^2 - 2x^\top x_1 + \|x_1\|^2 \leq \|x\|^2 - 2x^\top x_2 + \|x_2\|^2 \\ &\implies (x_1 - x_2)^\top x \geq \frac{\|x_1\|^2 - \|x_2\|^2}{2}\end{aligned}$$

Next, we will define 2 important quantities:

$$e := \frac{x_1 - x_2}{\|x_1 - x_2\|}, w := \frac{x_1 + x_2}{2}$$

So the above equation can be rewritten as:

$$e^\top x \geq e^\top w \Leftrightarrow p_1(x) \geq p_2(x)$$

This expression looks familiar: it is the equation for testing which side of halfspace we are on. The vectors involved have a nice intuitive interpretation as well: e is the unit vector extending from x_2 towards x_1, and w is the midpoint between x_1 and x_2. Notably, $e^\top w$ is a hyperplane separating x_1 and x_2. Therefore, we can rewrite the decision equations above in terms of halfspaces and vectors:

$$T^*(x) = \begin{cases} 1 & e^\top x > e^\top w \\ 2 & e^\top x \leq e^\top w \end{cases} \tag{2}$$

Now, let's look at the risk. We will utilize the fact that, given the parameterization of each candidate distribution, each observation X is simply a $Z \sim \mathcal{N}(0, I_d)$ shifted by a candidate mean x_i:

$$\begin{aligned}R_1(T^*|H_1, H_2) &= P_1(e^\top X \leq e^\top w) \\ &= P_1(e^\top x_1 + e^\top Z \leq e^\top w) \\ &= P_1(e^\top Z \leq e^\top (w - x_1)) \\ &= P_1\left(e^\top Z \leq -\frac{1}{2} e^\top (x_1 - x_2)\right) \\ &= P_1\left(e^\top Z \leq -\frac{1}{2}\|x_1 - x_2\|\right)\end{aligned}$$

Since $e^\top Z$ is a linear combination of standard normals, it is also normally distributed. Given that e is a unit vector, $e^\top Z \sim \mathcal{N}(e^\top 0, e^\top I_d e) = \mathcal{N}(0, \|e\|^2) = \mathcal{N}(0,1)$. So, $e^\top Z$ is a standard normal random variable, which allows us to conclude:

$$R_1(T^*|H_1, H_2) = 1 - \Phi\left(\frac{1}{2}\|x_1 - x_2\|\right),$$

where $\Phi(\cdot)$ is the Gaussian cdf. We can follow a similar deduction for $R_2(T^*)$:

$$\begin{aligned}
R_2(T^*|H_1, H_2) &= P_2(e^\top X > e^\top w) \\
&= P_2(e^\top Z > e^\top (w - x_2)) \\
&= P_2\left(e^\top Z > \frac{1}{2} e^\top (x_1 - x_2)\right) \\
&= P_2\left(e^\top Z > \frac{1}{2}\|x_1 - x_2\|\right) \\
&= 1 - \Phi\left(\frac{1}{2}\|x_1 - x_2\|\right).
\end{aligned}$$

Thus, $R_2(T^*) = R_1(T^*)$, so T^* is optimal. This expression for the risk makes sense intuitively: the closer the candidate distributions are, the harder they are to distinguish, so the risk of misclassifying will be higher.

Thus, we can see that not only is the most intuitive hypothesis test optimal, but it has a nice geometric interpretation as well.

3 Composite Case

For the composite case, we consider the special case when $\mathcal{P}_1, \mathcal{P}_2$ are sets of probability density functions (pdfs) for Gaussian random variables, though similar analysis can be extended to other types of probability distributions. Let $\mathcal{X}_1, \mathcal{X}_2 \subset \mathbb{R}^d$ be closed, convex, and disjoint sets, the composite hypotheses for Gaussian random variables is then:

$$\begin{aligned}
H_1 &: X \sim \mathcal{N}(x, I_d),\ x \in \mathcal{X}_1 \\
H_2 &: X \sim \mathcal{N}(x, I_d),\ x \in \mathcal{X}_2.
\end{aligned} \tag{3}$$

As in the simple case, an optimal test T for such composite hypotheses seeks to minimize the risk $R(T|H_1, H_2)$. Define $f : \mathcal{X}_1 \times \mathcal{X}_2 \mapsto \mathbb{R}_+$ to be $f(x_1, x_2) := \frac{1}{2}|x_1 - x_2|^2$ and let $x_1^* \in \mathcal{X}_1, x_2^* \in \mathcal{X}_2$ be the closest points between the two convex sets:

$$f(x_1^*, x_2^*) = \min_{x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2} f(x_1, x_2). \tag{4}$$

Then it turns out that the optimal test for the composite case is the optimal test for the simple case where the two Gaussian distributions has means x_1^*, x_2^*. Intuitively, this means that minimizing the

risk for the composite case is equivalent to minimizing the risk of the following *worst-case* simple hypothesis testing problem:

$$H_1^* : X \sim \mathcal{N}(x_1^*, I_d) \tag{5}$$
$$H_2^* : X \sim \mathcal{N}(x_2^*, I_d). \tag{6}$$

This simple hypothesis testing problem is the worst case since x_1^*, x_2^* are the closest centroids in the two sets, making them the hardest to distinguish.

To prove the result, we first need to prove a property of $\mathcal{X}_1$ (or equivalently, $\mathcal{X}_2$) in the following lemma.

Lemma 3.1 *Let $x_1^* \in \mathcal{X}_1, x_2^* \in \mathcal{X}_2$ be the minimizers in (4), then for any $x_1 \in \mathcal{X}_1$,*

$$(x_1 - x_1^*)^\top (x_1^* - x_2^*) \geq 0. \tag{7}$$

Proof: By convexity of $\mathcal{X}_1$, for any $\theta \in (0,1)$, we have $\theta x_1 + (1-\theta)x_1^* \in \mathcal{X}_1$ and $f(\theta x_1 + (1-\theta)x_1^*, x_2^*) \geq f(x_1^*, x_2^*)$. Let $h := x_1 - x_1^*$, this becomes:

$$\begin{aligned} f(x_1^* + \theta h, x_2^*) - f(x_1^*, x_2^*) &= \|x_1^* + \theta h - x_2^*\|^2 - \|x_1^* - x_2^*\|^2 \\ &= \theta^2 \|h\|^2 + 2\theta h^\top (x_1^* - x_2^*) \geq 0. \end{aligned}$$

Therefore, for any $\theta \in (0,1)$,

$$h^\top (x_1^* - x_2^*) = (x_1 - x_1^*)^\top (x_1^* - x_2^*) \geq -\frac{1}{2}\theta \|h\|^2.$$

Let $\theta \downarrow 0$ proves the lemma. ■

Let $e := \frac{x_1^* - x_2^*}{|x_1^* - x_2^*|}$, then by Lemma 3.1, we have:

$$e^\top x_1 \geq e^\top x_1^*,$$

and by symmetry:

$$e^\top x_2 \leq e^\top x_2^*.$$

We now prove the following theorem.

Theorem 3.1 *For any test T', the test $T : \mathcal{X} \to \{1,2\}$ defined as*

$$T(x) := \begin{cases} 1 & e^\top x \geq e^\top w \\ 2 & e^\top x < e^\top w, \end{cases} \tag{8}$$

where $w := \frac{x_1^* + x_2^*}{2}$, satisfies

$$R_i(T|H_1, H_2) \leq R_i(T'|H_1, H_2),\ i = 1, 2, \tag{9}$$

and therefore is the optimal test for the hypotheses in (3). Further, it is also the optimal test for the hypotheses in (5).

Proof: First, consider any probability distribution $P \in \mathcal{P}_1$ such that $P = \mathcal{N}(x_1, I_d), x_1 \in \mathcal{X}_1$, by definition of T, its risk for H_1 is then:

$$\begin{aligned} P\left\{T(X) = 2\right\} &= P\left\{e^\top X \le e^\top w\right\} \\ &= P\left\{e^\top x_1 + e^\top Z \le e^\top w\right\} \\ &= P\left\{e^\top Z \le e^\top (w - x_1)\right\}, \end{aligned}$$

where $Z \sim \mathcal{N}(0, I_d)$. Since e is a unit vector, $e^\top Z \sim \mathcal{N}(0,1)$ is a standard Gaussian random variable, and by Lemma 3.1,

$$\begin{aligned} e^\top (w - x_1) &= e^\top (w - x_1^*) - e^\top (x_1 - x_1^*) \\ &\le e^\top (w - x_1^*) = -\frac{\|x_1^* - x_2^*\|}{2}. \end{aligned}$$

Therefore,

$$P\left\{e^\top Z \le e^\top (w - x_1)\right\} \le 1 - \Phi\left(\frac{\|x_2^* - x_1^*\|}{2}\right).$$

As a result,

$$\begin{aligned} R_1(T|H_1, H_2) = \sup_{P \in \mathcal{P}_1} P\left\{T(X) = 2\right\} &\le 1 - \Phi\left(\frac{\|x_1^* - x_2^*\|}{2}\right) \\ &\le \inf_{\bar{T}} R(\bar{T}|H_1, H_2) \le R_1(T'|H_1, H_2). \end{aligned}$$

Notice that the left-hand side of the second inequality is the minimal risk for the probability distributions in (5), therefore it is less than the minimal risk over the entire sets of distributions from $\mathcal{P}_1$ and $\mathcal{P}_2$, $\inf_{\bar{T}} R(\bar{T}|H_1, H_2)$. By a symmetric argument, we have

$$R_2(T|H_1, H_2) = \sup_{P \in \mathcal{P}_2} P\left\{T(X) = 1\right\} \le 1 - \Phi\left(\frac{\|x_1^* - x_2^*\|}{2}\right) \le R_2(T'|H_1, H_2).$$

■

4 Detectors

We can generalize the hypothesis test above with a function called a **detector** $\phi : \mathcal{X} \to \mathbb{R}$, such that

$$T_\phi(x) = \begin{cases} 1 & \phi(x) \ge 0 \\ 2 & \phi(x) < 0 \end{cases}. \tag{10}$$

Note that the detector used in previous sections for hypothesis testing between Gaussian random variables is the affine function $\phi(x) = e^\top w - e^\top x$. The risks for such hypothesis test is defined as before:

$$R_1(T_\phi) = \sup_{P \in \mathcal{P}_1} P(\phi(x) < 0) \tag{11}$$

$$R_2(T) = \sup_{P \in \mathcal{P}_2} P(\phi(x) \geq 0). \tag{12}$$

Upon closer inspection, we see that $P(\phi(x) < 0) = \mathbf{E}_P[\mathbf{1}\{\phi(x) < 0\}]$, which is not differentiable with respect to the detector ϕ and as a result, difficult to analyze.

Instead, we can make the risks differentiable by replacing the indicator function inside the expectations with the differentiable exponential function $e^{\phi(x)}$:

$$R_-(\phi) = \sup_{P \in \mathcal{P}_1} \mathbf{E}_P[e^{-\phi(x)}] \tag{13}$$

$$R_+(\phi) = \sup_{P \in \mathcal{P}_2} \mathbf{E}_P[e^{\phi}(x)]. \tag{14}$$

Such risks are called *relaxed risks*. Using the fact that $\mathbf{1}\{x > 0\} \leq e^x$, we can show that the relaxed risks are upper bounds of the original risks

$$R_1(T_\phi) = \sup_{P \in \mathcal{P}_1} \mathbf{E}_P[\mathbf{1}\{\phi(x) < 0\}] \leq \sup_{P \in \mathcal{P}_1} \mathbf{E}_P\{e^{-\phi(x)}\} = R_-(\phi) \tag{15}$$

$$R_2(T_\phi) = \sup_{P \in \mathcal{P}_2} \mathbf{E}_P[\mathbf{1}\{\phi(x) > 0\}] \leq \sup_{P \in \mathcal{P}_2} \mathbf{E}_P\{e^{\phi(x)}\} = R_+(\phi). \tag{16}$$

As a result, the overall risk of the any detector-based risk satisfies:

$$\min_T \max\{R_1(T), R_2(T)\} \leq \min_\phi \max\{R_1(T_\phi), R_2(T_\phi)\} \leq \min_\phi \max\{R_-(\phi), R_+(\phi)\} \tag{17}$$

Although the detector-based test may not be optimal, the following lemma shows that there always exists a detector with an overall risk comparable to the optimal test.

Lemma 4.1 *Suppose there exists a test $T : \mathcal{X} \to \{1, 2\}$ such that $R_1(T), R_2(T) \leq \epsilon$ for some $\epsilon \in (0, 1/2]$. Then, $\exists \phi : \mathcal{X} \to \mathbb{R}$ such that:*

$$R_1(T_\phi), R_2(T_\phi) \leq 2\sqrt{\epsilon(1-\epsilon)}. \tag{18}$$

In addition, the detector ϕ can be applied to more complex cases when multiple samples are available:

$$\phi^n(x_1, \ldots, x_n) = \sum_{i=1}^n \phi(x_i) \tag{19}$$

Proof: Let $\mathcal{X}_1 := \{x \in \mathcal{X} : T(x) = 1\}, \mathcal{X}_2 := \{x \in \mathcal{X} : T(x) = 2\}$ be the decision regions of T. Notice that $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ and thus $P(\mathcal{X}_1) + P(\mathcal{X}_2) = P(\mathcal{X}) = 1$. By definition, the risks associated with T can be expressed in terms of these sets as:

$$R_1(T) = \sup_{P \in \mathcal{P}_1} P(\mathcal{X}_2) \leq \epsilon \tag{20}$$

$$R_2(T) = \sup_{P \in \mathcal{P}_2} P(\mathcal{X}_1) \leq \epsilon. \tag{21}$$

Consider the following detector ϕ:

$$\phi(x) := \begin{cases} \frac{1}{2}\log\frac{1-\epsilon}{\epsilon} & x \in \mathcal{X}_1 \\ \frac{1}{2}\log\frac{\epsilon}{1-\epsilon} & x \in \mathcal{X}_2 \end{cases}$$

Let T_ϕ be the test generated by ϕ. For some $P \in \mathcal{P}_1$, we have:

$$\begin{aligned}
R_1(T_\phi) &\leq R_-(\phi) \\
&= \sup_{P\in\mathcal{P}_1} \mathbf{E}_P[e^{-\phi(x)}] \\
&= \sup_{P\in\mathcal{P}_1} \int_{x\in\mathcal{X}_1} e^{-\phi(x)}p(x)dx + \int_{x\in\mathcal{X}_2} e^{-\phi(x)}p(x)dx \\
&= \sup_{P\in\mathcal{P}_1} \int_{\mathcal{X}_1} \sqrt{\frac{\epsilon}{1-\epsilon}}dP + \int_{\mathcal{X}_2} \sqrt{\frac{1-\epsilon}{\epsilon}}dP \\
&=: \sup_{P\in\mathcal{P}_1} \frac{1}{\delta}P(\mathcal{X}_1) + \delta P(\mathcal{X}_2) && \left(\delta := \sqrt{\frac{1-\epsilon}{\epsilon}}\right) \\
&= \sup_{P\in\mathcal{P}_1} \frac{1}{\delta}(1-P(\mathcal{X}_2)) + \delta P(\mathcal{X}_2) \\
&= \frac{1}{\delta} + \left(\delta - \frac{1}{\delta}\right) \sup_{P\in\mathcal{P}_1} P(\mathcal{X}_2) \\
&\leq \frac{1}{\delta} + \left(\delta - \frac{1}{\delta}\right)\epsilon = \frac{\epsilon + (1-2\epsilon)\epsilon}{\sqrt{\epsilon(1-\epsilon)}} = 2\sqrt{\epsilon(1-\epsilon)},
\end{aligned}$$

where the last inequality uses (20). The argument for $R_2(T_\phi)$ is the same, except it uses $R_+(\phi)$ instead. Therefore, we can find a detector whose corresponding hypothesis test has a risk not that much larger than that of the original test. ■

Next, we'll look at a larger question: what about the optimal detector? We'll consider the generalized composite case of choosing from sets Θ_1, Θ_2. For optimality, we want to minimize the risk, or equivalently:

$$\min_{\phi\in\mathcal{F}} \max\left\{\sup_{\theta_1\in\Theta_1} \mathbf{E}_{\theta_1}\{e^{-\phi(x)}\}, \sup_{\theta_2\in\Theta_2} \mathbf{E}_{\theta_2}\{e^{\phi(x)}\}\right\}$$

We'll denote this quantity as Δ.

5 Hypothesis Testing with Detectors

Define $\mathcal{P} := \{P_\theta\}_{\theta\in\Theta}$ as the set of all probability distributions over the parameter space Θ, and let $\Theta_1 \subset \Theta$ and $\Theta_2 \subset \Theta$ so that $\mathcal{P}_1$ and $\mathcal{P}_2$ can be defined as $\mathcal{P}_1 := \{P_\theta\}_{\theta\in\Theta_1}$ and $\mathcal{P}_2 := \{P_\theta\}_{\theta\in\Theta_2}$.

Recall that

$$\mathrm{R}_{-}\left[\phi \mid \mathcal{P}_1\right] = \sup_{P\in\mathcal{P}_1} \int_{\Omega} \exp\{-\phi(\omega)\} P(d\omega) \tag{22}$$

$$\mathrm{R}_{+}\left[\phi \mid \mathcal{P}_2\right] = \sup_{P\in\mathcal{P}_2} \int_{\Omega} \exp\{\phi(\omega)\} P(d\omega) \tag{23}$$

$$\mathrm{R}\left[\phi \mid \mathcal{P}_1, \mathcal{P}_2\right] = \max\left[\mathrm{R}_{-}\left[\phi \mid \mathcal{P}_1\right], \mathrm{R}_{+}\left[\phi \mid \mathcal{P}_2\right]\right] \tag{24}$$

We wish to find a detector $\phi : \mathcal{X} \mapsto \mathbb{R} \in \mathcal{F}$ that minimizes $\mathrm{R}\left[\phi \mid \mathcal{P}_1, \mathcal{P}_2\right]$. Let

$$\begin{aligned}\Delta &:= \min_{\phi\in\mathcal{F}} \mathrm{R}\left[\phi \mid \mathcal{P}_1, \mathcal{P}_2\right] \\ &= \min_{\phi\in\mathcal{F}} \max\left\{\sup_{P_1\in\mathcal{P}_1} \mathbb{E}_{P_1}\left[e^{-\phi}\right], \sup_{P_2\in\mathcal{P}_2} \mathbb{E}_{P_2}\left[e^{\phi}\right]\right\}\end{aligned} \tag{25}$$

we can rewrite the optimization problem as

$$\log\Delta = \min_{\phi\in\mathcal{F}} \max_{P_1\in\mathcal{P}_1, P_2\in\mathcal{P}_2} \underbrace{\frac{1}{2}\left[\log \mathbb{E}_{P_1}\left[e^{-\phi}\right] + \log \mathbb{E}_{P_2}\left[e^{\phi}\right]\right]}_{\Phi\left(\phi;(P_1,P_2)\right)} \tag{26}$$

This is because when shifting ϕ by a constant: $\phi(\cdot) \mapsto \phi(\cdot) - a$, the positive quantities $F[\phi] := \sup_{P_1\in\mathcal{P}_1} \mathbb{E}_{P_1}\left[e^{-\phi}\right]$ and $G[\phi] := \sup_{P_2\in\mathcal{P}_2} \mathbb{E}_{P_2}\left[e^{\phi}\right]$ are multiplied by e^{a} and e^{-a}, respectively, and their product remains intact. Therefore, a candidate solution ϕ to the problem $\min_{\phi} H[\phi]$, where $H[\phi] := \sqrt{F[\phi]G[\phi]}$, can be shifted by a constant to ensure $F[\phi] = G[\phi]$, and this operation does not change $H[\cdot]$; as a result, minimizing H over all ϕ is the same as minimizing H over ϕ that gives $F[\phi] = G[\phi]$. It therefore follows that 25 is exactly the same as to minimize over ϕ for the quantity $H[\phi]$ [JN20].

We wonder whether we could have

$$\min_{\phi\in\mathcal{F}} \max_{P_1\in\mathcal{P}_1, P_2\in\mathcal{P}_2} \Phi\left(\phi;(P_1,P_2)\right) \stackrel{?}{=} \max_{P_1\in\mathcal{P}_1, P_2\in\mathcal{P}_2} \min_{\phi\in\mathcal{F}} \Phi\left(\phi;(P_1,P_2)\right) \tag{27}$$

The following lemma solves for $\min_{\phi\in\mathcal{F}} \Phi\left(\phi;(P_1,P_2)\right)$ given $P_1(x)$ and $P_2(x)$

Lemma 5.1 *Given $P_1 \in \mathcal{P}_1$ and $P_2 \in \mathcal{P}_2$, we have*

$$\min_{\phi\in\mathcal{F}} \Phi\left(\phi;(P_1,P_2)\right) = \log \int \sqrt{P_1(x)P_2(x)}dx \tag{28}$$

The minimum is obtained by functions of the form

$$\phi(x) = \frac{1}{2}\log\frac{P_1(x)}{P_2(x)} + c$$

Proof: We first define

$$\bar{\phi}(x) := \frac{1}{2}\log\frac{P_1(x)}{P_2(x)}$$

Plugging $\bar{\phi}(x)$ into $\Phi\left(\phi;(P_1,P_2)\right)$ gives

$$\begin{aligned}\Phi\left(\bar{\phi};(P_1,P_2)\right) &= \frac{1}{2}\left[\log\mathbb{E}_{P_1}\left[e^{-\bar{\phi}}\right]+\log\mathbb{E}_{P_2}\left[e^{\bar{\phi}}\right]\right]\\ &= \log\int\sqrt{P_1(x)P_2(x)}dx\\ &= \Phi\left(\bar{\phi}+c;(P_1,P_2)\right)\end{aligned}$$

where c is a constant and does not change the value of Φ. Define $g(x) := \sqrt{P_1(x)P_2(x)}$ and rewrite $\phi(x)$ as $\phi(x) = \bar{\phi}(x)+\delta(x)$. Then

$$\begin{aligned}\log\int g(x)dx &= \int\sqrt{g(x)}e^{-\frac{\delta(x)}{2}}\sqrt{g(x)}e^{\frac{\delta(x)}{2}}dx\\ &\le \left(\int g(x)e^{-\delta(x)}dx\right)^{\frac{1}{2}}\left(\int g(x)e^{\delta(x)}dx\right)^{\frac{1}{2}} \qquad (29)\\ &= \left(\int\sqrt{P_1(x)P_2(x)}e^{\bar{\phi}(x)-\phi(x)}dx\right)^{\frac{1}{2}}\left(\int\sqrt{P_1(x)P_2(x)}e^{\phi(x)-\bar{\phi}(x)}dx\right)^{\frac{1}{2}}\\ &= \left(\int P_1(x)e^{-\phi(x)}dx\right)^{\frac{1}{2}}\left(\int P_2(x)e^{\phi(x)}dx\right)^{\frac{1}{2}}\\ &= e^{\Phi\left(\phi;(P_1,P_2)\right)}\end{aligned}$$

Therefore, the right hand side of 28 lower bounds $\Phi\left(\phi;(P_1,P_2)\right)$ and the lower bound is obtained with $\bar{\phi}(x)+c$. Note that the inequality in 29 is obtained via Cauchy-Schwarz and holds with equality when $\sqrt{g(x)}e^{-\frac{\delta(x)}{2}} = \lambda\sqrt{g(x)}e^{\frac{\delta(x)}{2}}$, or in other words, when $\delta(x)$ is a constant. ■

Remark 5.1 The right hand side of 28 is related to the notion of Hellinger distance $H^2(P_1,P_2)$ between two probability density functions $P_1(x)$ and $P_2(x)$:

$$\begin{aligned}H^2(P_1,P_2) &= \frac{1}{2}\int(\sqrt{P_1(x)}-\sqrt{P_2(x)})^2dx\\ &= 1-\int\sqrt{P_1(x)P_2(x)}dx \qquad (30)\end{aligned}$$

Intuitively, when $P_1(x)$ "matches" with $P_2(x)$, $\int\sqrt{P_1(x)P_2(x)}dx$ will be large and the Hellinger distance will be small.

If the equality sign in 27 holds, we could solve

$$\min_{\phi\in\mathcal{F}}\max_{P_1\in\mathcal{P}_1,P_2\in\mathcal{P}_2}\Phi\left(\phi;(P_1,P_2)\right)$$

by first finding $P_1^*(x) \in \mathcal{P}_1$ and $P_2^*(x) \in \mathcal{P}_2$ such that

$$\sqrt{P_1^*(x)P_2^*(x)}dx = \max_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \int \sqrt{P_1(x)P_2(x)}dx$$

and then setting

$$\phi^*(x) := \frac{1}{2} \log \frac{P_1(x)}{P_2(x)}$$

6 Min-max Theorem

6.1 Introduction

Let $A \subset X$, $B \subset Y$. We are given a function $f : A \times B \mapsto \mathbb{R}$. We define two quantities (or, optimization problems) $\bar{f}(A, B)$ and $\underline{f}(A, B)$ as

$$\bar{f}(A, B) := \inf_{x \in A} \sup_{y \in B} f(x, y) \tag{31}$$

$$\underline{f}(A, B) := \sup_{y \in B} \inf_{x \in A} f(x, y) \tag{32}$$

Clearly, we have $\underline{f}(A, B) \leq \bar{f}(A, B)$.

We first give the definition of a saddle point:

Definition 6.1 *A point $(x_*, y_*) \in A \times B$ is called a saddle point of function $f(x, y) : A \times B \mapsto \mathbb{R}$, if f as a function of $x \in A$ attains at this point its minimum, and as a function of $y \in B$ - its maximum, that is, if*

$$f(x, y_*) \geq f(x_*, y_*) \geq f(x_*, y) \, \forall (x \in A, y \in B)$$

The following proposition shows how the saddle point is related to the two optimization problems $\bar{f}$ and $\underline{f}$

Proposition 6.1 *[JN20] f has saddle point if and only if $\bar{f}(A, B)$ and $\underline{f}(A, B)$ are solvable with equal optimal values:*

$$\inf_{x \in A} \sup_{y \in B} f(x, y) = \sup_{y \in B} \inf_{x \in A} f(x, y)$$

Whenever this is the case, the saddle points of f are exactly the pairs (x_*, y_*) comprised of optimal solutions to problems $\bar{f}(A, B)$ and $\underline{f}(A, B)$, and the value of f at every one of these points is a common value.

6.2 Von Neumann's Minimax Theorem

We now show the *Von Neumann's Minimax Theorem* [Lue68] as a particular instance where $\underline{f}(A, B) = \bar{f}(A, B)$, with certain conditions on X, Y, A, B and f.

Theorem 6.1 *Let X be a normed space which is reflective, i.e., $X^{**} = X$. Let $A \subset X$ and $B \subset X^*$ be compact and convex. Then*

$$\min_{x\in A}\max_{y^*\in B}\left\langle x, y^*\right\rangle = \max_{y^*\in B}\min_{x\in A}\left\langle x, y^*\right\rangle \tag{33}$$

Proof: Let

$$f(x) := \sup_{y^*\in B}\left\langle y^*, x\right\rangle = \max_{y^*\in B}\left\langle y^*, x\right\rangle$$

The maximum exists as for each $x \in X$, $y^* \mapsto \langle y^*, x\rangle$ is continuous because

$$\left|\left\langle y_1^*, x\right\rangle - \left\langle y_2^*, x\right\rangle\right| \leq \left\|y_1^* - y_2^*\right\| \|x\|$$

and B is compact. By the Weierstrass extreme-value theorem, $\max_{y^*\in B}\langle y^*, x\rangle$ exists.

The function $f : A \mapsto \mathbb{R}$ is convex, as it is the maximum over affine functions. f is also continuous as

$$\begin{aligned} f(x) - f(x^{'}) &= \max_{y^*\in B}\left\langle y^*, x\right\rangle - \max_{y^*\in B}\left\langle y^*, x^{'}\right\rangle && (34)\\ &= \max_{y^*\in B}\min_{z^*\in B}\left(\left\langle y^*, x\right\rangle - \left\langle z^*, x^{'}\right\rangle\right) && (35)\\ &\leq \max_{y^*\in B}\left\langle y^*, x - x^{'}\right\rangle && (36)\\ &\leq \max_{y^*\in B}\left\|y^*\right\|\left\|x - x^{'}\right\| && (37) \end{aligned}$$

Therefore, as A is compact, $\min_{x\in A} f(x)$ exists by the Weierstrass extreme-value theorem.

We now apply the Fenchel duality theorem with the associations: $f \to f(x) = \max_{y^*\in B}\langle y^*, x\rangle$, $C \to X$, $g \to 0$, and $D \to A$:

$$\begin{aligned} \min_{x\in A} f(x) &= \min_{x\in C\cap D}\left\{f(x) + g(x)\right\} && (38)\\ &= \max_{C^*\cap D^*}\left\{-g^*\left(-x^*\right) - f^*\left(x^*\right)\right\} && (39) \end{aligned}$$

We immediately have

$$\begin{aligned} D^* &= \left\{x^* \in X^* : \sup_{x\in A}\left\langle x^*, x\right\rangle < \infty\right\} \equiv X^* && (40)\\ -g^*\left(-x^*\right) &= \min_{x\in A}\left\langle x^*, x\right\rangle && (41)\\ C^* &= \left\{x^* \in X^* : \sup_{x\in X}\left\{\left\langle x^*, x\right\rangle - f(x)\right\} < \infty\right\} \equiv B && (42)\\ f^*\left(x^*\right) &= \begin{cases} 0, & x^* \in B;\\ \infty, & \text{otherwise.} \end{cases} && (43) \end{aligned}$$

To prove 42 and 43 we note that for the two cases, we have

1. Let $x^* \notin B$. The corollary to the Eidelheit separation theorem (in week 10's notes) shows that there exists $x_1 \in X$ and c such that $\langle x^*, x_1\rangle - \langle y^*, x_1\rangle > c > 0$ for all $y^* \in B$. Therefore, $\langle x^*, x\rangle - \max_{y^*\in B}\langle y^*, x\rangle$ can be made arbitrarily large by taking $x = \lambda x_1$ with $\lambda > 0$.

2. Conversely, let $x^* \in B$. Then $\langle x^*, x\rangle - \max_{y^*\in B}\langle y^*, x\rangle \leq 0$, and the equality can be achieved with $x = 0$.

The theorem is easily proved with

$$\min_{x\in A}\max_{y^*\in B}\left\langle y^*, x\right\rangle =: \min_{x\in A} f(x) = \max_{y^*\in B\cap X^*} -g^*\left(-y^*\right) = \max_{y^*\in B}\min_{x\in A}\left\langle y^*, x\right\rangle \tag{44}$$

■

6.3 Examples

Example 6.1 (Two-player games with mixed strategies) Consider the two-player game with the payoff matrix $P \in \mathbb{R}^{M\times N}$, where one of the players tries to minimize the payoff by choosing a row m from $R := \{1, \ldots, M\}$, and the other player tries to maximize the payoff by choosing a column n from $C := \{1, \ldots, N\}$. We are interested in the cases where the following condition holds

$$\min_{m\in R}\max_{n\in C} P_{mn} = \max_{n\in C}\min_{m\in R} P_{mn}. \tag{45}$$

Unfortunately, one can easily come up with examples in which this condition does not hold. However, if we relax the conditions a bit and consider mixed strageties, we get a situation where euqation 45 must hold.

Put formally, let $A := \mathcal{P}(R)$ and $B := \mathcal{P}(C)$ be the set of all probability distributions on R and C respectively. Instead of forcing the two players to make deterministic choices from R and C, we now allow them pick distributions from A and B. The payoff we are interested in then becomes

$$f(\mu, \nu) := \sum_{m\in R}\sum_{n\in C}\mu(m)\nu(n)P_{mn}. \tag{46}$$

Since both A and B are compact and convex and are subsets of a reflective space, by von Neumann's minimax theorem, we have

$$\inf_{\mu\in A}\sup_{\nu\in B} f(\mu, \nu) = \sup_{\nu\in B}\inf_{\mu\in A} f(\mu, \nu). \tag{47}$$

In fact, this is what originately motivated von Numann's minimax theorem.

Example 6.2 (Statistical procedures) The process of choosing a statistical procedure can also be formulated as a two-player game, where the statistician chooses a procedure $f \in \mathcal{F}$ and the nature chooses a distribution $P \in \mathcal{P}$. The overall objective can be expressed as

$$\min_{f\in\mathcal{F}}\max_{P\in\mathcal{P}} R_{P,f}. \tag{48}$$

This corresponds to the worst-case risk of the best procedure. Now suppose that the minimax condition holds, i.e.

$$\min_{f\in\mathcal{F}}\max_{P\in\mathcal{P}} R_{P,f} = \max_{P\in\mathcal{P}}\min_{f\in\mathcal{F}} R_{P,f},$$

then the overall optimal strategy would correspond to the optimal strategy against the least favorable distribution. Notice that this is what we did in detector based testing, where we picked a pair of the pair of gaussians that are closest to each other, making them the hardest to tell apart.

This notion appeared in [Wal51] and [BG55].

Example 6.3 (General constrained optimization problems) When faced with a constrained optimization problem, assuming the minimax condition and interchange mins and maxs are often a good way to proceed. For example, when deriving Kantorovich's duality, we are presented with the objective

$$\inf_{\pi \in \Pi(\mu,\nu)} I(\pi). \tag{49}$$

One way to proceed is by introducing the function

$$\chi(\pi) := \begin{cases} 0 & \text{if } \pi \in \Pi(\mu,\nu) \\ \infty & \text{otherwise} \end{cases} \tag{50}$$

and rewriting the objective as

$$\inf_{\pi} \big(I(\pi) + \chi(\pi) \big) . \tag{51}$$

Now, note that

$$\inf_{\pi} \chi(\pi) = \inf_{\pi} \sup_{\varphi,\phi} \left\{ I(\pi) + \int \varphi \, d\mu + \int \phi \, d\nu - \int \big(\varphi(x) + \phi(y) \big) \, \pi(\, dx, \, dy) \right\}, \tag{52}$$

Now, if we assume the minimax condition and interchange the inf and sup, we get pretty close to the final answer. Although it still remains to prove the minimax condition rigorously, we can see how this can be a useful tool for simplifying problems.

6.4 Generalizations

In 1953, Ky Fan published a generalization to von Neumann's minimax theorem (Theorem 2 in [Fan53]). The theorem can be stated as follows

Theorem 6.2 (Fan's Minimax Theorem) *Suppose A is convex and B is convex and compact. Suppose $f : A \times B \to \mathbb{R}$ satisfies*

1. *$y \mapsto f(x, y)$ is upper semicontinuous and concave for all $x \in A$ and*
2. *$x \mapsto f(x, y)$ is lower semicontinuous and convex for all $y \in B$,*

then

$$\inf_{x \in A} \max_{y \in B} f(x, y) = \max_{y \in B} \inf_{x \in A} f(x, y). \tag{53}$$

Another generalization is published by Maurice Sion in 1958 [Sio58] (originally appeared as Corollary 3.4).

Theorem 6.3 (Sion's Minimax Theorem) *Suppose A and B are convex and one of them is compact. Suppose $f : A \times B \to \mathbb{R}$ satisfies*

1. *$y \mapsto f(x, y)$ is upper semicontinuous and quasi-concave for all $x \in A$ and*
2. *$x \mapsto f(x, y)$ is lower semicontinuous and quasi-convex for all $y \in B$,*

then

$$\inf_{x \in A} \sup_{y \in B} f(x, y) = \sup_{y \in B} \inf_{x \in A} f(x, y). \tag{54}$$

References

[BG55] David Blackwell and M.A. Girshick. *Theory of Games and Statistical Decisions.* Wiley, 1955.

[Fan53] Ky Fan. Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America*, 39(1):42–47, 1953.

[JN20] Anatoli B. Juditsky and Arkadi Nemirovski. *Statistical Inference via Convex Optimization.* Princeton University Press, 2020.

[Lue68] David G. Luenberger. *Optimization by Vector Space Methods.* Wiley, 1968.

[Sio58] Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8:171–176, 1958.

[Wal51] Abraham Wald. *Statistical Decision Functions.* Wiley, 1951.