# Don't Match the Noise: Distribution Matching under Unknown Measurement Error with an Audit Sample

**Arya Farahi** **Ritwik Vashistha**
Department of Statistics and Data Science, University of Texas at Austin, Austin, Texas 78712, USA
{arya.farahi, ritwik.v}@utexas.edu

## Abstract

Real-world machine learning systems often operate on imperfect measurements: sensors are noisy, annotations are produced by non-experts, and scientific observations are often indirect. In these settings, classic distribution comparison tools such as maximum mean discrepancy (MMD) can be systematically biased, matching artifacts of the measurement process rather than the underlying mechanisms of interest. We study a practical regime in which noisy observations are abundant, but only a small audited subset contains ground-truth measurements. We propose measurement-error-corrected MMD (MEC-MMD), an audit-based estimator that debiases kernel evaluations without requiring a noise model. MEC-MMD is exactly unbiased for the oracle MMD under simple random auditing, admits an explicit variance decomposition, and is consistent for the population MMD, achieving the standard $O_p(N^{-1/2})$ rate when the audit size scales with the dataset. We demonstrate the utility of distributional matching with MEC-MMD loss for both parameter estimation and unsupervised applications, showing that MEC-MMD consistently outperforms MMD and alternative methods. These results provide a principled foundation for robust distribution comparison, simulation-based inference, and learning in noisy real-world machine learning pipelines.

## 1 Introduction

In many scientific applications, the quantities of interest are not observed directly but only through noisy or imperfect measurements of latent variables. This setting is pervasive across domains including astronomy [1–3], remote sensing [4], nutrition sciences [5], experimental biology [6, 7], economics [8, 9], sports analytics [10], and differential privacy [11]. The data-generation process often involves instrument noise, preprocessing artifacts, discretization (e.g., rounding or binning), or the use of proxies. Hence, the analyst observes a proxy $\widetilde{Y}$ rather than the scientific target $Y$. Naïve procedures that ignore this discrepancy typically learn the distribution of $\widetilde{Y}$ instead of that of $Y$, leading to biased or scientifically misleading conclusions [12]. This issue is particularly acute in simulation-based inference, where parameters are estimated by matching simulated and observed data distributions [13, 14]. When observations are corrupted, discrepancies computed directly on $\widetilde{Y}$ favor simulators that reproduce the measurement process rather than the underlying latent mechanism. Properly accounting for measurement error is therefore essential for valid scientific inference.

In this work, we focus on *measurement error* [15, 16], that differ from data contamination [17, 18]. Under contamination [e.g., 19, 20], a subset of observations arises from an extraneous distribution; under measurement error, each observation corresponds to the correct latent unit but is recorded imperfectly. Classical approaches to measurement error include deconvolution density estimation [21–24], errors-in-variables regression [15, 25], simulation-extrapolation methods [SIMEX; 26–28], and likelihood-based latent-variable models [29–31]. While these methods have led to major theoretical and practical advances, they also face well-known challenges. Deconvolution is often ill-posed and sensitive to assumptions on the noise distribution; errors-in-variables methods typically rely on parametric structure, instrumental variables, or replicate measurements for identifiability; SIMEX can be sensitive to extrapolation error and model misspecification; and likelihood-based approaches require a correctly specified and tractable joint model of the latent variables and noise mechanism, which is difficult to justify in modern high-dimensional or weakly supervised settings.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Many modern data pipelines provide an additional source of information that is not fully leveraged by these classical frameworks: a small *audited* or *validation* subset in which the latent quantity is observed exactly [e.g., 32–37]. One often has access to a large sample of noisy observations $\widetilde{Y}$ together with a much smaller subset for which both $\widetilde{Y}$ and the corresponding clean value $Y$ are available, allowing the realized measurement error $\varepsilon = \widetilde{Y} - Y$ to be observed on audited points. This regime (abundant noisy data paired with sparse gold-standard measurements) arises naturally in scientific workflows, for example, through manual annotation, high-precision subsampling, or costly experimental validation (see Table 5 in Supplementary Materials). It lies between fully supervised settings (where clean labels are widely available) and unsupervised settings (where only proxies are observed), and calls for methods that can combine statistical efficiency from large noisy samples with principled bias correction from limited clean data [4, 38–40].

We propose to address this problem using the maximum mean discrepancy [MMD; 41], a kernel-based distance between distributions that is well-suited to likelihood-free testing and inference [42–50]. The proposed method's key advantage is that it is defined through expectations of kernel evaluations, which enables direct correction for measurement error at the level of the empirical loss. This avoids explicit deconvolution or full likelihood specification while remaining compatible with flexible simulators and gradient-based optimization. Building on this idea, we develop corrected MMD objectives for parameter estimation under unknown measurement error. Assuming access to many noisy observations and a smaller, representative audited subset, we construct debiased empirical MMD estimators that target the latent distribution of $Y$ rather than the proxy distribution of $\widetilde{Y}$. The approach is distribution-free, scalable, and readily applicable to modern scientific problems in which the measurements are made with noise. See positioning statement in Supplementary Materials.

## 2 Additive Measurement Error with an Audit Sample

**Background.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a measurable space on which the probability distributions are defined. We denote $X$, a random variable taking values in $\mathcal{X}$ according to a probability distribution $p$, and by $x \in X$, a realization of $X$. Similarly, we denote $Y$, a random variable taking values in $\mathcal{X}$ according to a probability distribution $q$, and by $y \in Y$, a realization of $Y$. Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space (RKHS, [51, 52]) of real-valued functions defined on $\mathcal{X}$. The MMD is then defined as

$$\mathrm{MMD}(p, q) = \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq 1} |\mathbb{E}_p(f(X)) - \mathbb{E}_q(f(Y))|. \tag{1}$$

MMD quantifies dissimilarity between two distributions by measuring the distance between their respective mean embeddings in the RKHS. The MMD statistic can be simplified using kernels, which provide a rich structure for analysis and computation.

**Assumption 2.1.** There is $K < \infty$ with $\sup_{x,x' \in \mathcal{X}} |k(x, x')| \leq K$.

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded positive definite kernel associated with $\mathcal{H}$ (Assumption 2.1). A closed computational form for $\mathrm{MMD}^2$ is $\mathrm{MMD}^2(p, q) = \mathbb{E}_{X,X'}\left[k\left(X, X'\right)\right] + \mathbb{E}_{Y,Y'}\left[k\left(Y, Y'\right)\right] - 2\mathbb{E}_{X,Y}\left[k\left(X, Y\right)\right]$, with $X, X' \overset{iid}{\sim} p$ and $Y, Y' \overset{iid}{\sim} q$. For characteristic kernels [53], it has been shown that $\mathrm{MMD}(p, q) = 0$ if and only if $p = q$, thus giving a consistent test [41]. This property justifies MMD as a loss function in learning and inference. See Table 6 in the Supplementary Materials for a notation summary.

**Problem setup.** We study a two-sample problem in which the reference sample is observed without error, while the target sample is observed only through noisy surrogates. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p$ and $Y_1, \ldots, Y_N \overset{\text{iid}}{\sim} q$, where $p$ and $q$ are probability distributions on $\mathcal{X}$. The sample $\{X_i\}_{i=1}^n$ is observed directly. By contrast, the latent target variables $\{Y_j\}_{j=1}^N$ are not fully observed. Instead, for each $j$, the analyst observes a noisy proxy

$$\widetilde{Y}_j = Y_j + \varepsilon_j, \qquad \varepsilon_j \sim P(\cdot \mid \phi_j), \qquad \phi_j \sim \Pi, \tag{2}$$

where $\phi_j$ is a unit-specific noise parameter (e.g., a noise scale). Thus, the law of the measurement error may vary across observations, allowing for heteroscedastic or otherwise structured noise, for instance when it depends on unit-level covariates. Such heterogeneity is common in practice [e.g., 54–60]. We assume that the tuples $(Y_j, \phi_j, \varepsilon_j)$ are independent across $j$, and independent of the clean sample $\{X_i\}_{i=1}^n$. Our method does not model the noise or any covariate driving it.

**Target Quantity.** Our goal is to estimate or test the squared maximum mean discrepancy between the clean reference distribution $p$ and the latent target distribution $q$,

$$\mathrm{MMD}^2(p, q; k) = \mathbb{E}[k(X, X')] + \mathbb{E}[k(Y, Y')] - 2\,\mathbb{E}[k(X, Y)]. \tag{3}$$

We recall the fact that when the observed target sample is corrupted by measurement error

$$\mathbb{E}[k(X,X')] + \mathbb{E}[k(Y,Y')] - 2\,\mathbb{E}[k(X,Y)] \neq \mathbb{E}[k(X,X')] + \mathbb{E}[k(\widetilde{Y},\widetilde{Y}')] - 2\,\mathbb{E}[k(X,\widetilde{Y})]. \quad (4)$$

The main challenge is structural. We formalize an estimand that is defined in terms of the latent distribution $q$. MMD depends nonlinearly on pairwise kernel evaluations involving the unobserved latent variables $Y_j$. In contrast to settings with erroneous labels in 0-1 loss [61], we pointwise reweight the observed proxies to recover the target quantity and correct the kernel terms directly.

**Oracle Estimator.** The natural (best case scenario) benchmark would be the usual empirical MMD estimator formed from the clean sample $\{X_i\}_{i=1}^n$ and the latent sample $\{Y_j\}_{j=1}^N$. Define

$$A_{pp}^\star := \sum_{i\neq i'} k(X_i, X_{i'}), \quad A_{qq}^\star := \sum_{j\neq j'} k(Y_j, Y_{j'}), \quad A_{pq}^\star := \sum_{i=1}^n \sum_{j=1}^N k(X_i, Y_j). \quad (5)$$

The resulting oracle estimator is

$$\widehat{\mathrm{MMD}}_\star^2 = \frac{A_{pp}^\star}{n(n-1)} + \frac{A_{qq}^\star}{N(N-1)} - 2\,\frac{A_{pq}^\star}{nN}. \quad (6)$$

This statistic is not directly available because most latent values $Y_j$ are unobserved. Nevertheless, it is the finite-sample target of the correction procedure below. A useful feature of the present setting is that every $\widetilde{Y}_j$ still corresponds to exactly one genuine latent draw $Y_j \sim q$, even when the measurement error depends on covariates. Thus, the target sample size remains $N$. The difficulty lies entirely in recovering the latent kernel evaluations that appear in $A_{qq}^\star$ and $A_{pq}^\star$.

**Audit Sample.** We further assume access to an *audit sample*, a small subset of target indices for which the latent values are measured exactly. Formally, let $L \subset \{1,\dots,N\}$, where $|L| = m \ll N$, and suppose that for every $j \in L$ the analyst observes $Y_j$ in addition to $\widetilde{Y}_j$. Equivalently, the realized measurement error $\varepsilon_j = \widetilde{Y}_j - Y_j$ is observed on the audited units. We, hence, assume the analyst has access to $\{\widetilde{Y}_j\}_{j=1}^N$, $\{Y_j : j \in L\}$ and $\{X_i\}_{i=1}^n$.

**Assumption 2.2.** The audit set $L \subset \{1,\dots,N\}$ is a random sample without replacement of fixed size $m$ with $2 \le m \le N$, drawn independently of $\{X_i\}_{i=1}^n$ and $\{(Y_j, \widetilde{Y}_j)\}_{j=1}^N$ where selection is independent of the measurement error. Hence $\Pr(j \in L) = \frac{m}{N}$ and $\Pr(j, j' \in L) = \frac{m(m-1)}{N(N-1)}$ for $j \neq j'$.

Unless stated otherwise, expectations of the form $\mathbb{E}_L[\cdot]$ are taken with respect to this audit-design randomness only, conditional on the realized finite population. The role of the audit sample is to reveal enough latent information to reconstruct, in a design-unbiased manner, the kernel contributions that would have been available if the full target sample had been observed without error. Section E relaxes this to known, unequal inclusion probabilities.

**Assumption 2.3.** $n, N, m \to \infty$ with $m/N \to \rho \in (0,1]$ and $n/N \to \gamma \in (0,\infty)$.

## 3 Method and Theoretical Guarantees

While the noisy observations $\widetilde{Y}_j$ provide a cheap proxy for the latent $Y_j$, $\widehat{\mathrm{MMD}}^2(X,Y,k) \neq \widehat{\mathrm{MMD}}^2(X,\widetilde{Y},k)$ even when $p = q$. To construct an estimator of $\mathrm{MMD}^2(p,q,k)$, we begin with the plug-in MMD on $\widetilde{Y}$, and then use the audited subset to correct it.

### 3.1 Estimating Oracle MMD

We have the following algebraic identities. For every $j \neq j'$,

$$k(Y_j, Y_{j'}) = k(\widetilde{Y}_j, \widetilde{Y}_{j'}) + \Gamma_{jj'}, \qquad j, j' = 1,\dots,N, \quad (7)$$

where $\Gamma_{jj'} := k(Y_j, Y_{j'}) - k(\widetilde{Y}_j, \widetilde{Y}_{j'})$. Likewise, for every $i, j$,

$$k(X_i, Y_j) = k(X_i, \widetilde{Y}_j) + \Lambda_{ij}, \qquad j = 1,\dots,N, \text{ and } \quad i = 1,\dots,n, \quad (8)$$

where $\Lambda_{ij} := k(X_i, Y_j) - k(X_i, \widetilde{Y}_j)$. The quantities $\Gamma_{jj'}$ and $\Lambda_{ij}$ exist for every $j$ and $j'$ but are only observed when $j, j' \in L$. Since these are exact algebraic identities, no approximation is involved. Now, define the corrected $qq$ kernel sum by

$$\widehat{A}_{qq} := \sum_{j\neq j'} k(\widetilde{Y}_j, \widetilde{Y}_{j'}) + \frac{N(N-1)}{m(m-1)} \sum_{\substack{j,j'\in L \\ j\neq j'}} \Gamma_{jj'}, \quad (9)$$

and the corrected $pq$ kernel sum by

$$\widehat{A}_{pq} := \sum_{i=1}^{n}\sum_{j=1}^{N} k(X_i, \widetilde{Y}_j) + \frac{N}{m}\sum_{i=1}^{n}\sum_{j\in L}\Lambda_{ij}. \tag{10}$$

Because the denominators are known, the measurement-error-corrected MMD estimator is simply

$$\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} = \frac{A^\star_{pp}}{n(n-1)} + \frac{\widehat{A}_{qq}}{N(N-1)} - 2\,\frac{\widehat{A}_{pq}}{nN} \tag{11}$$

Computing $\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}$ takes $O(n^2+N^2+nN)$ kernel evaluations, as for the plug-in estimator, plus $O(m^2+nm)$ for the audit correction. Appendix D reports runtime and memory.

### 3.2 Design-Based Guarantees and Asymptotics

Recall the Horvitz–Thompson identities (Lemma B.1) under Assumption 2.2 [62].

**Theorem 3.1.** *Conditional on the realized samples $\{X_i\}_{i=1}^n$ and $\{(Y_j, \widetilde{Y}_j)\}_{j=1}^N$. Under Assumption 2.2, $\mathbb{E}_L[\widehat{A}_{qq}] = A^\star_{qq}$ and $\mathbb{E}_L[\widehat{A}_{pq}] = A^\star_{pq}$. Consequently, $\mathbb{E}_L\left[\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}\right] = \widehat{\mathrm{MMD}}^2_{\star}$.*

Theorem 3.1 proves that $\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}$ is indeed an unbiased estimator of the oracle $\widehat{\mathrm{MMD}}^2_{\star}$. Now, we compute the variance.

**Theorem 3.2.** *Conditional on the realized samples $\{X_i\}_{i=1}^n$ and $\{(Y_j, \widetilde{Y}_j)\}_{j=1}^N$. Suppose Assumptions 2.2 and 2.1 hold and $N \geq 4$. Define*

$$\Lambda_j := \frac{1}{n}\sum_{i=1}^{n}\Lambda_{ij},\ \bar{\Lambda} := \frac{1}{N}\sum_{j=1}^{N}\Lambda_j,\ \bar{\Gamma} := \binom{N}{2}^{-1}\sum_{1\leq j<j'\leq N}\Gamma_{jj'},\ \bar{\Gamma}_j := \frac{1}{N-1}\sum_{j'\neq j}\Gamma_{j\wedge j', j\vee j'}.$$

*Let*

$$S^2_{\mathrm{proj}} := \frac{1}{N-1}\sum_{j=1}^{N}\left[\frac{N-1}{N-2}(\bar{\Gamma}_j - \bar{\Gamma}) - (\Lambda_j - \bar{\Lambda})\right]^2$$

*and*

$$R^2_{\mathrm{deg}} := \binom{N}{2}^{-1}\sum_{1\leq j<j'\leq N}\left[\Gamma_{jj'} - \bar{\Gamma} - \frac{N-1}{N-2}(\bar{\Gamma}_j - \bar{\Gamma}) - \frac{N-1}{N-2}(\bar{\Gamma}_{j'} - \bar{\Gamma})\right]^2.$$

*Then*

$$\mathbb{V}\mathrm{ar}_L\left(\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}\right) = \frac{4(N-m)}{mN}\,S^2_{\mathrm{proj}} + \frac{2(N-m)(N-m-1)}{m(m-1)(N-2)(N-3)}\,R^2_{\mathrm{deg}}. \tag{12}$$

The theorem shows that the audit-design variance of the MEC estimator separates cleanly into a first-order projection term and a degenerate second-order remainder, exactly mirroring a Hoeffding decomposition for U-statistics. The dominant contribution is the finite-population variance $S^2_{\mathrm{proj}}$ of the projected unit-level correction $\frac{N-1}{N-2}(\bar{\Gamma}_j - \bar{\Gamma}) - (\Lambda_j - \bar{\Lambda})$, which captures how much each audited point changes the corrected $qq$ and $pq$ kernel terms. This leading component scales as $O(m^{-1})$, exactly as in standard finite-population mean estimation. The second component, $R^2_{\mathrm{deg}}$, measures residual pairwise interaction. It is not reducible to unit-level effects because its coefficient is order $m^{-2}$ and $R^2_{\mathrm{deg}}$ is bounded under Assumption 2.1, therefore, it is asymptotically negligible.

**Corollary 3.3.** *Under the conditions of Theorem 3.2 and $\frac{m}{N} \to \rho \in (0, 1]$, we get*

$$\mathbb{V}\mathrm{ar}_L\left(\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}\right) = \frac{4(1-\rho)}{\rho N}\,S^2_{\mathrm{proj}} + \frac{2(1-\rho)^2}{\rho^2 N^2}\,R^2_{\mathrm{deg}} + o(N^{-1}). \tag{13}$$

*In particular, the leading term is typically*

$$\mathbb{V}\mathrm{ar}_L\left(\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}\right) \approx \frac{4(N-m)}{mN}\,S^2_{\mathrm{proj}}.$$

Under Assumption 2.1, $S^2_{\mathrm{proj}} \le 134K^2$ and $R^2_{\deg} \le 4K^2$ for all $N \ge 4$ (Corollary 3.3). Conceptually, the results here imply that the measurement-error problem behaves, to first order, like estimating an average unit-level correction, even though the underlying MMD functional is nonlinear in the latent sample. This makes the corrected procedure much more interpretable: to first order, increasing the audit size reduces variance at the familiar $1/m$ rate, while genuinely nonlinear pairwise effects appear only as a smaller second-order refinement.

### 3.3 Oracle Recovery and Consistency

**Theorem 3.4** (Design consistency and CLT)**.** *Suppose Assumptions 2.2, 2.3, and 2.1 hold. Conditional on the realized finite population $\{(Y_j, \widetilde{Y}_j)\}_{j=1}^N$ and $\{X_i\}_{i=1}^n$. Then*

$$\frac{\widehat{A}_{pq} - A^\star_{pq}}{nN} = O_{\mathbb{P}_L}(m^{-1/2}), \qquad \frac{\widehat{A}_{qq} - A^\star_{qq}}{N(N-1)} = O_{\mathbb{P}_L}(m^{-1/2}). \tag{14}$$

*Moreover, if the corresponding finite-population asymptotic variances converge to finite limits, then the vector*

$$\sqrt{m}\begin{pmatrix} (\widehat{A}_{pq} - A^\star_{pq})/(nN) \\ (\widehat{A}_{qq} - A^\star_{qq})/(N(N-1)) \end{pmatrix}$$

*is asymptotically Gaussian under the audit-design randomness. Consequently,*

$$\sqrt{m}\left(\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} - \widehat{\mathrm{MMD}}^2_\star\right) \Rightarrow \mathcal{N}(0, V_{\mathrm{design}}) \tag{15}$$

*for some finite $V_{\mathrm{design}} \ge 0$ given by the asymptotic variance of the corresponding linear combination.*

Theorem 3.4 is essentially a finite population CLT for HT estimators [63] and shows that, under simple random auditing, the correction error of MEC is asymptotically negligible at the standard $m^{-1/2}$ rate and satisfies a design-based CLT around the oracle empirical MMD. In particular, once the audit size $m$ grows, the dominant uncertainty induced by partial auditing behaves like ordinary sampling noise from the audited subset rather than a more pathological nonlinear error arising from measurement correction. Practically, it implies that inference for the corrected statistic can be based on the same first-order Gaussian approximations that underlie classical survey-sampling and U-statistic theory, with the oracle empirical MMD serving as the finite-sample target and the audit design contributing a transparent additional variance term. A simulation in Appendix D (Table 8) confirms the unbiasedness and the variance formula and shows that the normal approximation is standardized by the true variance.

**Theorem 3.5** (Consistency)**.** *Suppose Assumptions A.1, 2.2, 2.3, and 2.1 hold. Then*

$$\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} \xrightarrow{\mathbb{P}} \mathrm{MMD}^2(p, q; k), \tag{16}$$

*where the probability is with respect to all sources of randomness $(X, Y, \phi, \varepsilon, L)$.*

*More precisely,*

$$\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} - \mathrm{MMD}^2(p, q; k) = O_{\mathbb{P}}(n^{-1/2} + N^{-1/2} + m^{-1/2}). \tag{17}$$

*Under Assumption 2.3, this simplifies to*

$$\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} - \mathrm{MMD}^2(p, q; k) = O_{\mathbb{P}}(N^{-1/2}). \tag{18}$$

The theorem decomposes the estimation error into an audit term of order $m^{-1/2}$ and the usual empirical MMD sampling terms of order $n^{-1/2}$ and $N^{-1/2}$. Under Assumption 2.3, all three are of order $N^{-1/2}$. Thus, if the audit size grows proportionally with the target sample size, partial auditing, while increasing variance, asymptotically does not degrade the first-order convergence rate.

## 4 Experiments

**Experiment 1 – Bias–Variance Trade-off.** We study the bias–variance trade-off of measurement-error correction in a one-dimensional Gaussian example. We generate $X_{1:n} \sim \mathcal{N}(3, 1)$ and $Y_{1:n} \sim \mathcal{N}(0, 1)$ with $n = 500$, but observe only noisy targets $\widetilde{Y}_j = Y_j + \varepsilon_j$, where $\varepsilon_j \sim \mathcal{N}(0.1, 0.4^2)$. Figure 1

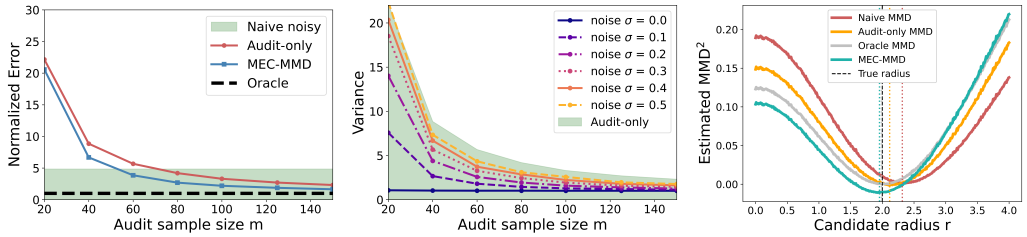



Figure 1: **Left Panel.** Normalized error (MSE divided by the oracle's) versus audit size. The naive estimator is low-variance but biased, the audit-only estimator is unbiased but high-variance, and MEC-MMD removes bias while reducing variance. **Middle Panel.** Effect of measurement noise on MMD estimation variance for several noise levels $\sigma$. The shaded region is the audit-only variance, independent of $\sigma$. **Right Panel.** MEC-MMD closely matches the oracle objective and correctly identifies the true parameter with lower bias and variance. See Supplementary Materials for details.

(left) compares four estimators: the oracle using the full latent sample $Y$, the naive estimator using noisy proxies $\widetilde{Y}$, the audit-only estimator using only $m$ audited latent points, and the proposed MEC-MMD estimator. The naive estimator has low variance but is biased, whereas the audit-only estimator removes this bias at the cost of substantially higher variance because $m \ll n$ (Figure 5). MEC-MMD combines the noisy sample with an audit-based correction, removing bias while retaining much of the efficiency of the full sample. See Supplementary Materials for the experiment details.

**Experiment 2 – Measurement Error Effect.** We study how measurement noise affects the efficiency of MEC-MMD. Figure 1 (middle) shows the normalized variance of MEC-MMD versus audit size $m$ across several noise levels. Variance decreases with $m$, consistent with the $m^{-1}$ scaling in Corollary 3.3, but increases with the noise magnitude because larger corrections are required. Increasing measurement noise raises the variance of MEC-MMD, since the correction must compensate for larger discrepancies between latent and noisy kernel evaluations. For moderate noise ($\sigma \leq 0.5$) MEC-MMD is more efficient than the audit-only estimator, but this advantage shrinks as $\sigma$ grows. Once the noise is comparable to the signal scale ($\sigma \gtrsim 1$ here), the correction terms carry little information about the latent kernel values and mostly add variance, so the audit-only estimator is preferable. As the noise level approaches zero, MEC-MMD converges to the oracle variance. The results provide evidence that MEC-MMD remains effective under moderate measurement error.

**Experiment 3 – Parameter Estimation.** We study parameter estimation in a geometric model where the latent target distribution is a noisy circle with unknown radius $r_\star$. Latent points are drawn from random angles and perturbed radii $R = r_\star + \eta$, with $\eta \sim \mathcal{N}(0, \sigma_r^2)$, but the analyst observes only noisy proxies $\widetilde{Y} = Y + \varepsilon$, except for a small audited subset with exact latent values. To estimate $r_\star$, we generate candidate samples from noisy circles with radius $r$ and evaluate empirical MMD between each candidate model and the observed data. Figure 1 (right) compares the resulting objective curves for four estimators: oracle, naive, audit-only, and MEC-MMD. The naive estimator is biased by measurement error and therefore attains its minimum at a shifted radius. The audit-only estimator removes this bias, but it has high variance because it relies on only a small audit sample. MEC-MMD combines the full noisy sample with an audit-based correction, yielding an objective whose minimum closely tracks the oracle while remaining substantially more stable than audit-only estimation. This illustrates MEC-MMD's potential in likelihood-free model parameter estimation settings [45].

**Experiment 4a – Density Estimation.** We illustrate the performance of our method for density estimation. The true latent distribution resembles a paw of an animal that was generated using a mixture of Gaussians. We consider three classes of heteroskedastic noise models in our experiment: (1) Gaussian, (2) Laplace, and (3) dependent measurement noise. In the case of Gaussian and Laplace measurement noise, the scale parameters are sampled from a uniform distribution, whereas in the case of dependent measurement noise, the Laplace distribution's scale parameters depend on the absolute values of the variables. We fix the sample size ($N = 2000$) and vary the audit sample size $m$ to compare the performance of our method for estimating the density in terms of Integrated Square Error (ISE). All MMD-based methods fit the same model, a 10-component Gaussian mixture model. The model supplies the clean sample $X$ of Section 2 (via reparameterised draws), and the methods differ only in the data side of the loss. Oracle uses the clean $Y$, Naive uses $\widetilde{Y}$, MEC-MMD uses $\widetilde{Y}$ with the audit correction of Eq. (11), and convMMD adds resampled audit residuals to the model draws before comparing with $\widetilde{Y}$. The fitted mixture is the density estimate, and ISE is computed on a

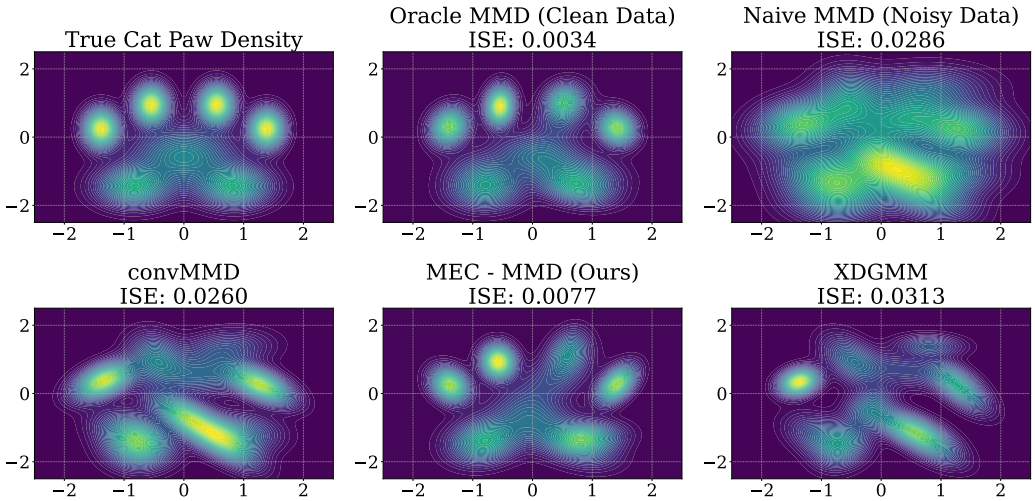


Figure 2: Comparison of Density Estimation in the presence of mean-zero Gaussian noise.

grid against the true density (details in Appendix D). Figure 2 and Table 1 report the experimental results. Figure 2 shows the estimated density from a single run under Gaussian noise with audit sample size $m = 400$, comparing our method against oracle MMD (best-case) [45], convMMD [50], XDGMM [21], and naive MMD trained directly on noisy data without noise correction. Table 1 summarizes performance across 10 runs for varying audit sizes and noise distributions. MEC-MMD improves consistently as the audit sample size increases, progressively closing the gap to oracle performance. Across all three noise regimes, it substantially outperforms both XDGMM and naive MMD. Notably, for larger audit samples ($m \geq 300$), MEC-MMD also surpasses convMMD and achieves the lowest ISE among all non-oracle methods. The gains are especially pronounced under Gaussian and dependent noise, highlighting the method's ability to recover the underlying density accurately even in complex, observation-dependent heteroscedastic settings.

Table 1: Mean and Standard Error of ISE across varying noise distributions and audit sample sizes.

| Noise | $m$ | Oracle | MEC-MMD (Ours) | Naïve MMD [45] | convMMD [50] | XDGMM [21] |
|---|---|---|---|---|---|---|
| Gaussian | 200 | $0.0025 \pm 0.0002$ | $0.0250 \pm 0.0014$ | $0.0292 \pm 0.0004$ | $\mathbf{0.0225 \pm 0.0015}$ | $0.0293 \pm 0.0033$ |
| | 300 | $0.0025 \pm 0.0002$ | $\mathbf{0.0160 \pm 0.0017}$ | $0.0292 \pm 0.0004$ | $0.0226 \pm 0.0016$ | $0.0288 \pm 0.0027$ |
| | 400 | $0.0025 \pm 0.0002$ | $\mathbf{0.0117 \pm 0.0011}$ | $0.0292 \pm 0.0004$ | $0.0228 \pm 0.0016$ | $0.0287 \pm 0.0026$ |
| Laplace | 200 | $0.0032 \pm 0.0002$ | $0.0227 \pm 0.0020$ | $0.0238 \pm 0.0006$ | $\mathbf{0.0150 \pm 0.0009}$ | $0.0387 \pm 0.0061$ |
| | 300 | $0.0032 \pm 0.0002$ | $\mathbf{0.0162 \pm 0.0016}$ | $0.0238 \pm 0.0006$ | $\mathbf{0.0151 \pm 0.0008}$ | $0.0382 \pm 0.0051$ |
| | 400 | $0.0032 \pm 0.0002$ | $\mathbf{0.0104 \pm 0.0010}$ | $0.0238 \pm 0.0006$ | $0.0150 \pm 0.0008$ | $0.0392 \pm 0.0044$ |
| Dependent | 200 | $0.0033 \pm 0.0003$ | $0.0187 \pm 0.0018$ | $\mathbf{0.0173 \pm 0.0004}$ | $0.0188 \pm 0.0013$ | $0.0710 \pm 0.0060$ |
| | 300 | $0.0033 \pm 0.0003$ | $\mathbf{0.0125 \pm 0.0009}$ | $0.0173 \pm 0.0004$ | $0.0182 \pm 0.0010$ | $0.0708 \pm 0.0044$ |
| | 400 | $0.0033 \pm 0.0003$ | $\mathbf{0.0091 \pm 0.0009}$ | $0.0173 \pm 0.0004$ | $0.0182 \pm 0.0011$ | $0.0699 \pm 0.0036$ |

**Experiment 4b – Non-uniform audits.** We repeat Experiment 4 with an audit drawn with unequal inclusion probabilities. Each mixture component $k$ receives a multiplier $c_k$ (log-uniform on $[0.3, 3]$, re-centred to geometric mean one, the same for all seeds); units are audited independently with probability $\pi_j \propto c_{k(j)}$, so the realised $m$ is random with $\mathbb{E}[m] \approx m_{\text{target}}$, and the $\pi_j$ are known. MEC-MMD-IPW uses the Hájek weights $1/\pi_j$ (Section E) and convMMD and XDGMM receive the same audited samples without weights. Table 2 shows that the weighted correction remains effective under the biased design. At $m = 200$ the reweighting adds variance and MEC-MMD-IPW is on par with naive MMD (0.030) and slightly behind XDGMM (0.028) and convMMD (0.024). From $m = 300$ it attains the lowest ISE among all non-oracle methods, and its error keeps decreasing with $m$, narrowing the gap to the oracle (0.004).

**Experiment 5 – Denoising with Empirical Bayes in the Presence of Outliers.** We now demonstrate the application of MEC-MMD in denoising using empirical Bayes in a setting where the noise distribution is unknown to the user, in addition to a small amount of contamination. Concretely, we observe $N = 2000$ noisy measurements $X_i = \theta_i + \varepsilon_i$ drawn from some latent distribution $p_\theta$, but the noise levels $\sigma_i$ are not provided. Instead, we assume access to a small audit sample of $m = 300$ paired observations $\{(\theta_i, X_i)\}_{i=1}^m$ drawn from the same data-generating process; the remaining $N - m$ training points are unpaired noisy observations only. Our goal is to learn a density $q_{\widehat{\theta}_N} \approx p_\theta$ and

Table 2: ISE under a non-uniform (stratified) audit with known inclusion probabilities (Gaussian noise, $N = 2000$; mean $\pm$ standard error over 10 seeds). MEC-MMD-IPW uses inverse-probability (Hájek) weights; bold marks the best non-oracle method.

| $m$ | Oracle | MEC-MMD-IPW (Ours) | Naïve MMD | convMMD | XDGMM |
|---|---|---|---|---|---|
| 200 | $0.0038 \pm 0.0014$ | $0.0304 \pm 0.0053$ | $0.0298 \pm 0.0009$ | $\mathbf{0.0241 \pm 0.0019}$ | $0.0282 \pm 0.0034$ |
| 300 | $0.0038 \pm 0.0014$ | $\mathbf{0.0203 \pm 0.0026}$ | $0.0298 \pm 0.0009$ | $0.0241 \pm 0.0022$ | $0.0275 \pm 0.0029$ |
| 400 | $0.0038 \pm 0.0014$ | $\mathbf{0.0166 \pm 0.0025}$ | $0.0298 \pm 0.0009$ | $0.0238 \pm 0.0022$ | $0.0273 \pm 0.0029$ |

use it as a data-driven prior to estimate the posterior mean $\mathbb{E}[\theta_i \mid X_i]$ via empirical Bayes. Latent points are corrupted by heteroscedastic Gaussian noise with per-sample standard deviations drawn uniformly, and $3\%$ of the latent distribution is contaminated with outliers to assess robustness. See Appendix D for details and a sensitivity analysis of the kernel bandwidth.

We evaluate on three synthetic datasets (Figure 3). We compare against two standard deconvolution baselines: NPEB [24], a nonparametric empirical Bayes method based on the NPMLE, and XDGMM [21], which fits a Gaussian-mixture prior. Since both baselines require a noise level as input, we supply them with a single global estimate $\widehat{\sigma} = \mathrm{std}(\theta_i - X_i)$ computed from the audit pairs; audit observations themselves are treated as exactly observed $(\sigma = 10^{-6})$. An Oracle baseline is included, which uses the true latent distribution estimated via KDE from a large, clean sample and thus provides an approximate lower bound on denoising error. Results are evaluated on a held-out test set of 2000 points unseen during training, with results averaged across five random seeds.

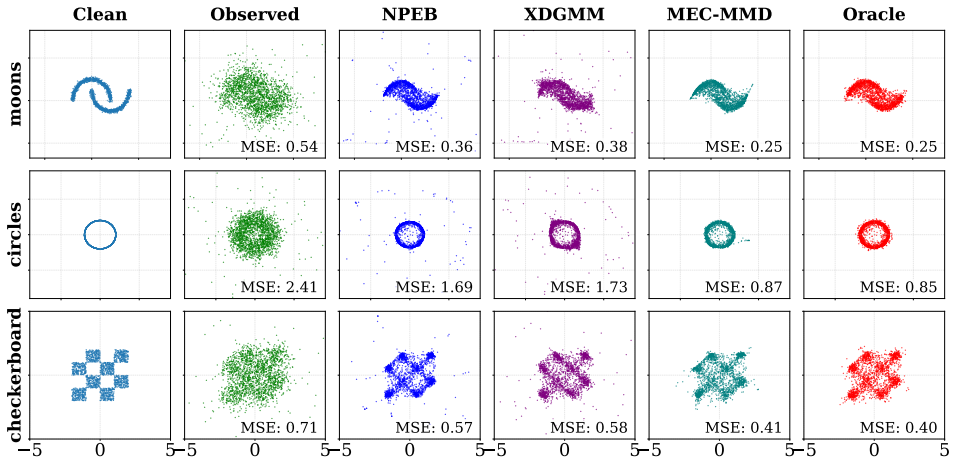



Figure 3: A random run of denoising on synthetic data.

Table 3 reports MSE (between denoised estimates and latent ground truth) and Wasserstein-1 (W1) distance (between each method's learned prior and the test-set latent distribution). Figure 3 shows denoised scatter plots from a representative run. We observed that NPEB places the prior on a discrete support, producing clustered posterior means. XDGMM is smoother but is constrained by its Gaussian-mixture parameterization, leading to oversmoothing and poor recovery of sharp boundaries, particularly for Checkerboard, and the hollow geometry of Circle. Both baselines are further penalized by relying on a scalar estimate of noise rather than the true per-sample heteroscedastic levels. MEC-MMD sidesteps noise estimation entirely during training: the audit correction absorbs the bias introduced by measurement error without requiring its parametric form. It faithfully recovers the continuous geometry of all three manifolds across seeds, achieves the best MSE and W1 among non-oracle methods on every dataset, and remains robust to outlier contamination, consistent with MMD's robustness to small amounts of corrupted mass [46, 64]. On Moons and Checkerboard,

Table 3: Performance comparison on synthetic datasets.

| Dataset | Method | MSE (↓) | W1 Distance (↓) |
|---|---|---|---|
| Moons | NPEB | $0.3394 \pm 0.0139$ | $0.1187 \pm 0.0102$ |
| | XDGMM | $0.3511 \pm 0.0144$ | $0.1600 \pm 0.0094$ |
| | MEC-MMD | $\mathbf{0.2546 \pm 0.0041}$ | $\mathbf{0.0957 \pm 0.0086}$ |
| | Oracle | *0.2524 ± 0.0042* | *0.0398 ± 0.0087* |
| Circles | NPEB | $1.6596 \pm 0.0805$ | $0.3439 \pm 0.0257$ |
| | XDGMM | $1.6609 \pm 0.0668$ | $0.3746 \pm 0.0298$ |
| | MEC-MMD | $\mathbf{0.8929 \pm 0.0269}$ | $\mathbf{0.2008 \pm 0.0261}$ |
| | Oracle | *0.8446 ± 0.0065* | *0.0768 ± 0.0125* |
| Checkerboard | NPEB | $0.5178 \pm 0.0311$ | $0.1866 \pm 0.0066$ |
| | XDGMM | $0.5101 \pm 0.0493$ | $0.1834 \pm 0.0233$ |
| | MEC-MMD | $\mathbf{0.4039 \pm 0.0116}$ | $\mathbf{0.1505 \pm 0.0103}$ |
| | Oracle | *0.3958 ± 0.0057* | *0.0700 ± 0.0044* |

MEC-MMD nearly closes the gap to the Oracle, demonstrating that a modest audit sample of 300 pairs is sufficient to recover near-optimal denoising even when the full noise distribution is withheld.

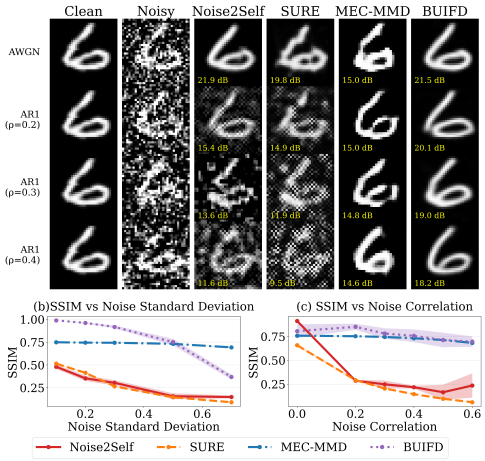



Figure 4: Image denoising performance.

**Experiment 6 – Robustness in Image Denoising.** While the method is not explicitly designed and optimized for image denoising, we demonstrate its applicability for unsupervised image denoising. The observed MNIST dataset contains only corrupted images, $\tilde{x} = x + u$, with known noise law $p(u)$, together with a small audit set of paired clean/noisy examples. We train a generator $G_\theta(z)$ as a clean-image prior by simulating the forward corruption process on generated samples and minimizing MEC-MMD between generated clean images and the observed noisy dataset. The learned prior is then used for empirical Bayes denoising by approximating $\mathbb{E}[x \mid \tilde{x}]$ with importance sampling on the generative manifold. We test on additive white Gaussian noise with $\sigma = 0.5$ and spatially correlated Gaussian noise with the same total variance. Figure 4 compares MEC-MMD with Noise2Self [65], BUIFD [66], and SURE [67]. BUIFD and Noise2Self perform better under white noise but degrade as spatial correlation increases. MEC-MMD remains stable across correlation and noise levels, indicating that our method yields more robust denoising in the presence of structured measurement error. See limitations for discussion.

**Experiment 7 – Scaling with dimension.** We investigate the efficacy of our method in high dimensions by considering a deconvolution task on a one-dimensional spiral embedded in $\mathbb{R}^d$, $d \in \{3, 7, 11, 15, 20, 30\}$, with $N = 2000$ noisy observations and heteroscedastic Gaussian noise. All MMD-based methods fit the same normalizing flow with a Laplace kernel. All noise-aware methods share the same audit of $m = 400$ points, and none is given the true noise level. convMMD resamples the audited residuals and XDGMM uses the audited residual variance, so the comparison is like-for-like in the unknown-noise setting. We measure performance by using the sliced Wasserstein distance on held-out latent points multiplied by $\sqrt{d}$ (Table 4; details in Appendix D). Between $d = 3$ and $d = 30$, naive MMD, convMMD and XDGMM degrade from $0.048$ to $0.208$, from $0.031$ to $0.062$ and from $0.030$ to $0.087$, whereas MEC-MMD stays essentially flat ($0.029$ to $0.034$). At $d = 30$, MEC-MMD is about $1.8\times$ and $2.6\times$ better than convMMD and XDGMM, respectively. A plausible explanation is that MEC-MMD never models the noise and it corrects only the scalar kernel expectations that define $\mathrm{MMD}^2$, whereas others must estimate quantities that grow with $d$ (a $d \times d$ noise covariance or the $d$-dimensional residual distribution).

Table 4: Sliced Wasserstein distance $\times\sqrt{d}$ (lower is better) on the tangled-ribbon deconvolution task across dimensions $d$; $N = 2000$, one audit of size $m = 400$ shared by all noise-aware methods. Entries are mean $\pm$ standard error over 5 seeds. No method is given the true noise level.

| Method | $d = 3$ | $d = 7$ | $d = 11$ | $d = 15$ | $d = 20$ | $d = 30$ |
|---|---|---|---|---|---|---|
| Naïve MMD | $0.048 \pm 0.002$ | $0.071 \pm 0.001$ | $0.098 \pm 0.002$ | $0.123 \pm 0.001$ | $0.151 \pm 0.002$ | $0.208 \pm 0.002$ |
| convMMD | $0.031 \pm 0.002$ | $\mathbf{0.031 \pm 0.001}$ | $0.037 \pm 0.003$ | $0.045 \pm 0.002$ | $0.047 \pm 0.001$ | $0.062 \pm 0.001$ |
| XDGMM | $0.030 \pm 0.003$ | $0.033 \pm 0.001$ | $0.043 \pm 0.001$ | $0.053 \pm 0.001$ | $0.061 \pm 0.002$ | $0.087 \pm 0.001$ |
| MEC-MMD (Ours) | $\mathbf{0.029 \pm 0.005}$ | $0.033 \pm 0.006$ | $\mathbf{0.035 \pm 0.005}$ | $\mathbf{0.030 \pm 0.002}$ | $\mathbf{0.034 \pm 0.003}$ | $\mathbf{0.034 \pm 0.006}$ |

## 5 Discussion

**Bias–Variance Trade-off.** Although the proposed correction reduces MMD bias, it is not uniformly optimal because it trades bias reduction for variance inflation. Measurement error induces systematic bias in kernel evaluations and empirical MMD estimates; the audit-based Horvitz–Thompson correction removes this bias using the audited subset. However, because the audit sample is small, the correction adds sampling variability (Theorem 3.2) and can have higher variance than the naive estimator. Thus, when ME bias is small relative to correction variance, the naive biased estimator may achieve lower MSE, as in the left panel of Figure 1 for $m < 40$. Similarly, the audit-only estimator can outperform the combined estimator when the audit sample is sufficiently informative, or the noisy sample has low SNR, as in the middle panel of Figure 1. Thus, the proposed estimator is best viewed as part of an estimator-selection problem rather than a universally optimal choice. The preferred

estimator depends on the relative sizes of ME bias, audit-sample variance, and correction-induced variance. In practice, cross-validation or sample splitting can compare the naive, audit-only, and corrected estimators and select the one with the lowest predictive or resampling-based risk.

**Non-uniform Audit Sample.** We studied settings where an *audit* or *validation* sample can be obtained through simple random sampling. However, in many practical scenarios, audit samples are collected through non-random mechanisms, such as stratified, outcome-dependent, or uncertainty-guided validation designs [38, 68]. Section E and Appendix C extend MEC-MMD with inverse-probability weights, and Experiment 4b shows that, when the inclusion probabilities are known, MEC-MMD performs well. While this extension preserves the basic kernel-level correction principle, it introduces additional methodological challenges, including estimating audit inclusion probabilities, controlling variance inflation from large weights, and developing stable or doubly robust alternatives. We therefore view non-uniform audit sampling as an important direction for future work.

**Beyond Additive ME.** Although the main formulation and experimental results are presented under an additive ME model, the audit-based correction is not intrinsically tied to additivity. The key requirement is that, for audited units, both the proxy observation $\widetilde{Y}_j$ and the corresponding latent value $Y_j$ are observed, such that $\Gamma_{jj'} = k(Y_j, Y_{j'}) - k(\widetilde{Y}_j, \widetilde{Y}_{j'})$ and $\Lambda_{ij} = k(X_i, Y_j) - k(X_i, \widetilde{Y}_j)$ can be computed. Thus, the same correction principle can in principle be applied to multiplicative noise, rounding, nonlinear sensor distortion, or other proxy mechanisms. However, this statistical generality should be distinguished from formal identifiability. The audit sample must be representative, or else properly weighted; each proxy observation should correspond to a well-defined latent unit; and identifiability of the latent target distribution under the measurement mechanism must be established separately. In settings involving many-to-one coarsening, censoring or truncation, selection-biased audits, or extremely low signal-to-noise ratios, MEC-MMD may remain well-defined but can suffer from high variance or may fail to recover the desired latent target without additional assumptions.

**Limitations and Open Challenges.** First, the proposed method has not been systematically evaluated how the required audit sample size should scale with dimension and this remains an open question. Second, the current approach does not explicitly exploit structural information about the ME mechanism when such information is available. The method is intentionally designed for a *distribution-free noise setting*, where the analyst may not know the ME distribution. Although this robustness is attractive, it necessarily leaves potentially informative structure unused. In settings where the noise distribution is known, can be consistently estimated, or sampled [e.g., 9, 10, 59, 69–71], methods that model the measurement process directly may be preferable [e.g., 50] and achieve potentially a lower variance and greater statistical efficiency. A further open question concerns kernel choice and bandwidth selection. In particular, there is no clear reason to expect that the optimal kernel bandwidth for the clean sample appearing in the audit-correction term should coincide with the bandwidth used elsewhere in the estimator. This suggests that kernel optimization, carried out with appropriate care, may provide an additional avenue for improving performance.

Although we include an image denoising case study, it is not the primary focus of this work. Image denoising is a mature field with specialized benchmarks, methods, and noise models [e.g., see, 72–76]. Our method is designed for general measurement-error correction and does not exploit image-specific structure or common imaging-noise assumptions. Thus, although we showed that MEC-MMD is robust to the noise structure, we do not claim it is optimized for image denoising. We therefore leave image-specific extensions and optimization to future work.

**Contributions.** We formulate MMD-based distribution matching under unknown ME with abundant noisy data and a small audit sample. We propose MEC-MMD, which corrects kernel evaluations to target the latent rather than proxy distribution. We prove design-unbiasedness, derive its variance, establish consistency, and demonstrate empirical gains in estimation, density learning, and denoising. Our results provide a practical foundation for robust distribution matching in noisy settings.

## Acknowledgments and Disclosure of Funding

The authors acknowledge support from the National Science Foundation under Cooperative Agreement No. 2421782 and from the Simons Foundation under Award No. MPS-AI-00010515, through the NSF–Simons AI Institute for Cosmic Origins (CosmicAI, `https://cosmicai.org/`). Additionally, the authors acknowledge the Texas Advanced Computing Center (TACC, `http://www.tacc.utexas.edu`) at The University of Texas at Austin for providing computational resources that have contributed to the research results reported within this paper.

## References


[1] Arya Farahi, Daisuke Nagai, and Yang Chen. Pope: A population-based approach to model the spatial structure of astronomical systems. *The Astronomical Journal*, 161(1):30, 2021.

[2] Herman L Marshall, Yang Chen, Jeremy J Drake, Matteo Guainazzi, Vinay L Kashyap, Xiao-Li Meng, Paul P Plucinsky, et al. Concordance: In-flight calibration of x-ray telescopes without absolute references. *The Astronomical Journal*, 162:254, 2021.

[3] Sarah Shy, Hyungsuk Tak, Eric D Feigelson, John D Timlin, and G Jogesh Babu. Incorporating measurement error in astronomical object classification. *The Astronomical Journal*, 164(1):6, 2022.

[4] Russell G Congalton. Accuracy assessment and validation of remotely sensed and other spatial information. *International journal of wildland fire*, 10(4):321–328, 2001.

[5] Anne CM Thiebaut, Laurence S Freedman, Raymond J Carroll, and Victor Kipnis. Is it necessary to correct for measurement error in nutritional epidemiology? *Annals of Internal Medicine*, 146(1):65–67, 2007.

[6] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11:733–739, 2010.

[7] Kevin McCoy, Zachary Wooten, and Christine B Peterson. Tree-based predictive models for noisy input data. *arXiv preprint arXiv:2603.07409*, 2026.

[8] John Bound and Alan B Krueger. The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? *Journal of Labor Economics*, 9:1–24, 1991.

[9] Nico Thurow. Characterizing measurement error in the german socio-economic panel using linked survey and administrative data. *arXiv preprint arXiv:2501.03015*, 2025.

[10] Amanda K. Glazer, Layla Parast, and Mevin B. Hooten. Beyond the yard line: Accommodating rounded sports data in statistical models. *The American Statistician*, 80(3):443–452, 2026.

[11] Yangdi Jiang, Yi Liu, Xiaodong Yan, Anne-Sophie Charest, Linglong Kong, and Bei Jiang. Analysis of differentially private synthetic data: A measurement error approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21206–21213, 2024.

[12] Deaglan J Bartlett and Harry Desmond. Marginalised normal regression: Unbiased curve fitting in the presence of x-errors. *The Open Journal of Astrophysics*, 6:42, 2023.

[13] Jarno Lintusaari, Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Fundamentals and recent developments in approximate bayesian computation. *Systematic biology*, 66(1):e66–e82, 2017.

[14] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

[15] Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

[16] John P Buonaccorsi. *Measurement error: Models, methods, and applications*. Chapman and Hall/CRC, 2010.

[17] Peter J Huber and Elvezio M Ronchetti. *Robust Statistics*. John Wiley & Sons, 2011.

[18] David Ruppert. Robust statistics: The approach based on influence functions, 1987.

[19] Arya Farahi. Two-point correlation function estimation with contaminated data. *Physical Review D*, 113(12):123522, 2026.

[20] Nevena Gligić and Arya Farahi. Robust-by-design distributional learning from contaminated samples. In *The Fortieth Annual Conference on Neural Information Processing Systems*, 2026.

[21] Jo Bovy, David W Hogg, and Sam T Roweis. Extreme deconvolution: Inferring complete distribution functions from noisy, heterogeneous and incomplete observations. *Annals of Applied Statistics*, 5(2):1657–1677, 2011.

[22] Abhra Sarkar, Bani K Mallick, John Staudenmayer, Debdeep Pati, and Raymond J Carroll. Bayesian semiparametric density deconvolution in the presence of conditionally heteroscedastic measurement errors. *Journal of Computational and Graphical Statistics*, 23(4):1101–1125, 2014.

[23] Peter Melchior and Andy D Goulding. Filling the gaps: Gaussian mixture models from noisy, truncated or incomplete samples. *Astronomy and computing*, 25:183–194, 2018.

[24] Jake A Soloff, Adityanand Guntuboyina, and Bodhisattva Sen. Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(1):1–32, 2025.

[25] Wayne A Fuller. *Measurement error models*. John Wiley & Sons, 2009.

[26] John R Cook and Leonard A Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328, 1994.

[27] Raymond J Carroll, Jeffrey D Maca, and David Ruppert. Nonparametric regression in the presence of measurement error. *Biometrika*, 86(3):541–554, 1999.

[28] John Staudenmayer and David Ruppert. Local polynomial regression and simulation–extrapolation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1):17–30, 2004.

[29] Brandon C Kelly. Some aspects of measurement error in linear regression of astronomical data. *The Astrophysical Journal*, 665(2):1489–1506, 2007.

[30] Abhra Sarkar, Bani K Mallick, and Raymond J Carroll. Bayesian semiparametric regression in the presence of conditionally heteroscedastic measurement and regression errors. *Biometrics*, 70(4):823–834, 2014.

[31] Mauro Sereno. A bayesian approach to linear regression in astronomy. *Monthly Notices of the Royal Astronomical Society*, 455(2):2149–2162, 2016.

[32] Mark Cartwright, Graham Dove, Ana Elisa Méndez Méndez, Juan P Bello, and Oded Nov. Crowdsourcing multi-label audio annotation tasks with citizen scientists. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–11, 2019.

[33] Michael A Hedderich, Dawei Zhu, and Dietrich Klakow. Analysing the noise model error for realistic noisy label data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7675–7684, 2021.

[34] Tetsuya Sakai, Sijie Tao, Nuo Chen, Yujing Li, Maria Maistro, Zhumin Chu, and Nicola Ferro. On the ordering of pooled web pages, gold assessments, and bronze assessments. *ACM Transactions on Information Systems*, 42(1):1–31, 2023.

[35] Tijana Zrnic and Emmanuel J Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024.

[36] Enshuo Hsu and Kirk Roberts. Leveraging large language models for knowledge-free weak supervision in clinical natural language processing. *Scientific reports*, 15(1):8241, 2025.

[37] Jenny Shen, Dane Isenberg, Kristin A Linn, and Rebecca A Hubbard. Integrating misclassified ehr outcomes with validated outcomes from a non-probability sample. *Statistics in Medicine*, 44(15-17):e70127, 2025.

[38] Gustavo Amorim, Ran Tao, Sarah Lotspeich, Pamela A Shaw, Thomas Lumley, and Bryan E Shepherd. Two-phase sampling designs for data validation in settings with covariate measurement error and continuous outcome. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(4):1368–1389, 2021.

[39] Lillian A Boe, Pamela A Shaw, Douglas Midthune, Paul Gustafson, Victor Kipnis, Eunyoung Park, Daniela Sotres-Alvarez, Laurence Freedman, on behalf of the Measurement Error of the STRATOS Initiative, and Misclassification Topic Group. Issues in implementing regression calibration analyses. *American journal of epidemiology*, 192(8):1406–1414, 2023.

[40] Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

[41] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

[42] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

[43] Mijung Park, Wittawat Jitkrittum, and Dino Sejdinovic. K2-abc: Approximate bayesian computation with kernel embeddings. In *Artificial intelligence and statistics*, pages 398–407. PMLR, 2016.

[44] Jovana Mitrovic, Dino Sejdinovic, and Yee-Whye Teh. Dr-abc: Approximate bayesian computation with kernel-based distribution regression. In *International Conference on Machine Learning*, pages 1482–1491. PMLR, 2016.

[45] Francois-Xavier Briol, Alessandro Barp, Andrew B Duncan, and Mark Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv preprint arXiv:1906.05944*, 2019.

[46] Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Mmd-bayes: Robust bayesian estimation via maximum mean discrepancy. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–21. PMLR, 2020.

[47] Ayush Bharti, Masha Naslidnyk, Oscar Key, Samuel Kaski, and François-Xavier Briol. Optimally-weighted estimators of the maximum mean discrepancy for likelihood-free inference. In *International Conference on Machine Learning*, pages 2289–2312. PMLR, 2023.

[48] Patrik Róbert Gerber, Tianze Jiang, Yury Polyanskiy, and Rui Sun. Kernel-based tests for likelihood-free hypothesis testing. *Advances in Neural Information Processing Systems*, 36:15680–15715, 2023.

[49] Lorenzo Pacchiardi, Sherman Khoo, and Ritabrata Dutta. Generalized bayesian likelihood-free inference. *Electronic Journal of Statistics*, 18(2):3628–3686, 2024.

[50] Ritwik Vashistha, Jeff M Phillips, Abhra Sarkar, and Arya Farahi. Simulation-based inference with convolutional maximum mean discrepancy for noisy data. *arXiv preprint arXiv:2604.12022*, 2026.

[51] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[52] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions:: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[53] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

[54] Vernon M Chinchilli, Juliann K Martel, Shiriki Kumanyika, and Tom Lloyd. A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics*, pages 341–353, 1996.

[55] Emil Kupek. Bias and heteroscedastic memory error in self-reported health behavior: an investigation using covariance structure analysis. *BMC Medical Research Methodology*, 2(1):14, 2002.

[56] Donna Spiegelman, Roger Logan, and Douglas Grove. Regression calibration with heteroscedastic error variance. *The international journal of biostatistics*, 7(1):4, 2011.

[57] Arthur Lewbel. Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of business & economic statistics*, 30(1):67–80, 2012.

[58] Arya Farahi, Valentina Guglielmo, August E Evrard, Bianca M Poggianti, Christophe Adami, Stefano Ettori, Fabio Gastaldello, Paul A Giles, Ben J Maughan, David Rapetti, et al. The xxl survey-xxiii. the mass scale of xxl clusters from ensemble spectroscopy. *Astronomy & Astrophysics*, 620:A8, 2018.

[59] Arya Farahi, X Chen, AE Evrard, DL Hollowood, Reese Wilkinson, Sunayana Bhargava, Paul Giles, AK Romer, T Jeltema, M Hilton, et al. Mass variance from archival x-ray properties of dark energy survey year-1 galaxy clusters. *Monthly Notices of the Royal Astronomical Society*, 490(3):3341–3354, 2019.

[60] Sally P Waterworth, Catherine J Kerr, Christopher J McManus, Rianne Costello, and Gavin RH Sandercock. Obese individuals do not underreport dietary intake to a greater extent than nonobese individuals when data are allometrically-scaled. *American Journal of Human Biology*, 34(7):e23743, 2022.

[61] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

[62] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[63] Robert J Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, pages 39–48, 1974.

[64] Ron Nafshi and Maggie Makar. Partial identification of the maximum mean discrepancy with mismeasured data. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

[65] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pages 524–533. PMLR, 2019.

[66] Majed El Helou and Sabine Süsstrunk. Blind universal bayesian image denoising with gaussian noise level learning. *IEEE Transactions on Image Processing*, 29:4885–4897, 2020.

[67] Florian Luisier, Thierry Blu, and Michael Unser. A new sure approach to image denoising: Interscale orthonormal wavelet thresholding. *IEEE Transactions on Image Processing*, 16(3):593–606, 2007.

[68] Linda Nab, Maarten Van Smeden, Reneé De Mutsert, Frits R Rosendaal, and Rolf HH Groenwold. Sampling strategies for internal validation samples for exposure measurement–error correction: A study of visceral adipose tissue measures replaced by waist circumference measures. *American Journal of Epidemiology*, 190(9):1935–1947, 2021.

[69] Yang Chen, Xiao-Li Meng, Xufei Wang, David A van Dyk, Herman L Marshall, and Vinay L Kashyap. Calibration concordance for astronomical instruments via multiplicative shrinkage. *Journal of the American Statistical Association*, 114(527):1018–1037, 2019.

[70] Karl Gebhardt, Erin Mentuch Cooper, Robin Ciardullo, Viviana Acquaviva, Ralf Bender, William P Bowman, Barbara G Castanheira, Gavin Dalton, Dustin Davis, Roelof S de Jong, et al. The hobby–eberly telescope dark energy experiment (hetdex) survey design, reductions, and detections. *The Astrophysical Journal*, 923(2):217, 2021.

[71] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2758–2767, 2020.

[72] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007.

[73] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A review of image denoising algorithms, with a new one. *Multiscale modeling & simulation*, 4(2):490–530, 2005.

[74] Qiang Zhang, Yaming Zheng, Qiangqiang Yuan, Meiping Song, Haoyang Yu, and Yi Xiao. Hyperspectral image denoising: From model-driven, data-driven, to model-data-driven. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):13143–13163, 2023.

[75] Michael Elad, Bahjat Kawar, and Gregory Vaksman. Image denoising: The deep learning revolution and beyond—a survey paper. *SIAM Journal on Imaging Sciences*, 16(3):1594–1654, 2023.

[76] Saeed Izadi, Darren Sutton, and Ghassan Hamarneh. Image denoising in the deep learning era. *Artificial Intelligence Review*, 56(7):5929–5974, 2023.

[77] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727. PMLR, 2015.

[78] Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *International conference on artificial intelligence and statistics*, pages 343–351. PMLR, 2021.

[79] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. *Advances in Neural Information Processing Systems*, 30:1–11, 2017.

[80] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning*, pages 2208–2217. PMLR, 2017.

[81] Charita Dellaporta, Jeremias Knoblauch, Theodoros Damoulas, and François-Xavier Briol. Robust bayesian inference for simulator-based models via the mmd posterior bootstrap. In *International Conference on Artificial Intelligence and Statistics*, pages 943–970. PMLR, 2022.

[82] Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272, 1991.

[83] Jyotishka Datta and Nicholas G Polson. Prediction-powered inference with inverse probability weighting. *arXiv preprint arXiv:2508.10149*, 2025.

[84] Yilin Song, Dan M. Kluger, Harsh Parikh, and Tian Gu. Demystifying Prediction-Powered Inference. *Harvard Data Science Review*, 8(4), oct 8 2026. https://hdsr.mitpress.mit.edu/pub/dwrc1xcp.

[85] Dan M Kluger and Stephen Bates. M-estimation under two-phase multiwave sampling with applications to prediction-powered inference. *arXiv preprint arXiv:2602.16933*, 2026.

[86] Antonin Schrab. A practical introduction to kernel discrepancies: Mmd, hsic & ksd. *arXiv preprint arXiv:2503.04820*, 2025.

[87] Felix Biggs, Antonin Schrab, and Arthur Gretton. Mmd-fuse: Learning and combining kernels for two-sample testing without data splitting. *Advances in Neural Information Processing Systems*, 36:75151–75188, 2023.

[88] Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023.

[89] Dan M Kluger, Kerri Lu, Tijana Zrnic, Sherrie Wang, and Stephen Bates. Prediction-powered inference with imputed covariates and nonuniform sampling. *arXiv preprint arXiv:2501.18577*, 2025.

## Positioning

We consider the problem of parameter estimation with an MMD loss when the scientific target is observed only through noisy measurements, together with a small audited subset of clean latent values. This problem sits at the intersection of kernel mean embeddings and MMD-based estimation, simulation-based inference, and classical measurement-error methodology.

**Noisy and Gold-Standard Data Regimes in Scientific Practice.** Modern scientific workflows frequently operate in a regime where two distinct data sources coexist: a large-scale, inexpensive, and error-prone measurement process, and a smaller, high-quality gold-standard measurement. The scalable data source enables broad coverage and statistical power, arising from mechanisms such as automated labeling, crowd-sourcing, proxy variables, passive data collection, or low-cost instrumentation. In contrast, gold-standard data are typically obtained through expert annotation, controlled experimental procedures, or high-precision instruments, and are therefore substantially more costly and time-intensive to acquire.

| Domain | Noisy / Scalable Measurement | Gold-Standard Measurement |
|---|---|---|
| Natural Language Processing | Heuristic or LLM-generated labels | Expert manual annotation |
| Citizen Science (Astronomy/Ecology) | Crowd-sourced labels or reports | Expert-reviewed classifications |
| Computer Vision | Web-scraped labels | Curated human annotations |
| Speech Processing | Automatically transcribed audio (ASR) | Human-transcribed speech |
| Medicine / Epidemiology | Proxy variables (e.g., ICD codes, questionnaires) | Clinical chart review or biomarkers |
| Genomics | High-throughput sequencing (noisy reads) | Sanger sequencing / validated assays |
| Astronomy / Physics | Photometric imaging | Spectroscopic measurements |
| Environmental Monitoring | Low-cost sensors | Regulatory-grade instruments |
| Remote Sensing | Satellite-derived proxies | Ground-truth field measurements |
| Economics / Social Science | Survey self-reports or administrative proxies | Audited records / verified outcomes |
| Recommender Systems | Implicit feedback (clicks, views) | Explicit ratings or curated preferences |
| Education Research | Automated grading / engagement logs | Expert grading or standardized assessments |

Table 5: Scientific domains exhibiting a combination of abundant noisy measurements and sparse gold-standard observations.

As a result of this cost asymmetry, gold-standard measurements are typically available only for a subset of observations, forming what is often referred to as a validation or audit sample. This pattern arises consistently across disciplines (see Table 5 for examples). In machine learning, for example, large datasets are often labeled using heuristics, weak supervision, or automated systems, with a smaller subset manually verified. In the natural sciences, high-throughput but noisy measurement technologies (e.g., remote sensing, imaging, or sequencing) are complemented by precise but lower-throughput methods like spectroscopy, field sampling, or targeted assays. In the social sciences and economics, self-reported or administrative data are frequently validated against audited records or carefully curated ground-truth datasets.

This setting represents a fundamental trade-off between scale and accuracy. Noisy measurements provide breadth but may introduce systematic bias, while gold-standard data provide accuracy but are limited in scope. Consequently, many modern empirical pipelines are designed to integrate these complementary sources, leveraging gold-standard subsets to calibrate, validate, or correct large-scale noisy datasets. The prevalence of this paradigm across fields suggests that it constitutes a general and unifying framework for understanding data quality, experimental design, and inference under imperfect measurement.

**MMD as a Discrepancy for Estimation and SBI.** Kernel mean embeddings provide a flexible nonparametric representation of distributions in RKHSs, and MMD is the induced discrepancy obtained by comparing these embeddings [52, 41]. Beyond two-sample testing, MMD is now widely used as an estimation objective because it avoids explicit density estimation and does not require tractable likelihoods. This has made it attractive in generative modeling, minimum-distance estimation, and simulation-based inference [77, 45, 14, 78]. MMD has been widely used in hypothesis testing [41], generative modeling [79], domain adaptation [80], and simulation-based inference [45, 47]. In SBI specifically, kernel and MMD-based discrepancies have been used in approximate Bayesian computation, robust generalized Bayesian procedures, and optimally weighted likelihood-free estimators [43, 81, 47]. However, this literature typically assumes that the observed sample is itself the inferential target. In our setting, the observed sample is a noisy proxy for a latent target

distribution, so the naive empirical MMD loss is systematically biased toward the measurement process.

**Measurement Error and Latent-variable Correction.** Classical measurement-error methods address this mismatch through deconvolution, errors-in-variables regression, SIMEX, and likelihood-based latent-variable modeling [82, 26, 15, 29]. These approaches provide powerful tools for density estimation and regression under noisy observation, but they often rely on strong assumptions on the noise law, parametric structure, replicate measurements, or identifiability conditions, and many become unstable in high-dimensional or weakly supervised settings. More importantly for our purposes, they are usually designed to recover the latent distribution or regression function directly, rather than to correct a distribution-matching loss used inside a simulator-based estimation pipeline. Our approach is different: rather than solving the full deconvolution problem, we correct the MMD objective itself so that optimization remains targeted at the latent scientific quantity.

**Audit-based Correction.** A growing line of work, such a prediction-powered inference paradigm [40, 35, 83–85], studies how to combine a large sample of inexpensive proxy measurements with a smaller set of gold-standard observations to enable valid inference. These approaches exploit a simple but powerful principle: the proxy data provide statistical efficiency and coverage, while the audited sample supplies a correction signal that mitigates bias induced by measurement error or model misspecification. Importantly, they avoid requiring the proxy to be fully accurate, instead using the gold-standard subset to adjust downstream estimators in a data-driven manner.

Existing methods in this framework focus primarily on low-dimensional targets such as means, quantiles, or regression parameters, where corrections can be expressed through bias-adjusted estimating equations or reweighting schemes. In contrast, our setting involves a distributional objective (MMD) defined through expectations of kernel features. This introduces additional challenges, as the discrepancy is nonlinear in the data-generating distribution and is evaluated implicitly through pairwise comparisons.

Our approach extends the audit-based paradigm to this setting by constructing corrected estimators of the kernel expectations that define MMD. Rather than modifying the downstream estimator after the fact, we directly adjust the loss function used for estimation. This preserves the computational and modeling advantages of MMD-based procedures while ensuring that optimization remains targeted at the latent distribution of scientific interest, rather than the noisy proxy observations.

**Positioning.** The existing literatures therefore provide only part of the solution. MMD-based SBI supplies a flexible likelihood-free objective but generally assumes clean data; classical measurement-error methods address noisy observations but are not tailored to kernel distribution-matching losses; and audit-based inference provides a correction template but has not been developed for MMD-based parameter estimation with latent targets. Our contribution is to combine these strands: we use MMD as the estimation loss, target latent density and regression parameters under measurement error, and exploit a small audited subset to debias the loss without requiring full deconvolution or exact specification of the measurement process.

# A Assumptions

We restate Assumptions 2.1 - 2.3 from the main text here for completness, along with additional assumptions.

**Assumption A.1** (Additive measurement-error generative model)**.** The latent target sample $Y_1, \dots, Y_N$ is drawn i.i.d. from $q$, and for each $j \in \{1, \dots, N\}$,

$$\phi_j \overset{iid}{\sim} \Pi, \qquad \varepsilon_j \mid \phi_j \sim P(\cdot \mid \phi_j), \qquad \widetilde{Y}_j = Y_j + \varepsilon_j.$$

The triples $(Y_j, \phi_j, \varepsilon_j)$ are independent across $j$, and are independent of the clean reference sample $X_1, \dots, X_n \overset{iid}{\sim} p$.

**Assumption A.2** (Simple random auditing)**.** The audit set $L \subset \{1, \dots, N\}$ is a simple random sample without replacement of fixed size $m$ with $2 \le m \le N$, drawn independently of the data-generating process. In particular,

$$\Pr(j \in L) = \frac{m}{N}, \qquad \Pr(j, j' \in L) = \frac{m(m-1)}{N(N-1)} \quad \text{for } j \neq j'.$$

**Assumption A.3** (Asymptotic regime)**.** We consider a sequence of problems with $n, N, m \to \infty$ such that

$$\frac{m}{N} \to \rho \in (0, 1], \qquad \frac{n}{N} \to \gamma \in (0, \infty).$$

**Assumption A.4** (Bounded kernel)**.** There exists $K < \infty$ such that

$$\sup_{x,x' \in \mathcal{X}} |k(x, x')| \leq K \quad \forall x, x' \in \mathcal{X}.$$

### A.1 Notation

Table 6 summarizes the notation used throughout the paper. We group symbols into four categories: (i) probabilistic objects, (ii) RKHS and MMD-related quantities, (iii) sampling and measurement-error components, and (iv) estimator-specific terms. This table is intended as a reference guide; we introduce symbols locally where needed, but maintain consistency with the definitions summarized here.

Table 6: Summary of notation.

| Notation | Type | Status | Description |
|---|---|---|---|
| $\mathcal{X} \subseteq \mathbb{R}^d$ | Set | Constant | Sample space. |
| $d$ | Scalar | Constant | Ambient dimension. |
| $p, q$ | Distribution | Constant | Reference and latent target distributions. |
| $X, X'$ | Vector | Random variable | iid draws from $p$. |
| $Y, Y'$ | Vector | Random variable | iid latent draws from $q$. |
| $x, y$ | Vector | Realization | Realized values of $X, Y$. |
| $\widetilde{Y}$ | Vector | Random variable | Noisy proxy for $Y$. |
| $\varepsilon$ | Vector | Random variable | Additive measurement error. |
| $Z$ | Vector | Random variable | Covariate affecting noise. |
| $\phi$ | Vector/other | Random variable | Unit-specific noise parameter. |
| $P(\cdot \mid \phi)$ | Distribution | Model | Conditional noise distribution. |
| $\Pi$ | Distribution | Model | Distribution over noise parameters. |
| $\mathcal{H}$ | Function space | Constant | RKHS induced by kernel $k$. |
| $f$ | Function | Variable | RKHS test function. |
| $k$ | Function | Constant | Positive definite kernel. |
| $\mathrm{MMD}^2(p, q; k)$ | Scalar | Target | Squared MMD. |
| $n, N, m$ | Scalars | Constants | Sample sizes (clean, target, audit). |
| $i, j, j'$ | Indices | Deterministic | Observation indices. |
| $X_i$ | Vector | Random variable | Clean reference observation. |
| $Y_j$ | Vector | Random variable | Latent target observation. |
| $\widetilde{Y}_j$ | Vector | Random variable | Noisy target proxy. |
| $\varepsilon_j$ | Vector | Random variable | Measurement error. |
| $L$ | Set | Random set | Audited indices. |
| $\mathbb{E}_L, \mathbb{P}_L$ | Operators | Design-based | Expectation/probability over audit design. |
| $A^\star_{pp}$ | Scalar | Oracle statistic | Clean–clean kernel sum. |
| $A^\star_{qq}$ | Scalar | Oracle statistic | Latent–latent kernel sum. |
| $A^\star_{pq}$ | Scalar | Oracle statistic | Clean–latent kernel sum. |
| $\widehat{\mathrm{MMD}}^2_\star$ | Scalar | Oracle estimator | Empirical MMD using latent data. |
| $\Gamma_{jj'}$ | Scalar | Realization | Correction for $qq$ kernel pair. |
| $\Lambda_{ij}$ | Scalar | Realization | Correction for $pq$ kernel pair. |
| $\widehat{A}_{qq}$ | Scalar | Estimator | Corrected latent–latent sum. |
| $\widehat{A}_{pq}$ | Scalar | Estimator | Corrected clean–latent sum. |
| $\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}$ | Scalar | Estimator | Measurement-error-corrected MMD. |
| $\Lambda_j$ | Scalar | Realization | Unit-level $pq$ correction. |
| $\bar{\Lambda}$ | Scalar | Realization | Mean of $\Lambda_j$. |
| $\bar{\Gamma}$ | Scalar | Realization | Mean of $\Gamma_{jj'}$. |
| $\bar{\Gamma}_j$ | Scalar | Realization | Unit-level $qq$ correction. |
| $S^2_{\mathrm{proj}}$ | Scalar | Realization | First-order projection variance. |
| $R^2_{\mathrm{deg}}$ | Scalar | Realization | Second-order remainder. |
| $\rho$ | Scalar | Limit | Audit fraction $m/N$. |
| $V_{\mathrm{design}}$ | Scalar | Limit | Asymptotic design variance. |

## B Proofs

**Lemma B.1.** *Fix any finite population and let $L$ be a simple random sample without replacement of size $m$ from $\{1, \ldots, N\}$. For any real-valued function $g : \{1, \ldots, N\} \to \mathbb{R}$,*

$$\mathbb{E}_L\left[\frac{N}{m}\sum_{j\in L} g(j)\right] = \sum_{j=1}^{N} g(j). \tag{19}$$

*For any real-valued function $h : \{1, \ldots, N\}^2 \to \mathbb{R}$ with $h(j, j) = 0$,*

$$\mathbb{E}_L\left[\frac{N(N-1)}{m(m-1)}\sum_{\substack{j,j'\in L\\ j\neq j'}} h(j, j')\right] = \sum_{\substack{j,j'=1\\ j\neq j'}}^{N} h(j, j'). \tag{20}$$

*Proof of Lemma B.1.* Let

$$I_j := \mathbf{1}\{j \in L\}, \qquad j = 1, \ldots, N,$$

denote the audit inclusion indicator for unit $j$. Under simple random sampling without replacement of size $m$,

$$\mathbb{E}_L[I_j] = \Pr(j \in L) = \frac{m}{N}.$$

Likewise, for distinct $j \neq j'$,

$$\mathbb{E}_L[I_j I_{j'}] = \Pr(j, j' \in L) = \frac{m(m-1)}{N(N-1)}.$$

We prove the two identities separately.

Single-index identity. Write

$$\sum_{j\in L} g(j) = \sum_{j=1}^{N} I_j\, g(j).$$

Therefore, by linearity of expectation,

$$\begin{aligned}
\mathbb{E}_L\left[\frac{N}{m}\sum_{j\in L} g(j)\right] &= \frac{N}{m}\,\mathbb{E}_L\left[\sum_{j=1}^{N} I_j g(j)\right] \\
&= \frac{N}{m}\sum_{j=1}^{N} g(j)\,\mathbb{E}_L[I_j] \\
&= \frac{N}{m}\sum_{j=1}^{N} g(j)\,\frac{m}{N} \\
&= \sum_{j=1}^{N} g(j),
\end{aligned}$$

which proves the claim.

*Pairwise identity.* Assume $m \geq 2$, so that the normalizing factor $m(m-1)$ is well defined. Since $h(j, j) = 0$, only off-diagonal pairs matter, and we may write

$$\sum_{\substack{j,j'\in L\\ j\neq j'}} h(j, j') = \sum_{\substack{j,j'=1\\ j\neq j'}}^{N} I_j I_{j'}\, h(j, j').$$

Taking expectation and using linearity,

$$\mathbb{E}_L\left[\frac{N(N-1)}{m(m-1)}\sum_{\substack{j,j'\in L\\ j\neq j'}} h(j,j')\right] = \frac{N(N-1)}{m(m-1)}\mathbb{E}_L\left[\sum_{\substack{j,j'=1\\ j\neq j'}}^{N} I_j I_{j'}\, h(j,j')\right]$$
$$= \frac{N(N-1)}{m(m-1)}\sum_{\substack{j,j'=1\\ j\neq j'}}^{N} h(j,j')\,\mathbb{E}_L[I_j I_{j'}].$$

For each $j \neq j'$,

$$\mathbb{E}_L[I_j I_{j'}] = \Pr(j, j' \in L) = \frac{m(m-1)}{N(N-1)}.$$

Substituting this into the previous display gives

$$\mathbb{E}_L\left[\frac{N(N-1)}{m(m-1)}\sum_{\substack{j,j'\in L\\ j\neq j'}} h(j,j')\right] = \frac{N(N-1)}{m(m-1)}\sum_{\substack{j,j'=1\\ j\neq j'}}^{N} h(j,j')\,\frac{m(m-1)}{N(N-1)}$$
$$= \sum_{\substack{j,j'=1\\ j\neq j'}}^{N} h(j,j'),$$

which proves the claim. □

*Proof of Theorem 3.1.* For the $pq$ term, write

$$A^\star_{pq} = \sum_{i=1}^{n}\sum_{j=1}^{N} k(X_i, Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{N} k(X_i, \widetilde{Y}_j) + \sum_{j=1}^{N} g(j),$$

where

$$g(j) := \sum_{i=1}^{n}\Big(k(X_i, Y_j) - k(X_i, \widetilde{Y}_j)\Big) = \sum_{i=1}^{n}\Lambda_{ij}.$$

By Lemma B.1,

$$\mathbb{E}_L\left[\frac{N}{m}\sum_{j\in L} g(j)\right] = \sum_{j=1}^{N} g(j),$$

which implies $\mathbb{E}_L[\widehat{A}_{pq}] = A^\star_{pq}$.

For the $qq$ term, write

$$A^\star_{qq} = \sum_{j\neq j'} k(Y_j, Y_{j'}) = \sum_{j\neq j'} k(\widetilde{Y}_j, \widetilde{Y}_{j'}) + \sum_{j\neq j'} h(j,j'),$$

where

$$h(j,j') := k(Y_j, Y_{j'}) - k(\widetilde{Y}_j, \widetilde{Y}_{j'}) = \Gamma_{jj'}, \qquad h(j,j) := 0.$$

Again by Lemma B.1,

$$\mathbb{E}_L\left[\frac{N(N-1)}{m(m-1)}\sum_{\substack{j,j'\in L\\ j\neq j'}} h(j,j')\right] = \sum_{j\neq j'} h(j,j'),$$

so $\mathbb{E}_L[\widehat{A}_{qq}] = A^\star_{qq}$.

Finally, $A^\star_{pp}$, $n(n-1)$, $N(N-1)$, and $nN$ are non-random under the audit design, so taking expectations term-by-term in (11) yields

$$\mathbb{E}_L[\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}] = \widehat{\mathrm{MMD}}^2_\star.$$

□

*Proof of Theorem 3.2.* Starting from (9), (10), and (11),

$$\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} - \widehat{\mathrm{MMD}}^2_{\star} = \left[ \binom{m}{2}^{-1} \sum_{\substack{1 \le j < j' \le N \\ j,j' \in L}} \Gamma_{jj'} - \bar{\Gamma} \right] - 2 \left[ \frac{1}{m} \sum_{j \in L} \Lambda_j - \bar{\Lambda} \right]. \tag{21}$$

Now decompose the finite-population kernel residuals as

$$\Gamma_{jj'} = \bar{\Gamma} + \alpha_j + \alpha_{j'} + \beta_{jj'}, \qquad 1 \le j < j' \le N,$$

where

$$\alpha_j := \frac{N-1}{N-2}(\bar{\Gamma}_j - \bar{\Gamma}), \qquad \beta_{jj'} := \Gamma_{jj'} - \bar{\Gamma} - \alpha_j - \alpha_{j'}.$$

By construction,

$$\sum_{j=1}^{N} \alpha_j = 0, \qquad \sum_{j' \ne j} \beta_{j \wedge j',\, j \vee j'} = 0 \quad \text{for every } j.$$

Substituting this decomposition into (21) gives

$$\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} - \widehat{\mathrm{MMD}}^2_{\star} = \frac{2}{m} \sum_{j \in L} \big[ \alpha_j - (\Lambda_j - \bar{\Lambda}) \big] + \binom{m}{2}^{-1} \sum_{\substack{1 \le j < j' \le N \\ j,j' \in L}} \beta_{jj'}.$$

Define

$$u_j := \alpha_j - (\Lambda_j - \bar{\Lambda}), \qquad V_L := \binom{m}{2}^{-1} \sum_{\substack{1 \le j < j' \le N \\ j,j' \in L}} \beta_{jj'}.$$

Then

$$\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} - \widehat{\mathrm{MMD}}^2_{\star} = \frac{2}{m} \sum_{j \in L} u_j + V_L. \tag{22}$$

Moreover, $\sum_{j=1}^{N} u_j = 0$, and by definition

$$S^2_{\mathrm{proj}} = \frac{1}{N-1} \sum_{j=1}^{N} u_j^2, \qquad R^2_{\mathrm{deg}} = \binom{N}{2}^{-1} \sum_{1 \le j < j' \le N} \beta_{jj'}^2.$$

By Assumption 2.1 and the triangle inequality (each of $\Lambda_{ij}, \Gamma_{jj'}$ is a difference of two kernel values, each bounded by $K$),

$$|\Lambda_{ij}| \le 2K, \qquad |\Gamma_{jj'}| \le 2K \qquad \text{for all } i,\ j \ne j'. \tag{23}$$

The variance of the linear term is the standard finite-population sampling variance:

$$\mathbb{V}\mathrm{ar}_L \left( \frac{1}{m} \sum_{j \in L} u_j \right) = \frac{N-m}{mN} S^2_{\mathrm{proj}}.$$

Hence

$$\mathbb{V}\mathrm{ar}_L \left( \frac{2}{m} \sum_{j \in L} u_j \right) = \frac{4(N-m)}{mN} S^2_{\mathrm{proj}}. \tag{24}$$

For the degenerate term, the row-sum condition

$$\sum_{j' \ne j} \beta_{j \wedge j',\, j \vee j'} = 0$$

implies the standard finite-population variance formula for a degenerate second-order U-statistic:

$$\mathbb{V}\mathrm{ar}_L(V_L) = \frac{2(N-m)(N-m-1)}{m(m-1)(N-2)(N-3)} R^2_{\mathrm{deg}}. \tag{25}$$

The same degeneracy implies orthogonality between the first-order and second-order parts,

$$\mathrm{Cov}_L\left(\frac{2}{m}\sum_{j\in L} u_j,\ V_L\right) = 0. \tag{26}$$

Combining (22), (24), (25), and (26) yields (12). □

*Proof of Corollary 3.3. Boundedness.* By Assumption 2.1 and (23), $|\Gamma_{jj'}|, |\Lambda_{ij}| \le 2K$, hence $|\bar{\Gamma}|, |\bar{\Gamma}_j|, |\Lambda_j|, |\bar{\Lambda}| \le 2K$. Let $\alpha_j, \beta_{jj'}, u_j$ be as in the proof of Theorem 3.2. Since $\frac{N-1}{N-2} \le \frac{3}{2}$ for $N \ge 4$, $|\alpha_j| \le 6K$ and $|u_j| \le |\alpha_j| + |\Lambda_j - \bar{\Lambda}| \le 10K$, so

$$S^2_{\mathrm{proj}} \le \tfrac{N}{N-1}\, 100K^2 \le 134K^2.$$

Moreover $\sum_j \alpha_j = 0$ and $\sum_{j'\neq j} \beta_{j\wedge j', j\vee j'} = 0$ for every $j$, so $\sum_{j<j'} \beta_{jj'}(\bar{\Gamma} + \alpha_j + \alpha_{j'}) = 0$. Hence $\sum_{j<j'} \Gamma^2_{jj'} = \sum_{j<j'} (\bar{\Gamma} + \alpha_j + \alpha_{j'})^2 + \sum_{j<j'} \beta^2_{jj'} \ge \sum_{j<j'} \beta^2_{jj'}$, and

$$R^2_{\mathrm{deg}} \le \binom{N}{2}^{-1} \sum_{j<j'} \Gamma^2_{jj'} \le 4K^2.$$

Thus $S^2_{\mathrm{proj}}$ and $R^2_{\mathrm{deg}}$ are $O(1)$ uniformly in $N$. By Theorem 3.2,

$$\mathbb{V}\mathrm{ar}_L\left(\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}\right) = \frac{4(N-m)}{mN}\, S^2_{\mathrm{proj}} + \frac{2(N-m)(N-m-1)}{m(m-1)(N-2)(N-3)}\, R^2_{\mathrm{deg}}.$$

Since $m/N \to \rho \in (0, 1]$,

$$\frac{N-m}{mN} = \frac{1-m/N}{m} = \frac{1-\rho}{\rho N} + o(N^{-1}),$$

and therefore

$$\frac{4(N-m)}{mN} = \frac{4(1-\rho)}{\rho N} + o(N^{-1}).$$

Similarly,

$$\frac{(N-m)(N-m-1)}{m(m-1)(N-2)(N-3)} = \frac{(1-m/N)^2}{(m/N)^2\, N^2} + o(N^{-2}) = \frac{(1-\rho)^2}{\rho^2 N^2} + o(N^{-2}),$$

so

$$\frac{2(N-m)(N-m-1)}{m(m-1)(N-2)(N-3)} = \frac{2(1-\rho)^2}{\rho^2 N^2} + o(N^{-2}).$$

Substituting these two expansions into the exact variance formula yields

$$\mathbb{V}\mathrm{ar}_L\left(\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}\right) = \left(\frac{4(1-\rho)}{\rho N} + o(N^{-1})\right) S^2_{\mathrm{proj}} + \left(\frac{2(1-\rho)^2}{\rho^2 N^2} + o(N^{-2})\right) R^2_{\mathrm{deg}}.$$

Because $S^2_{\mathrm{proj}} = O(1)$ and $R^2_{\mathrm{deg}} = O(1)$ by the bounds above, the first remainder term is $o(N^{-1})$ and the second remainder term is $o(N^{-2})$, hence in particular $o(N^{-1})$. Therefore

$$\mathbb{V}\mathrm{ar}_L\left(\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}\right) = \frac{4(1-\rho)}{\rho N}\, S^2_{\mathrm{proj}} + \frac{2(1-\rho)^2}{\rho^2 N^2}\, R^2_{\mathrm{deg}} + o(N^{-1}),$$

which proves (13).

The final display follows because the second term is $O(N^{-2})$ by the bounds above, which is of smaller order than the first term whenever $\rho < 1$ and $NS^2_{\mathrm{proj}} \to \infty$ (e.g. $S^2_{\mathrm{proj}}$ bounded away from 0) and therefore, the first-order approximation is

$$\mathbb{V}\mathrm{ar}_L\left(\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}\right) \approx \frac{4(N-m)}{mN}\, S^2_{\mathrm{proj}}.$$ □

*Proof of Theorem 3.4.* We first prove the rate statements.

For the $pq$ term, define

$$g_j := \frac{1}{n}\sum_{i=1}^{n}\Lambda_{ij} = \frac{1}{n}\sum_{i=1}^{n}\Big(k(X_i, Y_j) - k(X_i, \widetilde{Y}_j)\Big), \qquad j = 1, \dots, N.$$

By Equation (10),

$$\widehat{A}_{pq} - A^{\star}_{pq} = N\left(\frac{1}{m}\sum_{j\in L} g_j - \frac{1}{N}\sum_{j=1}^{N} g_j\right) n,$$

hence

$$\frac{\widehat{A}_{pq} - A^{\star}_{pq}}{nN} = \frac{1}{m}\sum_{j\in L} g_j - \frac{1}{N}\sum_{j=1}^{N} g_j. \tag{27}$$

Because Assumption 2.1 gives

$$|\Lambda_{ij}| \le |k(X_i, Y_j)| + |k(X_i, \widetilde{Y}_j)| \le 2K,$$

we have $|g_j| \le 2K$ for every $j$. Under Assumption 2.2, the right-hand side of Equation (27) is the centered sample mean under simple random sampling without replacement from a bounded finite population. Its design variance is therefore

$$\mathbb{V}\mathrm{ar}_L\left(\frac{\widehat{A}_{pq} - A^{\star}_{pq}}{nN}\right) = \frac{N-m}{mN}\, S^2_{g,N},$$

where

$$S^2_{g,N} := \frac{1}{N-1}\sum_{j=1}^{N}(g_j - \bar{g}_N)^2, \qquad \bar{g}_N := \frac{1}{N}\sum_{j=1}^{N} g_j.$$

Since $|g_j| \le 2K$, we have $S^2_{g,N} \le 4\frac{N}{N-1}K^2$, and Assumption 2.3 implies $(N-m)/(mN) = O(m^{-1})$. Hence

$$\mathbb{V}\mathrm{ar}_L\left(\frac{\widehat{A}_{pq} - A^{\star}_{pq}}{nN}\right) = O(m^{-1}),$$

which yields

$$\frac{\widehat{A}_{pq} - A^{\star}_{pq}}{nN} = O_{\mathbb{P}_L}(m^{-1/2}).$$

For the $qq$ term, define for $j < j'$,

$$u_{jj'} := \Gamma_{jj'} = k(Y_j, Y_{j'}) - k(\widetilde{Y}_j, \widetilde{Y}_{j'}).$$

Then by Equation (9),

$$\frac{\widehat{A}_{qq} - A^{\star}_{qq}}{N(N-1)} = \binom{m}{2}^{-1} \sum_{\substack{1\le j<j'\le N\\ j,j'\in L}} u_{jj'} - \binom{N}{2}^{-1} \sum_{1\le j<j'\le N} u_{jj'}.$$

Again Assumption 2.1 implies $|u_{jj'}| \le 2K$. Thus this is the centered sample mean of a bounded finite population of pair values under simple random sampling of units. Equivalently, it is a finite-population U-statistic of order two with bounded kernel. By the exact variance decomposition in Theorem 3.2, its design variance is $O(m^{-1})$, because the leading projection term is $O(m^{-1})$ and the degenerate remainder is $O(m^{-2})$. Therefore

$$\frac{\widehat{A}_{qq} - A^{\star}_{qq}}{N(N-1)} = O_{\mathbb{P}_L}(m^{-1/2}).$$

This proves Equation (14).

We now turn to asymptotic normality. The first component in Equation (27) is a centered sample mean under simple random sampling without replacement, so any standard finite-population central limit theorem for bounded populations applies. For the second component, the Hoeffding decomposition used in Theorem 3.2 shows that

$$\frac{\widehat{A}_{qq} - A^{\star}_{qq}}{N(N-1)} = \frac{2}{m}\sum_{j\in L}\alpha_j + R_{m,N},$$

where $\{\alpha_j\}$ is the first-order projection with bounded second moments and

$$\mathbb{V}\mathrm{ar}_L(R_{m,N}) = O(m^{-2}).$$

Hence

$$\sqrt{m}\, R_{m,N} \to 0 \qquad \text{in } \mathbb{P}_L\text{-probability},$$

so the asymptotic distribution of the $qq$ term is governed by its projected sample-mean part. Therefore both normalized errors reduce asymptotically to linear statistics of the audit sample, and by the multivariate finite-population CLT for simple random sampling, the vector

$$\sqrt{m}\begin{pmatrix} (\widehat{A}_{pq} - A^{\star}_{pq})/(nN) \\ (\widehat{A}_{qq} - A^{\star}_{qq})/(N(N-1)) \end{pmatrix}$$

is asymptotically Gaussian whenever its finite-population covariance matrix converges to a finite limit.

Finally,

$$\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} - \widehat{\mathrm{MMD}}^2_{\star} = \frac{\widehat{A}_{qq} - A^{\star}_{qq}}{N(N-1)} - 2\,\frac{\widehat{A}_{pq} - A^{\star}_{pq}}{nN}.$$

Thus Equation (15) follows immediately from the joint Gaussian convergence and the continuous mapping theorem, with $V_{\mathrm{design}}$ equal to the asymptotic variance of this linear combination. □

*Proof of Theorem 3.5.* Decompose

$$\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} - \mathrm{MMD}^2(p,q;k) = \underbrace{\left(\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} - \widehat{\mathrm{MMD}}^2_{\star}\right)}_{\text{(I)}} + \underbrace{\left(\widehat{\mathrm{MMD}}^2_{\star} - \mathrm{MMD}^2(p,q;k)\right)}_{\text{(II)}}. \quad (28)$$

We first control term (I). By Theorem 3.4,

$$\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} - \widehat{\mathrm{MMD}}^2_{\star} = O_{\mathbb{P}_L}(m^{-1/2})$$

conditional on the realized finite population $\{(Y_j, \widetilde{Y}_j)\}_{j=1}^N$ and $\{X_i\}_{i=1}^n$. Since the conditional $O_{\mathbb{P}_L}(m^{-1/2})$ bound is uniform under Assumption 2.1, it follows by iterated expectation that

$$\text{(I)} = O_{\mathbb{P}}(m^{-1/2}). \quad (29)$$

Now consider term (II). By Assumption A.1, the oracle statistic $\widehat{\mathrm{MMD}}^2_{\star}$ is exactly the standard unbiased empirical MMD estimator computed from the clean samples

$$X_1,\dots,X_n \overset{iid}{\sim} p, \qquad Y_1,\dots,Y_N \overset{iid}{\sim} q.$$

Write

$$\widehat{\mathrm{MMD}}^2_{\star} = \frac{A^{\star}_{pp}}{n(n-1)} + \frac{A^{\star}_{qq}}{N(N-1)} - 2\,\frac{A^{\star}_{pq}}{nN}.$$

Under Assumption 2.1, each kernel term is bounded.

First, $A^{\star}_{pp}/[n(n-1)]$ is a bounded U-statistic of order two based on $\{X_i\}_{i=1}^n$, so

$$\frac{A^{\star}_{pp}}{n(n-1)} - \mathbb{E}[k(X,X')] = O_{\mathbb{P}}(n^{-1/2}).$$

Second, $A^{\star}_{qq}/[N(N-1)]$ is a bounded U-statistic of order two based on $\{Y_j\}_{j=1}^{N}$, so

$$\frac{A^{\star}_{qq}}{N(N-1)} - \mathbb{E}[k(Y, Y')] = O_{\mathbb{P}}(N^{-1/2}).$$

Third, $A^{\star}_{pq}/(nN)$ is a bounded two-sample V-statistic average of iid terms $k(X_i, Y_j)$, hence

$$\frac{A^{\star}_{pq}}{nN} - \mathbb{E}[k(X, Y)] = O_{\mathbb{P}}(n^{-1/2} + N^{-1/2}).$$

Combining these three displays yields

$$\text{(II)} = O_{\mathbb{P}}(n^{-1/2} + N^{-1/2}). \tag{30}$$

In particular, (II) $\xrightarrow{\mathbb{P}} 0$, which is the usual consistency of the clean-sample empirical MMD (see [41]).

Finally, combining (29), (30), and (28),

$$\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} - \mathrm{MMD}^2(p, q; k) = O_{\mathbb{P}}(n^{-1/2} + N^{-1/2} + m^{-1/2}),$$

which proves (17). Since $n/N \to \gamma \in (0, \infty)$ and $m/N \to \rho \in (0, 1]$ by Assumption 2.3, the three rates are all of order $N^{-1/2}$, and therefore

$$\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} - \mathrm{MMD}^2(p, q; k) = O_{\mathbb{P}}(N^{-1/2}).$$

This implies (16). □

# C Non-uniform and Biased Audit Sampling

The present analysis assumes that the audit set is drawn uniformly at random, independently of the realized data. Under this simple random audit design,

$$\Pr(j \in L) = \frac{m}{N}, \qquad \Pr(j, j' \in L) = \frac{m(m-1)}{N(N-1)},$$

for all distinct $j \neq j'$. This assumption is useful because it leads to a particularly simple Horvitz–Thompson correction: each audited unit receives the same weight $N/m$, and each audited pair receives the same weight $N(N-1)/\{m(m-1)\}$.

In many scientific workflows, however, audits are not collected uniformly. Validation measurements may be targeted toward observations with high uncertainty, large residuals, unusual covariates, high scientific value, or greater ease of manual verification. Such designs can be statistically efficient, since they concentrate audit effort on informative observations, but they can also introduce selection bias if the audit mechanism is ignored. In particular, the empirical distribution of audited units may no longer represent the full target sample. Consequently, applying the uniform-audit MEC-MMD correction to a non-uniform audit sample may over-correct or under-correct the kernel sums.

A natural extension is to replace the uniform correction factors by inverse-probability weights. Let

$$\pi_j = \Pr(j \in L), \qquad \pi_{jj'} = \Pr(j, j' \in L)$$

denote the first- and second-order inclusion probabilities under the audit design. We assume that these probabilities are known, or can be consistently estimated, and satisfy the positivity conditions

$$\pi_j > 0, \qquad \pi_{jj'} > 0$$

for all units and all distinct pairs whose corrections may be needed. Then the clean–target kernel sum can be corrected by

$$\widehat{A}^{\mathrm{IPW}}_{pq} = \sum_{i=1}^{n}\sum_{j=1}^{N} k(X_i, \widetilde{Y}_j) + \sum_{i=1}^{n}\sum_{j \in L} \frac{1}{\pi_j}\left\{k(X_i, Y_j) - k(X_i, \widetilde{Y}_j)\right\}.$$

Similarly, the target–target kernel sum can be corrected by

$$\widehat{A}_{qq}^{\mathrm{IPW}} = \sum_{j \neq j'} k(\widetilde{Y}_j, \widetilde{Y}_{j'}) + \sum_{\substack{j,j' \in L \\ j \neq j'}} \frac{1}{\pi_{jj'}} \left\{ k(Y_j, Y_{j'}) - k(\widetilde{Y}_j, \widetilde{Y}_{j'}) \right\}.$$

The resulting IPW version of MEC-MMD is

$$\widehat{\mathrm{MMD}}^2_{\mathrm{IPW}} = \frac{A_{pp}^{\star}}{n(n-1)} + \frac{\widehat{A}_{qq}^{\mathrm{IPW}}}{N(N-1)} - 2\frac{\widehat{A}_{pq}^{\mathrm{IPW}}}{nN}.$$

The design-unbiasedness argument is the same as in the uniform case. Conditional on the realized finite population,

$$\mathbb{E}_L \left[ \sum_{j \in L} \frac{1}{\pi_j} \left\{ k(X_i, Y_j) - k(X_i, \widetilde{Y}_j) \right\} \right] = \sum_{j=1}^{N} \left\{ k(X_i, Y_j) - k(X_i, \widetilde{Y}_j) \right\},$$

and

$$\mathbb{E}_L \left[ \sum_{\substack{j,j' \in L \\ j \neq j'}} \frac{1}{\pi_{jj'}} \left\{ k(Y_j, Y_{j'}) - k(\widetilde{Y}_j, \widetilde{Y}_{j'}) \right\} \right] = \sum_{j \neq j'} \left\{ k(Y_j, Y_{j'}) - k(\widetilde{Y}_j, \widetilde{Y}_{j'}) \right\}.$$

Therefore,

$$\mathbb{E}_L \left[ \widehat{A}_{pq}^{\mathrm{IPW}} \right] = A_{pq}^{\star}, \qquad \mathbb{E}_L \left[ \widehat{A}_{qq}^{\mathrm{IPW}} \right] = A_{qq}^{\star},$$

and hence

$$\mathbb{E}_L \left[ \widehat{\mathrm{MMD}}^2_{\mathrm{IPW}} \right] = \widehat{\mathrm{MMD}}^2_{\star}.$$

Thus, the simple-random-audit estimator studied in this paper is a special case of an IPW correction with

$$\pi_j = \frac{m}{N}, \qquad \pi_{jj'} = \frac{m(m-1)}{N(N-1)}.$$

**Self-normalised (Hájek) version.** Let $w_j = 1/\pi_j$ and, for independent (Poisson) inclusion, $1/\pi_{jj'} = w_j w_{j'}$. Define $S_1 = \sum_{j \in L} w_j$ and $S_2 = \sum_{j,j' \in L,\, j \neq j'} w_j w_{j'} = S_1^2 - \sum_{j \in L} w_j^2$. The Hájek version replaces the normalisers $N$ and $N(N-1)$ of the audit corrections by $S_1$ and $S_2$:

$$\widehat{\mathrm{MMD}}^2_{\text{IPW-H}} = \frac{A_{pp}^{\star}}{n(n-1)} + \frac{1}{N(N-1)} \sum_{j \neq j'} k(\widetilde{Y}_j, \widetilde{Y}_{j'}) + \frac{1}{S_2} \sum_{\substack{j,j' \in L \\ j \neq j'}} w_j w_{j'}\, \Gamma_{jj'}$$
$$- \frac{2}{nN} \sum_{i=1}^{n} \sum_{j=1}^{N} k(X_i, \widetilde{Y}_j) - \frac{2}{nS_1} \sum_{i=1}^{n} \sum_{j \in L} w_j\, \Lambda_{ij}.$$

With equal weights it equals $\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}$. It is a ratio estimator, so it is not exactly unbiased in finite samples (exact unbiasedness is for the plain IPW version above), and we do not analyse it theoretically. This extension would broaden MEC-MMD to two-phase sampling, stratified validation, active auditing, and outcome-dependent audit designs. For example, one could oversample units with large estimated measurement uncertainty, rare covariate values, or high leverage in the MMD objective, while still preserving design-unbiasedness by weighting each audited correction by its inverse inclusion probability. Such designs may be more efficient than uniform auditing when audit resources are limited.

The main challenge is variance. If some inclusion probabilities are small, the corresponding inverse-probability weights can be large, leading to unstable corrections. This issue is especially important for the $qq$ term, since it depends on second-order inclusion probabilities $\pi_{jj'}$. Another open question is how to estimate the inclusion probabilities when the audit design is not fully known, for example, when validation data are collected opportunistically or through expert-driven selection.

Developing stable IPW, stabilized-weight, and doubly robust versions of MEC-MMD under non-uniform audit sampling is therefore an important direction for future work. Such extensions would preserve the central idea of MEC-MMD, i.e., correcting MMD at the kernel level using audited latent values, while allowing the audit mechanism to reflect realistic scientific data-collection workflows.

# D Detail of Experiments

**Computational cost.** Let $n$ be the number of generated samples, $N$ the number of noisy observations and $m$ the audit size. The plug-in terms of $\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}$ need $O(n^2 + N^2 + nN)$ kernel evaluations, the same as the naive estimator. The audit correction adds $m(m-1)/2$ evaluations of $\Gamma_{jj'}$ (two kernel evaluations each) and $2nm$ evaluations for the $\Lambda_{ij}$, i.e. $O(m^2 + nm)$ with $m \ll N$. Using a minibatch of $B$ noisy observations per step with the full audit correction is unbiased for the same objective and costs $O(n^2 + nB + nm)$.

*Benchmark.* We time one evaluation of the objective and its gradient with respect to $n = 1000$ generated samples in $d = 2$ dimensions (Gaussian kernel, $m = 0.1N$, float32) in a standalone PyTorch implementation on Mac M1 Pro, reporting the median and range of 4 repeated runs of 12 evaluations each after warm-up. "Recompute" evaluates every term at each step (naive MMD and MEC-MMD), "cached" precomputes the terms that do not involve the generated samples and "minibatch" additionally uses $B = 2048$ noisy observations per step. The model's own forward and backward cost is not included and it is identical for naive MMD and MEC-MMD.

*Results.* Table 7 shows that precomputing the terms that do not involve the generated sample makes a MEC-MMD step about $2.7\times$ faster at $N = 5000$ and about $5\times$ faster at $N = 10^4$, and that a minibatch of $B = 2048$ observations is faster by a further factor of about 1.5–2. The extra work of the audit correction ($O(m^2+nm)$ kernel evaluations) is not distinguishable from run-to-run variation. The time of a MEC-MMD step differed from that of naive MMD by $+7\%$ to $+13\%$ at $N = 5000$ and by $-28\%$ to $+24\%$ at $N = 10^4$ when both recompute every term.

Table 7: Time per evaluation of the objective and its gradient (ms; median over 4 runs, with the range below) on one thread of an Apple M1 Pro CPU ($n = 1000$ generated samples, $d = 2$, $m = 0.1N$, float32). Recompute: every term at each step; cached: model-independent terms precomputed; minibatch: $B = 2048$ noisy observations per step.

| | | Recompute | | Cached | Minibatch |
|---|---|---|---|---|---|
| $N$ | $m$ | Naïve | MEC-MMD | MEC-MMD | MEC-MMD |
| 5000 | 500 | 282 [259–294] | 304 [289–320] | 113 [101–159] | 76 [55–106] |
| 10000 | 1000 | 943 [855–1373] | 992 [940–1057] | 190 [152–285] | 89 [77–115] |

**Numerical check of the design-based results.** *Setting.* $X_i \sim \mathcal{N}(3,1)$ ($n = 500$), $Y_j \sim \mathcal{N}(0,1)$ ($N = 500$), $\widetilde{Y}_j = Y_j + \varepsilon_j$ with $\varepsilon_j \sim \mathcal{N}(0.1, 0.4^2)$, and the Gaussian kernel $k(a,b) = \exp\{-(a-b)^2/(2h^2)\}$ with $h = 1$. We generate 40 independent populations. For each population we compute the oracle estimate $\widehat{\mathrm{MMD}}^2_\star$ and the variance $\mathrm{Var}_{\mathrm{theo}}$ of Theorem 3.2 from the realised $\Gamma_{jj'}$ and $\Lambda_{ij}$, and then draw 3000 simple random audits of each size $m \in \{5, 20, 150\}$.

*Statistics.* For each audit we compute $\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}$ and $Z = (\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} - \widehat{\mathrm{MMD}}^2_\star)/\sqrt{\mathrm{Var}_{\mathrm{theo}}}$. Table 8 reports, averaged over the 40 populations, the mean of $\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}$ over audits, its empirical variance over audits, $\mathrm{Var}_{\mathrm{theo}}$, the standard deviation and skewness of $Z$, and the fraction of audits with $|Z| \le 1.96$.

*Estimated variance.* The last column replaces $\mathrm{Var}_{\mathrm{theo}}$ by $\widehat{\mathrm{Var}} = \frac{4(N-m)}{mN}\widehat{S}^2_{\mathrm{proj}}$, the leading term of Theorem 3.2, with $\widehat{S}^2_{\mathrm{proj}} = \frac{1}{m-1}\sum_{j\in L}\hat{u}_j^2$ and $\hat{u}_j = \frac{N-1}{N-2}(\hat{\Gamma}_j - \hat{\Gamma}) - (\Lambda_j - \bar{\Lambda}_L)$, where $\hat{\Gamma}_j =$

$\frac{1}{m-1}\sum_{j'\in L\setminus\{j\}}\Gamma_{jj'}$, $\hat{\Gamma}=\frac{1}{m}\sum_{j\in L}\hat{\Gamma}_j$ and $\bar{\Lambda}_L=\frac{1}{m}\sum_{j\in L}\Lambda_j$. It uses only audited quantities and ignores the degenerate term, which is of order $m^{-2}$.

*Results.* In this setting the mean of $\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}$ matches the oracle at every audit size (Theorem 3.1), the empirical variance matches $\mathrm{Var}_{\mathrm{theo}}$ (Theorem 3.2), and $Z$ has unit standard deviation, negligible skewness and 95% coverage already at $m=5$ (Theorem 3.4). These are the true-variance quantities of the theorems. With the variance estimated from the audit, coverage is [0.84, 0.92, 0.95] at $m=5,20,150$, so very small audits give less reliable intervals.

Table 8: Numerical check of Theorems 3.1, 3.2 and 3.4; 40 populations, 3000 simple random audits each, mean over populations. $Z=(\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}-\widehat{\mathrm{MMD}}^2_{\star})/\sqrt{\mathrm{Var}_{\mathrm{theo}}}$, with $\mathrm{Var}_{\mathrm{theo}}$ the variance of Theorem 3.2 on the realised population; the last column standardises by a variance estimated from the audit only. Oracle $\widehat{\mathrm{MMD}}^2_{\star}=0.894$.

| $m$ | $\mathbb{E}[\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}}]$ | $\mathrm{Var}_{\mathrm{emp}}$ | $\mathrm{Var}_{\mathrm{theo}}$ | $\mathrm{sd}(Z)$ | skew | 95% cov. | 95% cov. (est. var.) |
|---|---|---|---|---|---|---|---|
| 5 | 0.8937 | $1.62\times10^{-2}$ | $1.63\times10^{-2}$ | 1.00 | 0.01 | 0.948 | 0.836 |
| 20 | 0.8941 | $3.29\times10^{-3}$ | $3.27\times10^{-3}$ | 1.00 | 0.00 | 0.949 | 0.924 |
| 150 | 0.8942 | $3.02\times10^{-4}$ | $3.03\times10^{-4}$ | 1.00 | −0.01 | 0.950 | 0.947 |

**Experiment 1.** This experiment is designed to explicitly separate bias and variance in the additive measurement-error setting. We generate a clean reference sample

$$X_1,\dots,X_n\overset{\mathrm{iid}}{\sim}\mathcal{N}(3,1)$$

and a latent target sample

$$Y_1,\dots,Y_N\overset{\mathrm{iid}}{\sim}\mathcal{N}(0,1),$$

with $n=N=500$. The target quantity is the squared population discrepancy $\mathrm{MMD}^2(p,q;k)$, where $k$ is a Gaussian kernel with bandwidth $0.1$. The latent target sample is not directly available to the analyst. Instead, for each $j$,

$$\widetilde{Y}_j=Y_j+\varepsilon_j,\qquad \varepsilon_j\overset{\mathrm{iid}}{\sim}\mathcal{N}(0.1,0.4^2),$$

so the observed target sample is systematically shifted and corrupted by additive noise. We additionally draw an audit subset $L\subset\{1,\dots,N\}$ uniformly without replacement, with

$$m\in\{20,40,60,80,100,125,150\},$$

and reveal the latent values $\{Y_j:j\in L\}$ for those audited points.

For each audit size $m$, we compare four estimators. The *oracle* estimator computes empirical MMD using the full latent sample $Y$, and is unavailable in practice but provides a benchmark corresponding to the smallest variance achievable when all target values are observed cleanly. The *naive noisy* estimator computes empirical MMD directly from $(X,\widetilde{Y})$, ignoring measurement error entirely. Because this estimator uses all $N$ target observations, it can have relatively small variance, but it is biased for $\mathrm{MMD}^2(p,q;k)$ since it targets the distribution of the noisy proxies rather than the latent distribution $q$. The *audit-only* estimator uses only the audited latent target points together with the full clean sample $X$. This removes the measurement-error bias, but its effective target sample size is only $m$, so its variance is substantially larger than that of the oracle benchmark. Finally, the proposed MEC-MMD estimator combines the full noisy sample with an audit-based Horvitz–Thompson correction. Its role is precisely to exploit the low-variance information in the noisy sample while using the audit sample only to remove the induced bias.

We repeat the experiment over $5000$ independent Monte Carlo realizations. For each method and each audit size, we compute the empirical bias and variance of the resulting estimator. The appendix figure reports these two components in separate panels. This decomposition makes the tradeoff transparent. The naive noisy estimator occupies the low-variance but biased extreme. The audit-only estimator is essentially unbiased, but its variance is much larger because it discards almost all target observations. MEC-MMD improves on both in the expected direction: compared with the naive

estimator, it eliminates most or all of the measurement-error bias; compared with the audit-only estimator, it achieves substantially lower variance by leveraging the full noisy sample. Across all audit sizes, the oracle curve provides the natural lower benchmark for variance, while the gap between the naive and corrected methods in the bias panel visualizes the cost of ignoring measurement error.

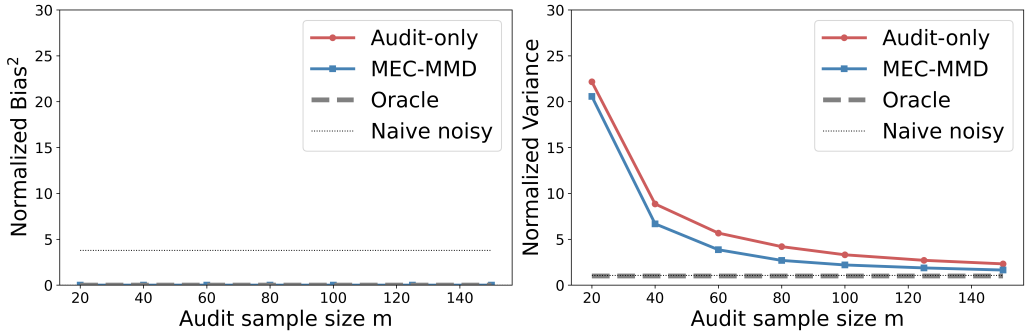



Figure 5: Bias (left panel) and variance (right panel) for the Gaussian measurement-error experiment in Figure 1. The left panel reports empirical bias relative to $\mathrm{MMD}^2(p, q; k)$, and the right panel reports empirical variance. The naïve noisy estimator has the smallest variance among practical methods because it uses all $N$ observed target points, but it is biased because it ignores measurement error. The audit-only estimator is unbiased but has substantially larger variance since its effective target sample size is only $m$. MEC-MMD uses the noisy sample to retain efficiency and the audit sample to remove bias, yielding a markedly better bias–variance tradeoff: much lower variance than audit-only estimation and far smaller bias than the naive noisy estimator. The oracle variance serves as the lower benchmark corresponding to full access to the latent target sample.

**Experiment 2 – Effect of measurement noise.** To study the effect of measurement noise on estimator efficiency, we consider a one-dimensional Gaussian benchmark. We generate a clean reference sample

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(3, 1),$$

and a latent target sample

$$Y_1, \ldots, Y_N \overset{\text{iid}}{\sim} \mathcal{N}(0, 1),$$

with $n = N = 500$. The estimand is the squared population discrepancy $\mathrm{MMD}^2(p, q; k)$, computed using a Gaussian kernel with bandwidth $0.1$.

The analyst does not observe the latent targets $Y_j$ directly. Instead, noisy surrogates are observed,

$$\widetilde{Y}_j = Y_j + \varepsilon_j, \qquad \varepsilon_j \sim \mathcal{N}(\mu_\varepsilon, \sigma_\varepsilon^2).$$

Throughout the experiment we fix $\mu_\varepsilon = 0.1$ and vary the noise standard deviation across $\sigma_\varepsilon \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. An audit subset $L \subset \{1, \ldots, N\}$ is drawn uniformly without replacement with sizes $m \in \{20, 40, 60, 80, 100, 125, 150\}$. For audited points, the latent values $Y_j$ are revealed. The experiment is repeated over $5000$ independent Monte Carlo realizations. For each configuration, we evaluate three practical estimators together with an oracle benchmark:

- **Oracle MMD:** empirical MMD computed using the full latent sample $Y_1, \ldots, Y_N$. This estimator is not available in practice but provides the minimum achievable variance when all targets are observed cleanly.
- **Audit-only:** empirical MMD computed using the clean reference sample $X$ and only the audited latent targets $\{Y_j : j \in L\}$. This estimator is unbiased but its effective target sample size is $m$, leading to substantially higher variance when $m \ll N$.
- **MEC-MMD:** the proposed measurement-error-corrected estimator, which uses the full noisy sample $\widetilde{Y}$ together with audit-based Horvitz–Thompson corrections to recover unbiased kernel sums.

For each estimator we compute empirical mean, bias, and variance across the $5000$ realizations.

The results illustrate how measurement noise affects the efficiency of the correction procedure. As the noise level increases, the variance of MEC-MMD increases because the correction terms become larger and more variable. However, even at the highest noise level considered, MEC-MMD remains substantially more efficient than the audit-only estimator. At the same time, when the noise level approaches zero, MEC-MMD becomes nearly identical to the oracle estimator since the observed proxies coincide with the latent targets. This experiment therefore highlights the key advantage of MEC-MMD: it leverages the large noisy dataset for efficiency while using the audit sample only to remove the bias induced by measurement error.

**Experiment 3 – Parameter estimation under measurement error.** We consider a two-dimensional generative model in which latent target samples lie on a noisy circle with unknown radius $r_\star$. The true latent points are generated as

$$\theta_i \sim \mathrm{Unif}(0, 2\pi), \qquad R_i = r_\star + \eta_i, \qquad \eta_i \sim \mathcal{N}(0, \sigma_r^2),$$
$$Y_i = (R_i \cos\theta_i,\ R_i \sin\theta_i),$$

with $r_\star = 2.0$ and radial noise standard deviation $\sigma_r = 1.0$. This produces a circular ring whose thickness reflects variability in the underlying radius.

The analyst does not observe the latent targets $Y_i$ directly. Instead, noisy proxies are observed,

$$\widetilde{Y}_i = Y_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_2),$$

with observation noise standard deviation $\sigma_\varepsilon = 1.0$. The observed target sample size is $N = 1000$. Additionally, a small audit subset of size $m = 50$ is drawn uniformly without replacement, revealing the latent values for those points.

*Model estimation procedure.* To estimate the unknown radius, we consider a grid of candidate radii $r \in [0, 4]$, discretized using $400$ evenly spaced values. For each candidate radius, we generate a reference sample of size $n_{\text{model}} = 20{,}000$ from the same noisy-circle model but with radius fixed at $r$. We then compute the empirical MMD between the reference sample and the target data using a Gaussian kernel with bandwidth $2.0$. Four estimators are compared:

- **Naïve MMD:** empirical MMD between the candidate model sample and the noisy observations $\widetilde{Y}$.
- **Audit-only MMD:** empirical MMD between the candidate model sample and the audited latent targets.
- **Oracle MMD:** empirical MMD between the candidate model sample and the full latent sample $Y$.
- **MEC-MMD:** the proposed measurement-error-corrected estimator that combines the noisy observations with Horvitz–Thompson corrections computed from the audit sample.

For each estimator, the estimated radius is obtained by minimizing the corresponding empirical MMD curve over the candidate grid.

The right panel of Figure 1 shows the MMD objective as a function of the candidate radius. The oracle curve reaches its minimum near the true radius $r_\star = 2.0$. The naïve estimator produces a shifted minimum due to measurement error in the observations, demonstrating the bias induced by ignoring noise. The audit-only estimator reduces this bias but produces a noisier objective because it relies on only $m = 50$ samples. The MEC-MMD estimator combines the information from the noisy dataset and the audit subset, producing an objective curve that closely matches the oracle curve and correctly identifies the true radius. This experiment highlights how measurement-error correction can enable reliable parameter estimation even when the majority of observations are corrupted.

**Experiment 4 – Density Estimation under measurement error** We consider a two-dimensional density estimation task in which the latent target samples follow a 10-component Gaussian Mixture Model (GMM) designed to resemble a cat paw. The true latent points are generated as

$$Y_i \sim \sum_{k=1}^{10} \pi_k \mathcal{N}(\mu_k, \Sigma_k),$$

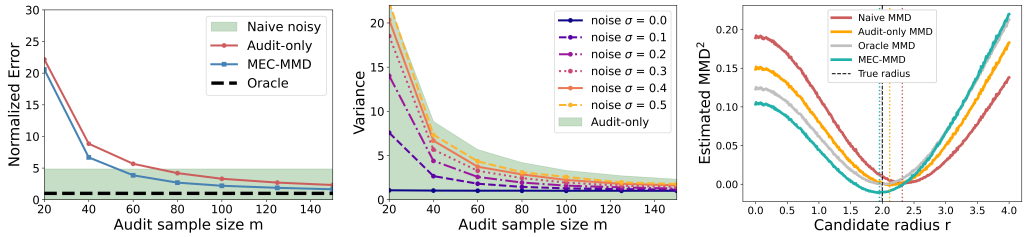



Figure 6: **Left Panel.** Estimation error. The oracle estimator uses the full latent sample and yields a lower-variance bound. The naive estimator based on the full noisy sample has low variance but is biased for $\mathrm{MMD}^2(p, q; k)$. The audit-only estimator removes this bias but has a much larger variance because it uses only $m$ latent target points. MEC-MMD combines noisy samples with audit-based correction, thereby removing bias induced by noisy measurements while achieving substantially lower variance than audit-only estimation. **Middle Panel.** Measurement error effect on estimator variance. The horizontal axis shows the audit sample size $m$, and the vertical axis shows the variance of the estimator. Colored curves correspond to different noise levels $\sigma \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. As measurement noise increases, the variance of the MEC-MMD estimator rises substantially, especially when the audit sample is small, because the correction must be estimated from a limited number of latent target points. For every noise level, increasing $m$ reduces variance, reflecting the improved precision of the audit-based correction. When $\sigma = 0$, the variance equals oracle variance and is flat in $m$, since no measurement-error correction is needed. The results illustrate that measurement noise primarily affects statistical efficiency rather than identifiability: larger noise inflates uncertainty, but this effect can be mitigated by increasing the audit sample size. **Right Panel.** Parameter estimation. The horizontal axis shows candidate radii, and the vertical axis shows the corresponding empirical MMD objective. Vertical lines indicate the minimizing radius for each estimator. The naive estimator (red) is biased because it compares candidate models to noisy observations. The audit-only estimator (orange) removes this bias but has higher variability due to the small audit sample. The proposed MEC-MMD estimator (teal) combines the full noisy dataset with audit-based corrections and closely matches the oracle estimator (gray), correctly identifying the true radius.

where the mixture components represent four elliptical toe pads on an upper arc and three lower palm pads. The mixture weights $\pi_k$ are set to approximately $0.113$ for each toe, $0.151$ for the side pads, and $0.245$ for the central pad, with corresponding geometric means and diagonal covariances. The analyst does not observe the latent targets $Y_i$ directly. Instead, noisy samples are observed,

$$\widetilde{Y}_i = Y_i + \varepsilon_i,$$

We consider three distinct heteroscedastic noise settings. For (1) **Gaussian noise**, $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2 I_2)$ with individual scale parameters sampled as $\sigma_i \sim \mathrm{Unif}(0.2, 0.6)$. For (2) **Laplace noise**, the noise follows a zero-mean Laplace distribution with scale $b_i = \sigma_i/\sqrt{2}$, where again $\sigma_i \sim \mathrm{Unif}(0.2, 0.6)$. For (3) **Dependent noise**, the Laplace scale is spatially dependent on the underlying latent variable, setting the noise standard deviations for each dimension $d \in \{1, 2\}$ to $\sigma_{i,d} = 0.3|Y_{i,d}|$. The sample size is fixed at $N = 2000$. Additionally, an audit subset of size m, varying across 200, 300, and 400, is drawn uniformly without replacement, revealing the true latent values for those points. *Model estimation procedure.* All MMD-based methods fit the same density model, which is a mixture of 10 full-covariance Gaussians in $\mathbb{R}^2$ with trainable weights, means and Cholesky factors. Every method, including the oracle, starts from the same scikit-learn `GaussianMixture` fit (full covariance, 5 restarts) to the noisy sample $\widetilde{Y}$. To estimate the underlying density, we evaluate the methods based on the Integrated Square Error (ISE). For MMD-based methods, the kernel bandwidth $h$ is selected dynamically using a stabilized quantile-based heuristic over pairwise distances. For $q \in \{0.1, \ldots, 0.9\}$, $h_q = \sqrt{Q_q}/(\log_{10}(2n)\sqrt{d})$, where $Q_q$ is the $q$-quantile of the pairwise squared distances above $10^{-5}$ among the training sample ($Y$ for the oracle, $\widetilde{Y}$ otherwise) pooled with $n$ draws from the initial model; $n = 2000$, $d = 2$.

Five estimators are compared:

- **Naive MMD:** empirical MMD trained directly on the noisy observations $\widetilde{Y}$ without accounting for measurement error.

- **Oracle MMD [45]:** empirical MMD trained directly on the full, unobserved clean latent sample $Y$.
- **convMMD [50]:** an estimator that accounts for noise through a resampling scheme, simulating the noise distribution by bootstrapping residuals $(\widetilde{Y} - Y)$ from the audit set.
- **XDGMM [21]:** Extreme Deconvolution GMM, which analytically models measurement error by assuming zero-mean Gaussian noise. We set the noise covariance to the empirical variance of the audit residuals.
- **MEC-MMD (Ours):** the proposed measurement-error-corrected estimator utilizing the available audit samples.

Figure 2 and Table 1 illustrate the performance of these estimators. In Figure 2, visual comparisons of the estimated density under Gaussian noise (with an audit size of $m = 400$) reveal that the naive estimator suffers from severe blurring due to the uncorrected heteroscedastic noise. Table 1 summarizes the ISE across multiple runs and noise distributions. The naive estimator and XDGMM demonstrate significant bias, with XDGMM particularly struggling under the complex, non-Gaussian noise conditions. The convMMD method offers a strong baseline but relies heavily on empirical bootstrapping. Our MEC-MMD estimator effectively combines the full noisy dataset with the audit subset, producing an estimate that closely matches the oracle performance. As the audit sample size increases to $m = 300$ and above, MEC-MMD consistently achieves the lowest ISE among the non-oracle methods across all three noise settings, highlighting its capability to accurately recover the underlying density even under complex, dependent noise. Error bars are computed based on 10 random realizations, mean and standard deviations are reported.

**Experiment 4b – Non-uniform audits.** We repeat Experiment 4 with non-random audit samples. We vary the audit size as $m_{\text{target}} \in \{200, 300, 400\}$.

*Audit design.* Let $K = 10$ be the number of mixture components and $k(j)$ the component that generated unit $j$. Each component receives an audit-intensity multiplier $c_k = \exp(u_k - \bar{u})$ with $u_1, \dots, u_K \overset{\text{iid}}{\sim} \text{Unif}(\log 0.3, \log 3)$ and $\bar{u} = K^{-1} \sum_k u_k$, so the multipliers have geometric mean one. Components with $c_k < 1$ are under-sampled, those with $c_k > 1$ over-sampled, and the ratio of the largest to the smallest multiplier is at most 10. The multipliers are drawn once and kept fixed across seeds. Unit $j$ is audited independently with probability

$$\pi_j = \min\Big\{1,\ \max\Big\{10^{-3},\ \frac{m_{\text{target}}\, c_{k(j)}}{\sum_{i=1}^{N} c_{k(i)}}\Big\}\Big\},$$

so that $\pi_{jj'} = \pi_j \pi_{j'}$ for $j \neq j'$, the realised audit size $m$ is random with $\mathbb{E}[m] \approx m_{\text{target}}$, and the inclusion probabilities are known exactly. The component label acts as an observed stratum.

*Estimators.* MEC-MMD-IPW is the self-normalised (Hájek) estimator of Appendix C with weights $w_j = 1/\pi_j$ and no weight truncation. convMMD bootstraps the audited residuals $\widetilde{Y}_j - Y_j$, and XDGMM uses a noise covariance with the pooled variance of the audited residuals on the diagonal; both use the audited pairs without weights. Naïve MMD and the oracle do not use the audit.

*Protocol.* Each configuration is repeated over 10 seeds; we report the mean $\pm$ standard error of the integrated squared error on an $80 \times 80$ grid over $[-2.5, 2.5]^2$.

**Experiment 5 – Denoising with Empirical Bayes in the Presence of Outliers.** In this experiment, we demonstrate an application of MEC-MMD in empirical Bayes denoising. Given the learned density $q_{\widehat{\theta}_N}$, we use it as a data-driven prior to estimate the posterior mean $\mathbb{E}[\theta_i \mid X_i]$. We evaluate this procedure on three structured two-dimensional datasets (Two Moons, Circle, and Checkerboard), in which latent points are corrupted by heteroscedastic Gaussian noise. To assess robustness, we further contaminate the latent distribution with $3\%$ outliers. We compare our results using a normalizing-flow-based density estimation against two standard deconvolution baselines: NPEB, a nonparametric empirical Bayes method based on the NPMLE [24], and XDGMM, which fits a Gaussian-mixture prior by expectation maximization [21]. An Oracle baseline is also included, which uses the true latent distribution estimated via KDE from a large, clean sample and thus provides an approximate upper bound on denoising performance.

*Data Generating Process and Outlier Corruption.* For each of the three topologies (Two Moons, Circle, Checkerboard), we generate $N = 2000$ latent samples $\theta_i \in \mathbb{R}^2$. To evaluate robustness,

we randomly select $3\%$ of the latent samples and displace them using a large uniform offset (e.g., $\mathcal{U}(-4.0, 4.0)$). This creates a set of "true" outliers that exist far outside the primary manifold.

We generate independent per-sample noise standard deviations $\sigma_i \sim \mathcal{U}(a, b)$. For the Moon and Checkerboard datasets, we use noise bounds $(a, b) = (0.2, 0.6)$. Because the Circle has a larger geometric footprint, we increase the noise bounds to $(a, b) = (0.5, 1.0)$. The final observed data is generated as $X_i = \theta_i + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2 I_2)$.

*MEC-MMD.* We utilize a Masked Autoregressive Flow (MAF) as the latent model. The flow consists of 4 layers, where each layer applies a masked affine autoregressive transform parameterized by a residual network containing 2 blocks and a hidden dimension of 512, using ReLU activations. Each autoregressive layer is followed by a reverse permutation and an LU linear transformation. The base distribution is a standard bivariate Gaussian. The model is trained for 3000 epochs. We use the AdamW optimizer with a learning rate of $10^{-4}$ and a cosine decay schedule. Kernel bandwidths are chosen automatically using the median heuristic [86, 87]. Once trained, we draw $M = 30,000$ samples from the learned prior $q_{\widehat{\theta}_N}$ and compute the posterior mean for each observation using importance sampling:

$$\mathbb{E}[\theta_i \mid X_i] \approx \sum_{m=1}^{M} w_{i,m} \tilde{\theta}_m, \quad \text{where} \quad w_{i,m} \propto \exp\left(-\frac{\|X_i - \tilde{\theta}_m\|^2}{2\sigma_i^2}\right)$$

*NPEB.* We use the Nonparametric Empirical Bayes framework via the ‘npeb‘ package. This method computes the Nonparametric Maximum Likelihood Estimator (NPMLE) using the ‘GLMixture‘ solver. The algorithm places a fine grid over the observational space and optimizes the mixture weights using convex optimization (via CVXPY). We use diagonal precision matrices parameterized by the known heteroscedastic noise variances. The denoised estimates are then extracted using the exact discrete posterior mean (GMLEB).

*XDGMM.* Extreme Deconvolution (XDGMM) fits a continuous Gaussian Mixture Model to data with known heteroscedastic measurement noise. To ensure a fair comparison and avoid bottlenecks, we implemented a custom JAX-accelerated, JIT-compiled version of the XDGMM Expectation-Maximization algorithm. We set the number of Gaussian components to $K = 10$ and regularize the covariances by a factor of $10^{-6}$ to prevent singularities caused by outliers. The model is trained for 100 EM iterations (or until convergence with a tolerance of $10^{-5}$). Once fit, we calculate the exact closed-form Gaussian posterior mean for each observation.

*Evaluation.* Performance is measured using the standard Mean Squared Error (MSE) between the denoised estimates $\widehat{\theta}_i$ and the clean latent truths $\theta_i$ (excluding the outliers):

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} \|\widehat{\theta}_i - \theta_i\|^2$$

The MSE results presented in Figure 3 are representative of the average performance across 5 independent random seeds.

**Kernel Bandwidth Sensitivity Analysis.** In this experiment, we test bandwidth sensitivity for MEC-MMD. We use data from Experiment 5 to compare performance across different bandwidth values. We use a multi-scale Gaussian kernel with bandwidths $\{\sigma_j\}_{j=1}^{9}$ selected via the median heuristic. This provides the user with a principled data-driven default. Here, we verify the sensitivity of the learned prior and downstream denoising performance on bandwidth choice.

*Setup.* We hold the full training procedure fixed and vary only a single multiplicative scale factor $s \in \{0.25, 0.50, 1.00, 3.00, 5.00\}$ applied uniformly to all nine bandwidths, i.e. $\tilde{\sigma}_j = s \cdot \sigma_j^{\text{heuristic}}$. The value $s = 1$ recovers the default heuristic. For each $(s, \text{dataset}, \text{seed})$ triplet, we train a fresh normalizing flow from the same initialization using the scaled bandwidths and evaluate on the held-out test set of 2000 points. Base bandwidths are computed once per seed before training the model. All other hyperparameters (architecture, learning rate, and number of gradient steps) match those of the main experiment. Results are averaged over three seeds, and we report the mean $\pm$ standard deviation. As a bandwidth-independent reference, we include the Oracle baseline (KDE fitted on a large clean sample with no outliers), whose MSE and W1 do not depend on $s$ and are shown as horizontal dotted lines in Figure 7.

*Results.* Figure 7 plots test MSE (left) and prior W1 distance (right) as a function of $s$ for all three datasets, with shaded $\pm 1$ std bands across seeds. We see that performance is remarkably stable for $s \in [0.25, 2.00]$ for MSE. MSE remains within one standard deviation of the $s = 1$ value across this range, indicating that MEC-MMD is not sensitive to moderate variations of the bandwidth around the median heuristic for denoising. However, in case of prior estimation, the performance is more sensitive to bandwidth scale. The performance remains stable for $s \in [0.25, 1.00]$ but as bandwidth scale take extreme values, the performance degrades. These results justify using the median heuristic as a fixed, tuning-free component of the MEC-MMD pipeline.

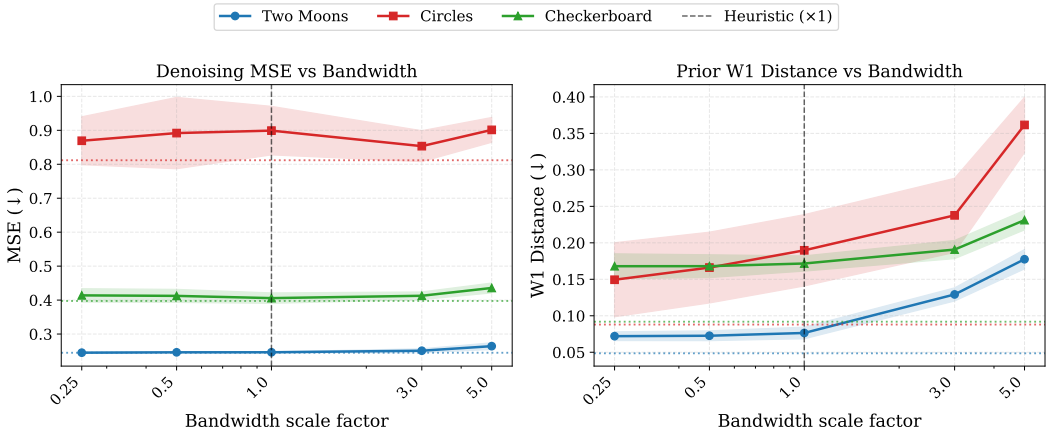



Figure 7: Kernel bandwidth sensitivity analysis.

**Experiment 6 – Image Denoising.** We now evaluate the efficacy of MEC-MMD for denoising noisy images. In this experiment, we observe a dataset of corrupted images $\mathcal{D}_{obs} = \{\tilde{x}_i\}_{i=1}^N$, where $\tilde{x} = x + u$, and $u$ is drawn from a known noise distribution $p(u)$. We also have an audit sample which contains both clean and noisy images $\mathcal{D}_{\text{audit}} = \{x_j, \tilde{x}_j\}_{j=1}^m$. Our framework operates in two stages:

- Prior Learning: We train a generative model $G_\theta(z)$ to map latent variables $z \sim \mathcal{N}(0, I)$ to the clean data manifold. The generator is optimized by minimizing the MEC-MMD between the empirical distribution of the generated noisy samples and the observed noisy dataset.
- Empirical Bayes Inference: Once $G_\theta$ successfully models the clean prior $q_\theta(x)$, we denoise novel test images by estimating the posterior mean $\mathbb{E}[x|\tilde{x}]$ via Importance Sampling (IS) over the learned generative manifold.

We benchmark our methods on the MNIST dataset. We evaluate performance across two distinct noise regimes: (1) Additive White Gaussian Noise (AWGN) with $\sigma = 0.5$ and (2) Spatially Correlated Noise (generated by applying a spatial Gaussian filter to AWGN). The filtered noise is re-normalized to maintain the exact total noise variance ($\sigma = 0.5$), ensuring the noise energy remains constant while the spatial structure varies from independent to highly correlated.

*Network Architecture.* We utilize a standard Convolutional Generator. The latent vector $z \in \mathbb{R}^{128}$ is projected via a fully-connected layer to a $256 \times 7 \times 7$ feature map, followed by three transposed convolutional blocks (kernel size 4, stride 2, padding 1) with SELU activations and BatchNorm. The final layer uses a $3 \times 3$ convolution with a Sigmoid activation to output $28 \times 28$ images.

*Noise.* We model the corruption as $y = x + n$. We utilize a 2D separable Autoregressive AR(1) Gaussian random field. The spatial correlation is controlled by $\rho \in [0, 1)$. The discrete Power Spectral Density (PSD) for an $H \times W$ grid is defined as:

$$S(\omega_x, \omega_y) = \frac{1}{(1 - 2\rho \cos \omega_x + \rho^2)(1 - 2\rho \cos \omega_y + \rho^2)}$$

where $\omega_x = \frac{2\pi i}{H}$ and $\omega_y = \frac{2\pi j}{W}$. To generate a noise sample, we draw standard white noise $w \sim \mathcal{N}(0, \mathbf{I})$, transform it via 2D FFT, scale it by $\sqrt{S(\omega_x, \omega_y)}$, and apply the inverse FFT. Finally, the noise field is globally scaled such that its standard deviation exactly matches $\sigma_{\text{noise}} = 0.5$.

*Kernel Selection.* We utilize an unbiased squared MMD formulation using a mixture of Radial Basis Function (mix-RBF) kernels:

$$k(x, y) = \sum_{s \in \mathcal{S}} \exp\left(-\frac{||x - y||^2}{2\sigma_s^2}\right)$$

In pixel space, the bandwidths were set using the median bandwidth heuristic.

*Empirical Bayes Inference.* Denoising is formulated as finding the posterior mean $\hat{x} = \mathbb{E}_{p(x|\tilde{x}_{obs})}[x]$. Because the true posterior is intractable, we approximate it using Importance Sampling over $K$ candidate samples generated by the trained prior $q_\theta$:

$$\hat{x} \approx \sum_{k=1}^{K} w_k x_{gen}^{(k)}, \quad \text{where} \quad w_k = \frac{p(\tilde{x}_{obs}|x_{gen}^{(k)})}{\sum_j p(\tilde{x}_{obs}|x_{gen}^{(j)})}$$

We draw $K = 30,000$ samples per test image. The likelihood $p(\tilde{x}_{obs}|x_{gen})$ is computed assuming a known Gaussian noise model. To prevent numerical underflow in high dimensions, weights are computed in log-space using the LogSumExp trick before softmax normalization.

*Training Hyperparameters.* All models were trained using the Adam optimizer with a learning rate of $5 \times 10^{-4}$. The DCGAN was trained for 40,000 iterations with a batch size of 1,000. All experiments were executed on an NVIDIA H200 GPU.

**Experiment 7 – Deconvolution across dimensions.** *Data.* We assume that the latent data lies on a one-dimensional spiral (the "tangled ribbon"): $t \sim \text{Unif}(1.5\pi, 4.5\pi)$ and $(x, y) = (t\cos t,\, t\sin t)/(4.5\pi)$. The curve is embedded in $\mathbb{R}^d$ by appending $d - 2$ zero coordinates and applying a random orthogonal matrix drawn for each dataset, with $d \in \{3, 7, 11, 15, 20, 30\}$. Observations are $\widetilde{Y}_j = Y_j + \sigma_j \xi_j$ with $\xi_j \sim \mathcal{N}(0, I_d)$ and $\sigma_j \sim \text{Unif}(0.0848, 0.1272)$ (mean 0.106, $\pm 20\%$). We note that the noise level does not change with $d$, so the total noise power grows with $d$. Each seed draws $N = 2000$ noisy training points and 10,000 held-out latent points used for evaluation.

*Audit and methods.* A simple random audit of $m = 400$ of the $N = 2000$ training points (20%) is shared by all noise-aware methods and no method is given the true noise level. Naïve MMD ignores the measurement error and uses no audit. convMMD bootstraps the audited residual vectors $\widetilde{Y}_j - Y_j$. XDGMM (25 components, 100 EM iterations) uses the noise covariance $\hat{\sigma}^2 I_d$, with $\hat{\sigma}$ the standard deviation of the audited residual entries. MEC-MMD applies the audit correction of Section 3 with the same audit.

*Model and training.* All MMD-based methods fit the same normalising flow (4 rational-quadratic-spline blocks, 16 bins, hidden width 64, tail bound 1) with the Laplace kernel $k(x, y) = \exp(-\|x - y\|_1/h_q)$ averaged over nine bandwidths $h_q = \sqrt{Q_q}/\log_{10}(2n)$, $q = 0.1, \ldots, 0.9$, computed once at initialisation from pairwise distances between the observed sample and model samples (augmented with bootstrapped residuals for convMMD). Training uses 3000 full-batch epochs of AdamW (learning rate $10^{-2}$ with cosine decay, weight decay 0.1, gradient clipping at 1).

*Evaluation.* We report the sliced Wasserstein distance (Euclidean ground cost, 1000 random projections) between the held-out latent points and 10,000 samples from the fitted model, multiplied by $\sqrt{d}$ and averaged over 5 seeds.

## E Non-Uniform Audits

Simple random auditing (Assumption 2.2) can fail in cases where audits are stratified, outcome-dependent, or enriched for uncertain units. Let $\pi_j = \Pr(j \in L)$ and $\pi_{jj'} = \Pr(j, j' \in L)$ be the first- and second-order inclusion probabilities of the audit design, assumed known and positive. Replacing the factors $N/m$ and $N(N-1)/[m(m-1)]$ in (10) and (9) by $1/\pi_j$ and $1/\pi_{jj'}$ gives an

inverse-probability-weighted estimator $\widehat{\mathrm{MMD}}^2_{\mathrm{IPW}}$, which is design-unbiased for $\widehat{\mathrm{MMD}}^2_{\star}$ by the same Horvitz–Thompson argument as Theorem 3.1(Appendix C). Because small $\pi_j$ produce large weights, our experiments use the self-normalised (Hájek) version, which divides the audit corrections by the realised weight sums instead of $N$ and $N(N-1)$. It coincides with (11) when all weights are equal and trades exact unbiasedness for lower variance. We note that the variance and CLT results above are proved for simple random auditing only.

# F Suggestions for Future Direction and Improvements

**Utilizing the characteristics of ME.** The plug-in proxy $k(X_i, \widetilde{Y}_j)$ and $k(\widetilde{Y}_j, \widetilde{Y}_{j'})$ is only one possible choice. More generally, let

$$\widetilde{K}^{pq}_{ij} = \psi_{pq}(X_i, \widetilde{Y}_j), \qquad \widetilde{K}^{qq}_{jj'} = \psi_{qq}(\widetilde{Y}_j, \widetilde{Y}_{j'}),$$

be *any* proxy kernel predictions, possibly built from a denoiser, a deconvolution model, or an estimate $\widehat{\theta}$ of the noise law.

Define audited residuals

$$R^{pq}_{ij} := k(X_i, Y_j) - \widetilde{K}^{pq}_{ij}, \qquad R^{qq}_{jj'} := k(Y_j, Y_{j'}) - \widetilde{K}^{qq}_{jj'}.$$

Then the estimators

$$\widehat{A}^{(\psi)}_{pq} = \sum_{i=1}^{n}\sum_{j=1}^{N} \widetilde{K}^{pq}_{ij} + \frac{N}{m}\sum_{i=1}^{n}\sum_{j\in L} R^{pq}_{ij},$$

and

$$\widehat{A}^{(\psi)}_{qq} = \sum_{j\neq j'} \widetilde{K}^{qq}_{jj'} + \frac{N(N-1)}{m(m-1)} \sum_{\substack{j,j'\in L\\ j\neq j'}} R^{qq}_{jj'}$$

remain design-unbiased for $A^{\star}_{pq}$ and $A^{\star}_{qq}$ by exactly the same Horvitz–Thompson proof.

Thus, the hierarchical noise model $\varepsilon \sim P(\cdot \mid \phi)$, $\phi \sim \Pi$ may be used to construct *better* proxy kernel terms, but the correctness of that model is not required for the validity of the audit correction. As in prediction-powered inference, the proxy affects efficiency rather than validity.

**Adaptive interpolation and shrinkage.** The proposed MEC-MMD estimator corresponds to applying the full audit-based correction and is therefore design-unbiased for the oracle empirical MMD. However, as discussed in Section 5, this unbiasedness comes at the cost of increased variance when the audit sample is small. This suggests a natural extension based on shrinkage or interpolation between the naive and corrected estimators. In particular, one may consider the family

$$\widehat{\mathrm{MMD}}^2_{\lambda} = \widehat{\mathrm{MMD}}^2_{\mathrm{naive}} + \lambda\Big(\widehat{\mathrm{MMD}}^2_{\mathrm{MEC}} - \widehat{\mathrm{MMD}}^2_{\mathrm{naive}}\Big), \qquad \lambda \in [0,1],$$

which interpolates between the low-variance but biased naive estimator ($\lambda = 0$) and the unbiased but higher-variance MEC-MMD estimator ($\lambda = 1$). Intermediate values of $\lambda$ introduce bias but may substantially reduce mean-squared error when the audit-based correction is noisy.

Unlike the Horvitz–Thompson correction, the choice of $\lambda$ is not dictated by unbiasedness and must instead be selected to balance bias and variance. In principle, $\lambda$ could be estimated using sample splitting, cross-validation, or bootstrap-based risk estimation, analogously to recent work on prediction-powered inference and optimally weighted estimators [88, 89]. We emphasize that this interpolation should be viewed as an efficiency-tuning mechanism rather than a correction required for unbiasedness. A systematic study of optimal weighting strategies and their theoretical guarantees is an interesting direction for future work.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Theoretical guarantees are discussed in Section 3 and empirical performance and benchmarking is shown in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: There is an explicit limitation paragraph in Section 5. All the assumptions used in the theoretical results are listed in the Supplementary Materials.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All the assumptions are clearly stated and listed in the Supplementary Materials. All the proofs are provided in the Supplementary Materials. All the theorems, formulas, and proofs in the paper are numbered and cross-referenced; and all assumptions are referenced in the statement of any theorems.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The generative models used in simulation settings, data sets used in this work, and the hyperparameters are discussed in detail in Supplementary Materials. The benchmark methods used in this work are cited. The key contribution is an estimator.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release the codes used to produce all the results and figures in this work. The data sets used in this work are all publicly available and described in the supplemental material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: In addition to releasing the code, we believe all the necessary details required to replicate the results are provided, and the choice of random seed should not have an effect on the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The error bars are provided for the key experiments. To analyze bias and variance and our experimental results $n$ random realizations are generated. $n$ varies depending on the experiment and its computational requirements. The results should be independent of the choice of random seed.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: For the computationally heavy experiments, the hardware setup is provided. However, for the experiments that are not computationally heavy, we do not report the infrastructure; those experiments are executed on Google Colab.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We did not run human subject experiments; hence, human subject-related items do not apply to this work. We only used publicly available data, and the key contribution of this work is methodological.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This work proposed an estimator that can, in principle, be used in many applications. It does not have a direct positive or negative social impact; however, like any other methodological work, it can be used in applications with positive or negative social impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This work proposes a novel estimator for distribution matching. The method itself does not carry an inherent risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The data used in this work are all publicly available, and we cite them properly.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: We do not release a new asset.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We only use LLMs for writing, editing, formatting, and coding purposes. We evaluated the paper and checked the proofs using LLMs, but the ideas, theoretical results, experimental setups, research questions, and motivations are original, and LLMs were not used for these purposes.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.