NOTES ON BLACKWELL DOMINANCE

Tilman Börgers, December 3, 2023

1. Set-Up

There is a finite set of states: $\Omega = \{\omega_1, \ldots, \omega_n\}$ where $n \in \mathbb{N}$ is the number of states. An experiment P is an $n \times N$ matrix $\{P_{ij}\}_{i=1,\ldots,n;j=1,\ldots,N}$ such that $P_{ij} \geq 0$ for all i, j and $\sum_{j=1}^{N} P_{ij} = 1$ for all $i = 1, \ldots, n$. We shall call such a matrix a "Markov matrix." The interpretation is here that $N \in \mathbb{N}$ is the number of possible outcomes of the experiment, and that P_{ij} is the probability of observing experiment outcome j in state ω_i.

We shall think of the elements of Ω as the "payoff-relevant states." Specifically, there is a decision maker who has to choose an action from a set non-empty set A which is a subset of $\mathbb{R}^n$. Suppose $a \in A$. Let the components of a be denoted by a_i, for $i = 1, 2, \ldots, n$. Then a_i is the loss that the decision maker suffers if she chooses action a and the state turns out to be ω_i.[1] The decision maker has to choose a without knowing Ω, but after observing the outcome of one or more experiments. The decision maker seeks to minimize expected losses.

If the decision maker first observes the outcome of experiment P and then chooses an action from A, then a decision rule maps outcomes of P into elements of A. We can describe such decision rules using a $N \times n$ matrix D where every row of D is an element of A. The interpretation is that the decision maker plans to take the action that is the j-th row of D if she observes outcome j of the experiment.

Suppose the experiment is P, and the decision maker chooses decision rule D. Then the diagonal elements of the $n \times n$ matrix PD describe the expected loss in each state i of the world. This is called the "risk" in state i. The vector of diagonal elements

These notes are based on Chapter12 of David Blackwell and M. A. Girshick, *Theory of Games and Statistical Decisions*, Wiley, New York, 1954. (Republished by Dover Publications, Mineola, New York, 1979.)

[1] We describe the decision maker's objective as loss minimization rather than utility maximization because, historically, some of the literature on which these notes build have adopted this perspective. Of course, the two descriptions of the decision maker's objectives are equivalent.

of PD is $b(P, D)$. We call $b(P, D)$ the "risk vector." As D varies over all possible decision rules, we obtain the set of all possible risk vectors, denoted by $B(P, A)$. This set describes the feasible risk vectors, i.e. the risk vectors that the decision maker can achieve by choosing from A if she has experiment P available. Some of the criteria for comparing experiments below will focus on comparing the set $B(P, A)$ for different experiments. This is a somewhat agnostic approach: we evaluate an experiment by studying the implied feasible set rather than the optimal choice from that feasible set. Other criteria will compare, say, the minimized expected cost for some given prior over Ω. We shall introduce the notation needed for such comparisons later.

Suppose there were two experiments: P and Q. We shall be interested in formal definitions and characterizations of the following three concepts:

- P is more informative than Q;
- P and Q are complements;
- P and Q are substitutes.

To investigate these relations in detail, we might need to know not just the marginal distributions of the outcomes of P and the outcomes of Q conditional on the state, but also the joint distribution of outcomes of P and Q conditional on the state. Let us denote the probability that we observe outcome j of experiment P and outcome k of experiment Q in state ω_i by R_{ijk}. If the number of outcomes of P is N_1, and the number of outcomes of Q is N_2, then we can think of R as an experiment with $N_1 \cdot N_1$ outcomes.

2. P Is More Informative Than Q

Perhaps the simplest way of defining that P is more informative than Q is to say that this is the case when knowledge of the outcome of P implies knowledge of the outcome of Q. This is captured by the following definition:

Definition 0. *P implies Q, written as "$P \to Q$," if for all j, k we have: if there exists a state i such that: $R_{i,j,k} > 0$ then $R_{i',j,k'} = 0$ for all states i' and for all experiment outcomes $k' \neq k$.*

Note that this definition refers directly to the joint distribution of the outcomes of P and Q, and that no reference is made to the usefulness of the experiments in decision problems. Our focus in these notes will be the connection between definitions of relations among experiments that refer directly to the distributions of these

experiments, and definitions that refer to the distribution of the experiments' outcomes. We shall therefore now propose another definition that is in spirit similar to Definition 0, but that refers to decision problems.

Definition 1. *P makes Q redundant, written as "$P \twoheadrightarrow Q$," if $B(P, A) = B(R, A)$ for all non-empty, compact and convex subsets A of $\mathbb{R}^n$.*

Thus, in words, we say that P makes Q redundant when the opportunities that are available to a decision maker who knows P are the same as the opportunities that are available to a decision maker who knows P and Q, independent of the decision problem that the decision maker faces. The decision problem is captured by the set A. Note we don't quite consider *all* non-empty subsets A of $\mathbb{R}^n$, but only those subsets that are compact and convex. Compactness is assumed to make sure that, when later we introduce the objective of expected loss minimization, a solution to the decision maker's optimization problem exists. Convexity reflects the idea that, if two actions are available to the decision maker, also all probability distributions over these actions are available. Intuitively, the decision maker has access to a randomization device that is independent of Ω, and independent of the experiment outcomes. Without this assumption, pure noise could allow the decision maker to expand her opportunities, and would therefore appear useful if we focus on the size of the set $B(P, A)$ only.

Are Definitions 0 and 1 equivalent? Obviously, if P implies Q, then P also makes Q redundant. The converse is, however, not true. For example, when Q is random noise, that is, outcomes of P are not correlated with ω, then P makes Q redundant, but clearly P does not imply Q. It thus becomes interesting to search for a condition that is equivalent to redundancy, but that refers only to the distribution of the outcomes of P and Q, and does not require the explicit consideration of decision problems. In particular, we might ask whether such a condition, like Definition 0, has to refer to the joint distribution of the outcomes of the two experiments, or whether the redundancy condition can be characterized in terms of the marginal distributions only. We shall postpone answering these questions until later.

Instead, we now turn to third formalization of the idea that P is more informative than Q. The following definition is due to David Blackwell.

Definition 2. *P is more informative than Q, written as "$P \supset Q$," if $B(Q, A) \subseteq B(P, A)$ for all closed, bounded and convex subsets A of $\mathbb{R}^n$.*

Thus, P is more informative than Q if, whatever the decision problem is, provided that A is closed, bounded, and convex, the set of risk vectors that the decision maker

can implement basing her decision on the outcome of P, is a superset of the set of risk vectors that the decision maker can implement if she bases her decision on the outcome of Q. Of course, this implies that, if the decision maker minimizes expected loss with some subjective prior over the state space, she will achieve at least as low an expected loss if the experiment on which she bases her decision is P as she does if it is Q. One might ask whether the converse is also true, that is: Is it the case that P is more informative than Q whenever the decision maker will achieve at least as low an expected loss if the experiment on which she bases her decision is P as she does if it is Q, for all decision problems and all subjective priors? This is indeed the case. We shall prove it in the proof of Proposition 2.

Blackwell informativeness is a much studied concept. Early research on the subject is summarized in Chapter 12 of Blackwell and Girshick (1954). In particular, their Theorem 12.2.2 offers five equivalent characterizations of the concept of "more informative." Here, we break down this result into two Propositions which we state and prove separately.

Proposition 1. *Let P and Q be two $n \times N_1$ and $n \times N_2$ Markov matrices. Each of the following conditions is equivalent to $P \supset Q$:*

(1) There is an $N_1 \times N_2$ Markov matrix M with $PM = Q$.

(2) For every $N_2 \times n$ matrix D there is an $N_1 \times N_2$ Markov matrix M such that every diagonal element of PMD equals the corresponding diagonal element of QD.

(3) For every $N_2 \times n$ matrix D there is an $N_1 \times N_2$ Markov matrix M with

$$\textit{Trace of } PMD \leq \textit{ Trace of } QD.$$

Among the three conditions in Proposition 1, the first one is the most familiar one. The matrix M describes how the outcome of experiment P gets "garbled" in a way that is independent of the state to yield the outcome of experiment Q.

The second and third conditions in Proposition 1 both have to do with risk vectors. In condition 2 the matrix D describes a decision rule of the decision maker in case the decision maker observes experiment Q. Thus, the diagonal elements of QD constitute the risks in each state. The condition says that the same risks could be obtained by garbling the experiment Q using the matrix M, and then choosing actions D.

The second condition is at first sight weaker than condition 1 for two reasons. Firstly, condition 2 allows the garbling matrix M to depend on D. Secondly, whereas condition 1 requires the precise conditional probability distributions of the garbled

experiment PM to be the same as the conditional probability distributions of the experiment Q, condition 2 only requires the risk vectors to be the same (given that decision rule D is chosen).

There is a different way of reading condition 2. Instead of interpreting the diagonal of PMD as the risk vector that the decision maker faces when observing the outcome of experiment PM and then choosing decisions according to D, we can interpret it as the risk vector that the decision maker faces when observing the outcome of experiment P and then choosing decisions according to MD. The decision rule prescribes for each observed outcome of P an action that is a convex combination of actions prescribed by D. Thus, the second condition says that the decision maker can achieve the same risk vector when observing P as she achieves when observing Q, and that to do so the decision maker only chooses actions that are convex combinations of the actions chosen when observing Q.

Condition 3 refers to the trace of matrices PMD and QD. The trace is the sum of the diagonal elements of these matrices, that is, the sum of the risks in all states. We can think of this sum as the ex ante expected loss where all states are equally likely. More precisely, the expected loss is $1/n$ times the trace. The condition leaves out the normalizing factor $1/n$ but that doesn't matter. Condition 3 is in two ways weaker than condition 2: Firstly, condition 3 only refers to the sum of the diagonal elements of PMD and QD, whereas condition 2 refers to each of these elements. Secondly, condition 3 is a weak inequality, whereas condition 2 is an equality.

Having interpreted the conditions, we can immediately notice that they get successively weaker: 1. $\Rightarrow$ 2. $\Rightarrow$ 3. To prove the proposition we first show the equivalence of all three conditions. Then we prove that the three conditions are also equivalent to experiment P being more informative than experiment Q. Given the first step, it is sufficient to prove this for any one of the conditions, and we prove it for condition 2.

Proof of the equivalence of 1., 2. and 3. We prove that condition 3 implies condition 1. Consider the following two player zero sum game. Player 1 chooses an $N_2 \times n$ matrix D where $0 \leq d_{ji} \leq 1$ for all j, i. Thus, player 1 chooses a decision rule based on observing experiment Q. Player 2 chooses an $N_1 \times N_2$ Markov matrix M. That is, player 2 chooses how experiment P is garbled. Player 1 seeks to maximize, and player 2 seeks to minimize, the trace of the matrix $(PM - Q)D$. This is the difference between the expected loss that a player would get when choosing D based on the garbled experiment PM and when choosing D based on Q. Expected values are

calculated using the uniform distribution, and omitting the normalization $1/n$. We denote this trace by $\Pi(D, M)$.

Nash's theorem on the existence of pure strategy Nash equilibria applies to this game. Denote by D_0 and M_0 equilibrium strategies, and define $v_o \equiv \Pi(D_0, M_0)$. By the definition of a pure strategy Nash equilibrium we have:

$$\Pi(D, M_0) \leq v_0 \leq \Pi(D_0, M)$$

for all D and all M.

Now apply condition 3 of Proposition 1 to decision rule D_0. There must then be some garbling of P that gives a loss not larger than QD_0. Therefore, given any decision rule, player 2 can achieve an expected payoff of no more than zero. This means that we have to have: $v_0 \leq 0$.

This means that for every decision rule D of player 1 we must have: $\Pi(D, M_0) \leq 0$. Now consider the matrix $PM_0 - Q$, that is, the difference between the probability distribution generated by the scrambled experiment PM_0 and the probability distribution generated by Q. Suppose any entry in this matrix were positive, that is, in some state, some experiment outcome were more likely under the scrambling than under Q. Then player 1 would be able to choose the decision rule D such that positive losses only occur only for that signal state combination, and other losses are zero. This would make $\Pi(D, M_0) > 0$, and would thus contradict that $\Pi(D, M_0) \leq 0$ for all D.

We conclude that all entries of $PM - Q$ must be non-negative. But then, because we are subtracting probabilities from each other, $PM = Q$. □

Proof that condition 2. is necessary and sufficient for $P \supset Q$.

$P \supset Q \Rightarrow$ Condition 2. Consider any decision rule D, that is, any $N_2 \times n$ matrix D. Let A be the set of all convex combinations of rows of D. Now apply the definition of $P \supset Q$ to conclude: $B(Q, A) \subseteq B(P, A)$. Note that D prescribes actions in A, and hence $b(Q, D) \in B(Q, A)$. Therefore, $b(Q, D) \in B(P, A)$. That means that there is a decision rule D' that is based on P such that $b(P, D') = b(Q, D)$. Every row of D' is a linear combination of rows of D, because D' only prescribes actions in A, and A is the convex hull of the rows of D. Therefore, there is an $N_1 \times N_2$ matrix M such that $D' = MD$. Hence, $PD' = PMD$ and therefore $b(P, MD) = b(Q, D)$, as had to be shown.

CONDITION 2 $\Rightarrow$ $P \supset Q$. Let A be any closed, bounded, and convex set of actions, and let D be a decision rule for experiment Q. We want to show that there is a decision rule D' for experiment P such that $b(Q, D) = b(P, D')$. By condition 2 there is an $N_1 \times N_2$ Markov matrix M such that $b(Q, D) = b(P, MD)$. Define $D' \equiv MD$. Because A is convex, MD is a decision rule for P, and hence has the required properties. □

Blackwell and Girshick's (1954) Theorem 12.2.2 offers two further conditions that are equivalent to $P \supset Q$. These conditions refer to the decision maker's posterior beliefs about Ω, if the decision maker's prior is uniform. Define Δ to be the set of all probability distributions over the state space. For every outcome $j = 1, \ldots, N_1$ of experiment P we define:

$$\alpha_j \equiv \sum_{i=1}^{n} P_{ij}$$

to be the ex ante probability of observing j. We omit here the normalizing factor $1/n$. We assume that for every j the probability α_j is strictly positive. The vector of the agent's posterior beliefs after seeing outcome j is then given by:

$$e_j \equiv \left(\frac{P_{1j}}{\alpha_j}, \ldots, \frac{P_{nj}}{\alpha_j} \right).$$

Notice, that we have left out the normalizing factor $1/n$ in the numerator as well as the denominator, so that it doesn't matter. Had we included it, it would have cancelled out. We define similarly for every outcome $k = 1, \ldots, N_2$ of experiment Q:

$$\beta_k \equiv \sum_{i=1}^{n} Q_{ik}$$

and

$$f_k \equiv \left(\frac{Q_{1k}}{\beta_k}, \ldots, \frac{Q_{nk}}{\beta_k} \right).$$

Proposition 2. *Let P and Q be two $n \times N_1$ and $n \times N_2$ Markov matrices. Each of the following conditions is equivalent to $P \supset Q$:*

4. There is an $N_2 \times N_1$ Markov matrix T such that:

(a) $\sum_{j=1}^{N_1} t_{kj} e_j = f_k$ *for all* $k = 1, \ldots, N_2$;

(b) $\sum_{k=1}^{N_2} \beta_k t_{kj} = \alpha_j$ *for all* $j = 1, \ldots, N_1$.

5. For every continuous convex function $\phi : \Delta \to \mathbb{R}$ we have:

$$\sum_{j=1}^{N_1} \alpha_j \phi(e_j) \geq \sum_{k=1}^{N_2} \beta_k \phi(f_k).$$

Condition 4 says that the posteriors that the decision maker may hold after observing the outcome of experiment P are a mean-preserving spread of the posteriors that the decision maker may hold after observing the outcome of experiment Q. Thus, a vector with expected value zero is added to the posterior after observing Q to obtain the posterior after observing P. Specifically, after observing outcome k of experiment Q and thus forming posterior f_k, with probability t_{kj} the vector $e_j - f_k$ is added to the posterior to obtain the posterior e_j. Condition (a) says that the expected value of this modification of e_j is zero. Condition (b) says that the distribution of the posterior that is obtained from e_j by adding noise with expected value zero is the same as the distribution of the posterior after observing P.

To understand why condition 5 is of interest we consider a special case of the convex functions ϕ referred to in condition 5. We obtain these functions as the result of optimally choosing actions. Imagine that the decision maker had a closed and bounded set of actions A available. Suppose the decision maker seeks to maximize expected utility, and the entries of each vector in A represented utilities rather than losses. Now imagine the decision maker observed the outcome of experiment P before having to choose. If the decision maker observed outcome j, then the decision maker would maximize $e_j \cdot a$. Define the value of this maximum to be:

$$\psi(e_j) \equiv \max_{a \in A} e_j \cdot a.$$

This is well defined because A is compact. Now we observe that ψ is a continuous and convex function of e_j. Continuity follows from the maximum theorem. To see convexity, suppose $e, e' \in \Delta$ and $\lambda \in [0, 1]$. We want to show:

$$\psi(\lambda e + (1 - \lambda)e') \leq \lambda \psi(e) + (1 - \lambda)\psi(e').$$

To see that this is true denote by $\hat{a}$ the $\arg\max_{a \in A}(\lambda e + (1\lambda)e') \cdot a$, and observe that:

$$\psi(\lambda e + (1 - \lambda)e') = \lambda e \cdot \hat{a} + (1 - \lambda)e' \cdot \hat{a}.$$

The assertion now follows from:

$$\psi(e) \geq e \cdot \hat{a} \text{ and } \psi(e') \geq e' \cdot \hat{a},$$

which are true by definition of ψ. Thus, ψ is an example of the functions to which condition 5 applies.

Condition 5 now says that the maximum expected payoff that the decision maker achieves when choosing actions after observing the outcome of P is at least as large as the maximum expected payoff that the decision maker achieves when choosing actions after observing the outcome of Q. This is the familiar condition which is often regarded as the decision theoretic definition of Blackwell comparisons. Note that condition 5 (implicitly) requires the assertion to be true for all decision problems where A is compact, and indeed it requires it to be true also for all convex functions ϕ that are not constructed as maximized expected utilities. However, as the proof below shows, the condition would also be necessary and sufficient if attention were restricted to convex functions ϕ that are constructed as maximized expected utilities. Thus, the proof of Proposition 2 and Proposition 1 together show the equivalence of the standard decision theoretic definition of Blackwell comparisons and condition 1. This is the most frequently cited version of Blackwell's result.

Proof. Proof that condition 1 implies condition 4: By definition:

$$f_k \equiv \left(\frac{Q_{1k}}{\beta_k}, \ldots, \frac{Q_{nk}}{\beta_k} \right).$$

Now by condition 1 for every i we have:

$$Q_{ik} = \sum_{j=1}^{N_1} P_{ij} M_{jk}.$$

Therefore:

$$\begin{aligned} \frac{Q_{ik}}{\beta_k} &= \frac{\sum_{j=1}^{N_1} P_{ij} M_{jk}}{\beta_k} \\ &= \sum_{j=1}^{N_1} \frac{\alpha_j M_{jk}}{\beta_k} \frac{P_{ij}}{\alpha_j}. \end{aligned}$$

This can be written as:

$$f_k = \sum_{j=1}^{N_1} t_{kj} e_j$$

where

$$t_{kj} = \frac{\alpha_j M_{jk}}{\beta_k}.$$

The interpretation of t_{kj} is as follows: Imagine that experiment Q is generated by scrambling P using M. Then t_{kj} is the conditional probability that experiment P had outcome j if experiment Q had outcome k. Note that clearly $\sum_{j=1}^{N_1} t_{kj} = 1$ as required

by the definition of a Markov matrix. We have thus shown part (a) of condition 4. Part (b) follows from this calculation:

$$\sum_{k=1}^{N_2} \beta_k t_{kj} = \sum_{k=1}^{N_2} \alpha_j M_{jk} = \alpha_j.$$

Proof that condition 4 implies condition 5: For every continuous convex function ϕ we have:

$$\begin{aligned}
\sum_{j=1}^{N_1} \alpha_j \phi(e_j) &= \sum_{j=1}^{N_1} \left(\sum_{k=1}^{N_2} \beta_k t_{kj} \right) \phi(e_j) \\
&= \sum_{k=1}^{N_2} \beta_k \sum_{j=1}^{N_1} t_{kj} \phi(e_j) \\
&\geq \sum_{k=1}^{N_2} \beta_k \phi \left(\sum_{j=1}^{N_1} t_{kj} e_j \right) \\
&= \sum_{k=1}^{N_2} \beta_k \phi(f_k).
\end{aligned}$$

where the first equality follows from part (2) of condition 4, the second equality rearranges the summation signs, the third inequality follows from the convexity of ϕ, and the last equality follows by definition of f_k.

Proof that condition 5 implies condition 3: Consider any decision rule D that is based on the outcome of experiment Q. We will construct the matrix M to which condition 3 refers. Define C to be the set of rows of D, and define $\phi : \Delta \to \mathbb{R}$ to be the function that assigns to every conditional belief about the state spaces the negative of the lowest expected loss (that is, the highest expected "payoff") that the decision maker can achieve if she is restricted to pick one of the actions in C. By the argument in the text following Proposition 2 this function ϕ is convex. Because C is not necessarily based on optimized choices given the observed Q we have:

$$\textit{Trace of } QD \geq -\sum_{k=1}^{N_2} \beta_k \phi(f_k).$$

Now construct the matrix M such that every outcome j of P is transformed with probability 1 into that outcome of Q that triggers the conditionally optimal choice from C given beliefs e_j. Then

$$\textit{Trace of } PMD = -\sum_{j=1}^{N_1} \alpha_j \phi(e_j).$$

Because ϕ is convex, condition 5 implies:

$$\sum_{j=1}^{N_1} \alpha_j \phi(e_j) \geq \sum_{k=1}^{N_2} \beta_k \phi(f_k),$$

and hence:

$$\textit{Trace of } PMD \leq \textit{Trace of } QD,$$

as required. □

Blackwell and Girshick offer two characterizations of the notion of "more informative" when the number of state is $n = 2$. The first is meant to provide a "systematic method" for deciding whether $P \subset Q$. Recall that e_j denotes the agent's posterior if outcome j of experiment P was observed. If there are only two states, this posterior can be identified with e_{1j} which we define to be the posterior probability of state 1. For any $t \in [0, 1]$ define $F_P(t)$ to be the probability that e_{1j} is at most t:

$$F_P(t) = \sum_{\{j | e_{1j} \leq t\}} \alpha_j.$$

For experiment Q we define $F_Q(t)$ analogously.

Proposition 3. *Suppose $n = 2$, and Let P and Q be two $n \times N_1$ and $n \times N_2$ Markov matrices. The following condition is equivalent to $P \supset Q$:*

6. $\int_0^t F_P(u)du \geq \int_0^t F_Q(u)du$ *for all* $t \in [0, 1]$.

Proof. Proof that condition 5 implies condition 6: We calculate the integrals in condition 6 and show that for fixed t each of these integrals is equal to the expected value of a convex function of the posterior. To calculate the integrals we use integration by parts:

$$\begin{aligned} \int_0^t F_P(u)du &= [uF_P(u)]_0^t - \int_0^t u dF_P(u) \\ &= tF_P(t) - \sum_{\{j | e_{1j} \leq t\}} \alpha_j e_{1j} \\ &= \sum_{\{j | e_{1j} \leq t\}} \alpha_j \left(t - e_{1j}\right). \end{aligned}$$

Now define the function $\phi : \Delta \to \mathbb{R}$ as follows:

$$\phi(e_j) \equiv \begin{cases} t - e_{1j} & \text{if } e_{1j} \leq t \\ 0 & \text{if } e_{1j} > t \end{cases}$$

Our calculation of the integral then shows that the integral equals the expected value of $\phi(e_j)$:

$$\int_0^t F_P(u)du \;\; = \;\; \sum_{j=1}^{N_1} \alpha_j \phi(e_j).$$

Thus, we can complete the proof by showing that ϕ is convex. This can be verified by drawing the graph of ϕ as a function of e_{1j} only.

Proof that condition 6 implies condition 5: The argument in the first part of the proof shows that condition 5 is equivalent to the requirement that for every function $\phi : \Delta \to \mathbb{R}$ where for some $t \in [0, 1]$:

$$\phi(e_j) \equiv \begin{cases} t - e_{1j} & \text{if } e_{1j} \leq t \\ 0 & \text{if } e_{1j} > t \end{cases}$$

we have:

$$\sum_{j=1}^{N_1} \alpha_j \phi(e_j) \geq \sum_{k=1}^{N_2} \alpha_j \phi(f_k).$$

Thus, condition 6 is the same as condition 5, but only for a subclass of convex functions. Our task is to verify that this subclass is sufficiently large to infer that the inequality holds for all convex functions ϕ.

First, we note that the inequality remains true if we multiply ϕ by a non-negative constant c, and add a function ℓ that is linear in e_j:

$$\sum_{j=1}^{N_1} \alpha_j \left[c\phi(e_j) + \ell(e_j)\right] \geq \sum_{k=1}^{N_2} \alpha_j \left[c\phi(f_k) + \ell(f_k)\right].$$

The reason is that obviously the non-negative constant leaves the inequality unchanged. Moreover, the expected value of a linear function equals the linear function evaluated at the expected value of the argument, and the expected values of e_j and f_k both equal the prior probability of state 1.

Next, we note that the inequality remains true if we consider a finite linear combination of functions of the form $c\phi(\cdot) + \ell(\cdot)$. We conclude the proof by noting that

all convex functions can be uniformly approximated as a finite linear combination of functions of this form.[2] □

[2]This is asserted in Blackwell and Girshick without proof. It sounds right.