

## APPENDIX

## A PRELIMINARY: MATRICES AND KERNELS

This section provides basic definitions and theorems for matrices and kernels. Given a  $d \times d$  matrix  $\mathbf{A}$ , the determinant of  $\mathbf{A}$  is

$$\det(\mathbf{A}) = \sum_{\sigma \in \text{symm}(m)} \text{sgn}(\sigma) \prod_{i=1}^d A(i, \sigma(i)),$$

where  $\text{symm}(m)$  is the set of all permutations of  $[d]$  and  $\text{sgn}(\sigma)$  is the sign function of a permutation.

Given  $d \times d$  matrices  $\mathbf{A}$  and  $\mathbf{B}$ , the Frobenius inner product between them is  $\langle \mathbf{A}, \mathbf{B} \rangle_F := \sum_{i,j \in [d]} \mathbf{A}(i, j) \mathbf{B}(i, j)$ .

We introduce two approximation results for determinants. The first one shows that  $\det(\mathbf{A})$  can be approximated by the determinant of its diagonal matrix, and the second shows that the determinant is smooth under small perturbation.

**Theorem A.1** ((Ipsen & Lee, 2011)). *Let  $\mathbf{A}$  be a  $d$ -dimensional squared matrix,  $\mathbf{A}_D$  be the associated diagonal matrix, and  $\mathbf{A}_E = \mathbf{A} - \mathbf{A}_D$ . If  $\mathbf{A}_D$  is non-singular and spectral norm  $\rho := \|\mathbf{A}_D^{-1} \mathbf{A}_E\|_2 < 1$  then*

$$\frac{|\det(\mathbf{A}) - \det(\mathbf{A}_D)|}{|\det(\mathbf{A}_D)|} \leq c \rho e^{c\rho}, \text{ where } c = -d \ln(1 - \rho)$$

Moreover, if  $c\rho < 1$ ,  $\frac{|\det(\mathbf{A}) - \det(\mathbf{A}_D)|}{|\det(\mathbf{A}_D)|} \leq \frac{7}{4} c\rho$ .

**Theorem A.2** ((Ipsen & Rehman, 2008)). *Let  $\mathbf{A}$  and  $\mathbf{E}$  be  $d \times d$  matrices. If  $\mathbf{A}$  is nonsingular, then*

$$\frac{|\det(\mathbf{A} + \mathbf{E}) - \det(\mathbf{A})|}{|\det(\mathbf{A})|} \leq \left(1 + \kappa \frac{\|\mathbf{E}\|_2}{\|\mathbf{A}\|_2}\right)^d - 1$$

where  $\kappa = \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2$  and  $\|\cdot\|_2$  is the spectral norm.

**Lemma A.3.** *Given  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times d}$ , if  $\mathbf{B} \neq \mathbb{I}$  is column stochastic and  $\mathbf{A}, \mathbf{B}\mathbf{A}$  are column diagonally maximal,  $\mathbf{B}$  is not a permutation matrix.*

*Proof of lemma A.3.* Suppose not and there exists a permutation  $\sigma : [d] \rightarrow [d]$  and  $\iota \in [d]$  so that  $B(i, j) = 1[j = \sigma(i)]$  and  $\sigma(\iota) \neq \iota$ . Because  $\mathbf{A}$  is column diagonally maximal

$$(\mathbf{B}\mathbf{A})(\iota, \iota) = \sum_j B(\iota, j) \mathbf{A}(j, \iota) = \mathbf{A}(\sigma(\iota), \iota) < \mathbf{A}(\iota, \iota).$$

Additionally,

$$(\mathbf{B}\mathbf{A})(\sigma^{-1}(\iota), \iota) = \mathbf{B}(\sigma^{-1}(\iota), \iota) \mathbf{A}(\iota, \iota) = \mathbf{A}(\iota, \iota) > (\mathbf{B}\mathbf{A})(\iota, \iota).$$

Therefore,  $\mathbf{B}\mathbf{A}$  column diagonally maximal which is a contradiction.  $\square$

Now we introduce kernel.

**Definition A.4.** *A function  $K : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$  is positive definite kernel if for all  $\{y_1, \dots, y_m\} \subseteq \mathcal{Y}$ , the matrix  $[K(y_i, y_j)]_{i,j} \in \mathbb{R}^{m \times m}$  is symmetric positive semi definite. Additionally, it is strictly positive definite if the matrix is positive definite.*

By Moore-Aronszajn theorem (Aronszajn, 1950), given a positive definite kernel  $K$ , there exists a Hilbert space  $\mathcal{H}$  known as reproducing kernel Hilbert space so that for any  $y \in \mathcal{Y}$ ,  $K(\cdot, y) \in \mathcal{H}$  and for all  $h \in \mathcal{H}$ ,  $h(y) = \langle h, K(y, \cdot) \rangle$ . This allows us to think of a kernel defines a feature map  $\phi : y \mapsto K(\cdot, y) \in \mathcal{H}$  where the inner product in the embedded space reduces to kernel evaluation, because  $\langle K(\cdot, y), K(\cdot, y') \rangle = K(y, y')$

Moreover, given a measurable kernel, we can define the *kernel mean embedding* (Berlinet & Thomas-Agnan, 2011) of probability measures on  $\mathcal{Y}$ ,  $P \in \Delta(\mathcal{Y})$ , into  $\mathcal{H}$  where

$$\phi(P) := \int K(\cdot, y) dP(y) = \mathbb{E}_{y \sim P}[\phi(y)].$$

Here we slightly abuse the notations, and note that  $\phi$  is linear in  $P$  by linearity of integration. We can further extend this to signed measures  $\phi(\mu) := \int K(\cdot, y) d\mu(y)$ . Finally, a kernel  $K$  is *integrally strictly positive definite* if the  $\iint_{\mathcal{Y}} K(y, y') d\mu(y) d\mu(y') > 0$  for all finite non-zero signed measures  $\mu$ .

## B PROOFS AND DETAILS IN SECTION 2

We shows that the reliability orderings are well-defined ordering. Formally, a binary relationship  $\succ$  on  $\Omega$  is a *strict partial order* if it satisfies the following conditions for all  $a, b, c \in \Omega$

1. anti-reflexive: no element is larger than itself
2. asymmetry: if  $a \succ b$  then not  $b \succ a$
3. Transitivity: if  $a \succ b$  and  $b \succ c$ , then  $a \succ c$ .

Next, we show that the reliability orderings defined in section 2 form a strict partial order over reports, given a fixed true data.

**Proposition B.1.** *For any  $x \in \mathcal{X}^N$ , the exact match ordering  $\succ_{\text{EXACT}}^x$  is a strict partial order on all  $\hat{x}$  and  $\hat{x}' \in \mathcal{X}^N$ .*

*Proof.* The first two are trivial. For transitivity, if  $\hat{x}_1 \succ_{\text{EXACT}}^x \hat{x}_2$ , then  $\hat{x}_2 \neq x$  so there is no  $\hat{x}_3$  with  $\hat{x}_2 \succ_{\text{EXACT}}^x \hat{x}_3$ .  $\square$

The following shows that Blackwell dominant ordering is a strict partial order over subsets of reports under the invertible and diagonally maximal conditions. Those conditions are essential. If the misreport matrices are not invertible, the Blackwell ordering may fail to be asymmetric: it is possible for two distinct reports to Blackwell-dominate each other, violating the strictness of the relation. Similarly, if the misreport matrices are not diagonally maximal, the ordering also fails asymmetry via non-trivial permutation.

**Proposition B.2.** *For any  $x \in \mathcal{X}^N$ , Blackwell dominant ordering  $\succ_{\text{Blackwell}}^x$  is a strict partial order on all  $\hat{x}$  and  $\hat{x}' \in \mathcal{X}^N$  so that the associated misreport matrices  $Q, Q' \in \mathcal{Q}_{\text{reg}}$  are invertible and diagonally maximal.*

*Proof.* Suppose  $\succ_{\text{Blackwell}}^x$  is not anti-reflective. There exists  $\hat{x} \succ_{\text{Blackwell}}^x \hat{x}$  with misreport matrix  $Q$  and a column stochastic matrix  $T \neq \mathbb{I}$  so that

$$TQ_{\hat{x}|x} = Q_{\hat{x}|x}.$$

Because  $Q = (Q_{\hat{x}|x} Q_x)^\top$  is invertible,  $Q_{\hat{x}|x}$  is also invertible and  $T = \mathbb{I}$  which is a contradiction.

For asymmetry, if  $\hat{x} \succ_{\text{Blackwell}}^x \hat{x}'$  and  $\hat{x}' \succ_{\text{Blackwell}}^x \hat{x}$ , there exist column stochastic matrices  $T$  and  $T'$  so that

$$TQ_{\hat{x}|x} = Q'_{\hat{x}|x} \text{ and } T'Q'_{\hat{x}|x} = Q_{\hat{x}|x}.$$

Because  $Q, Q'$  are invertible,  $TT' = \mathbb{I}$ , and both  $T$  and  $T'$  are permutation matrices. ([https://math.stackexchange.com/users/436618/angina seng](https://math.stackexchange.com/users/436618/angina%20seng)) However, because  $Q$  and  $Q'$  are (row) diagonally maximal,  $Q_{\hat{x}|x}$  and  $Q'_{\hat{x}|x}$  are column diagonally maximal. Therefore by lemma A.3,  $T = T' = \mathbb{I}$  which is a contradiction.

Transitivity is trivial, because the product of column stochastic matrices is still stochastic.  $\square$

**Proposition B.3.** *For any  $x \in \mathcal{X}^N$ , dist ordering  $\succ_{\text{dist}}^x$  is a strict partial order on all  $\hat{x}$  and  $\hat{x}' \in \mathcal{X}^N$ .*

*Proof.* The first two are trivial. For transitivity, given  $x, x'$  let  $\text{dist}(x, x') := \sum_n \text{dist}(x_n, x'_n)$ . If  $\hat{x}_1 \succ_{\text{dist}}^x \hat{x}_2$  and  $\hat{x}_2 \succ_{\text{dist}}^x \hat{x}_3$  then  $\text{dist}(x, \hat{x}_1) < \text{dist}(x, \hat{x}_2)$  and  $\text{dist}(x, \hat{x}_2) < \text{dist}(x, \hat{x}_3)$ . Therefore,  $\hat{x}_1 \succ_{\text{dist}}^x \hat{x}_3$ .  $\square$

### B.1 PROOF OF PROPOSITION 2.1

*Proof of proposition 2.1.* Given  $x, \hat{x}$ , and  $\hat{x}'$ , if  $\hat{x} \succ_{\text{EXACT}}^x \hat{x}'$ ,  $Q_{\hat{x}|x} = \mathbb{I}$  and  $Q'_{\hat{x}|x} \neq \mathbb{I}$ . If we set a column stochastic  $T = Q'_{\hat{x}|x}$ ,  $Q'_{\hat{x}|x} = TQ_{\hat{x}|x}$ . Therefore,  $\hat{x} \succ_{\text{Blackwell}}^x \hat{x}'$ .

If  $\hat{x} \succ_{\text{Blackwell}}^x \hat{x}'$ , there is  $T \neq \mathbb{I}$  so that  $Q'_{\hat{x}|x} = TQ_{\hat{x}|x}$ . With eq. (1) we have  $Q' = (Q'_{\hat{x}|x} Q_x)^\top = (TQ_{\hat{x}|x} Q_x)^\top = QT^\top$ , and

$$\begin{aligned} \text{Tr}(Q') &= \text{Tr}(QT^\top) = \sum_{i,j} Q(i,j)T(i,j) \\ &= \sum_i Q(i,i)T(i,i) + \sum_{i,j:i \neq j} Q(i,j)T(i,j) \\ &\leq \sum_i Q(i,i)T(i,i) + \sum_{i,j:i \neq j} Q(i,i)T(i,j) \quad (Q \text{ is diagonally maximal and } T \neq \mathbb{I}) \\ &= \sum_i Q(i,i) = \text{Tr}(Q) \quad (T \text{ is column stochastic}) \end{aligned}$$

Therefore,  $\hat{x} \succ_{\text{Hamming}}^x \hat{x}'$ . The third is straightforward.  $\square$

### C PROOFS AND DETAILS IN SECTION 3

We discuss the connection between detail-free setting and partial knowledge setting. First note that as the order of data is not relevant, given  $\hat{x}, y$  of size  $N$ , it is sufficient to consider the histogram of  $R \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|}$  and  $N$  where

$$R(y, x) = \frac{1}{N} \sum_n \mathbf{1}[y_n = y, \hat{x}_n = x].$$

By symmetrization, we can write a reliability score in detail-free setting as a stochastic function on the histogram  $R$  and  $N$  that have the same expected score. (Zheng et al., 2021) The expectation of  $R$  over the randomness of experiment is  $\mathbb{E}[R] = PQ$ . This leads to two key implications. First when the data size  $N$  is large,  $R$  converges to  $PQ$  so that the expectation of any smooth reliability score

$$\mathbb{E}[S(R)] \rightarrow S(\mathbb{E}[R]) = S(PQ).$$

Second, if we consider any *empirical risk-based scores* so that has  $\ell : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  so that

$$S(\hat{x}, y) = \frac{1}{N} \sum_n \ell(\hat{x}_n, y_n).$$

This includes common metrics like empirical risk and log-likelihood function. We can rewrite it as a linear function of  $R$

$$\begin{aligned} S(\hat{x}, y) &= \frac{1}{N} \sum_n \ell(\hat{x}_n, y_n) \\ &= \frac{1}{N} \sum_{x,y} \sum_n \mathbf{1}[\hat{x}_n = x, y_n = y] \ell(x, y) \\ &= \sum_{x,y} R(y, x) \ell(x, y) \end{aligned}$$

which is simply the Frobenius inner product between  $R$  and the score matrix based on  $\ell$ .

Finally, as definition 4.1, our Gram determinant score is also a function of  $PQ$ . Consequently, the impossibility results presented in section 3 for the partial knowledge setting apply not only to the Gram determinant score but also to any empirical risk-based score.

We provide the proof of proposition 3.1 consists of three parts: exact, Blackwell, and Hamming orderings.

**Proof of Exact orderings in proposition 3.1** For the exact ordering setting, we motivate the independence condition on experiments and non-permutation condition on misreport matrix. First we show that we need additional condition on experiments  $\mathcal{P}$ , even restricting to  $\mathcal{Q}_{\text{nonperm}}$ . Second, we show that  $\mathcal{Q}_{\text{nonperm}}$  is the maximal set of misreport matrices to have a reliability score that respects the exact match ordering on  $\mathcal{P}_{\text{indep}}$ .

Both parts use the idea that if two labels in  $\mathcal{X}$  induce the same distribution over observations, it becomes impossible to determine whether the reports agree with the true data.

For the first part, if  $P$  consists of identical columns, we can find a diagonal matrix  $Q_x$  and a doubly stochastic  $Q_{\hat{x}|x} \neq \mathbb{I}$  so that  $P(Q_{\hat{x}|x}Q_x)^\top = PQ_xQ_{\hat{x}|x}^\top = PQ_x$ . Hence, we can set  $x, \hat{x}$  with misreport matrices  $Q = (Q_{\hat{x}|x}Q_x)^\top$  so that  $x \succ_{\text{EXACT}}^x \hat{x}$ , but have the same joint distribution between reports and observations. Therefore, no score in the partial knowledge setting can distinguish them and preserves exact match ordering.

For the second part, because we can only observe the observations and reports, it would be impossible to always score true data over relabeled reports (permutation). Suppose not and there exists a score  $S$  in partial knowledge setting that preserves all misreport matrices. Given any  $P \in \mathcal{P}_{\text{indep}}$ , the uniform marginal distribution  $Q_x := \frac{1}{d}\mathbb{I}$ , and permutation  $T \neq \mathbb{I}$ , there exist  $x$  and  $\hat{x}$  so that the misreport matrix equals  $TQ_x = \frac{1}{d}T$  and  $x \succ_{\text{EXACT}}^x \hat{x}$ . Because the joint distribution between reports and observations is  $\frac{1}{d}P$  for  $(x, y)$ , and  $\frac{1}{d}PT^\top$  for  $(\hat{x}, y)$ , we have

$$S\left(\frac{1}{d}P\right) > S\left(\frac{1}{d}PT^\top\right).$$

Conversely, we can set an new experiment  $P' = PT^\top$  and  $x' = \hat{x}$  and  $\hat{x}' = x$  so that the misreport matrix equals  $\frac{1}{d}T^\top$  and  $x' \succ_{\text{EXACT}}^{x'} \hat{x}'$ . First, because  $T$  is a permutation  $P' = PT^\top \in \mathcal{P}_{\text{indep}}$  and the joint distributions becomes  $\frac{1}{d}P' = \frac{1}{d}PT^\top$  for  $(x', y')$  and  $\frac{1}{d}P'T = \frac{1}{d}PT^\top T = \frac{1}{d}P$  for  $(\hat{x}', y')$ . Therefore,

$$S\left(\frac{1}{d}PT^\top\right) > S\left(\frac{1}{d}P\right)$$

which is a contradiction.

**Proof of Blackwell orderings in proposition 3.1** For Blackwell ordering, we further show that the existence of *any* linearly dependent experiment  $P$  (i.e. columns of  $P$  are linearly dependent) in  $\mathcal{P}$  makes it impossible to preserve Blackwell ordering on  $\mathcal{P}$  and  $\mathcal{Q}_{\text{reg}}$ . The  $\mathcal{Q}_{\text{reg}}$  restriction rules out the possibility of improving data reliability through simple post-processing operations like (noisy) relabeling. In addition, recall that Blackwell ordering requires  $\mathcal{Q}_{\text{reg}}$  to be a strict partial ordering.

The proof idea is similar to that of the exact ordering setting: it is impossible to detect misreporting when two labels induce identical observation distributions—i.e., when  $P$  has identical columns. The main challenge in proposition 3.1, however, is to show that for any linearly dependent  $P$  (which may not have identical columns), we can construct a misreport matrix  $Q$  such that  $PQ$  has identical columns.<sup>7</sup>

If we can find  $P \in \mathcal{P}$ , a misreport matrix  $Q$ , and column stochastic  $T \neq \mathbb{I}$  with  $PQ = PQT^\top$ , we have  $x, \hat{x}, \hat{x}'$  with misreport matrices  $Q$  and  $QT^\top$  so that  $\hat{x} \succ_{\text{Blackwell}}^x \hat{x}'$ , but have the same joint distribution between reports and observations. Therefore, no score in the partial knowledge (and detail-free) setting can distinguish them and preserves the Blackwell ordering.

Now we construct  $P, Q$ , and  $T$ . By the condition in proposition 3.1 there exists  $P \in \mathcal{P}$  and  $v \neq 0 \in \mathbb{Q}^d$  so that  $Pv = 0$ . We decompose  $v$  as  $v = v_+ - v_-$  where  $v_+$  and  $v_-$  are nonnegative and have disjoint support, so

$$Pv_+ = Pv_- \quad (7)$$

and  $v_+, v_- \neq 0$  because  $P$  is a collection of distributions. Let  $\iota_+ \in [d]$  be the index of the largest entry in  $v_+$ , and  $\iota_-$  for  $v_-$  similarly, breaking ties arbitrarily. Note that  $\iota_+ \neq \iota_-$ , because  $v_+$  and  $v_-$  have disjoint supports. We first construct  $A$  by replacing the  $\iota_+$  column of the identity

<sup>7</sup>We require  $v \in \mathbb{Q}^d$  to have rational coefficients to ensure the resulting  $Q$  has rational coefficients to be a valid misreport matrix.

matrix  $\mathbb{I} \in \mathbb{R}^d$  by  $v_+$  and  $v_-$  column by  $v_-$ , and set  $Q = \frac{1}{Z}A$  where  $Z = \sum_{i,j} A(i,j)$ . This normalization ensures that  $Q$  forms a distribution as  $v_+$  and  $v_-$  are non-negative. By construction,  $Q$  is diagonally maximized by the choice of  $v_+, v_-$ , and invertible because  $v_+, v_- \neq 0$  and using Gaussian elimination. Most importantly, the  $v_+$  and  $v_-$  columns of  $PQ$  are identical by eq. (7).

To complete the construction, given  $\epsilon > 0$  we set  $T \neq \mathbb{I}$  so that

$$T(i,j) = \begin{cases} 1 & \text{if } i = j \text{ and } \{i,j\} \cap \{v_+, v_-\} = \emptyset \\ 0 & \text{if } i \neq j \text{ and } \{i,j\} \cap \{v_+, v_-\} = \emptyset \\ \epsilon & \text{if } i = v_+, j = v_- \text{ or } i = v_-, j = v_+ \\ 1 - \epsilon & \text{if } i = j \in \{v_+, v_-\} \\ 0 & \text{if } i \neq j \text{ and } |\{i,j\} \cap \{v_+, v_-\}| = 1 \end{cases}$$

which is the identical matrix excepts for the  $v_+$  and  $v_-$  columns and rows. Note that  $T$  is a column stochastic matrix,  $QT^T$  is still invertible and diagonally maximal when  $\epsilon$  is small enough. Finally,  $PQT^T$  mixes the  $v_+$  and  $v_-$  columns. However, because the  $v_+$  and  $v_-$  columns of  $PQ$  are identical,  $PQ = PQT^T$  which completes our proof.

**Proof of Hamming and dist orderings in proposition 3.1** Finally, we show that there does not exist a reliability score that preserves the Hamming and dist distance ordering, even restricting to diagonally dominant misreport matrices  $Q_{\text{dom}} \subset Q_{\text{reg}}$ .

We begin the proof with the Hamming ordering. Suppose we can find two settings: one has  $Q_1, Q'_1 \in Q_{\text{dom}}$  and  $P_1 \in \mathcal{P}_{\text{indep}}$ , the other has  $Q_2, Q'_2 \in Q_{\text{dom}}$  and  $P_2 \in \mathcal{P}_{\text{indep}}$  so that

$$\text{Tr}(Q_1) > \text{Tr}(Q'_1), \text{Tr}(Q_2) < \text{Tr}(Q'_2), \text{ but } P_1 Q_1 = P_2 Q_2, P_1 Q'_1 = P_2 Q'_2.$$

Then we can find  $x_1, \hat{x}_1, \hat{x}'_1, x_2, \hat{x}_2, \hat{x}'_2$  so that  $\hat{x}_1 \succ_{\text{Hamming}}^{x_1} \hat{x}'_1$  and  $\hat{x}'_2 \succ_{\text{Hamming}}^{x_2} \hat{x}_2$  by setting the misreport matrix of  $x_1, \hat{x}_1$  be  $Q_1$ , the misreport matrix  $x_1, \hat{x}'_1$  as  $Q'_1$ , the misreport matrix of  $x_2, \hat{x}_2$  be  $Q_2$ , the misreport matrix  $x_2, \hat{x}'_2$  as  $Q'_2$ . If there is a reliability score that preserves the Hamming ordering on  $\mathcal{P}_{\text{indep}}, Q_{\text{dom}}$ ,

$$\mathbb{E}[S(P_1 Q_1)] > \mathbb{E}[S(P_1 Q'_1)] \text{ and } \mathbb{E}[S(P_2 Q_2)] < \mathbb{E}[S(P_2 Q'_2)] \quad (8)$$

which reaches a contradiction as  $P_1 Q_1 = P_2 Q_2$  and  $P_1 Q'_1 = P_2 Q'_2$

We construct

$$P_1 = \begin{pmatrix} 0.74 & 0 & 0.26 \\ 0.26 & 0.74 & 0 \\ 0 & 0.26 & 0.74 \end{pmatrix}, Q_1 = \frac{1}{3} \begin{pmatrix} 0.8 & 0 & 0.2 \\ 0.2 & 0.8 & 0 \\ 0 & 0.2 & 0.8 \end{pmatrix}, Q'_1 = \frac{1}{3} \begin{pmatrix} 0.7 & 0.3 & 0 \\ 0 & 0.7 & 0.3 \\ 0.3 & 0 & 0.7 \end{pmatrix}.$$

For the second setting, we define  $P_2 = \mathbb{I}$ , and

$$Q_2 = P_1 Q_1 = \frac{1}{3} \begin{pmatrix} 0.592 & 0.052 & 0.356 \\ 0.356 & 0.592 & 0.052 \\ 0.052 & 0.356 & 0.592 \end{pmatrix}$$

$$Q'_2 = P_1 Q'_1 = \frac{1}{3} \begin{pmatrix} 0.596 & 0.222 & 0.182 \\ 0.182 & 0.596 & 0.222 \\ 0.222 & 0.182 & 0.596 \end{pmatrix}$$

Therefore,  $P_1 Q_1 = P_2 Q_2$  and  $P_1 Q'_1 = P_2 Q'_2$ . By direct computation, we have  $\text{Tr}(Q_1) = \frac{24}{30} > \text{Tr}(Q'_1) = \frac{21}{30}$  and  $\text{Tr}(Q_2) = \frac{1776}{3000} < \text{Tr}(Q'_2) = \frac{1788}{3000}$ . Finally, note that we can easily generalize this construction beyond three dimensions by padding the other dimension with identity.

Interestingly, the same construction works for general dist-ordering, due to the symmetry in  $Q_1, Q'_1, Q_2$  and  $Q'_2$ . First note that  $\sum_{n=1}^N \text{dist}(\hat{x}_n, x_n) = N \sum_{i,j \in [d]} Q(i,j) \text{dist}(i,j) = N \langle Q, \text{dist} \rangle_F$  where  $\langle \cdot, \cdot \rangle_F$  is the Frobenius inner product defined in appendix A. Hence, with eq. (8), it is sufficient to show the above construction satisfies

$$\langle Q_1, \text{dist} \rangle_F > \langle Q'_1, \text{dist} \rangle_F \text{ and } \langle Q_2, \text{dist} \rangle_F < \langle Q'_2, \text{dist} \rangle_F.$$

Let  $A = \text{dist}(1, 2) + \text{dist}(2, 3) + \text{dist}(3, 1) = \text{dist}(1, 3) + \text{dist}(2, 1) + \text{dist}(3, 2) > 0$  as  $\text{dist}(x, x') = \text{dist}(x', x)$  for all  $x, x'$ . By symmetry, we note the Frobenius inner product only depends on  $A$ ,

$$\begin{aligned}\langle \mathbf{Q}_1, \text{dist} \rangle_F - \langle \mathbf{Q}'_1, \text{dist} \rangle_F &= \frac{1}{3}(0.2A - 0.3A) < 0 \quad (\text{dist}(x, x) = 0 \text{ for all } x) \\ \langle \mathbf{Q}_2, \text{dist} \rangle_F - \langle \mathbf{Q}'_2, \text{dist} \rangle_F &= \frac{1}{3}(0.408A - 0.404A) > 0\end{aligned}$$

which completes the proof.

## D PROOFS AND DETAILS IN SECTION 4.1

*Proof of eq. (4).* For all  $\hat{x}, \hat{x}' \in \mathcal{X}$ ,

$$\begin{aligned}\hat{\mathbf{G}}(\hat{x}, \hat{x}') &= \frac{1}{N^2} \sum_{n, n': \hat{x}_n = x, \hat{x}_{n'} = x'} \langle P_{x_n}, P_{x_{n'}} \rangle \\ &= \frac{1}{N^2} \sum_{x, x' \in \mathcal{X}} \sum_{\substack{n, n': \\ \hat{x}_n = x, \hat{x}_{n'} = x', \\ x_n = x, x_{n'} = x'}} \mathbf{G}(x, x') \\ &= \sum_{x, x' \in \mathcal{X}} \mathbf{Q}(x, \hat{x}) \mathbf{G}(x, x') \mathbf{Q}(x', \hat{x}')\end{aligned}$$

which proves eq. (4).  $\square$

### D.1 AN EXAMPLE OF GRAM DETERMINANT SCORE

Here we provide a simple example for Gram determinant score. Consider  $\mathcal{X} = \mathcal{Y} = \{1, 2\}$ ,  $\mathbf{P} = \begin{pmatrix} 1 - p_1 & 1 - p_2 \\ p_1 & p_2 \end{pmatrix}$ , and the misreport matrix  $\mathbf{Q} = \begin{pmatrix} \frac{1-\delta}{4} & \frac{\delta}{4} \\ \frac{\delta}{4} & \frac{1-\delta}{4} \end{pmatrix}$  with  $\delta \geq 0$  where  $x = \hat{x}$  if  $\delta = 0$  whereas increasing  $\delta$  makes the reports less reliable. By direct computation, the Gram determinant score is

$$\det(\hat{\mathbf{G}}) = \det(\mathbf{P})^2 \det(\mathbf{Q})^2 = \frac{1}{2^8} (p_1 - p_2)^2 (1 - 2\delta)^2. \quad (9)$$

Given a fixed experiment  $\mathbf{P}$ , the Gram determinant score eq. (9) decreases as  $\delta$  increases from  $\delta = 0$  to  $1/2$ . In particular, it maximizes at  $\delta = 0$ , when the reported data exactly match the true data, and drops to zero at  $\delta = 1/2$ , where all reports contain the same uniform mixture of the true labels. Additionally, the score also depends the quality of the experiment  $\mathbf{P}$ . If  $p_1 = p_2$ , columns of  $\mathbf{P}$  are linearly dependent and Gram determinant score become zero. In contrast, if  $p_1 \neq p_2$  and  $\delta < 1/2$ , the score is strictly positive.

### D.2 LEMMAS AND PROOFS FOR THEOREM 4.2

*Proof of theorem 4.2.* As discussed in the main body. The key idea of proving theorem 4.2 that the determinant has multiplicative property eq. (4) which allows us to decouple the misreport matrix  $\mathbf{Q}$  from the quality of the experiment  $\mathbf{P}$ , and  $\det(\mathbf{G}) = \det(\mathbf{P}^\top \mathbf{P}) > 0$ , for all  $\mathbf{P} \in \mathcal{P}_{\text{indep}}$ . Therefore,

$$\Gamma > \Gamma' \text{ if and only if } \det(\mathbf{Q}^\top \mathbf{Q}) > \det((\mathbf{Q}')^\top \mathbf{Q}').$$

Thus, the following lemmas D.1 and D.2 prove the first and second case. For the third case, lemma D.3 proves the score preserves the approximate Hamming ordering, as  $\Delta = 1$  for Hamming distance. For general distance, let  $\text{Hamming}(x, \hat{x}) = \sum_n \mathbf{1}[\hat{x}_n \neq x_n]$  and  $\text{dist}(x, \hat{x}) = \sum_n \text{dist}(\hat{x}_n, x_n)$  be the Hamming distance and dist between  $\hat{x}$  and  $x$ . Because

$$\min_{x \neq x'} \text{dist}(x, x') \text{Hamming}(\hat{x}, x) \leq \text{dist}(x, \hat{x}) \leq \max_{x, x'} \text{dist}(x, x') \text{Hamming}(\hat{x}, x),$$

and  $\Delta = \frac{\max_{x, x'} \text{dist}(x, x')}{\min_{x \neq x'} \text{dist}(x, x')}$ ,  $\hat{x} \succ_{\text{dist}, 1/(4\Delta L)}^x \hat{x}'$  implies  $\hat{x} \succ_{\text{Hamming}, 1/(4L)}^x \hat{x}'$ , which completes the proof.  $\square$

**Lemma D.1.** For all  $x, \hat{x}, \hat{x}'$  if  $\hat{x} \succ_{\text{EXACT}}^x \hat{x}'$  and  $Q, Q' \in \mathcal{Q}_{\text{nonperm}}$ ,  $\det(Q^\top Q) > \det((Q')^\top Q')$ .

*Proof of lemma D.1.* As  $x, \hat{x}$ , and  $\hat{x}'$  with  $\hat{x} \succ_{\text{EXACT}}^x \hat{x}'$ ,  $Q_{\hat{x}|x} = \mathbb{I}$  and there is  $T \neq \mathbb{I}$  so that  $Q'_{\hat{x}|x} = TQ_{\hat{x}|x} = T$ . By eq. (1) we have  $Q' = QT^\top = Q_x T^\top$  and  $Q = Q_x$ . Therefore

$$\det((Q')^\top Q') = \det(TQ^\top QT^\top) = \det(TT^\top) \det(Q^\top Q) \quad (10)$$

Because the diagonal matrix  $Q = Q_x$  has positive diagonals, and  $T$  is column stochastic and not a permutation matrix, the Perron–Frobenius theorem (or (Kong, 2020)) implies  $|\det(T)| < 1$  and  $\det((Q')^\top Q') = \det(TT^\top) \det(Q^\top Q) < \det(Q^\top Q)$ .  $\square$

**Lemma D.2.** For all  $x, \hat{x}, \hat{x}'$  if  $\hat{x} \succ_{\text{Blackwell}}^x \hat{x}'$  and  $Q, Q' \in \mathcal{Q}_{\text{reg}}$ ,  $\det(Q^\top Q) > \det((Q')^\top Q')$ .

*Proof of lemma D.2.* As  $\hat{x} \succ_{\text{Blackwell}}^x \hat{x}'$ , there is a column stochastic  $T \neq \mathbb{I}$  so that  $Q'_{\hat{x}|x} = TQ_{\hat{x}|x}$ . By eq. (10),

$$\det((Q')^\top Q') = \det(TQ^\top QT^\top) = \det(TT^\top) \det(Q^\top Q)$$

Because  $Q \in \mathcal{Q}_{\text{reg}}$  is invertible,  $\det(Q) \neq 0$ . By lemma A.3,  $T$  is not a permutation matrix, so  $|\det(T)| < 1$ , and  $\det((Q')^\top Q') = \det(TT^\top) \det(Q^\top Q) < \det(Q^\top Q)$ .  $\square$

**Lemma D.3.** Given  $\mathcal{X} = [d]$  and  $L \geq 1$ , for all  $x, \hat{x}, \hat{x}'$  if  $\hat{x} \succ_{\text{Hamming}, \frac{1}{4L}}^x \hat{x}'$  and  $Q, Q' \in \mathcal{Q}_{L, 1/(64L^2 d^2)}$ ,  $\det(Q^\top Q) > \det((Q')^\top Q')$ .

**Proof of lemma D.3** Lemma D.3 establishes that the Gram–determinant score approximately preserves the Hamming ordering under balancedness and small Hamming distance conditions. The main technical challenge lies in deriving upper and lower bounds on the Gram determinant in terms of the Hamming distance lemma D.4.

**Lemma D.4.** For all  $\delta \geq 0$ , and  $x, \hat{x}$  with diagonally dominant  $Q$ , if  $\delta = 1 - \text{Tr}(Q)$  and  $\delta < \frac{\min_i q_x(i)}{4}$ ,

$$\left(1 - \frac{8d\delta^2}{\min_i q_x(i)^2}\right) \left(1 - \frac{\delta}{\min_i q_x(i)}\right) \leq \frac{\det(Q)}{\prod_i q(i)} \leq \left(1 + \frac{8d\delta^2}{\min_i q_x(i)^2}\right) \left(1 - \frac{\delta}{2 \max_i q_x(i)}\right).$$

**Lemma D.5.** Given  $L \geq 1$ , if  $a_1, \dots, a_d \geq 0$ ,  $\sum_{i \in [d]} a_i = 1$  and  $a_i \leq La_j$  for all  $i, j \in [d]$ , then

$$\frac{1}{Ld - L + 1} \leq a_i \leq \frac{L}{d + L - 1}, \text{ for all } i$$

*Proof of lemma D.3.* If  $x, \hat{x}, \hat{x}'$  with  $Q, Q' \in \mathcal{Q}_{L, 1/(64L^2 d^2)}$ , the true labels are  $L$  balanced, and Hamming distances  $\delta = 1 - \text{Tr}(Q)$ ,  $\delta' = 1 - \text{Tr}(Q')$  are less than  $\frac{1}{64L^2 d^2}$ . If  $\hat{x} \succ_{\text{Hamming}, 1/(4L)}^x \hat{x}'$ , we want to show  $\left(\frac{\det(Q)}{\det(Q')}\right)^2 > 1$ .

Note that as  $Q, Q'$  are diagonally dominant  $\det(Q), \det(Q') > 0$ , and we use lemma D.4 to show that the lower bound of  $\det(Q)$  is larger than the upper bound of  $\det(Q')$ ,

$$\left(1 + \frac{8d(\delta')^2}{\min_i q_x(i)^2}\right) \left(1 - \frac{\delta'}{2 \max_i q_x(i)}\right) < \left(1 - \frac{8d\delta^2}{\min_i q_x(i)^2}\right) \left(1 - \frac{\delta}{\min_i q_x(i)}\right).$$

By taking the difference, we have

$$\begin{aligned}
& \left(1 - \frac{8d\delta^2}{\min_i q_{\mathbf{x}}(i)^2}\right) \left(1 - \frac{\delta}{\min_i q_{\mathbf{x}}(i)}\right) - \left(1 + \frac{8d(\delta')^2}{\min_i q_{\mathbf{x}}(i)^2}\right) \left(1 - \frac{\delta'}{2 \max_i q_{\mathbf{x}}(i)}\right) \\
& > \frac{\delta'}{2 \max_i q_{\mathbf{x}}(i)} - \frac{8d\delta^2}{\min_i q_{\mathbf{x}}(i)^2} - \frac{\delta}{\min_i q_{\mathbf{x}}(i)} - \frac{8d(\delta')^2}{\min_i q_{\mathbf{x}}(i)^2} \\
& \quad \text{(The second order terms are positive)} \\
& \geq \frac{\delta'}{2 \max_i q_{\mathbf{x}}(i)} - \frac{\delta}{\min_i q_{\mathbf{x}}(i)} - \frac{16d(\delta')^2}{\min_i q_{\mathbf{x}}(i)^2} \quad (\delta < \delta') \\
& \geq \frac{\delta'}{4 \max_i q_{\mathbf{x}}(i)} - \frac{16d(\delta')^2}{\min_i q_{\mathbf{x}}(i)^2} \quad (\delta' > 4L\delta) \\
& = \frac{\delta'}{4 \max_i q_{\mathbf{x}}(i)} \left(1 - \frac{64d \max_i q_{\mathbf{x}}(i) \delta'}{\min_i q_{\mathbf{x}}(i)^2}\right) \\
& \geq \frac{\delta'}{4 \max_i q_{\mathbf{x}}(i)} \left(1 - \frac{64Ld}{\min_i q_{\mathbf{x}}(i)} \delta'\right) \quad (\max_i q_{\mathbf{x}}(i) < L \min_i q_{\mathbf{x}}(i)) \\
& > 0 \quad (\delta' < \frac{1}{64L^2d^2} \text{ and } \min_i q_{\mathbf{x}}(i) \geq \frac{1}{Ld} \text{ by lemma D.5})
\end{aligned}$$

□

*Proof of lemma D.4.* We want to estimate  $\det(\mathbf{Q})$  by the Hamming distance. Let  $\mathbf{Q} = \mathbf{D} + \mathbf{E}$  where  $\mathbf{D}$  is a diagonal matrix and  $\mathbf{E}$  has zero diagonal, and  $\delta_i = \sum_{j \neq i} \mathbf{E}(i, j) = q_{\mathbf{x}}(i) - \mathbf{D}(i, i) \geq 0$  for all  $i \in \mathcal{X}$  which is the off-diagonal weight of row  $i$ . With above notations,  $1 - \text{Tr}(\mathbf{Q}) = \sum_{i \in \mathcal{X}} \delta_i = \delta$  and  $\det(\mathbf{D}) = \prod (q_{\mathbf{x}}(i) - \delta_i)$ . If  $\rho = \|\mathbf{D}^{-1} \mathbf{E}\|_2$  and  $\delta_Q = -\rho d \ln(1 - \rho)$ , by theorem A.1

$$1 - \delta_Q \leq \frac{\det(\mathbf{Q})}{\det(\mathbf{D})} \leq 1 + \delta_Q. \quad (11)$$

As  $\mathbf{D}^{-1} \mathbf{E}$  is a nonnegative matrix, by Gershgorin theorem, the spectral radius  $\rho$  can be bounded by the row sum  $\delta_i / \mathbf{Q}(i, i) \leq \frac{2\delta}{\min_i q_{\mathbf{x}}(i)}$  since  $\mathbf{Q}$  is diagonally dominant. Because  $-\ln(1 - t) \leq 2t$  for all  $t < 1/2$  and  $\delta \leq \frac{\min_i q_{\mathbf{x}}(i)}{4}$ , we have

$$\delta_Q \leq 2d\rho^2 \leq \frac{8d\delta^2}{\min_i q_{\mathbf{x}}(i)^2} \quad (12)$$

Now we bound the ratio  $\frac{\det(\mathbf{D})}{\prod_i q_{\mathbf{x}}(i)} = \prod_i \left(1 - \frac{\delta_i}{q_{\mathbf{x}}(i)}\right)$ . By union bound,

$$\prod_i \left(1 - \frac{\delta_i}{q_{\mathbf{x}}(i)}\right) \geq 1 - \sum_i \frac{\delta_i}{q_{\mathbf{x}}(i)} \geq 1 - \frac{\delta}{\min_i q_{\mathbf{x}}(i)} \quad (\delta = \sum \delta_i)$$

On the other hand,

$$\begin{aligned}
\prod_i \left(1 - \frac{\delta_i}{q_{\mathbf{x}}(i)}\right) & \leq \exp\left(-\sum \frac{\delta_i}{q_{\mathbf{x}}(i)}\right) \quad (1 - t \leq e^{-t} \text{ for all } t) \\
& \leq \exp\left(-\frac{\delta}{\max_i q_{\mathbf{x}}(i)}\right) \quad (\delta = \sum \delta_i) \\
& \leq 1 - \frac{\delta}{2 \max_i q_{\mathbf{x}}(i)} \quad (\delta < \max q_{\mathbf{x}}(i) \text{ and } e^{-t} \leq 1 - \frac{1}{2}t \text{ if } 0 \leq t \leq 1)
\end{aligned}$$

Therefore,

$$1 - \frac{\delta}{\min_i q_{\mathbf{x}}(i)} \leq \frac{\det(\mathbf{D})}{\prod_i q_{\mathbf{x}}(i)} \leq 1 - \frac{\delta}{2 \max_i q_{\mathbf{x}}(i)} \quad (13)$$

By eqs. (11) and (13), we have

$$(1 - \delta_Q) \left(1 - \frac{\delta}{\min_i q_{\mathbf{x}}(i)}\right) \leq \frac{\det(\mathbf{Q})}{\prod_i q_{\mathbf{x}}(i)} \leq (1 + \delta_Q) \left(1 - \frac{\delta}{2 \max_i q_{\mathbf{x}}(i)}\right)$$

which completes the proof by plugging in eq. (12) □



*Proof of lemma D.5.* Because  $a_j \geq \frac{1}{L}a_i$  for all  $i \neq j$ ,  $1 = \sum a_j \geq a_i + \frac{d-1}{L}a_i \leq \frac{L+d-1}{L}a_i$ , and

$$a_i \leq \frac{L}{L+d-1}.$$

On the other hand, because  $a_j \leq La_i$ ,  $1 = \sum a_j \leq a_i + (d-1)La_i$ , and

$$a_i \geq \frac{1}{Ld-L+1}$$

□

### D.3 PROOFS FOR EXPERIMENT AGNOSTIC

*Proof of proposition 4.3.* We first show that there is  $\alpha = 1/S(\mathbb{I}) > 0$  so that for all  $\mathbf{P}, \mathbf{Q} \in GL_d$

$$S(\mathbf{PQ}) = \alpha S(\mathbf{P})S(\mathbf{Q}). \quad (14)$$

Since  $S$  is experiment agonistic, given any  $\mathbf{P}$ ,  $S(\mathbf{PQ})$  is increasing in  $S(\mathbf{Q})$ , and there exists an increasing function  $g_{\mathbf{P}}$  so that  $S(\mathbf{PQ}) = g_{\mathbf{P}}(S(\mathbf{Q}))$ . Because for any  $s, t > 0$  and  $\mathbf{Q}$ ,  $S(st\mathbf{Q}) = c(st)S(\mathbf{Q}) = c(s)c(t)S(\mathbf{Q})$  and  $S(\mathbf{Q}) > 0$ , we have  $c(st) = c(s)c(t)$  for all  $s, t > 0$ . Therefore,

$$c(t) = t^\gamma \text{ for some } \gamma \in \mathbb{R}. \quad (15)$$

For any  $t > 0$  and  $\mathbf{P}, \mathbf{Q}$ , we have  $S(\mathbf{PtQ}) = c(t)S(\mathbf{PQ}) = c(t)g_{\mathbf{P}}(S(\mathbf{Q}))$ , and  $S(\mathbf{PtQ}) = g_{\mathbf{P}}(S(t\mathbf{Q})) = g_{\mathbf{P}}(c(t)S(\mathbf{Q}))$ . Hence

$$g_{\mathbf{P}}(c(t)S(\mathbf{Q})) = c(t)g_{\mathbf{P}}(S(\mathbf{Q})).$$

For any  $\mathbf{P}$  and  $\mathbf{Q}$ , we have

$$S(\mathbf{PQ}) = g_{\mathbf{P}}\left(S(\mathbf{Q}) \cdot \frac{1}{S(\mathbf{Q})}S(\mathbf{Q})\right) = S(\mathbf{Q})g_{\mathbf{P}}(1)$$

by eq. (15) and taking  $t = S(\mathbf{Q})^{-\gamma}$ . By taking  $\mathbf{Q} = \mathbb{I}$  we have  $g_{\mathbf{P}}(1) = \frac{S(\mathbf{PQ})}{S(\mathbf{Q})} = \frac{S(\mathbf{P})}{S(\mathbb{I})}$ , and prove eq. (14).

By eq. (14),  $\tilde{S}(\mathbf{Q}) := \alpha S(\mathbf{Q})$  is a continuous homomorphism between  $GL_d$  and  $(\mathbb{R}_{>0}, \cdot)$  so that for all  $\mathbf{P}, \mathbf{Q}$   $\tilde{S}(\mathbf{PQ}) = \tilde{S}(\mathbf{P})\tilde{S}(\mathbf{Q})$ . Thus, there exists a continuous  $f : \mathbb{R} \setminus \{0\} \rightarrow \mathbb{R}_{>0}$  so that  $\tilde{S}(\mathbf{Q}) = f(\det(\mathbf{Q}))$ . (user856) We now pin down the function  $f$ . First,  $S(t\mathbf{Q}) = \alpha f(t^d \det(\mathbf{Q}))$  and by eq. (15),  $S(t\mathbf{Q}) = \alpha c(t)f(\det(\mathbf{Q})) = \alpha t^\gamma f(\det(\mathbf{Q}))$  for all  $t > 0$  and  $\mathbf{Q}$ . Given  $\beta = \gamma/(2d)$ , for all  $t > 0$ ,  $f(t) = t^{2\beta}f(1)$  and  $f(-t) = t^{2\beta}f(-1)$ . Moreover, because  $f$  is a homomorphism  $f(-1)^2 = f((-1)^2) = f(1) = 1$  and  $f(-1) > 0$ , we have for all  $z \neq 0$ ,  $f(z) = |z|^{2\beta}$  and

$$S(\mathbf{Q}) = \alpha f(\det(\mathbf{Q})) = \alpha |\det(\mathbf{Q})|^{2\beta} = \alpha \det(\mathbf{Q}^\top \mathbf{Q})^\beta.$$

□

## E LEMMAS AND PROOFS FOR SECTION 4.2

While the above plugged-in estimate can asymptotically preserve all reliability ordering in theorem 4.2, it lacks provable guarantees for data of finite size. In practice, only a limited number of observations are available, and the data source can be strategic and aims to maximize its reliability score. Definition E.1 provides an estimator that preserves the exact match ordering using finite data which rewards truthful reporting than any other reports.

**Definition E.1.** Given  $\mathcal{X} = [d]$ , and  $\hat{\mathbf{x}}, \mathbf{y}$  of size  $N$ , a stratified matching estimator for the Gram determinant score is defined as the following

1. Return 0 if the minimum occurrence  $\min_{x \in \mathcal{X}} |\{n \in [N] : \hat{x}_n = x\}|$  is less than 2. Otherwise, we randomly select two disjoint index sets  $Col, Row \subseteq [N]$  of size  $d$  where each label  $i \in \mathcal{X}$  occurs in each set exactly once. Then re-index them as two sequences of pairs  $(\hat{x}_{i,Col}, y_{i,Col})_{i \in [d]}$  and  $(\hat{x}_{i,Row}, y_{i,Row})_{i \in [d]}$  so that  $\hat{x}_{i,Col} = \hat{x}_{i,Row} = i$  for all  $i \in \mathcal{X}$ . We call the first as column sequence and the second as row sequence.

2. Randomly sample one permutations  $\sigma \in \text{sym}(d)$ , and return

$$\text{score}(\hat{\mathbf{x}}, \mathbf{y}) := d! \text{sgn}(\sigma) \prod_{i,j \in [d], j=\sigma(i)} \mathbf{1}[y_{i, \text{Row}} = y_{j, \text{Col}}] \mathbf{q}_{\hat{\mathbf{x}}}(i) \mathbf{q}_{\hat{\mathbf{x}}}(j). \quad (16)$$

Equation (16) approximates the second form of the Gram determinant in eq. (3) by summing over all permutations. The first step is a stratified sampling to collect one report of each label in *Col* and *Row* respectively. The term  $\mathbf{1}[y_{i, \text{Row}} = y_{j, \text{Col}}]$  approximates the inner product between the observation distributions of reports  $i$  and  $j$ , and the extra  $\mathbf{q}_{\hat{\mathbf{x}}}(i) \mathbf{q}_{\hat{\mathbf{x}}}(j)$  offset the stratified sampling.

The stratified-matching estimator only require each label has at least two true data. If any label occurs fewer than two times, the estimator returns zero and yields a worse score than truthful data. The following result shows that under mild balance conditions, the stratified-matching estimator preserves exact match ordering over linearly independent experiments. The proof is deferred to the appendix.

**Proposition E.2.** *Given  $\mathcal{X} = [d]$  and  $L \geq 1$ , the stratified-matching estimator in definition E.1 preserves exact ordering on  $\mathcal{P}_{\text{indep}}$ ,  $\mathcal{Q}_L$ , and  $N = 2Ld$ .*

### Proofs for proposition 4.5

**Lemma E.3.** *Given  $\delta > 0$  and report size  $N$ ,*

$$\Pr \left[ \|\bar{\mathbf{G}} - \hat{\mathbf{G}}\|_2 \leq 4\sqrt{\frac{\log 2d/\delta}{N}} + 2\frac{\log 2d/\delta}{N} \right] \geq 1 - \delta.$$

*Proof of proposition 4.5.* By theorem A.2, we have

$$\frac{|\det(\bar{\mathbf{G}}) - \det(\hat{\mathbf{G}})|}{\det(\hat{\mathbf{G}})} \leq \left( 1 + \frac{\|\bar{\mathbf{G}} - \hat{\mathbf{G}}\|_2}{\|\hat{\mathbf{G}}^{-1}\|_2} \right)^d - 1.$$

Hence with lemma E.3 and  $\delta = 1/N$ , we have  $\frac{|\det(\bar{\mathbf{G}}) - \det(\hat{\mathbf{G}})|}{\det(\hat{\mathbf{G}})} = o(1)$ , with probability greater than  $1 - 1/N$ . Additionally, because the random variable  $\det(\bar{\mathbf{G}})$  is always bounded by 1, the expectation

$$\mathbb{E}[\det(\bar{\mathbf{G}})] = (1 + o(1)) \det(\hat{\mathbf{G}}). \quad (17)$$

For all  $\mathbf{P}$  and  $\mathbf{Q}, \mathbf{Q}'$ , if  $\det(\hat{\mathbf{G}}) = \det(\mathbf{Q}^\top \mathbf{G} \mathbf{Q}) > \det((\mathbf{Q}')^\top \mathbf{G} \mathbf{Q}') = \det(\hat{\mathbf{G}}') > 0$ , by eq. (17) there exists a large enough  $N_0$  so that any  $\mathbf{x}, \hat{\mathbf{x}}, \hat{\mathbf{x}}'$  with length at least  $N_0$  and  $\mathbf{Q}, \mathbf{Q}'$  so that  $\mathbb{E}[\det(\bar{\mathbf{G}})] > \mathbb{E}[\det(\bar{\mathbf{G}}')]$ . Therefore, the plugged-in estimator asymptotically preserves all reliability as theorem 4.2.  $\square$

*Proof of lemma E.3.* Let  $N_i = Nq_{\hat{\mathbf{x}}}(i)$  be the number of report  $i$  which is nonzero as  $\mathbf{Q} \in \mathcal{Q}_{\text{reg}}$ . Let  $|\mathcal{Y}| = k$ , and we can set  $\phi : \mathcal{Y} \rightarrow \mathbb{R}^k$  be the delta vector  $y \mapsto \mathbf{1}_y$ . We define  $\bar{\mathbf{v}}_i = \sum_{n: \hat{x}_n=i} \phi(y_n)$  and  $\mathbf{v}_i = \mathbb{E}[\bar{\mathbf{v}}_i]$  as the sum of (empirical) mean of report  $i \in \mathcal{X}$ , and error  $\mathbf{e}_i = \bar{\mathbf{v}}_i - \mathbf{v}_i$ . Hence for all  $i, j$ ,  $\bar{\mathbf{G}}(i, j) = \frac{1}{N^2} \sum_{n, n': \hat{x}_n=i, \hat{x}_{n'}=j} \langle \phi(y_n), \phi(y_{n'}) \rangle = \frac{1}{N^2} \bar{\mathbf{v}}_i^\top \bar{\mathbf{v}}_j$ ,  $\hat{\mathbf{G}}(i, j) = \frac{1}{N^2} \mathbf{v}_i^\top \mathbf{v}_j$ , and

$$\bar{\mathbf{G}}(i, j) - \hat{\mathbf{G}}(i, j) = \frac{1}{N^2} (\mathbf{v}_i^\top \mathbf{e}_j + \mathbf{e}_i^\top \mathbf{v}_j + \mathbf{e}_i^\top \mathbf{e}_j) \quad (18)$$

To bound the spectral norm of  $\bar{\mathbf{G}} - \hat{\mathbf{G}} \in \mathbb{R}^{d \times d}$ , for any  $\mathbf{a} \in \mathbb{R}^d$  with  $\|\mathbf{a}\|_2 = 1$ , we define  $\mathbf{v}(\mathbf{a}) = \sum_i a_i \mathbf{v}_i$ ,  $\mathbf{e}(\mathbf{a}) = \sum_i a_i \mathbf{e}_i \in \mathbb{R}^k$ , and  $R_{\mathbf{v}} = \sup_{\|\mathbf{a}\|=1} \|\mathbf{v}(\mathbf{a})\|$ ,  $R_{\mathbf{e}} = \sup_{\|\mathbf{a}\|=1} \|\mathbf{e}(\mathbf{a})\|$ . By eq. (18)

$$\mathbf{a}^\top (\bar{\mathbf{G}} - \hat{\mathbf{G}}) \mathbf{a} = \frac{1}{N^2} (2\mathbf{v}(\mathbf{a})^\top \mathbf{e}(\mathbf{a}) + \mathbf{e}(\mathbf{a})^\top \mathbf{e}(\mathbf{a})) \leq \frac{1}{N^2} (2R_{\mathbf{v}}R_{\mathbf{e}} + R_{\mathbf{e}}^2). \quad (19)$$

We first bound  $R_v$ . For all  $a$  with  $\|a\| = 1$ , let  $V = N^2 \hat{G} \in \mathbb{R}^{d \times d}$  where  $V(i, j) = v_i^\top v_j$  which is positive semi definite

$$\begin{aligned}
\|v(a)\|^2 &= \sum_{i,j} a_i a_j v_i^\top v_j \\
&= a^\top V a \\
&\leq \sum_i v_i^\top v_i && \text{(Rayleigh quotient is upper bounded by the trace)} \\
&= \sum_i N^2 \hat{G}(i, i) && \text{(definition of } v_i) \\
&\leq \sum_i N^2 q_{\hat{x}}(i)^2 && \text{(because } \langle P_x, P_{x'} \rangle \leq 1 \text{ for any } x, x') \\
&\leq N^2 \max_i q_{\hat{x}}(i)
\end{aligned}$$

Therefore,

$$R_v \leq N \sqrt{\max_i q_{\hat{x}}(i)} \leq N \quad (20)$$

We bound  $R_e$  using Chernoff bound. For each  $i \in \mathcal{X}$ ,  $e_i = \bar{v}_i - v_i = \sum_{n:\hat{x}_n=i} \phi(y_n) - \mathbb{E}\phi(y_n)$  is sum of  $Nq_x(i)$  independent vectors in  $\mathbb{R}^k$ , and the norm of each vector is bounded by 1. Therefore, by (Pinelis, 1994, Theorem 3.5), for all  $r_i \geq 0$

$$\Pr[\|e_i\| \geq r_i] \leq 2 \exp\left(-\frac{r_i^2}{2Nq_x(i)}\right)$$

Given any  $\delta > 0$  and  $a$  with  $\|a\| = 1$ , we set  $r_i = \sqrt{2Nq_x(i) \ln(2d/\delta)}$ , and we have

$$\|e(a)\|^2 \leq \sum_i \|e_i\|^2 \leq \sum_i 2Nq_x(i) \ln\left(\frac{2d}{\delta}\right) = 2N \ln \frac{2d}{\delta}$$

Therefore,

$$R_e \leq \sqrt{2N \ln \frac{2d}{\delta}} \quad (21)$$

with probability at least  $1 - \delta$ . Plugging in eqs. (20) and (21) to eq. (19), we have

$$\|\bar{G} - \hat{G}\|_2 \leq \frac{1}{N^2} \left( 2N \sqrt{2N \ln \frac{2d}{\delta}} + 2N \ln \frac{2d}{\delta} \right) \leq \frac{4\sqrt{\ln 2d/\delta}}{\sqrt{N}} + \frac{2 \ln 2d/\delta}{N}.$$

□

**Proof of proposition E.2** The core idea relies on the multi-linearity of the determinant, and we can approximately get samples of  $\hat{G} = Q^\top G Q$  in the detail-free setting. However, one caveat is that we may not have access to multiple independent samples from  $\hat{G}$  as  $x, \hat{x}$  are deterministic. To circumvent this issue, we first observe that if  $\hat{x} = x$ , the observations are independently and identically distributed for each label, allowing an unbiased estimator for  $\hat{G}$  and thus  $\det(\hat{G})$ . If  $\hat{x} \neq x$ , our sampling scheme ensures that the expectation is bounded above by the Gram determinant score. This guarantees that exact match orderings are preserved, as the truthful reports yield higher or scores in expectation compared to any nontruthful reports.

*Proof of proposition E.2.* By the definition of exact ordering, it is sufficient to show for any  $x, \hat{x}$  with  $x \succ_{\text{EXACT}}^x \hat{x}$  and  $P \in \mathcal{P}_{\text{indep}}$ ,

$$\mathbb{E}_{y \sim P(x)}[\text{score}(x, y)] > \mathbb{E}_{y \sim P(x)}[\text{score}(\hat{x}, y)].$$

When the minimum occurrence is at least two, the expectation of eq. (16) involves three sources of randomness: observation  $y$ , permutations  $\sigma$ , and the choice of *Col* and *Row*. The expectation of  $\text{score}(x, y)$  only depends on the first two as difference indexing does not change the distribution of score. However, for  $\text{score}(\hat{x}, y)$ , the third part will kick in.

Given the index sets  $Col, Row$ , we define  $\mathbf{Q}^{Col}, \mathbf{Q}^{Row} \in \mathbb{R}^{d \times d}$  so that

$$\mathbf{Q}^{Col}(i, j) = q_{\hat{x}}(j) \sum_{n \in Col} \mathbf{1}[x_n = i, \hat{x}_n = j] = q_{\hat{x}}(j) \mathbf{1}[x_{j, Col} = i] \quad (22)$$

and  $\mathbf{Q}^{Row}(i, j)$  similarly which are the misreport matrix when restricting reports in  $Col$  and  $Row$  respectively. As  $Col$  can be seen as stratified sampling where each report has exactly one element in  $Col$ ,  $\sum_{n \in Col} \mathbf{1}[x_n = i, \hat{x}_n = j] = \mathbf{Q}_{x|\hat{x}}(i, j)$ , and the expectation over the choice of index is

$$\mathbb{E}[\mathbf{Q}^{Col}] = \mathbb{E}[\mathbf{Q}^{Row}] = \mathbf{Q}. \quad (23)$$

With the above notation, we first compute the expectation of eq. (16) conditional on  $Col$  and  $Row$ .

$$\begin{aligned} & \mathbb{E}[\text{score}(\hat{x}, y) \mid Col, Row] \\ &= \mathbb{E} \left[ d! \text{sgn}(\sigma) \prod_{k, l \in [d], l = \sigma(k)} \mathbf{1}[y_{k, Row} = y_{l, Col}] q_{\hat{x}}(k) q_{\hat{x}}(l) \mid Col, Row \right] \\ &= \mathbb{E} \left[ \sum_{\sigma \in \text{sym}(d)} \text{sgn}(\sigma) \prod_{k, l \in [d], l = \sigma(k)} \mathbf{1}[y_{k, Row} = y_{l, Col}] q_{\hat{x}}(k) q_{\hat{x}}(l) \mid Col, Row \right] \quad (\text{random } \sigma) \\ &= \mathbb{E} \left[ \sum_{\sigma \in \text{sym}(d)} \text{sgn}(\sigma) \prod_{k, l \in [d], l = \sigma(k)} \langle P_{x_{k, Row}}, P_{x_{l, Col}} \rangle q_{\hat{x}}(k) q_{\hat{x}}(l) \mid Col, Row \right] \quad (Col \cap Row = \emptyset) \\ &= \mathbb{E} \left[ \sum_{\sigma \in \text{sym}(d)} \text{sgn}(\sigma) \prod_{k, l \in [d], l = \sigma(k)} \sum_{i, j} \mathbf{Q}^{Row}(i, k) \mathbf{Q}^{Col}(j, l) \langle P_i, P_j \rangle \mid Col, Row \right] \quad (\text{by eq. (22)}) \\ &= \mathbb{E} \left[ \sum_{\sigma \in \text{sym}(d)} \text{sgn}(\sigma) \prod_{k, l \in [d], l = \sigma(k)} ((\mathbf{Q}^{Row})^\top \mathbf{G} \mathbf{Q}^{Col})(k, l) \mid Col, Row \right] \\ &= \mathbb{E} [\det((\mathbf{Q}^{Row})^\top \mathbf{G} \mathbf{Q}^{Col}) \mid Col, Row] \end{aligned}$$

Therefore,

$$\mathbb{E}[\text{score}(\hat{x}, y)] = \mathbb{E} [\det((\mathbf{Q}^{Row})^\top \mathbf{G} \mathbf{Q}^{Col})] = \mathbb{E} [\det((\mathbf{Q}^{Row})^\top \mathbf{Q}^{Col})] \det(\mathbf{G}) \quad (24)$$

First, when  $\hat{x} = x$ , because  $\mathbf{Q} \in \mathcal{Q}_L$ , and  $N \geq 2Ld$ , every label has at least  $N \min_i q_x(i) \geq 2Ld \frac{1}{Ld-L+1} \geq 2$  reports by lemma D.5, and the minimum occurrence is at least two. Moreover,  $\mathbf{Q}^{Col} = \mathbf{Q}^{Row} = \mathbf{Q}$  are identity matrices regardless the choice of  $Col$  and  $Row$ , by eq. (24), we have

$$\mathbb{E}[\text{score}(x, y)] = \det(\mathbf{G}). \quad (25)$$

On the other hand, for  $\hat{x}$  with  $x \succ_{\text{EXACT}}^x \hat{x}$ , if the minimum occurrence is less than two, the score would be zero and less than eq. (28). Otherwise, by Cauchy-Schwarz inequality, we have

$$\mathbb{E} [\det((\mathbf{Q}^{Row})^\top \mathbf{Q}^{Col})] \leq \mathbb{E} [\det((\mathbf{Q}^{Row}))] \mathbb{E} [\det(\mathbf{Q}^{Col})] \quad (26)$$

Formally, consider  $\mathcal{I}$  the collection of all possible index set of size  $d$  where each label occurs exactly once. Then we can generate  $Col$  and  $Row$  by sampling two distinct  $(i, j)$  element of  $\mathcal{I}$  uniformly at random. In particular, if we set  $a_i$  be the determinant of the misreporting matrix of the  $i$ -th index set in  $\mathcal{I}$ , the joint distribution of  $(\det(\mathbf{Q}^{Col}), \det(\mathbf{Q}^{Row}))$  equals  $(a_i, a_j)$  and

$$\begin{aligned} & \mathbb{E} [\det((\mathbf{Q}^{Row}))] \mathbb{E} [\det(\mathbf{Q}^{Col})] - \mathbb{E} [\det((\mathbf{Q}^{Row})^\top \mathbf{Q}^{Col})] \\ &= \left( \frac{1}{|\mathcal{I}|} \sum_i a_i \right) \left( \frac{1}{|\mathcal{I}|} \sum_j a_j \right) - \frac{1}{|\mathcal{I}|(|\mathcal{I}| - 1)} \sum_{i \neq j \in \mathcal{I}} a_i a_j \\ &= \frac{1}{|\mathcal{I}|^2} \sum_i a_i^2 - \frac{1}{|\mathcal{I}|^2(|\mathcal{I}| - 1)} \sum_{i \neq j} a_i a_j \\ &= \frac{1}{2|\mathcal{I}|^2(|\mathcal{I}| - 1)} \sum_{i \neq j} (a_i - a_j)^2 \geq 0 \end{aligned}$$

which proves eq. (26). Finally, using the first part of theorem 4.2 and eqs. (23) to (26)

$$\mathbb{E}[\text{score}(\hat{\mathbf{x}}, \mathbf{y})] \leq \det(\mathbf{Q}^\top \mathbf{G} \mathbf{Q}) = \det(\hat{\mathbf{G}}) < \det(\mathbf{G}) = \mathbb{E}[\text{score}(\mathbf{x}, \mathbf{y})]$$

□

## F DETAILS AND PROOFS FOR SECTION 4.3

Now we provide examples to motivate kernelized Gram determinant scores in definition 4.6.

1. Given any feature map  $\phi : \mathcal{Y} \rightarrow \mathbb{R}^k$  that maps observations to Euclidean space, we define  $K(y, y') = \langle \phi(y), \phi(y') \rangle$  as the standard inner product between the features. A feature map is *injective* if the vectors  $\{\phi(y)\}_{y \in \mathcal{Y}}$  are linearly independent. For instance, using the one-hot encoder  $\phi : y \mapsto \delta_y$  results in delta-kernel  $K(y, y') = \mathbf{1}[y = y']$  and reproduces definition 4.1.
2. More generally, we can consider implicit feature maps, e.g., Gaussian radial basis function where  $K(y, y') = \exp\left(-\frac{\|y - y'\|_2^2}{\sigma^2}\right)$  for  $\mathcal{Y} \subseteq \mathbb{R}^k$ , or general Hilbert space. (Ziegel et al., 2022)
3. We can use feature maps to incorporate special structure in  $\mathcal{Y}$ , e.g., predictions of true labels. Formally, given  $\mathbf{P}$ , we say an observation  $y$  is a pseudo-posterior with prior  $\tilde{q} \in \Delta(\mathcal{X})$  if  $y = \{\tilde{\text{Pr}}[x = x | y]\}_{x \in \mathcal{X}} = \left\{ \frac{P(y, x) \tilde{q}(x)}{\sum_{x'} P(y, x') \tilde{q}(x')} \right\}_{x \in \mathcal{X}}$  is the posterior of true label under prior  $\tilde{q}$ . (Kass & Wasserman, 1996) Rather than one-hot encoder, we may consider  $\phi(y) = y \in \mathbb{R}^d$  which has smaller and meaningful feature space. We call the associated kernel  $K(y, y') = y^\top y'$  with pseudo posterior experiment as *pseudo-posterior kernel*.

We show that kernelized Gram determinant reliability scores also preserve all reliability orderings in theorem 4.2 for general observation space  $\mathcal{Y}$  under three kernel families. First, the result holds for any integrally strictly positive-definite kernel, so admitting arbitrary (possibly infinite or continuous) observation spaces. When  $\mathcal{Y}$  is finite, one may use any kernel with injective feature maps. The guarantee also holds when the observations are pseudo-posteriors with arbitrary prior  $\tilde{q}$  with full support.

**Theorem F.1.** *Given  $\mathcal{X} = [d]$ ,  $\mathcal{Y}$  and  $L \geq 1$ , the Gram determinant score with any of the following kernels preserve in definition 4.6 preserves reliability orderings in theorem 4.2:*

1. *Integrally strictly positive definite kernels—in particular the Gaussian (RBF) kernel on any separable Hilbert space  $\mathcal{Y}$ .*
2. *Kernels with an injective feature map  $\phi : \mathcal{Y} \rightarrow \mathbb{R}^k$  and finite set  $\mathcal{Y}$ .*
3. *Pseudo posterior kernel  $K$  with full support  $\tilde{q}$*

The proof is mostly identical to theorem 4.2. As the kernel only changes the Gram matrix of labels  $\mathbf{G}$ , it is sufficient to show  $\mathbf{G}$  is positive definite to reuse lemmas D.1 and D.3. Similarly, those two estimators in section 4.2 can also adopt kernels. We provide details in the appendix.

*Proof of theorem F.1.* To use lemmas D.1 to D.3, it is sufficient to show that  $\mathbf{G}$  is positive definite for all  $\mathbf{P} \in \mathcal{P}_{\text{indep}}$  so that for any nonzero sequence  $a : \mathcal{X} \rightarrow \mathbb{R}$ , the quadratic form is positive,

$$\sum_{x, x' \in \mathcal{X}} a(x) G(x, x') a(x') > 0. \quad (27)$$

First for any integrally strictly positive definite kernel, recall that the kernel mean embedding of  $P_x$  is  $\phi(P_x) = \mathbb{E}_{y \sim P_x}[\phi(y)] \in \mathcal{H}$ , so  $\sum_{x, x'} a(x) G(x, x') a(x') = \|\sum_x a(x) \phi(P_x)\|^2 \geq 0$  which shows  $\mathbf{G}$  is positive semi-definite. If the equality happens, by linearity of integration,  $0 = \|\sum_x a(x) \phi(P_x)\|^2 = \iint_{\mathcal{Y}} K(y, y') d\mu(y) d\mu(y')$  where  $\mu = \sum_x a(x) P_x$  is a finite signed measure. Therefore,  $\mu = \sum_x a(x) P_x = 0$  because  $K$  is integrally strictly positive definite. Finally  $a(x) = 0$  as columns of  $\mathbf{P}$  are linearly independent. Therefore the statement holds for integrally strictly positive definite kernels. Additionally, by (Ziegel et al., 2022, Theorem 3.1), the Gaussian kernel is integrally strictly positive definite.

Second, given a feature map  $\phi : \mathcal{Y} \rightarrow \mathbb{R}^k$ , eq. (27) becomes  $\|\sum_{x,y} a(x)P(y,x)\phi(y)\|_2^2$ . Because  $\mathbf{P} \in \mathcal{P}_{\text{indep}}$  and  $\phi$  is injective, the quadratic form equals zero if and only if  $a(x) = 0$  for all  $x$ . Moreover, delta kernel is injective, so the statement also holds.

Finally, for any pseudo-posterior observations, eq. (27) can be written as

$$\begin{aligned} & \left\langle \sum_{x,y} a(x)P(y,x)y, \sum_{x',y'} a(x')P(y',x')y' \right\rangle \\ &= \sum_{x,x',y,y'} a(x)a(x')P(y,x)P(y',x')\langle \tilde{P}[x|y], \tilde{P}[x'|y'] \rangle \\ &= \sum_{x,x',y,y'} a(x)a(x')P(y,x)P(y',x') \sum_{x''} \tilde{P}[x=x''|y] \tilde{P}[x'=x''|y'] \end{aligned}$$

Let  $b(y) = \sum_x a(x)P(y,x)$ ,  $w(y) = \sum_x P(y,x)\tilde{q}(x)$ . Then  $\sum_x \tilde{P}[x=x|y]\tilde{P}[x=x|y'] = \sum_x P(y,x)\frac{\tilde{q}(x)}{w(y)}P(y',x)\frac{\tilde{q}(x)}{w(y')}$ <sup>8</sup> and

$$\begin{aligned} & \sum_{x,x',y,y'} a(x)a(x')P(y,x)P(y',x') \sum_{x''} \tilde{P}[x=x''|y] \tilde{P}[x'=x''|y'] \\ &= \sum_{y,y'} b(y)b(y') \sum_x P(y,x)\frac{\tilde{q}(x)}{w(y)}P(y',x)\frac{\tilde{q}(x)}{w(y')} \\ &= \sum_x \tilde{q}(x)^2 \left( \sum_y P(y,x)\frac{b(y)}{w(y)} \right)^2 \end{aligned}$$

Because  $\tilde{q}$  has full support, the quadratic form equals zeros if and only if  $\sum_y P(y,x)\frac{b(y)}{w(y)} = 0$  for all  $x$ . Equivalently, if we set vector  $\mathbf{b} = \mathbf{P}\mathbf{a} \in \mathbb{R}^{|\mathcal{Y}|}$  and  $\mathbf{D}_w$  be the diagonal matrix with  $w$ , we have  $\mathbf{0} = \mathbf{b}^\top \mathbf{D}_w \mathbf{P} = \mathbf{a}^\top \mathbf{P}^\top \mathbf{D}_w \mathbf{P}$ . Since  $\mathbf{P}$  has full column rank and  $w(y) = 0$  when  $\mathbf{P}(x,y) = 0$  for all  $x$ ,  $a(x) = 0$  for all  $x$ .  $\square$

**Definition F.2** (Plugged-in Kernelized Gram determinant reliability score). *Given a kernel  $K : \mathcal{Y}^2 \rightarrow \mathbb{R}$ ,  $\hat{\mathbf{x}}, \mathbf{y}$  of length  $N$ , let  $\bar{\mathbf{G}} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  where*

$$\bar{\mathbf{G}}(x, x') = \frac{1}{N^2} \sum_{n, n' \in [N] : \hat{x}_n = x, \hat{x}_{n'} = x'} K(y_n, y_{n'}).$$

*The plugged-in kernelized Gram determinant reliability score is  $\bar{S}(\hat{\mathbf{x}}, \mathbf{y}) = \det(\bar{\mathbf{G}})$*

**Theorem F.3.** *Given  $\mathcal{X} = [d]$ , finite set  $\mathcal{Y}$  and  $L \geq 1$ , the plugged-in Gram determinant score with any bounded kernels in theorem 4.2 asymptotically preserves reliability orderings in theorem 4.2.*

**Lemma F.4.** *Given  $|K| \leq 1$ ,  $\delta > 0$  and report length  $N$ ,*

$$\Pr \left[ \|\bar{\mathbf{G}} - \hat{\mathbf{G}}\|_2 \leq 4\sqrt{\frac{\log 2d/\delta}{N}} + 2\frac{\log 2d/\delta}{N} \right] \geq 1 - \delta$$

The above lemma shows that the empirical estimator  $\bar{\mathbf{G}}$  is close to its expectation  $\hat{\mathbf{G}}$  in spectral norm. The proof is mostly identical to lemma E.3. We only need concentration results of the sum of independent random elements in Hilbert spaces as (Pinelis, 1994, Theorem 3.5).

Finally, we can also design an estimator that preserves exact match ordering even with finite length data.

**Definition F.5.** *Given a kernel  $K : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ ,  $\hat{\mathbf{x}}, \mathbf{y}$  of length  $N$ , a stratified-matching estimator estimates the kernelized Gram determinant as the following*

1. *Return 0 if the minimum occurrence  $\min_{x \in \mathcal{X}} |\{n \in [N] : \hat{x}_n = x\}|$  is less than 2. Otherwise, we randomly select two disjoint index sets  $Col, Row \subseteq [N]$  of size  $d$  where each label  $i \in \mathcal{X}$  occurs in each set exactly once. Then re-index them as two sequences of pairs  $(\hat{x}_{i,Col}, y_{i,Col})_{i \in [d]}$  and  $(\hat{x}_{i,Row}, y_{i,Row})_{i \in [d]}$  so that  $\hat{x}_{i,Col} = \hat{x}_{i,Row} = i$  for all  $i \in \mathcal{X}$ . We call the first as column sequence and the second as row sequence.*

<sup>8</sup>Here we set  $0/0 = 0$  if  $w(y) = 0$

2. Randomly sample one permutations  $\sigma \in \text{sym}(d)$ , and return

$$\text{score}(\hat{\mathbf{x}}, \mathbf{y}) := d! \text{sgn}(\sigma) \prod_{i,j \in [d], j=\sigma(i)} K(y_{i, \text{Row}}, y_{j, \text{Col}}) q_{\hat{\mathbf{x}}}(i) q_{\hat{\mathbf{x}}}(j). \quad (28)$$

**Theorem F.6.** Given  $\mathcal{X} = [d]$  and  $L \geq 1$ , the stratified-matching estimator in definition E.1 with any of kernels in theorem 4.2 preserves exact ordering on  $\mathcal{P}_{\text{indep}}$ ,  $\mathcal{Q}_L$ , and  $N \geq 2Ld$ .

The proof is mostly identical to proposition E.2

## G EXPERIMENT DETAILS AND ADDITIONAL EXPERIMENTS

### G.1 EXPERIMENT DETAILS

Due to space limitations, we omit some settings in the main paper. Here, we provide the details of how we compute error bars and how we obtain the ranking-accuracy across sample sizes in fig. 2d.

**Error bars.** Let  $M$  be the number of independent trials. For each trial  $m \in [M]$ , let  $\text{score}^{(m)}$  denote the determinant score, and similarly let  $\text{Hamming}^{(m)}$  and  $\ell_2^{(m)}$  denote the Hamming distance and  $\ell_2$ -norm error, respectively. We compute the sample mean

$$\overline{\text{score}} = \frac{1}{M} \sum_{m=1}^M \text{score}^{(m)}$$

and the standard error of the mean

$$SE(\overline{\text{score}}) = \frac{1}{\sqrt{M(M-1)}} \sqrt{\sum_{m=1}^M (\text{score}^{(m)} - \overline{\text{score}})^2}.$$

Under approximate normality, we report a 95% confidence interval as  $\overline{\text{score}} \pm 1.96SE(\overline{\text{score}})$  in fig. 2a, fig. 2b and fig. 2c. The same procedure is applied to  $\text{Hamming}^{(m)}$  and  $\ell_2^{(m)}$  to yield their error bars in figs. 2b and 2c.

**Ranking accuracy across sample sizes.** In fig. 2d (right), we plot the fraction of trials in which the reversed ranking induced by the determinant score agrees with the ranking induced by each baseline metric—namely the reporting probability  $p$ , the Hamming distance, and the  $\ell_2$ -norm—over six noise levels  $\mathcal{P} = \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ . Concretely, in each trial  $m$  we form three vectors

$$(\text{score}_p^{(m)})_{p \in \mathcal{P}}, \quad (\text{Hamming}_p^{(m)})_{p \in \mathcal{P}}, \quad (\ell_2^{(m)})_{p \in \mathcal{P}}.$$

We then check whether the total order of  $(\text{score}_p^{(m)})$  in decreasing order matches the order of  $(\text{Hamming}_p^{(m)})$  in increasing order (and similarly for  $\ell_2$  and for  $p$  itself). If they coincide, trial  $m$  is counted as a “correct” ranking. The plotted accuracy is

$$\frac{1}{M} \sum_{m=1}^M \mathbf{1}\{\text{orders agree in trial } m\}.$$

A random guess among the  $6!$  possible orderings yields a baseline accuracy of  $1/6! \approx 0.0008$ .

### G.2 MORE EXPERIMENTS ON CATEGORICAL SYNTHETIC DATA

In this section, we still use the dataset and experiment settings from Experiment 1. However, besides the uniform random manipulation introduced in eq. (6), we consider *normal manipulation*:

$$\hat{\mathbf{x}}_k \sim \text{clip}(1, d, \text{round}(\mathcal{N}(\mathbf{x}, \sigma_0))). \quad (29)$$

This manipulation introduces localized perturbations to the data. We adopt  $\sigma_0 \in \{0.30, 0.37, 0.44, \dots, 1.00\}$  in our experiments and refer to this type of manipulation as normal

Figure	Experiments	Manipulation	Kernel
Figure 2	Random exp	uniform eq. (6)	delta
Figure 4	Random exp	normal eq. (29)	delta
Figure 5	Random exp	uniform	Gaussian
Figure 6	Random exp	normal	Gaussian
Figure 7	Gaussian	uniform	Gaussian
Figure 7	Gaussian	normal	Gaussian
Figure 8	Gaussian	uniform	delta
Figure 6	Gaussian	normal	delta

Table 1: Table for settings of the experiments on synthetic data.

manipulation. For the matched rankings results, the rankings are computed for 6 data points with  $\sigma_0 \in \{0.30, 0.44, 0.58, 0.72, 0.86, 1.00\}$ .

Moreover, we also use the *Gaussian kernel*

$$K(y, y') = \exp\left(-100 \|y - y'\|_2^2\right)$$

besides the delta kernel  $K(y, y') = \mathbf{1}[y = y']$ .

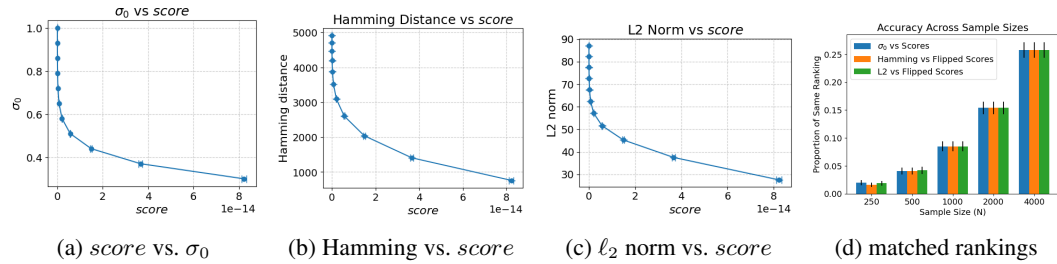


Figure 4: Experiments of plugged-in Gram determinant reliability score with delta kernel on categorical synthetic data with normal manipulation in eq. (29).

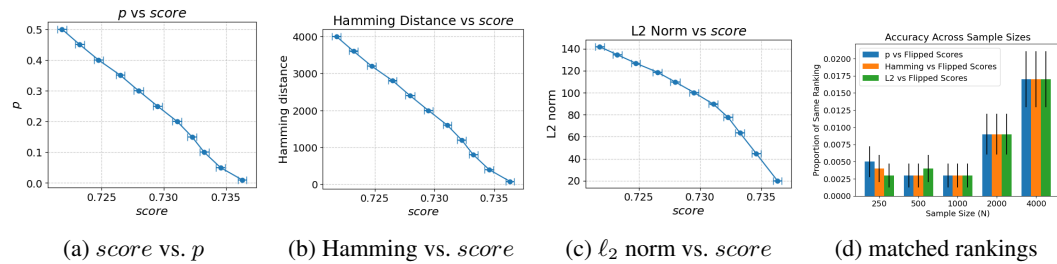


Figure 5: Experiments of plugged-in Gram determinant reliability score with Gaussian kernel on categorical synthetic data with uniformly random manipulation in eq. (6).

From figs. 4 to 6, we can conclude that regardless of the kernel used for the Gram determinant score, it consistently behaves well as a measure of reliability. It exhibits a negative correlation with all reliability metrics. As the sample size increases, the ability of the Gram determinant score to align with other reliability metrics also improves. For this categorical dataset, the delta kernel outperforms the Gaussian kernel, achieving both a smaller standard mean error and higher accuracy for matched rankings across sample sizes.

### G.3 EXPERIMENT 3: GRAM DETERMINANT SCORE ON GAUSSIAN SYNTHETIC DATA

In this section, we create a new synthetic dataset. Instead of random experiment, each  $y_i$  is sampled from a normal distribution  $\mathcal{N}(x_i, \sigma)$  centered at  $x_i$ . We adopt  $\sigma = 0.1$  and  $d = 4$  in this experiment, and all other experimental settings remain the same as in Experiment 1.



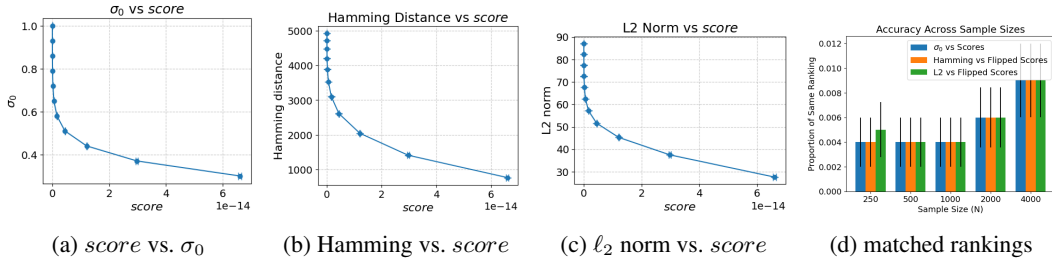


Figure 6: Experiments of plugged-in Gram determinant reliability score with Gaussian kernel on categorical synthetic data with normal manipulation in eq. (29).

Since  $\mathcal{Y}$  lies in the continuous space  $\mathbb{R}$  rather than being categorical, we cannot directly apply the plugged-in Gram determinant score with the delta kernel. Hence, we use an approximate scoring method: we create a new sequence  $\bar{y}$  by bucketing  $y$  into  $d$  bins using empirical quantiles, and then apply the plugged-in Gram determinant score on  $\hat{x}$  and  $\bar{y}$ .

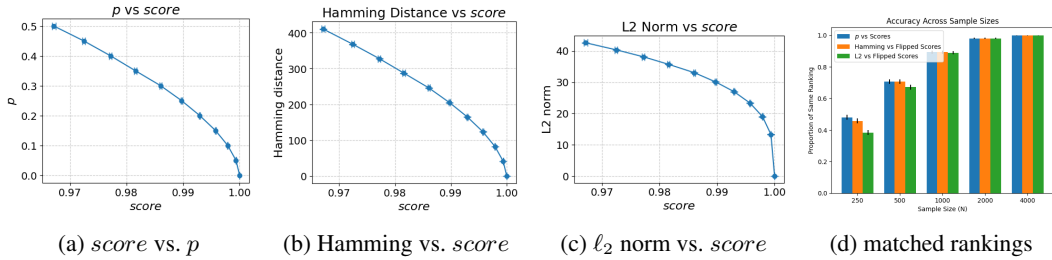


Figure 7: Experiments of plugged-in Gram determinant reliability score with Gaussian kernel on Gaussian synthetic data with uniformly random manipulation in eq. (6).

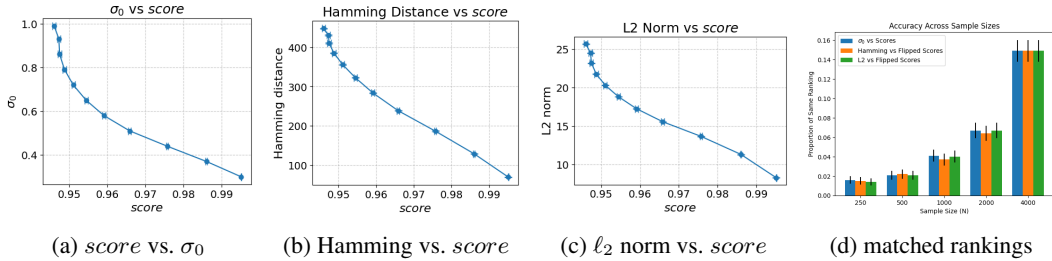


Figure 8: Experiments of plugged-in Gram determinant reliability score with Gaussian kernel on Gaussian synthetic data with normal manipulation in eq. (29).

From figs. 7 to 10, we observe that both the delta kernel and the Gaussian kernel perform well as reliability measures across all three metrics, under both normal and uniformly random manipulations. In particular, the delta kernel variant using bucketed  $y$  achieves consistently strong performance. Empirically, the plugged-in Gram determinant score with the delta kernel generally outperforms the version with the Gaussian kernel in most situations, despite the lack of theoretical guarantees for this delta kernel variant. For reported data with small  $\ell_2$  norm error, the Gaussian kernel outperforms the approximate delta kernel score.

## H ALTERNATIVES TO GRAM DETERMINANT SCORE

### H.1 COMPARISON TO KONG (2024)

Our Gram determinant score can be viewed as an application of the peer prediction mechanism introduced in (Kong, 2024), where one agent’s report is replaced with the observation  $y$ . In addition

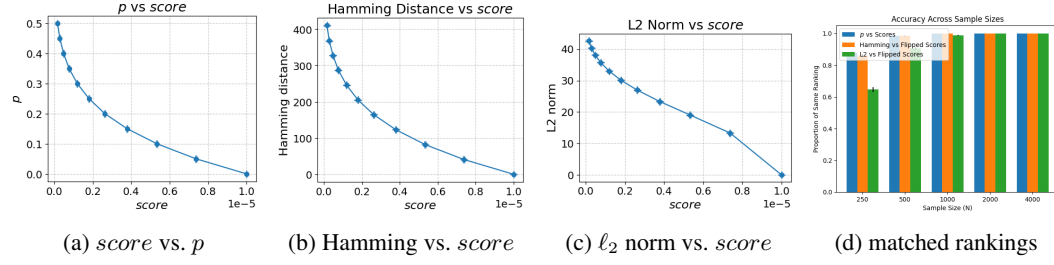


Figure 9: Experiments of plugged-in Gram determinant reliability score with delta kernel on Gaussian synthetic data with uniformly random manipulation in eq. (6).

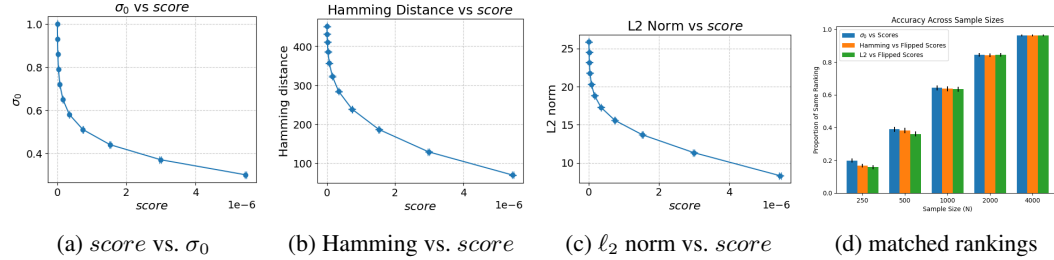


Figure 10: Experiments of plugged-in Gram determinant reliability score with delta kernel on Gaussian synthetic data with normal manipulation in eq. (29).

to offering a more fine-grained characterization of the Gram determinant score, as discussed in related work, we also introduce several technical improvements over the original determinant mutual information method. First, the prior approach requires  $\mathcal{Y} = \mathcal{X}$  and overlooks potential structure in the observations. As shown in section 4.3, we address this by introducing kernel methods, allowing us to generalize the score to arbitrary observation spaces  $\mathcal{Y}$ —a crucial extension for handling continuous observations such as Gaussian variables or image embeddings, as demonstrated in section 5. Second, our stratified-matching estimators in definitions E.1 and F.5 are unbiased in the multi-task peer prediction setting of (Kong, 2024), and they reduce the estimator’s range from order  $(d!)^2$  to  $d!$ .

## H.2 MORE DATA RELIABILITY SCORES

There is a long line of research on measuring the stochastic relationship between random variables. We may view them as data reliability scores applying to the reported data  $\hat{x}$  and observation  $y$ . In this section, we list some common candidates and illustrate the limitations and possibilities.

### $\Phi$ -mutual information

**Definition H.1** ( $\Phi$ -divergence (Csiszár, 1964; Morimoto, 1963; Ali & Silvey, 1966)). Let  $\Phi : [0, \infty) \rightarrow \mathbb{R}$  be a convex function with  $\Phi(1) = 0$ . Let  $P$  and  $Q$  be two probability distributions on a common measurable space  $(\Omega, \mathcal{F})$ . The  $\Phi$ -divergence of  $Q$  from  $P$  where  $P \ll Q^9$  is defined as  $D_\Phi(P\|Q) := \mathbb{E}_Q[\Phi(P/Q)]$ .<sup>10</sup>

We can use these divergences to measure how interdependent between two random variables  $x$  and  $y$ . Formally, Let  $P_{x,y}$  be a distribution over  $(x, y) \in \mathcal{X} \times \mathcal{Y}$ , and  $P_x$  and  $P_y$  be marginal distributions of  $x$  and  $y$  respectively. We set  $P_x P_y$  be the tensor product between  $P_x$  and  $P_y$  such that  $P_x P_y(x, y) = P_x(x)P_y(y)$ . We call  $D_\Phi(P_{x,y}\|P_x P_y)$  the  $\Phi$ -mutual information between  $x$  and  $y$ .

1. Total variation has  $\Phi(a)$  as  $\frac{1}{2}|a - 1|$ .

<sup>9</sup> $P$  is absolutely continuous with respect to  $Q$ : for any measurable set  $A \in \mathcal{F}$ ,  $Q(A) = 0 \Rightarrow P(A) = 0$ .

<sup>10</sup> $P/Q$  is the Radon-Nikodym derivative between measures  $P$  and  $Q$ , and it is equal to the ratio of density function.

2. KL-divergence has  $a \log a$
3.  $\chi^2$ -divergence has  $a^2 - 1$
4. Squared Hellinger distance has  $(1 - \sqrt{a})^2$

In the partial knowledge setting, we can access the  $\mathbf{J} := \mathbf{PQ}$  which can be seen as a joint distribution between reported data and observation  $\mathbf{J} = P_{x,y}$ , and set

$$S_\Phi(\mathbf{PQ}) = D_\Phi(P_{x,y} \| P_x P_y).$$

These family of score satisfy the data processing inequality, which is analogous to our *weak* Blackwell dominant ordering so that garbling the report can only decrease the score. Nevertheless, the impossibility results in section 3 still apply. In addition, they are generally not experiment-agnostic, and lack kernelized extensions as in section 4.3 for complicated observation space  $\mathcal{Y}$ .

**Family of symmetric gauge on singular values** Our Gram determinant is essentially a functional on the singular values of  $\mathbf{J} = \mathbf{PQ}$  and sub multiplicative under right multiplication by contraction. One may additionally consider functional on the singular values of the whitened matrix. Formally, given a joint distribution  $\mathbf{J} := \mathbf{PQ}$ , let

$$\bar{\mathbf{J}} = \mathbf{D}_y^{-1/2}(\mathbf{J} - \mu_y \mu_x^\top) \mathbf{D}_x^{-1/2}$$

where  $\mu_x$  and  $\mu_y$  are marginal distributions and  $\mathbf{D}_x, \mathbf{D}_y$  are diagonal matrix of them respectively. Given a matrix  $\mathbf{A}$ ,  $\sigma(\mathbf{A})$  denote the singular value list of  $\mathbf{A}$ , we can find a symmetric gauge  $\psi$  and define our score as

$$S_\psi(\mathbf{J}) = \psi(\sigma(\bar{\mathbf{J}})).$$

Let  $\bar{s} = \sigma(\bar{\mathbf{J}}) = (\bar{s}_1, \dots, \bar{s}_d)$  with  $\bar{s}_1 \geq \bar{s}_2 \geq \dots \bar{s}_d \geq 0$ .

1. Top- $k$  volume has  $\psi_{\wedge k}(\mathbf{s}) = \prod_{i=1}^k \bar{s}_i$
2. Maximal correlation  $\psi_{\max} = \bar{s}_1$ . The maximum correlation can be also written as

$$\max_{(f,g) \in \mathcal{S}} \mathbb{E}[f(\mathbf{x})g(\mathbf{y})]$$

where  $\mathcal{S}$  is the collection of real-valued random variables so that  $\mathbb{E}f(\mathbf{x}) = \mathbb{E}g(\mathbf{y}) = 0$  and  $\mathbb{E}f(\mathbf{x})^2 = \mathbb{E}g(\mathbf{y})^2 = 1$ .

3. Ky-Fan  $k$ -sum  $\sum_{i=1}^k \bar{s}_i$
4.  $\chi^2$ -mutual information  $I_{\chi^2}(\mathbf{x}, \mathbf{y}) = \sum_{x,y} \mu_x(x) \mu_y(y) \left( \frac{\mathbf{J}(y,x)}{\mu_x(x) \mu_y(y)} - 1 \right)^2 = \|\bar{\mathbf{J}}\|_F = \sum_i \bar{s}_i^2$

Similarly, the impossibility results in section 3 still apply and they are generally not experiment-agnostic.

### H.3 EXPERIMENTS ON SCORE COMPARISON

We follow the same data generation process and manipulation policies as in Experiment 1, and focus here on comparing a family of reliability scores computed from the empirical joint distribution  $\mathbf{J} = \mathbf{PQ}$ .

Across manipulations and keep-probabilities  $p$ , these scores exhibit highly consistent behavior: larger values indicate less corruption, and in practice they are inversely related to the corruption level as measured by  $1 - p$ , the Hamming distance, and the  $\ell_2$  norm between  $\mathbf{x}$  and  $\hat{\mathbf{x}}$  (see figs. 11 to 15). This alignment across multiple metrics demonstrates that the proposed scores all provide robust and informative signals of data quality.

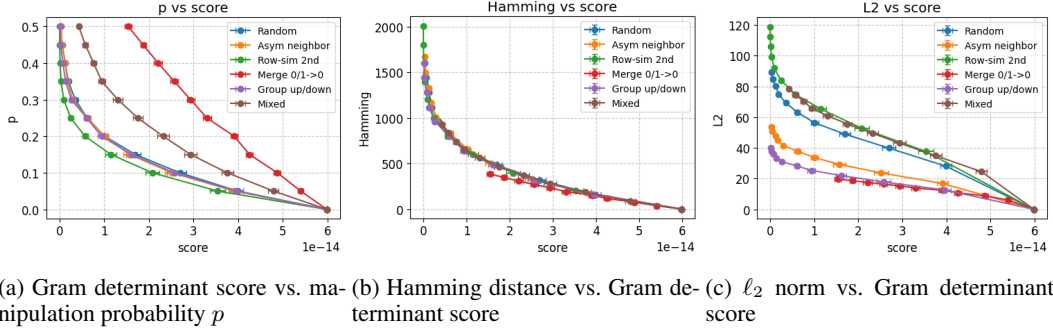


Figure 11: Comparison of the Gram determinant reliability score across different corruption levels and metrics.

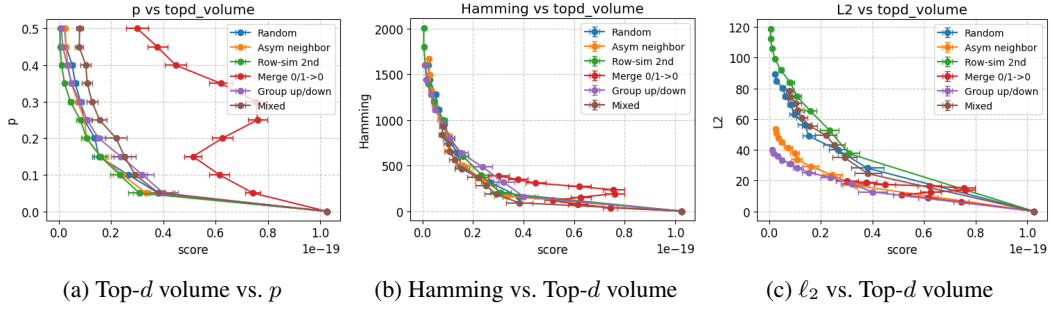


Figure 12: Comparison of the Top- $d$  volume score under different corruption levels.

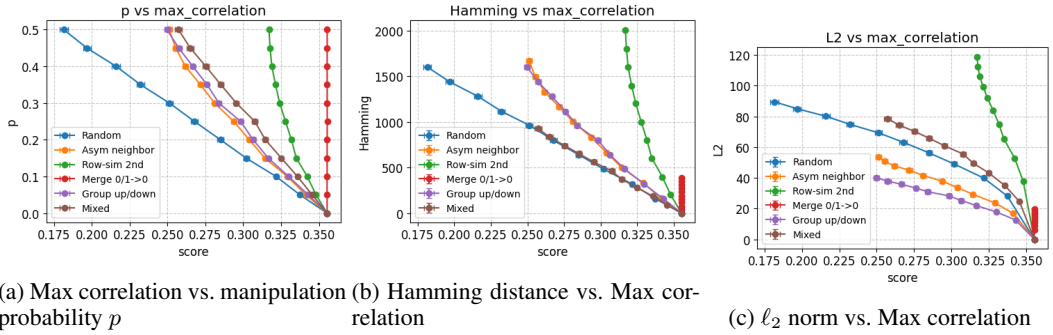


Figure 13: Comparison of the Max correlation score under different corruption levels and metrics.

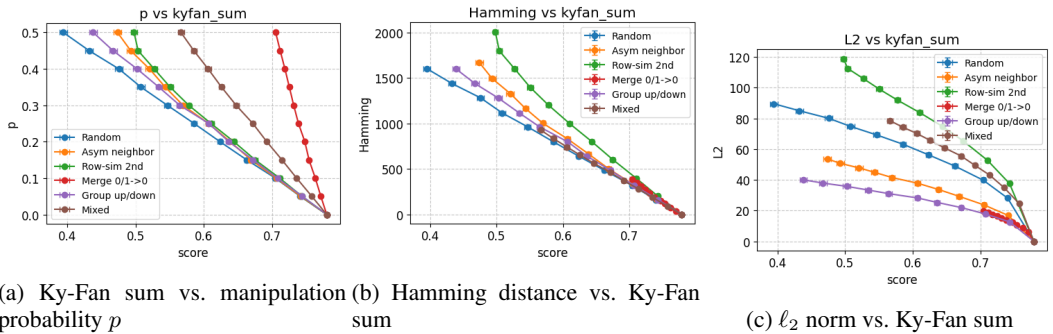


Figure 14: Comparison of the Ky-Fan sum score under different corruption levels and metrics.

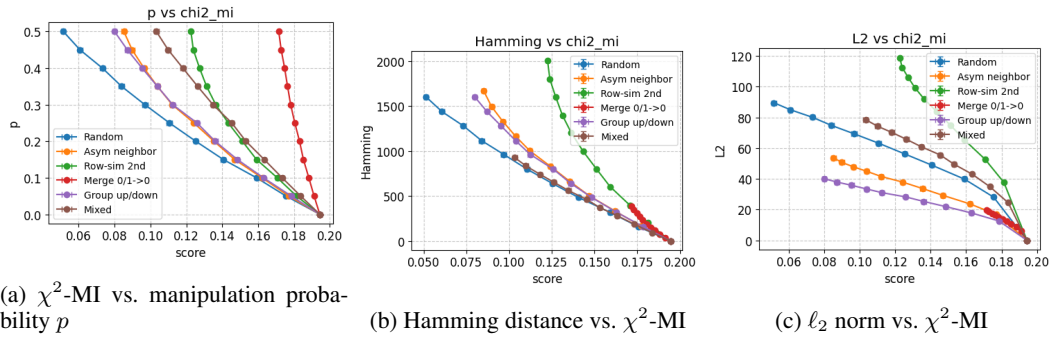


Figure 15: Comparison of the  $\chi^2$ -mutual information score under different corruption levels and metrics.