MONASH University

MONASH BUSINESS SCHOOL

DEPARTMENT OF ECONOMETRCS AND BUSINESS STATISTICS

ISSN 1440-771X

WORKING PAPER SERIES

Information Loss in Bayesian Decision Problems: A Partial-Matching Perspective

Rami V. Tabri

Working Paper No. 13/26
November 2026

Information Loss in Bayesian Decision Problems: A Partial-Matching Perspective

Rami V. Tabri

Department of Econometrics and Business Statistics
Monash University, Melbourne, Australia

November 2026

Abstract

Processing an observation can destroy information without changing the optimal decision. What distinguishes such processing from a genuine increase in Bayes risk? I approach this question through partial matching of action-specific loss measures. Before processing, the residual components of these measures cannot be matched simultaneously at identical observations. After processing, however, residual losses originating from different observations may become compatible with the same processed observation. I show that the mass newly matched in this way equals the increase in minimum Bayes risk. A joint partial-transport formulation reveals where this loss arises and which original observations contribute to it, while its dual recovers the optimal decision based on the processed data. This provides a geometric explanation of decision-relevant information loss.

Keywords: Bayes risk, statistical decision theory, Blackwell comparison, information loss, partial matching, optimal transport.

1 Introduction

Processing an observation can discard information without compromising an optimal statistical decision. A complete sample may be replaced by a summary statistic, a continuous measurement may be grouped into intervals, or a recorded signal may be corrupted by noise. In some cases, the resulting observation retains everything needed to select an optimal action. In others, processing makes observations that favor different actions indistinguishable, increasing the minimum Bayes risk. What distinguishes these situations, and how can the resulting loss be understood geometrically?

The classical framework for studying this question is the comparison of statistical experiments initiated by Blackwell (1951, 1953). Blackwell's result implies that processing an observation through a parameter-independent Markov kernel, commonly called a garbling, cannot reduce minimum Bayes risk. The familiar argument is operational: a decision maker with access to the original observation can simulate the processed one and reproduce any decision procedure based on it. The comparison also has established geometric interpretations. In risk-set geometry, processing restricts the attainable risk vectors, potentially raising the minimum of a prior-weighted risk functional (e.g., Young and Smith, 2005). In likelihood-ratio geometry, processing acts through conditional averaging, connecting the comparison to convexity (Blackwell, 1951; Le Cam and Yang, 2000). More recently, Williamson and Cranko (2024, Theorems 23 and 40) developed a support-function representation of Bayes risk and an information-processing equality.

In this note, I develop a complementary geometric interpretation based on partial matching of measures. For a fixed Bayesian decision problem with finitely many actions, each action

determines a finite loss measure. The greatest mass that can be matched simultaneously across these measures at identical observations equals the minimum Bayes risk. Removing this common component leaves residual measures whose densities represent the additional conditional losses from choosing suboptimal actions. These residual measures have no simultaneous common mass at the original observations. After processing, however, residual losses associated with different original observations may become compatible with the same processed observation.

This observation suggests a different way to understand information loss: rather than beginning with an inequality between Bayes risks, identify the residual mass that becomes newly matchable through processing. I show that this mass defines a nonnegative measure on the processed observation space whose total mass equals the increase in minimum Bayes risk. A joint partial-transport formulation makes the original sources of this mass explicit, while its dual recovers the optimal decision rule. In particular, the approach explains why processing can destroy information without affecting an optimal decision: the transformed residual measures need not acquire any common mass.

The resulting perspective complements the classical geometry of statistical decisions. Risk-set and likelihood-ratio representations describe the consequences of processing through attainable risks and conditional averaging, respectively; partial matching describes them through the structure of action-specific loss measures. The connection also extends to Le Cam's deficiency, whose reconstruction problem admits a related interpretation through common submeasures. Building on the maximal-matching perspective of my Neyman–Pearson proof (Tabri, 2026), this note develops a geometric account of how statistical decisions are affected by the loss of information.

2 Bayes risk and a maximal partial-matching problem

I first describe the decision problem. Let Θ be a measurable parameter space and let $(\mathcal{X}, \mathcal{B}_X)$ be a standard Borel sample space. The experiment $\mathcal{E} = \{P_\theta : \theta \in \Theta\}$ is a measurable family of probability distributions on $\mathcal{X}$. Fix a prior Π and a finite action set $\mathsf{A} = \{1, \dots, k\}$, where $k \geq 2$. For each action a, let $L(\theta, a) \geq 0$ be real-valued, measurable and satisfy $\int L(\theta, a)\Pi(d\theta) < \infty$. A nonrandomized measurable decision rule $\delta : \mathcal{X} \to \mathsf{A}$ has Bayes risk

$$r_{\mathcal{E}}(\Pi, \delta) = \int_\Theta \int_{\mathcal{X}} L(\theta, \delta(x)) P_\theta(dx)\Pi(d\theta). \tag{1}$$

Its minimum is $V_{\mathcal{E}} = \inf_{\delta:\mathcal{X}\to\mathsf{A},\, \text{measurable}} r_{\mathcal{E}}(\Pi, \delta)$.

For each action, introduce the finite *loss-weighted measure*

$$\nu_a(B) = \int_\Theta L(\theta, a) P_\theta(B)\Pi(d\theta), \qquad B \in \mathcal{B}_X. \tag{2}$$

This measure assigns to an observed set the prior-weighted expected loss incurred if action a is chosen there. It need not have unit mass. If D_a is the set where a decision rule chooses action a, i.e., $D_a = \{x \in \mathcal{X} : \delta(x) = a\}$, then $\{D_a : a \in \mathsf{A}\}$ is a partition of $\mathcal{X}$ and $r_{\mathcal{E}}(\Pi, \delta) = \sum_a \nu_a(D_a)$. Thus the decision problem asks how the observation space should be partitioned among the action-specific loss measures.

Consider the opposite operation: instead of assigning observations to different actions, seek mass that can be taken from *all* the loss measures at the same observation. A measure γ is a *submeasure* of ν_a, written $\gamma \leq \nu_a$, if $\gamma(B) \leq \nu_a(B)$ for every measurable set B. To formulate the matching problem, consider the k-fold product space $\mathcal{X}^k$. Each point $(x_1, \dots, x_k) \in \mathcal{X}^k$ has one coordinate for each of the k action-specific loss measures. The diagonal of this product space is the set

$$\Delta_{\mathcal{X}} = \{(x_1, \dots, x_k) \in \mathcal{X}^k : x_1 = \cdots = x_k\}.$$

Thus, a point lies on the diagonal precisely when all k coordinates represent the same observation.

Write $\mathcal{M}_+(\mathcal{X}^k)$ for the finite nonnegative measures on $\mathcal{X}^k$. A measure $\pi \in \mathcal{M}_+(\mathcal{X}^k)$ describes a joint matching of mass from the k loss measures. Its ath marginal, denoted by π_a, records the mass assigned to the ath coordinate. The restriction $\pi_a \le \nu_a$ ensures that the matching does not use more mass than is available from action a. In optimal transport terminology, such a measure π is called a partial transport plan. The maximal partial matching problem is

$$\sup_{\substack{\pi \in \mathcal{M}_+(\mathcal{X}^k) \\ \pi_a \le \nu_a,\ \forall a}} \int_{\mathcal{X}^k} \mathbf{1}_{\Delta_{\mathcal{X}}}\, d\pi. \tag{3}$$

The measure π need not use all the mass available from each action-specific loss measure. The objective counts only the mass assigned to the diagonal, where all coordinates correspond to the same observation. To understand this matching, consider the measure on $\mathcal{X}$ defined by

$$\gamma_\pi(B) = \pi\big(\{(x, \dots, x) : x \in B\}\big), \qquad B \in \mathcal{B}_{\mathcal{X}}.$$

Thus, $\gamma_\pi(B)$ records the mass that π matches across all k loss measures at identical observations belonging to B. Since every coordinate agrees on the diagonal, γ_π is dominated by each marginal π_a, and hence by each ν_a.

Conversely, any measure γ on $\mathcal{X}$ satisfying $\gamma \le \nu_a$ for every a defines a feasible joint measure concentrated on the diagonal, by assigning mass $\gamma(B)$ to the points $(x, \dots, x)$ with $x \in B$. The matching problem therefore reduces to finding the greatest common submeasure of the action-specific loss measures.

Let $\rho = \sum_a \nu_a$ and $f_a = d\nu_a/d\rho$. Denote the greatest common submeasure by $\gamma^* = \bigwedge_a \nu_a$; explicitly,

$$\mathrm{d}\gamma^* = \min_a f_a \,\mathrm{d}\rho, \qquad \gamma^*(\mathcal{X}) = \int_{\mathcal{X}} \min_a f_a \,\mathrm{d}\rho. \tag{4}$$

The symbol $\wedge$ denotes the greatest common submeasure, not an intersection of sets. Thus the common mass is the largest amount that can be matched *simultaneously* across all k loss measures along the diagonal.

This is a zero-cost instance of partial optimal transport. In the two-marginal Euclidean problem with a quadratic transportation cost, Figalli (2010) studies the problem of matching a *prescribed* amount of mass. The greatest common mass is the threshold up to which that amount can be matched on the diagonal at zero cost. Beyond that threshold, matching requires transport between distinct points; Figalli investigates this richer regime and the geometry of active transport regions. The present problem is different: I determine the maximal zero-cost mass for a finite collection of loss measures and ask how that threshold changes when observations are processed. No Euclidean metric is needed for this question.

Because a decision rule must select an action at each observation, it is natural to relate the maximal common mass to the minimum expected loss. The following proposition establishes that these quantities coincide and shows that the dual variables of the partial-matching problem have a direct statistical interpretation as randomized decision rules.

Proposition 1 (Matching and Bayes decisions)**.** *The optimal Bayes risk, $V_{\mathcal{E}}$, equals the maximal mass that can be matched simultaneously across the action-specific loss measures at identical observations. Moreover, the maximal partial-matching problem* (3) *admits the following normalized dual representation:*

$$\begin{aligned} V_{\mathcal{E}} &= \sup_{\substack{\pi \in \mathcal{M}_+(\mathcal{X}^k) \\ \pi_a \le \nu_a,\ a \in \mathsf{A}}} \int_{\mathcal{X}^k} \mathbf{1}_{\Delta_{\mathcal{X}}}\, \mathrm{d}\pi \\ &= \inf_{\substack{\psi_a : \mathcal{X} \to [0,1] \\ \sum_{a=1}^k \psi_a = 1}} \sum_{a=1}^k \int_{\mathcal{X}} \psi_a \,\mathrm{d}\nu_a \\ &= \gamma^*(\mathcal{X}). \end{aligned} \tag{5}$$

Both the primal and dual problems attain their optimal values. A normalized dual solution $\psi = (\psi_1, \ldots, \psi_k)$ is optimal if and only if, for ρ-almost every x, it assigns positive probability only to actions minimizing $f_a(x) = d\nu_a/d\rho(x)$. In particular, a nonrandomized optimal Bayes decision rule exists.

Proof. I first derive the dual problem. Let $u_1, \ldots, u_k$ be nonnegative measurable functions on $\mathcal{X}$ satisfying $\sum_{a=1}^k u_a(x) \geq 1,\ x \in \mathcal{X}$. For every $(x_1, \ldots, x_k) \in \mathcal{X}^k$,

$$\mathbf{1}_{\Delta_\mathcal{X}}(x_1, \ldots, x_k) \leq \sum_{a=1}^k u_a(x_a).$$

Indeed, the sum is at least one when all coordinates coincide, and is nonnegative otherwise. Consequently, every feasible joint measure π in (3) satisfies

$$\begin{aligned}\int_{\mathcal{X}^k} \mathbf{1}_{\Delta_\mathcal{X}} \,\mathrm{d}\pi &\leq \sum_{a=1}^k \int_{\mathcal{X}^k} u_a(x_a) \,\mathrm{d}\pi \\ &= \sum_{a=1}^k \int_\mathcal{X} u_a \,\mathrm{d}\pi_a \\ &\leq \sum_{a=1}^k \int_\mathcal{X} u_a \,\mathrm{d}\nu_a.\end{aligned}$$

The last inequality follows from the marginal restrictions $\pi_a \leq \nu_a$. Minimizing this upper bound gives the dual problem

$$D = \inf_{\substack{u_a \geq 0 \\ \sum_a u_a \geq 1}} \sum_{a=1}^k \int_\mathcal{X} u_a \,\mathrm{d}\nu_a.$$

This dual can be normalized without changing its value. For any feasible collection $(u_1, \ldots, u_k)$ with finite objective, let $s(x) = \sum_a u_a(x) \geq 1$ and define $\psi_a(x) = u_a(x)/s(x)$. Then $\psi_a \geq 0$, $\sum_a \psi_a = 1$, and $\psi_a \leq u_a$. Since the measures ν_a are nonnegative, normalization cannot increase the objective. Conversely, every normalized collection is feasible for the original dual. Thus

$$D = \inf_{\substack{\psi_a \geq 0 \\ \sum_a \psi_a = 1}} \sum_{a=1}^k \int_\mathcal{X} \psi_a \,\mathrm{d}\nu_a.$$

To evaluate this infimum, recall that $\rho = \sum_a \nu_a$ and $f_a = d\nu_a/d\rho$. For every normalized collection,

$$\begin{aligned}\sum_{a=1}^k \int_\mathcal{X} \psi_a \,\mathrm{d}\nu_a &= \int_\mathcal{X} \sum_{a=1}^k \psi_a(x) f_a(x) \,\mathrm{d}\rho(x) \\ &\geq \int_\mathcal{X} \min_a f_a(x) \,\mathrm{d}\rho(x).\end{aligned}$$

Equality is attained by assigning all weight to an action with the smallest density $f_a(x)$. Because the action set is finite, choosing the smallest index among the minimizers gives a measurable selection. Therefore,

$$D = \int_\mathcal{X} \min_a f_a \,\mathrm{d}\rho = \gamma^*(\mathcal{X}).$$

It remains to establish attainment in the primal problem. Let $d : \mathcal{X} \to \mathcal{X}^k$ be the diagonal mapping $d(x) = (x, \ldots, x)$, and define $\pi^* = \gamma^* \circ d^{-1}$. This joint measure is concentrated on $\Delta_\mathcal{X}$, and each of its marginals equals $\gamma^* \leq \nu_a$. Hence π^* is feasible and

$$\int_{\mathcal{X}^k} \mathbf{1}_{\Delta_\mathcal{X}} \,\mathrm{d}\pi^* = \gamma^*(\mathcal{X}) = D.$$

Together with the dual upper bound, this proves equality and attainment of the primal and dual values.

Finally, for a measurable decision rule δ, the Bayes risk equals $r_{\mathcal{E}}(\Pi, \delta) = \int_{\mathcal{X}} f_{\delta(x)}(x)\, \mathrm{d}\rho(x)$. Choosing an action with the smallest density at each observation therefore attains $\gamma^*(\mathcal{X})$ and establishes $V_{\mathcal{E}} = D$. □

The normalized potentials in (5) are randomized decision rules: $\psi_a(x)$ is the probability of choosing action a after observing x. A deterministic rule chooses one action with probability one. Since each action-specific loss measure contains the same component γ^*, this component contributes $\gamma^*(\mathcal{X})$ to the Bayes risk under every decision rule. The remaining contribution depends on the selected actions. An optimal rule eliminates this additional loss entirely, so its Bayes risk equals $\gamma^*(\mathcal{X})$.

Define the *residual loss measures*

$$R_a = \nu_a - \gamma^*, \qquad a \in \mathsf{A}. \tag{6}$$

Since the sets D_a defined above form a measurable partition of $\mathcal{X}$, the Bayes-risk objective admits the decomposition

$$r_{\mathcal{E}}(\Pi, \delta) = \sum_{a=1}^{k} \nu_a(D_a) = \gamma^*(\mathcal{X}) + \sum_{a=1}^{k} R_a(D_a). \tag{7}$$

Thus, the common component contributes the same amount to the Bayes risk under every decision rule. Only the residual contribution depends on the selected actions. Proposition 1 shows that an optimal rule makes this additional contribution zero.

The residual measures satisfy $\bigwedge_{a\in\mathsf{A}} R_a = 0$, where the right-hand side denotes the zero measure. Indeed, their densities satisfy $\dfrac{dR_a}{d\rho}(x) = f_a(x) - \min\limits_{b\in\mathsf{A}} f_b(x)$, and consequently $\min\limits_{a\in\mathsf{A}} \dfrac{dR_a}{d\rho}(x) = 0$ for ρ-almost every x. Thus, at almost every observation, at least one action has zero residual density, although the action may depend on the observation. Consequently, no positive mass can be matched simultaneously across all the residual measures at identical observations. For two actions, this condition is equivalent to mutual singularity of the residual measures, as in the partial-matching derivation of the Neyman–Pearson lemma in Tabri (2026). With three or more actions, pairwise overlap may remain even though no positive mass can be matched simultaneously across all residual measures.

To interpret the residual measures statistically, consider the joint prior–sampling probability measure

$$\mathbb{P}_\Pi(d\theta, dx) = \Pi(d\theta) P_\theta(dx)$$

on $\Theta \times \mathcal{X}$. Its marginal distribution on $\mathcal{X}$ is the prior predictive measure $Q_X(B) = \int_\Theta P_\theta(B)\Pi(d\theta)$. Relative to Q_X, the residual measures have densities

$$\begin{aligned} \frac{dR_a}{dQ_X}(x) &= m_a(x) - \min_{b\in\mathsf{A}} m_b(x), \text{ where} \\ m_a(x) &= \mathbb{E}_{\mathbb{P}_\Pi}[L(\theta, a) \mid X = x]. \end{aligned} \tag{8}$$

Thus, the residual density is the additional conditional expected loss, or *regret*, from selecting action a instead of an optimal action at observation x. It vanishes wherever action a is optimal, up to Q_X-null sets. This interpretation will be useful when the observation is processed.

3 Processing creates new residual matches

Let K be a Markov kernel from $\mathcal{X}$ to another standard Borel space $\mathcal{Y}$. Conditional on $X = x$, the reported observation Y has distribution $K(\cdot \mid x)$, independently of the parameter. For any finite nonnegative measure μ on $\mathcal{X}$, its image under K is the measure on $\mathcal{Y}$ defined by

$$(\mu K)(B) = \int_{\mathcal{X}} K(B \mid x)\, \mu(dx), \qquad B \in \mathcal{B}_{\mathcal{Y}}.$$

Since $K(\mathcal{Y} \mid x) = 1$ and $K(B \mid x) \geq 0$, the kernel preserves total mass and domination: $(\mu K)(\mathcal{Y}) = \mu(\mathcal{X})$, and $\mu \leq \nu$ implies $\mu K \leq \nu K$. The processed experiment $\mathcal{F} = \mathcal{E}K$ has distributions $P_\theta K$, and its action-specific loss measures are $\nu_a K$. The minimum Bayes risk under the processed experiment is

$$V_{\mathcal{F}} = \inf_{\delta:\mathcal{Y}\to\mathsf{A}} \int_\Theta \int_{\mathcal{Y}} L(\theta, \delta(y))(P_\theta K)(dy)\Pi(d\theta),$$

where the infimum is over measurable decision rules based on the reported observation.

Observing that $\gamma^* \leq \nu_a$ for every action a, the measure $\gamma^* K$ remains a common submeasure of the processed loss measures. However, it need not be their greatest common submeasure. To identify the additional common mass created by processing, define the finite measure

$$\Xi_K = \left(\bigwedge_{a\in\mathsf{A}} \nu_a K \right) - \gamma^* K. \tag{9}$$

This measure is nonnegative because $\gamma^* K \leq \nu_a K$ for every action. Moreover, applying K to the decomposition $\nu_a = \gamma^* + R_a$ gives $\nu_a K = \gamma^* K + R_a K$. Since the same measure $\gamma^* K$ is contained in every transformed loss measure, taking their greatest common submeasure yields

$$\bigwedge_{a\in\mathsf{A}} \nu_a K = \gamma^* K + \bigwedge_{a\in\mathsf{A}} (R_a K).$$

This identity follows by taking the pointwise minimum of their densities with respect to a common dominating measure. Consequently,

$$\Xi_K = \bigwedge_{a\in\mathsf{A}} (R_a K). \tag{10}$$

Thus, Ξ_K is the greatest common submeasure of the transformed residual loss measures. The aim is to show that its total mass $\Xi_K(\mathcal{Y})$ equals the increase in minimum Bayes risk, $V_{\mathcal{F}} - V_{\mathcal{E}}$, and to explain how residual losses associated with different original observations contribute to this increase.

To that end, I next represent this matching by a joint measure. For each action a, define $S_a(dx, dy) = R_a(dx)K(dy \mid x)$ on $\mathcal{X} \times \mathcal{Y}$. It records the residual loss associated with action a at the original observation x together with the report y that can be generated from it. Its marginal on the reported observation is $R_a K$.

A *joint matching measure* Λ is a nonnegative measure on $\mathcal{X}^k \times \mathcal{Y}$. A point $(x_1, \ldots, x_k, y)$ records one original observation for each action and a report shared by all of them. For action a, consider the coordinate projection

$$q_{aY}(x_1, \ldots, x_k, y) = (x_a, y).$$

This is a measurable, surjective map onto $\mathcal{X} \times \mathcal{Y}$ when $\mathcal{X}$ is nonempty. The *marginal* Λ_{aY} is the image of Λ under q_{aY}: for a measurable $B \subseteq \mathcal{X} \times \mathcal{Y}$, it assigns mass $\Lambda(q_{aY}^{-1}(B))$ to B. Constructing Λ with specified or bounded marginals is sometimes called *lifting* those marginals

to a joint space; it does not mean reconstructing the original observation from the report. I require $\Lambda_{aY} \le S_a$ for every action, so that no joint match uses more residual mass than is available from any action. The resulting problem is

$$P_K = \sup_{\substack{\Lambda \in \mathcal{M}_+(\mathcal{X}^k \times \mathcal{Y}) \\ \Lambda_{aY} \le S_a,\ \forall a}} \Lambda(\mathcal{X}^k \times \mathcal{Y}). \tag{11}$$

Unlike (3), this joint measure need not match identical original observations. It may draw the residual for each action from a different x_a, provided every contribution is compatible with the *same* report y. This is the geometric mechanism of information loss: the observation channel can make distinct sources of residual loss indistinguishable to the decision maker.

Theorem 1 (Residual matching under data processing)**.** *For the finite-action Bayesian problem above and any Markov kernel K, the following conclusions hold.*

The measure Ξ_K in (9) *is nonnegative and equals the greatest common transformed residual measure in* (10)*. The lifted partial-matching problem* (11) *attains its maximum, and*

$$P_K = \Xi_K(\mathcal{Y}) = V_{\mathcal{F}} - V_{\mathcal{E}}. \tag{12}$$

Moreover, its normalized dual is

$$D_K = \inf_{\substack{\psi_a : \mathcal{Y} \to [0,1] \\ \sum_a \psi_a = 1}} \sum_{a=1}^{k} \int_{\mathcal{Y}} \psi_a(y)\, (R_a K)(dy), \tag{13}$$

with $P_K = D_K$ and attainment on both sides. An optimal dual potential selects only actions minimizing the conditional residual loss given Y. For an optimal primal plan Λ^* and dual potential ψ^*,

$$\int_{\mathcal{X} \times \mathcal{Y}} \psi_a^*(y)\ \mathrm{d}(S_a - \Lambda_{aY}^*)(x, y) = 0 \quad (a \in \mathsf{A}). \tag{14}$$

Proof. The relation (10) follows by substituting $\nu_a = \gamma^* + R_a$ into (9). Taking total masses and using Proposition 1 gives $\Xi_K(\mathcal{Y}) = V_{\mathcal{F}} - V_{\mathcal{E}}$.

For any feasible Λ, its Y-marginal η is dominated by every $R_a K$. Hence $\eta \le \Xi_K$ and $P_K \le \Xi_K(\mathcal{Y})$. Conversely, the common reported marginal Ξ_K can be placed back on the joint space using the conditional distributions of the original observations. On standard Borel spaces, disintegrate S_a in the form

$$S_a(dx, dy) = H_a(dx \mid y)(R_a K)(dy).$$

To attain the upper bound, I construct a joint measure by gluing the conditional distributions $H_a(\cdot \mid y)$ over their greatest common reported submeasure Ξ_K. One such construction takes the original observations to be conditionally independent given the common report y, yielding

$$\Lambda^*(dx_1, \dots, dx_k, dy) = \Xi_K(dy) \prod_{a=1}^{k} H_a(dx_a \mid y). \tag{15}$$

For each action a, the corresponding (x_a, y)-marginal is

$$\Lambda_{aY}^*(dx_a, dy) = H_a(dx_a \mid y)\Xi_K(dy) \le S_a(dx_a, dy),$$

where the inequality follows from $\Xi_K \le R_a K$. Thus, Λ^* is feasible and has total mass $\Xi_K(\mathcal{Y})$. This establishes primal attainment. Conditional independence is used only to construct one optimal plan; it is not a requirement of the matching problem.

The normalized dual follows by the same marginal-domination argument as in Proposition 1. Let ψ_a be nonnegative measurable functions on $\mathcal{Y}$ satisfying $\sum_a \psi_a(y) = 1$. For any feasible Λ, its $\mathcal{Y}$-marginal η satisfies $\eta \leq R_a K$ for every action. Consequently,

$$\Lambda(\mathcal{X}^k \times \mathcal{Y}) = \sum_{a=1}^{k} \int_{\mathcal{Y}} \psi_a \, d\eta \leq \sum_{a=1}^{k} \int_{\mathcal{Y}} \psi_a \, d(R_a K). \tag{16}$$

This gives the normalized dual upper bound in (13).

Writing $s_a = d(R_a K)/dQ_Y$, the dual objective is minimized pointwise by selecting an action with the smallest $s_a(y)$, breaking ties measurably. Its minimum is

$$\int_{\mathcal{Y}} \min_a s_a(y) \, Q_Y(dy) = \Xi_K(\mathcal{Y}).$$

Since the primal construction attains this same value, the normalized dual is exact, and both problems attain their optima. Equality in (16) then yields the complementary-slackness condition (14). □

The dual has a direct statistical meaning. To make the conditioning explicit, let

$$\mathbb{P}_{\Pi,K}(d\theta, dx, dy) = \Pi(d\theta) P_\theta(dx) K(dy \mid x)$$

denote the joint prior–sampling–processing probability measure. Its marginal distribution on $\mathcal{Y}$ is $Q_Y = Q_X K$. Using (8), the transformed residual density satisfies

$$s_a(y) = \mathbb{E}_{\mathbb{P}_{\Pi,K}}\left[m_a(X) - \min_{b \in \mathsf{A}} m_b(X) \mid Y = y \right], \; Q_Y\text{-a.e.} \tag{17}$$

All actions share the same subtracted conditional term. Consequently, an action minimizing $s_a(y)$ also minimizes $\mathbb{E}_{\mathbb{P}_{\Pi,K}}[L(\theta, a) \mid Y = y]$. Thus, the optimal dual rule is the Bayes rule based on the reported observation. An action attaining $\min_a s_a(y)$ has the smallest transformed residual density at y and therefore limits the amount of mass that can be matched simultaneously across all actions. I refer to such an action as a *limiting action*. Complementary slackness shows that, wherever an optimal dual rule assigns positive probability to a limiting action, the matching plan exhausts the residual measure available for that action. Thus, the optimal statistical decision also identifies which residual loss measure limits the common matching.

The joint matching has two further properties. First, every optimal plan has $\mathcal{Y}$-marginal Ξ_K. Thus, the distribution of newly matched residual mass over the reported observations is uniquely determined, even though the joint distribution of the original observations need not be.

Second, the original observations contributing to a feasible plan cannot all coincide. To see this, let

$$G_a = \left\{ x \in \mathcal{X} : \frac{dR_a}{d\rho}(x) > 0 \right\}.$$

Each residual measure R_a is concentrated on G_a. As established in Section 2, $\rho\left(\bigcap_{a \in \mathsf{A}} G_a\right) = 0$, because at almost every observation at least one residual density vanishes. Moreover, the x_a-marginal of any feasible plan Λ is dominated by R_a. Consequently, Λ assigns mass only to tuples for which $x_a \in G_a$ for every action a, up to null sets. The coordinates therefore cannot all coincide on a set of positive Λ-measure.

Thus, simultaneous residual matching after processing must involve at least two distinct original observations that are compatible with the same report. For $k \geq 3$, the sets G_a may overlap pairwise; what is excluded is their simultaneous overlap on a set of positive ρ-measure.

4 Decision preservation and successive processing

Theorem 1 shows that data processing increases minimum Bayes risk precisely when it creates a common component among the transformed residual loss measures. However, processing may make distinct original observations indistinguishable without affecting the minimum Bayes risk for the decision problem at hand. The following proposition characterizes when the reported observation retains sufficient information to select an action that was optimal before processing.

Proposition 2 (Decision preservation)**.** *Let $A^*(x) = \{a : m_a(x) = \min_b m_b(x)\}$ be the set of optimal original-data actions, and let $J(dx, dy) = Q_X(dx)K(dy \mid x)$ be the joint prior predictive law. The following are equivalent: $V_{\mathcal{F}} = V_{\mathcal{E}}$; $\Xi_K = 0$; and there exists a measurable decision rule $\delta : \mathcal{Y} \to \mathsf{A}$ such that*

$$\delta(Y) \in A^*(X) \qquad J\text{-almost surely.} \tag{18}$$

Equivalently, for Q_Y-almost every y, the original observations conditionally compatible with y have at least one common optimal action, up to conditional null sets.

Proof. For each action a, let

$$r_a(x) = m_a(x) - \min_{b \in \mathsf{A}} m_b(x)$$

denote its nonnegative conditional regret. Since J is the (X, Y)-marginal of $\mathbb{P}_{\Pi,K}$, equation (17) gives

$$s_a(y) = \mathbb{E}_J[r_a(X) \mid Y = y], \qquad Q_Y\text{-a.e.}$$

By (10), the density of Ξ_K with respect to Q_Y is $\min_{a \in \mathsf{A}} s_a$. Consequently, $\Xi_K = 0$ if and only if

$$\min_{a \in \mathsf{A}} s_a(y) = 0 \qquad Q_Y\text{-a.e.}$$

Since A is finite, a measurable decision rule δ can be constructed by selecting an action minimizing $s_a(y)$, breaking ties by the smallest index. For this rule,

$$\mathbb{E}_J[r_{\delta(Y)}(X) \mid Y] = \min_{a \in \mathsf{A}} s_a(Y) = 0 \qquad J\text{-a.s.}$$

Because conditional regrets are nonnegative, this implies $r_{\delta(Y)}(X) = 0$ J-almost surely, or equivalently, $\delta(Y) \in A^*(X)$ J-almost surely.

Conversely, if a measurable rule δ satisfies (18), then $r_{\delta(Y)}(X) = 0$ J-almost surely. Taking conditional expectations given Y shows that $s_{\delta(y)}(y) = 0$ for Q_Y-almost every y, and hence $\Xi_K = 0$. Finally, the equivalence with $V_{\mathcal{F}} = V_{\mathcal{E}}$ follows from (12). □

Thus, preserving minimum Bayes risk does not require the reported observation to determine the original observation. It suffices that, for Q_Y-almost every report y, there exists an action that is optimal at almost every original observation conditional on that report. When this condition fails, processing creates positive simultaneous overlap among the transformed residual loss measures, increasing the minimum Bayes risk. Consequently, a transformation may be strictly less informative in Blackwell's ordering while preserving the minimum Bayes risk for a particular decision problem.

Preserving an optimal action under data processing is a weaker requirement than sufficiency of the processing kernel K for the experiment $\mathcal{E}$. The latter requires that the original experiment can be reconstructed from the processed experiment through a Markov kernel, and hence preserves minimum Bayes risk for every prior and loss function. Proposition 2, however, requires only that the processed observation permit the selection of an action that was optimal under the original observation, for the decision problem at hand.

A second consequence concerns successive data processing. An observation may first be summarized and then perturbed by noise, with each transformation potentially creating additional

common mass among the residual loss measures. The measure-valued formulation allows the contributions of the two stages to be distinguished.

Suppose the original observation is processed first by $K_1 : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and then by $K_2 : \mathcal{Y} \rightsquigarrow \mathcal{Z}$. Write $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)$ for the original action-specific loss measures and $\boldsymbol{\nu}K_1 = (\nu_1 K_1, \ldots, \nu_k K_1)$ for their images after the first transformation. I use $\Xi_{K_2}(\boldsymbol{\nu}K_1)$ to denote the measure in (9) formed from these first-stage images.

Let $\gamma_1 = \bigwedge_{a \in \mathsf{A}} (\nu_a K_1)$ be their greatest common submeasure. Adding and subtracting $\gamma_1 K_2$ gives

$$\begin{aligned} \Xi_{K_1 K_2}(\boldsymbol{\nu}) &= \bigwedge_{a \in \mathsf{A}} (\nu_a K_1 K_2) - \gamma^* K_1 K_2 \\ &= \left[\bigwedge_{a \in \mathsf{A}} (\nu_a K_1 K_2) - \gamma_1 K_2 \right] \\ &\quad + [\gamma_1 K_2 - \gamma^* K_1 K_2] \\ &= \Xi_{K_2}(\boldsymbol{\nu}K_1) + \Xi_{K_1}(\boldsymbol{\nu}) K_2. \end{aligned} \tag{19}$$

Both terms are nonnegative measures on $\mathcal{Z}$. The first represents the additional common residual mass created by the second transformation. The second represents the common residual mass created at the first stage and subsequently carried through K_2.

Taking total masses yields the sum of the increases in minimum Bayes risk at the two stages. More importantly, (19) describes how the distribution of newly matched residual mass over the reported observation space evolves through successive processing.

5 An illustration with three actions

Consider three equally likely states, $\Theta = \{1, 2, 3\}$. The original experiment $\mathcal{E}$ reveals the state: observation x_i occurs with probability one under state i. There are three actions, with loss matrix

	$a = 1$	$a = 2$	$a = 3$
x_1	0	0	1
x_2	1	0	0
x_3	0	1	0

.

At x_1, actions 1 and 2 are optimal; at x_2, actions 2 and 3 are optimal; and at x_3, actions 1 and 3 are optimal. The optimal-action sets intersect pairwise, but no action is optimal at all three observations.

Consider two successive, deterministic processing steps. First, K_1 combines x_1 and x_2 into the report y_{12} while retaining x_3 as y_3. The resulting experiment $\mathcal{F}_1 = \mathcal{E}K_1$ still permits zero Bayes risk: choose action 2 at y_{12} and action 1 or 3 at y_3. Next, K_2 combines y_{12} and y_3 into a single report y, producing $\mathcal{F}_2 = \mathcal{E}K_1K_2$. Now every action has expected loss 1/3. Therefore

$$V_{\mathcal{E}} = V_{\mathcal{F}_1} = 0, \text{ and } V_{\mathcal{F}_2} = \tfrac{1}{3}.$$

The first transformation reduces the set of available decisions without increasing the minimum Bayes risk; the second makes an additional loss unavoidable.

Risk-set geometry (e.g.,Young and Smith, 2005, Section 2.5) provides one way to see this distinction. For an experiment $\mathcal{G}$, let $\mathcal{S}_{\mathcal{G}}(L) \subseteq [0,1]^3$ be the set of vectors (r_1, r_2, r_3) of frequentist risks attainable by randomized rules under the displayed loss matrix. Direct calculation gives

$$\begin{aligned} \mathcal{S}_{\mathcal{E}}(L) &= [0,1]^3, \\ \mathcal{S}_{\mathcal{F}_1}(L) &= \{r \in [0,1]^3 : r_1 + r_2 \leq 1\}, \\ \mathcal{S}_{\mathcal{F}_2}(L) &= \{r \in [0,1]^3 : r_1 + r_2 + r_3 = 1\}. \end{aligned} \tag{20}$$

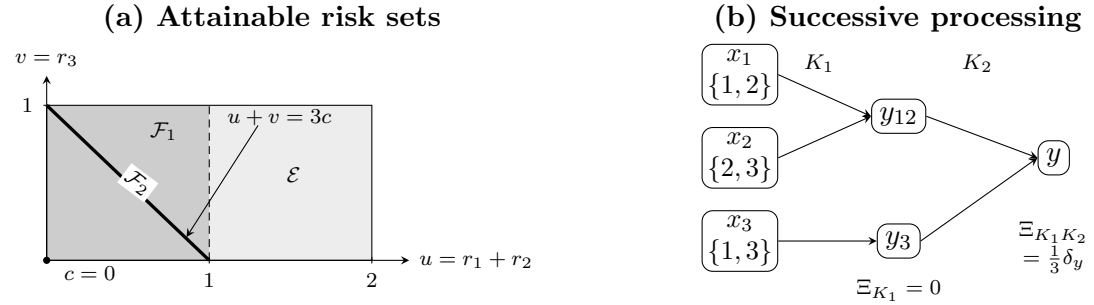

Figure 1: Two views of the same information loss. (a) Exact projections of the nested attainable risk sets under the map $(r_1, r_2, r_3) \mapsto (u, v) = (r_1 + r_2, r_3)$ for the stated loss matrix. The uniform-prior Bayes-risk levels are the lines $u + v = 3c$: the first two sets admit $c = 0$, while the third admits only $c = 1/3$. (b) Pooling x_1 with x_2 preserves an optimal action and creates no simultaneous residual match; combining the two remaining reports creates matched residual mass $1/3$. Braces display the optimal actions at each original observation.

With perfect observation, actions can be randomized separately at each state. Under $\mathcal{F}_1$, if $\psi_a(y_{12})$ is the probability of action a at the pooled report, then $r_1 = \psi_3(y_{12})$ and $r_2 = \psi_1(y_{12})$, while r_3 can be varied independently. Under $\mathcal{F}_2$, the risk vector is $(\psi_3(y), \psi_1(y), \psi_2(y))$, whose coordinates sum to one. These convex sets are nested. For the uniform prior, Bayes risk is $(r_1 + r_2 + r_3)/3$. Thus its level sets are parallel planes in three dimensions. To depict them in two dimensions, set $u = r_1 + r_2$ and $v = r_3$. The exact images of the three risk sets are the rectangles $[0, 2] \times [0, 1]$ and $[0, 1]^2$, and the line segment $\{(u, v) \in [0, 1]^2 : u + v = 1\}$, respectively. In Figure 1(a), the level lines $u + v = 3c$ touch the first two sets at the origin, but touch the third at $c = 1/3$.

The partial-matching calculation explains the same change in terms of the loss-weighted measures. The original common component vanishes, and

$$R_1 = \tfrac{1}{3}\delta_{x_2}, \qquad R_2 = \tfrac{1}{3}\delta_{x_3}, \qquad R_3 = \tfrac{1}{3}\delta_{x_1}.$$

After K_1, the three residual measures are, respectively, $\frac{1}{3}\delta_{y_{12}}$, $\frac{1}{3}\delta_{y_3}$, and $\frac{1}{3}\delta_{y_{12}}$. They still have no simultaneous common mass, so $\Xi_{K_1} = 0$. After both steps, all three become $\frac{1}{3}\delta_y$, giving

$$\Xi_{K_1K_2} = \tfrac{1}{3}\delta_y, \qquad \Xi_{K_1K_2}(\{y\}) = V_{\mathcal{F}_2} - V_{\mathcal{E}}.$$

One optimal joint matching measure places mass $1/3$ at (x_2, x_3, x_1, y): distinct original observations supply the three residual contributions to the same report. Figure 1(b) shows the successive pooling. In particular, this example illustrates the measure decomposition in (19): the first stage creates no new matching, whereas the second creates all of it.

6 Likelihood-ratio geometry and deficiency

The measure-valued representation of information loss can also be connected to the classical likelihood-ratio geometry of statistical experiments. Under data processing, likelihood-ratio vectors are averaged conditionally, and the resulting concavity gap coincides with the density of newly matched residual loss mass.

To make this connection precise, suppose $\Theta = \{1, \ldots, n\}$ and the prior weights π_i are positive. Let $S_{\mathcal{E}} = \sum_{i=1}^{n} P_i$ and define the normalized likelihood-ratio vector

$$U(x) = \left(\frac{dP_1}{dS_{\mathcal{E}}}(x), \ldots, \frac{dP_n}{dS_{\mathcal{E}}}(x) \right),$$

which lies in the probability simplex. The distribution of U under $S_{\mathcal{E}}$ is a finite-mass version of Blackwell's canonical measure. Define the concave, piecewise-affine function $h(u) = \min_{a\in\mathsf{A}} \sum_{i=1}^{n} \pi_i L(i,a) u_i$. Then the minimum Bayes risk can be written $V_{\mathcal{E}} = \int_{\mathcal{X}} h(U(x))\, S_{\mathcal{E}}(dx)$. Under garbling, the processed likelihood-ratio vector satisfies $U_Y(Y) = \mathbb{E}[U(X) \mid Y]$ under the joint law obtained by combining $S_{\mathcal{E}}/n$ with K. The posterior martingale description in Blackwell's Theorems 4 and 6 and the likelihood-ratio approach of Le Cam and Yang (2000, Section 2.3) are therefore directly applicable.

The measure-valued matching identity identifies the conditional concavity gap at each reported observation. With $S_{\mathcal{F}} = S_{\mathcal{E}}K$,

$$\frac{d\Xi_K}{dS_{\mathcal{F}}}(y) = h(U_Y(y)) - \mathbb{E}[h(U(X)) \mid Y = y],\ S_{\mathcal{F}}\text{-a.e.} \tag{21}$$

Indeed, the two terms on the right-hand side are the densities of $\bigwedge_{a\in\mathsf{A}}(\nu_a K)$ and $\gamma^* K$ with respect to $S_{\mathcal{F}}$, respectively. The right-hand side is nonnegative by conditional Jensen's inequality. Its integral is the increase in minimum Bayes risk, while the left-hand side identifies the same increase as newly matched residual loss mass. Thus, (21) connects the likelihood-ratio and partial-matching descriptions of information loss, rather than establishing a new convex-order result. In the binary zero–one-loss case, with $\Theta = \{0,1\}$, the greatest common mass before processing is

$$\pi_1(P_1 \wedge \lambda P_0)(\mathcal{X}), \qquad \lambda = \pi_0/\pi_1,$$

recovering the matching profile of Tabri (2026) as a special case.

Le Cam's deficiency (Le Cam, 1964) extends the comparison from a fixed Bayesian decision problem to a uniform assessment of how well one experiment can approximate another. Rather than fixing a prior and loss function, it considers reconstruction of the original sampling distributions uniformly across parameter values. For finite parameter and observation spaces, adopt the total-variation convention

$$\|P - Q\|_{\mathrm{TV}} = \sup_B |P(B) - Q(B)|$$

and define the directed deficiency from $\mathcal{F}$ to $\mathcal{E}$ by

$$\delta(\mathcal{F}, \mathcal{E}) = \inf_T \max_i \|Q_i T - P_i\|_{\mathrm{TV}},$$

where $Q_i = P_i K$ and T ranges over Markov kernels that attempt to reconstruct an $\mathcal{E}$-observation from an $\mathcal{F}$-observation.

For $\mathcal{F} = \mathcal{E}K$, the finite-experiment randomization criterion gives

$$\delta(\mathcal{F}, \mathcal{E}) = \sup_{\Pi, L, \mathsf{A}} \big(V_{\mathcal{F}} - V_{\mathcal{E}}\big) = \sup_{\Pi, L, \mathsf{A}} \Xi_K^{\Pi, L}(\mathcal{Y}), \tag{22}$$

where the supremum is over finite action sets, priors, and loss functions satisfying $0 \le L \le 1$. Here $\Xi_K^{\Pi,L}$ emphasizes the dependence of the residual-matching measure on the specified prior and loss function. Thus, deficiency is the largest increase in minimum Bayes risk, or equivalently the largest additional residual matching, over bounded Bayesian decision problems. This is the classical randomization characterization of deficiency (Torgersen, 1991, Theorems 6.3.4 and 6.4.1), expressed through the partial-matching identity.

In particular, for finite experiments, $\Xi_K^{\Pi,L} = 0$ for every bounded Bayesian decision problem if and only if $\delta(\mathcal{F}, \mathcal{E}) = 0$. By the randomization criterion (Torgersen, 1991, Theorem 6.4.1), the latter is equivalent to the existence of a Markov kernel $H : \mathcal{Y} \rightsquigarrow \mathcal{X}$ satisfying

$$P_i = (P_i K)H \qquad \text{for every } i.$$

Since $\mathcal{F} = \mathcal{E}K$, this establishes Blackwell equivalence of the two experiments. Thus, preservation of minimum Bayes risk across all bounded decision problems characterizes sufficiency of K for the experiment $\mathcal{E}$, whereas Proposition 2 characterizes preservation for a fixed decision problem.

The illustration in Section 5 shows why the maximization in (22) matters. For the displayed loss matrix and uniform prior, complete pooling raises minimum Bayes risk by $1/3$. However, the reverse deficiency of $\mathcal{F}_2$ relative to the perfectly revealing experiment $\mathcal{E}$ is $2/3$. Indeed, a reconstruction kernel applied to the constant report y chooses the original observation x_j with probability t_j, where $t_j \geq 0$ and $\sum_j t_j = 1$. Its total-variation error in state i is $1 - t_i$, and hence

$$\delta(\mathcal{F}_2, \mathcal{E}) = \min_{\substack{t_j \geq 0 \\ \sum_j t_j = 1}} \max_i (1 - t_i) = \tfrac{2}{3}.$$

The larger risk gap is attained, for example, with the same uniform prior but the state-identification loss

$$L_{\mathrm{id}}(i, a) = \mathbf{1}\{i \neq a\}.$$

Perfect observation gives zero Bayes risk, whereas the constant report gives $2/3$. Thus, Ξ_K quantifies information loss for a specified Bayesian decision problem, while deficiency takes the largest such loss over bounded decision problems.

There is also an optimal-transport interpretation of deficiency. For the discrete metric $d_0(x, x') = \mathbf{1}\{x \neq x'\}$, the Wasserstein-1 distance coincides with total variation:

$$W_{1,d_0}(P, Q) = 1 - (P \wedge Q)(\mathcal{X}) = \|P - Q\|_{\mathrm{TV}}.$$

Consequently,

$$\delta(\mathcal{F}, \mathcal{E}) = \inf_T \max_i W_{1,d_0}(Q_i T, P_i).$$

Equivalently, the maximal-coupling identity gives

$$\delta(\mathcal{F}, \mathcal{E}) = 1 - \sup_T \min_i (P_i \wedge Q_i T)(\mathcal{X}).$$

Thus, deficiency admits a minimax optimal-transport interpretation: it is the complement of the greatest worst-case common mass attainable through a reconstruction kernel shared across parameter values. This differs from the partial-matching formulation of Ξ_K, which concerns the action-specific residual loss measures of a fixed Bayesian decision problem. Both interpretations, however, express statistical information through the mass that can be matched under the relevant constraints.

7 Discussion

Blackwell's data-processing result compares minimum Bayes risks before and after an observation is transformed. The partial-matching approach explains this comparison through the structure of the action-specific loss measures. Their greatest common submeasure determines the minimum Bayes risk before processing, while the residual measures represent the additional conditional losses associated with choosing suboptimal actions. Although these residual measures have no simultaneous common mass at the original observations, processing can make some of their contributions compatible with the same report. The measure Ξ_K records this newly created common mass. Its total mass equals the increase in minimum Bayes risk, while its distribution over the reported observations and the associated joint transport plans identify where this increase arises and which original observations contribute to it.

The primal–dual formulation also connects this geometric interpretation to optimal statistical decisions. The normalized dual recovers the Bayes rule based on the reported observation, while complementary slackness identifies the residual loss measures that limit simultaneous matching.

Decision preservation has a corresponding interpretation: processing leaves minimum Bayes risk unchanged precisely when the reported observation still permits the selection of an action that was optimal under the original observation. Furthermore, the decomposition (19) separates residual mass made matchable at an earlier processing stage from additional matching created subsequently.

The three-action illustration also connects the partial-matching approach to the familiar convex geometry of statistical decisions. Data processing produces nested sets of attainable risk vectors. A fixed prior determines a linear functional on these sets, whose minimum may increase as the set of attainable risks contracts. The partial-matching identity explains the same increase through the additional common mass of the transformed residual loss measures. These are complementary geometric descriptions: one concerns the risks attainable by decision rules, while the other concerns the measures underlying the decision problem. The likelihood-ratio representation provides a further connection. Under processing, likelihood-ratio vectors undergo conditional averaging, and (21) identifies the resulting conditional concavity gap with the density of Ξ_K.

Le Cam's deficiency extends the comparison beyond a fixed prior and loss function by considering the largest increase in minimum Bayes risk over bounded decision problems. Its representation through Wasserstein-1 distance with the discrete metric gives another connection to optimal transport. In that formulation, the objective is to maximize the worst-case common mass between the original and reconstructed sampling distributions, subject to using the same reconstruction kernel across parameter values. This differs from the matching of action-specific residual loss measures considered here, but both formulations describe statistical information through common submeasures under appropriate constraints.

Beyond its geometric interpretation, the partial-matching perspective suggests an alternative route to computing decision-relevant information loss. Entropic regularization of the lifted multimarginal partial-transport problem (11), with explicit treatment of both matched and unmatched residual measures, suggests an information-projection formulation. This connects the proposed computational approach to the duality and approximation methods for information projections in Tabri (2026b) and the sieve-based estimation framework for entropic optimal transport in Tabri (2025). Establishing convergence guarantees and determining whether this approach yields computational advantages over direct evaluation of the greatest common submeasure remain questions for future research.

The partial-matching perspective thus provides a complementary primal–dual interpretation of classical statistical decision theory, connecting optimal Bayes risk, decision preservation, and information loss through common submeasures and residual loss measures. Its connection to Figalli's prescribed-mass partial transport concerns the threshold for matching at zero cost, rather than the free-boundary geometry of positive-cost Euclidean transport. The present construction assumes finitely many actions, although the parameter and observation spaces need not be finite. It also complements the maximal-matching interpretation of the Neyman–Pearson lemma in Tabri (2026), shifting the focus from the optimality of a statistical test to the geometry of decision-relevant information loss.

References

Blackwell, D. (1951). Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 93–102. University of California Press.

Blackwell, D. (1953). Equivalent comparisons of experiments. *Annals of Mathematical Statistics* **24**, 265–272.

Figalli, A. (2010). The optimal partial transport problem. *Archive for Rational Mechanics and Analysis* **195**, 533–560.

Le Cam, L. (1964). Sufficiency and approximate sufficiency. *Annals of Mathematical Statistics* **35**, 1419–1455.

Le Cam, L. and Yang, G. L. (2000). *Asymptotics in Statistics: Some Basic Concepts*, 2nd ed. Springer.

Tabri, R. V. (2025). A sieve M-estimator for entropic optimal transport. Preprint, arXiv:2512.21981.

Tabri, R. V. (2026). A partial-matching proof of the Neyman–Pearson lemma. Preprint, arXiv:2611.07368.

Tabri, R. V. (2026b). The information projection in moment inequality models: Existence, dual representation, and approximation. *Mathematics of Operations Research* **51**(3), 2468–2494.

Torgersen, E. (1991). *Comparison of Statistical Experiments*. Cambridge University Press.

Williamson, R. C. and Cranko, Z. (2024). Information processing equalities and the information–risk bridge. *Journal of Machine Learning Research* **25**(103), 1–53.

Young, G. A. and Smith, R. L. (2005). *Essentials of Statistical Inference*. Cambridge University Press.