Private and Scalable: Hash Kernel Bit-flip Two-Sample Testing

Jongmin Mun [*]
Data Sciences and Operations Department
USC Marshall School of Business
jongmin.mun@marshall.usc.edu

Abstract

Mun et al. (2024) proposed a minimax optimal two-sample (A/B) testing algorithm under local differential privacy (LDP). However, a key challenge arises when dealing with high-dimensional datasets, such as images or text, where the domain size k is massive or unknown. The resulting multinomial distribution implies that the entry-wise perturbation method of Mun et al. (2024), which relies on a variant of Erlingsson et al. (2014), introduces excessive noise relative to the signal, significantly reducing testing power. Furthermore, conventional dimension reduction methods like PCA are not directly applicable in the LDP setting, as individual data owners cannot access the global covariance structure. Since the test statistic relies on the Euclidean inner product—a simple linear kernel—we propose utilizing the hash kernel approximation from Shi et al. (2009) to enhance scalability. We present a modified LDP algorithm that projects high-dimensional data into a lower-dimensional sketch before noise injection. This approach enables efficient computation and handles unknown alphabet sizes while preserving the core geometric structure required for the two-sample test.

1 Introduction

Large-scale internet services collect sensitive data from vast user bases, enabling companies to conduct cost-effective randomized experiments (A/B testing). By testing whether two independent sets of samples are drawn from the same distribution, organizations can statistically evaluate the impact of new interfaces or campaigns. While non-parametric statistical tests allow for assessing general distributional changes beyond simple mean differences, the sensitivity of user data mandates rigorous privacy protections.

Our focus is on the Local Differential Privacy (LDP) model. In this setting, data is privatized on the user's device before being sent to the central aggregator, ensuring that the aggregator never sees the raw data. While LDP provides strong privacy guarantees, it incurs a statistical cost, typically requiring larger sample sizes. Mun et al. (2024) quantified this cost by establishing minimax rates for two-sample testing under LDP.

However, existing optimal algorithms struggle with high-dimensional data. Standard LDP mechanisms for multinomial data, such as RAPPOR, require a perturbation of the entire domain vector. When the domain size k is large (e.g., the set of all possible URLs or pixel combinations), the noise variance scales with k, rendering the test powerless. To address this, we propose a dimension-reduction strategy based on Hash Kernels, which allows for testing in high-dimensional spaces without requiring prior knowledge of the full alphabet size.

2 Related Works

The problem of private hypothesis testing sits at the intersection of robust statistics and privacy-preserving data analysis.

[*] Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies. Funding acknowledgements go at the end of the paper.

Two-Sample Testing. Non-privately, the Maximum Mean Discrepancy (MMD) introduced by Gretton et al. (2012) serves as a standard framework for non-parametric two-sample testing. The test statistic used in Mun et al. (2024) can be viewed as an unbiased estimator of the squared MMD with a linear kernel. Chan et al. (2014) established optimal rates for multinomial testing in the non-private setting, serving as a baseline for measuring the cost of privacy.

Local Differential Privacy (LDP). LDP has been extensively studied for estimation tasks, such as mean estimation (Duchi et al., 2018) and frequency estimation (Bassily & Smith, 2015; Wang et al., 2017). In the context of hypothesis testing, Gaboardi et al. (2016) and Rogers & Kifer (2017) have explored goodness-of-fit tests under central DP. More recently, Acharya et al. (2019) and Sheffet (2018) have investigated testing under LDP, though often focusing on discrete distributions with small alphabets.

Dimensionality Reduction in Privacy. To handle high-dimensional data in LDP, various sketching and projection techniques have been proposed. Bassily et al. (2017) utilized the Johnson-Lindenstrauss transform (random projection) for private frequency estimation. However, random projections often require dense matrix multiplications which can be computationally expensive on client devices. Cormode & Muthukrishnan (2005) and Erlingsson et al. (2014) utilize hashing and Bloom filters, respectively. Our work specifically leverages the Hash Kernel framework of Shi et al. (2009), also known as Feature Hashing, which provides a computationally efficient sparse projection that preserves inner products—the critical component of our test statistic.

3 Minimax Analysis Summary

We briefly summarize the theoretical foundation provided by Mun et al. (2024). Let $\mathcal{P}_{\text{multi}}^{(k)}$ denote the set of pairs of probability vectors with k categories. We observe raw sample sets $\{Y_i\}_{i=1}^{n_1}$ and $\{Z_j\}_{j=1}^{n_2}$ drawn from distributions with probability vectors $(\mathbf{p}_Y, \mathbf{p}_Z)$. The curator receives ϵ-LDP views $\{\tilde{Y}_i\}_{i=1}^{n_1}$ and $\{\tilde{Z}_j\}_{j=1}^{n_2}$ and must decide between the null hypothesis $\mathbf{p}_Y = \mathbf{p}_Z$ and the alternative hypothesis defined by the separation distance ρ_{n_1,n_2}:

$$\mathcal{P}_{1,\text{multi}}(\rho_{n_1,n_2}) := \big\{(\mathbf{p}_Y, \mathbf{p}_Z) \in \mathcal{P}_{\text{multi}}^{(k)} : \|\mathbf{p}_Y - \mathbf{p}_Z\|_2 \geq \rho_{n_1,n_2}\big\}.$$

Mun et al. (2024) established that the critical separation rate ρ_{n_1,n_2} necessary to uniformly control Type I and Type II errors scales as:

$$\frac{k^{1/4}}{(n_1\epsilon^2)^{1/2}} \vee \frac{1}{{n_1}^{1/2}}.$$

This result highlights the "curse of dimensionality" in the private regime: the separation rate depends on $k^{1/4}$. As $k \to \infty$ (e.g., in text or image domains), the required sample size to detect a difference explodes.

4 Methodology

To overcome the dependency on the domain size k, we propose a modification to the testing procedure that interlaces a dimensionality reduction step *before* the privatization step.

4.1 The Limitation of Bit-Flipping on One-Hot Vectors

The standard mechanism in Mun et al. (2024) treats a sample $X_i \in \{1, \ldots, k\}$ as a standard basis vector $\mathbf{e}_{X_i} \in \mathbb{R}^k$. It then applies a randomized response mechanism that flips specific entries. While minimax optimal for small k, this approach fails practically when k is large or unknown because:

1. **Noise Accumulation:** The mechanism adds variance to every dimension. With large k, the aggregated noise drowns out the signal.
2. **Implementation Constraints:** Constructing a vector of size k is infeasible if k represents, for example, the space of all n-grams in a text corpus.

4.2 Proposed Method: Private Hash Kernel Test

We propose replacing the raw high-dimensional input with a low-dimensional sketch using Feature Hashing (Hash Kernels), followed by a Laplace perturbation. This method relies on the observation that the U-statistic defined in Eq. equation 3 is composed entirely of inner products $\mathbf{Y}^\top \mathbf{Z}$. If we can approximate these inner products privately, we can recover the test statistic.

4.2.1 Step 1: Feature Hashing (Client-Side)

Let $h : \mathcal{X} \to [m]$ be a hash function that maps the original high-dimensional categories (or features) into a lower-dimensional space of size m, where $m \ll k$. Ideally, h is drawn from a family of pairwise independent hash functions. For a data point $\mathbf{x}$ (represented as a sparse vector or a set of features), we define the feature map $\phi : \mathbb{R}^k \to \mathbb{R}^m$ such that the j-th entry of the hashed vector is:

$$\phi_j(\mathbf{x}) = \sum_{i \in [k] : h(i) = j} x_i. \tag{1}$$

In the specific case where our raw data X is a single category (multinomial), the input vector $\mathbf{x}$ is 1-sparse (a single '1'). Consequently, the hashed vector $\phi(\mathbf{x})$ is also a standard basis vector in $\mathbb{R}^m$, denoted as $\mathbf{e}_{h(X)}$. This effectively "aliases" the large alphabet onto a smaller support m.

According to Shi et al. (2009), this transformation preserves the inner product in expectation:

$$\mathbb{E}[\phi(\mathbf{x})^\top \phi(\mathbf{x}')] = \mathbf{x}^\top \mathbf{x}'.$$

This property allows us to compute the test statistic on the hashed vectors while introducing a bounded amount of collision error, which is controlled by the target dimension m.

4.2.2 Step 2: Privacy Mechanism (Client-Side)

Unlike the bit-flipping mechanism which operates on bits, we view the hashed vector $\phi(\mathbf{x})$ as a numeric vector in $\mathbb{R}^m$. To satisfy ϵ-LDP, we apply the Laplace mechanism. The L_1-sensitivity of the hashed vector is bounded. Since $\mathbf{x}$ represents a single category, $\phi(\mathbf{x})$ has exactly one non-zero entry with value 1. The maximum L_1 distance between any two hashed vectors $\phi(\mathbf{x})$ and $\phi(\mathbf{x}')$ is:

$$\Delta_1 = \max_{\mathbf{x}, \mathbf{x}'} \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_1 = \|\mathbf{e}_j - \mathbf{e}_k\|_1 = 2 \quad (\text{for } j \neq k).$$

Therefore, to achieve ϵ-LDP, the client adds independent Laplace noise to each of the m components of the hashed vector. The privatized view $\tilde{\mathbf{z}}_i$ is given by:

$$\tilde{\mathbf{z}}_i = \phi(\mathbf{x}_i) + \eta_i, \quad \text{where } \eta_{i,j} \sim \text{Lap}\left(\frac{2}{\epsilon}\right). \tag{2}$$

4.2.3 Step 3: Server-Side Testing

The server receives the privatized hashed vectors $\{\tilde{\mathbf{z}}_Y, \dots\}$ and $\{\tilde{\mathbf{z}}_Z, \dots\}$. The server then computes the U-statistic using these low-dimensional vectors:

$$U_{\text{hash}} := \frac{1}{n_1(n_1-1)} \sum_{i \neq k} \tilde{\mathbf{z}}_{Y_i}^\top \tilde{\mathbf{z}}_{Y_k} + \frac{1}{n_2(n_2-1)} \sum_{j \neq l} \tilde{\mathbf{z}}_{Z_j}^\top \tilde{\mathbf{z}}_{Z_l} - \frac{2}{n_1 n_2} \sum_{i,j} \tilde{\mathbf{z}}_{Y_i}^\top \tilde{\mathbf{z}}_{Z_j}. \tag{3}$$

Finally, the permutation test is applied to U_{hash} as described in Section 3 to determine the rejection threshold.

4.3 Advantages over Previous Methods

- **Fixed Dimensionality:** The noise added depends on m (the hash size) rather than k. We can choose m based on the available sample size n to optimize the bias-variance tradeoff, effectively decoupling the privacy cost from the raw data dimensionality.
- **Unknown Alphabet Support:** This method handles online settings where new categories appear dynamically. The hash function maps any new input to $[m]$ without requiring a dictionary update or protocol renegotiation.

- **Computational Efficiency:** Computing $\phi(\mathbf{x})$ requires $O(1)$ operations per token (sparse input), and the inner product computation scales with m rather than k.

By utilizing the hash kernel approximation, we transition from a regime where privacy costs are dictated by the massive universe of possible data values to a regime where costs are controlled by a user-selected parameter m, making private two-sample testing feasible for complex, high-dimensional data types.

REFERENCES

Jayadev Acharya, Clément L Canonne, Cody Freitag, and Himanshu Tyagi. Test without trust: Optimal locally private hypothesis testing. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2067–2076. PMLR, 2019.

Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 127–135, 2015.

Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Guha Thakurta. Practical locally private heavy hitters. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1193–1203. SIAM, 2014.

Graham Cormode and S Muthukrishnan. An improved data stream summary: The count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1054–1067, November 2014.

Marco Gaboardi, Hyunwu Lim, Ryan M Rogers, and Salil P Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *International Conference on Machine Learning*, pp. 2111–2120. PMLR, 2016.

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.

Jongmin Mun, Seungwoo Kwak, and Ilmun Kim. Minimax optimal two-sample testing under local differential privacy, 2024. URL `https://arxiv.org/abs/2411.09064`.

Ryan Rogers and Daniel Kifer. A new class of private chi-square tests. In *Artificial Intelligence and Statistics*, pp. 991–1000. PMLR, 2017.

Or Sheffet. Locally private hypothesis testing. In *International Conference on Machine Learning*, pp. 4605–4614. PMLR, 2018.

Qinfeng Shi, James Petterson, Gideon Dror, John Langford, Alex Smola, and S. V. N. Vishwanathan. Hash Kernels for Structured Data. *Journal of Machine Learning Research*, 10(90):2615–2637, 2009. ISSN 1533-7928. URL `http://jmlr.org/papers/v10/shi09a.html`.

Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Locally differentially private heavy hitter identification for high-dimensional data. *Proceedings of the VLDB Endowment*, 10(12): 1866–1879, 2017.