arXiv:2611.31208v1 [cs.LG] 23 Nov 2026

度量隐私下的匹配*

Hongyi B. Hu[1] Dung Nguyen[2,3] M. Amin Rahimian[1]

[1]匹兹堡大学[2]哈弗福德学院[3]圣
安东尼奥德克萨斯大学

2026年11月20日

摘要

我们研究如何表示并随机化 d维欧几里得空间中的点，以在度量隐私下获得准确的二分图匹配。给定 n 个私有点和 n 个公开点，我们分析对每个私有点或其到公开点的距离进行随机化的本地机制。利用公开点的仿射张成空间，我们构造了一种更小的表示，同时保留所有匹配代价。当可信服务器接收私有点时，我们为 Gumbel 匹配建立隐私和效用保证；该方法对边代价进行随机化，并仅发布分配。通过上界和最坏情况构造，我们确定了期望代价高于非隐私最优值的紧确速率$\Theta(n\min\{d,\log n\}/\varepsilon)$。基准实验表明， $d \approx \log n$ 可作为实用指南，帮助判断何时应从直接随机化点坐标转向使用 Gumbel 匹配。对平方欧几里得代价的扩展使私有非参数两样本假设检验成为可能。这些结果表明，应用的信任模型会影响表示方式和维度的选择，从而提升效用：在本地隐私下，低维点表示可降低超额代价；而在中心隐私下，即使处于高维，距离表示也能达到最优速率。

1 引言

最小代价匹配在组合优化领域有着悠久的历史，相关算法已十分成熟，并衍生出众多扩展 [11]。这些算法在分配等应用中十分有效 [22, 58], ，例如数据对齐和适配 [14, 15, 54],以及统计推断 [7, 17]。

尽管匹配算法种类繁多，但观测到的输出可与任何先验信息结合，用来推断并暴露输入；对于输入数据敏感的应用而言，这是不可接受的。差分隐私通过限制任意替换单个数据点后输出分布的变化幅度来提供保护 [20]。度量隐私则进一步根据输入之间的距离对可区分性的限制进行缩放：输入的微小变化应导致输出分布的微小变化 [12]。

在这项工作中，我们研究如何进行匹配，以最小化两点属于集合数据的距离 X 与 Y，每个包含 n 个对象，且只有 X 中的点属于敏感数据。距离是在满足度量隐私的同时，发布一个匹配$\sigma: \ i \mapsto j$，并使总匹配成本 $\sum_i \|\boldsymbol{x}_i - \boldsymbol{y}_{\sigma(i)}\|_2$ 尽可能小。当领域自适应中将机密目标数据与公开来源数据对齐，或根据工人的私有属性将其分配给公开任务时，这种设置自然会出现。

*代码可用性：https://github.com/hu-bryan/metric-private-matching。

关于隐私匹配的既有工作与我们主要有两点不同。首先，差分隐私文献研究的是对节点和边 [18], 边权重的元组 [52], 的存在性提供保护的匹配，或为分配提供联合差分隐私 [30, 32, 34]。这些工作都没有对点的底层数值进行隐私化。其次，研究度量隐私下匹配的空间众包和车辆分配工作仅限于二维位置和结构化的有限度量空间 [51, 57, 62, 68]。因此，它们的方法无法扩展到一般维度的欧几里得空间中的匹配。

我们通过考察欧几里得空间中的数据如何表示和随机化，以及这对满足度量隐私的输出匹配有何影响，来弥补这一空白。对欧几里得空间中的数据进行消敏的一种标准方法是使用拉普拉斯机制，通过 $\boldsymbol{x} \mapsto \boldsymbol{x} + \mathbf{Z}$扰动每个点，其中$\mathbf{Z} \sim p(\boldsymbol{z}) \propto \exp(-\varepsilon\|\boldsymbol{z}\|_2)$ [12, 29]。该机制满足度量隐私，且尤其能使重构误差最小化 $\mathbb{E}[\|\mathbf{Z}\|^2]$ [39]。以这种方式发布的数据尤其适合用于估计质心等聚合统计量 [4], ，但基于这些数据计算出的匹配可能会产生相对于非隐私最优值而言相当大的超额代价。

我们发现，公开点 Y 使我们能够减少需要添加噪声的坐标数量。将私有点 X 投影到 Y 的仿射张成空间上，并保留它们到该空间的距离，可以保持这些点之间的所有成对距离。因此，我们可以在更少的坐标上添加噪声，同时不丢失匹配所需的信息。

当可信平台计算匹配时，可以将私有点保留在内部，只发布在度量隐私保护下的分配。因此，隐私噪声只需用于防止从匹配中进行推断。我们基于 Gumbel-max 技巧改造 Gumbel 匹配 [46],, 将噪声添加到 n^2 边而非 n匹配中！据我们所知，我们是首个建立 Gumbel 匹配与度量隐私之间联系的团队。此外，我们构造了 X， Y 的最坏情况实例，表明我们的机制在度量隐私下实现了可能达到的最低渐近额外成本。

另一项优势是，即使成本为欧几里得距离平方 $\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$，我们的所有机制也都能提供隐私和效用保证。最小化这一目标等同于最小化均方误差并最大化点积之和，这两种做法在统计学和机器学习中都得到广泛应用。我们通过非参数双样本检验展示其下游价值：将样本与公开参考网格进行匹配，可以构造多元秩[17]。以私密方式计算这一匹配，可将检验扩展到机密样本。

总结我们的贡献：

- 通过对单个数据点进行隐私化处理，我们证明，对于某些 $s \le \min\{n, d\}$，可以实现误差为$O(ns/\varepsilon)$ 的匹配。
- 我们证明，任何 εD-隐私匹配机制在某个最坏情形实例上的误差都至少为 $\Omega(n\min\{d,\log\ n\}/\varepsilon)$; 这表明， $d \approx \log n$ 是区分低维与高维情形的理论阈值。
- 我们在低维和高维情形下分别得到了上界为 $O(nd/\varepsilon)$ 和 $O(n\log\ n/\varepsilon)$ 的机制，从而达到最优。
- 我们在 16 个基准数据集上对机制进行实证检验，考察匹配误差如何随隐私级别、样本量和数据维度变化。我们进一步将隐私匹配用于构建多元秩，以进行非参数两样本假设检验。在 Sonar 数据集上，Gumbel 匹配保留了最多拒绝原假设的证据，在 $\varepsilon = 16$ 处获得低于 0.05 的 p 值。

附录A 综述了度量隐私、私有优化、几何表示以及位置隐私分配方面的相关工作。

2 问题设定与隐私预备知识

匹配。 设 $X = (\boldsymbol{x}_i)_{i=1}^n$ 为私有数据集，$Y = (\boldsymbol{y}_j)_{j=1}^n$ 为公开数据集，其中的数据点属于 $\mathbb{R}^d$。每个排列 $\sigma \in S_n$ 都定义了私有点 $\boldsymbol{x}_i$与公共点 $\boldsymbol{y}_{\sigma(i)}$之间的一种匹配。我们将代价矩阵 $\boldsymbol{C} = (C_{ij})$、效用 $u(\sigma)$以及匹配问题的最优值定义为

$$C_{ij} = \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2, \qquad u(\sigma) = \sum_{i=1}^{n} C_{i,\sigma(i)}, \qquad \mathrm{OPT} = \min_{\sigma \in S_n} u(\sigma). \tag{1}$$

Th在本文中，我们分析额外匹配成本 $\mathrm{err}(\sigma) = u(\sigma) - \mathrm{OPT}$.

度量隐私。 设数据集之间的隐私度量定义为

$$D(X, X') = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{x}'_i\|_2. \tag{2}$$

给定一个固定且公开的 Y，若机制 M_Y 满足 εD-隐私 [12] ，则

$$\Pr[M_Y(X) \in A] \le \exp\big(\varepsilon D(X, X')\big) \Pr[M_Y(X') \in A] \tag{3}$$

对于每个 X， $X' \in (\mathbb{R}^d)^n$ 以及每个可测事件 A均成立。在我们的情形中， M_Y是定义在匹配集合上的概率分布。

局部度量隐私是一种更强的隐私概念，它要求数据集中的每个条目 i 都通过机制 $K_{i,\ Y}(\boldsymbol{x}_i)$ 独立地对自身的数据点进行隐私保护，其中

$$\Pr[K_{i,Y}(\boldsymbol{x}) \in A] \le \exp\big(\varepsilon \|\boldsymbol{x} - \boldsymbol{x}'\|_2\big) \Pr[K_{i,Y}(\boldsymbol{x}') \in A] \tag{4}$$

对每个 $\boldsymbol{x}$， $\boldsymbol{x}' \in \mathbb{R}^d$ 以及可测的 A 均成立。换言之，$K_{i,\ Y}$在单点数据集上满足 $\varepsilon\|\cdot\|_2$-隐私。随后，通过并行组合局部度量隐私机制定义的机制 $M_Y(X) = (K_{i,\ Y}(\boldsymbol{x}_i))_{i=1}^n$ 满足 εD-隐私。

此外，顺序组合两个隐私机制 M_1， M_2 时，它们的隐私预算会相加，从而满足 $(\varepsilon_1 + \varepsilon_2)D$-隐私。最后，对隐私机制输出进行任何后处理，结果仍然保持隐私。其证明遵循差分隐私中的标准论证 [20]。

附录 B.1 说明了与差分隐私的联系，以及如何解释和校准ε。

拉普拉斯机制。 实现局部度量隐私的一种方法是添加（广义）拉普拉斯噪声。对于任意 $\varepsilon > 0$ 和 $\mathbb{R}^s$ 上的范数 $\|\cdot\|$，拉普拉斯机制将 $\boldsymbol{v} \in \mathbb{R}^s$ 映射到 $\boldsymbol{v} + \mathbf{Z}$，其中噪声与输入相互独立，且其密度为

$$p_{\mathbf{Z}}(\boldsymbol{z}) \propto \exp(-\varepsilon\|\boldsymbol{z}\|).$$

该机制满足 $\varepsilon\|\cdot\|$-隐私。当范数为欧几里得范数时，噪声可分解为$\mathbf{Z} = R\boldsymbol{\Theta}$，其中 $R \sim \mathrm{Gamma}(s, \mathrm{rate}\ \varepsilon)$，且 $\boldsymbol{\Theta}$ 在单位球面 $\mathbb{S}^{s-1}$[5]上服从独立均匀分布。对于任意选定的范数，噪声的期望大小为 $\mathbb{E}\|\mathbf{Z}\| = s/\varepsilon$。

本文中的若干机制会先处理私有数据点，通过$h_Y(\boldsymbol{x}_i)$表示这些数据点，然后再添加噪声。 h_Y 的利普希茨常数决定了相对于原始度量、实现局部 ε-度量隐私所需的噪声尺度。

命题 1。 假设 h_Y： $\mathbb{R}^d \to \mathbb{R}^s$ 对于某个 $L > 0$是 L-利普希茨的，即对于所有 $\boldsymbol{x}$、 $\boldsymbol{x}' \in \mathbb{R}^d$， $\|h_Y(\boldsymbol{x}) - h_Y(\boldsymbol{x}')\| \leq L\|\boldsymbol{x} - \boldsymbol{x}'\|_2$。那么，将拉普拉斯机制应用于 $h_Y(\boldsymbol{x})$且 $\varepsilon' = \varepsilon/L$时，是 $\varepsilon\|\cdot\|_2$-隐私的。

2.1 扰动点与距离

最简单的方法是直接向每个私有点添加拉普拉斯噪声。每个人独立发布 $\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i + \mathbf{Z}_i$，其中 $\mathbf{Z}_i$ 的密度为 $p_{\mathbf{Z}_i}(\boldsymbol{z}) \propto \exp(-\varepsilon\|\boldsymbol{z}\|_2)$。我们计算$\widetilde{C}_{ij} = \|\widetilde{\boldsymbol{x}}_i - \boldsymbol{y}_j\|_2$ 并返回

$$\widehat{\sigma} \in \arg\min_{\sigma \in S_n} \sum_{i=1}^{n} \widehat{C}_{i,\sigma(i)},$$

使用任意一种最小成本匹配算法。

定理 2. 输出匹配满足 εD-隐私，并且

$$\mathbb{E}[\mathrm{err}(\widehat{\sigma})] \leq \frac{2nd}{\varepsilon}. \tag{5}$$

该保证随 d 线性增长，因此在高维情况下实用性较低。不过，匹配只需要边权重 $\|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2$，因此我们可以用距离向量替代每个私有点

$$\Phi_Y(\boldsymbol{x}) = \big(\|\boldsymbol{x} - \boldsymbol{y}_1\|_2, \ldots, \|\boldsymbol{x} - \boldsymbol{y}_n\|_2\big).$$

该向量包含 $\boldsymbol{C}$ 中一行的所有代价，并且有 n 个坐标，与 d 无关。对于 $p \in [1, \infty]$，每个人 i 独立发布

$$\widetilde{\boldsymbol{c}}_i = \Phi_Y(\boldsymbol{x}_i) + \mathbf{Z}_i, \qquad p_{\mathbf{Z}_i}(\boldsymbol{z}) \propto \exp\Big(-\frac{\varepsilon}{n^{1/p}}\|\boldsymbol{z}\|_p\Big).$$

$\widehat{}$我们设定 $C_{ij} = [\widetilde{\boldsymbol{c}}_i]_j$，并为 $\boldsymbol{C}$ 计算

最小代价匹配。

定理 3. 对于每个 $p \in [1, \infty]$，输出匹配满足 εD-隐私，并 d

$$\mathbb{E}[\mathrm{err}(\widehat{\sigma})] \leq 2n\,\mathbb{E}\|\mathbf{Z}\|_p = \frac{2n^{2+1/p}}{\varepsilon}, \tag{6}$$

其在 $p = \infty$ 处取最小值，得到

$$\mathbb{E}[\mathrm{err}(\widehat{\sigma})] \leq \frac{2n^2}{\varepsilon}. \tag{7}$$

证明见附录 C。距离向量界不再依赖于 d，但会随 n 二次增长。因此，在这两种方法之间，权衡在于要么向 d个坐标添加噪声，要么向 n 个距离添加噪声。然而，这 n个距离是由同一个私有点以及仿射张成空间维度至多为 d 的公开点计算得到的。这表明可以利用该几何结构，以更少的含噪坐标保留相同的距离信息。

3 匹配机制

3.1 投影表示

我们采用 Narayanan 和 Nelson [50] 提出的一种表示，该表示保留公开点与私有点之间的距离。它保留每个私有点在公共仿射包上的投影及其到该包的距离。在平移两个数据集使得 $\boldsymbol{y}_1 = 0$ 之后，令

$$W = \operatorname{span}\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}, \qquad r = \dim W, \qquad s = \min\{r+1, d\}.$$

令 P_W 与 $P_{W^\perp}$ 表示到 W 与 $W^\perp$ 上的正交投影。将投影向量写在 W 的一个固定公开正交基中，定义 $f_Y\colon \mathbb{R}^d \to \mathbb{R}^s$ 为

$$f_Y(\boldsymbol{x}) = \begin{cases} \big(P_W \boldsymbol{x}, \|P_{W^\perp}\boldsymbol{x}\|_2\big), & s < d, \\ \boldsymbol{x}, & s = d. \end{cases} \tag{8}$$

垂直于 W 的分量仅通过其模长影响每个公开到私有的距离，因此只需再增加一个坐标即可。当这并不能降低维度时，我们保留原始坐标。该表示保留了所有匹配代价：

$$\|f_Y(\boldsymbol{x}) - f_Y(\boldsymbol{y}_j)\|_2 = \|\boldsymbol{x} - \boldsymbol{y}_j\|_2.$$

由于 $r \le n-1$，该表示最多使用 $\min\{n, d\}$ 个坐标。

每个个体独立地发布

$$\widetilde{\boldsymbol{x}}_i = f_Y(\boldsymbol{x}_i) + \mathbf{Z}_i, \qquad p_{\mathbf{Z}_i}(\boldsymbol{z}) \propto \exp(-\varepsilon\|\boldsymbol{z}\|_2),$$

在 $\mathbb{R}^s$ 中加入噪声。W 匹配。 ^e计算 $C_{ij} = \|\widetilde{\boldsymbol{x}}_i - f_Y(\boldsymbol{y}_j)\|_2$ 并返回 $\boldsymbol{C}$的最小代价

定理 4. 输出匹配满足 εD-隐私，并且

$$\mathbb{E}[\operatorname{err}(\widehat{\sigma})] \le \frac{2ns}{\varepsilon} \le \frac{2n\min\{n, d\}}{\varepsilon}. \tag{9}$$

证明见附录C.4。

3.2 Gumbel 匹配

当计算匹配的平台可信时，可以用局部度量隐私换取更好的效用。我们可以直接随机化分配，而不是发布每个私有点的含噪表示。我们使用 Gumbel 匹配，向边分数添加独立的 Gumbel 噪声，并计算最大权匹配 [46]。标准 Gumbel 分布 Gumbel(0,1) 的累积分布函数为

$$\Pr[G \le t] = \exp(-e^{-t}), \qquad t \in \mathbb{R}.$$

对于我们的欧几里得代价，独立抽取 $G_{ij} \sim \text{Gumbel}(0,1)$，并返回

$$\widehat{\sigma} \in \arg\max_{\sigma \in S_n} \sum_{i=1}^{n} \left(-C_{i,\sigma(i)} + \frac{2}{\varepsilon} G_{i,\sigma(i)} \right). \tag{10}$$

定理 5。 对于每个 $X, Y \in (\mathbb{R}^d)^n$ 和 $\varepsilon > 0$，*Gumbel* 匹配满足 εD-隐私，并且

$$\mathbb{E}[\operatorname{err}(\widehat{\sigma})] \le \frac{2n\log n}{\varepsilon}. \tag{11}$$

证明见附录C.5。

4 隐私下界

前述保证并未说明是否存在其他机制可以改善对n 和 d的依赖关系。现在我们给出度量隐私所要求误差的下界。

定理 6. 对于每一个 $n \geq 2$、 $d \geq 1$ 和 $\varepsilon > 0$，存在一个公开数据集 $Y \in (\mathbb{R}^d)^n$ ，使得每一个 εD-隐私匹配机制 M_Y 都有一个私有输入 $X \in (\mathbb{R}^d)^n$ ，其输出$\widehat{\sigma} = M_Y(X)$ 满足

$$\mathbb{E}[\mathrm{err}(\widehat{\sigma})] = \Omega\left(\frac{n \min\{d, \log n\}}{\varepsilon}\right). \tag{12}$$

我们的证明采用打包方法（例如参见 Hardt 和 Talwar [29]）以获得隐私下界：取 Y $\Omega(h)$相距的点，我们构造一族 $\{X^\pi\}$，每个都具有最优匹配π，使得

$$\mathrm{err}(\sigma) \gtrsim h \, \#\{i : \sigma(i) \neq \pi(i)\}.$$

我们可以通过证明均匀随机选取的一个在期望下确实如此，来说明某个 X^π 能达到 $\Omega(n)$ 次此类不一致。这可以通过如下方式证明：$\Pr[\sigma(i) \neq \pi(i)]$ 在 $\Omega(n)$ 个下标 $i \in [n]$ 上具有常数下界。在 X^π 的范围内，$\pi(i)$ 取 k 个不同的值，且有 $\log k = \Theta(\min\{d, \log n\})$。具有$\pi(i) = a$ 的输入可以与具有 $\pi'(i) = b$ 的输入配对，只需在最优匹配中交换 a, b ，同时保持 $D(X^\pi, X^{\pi'}) = O(h)$。因此，度量隐私将这些情形之间任意被发布伙伴的概率比以 $e^{O(\varepsilon h)}$为界，包括在对 π 的其余选择取平均之后。将 h 取为 $\log k/\varepsilon$ 的一个足够小的常数倍，这些界与概率归一化便迫使得到所需的概率下界。

完整的构造和证明见附录C.6。

当 $d \leq \log n$ 时选择原始点扰动，否则采用 Gumbel 匹配，可得

$$\mathbb{E}[\mathrm{err}(\widehat{\sigma})] \leq \frac{2n \min\{d, \log n\}}{\varepsilon}.$$

该选择仅取决于公开的 n， d，并保持 εD-隐私。值得注意的是， $d \approx \log n$ 阈值划分了低维与高维情形。

5 扩展

平方欧几里得匹配。 当匹配代价为平方欧几里得距离时，同样适用这四种机制：

$$C_{ij} = \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2^2.$$

总成本展开为

$$\begin{aligned} u(\sigma) &= \sum_i C_{i,\sigma(i)} = \sum_i \|\boldsymbol{x}_i - \boldsymbol{y}_{\sigma(i)}\|_2^2 \\ &= \sum_i \|\boldsymbol{x}_i\|_2^2 + \sum_j \|\boldsymbol{y}_j\|_2^2 - 2\sum_i \langle \boldsymbol{x}_i, \boldsymbol{y}_{\sigma(i)} \rangle. \end{aligned} \tag{13}$$

由于双射中的每个点都恰好使用一次，前两项与匹配无关。因此，最小化目标等价于最大化匹配点积总和：

$$\arg\min_{\sigma \in S_n} u(\sigma) = \arg\max_{\sigma \in S_n} \sum_i \langle \boldsymbol{x}_i, \boldsymbol{y}_{\sigma(i)} \rangle.$$

均方误差和点积之和在统计学和机器学习中无处不在。附录D.1给出了该目标下相应的期望超额代价上界。

表 1：实验配置。维度与样本量扫描每次改变一个量，同时保持另一个量不变。

实验	数据集	n每侧	预算与目标
隐私扫描	全部十六个	64	$\varepsilon = 0.25, 0.50, \ldots, 16$；两种成本
固定预算柱状图	全部十六个	64	$\varepsilon = 8$；两种成本
维度扫描	Fashion-MNIST	800	$\varepsilon = 16$；欧几里得
样本量扫描	纸币	20、40、.....、200	$\varepsilon = 8$；欧几里得

两个数据集均为私有时的匹配。当 X 和 Y 均为敏感数据时，隐私机制必须保护任一数据集的变更。附录D.2 给出了欧几里得匹配对应的机制和上界。

6 数值实验

我们考察理论上隐私、维度和样本量对结果的影响如何体现在基准测试数据中。实验围绕三个问题展开：隐私预算增加时效用如何提升；维度变化时，点扰动与直接匹配之间的选择如何变化；以及增加匹配伙伴的数量会如何影响误差。我们采用几何形态各异的真实世界和合成基准测试，使这些比较不局限于任何单一分布。

6.1 实验设置

我们比较 原始点、投影表示、距离向量和$Gumbel$ ，所采用的欧几里得隐私度量与理论分析中的度量一致。隐私预算扫描同时使用欧几里得代价 $C_{ij} = \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2$和平欧几里得代价，$C_{ij} = \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2^2$；维度与样本量扫描则使用欧几里得代价。附录 E.3 给出了各目标的校准方法。

主要比较采用十六个数据集。其中十五个来自版本化的 聚类算法基准套件 [24, 25], ，涵盖合成形状、生物与物理测量数据以及图像，维度范围为 2 至 784。我们还加入了纸币认证 [44]：其四个特征和 1,372 条记录使我们能够在 $\log n = 4$ 两侧使用不相交样本进行样本量扫描。表 3 列出了每个数据集的原始来源和可用规模。类别标签不纳入特征，也不参与采样或匹配。

特征与距离尺度。 表格数据集可能混合具有不同单位和数值范围的特征。我们通过减去其均值并除以其标准差，对其各不相同的数值特征进行标准化，同时在最终的共同重新缩放之前，将二元指示变量映射为零和一。例如，一条大肠杆菌记录将数值序列得分与二元指示变量相结合；我们标准化这些得分，并将每个指示变量保留为一个坐标。这遵循机器学习中的标准输入归一化 [27] ，并设定匹配距离中的相对特征权重。鸢尾花、合成数据集和图像数据集在各坐标上具有一致的单位，因此我们保留其相对特征尺度。

经过这种特征处理后，每个实际匹配实例都会进行中心化，并除以合并后的均方根（RMS）半径，使均方根半径为 1。隐私保证适用于这些归一化输入，且预处理保持不变。

维度与样本量扫描。为考察误差如何随维度增长，我们固定$n = 800$，并将 Fashion-MNIST 图像投影到一系列嵌套的随机子空间中，维度逐渐降低，起始维度为全部 784 个坐标。为考察误差如何随样本量增长，我们固定 $d = 4$，并从 Banknote 数据集中依次抽取嵌套样本，每侧包含 $n = 20$、40、……、200 个点。两种扫描都会在每种设置下重复中心化和 RMS 重缩放，同时保持特征权重不变。因此，在总体距离尺度和隐私预算相同的情况下，误差变化反映的是维度或样本量的变化。附录E.2介绍了具体的变换，以及这些变换与附录B.1中的隐私校准之间的关系。

评估。 在每项实验中，我们抽取二十个独立的记录集合，并在每个集合内进行无放回采样。每个集合提供五个分区，每个分区具有互不相交的公开侧与私有侧；在维度扫描中，还提供五条随机投影链。我们在扰动后精确求解分配问题，并基于真实成本评估超额代价。我们按照 RAND $= n^{-1}\sum_{i,j} C_{ij}$ 的方式精确计算均匀随机匹配的基准。最优解与随机基准均针对每个实例和目标分别评估。

对于每个固定的匹配实例，我们对 50 次独立机制运行取平均，然后在每个集合内对分区和投影链取平均。所报告的均值是这二十个集合均值的平均。附录E.4 报告了它们的置信区间，计算细节见附录 E.3。

6.2 结果

隐私预算扫描与基准比较。 增大 ε 通常会降低额外匹配成本（附录 E.5）。图 1 表明，各机制的相对表现因数据集和目标函数而异。对于欧几里得代价，点方法在葡萄酒上给出 16.43，而 Gumbel 为 17.27；WDBC 则更有利于 Gumbel，为 16.48 对 21.26。对于平方代价，投影表示在大肠杆菌上给出 14.92，而 Gumbel 为 36.54；WDBC 同样更有利于 Gumbel。

维度扫描。在图2(a) 中，两种点方法的结果一致，且其超额代价随维度增加。Gumbel 在 $d = 19$ 和 24 之间超过了它们。Gumbel 和距离向量都逐渐趋于平台，但 Gumbel 始终较低。在测量维度范围内，综合理论保证始终低于随机基准的均值 $d = 14$。表11 给出了完整的数值比较。

样本量扫描。在图2(b) 中，点方法的每对超额成本始终低于 Gumbel。点方法和 Gumbel 的曲线逐渐趋于平缓，而距离向量则继续更明显地上升。表4给出了数值结果。

6.3 解释

Gumbel 的优势表明，改进表示并非提升隐私匹配效果的唯一途径。Gumbel 和距离向量对相同的匹配代价添加扰动，但将隐私保护施加于不同的输出。距离向量保护整个报告，而 Gumbel 将扰动后的代价保留在可信服务器内，仅发布分配。因此，后者可以使用专门校准的噪声来保护匹配，而不必使每个报告的距离都具有隐私性。

这一区别给出了对候选数量的对数依赖，而非对环境维度的依赖。$\log n$ 项来自每一行中n 冈贝尔变量的期望最大值。结合下界，这表明 Gumbel 达到

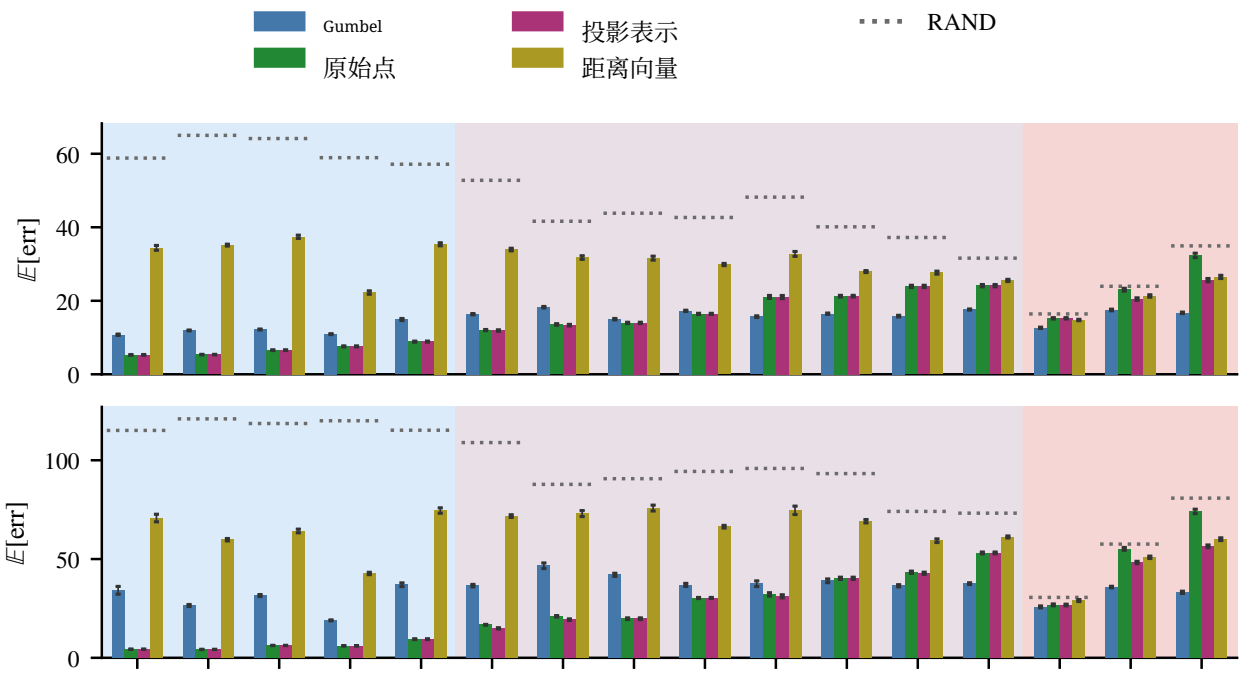

图 1：在 $\varepsilon = 8$, $n = 64$ 下，欧几里得（上）和平方欧几里得（下）目标函数的额外匹配成本。蓝色方块表示 $d \le \log n$，玫瑰色方块表示 $d \ge n$，两者的等权混合为 $\log n < d < n$。每个数据集内的柱形按图例顺序排列。虚线段表示该数据集和目标函数对应的随机基准。

高维下的最优超额代价阶。因此，其经验优势伴随着一项适用于已评估数据集之外的理论保证。

维度与样本量扫描从两个方向考察与点扰动的比较。减小 d 或增大 n 会趋近于 $d \le \log n$，此时原始点在局部度量隐私下已达到最优阶。若公共仿射包允许，投影表示还能进一步减少需要添加噪声的坐标数量。理论界限与实验比较共同表明，可将 $d \approx \log n$ 作为实用指南，用于判断优势何时会从点扰动转向 Gumbel 匹配。

7 非参数双样本假设检验

Deb 和 Sen [17] 基于多元秩提出了一种分布无关的非参数双样本检验。其经验秩映射通过最小化总平方欧几里得距离，将合并观测值匹配到固定参考网格。所得秩能量统计量比较两个经验秩分布。

当观测值属于机密时，对秩分配进行私有化提供了一种在度量隐私下执行该检验的自然方法。合并观测值构成匹配的私有侧，而同等规模的参考网格构成公开侧。我们描述两样本

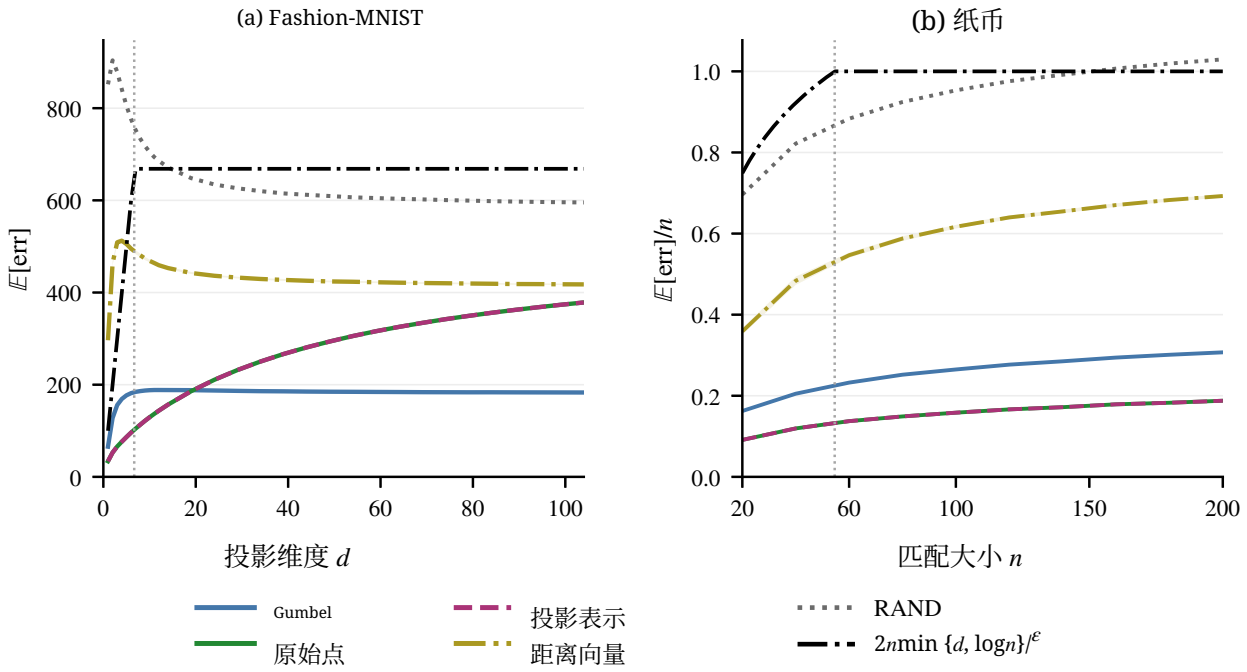

图 2：采用 RMS 重缩放的维度与样本量扫描。(a) Fashion-MNIST：$n = 800$,$\varepsilon = 16$，总超额代价。(b) 纸币：$d = 4$, $\varepsilon = 8$，每对超额成本。灰色点线表示相应的随机基准。黑色点划线在 (a) 中表示 $2n\min\{d,\log n\}/\varepsilon$，在 (b) 中表示 $2\min\{d,\log n\}/\varepsilon$。竖直虚线分别标出 $d = \log(800) \approx 6.68$ 和 $n = e^4 \approx 54.60$。

设定如下，然后在声呐上评估该私有检验。

7.1 双样本秩能量检验

设 $X = (\boldsymbol{x}_i)_{i=1}^N$ 将两个来自 $\mathbb{R}^d$ 的独立样本合并，其中 $\boldsymbol{x}_1$、……、$\boldsymbol{x}_a$ 取自 F，而 $\boldsymbol{x}_{a+1}$、……、$\boldsymbol{x}_N$ 取自 G，其中 $N = a + b$。队列标签是公开的。我们检验

$$H_0 : F = G \qquad \text{versus} \qquad H_1 : F \neq G.$$

固定 N 个互不相同的公共点 $Y = (\boldsymbol{y}_j)_{j=1}^N \subset [0,1]^d$，且不依赖于观测值。$\boldsymbol{x}_i$ 的经验多元秩是由以下方式分配的网格点 $\boldsymbol{y}_{\sigma^\star(i)}$：

$$\sigma^\star \in \arg\min_{\sigma \in S_N} \sum_{i=1}^{N} \|\boldsymbol{x}_i - \boldsymbol{y}_{\sigma(i)}\|_2^2. \tag{14}$$

对于任意分配 σ，将秩能量统计量定义为

$$\begin{aligned}\mathcal{E}(\sigma) = \frac{2}{ab}\sum_{i=1}^{a}\sum_{j=a+1}^{N}\|\boldsymbol{y}_{\sigma(i)} - \boldsymbol{y}_{\sigma(j)}\|_2 \\ - \frac{1}{a^2}\sum_{i,j=1}^{a}\|\boldsymbol{y}_{\sigma(i)} - \boldsymbol{y}_{\sigma(j)}\|_2 \\ - \frac{1}{b^2}\sum_{i,j=a+1}^{N}\|\boldsymbol{y}_{\sigma(i)} - \boldsymbol{y}_{\sigma(j)}\|_2, \qquad T(\sigma) = \frac{ab}{N}\mathcal{E}(\sigma).\end{aligned} \tag{15}$$

能量统计量使用未平方欧几里得距离来衡量经验秩分布之间的分离度。较大的 T 表明有证据反对 H_0。

我们将 $\sigma^\star$ 替换为 εD私有分配 $\widehat{\sigma}$，并根据其秩计算 $T(\widehat{\sigma})$。在 H_0下，精确秩统计量具有分布无关的零分布：其分布仅取决于 Y、a、b，而与观测值的未知共同分布无关（见 [17, 第 4 节2]）。当匹配规则忽略队列标签并对称地处理观测行时，私有化后也具有相同性质。我们根据该零分布计算上尾 p值，并在 $p \le 0.05$处拒绝原假设。附录F 介绍了我们如何近似零分布并计算 p值。该统计量和决策是对私有分配进行的后处理。

7.2 声呐数据集上的实验设计

遵循 Deb 和 Sen [17], 中的应用，我们检验金属圆柱体和岩石反射的声呐测量值是否具有相同的分布。UCI 声呐数据集包含 $a = 111$个金属圆柱体观测值和 $b = 97$ 个岩石观测值，每个观测值在 $[0, 1]$[53]中均有 $d = 60$ 个信号能量坐标。我们保留全部 $N = 208$ 个观测值及其原始单位，不进行标准化或特征选择。尽管 Sonar 是一个公开基准数据集，我们仍将每个测量向量视为机密信息；队列标签和名册则是公开的。

我们实现了上述平方代价秩匹配和秩能量检验，使用由 60 个维度中前 208 个 Halton 点组成的固定参考网格。同一网格用于每种方法和每个预算。我们比较 原始点、投影表示、距离向量和 $Gumbel$，并采用平方欧几里得匹配代价。

对于每种机制和 $\varepsilon \in \{1, 2, 4, 8, 16\}$，我们在该固定数据集上评估 1,000 次独立运行。我们使用从零分布中一次性生成、并在所有方法和运行之间共享的 9,999 次独立抽样来计算 p 值。我们报告 1,000 次运行的均值 p值，并以精确的非隐私秩作为基线。

7.3 结果与解释

我们实验中的精确非隐私匹配给出 $T = 7.064$ 和 $p = 0.0022$，拒绝 H_0 在 5% 水平上。图3显示均值 p值，这些值由隐私匹配产生，并以该非隐私结果作为固定基线。

在 $\varepsilon = 1$，2 时，四种方法得到的均值 p-值都接近 0.5，表明强隐私会掩盖两组之间的差异。随着 ε 增大，隐私检验会提供更多信息。Gumbel 的均值 p-值从 $\varepsilon = 4$ 时的 0.4290 降至 $\varepsilon = 8$ 时的 0.2386，以及 $\varepsilon = 16$ 时的 0.0463。在 $\varepsilon = 16$ 时，投影表示的结果为 0.2539，原始点为 0.2924，距离向量为 0.3791。因此，增大隐私预算有助于重新发现分布差异的证据，

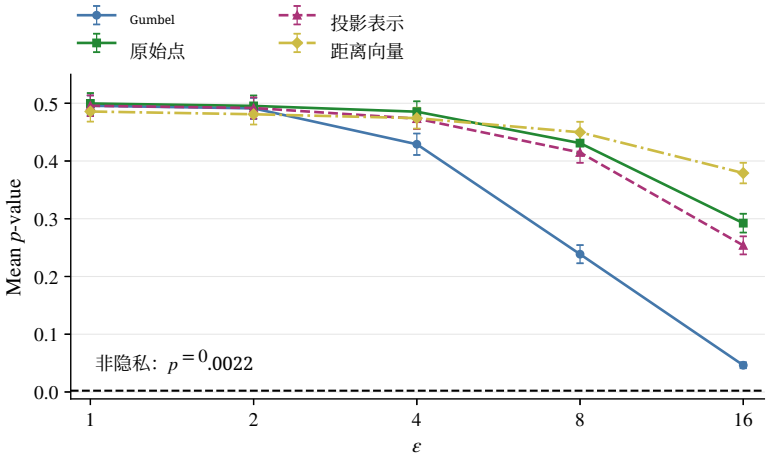

图 3：在声呐数据集上，使用相同的数据集和模拟的原假设分布得到的均值 rank-energy p-值。误差线表示均值的逐点95%置信区间。

在该数据集上，Gumbel 的结果最接近非隐私基线。附录B.2 从攻击者区分声呐数据点的能力角度解读这些隐私预算。

8 讨论

应用的信任模型应指导敏感数据的表示方式和维度选择。在局部度量隐私下，低维嵌入可减少需要添加噪声的坐标数量；公共仿射包则能在保留所有匹配代价的同时实现这种降维。对于可信服务器，Gumbel 匹配可以保留高维距离信息，并达到最优超额代价率。因此， $d \approx \log n$ 为机制选择和嵌入设计提供了指导。高维情形下最优的局部隐私率仍有待研究。

Gumbel 匹配还将度量隐私与潜在排列学习联系起来。Mena 等人 [46]提出了 Gumbel-Sinkhorn，作为通过随机指派进行学习的一种可微松弛。我们的分析揭示了离散 Gumbel 匹配的一项互补性质：控制边代价的变化，即使存在全局约束，也能保护整个分配。这表明，值得研究 Gumbel 构造在其他度量隐私优化问题以及涉及潜在对应关系的学习任务中的应用。若能将这些保证扩展到学习得到的分数和发布的软指派，度量隐私将更接近可微学习。

假设检验应用表明，隐私匹配可以保留对分配本身之外仍有用的信息。将额外匹配成本与统计功效联系起来，有助于把我们的优化保证与下游推断衔接起来。在线匹配提供了另一个方向，尤其适用于工人或任务依次到达且必须立即做出分配的分配问题。研究隐私噪声如何与不确定到达和不可撤销的决策相互作用，可以将本文提出的表示与机制选择扩展到动态分配。

人工智能使用声明

生成式人工智能工具贯穿了本研究及本文的准备过程。其用途包括：识别并总结相关文献；完善研究思路；发展数学论证；起草、检查和修订证明；设计实验；实现、调试和运行实验代码；处理数据；分析结果；准备图、表和可复现性材料；以及编辑手稿的措辞、语法和清晰度。

这些活动均在作者的指导下开展。作者对最终手稿承担全部责任，包括其中理论主张、证明、实验方法、结果以及所有其他由人工智能辅助完成内容的正确性和原创性。

参考文献

[1] Stefan Aeberhard 和 M. Forina。葡萄酒。UCI机器学习仓库，1992年。网址：https://archive.ics.uci.edu/dataset/109/wine。数据集。[2]Mário S. Alvim、Konstantinos Chatzikokolakis、Catuscia Palamidessi 和 Anna Pazii。度量空间上的局部差分隐私：优化效用权衡。载于2018年IEEE第31届计算机安全基础研讨会，第262–267页，2018年。doi：10.1109/CSF.2018.00026。网址：https://conferences.computer.org/sp/pdfs/csf/2018/InvitedPaperLocalDifferentialPrivacyonMetricS.pdf。[3] Balamurugan Anandan 和 Chris Clifton。安全最小权重二分匹配。载于IEEE可信与安全计算会议，2017年。doi：10.1109/DESEC.2017. 8073798。[4]Miguel E. Andrés、Nicolás E. Bordenabe、Konstantinos Chatzikokolakis 和 Catuscia Palamidessi。地理不可区分性：基于位置的系统中的差分隐私。载于ACM SIGSAC计算机与通信安全会议论文集，第901–914页，2013年。网址：https://arxiv.org/abs/1212.1984。[5] Jordan Awan 和 Aleksandra Slavković。差分隐私中的结构与敏感度：比较 K-范数机制。《美国统计学会杂志》，116(534)：935–954，2021年。doi：10.1080/01621459.2020.1773831。网址：https://arxiv.org/abs/1801.09236。[6] Arturs Backurs、Zinan Lin、Sepideh Mahabadi、Sandeep Silwal 和 Jakub Tarnawski。高效计算与私有数据集的相似度。载于国际学习表征会议，2024年。网址：https://arxiv.org/abs/2403.08917。[7] Yuehao Bai。随机对照试验中配对设计的最优性。《美国经济评论》，112(12)：3911–3940，2022。doi：10.1257/aer.20201856。网址：https://arxiv. org/abs/2206.07845。[8] Alexander Benvenuti、Brendan Bialy、Miriam Dennis 和 Matthew Hale。差分隐私线性规划：降低次优性并保证约束满足。AAAI人工智能会议论文集，40(24)：19702–19710，2026年。doi：10.1609/aaai.v40i24.39051。[9] March Boedihardjo、Thomas Strohmer 和 Roman Vershynin。隐私—效用权衡的度量几何。arXiv 预印本 arXiv:2405.00329，2024年。网址：https://arxiv.org/abs/2405. 00329。[10] Nicolás E. Bordenabe、Konstantinos Chatzikokolakis 和 Catuscia Palamidessi。位置隐私的最优地理不可区分机制。载于ACM SIGSAC计算机与通信安全会议论文集，第251–262页，2014。doi：10.1145/2660267. 2660345。[11] Rainer E. Burkard 和 Eranda Çela。线性指派问题及其扩展。载于《组合优化手册》，第75–149页。克吕维尔学术出版社，1999年。doi：10.1007/978-1-4757-3023-4_2。[12] Konstantinos Chatzikokolakis、Miguel E. Andrés、Nicolás Emilio Bordenabe 和 Catuscia Palamidessi。利用度量扩展差分隐私的适用范围。载于隐私增强技术，第82–102页。施普林格，2013年。doi：10.1007/978-3-642-39077-7_5。

[13] Joann Qiongna Chen、Tianhao Wang、Zhikun Zhang、Yang Zhang、Somesh Jha 和 Zhou Li。差分隐私资源分配。载于《年度计算机安全应用会议》，第772–786页，2023年。doi：10.1145/3627106.3627181。[14] Nicolas Courty、Rémi Flamary、Devis Tuia 和 Alain Rakotomamonjy。用于领域自适应的最优传输。《IEEE模式分析与机器智能汇刊》，39(9)：1853–1865，2017年。doi：10.1109/TPAMI.2016.2615921。URL https://arxiv.org/abs/1507.00504。[15] Osman E. Dai、Daniel Cullina 和 Negar Kiyavash。使用高斯特征进行数据库对齐。载于 Kamalika Chaudhuri 和 Masashi Sugiyama（编），《第二十二届国际人工智能与统计会议论文集》，《机器学习研究论文集》第89卷，第3225–3233页。PMLR，2019年4月16–18日。URL https:// proceedings.mlr.press/v89/dai19b.html。[16] Panayiotis Danassis、Aleksei Triastcyn 和 Boi Faltings。在无限大规模场景下用于资源分配的分布式差分隐私算法。载于《第21届国际自主智能体与多智能体系统会议论文集》，第327–335页。国际自主智能体与多智能体系统基金会，2022年。URL https://ifaamas.org/Proceedings/aamas2022/pdfs/p327.pdf。[17] Nabarun Deb 和 Bodhisattva Sen。使用测度传输的基于多元秩的分布无关非参数检验。《美国统计学会杂志》，118(541)：192–207，2023年。doi：10.1080/01621459.2021.1923508。URL https://doi.org/10.1080/ 01621459.2021.1923508。[18] Michael Dinitz、George Z. Li、Quanquan C. Liu 和 Felix Zhou。差分隐私匹配。arXiv预印本 arXiv:2501.00926，2025年。URL https://arxiv.org/abs/2501.00926。[19] Leilei Du、Peng Cheng、Libin Zheng、Wei Xi、Xuemin Lin、Wenjie Zhang 和 Jing Fang。差分隐私下的动态隐私任务分配。载于《IEEE第39届国际数据工程会议》，第2740–2752页，2023年。URL https://arxiv.org/abs/2302.09511。[20] Cynthia Dwork 和 Aaron Roth。《差分隐私的算法基础》。Now出版社，2014年。URL https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf。[21] Alina Ene、Huy Le Nguyen、Ta Duy Nguyen 和 Adrian Vladu。使用差分隐私求解线性规划。arXiv预印本 arXiv:2507.10946，2025年。URL https://arxiv.org/abs/ 2507.10946。[22] Matthew Fahrbach、Zhiyi Huang、Runzhou Tao 和 Morteza Zadimoghaddam。边加权在线二分图匹配。《美国计算机学会期刊》，69(6)：1–35，2022年。[23] Ronald A. Fisher。在分类学问题中使用多重测量。《优生学年鉴》，7(2)：179–188，1936年。doi：10.1111/j.1469-1809.1936.tb02137.x。[24] Marek Gagolewski。聚类算法基准测试框架。SoftwareX，20：101270，2022年。doi：10.1016/j.softx.2022.101270。URL https://clustering-benchmarks. gagolewski.com/weave/suite-v1.html。[25] Marek Gagolewski 等。聚类算法基准套件。1.2.0版，2026年。URL https://github.com/gagolews/clustering-data-v1/releases/tag/v1.2.0。数据集 sna

p镜头。

[26] B. German。玻璃识别。UCI机器学习仓库，1987年。网址：https: // archive.ics.uci.edu/dataset/42/glass+识别。数据集。[27] Ian Goodfellow、Yoshua Bengio 和 Aaron Courville。深度学习。麻省理工学院出版社，2016年。网址：https://www.deeplearningbook.org/。[28] Daniel Graves 和 Witold Pedrycz。基于核的模糊聚类与模糊聚类：一项比较性实验研究。《模糊集与系统》，161(4)：522–543，2010年。doi：10.1016/ j.fss.2009.10.021。[29] Moritz Hardt 和 Kunal Talwar。差分隐私的几何学。载于《第四十二届ACM计算理论研讨会论文集》，第705–714页，2010年。网址：https://arxiv.org/abs/0907.3754。[30] Justin Hsu、Zhiyi Huang、Aaron Roth、Tim Roughgarden 和 Zhiwei Steven Wu。隐私匹配与分配。载于《第四十六届ACM计算理论研讨会论文集》，第21–30页，2014年。doi：10.1145/2591796.2591826。网址：https:// arxiv.org/abs/1311.2828。[31] Justin Hsu、Aaron Roth、Tim Roughgarden 和 Jonathan Ullman。以隐私方式求解线性规划。载于《自动机、语言与程序设计》，第612–624页。施普林格，2014年。doi：10.1007/978-3-662-43948-7_51。网址：https://arxiv.org/abs/1402.3631。[32] Justin Hsu、Zhiyi Huang、Aaron Roth 和 Zhiwei Steven Wu。联合差分隐私凸规划。载于《第二十七届ACM-SIAM离散算法研讨会论文集》，第580–599页，2016年。doi：10.1137/1.9781611974331.ch43。网址：https: //arxiv.org/abs/1411.0998。[33] Zhiyi Huang 和 Aaron Roth。利用度量结构实现高效的隐私查询发布。载于《第二十五届ACM-SIAM离散算法研讨会论文集》，第523–534页，2014年。doi：10.1137/1.9781611973402.39。网址：https://arxiv.org/abs/1211.7302。[34] Zhiyi Huang 和 Xue Zhu。可扩展的联合差分隐私装箱。载于《第46届自动机、语言与程序设计国际研讨会》，《信息学领域莱布尼茨国际会议论文集》第132卷，第73:1–73:12页。达格斯图尔城堡—莱布尼茨信息学中心，2019年。doi：10.4230/LIPIcs.ICALP.2019.73。网址：https://drops.dagstuhl.de/ entities/document/10.4230/LIPIcs.ICALP.2019.73。[35]Jacob Imola、Shiva Kasiviswanathan、Stephen White、Abhinav Aggarwal 和 Nathanael Teissier。在度量差分隐私中平衡效用与可扩展性。载于《第三十八届人工智能不确定性会议论文集》，《机器学习研究论文集》第180卷，第885–894页。PMLR，2022年。网址：https://proceedings. mlr.press/v180/imola22a.html。[36] Jacob Imola、Amrita Roy Chowdhury 和 Kamalika Chaudhuri。通过推土机距离实现用户级度量差分隐私。载于《ACM SIGSAC计算机与通信安全会议论文集》，第348–362页，2024年。doi：10.1145/3658644.3690363。[37] Jacob Imola、Rasmus Pagh 和 Lukas Retschmeier。利用重构攻击研究隐私图优化问题的下界。arXiv预印本，arXiv:2609.10877，2026年。网址：https://arxiv.org/abs/2609.10877。

[38]Krishnaram Kenthapadi、Aleksandra Korolova、Ilya Mironov 和 Nina Mishra。通过 Johnson–Lindenstrauss 变换实现隐私保护。《隐私与保密期刊》，5(1):39–71，2013年。URL https://cs.stanford.edu/people/kngk/papers/privacyViaJohnsonLindenstraussTransform-JPC2013.pdf。[39] Fragkiskos Koufogiannis、Shuo Han 和 George J. Pappas。差分隐私中拉普拉斯机制的最优性，2015年。URL https://arxiv.org/abs/1504.00065。[40] Yann LeCun、Léon Bottou、Yoshua Bengio 和 Patrick Haffner。应用于文档识别的基于梯度的学习。《IEEE 会议录》，86(11):2278–2324，1998年。doi: 10. 1109/5.726791。[41] Nam LeTien、Amaury Habrard 和 Marc Sebban。差分隐私最优传输：应用于领域自适应。载于《第二十八届国际人工智能联合会议论文集》，第2852–2858页，2019。doi: 10.24963/ijcai.2019/395。URL https://www.ijcai.org/proceedings/2019/395。[42] Maocheng Li、Jiachuan Wang、Libin Zheng、Han Wu、Peng Cheng、Lei Chen 和 Xuemin Lin。具有不可信服务器的空间众包中隐私保护的批量任务分配。载于《第三十届 ACM 信息与知识管理国际会议论文集》，第947–956页，2021年。doi: 10.1145/3459637.3482288。[43] Svante Linusson 和 Johan Wästlund。随机指派问题中 Parisi 猜想的证明。《概率论及其相关领域》，128:419–440，2004年。URL https://arxiv.org/ abs/math/0303214。[44] Volker Lohweg。纸币认证。UCI机器学习仓库，2012年。URL https://archive.ics.uci.edu/dataset/267/banknote+authentication。数据集。[45] Chaojie Lv、Lan Zhang 和 Xiang-Yang Li。空间众包中个性化差分隐私在线最小二分匹配。载于第八届大数据计算与通信国际会议，第134–143页，2022。doi: 10.1109/BIGCOM57025.2022. 00025。[46] Gonzalo E. Mena、David Belanger、Scott Linderman 和 Jasper Snoek。利用 Gumbel-Sinkhorn 网络学习潜在排列。载于国际学习表征会议，2018年。URL https://arxiv.org/abs/1802.08665。[47] Francesco Mezzadri。如何从经典紧群生成随机矩阵。《美国数学会通报》，54(5):592–604，2007年。URL https://arxiv.org/abs/math-ph/0609050。[48] Kenta Nakai。酵母。UCI机器学习仓库，1991年。URL https://archive.ics. uci.edu/dataset/110/yeast。数据集。[49] Kenta Nakai。大肠杆菌。UCI机器学习仓库，1996年。URL https://archive.ics.uci.edu/dataset/39/ecoli。数据集。[50] S hyam Narayanan 和 Jelani Nelson。欧几里得空间中的最优终端维数缩减。载于第五十一届 ACM SIGACT 理论计算年度研讨会论文集，第1064–1069页，2019年。URL https://arxiv.org/abs/1810.09250。

[51] Amanda Prorok 和 Vijay Kumar。面向按需出行系统的隐私保护车辆分配。载于*IEEE/RSJ*智能机器人与系统国际会议，第1869–1876页，2017年。doi: 10.1109/IROS.2017.8206003。URLhttps://arxiv.org/abs/1703.04738。

[52] Adam Sealfon。差分隐私下的最短路径与距离。载于 第三十五届 *ACM SIGMOD-SIGACT-SIGAI*数据库系统原理研讨会论文集，第 29–41 页，2016 年。doi：10.1145/2902251.2902291。网址：https://arxiv.org/abs/1511.04631。

[53] Terry Sejnowski 和 R. Gorman。连接主义基准（声呐、水雷与岩石）。UCI机器学习仓库，1988 年。网址：https://doi.org/10.24432/C5T01Q。数据集。

[54]Changho Shin、Jitian Zhao、Sonia Cromp、Harit Vishwakarma和Frederic Sala。OTTER：零样本模型的轻松标签分布自适应。《神经信息处理系统进展》，37：44064–44101，2024。doi: 10.52202/ 079017-1399。URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ 4df1cc5a7528b7197ad8ae76ff30107a-Abstract-Conference.html。[55] Sami Sieranoja和Pasi Fränti。快速且通用的密度峰值聚类。《模式识别快报》，128：551–558，2019。doi: 10.1016/j.patrec.2019.10.019。[56] V. Sigillito、S. Wing、L. Hutton和K. Baker。电离层。UCI机器学习仓库，1989年。URL https://archive.ics.uci.edu/dataset/52/ionosphere。数据集。[57] Qian Tao、Yongxin Tong、Zimu Zhou、Yexuan Shi、Lei Chen和Ke Xu。空间众包中的差分隐私在线任务分配：一种基于树的方法。见IEEE第36届数据工程国际会议，第517–528页，2020年。doi: 10.1109/ICDE48307. 2020.00051。[58] Hien To、Gabriel Ghinita和Cyrus Shahabi。空间众包中保护工人位置隐私的框架。《VLDB基金会论文集》，7(10)：919–930，2014。URL https://www.vldb.org/pvldb/vol7/p919-to.pdf。[59]Aleksander Truszczyński。圆，2019。URL https://github.com/gagolews/ clustering-data-v1/blob/v1.2.0/wut/circles.txt。华沙理工大学基准数据集。[60]UCI机器学习仓库。Statlog（图像分割），1990年。URL https: // archive.ics.uci.edu/dataset/147/statlog+image+segmentation。数据集。[61] Alfred Ultsch。使用自组织映射进行聚类：U*C。见《自组织映射研讨会论文集》，第75–82页，2005年。[62] Hengzhi Wang、En Wang、Yongjian Yang、Jie Wu和Falko Dressler。空间众包中的隐私保护在线任务分配：一种基于图的方法。见IEEE INFOCOM 2022——IEEE计算机通信会议，第570–579页，2022。doi: 10.1109/ INFOCOM48880.2022.9796827。[63] Jiaqi Wang和Ziv Goldfeld。差分隐私Sinkhorn算法。见第60届Allerton通信、控制与计算会议年会，2024。doi: 10.1109/ ALLERTON63246.2024.10735319。URL https://par.nsf.gov/servlets/purl/10577813。

[64] William H. Wolberg、W. Nick Street和Olvi L. Mangasarian。威斯康星乳腺癌（诊断）。UCI机器学习仓库，1993年。URLhttps://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic。数据集。[65] Han Xiao、Kashif Rasul和Roland Vollgraf。Fashion-MNIST：用于基准测试机器学习算法的新型图像数据集。arXiv:1708.07747，2017年。URL https://arxiv.org/abs/1708.07747。[66] Xinpeng Xie、Chenyang Yu、Yan Huang、Yang Cao和Chenxi Qiu。SoK：理论与实践中的度量差分隐私。《隐私增强技术会议论文集》，2026(4)：415–437，2026年。doi: 10.56553/popets-2026-0128。URLhttps://petsymposium.org/popets/2026/popets-2026-0128.php。[67] Yifan Xu、Jun Tao、Xinyu Chen、Jun Yan和Jun Shen。高效的隐私保护拼车：一种基于在线匹配的方法。《*IEEE*信息取证与安全汇刊》，21：5485–5498，2026年。doi: 10.1109/TIFS.2026.3698576。[68]Pengfei Zhang、Xiang Cheng、Sen Su和Ning Wang。通过基于群组的噪声添加实现地理不可区分性下的任务分配。《*IEEE*大数据汇刊》，9(3)：860–877，2023年。doi: 10.1109/TBDATA.2022.3215467。[69] Zongrui Zou、Alessandro Epasto、Chenglin Fan和Rudrajit Das。欧几里得图中的差分隐私最小生成树。见《第二十九届人工智能与统计国际会议论文集》，第300卷出自《机器学习研究论文集》，第379–387页。PMLR，2026年。URLhttps://proceedings.mlr.press/v300/zou26a.html。

附录目录

A 相关工作 **21**
A.1 度量隐私 .21A.2 隐私优化与匹配.21A.3 私有几何表示与距离查询 .22A.4 位置隐私分配与空间众包 .22

B 差分隐私与隐私校准 **24**B.1 差分隐私及其含义 ε .24B.2 在实验数据集上的校准 .24

C 结果证明 **25**
C.1 匹配稳定性 .25C.2 原始点 .26C.3 距离向量 .26C.4 投影表示 .27C.5 Gumbel 匹配.27C.6 隐私下界.29

D 对其他情形的推广 **30**D.1 平方欧几里得代价的上界 .30D.2 两个数据集均为私有时的匹配 .32

E 数值实验的详细内容 数值实验 **34**
E.1 数据集与特征处理 .34E.2 距离尺度、投影链与嵌套样本.35E.3 机制校准与重复运行 .36E.4 数值 .36E.5 隐私预算扫描.41

F 私有秩能量实验的详细内容 **46**F.1 参考网格和零分布 .46

A 相关工作

A.1 度量隐私

度量隐私根据输入之间的距离来界定其可区分性，在保护精确值的同时允许披露较粗略的信息。Chatzikokolakis 等人 [12] 提出了通用框架，而 Andrés 等人 [4] 则为基于位置的系统发展了地理不可区分性。关于度量隐私的发展、其机制类别及其应用的更全面介绍，请读者参阅 Xie 等人撰写的知识体系化综述。 [66]。

后续研究探讨了机制如何优化隐私与效用之间的权衡。Bordenabe 等人 [10]设计了满足地理不可区分性的地理位置发布机制，Imola 等人 [35] 则研究了如何大规模构建实用的度量隐私机制所面临的计算挑战。更一般地说，几何方法将隐私—效用权衡与发布统计数据及输入度量的几何性质联系起来 [9, 29]。输入度量的选择也决定了哪些信息受到保护：例如，Imola 等人 [36] 使用推土机距离，为记录集合提出了用户级度量隐私的定义。将运输用作输入隐私度量，与最小化运输目标或匹配目标截然不同。Alvim 等人 [2]研究了局部度量隐私的效用；Koufogiannis 等人 [39] 分析了在利普希茨隐私下用于数值重构的加性机制。后一种准则不同于本文研究的分配损失。

A.2 隐私优化与匹配

隐私匹配属于针对敏感问题数据进行优化这一更广泛的研究范畴。Hsu 等人 [31] 根据敏感信息出现在目标中还是约束条件中，对各类隐私线性规划加以区分，并为这些不同表述提供了算法与不可能性结果。近期工作进一步研究隐私线性规划可行性，以及约束满足与优化精度之间的权衡 [8, 21]。这些区分与分配直接相关：当匹配约束公开时，其线性规划表述具有固定的可行域，而隐私位置决定目标系数。这些系数之间的几何依赖提供了超出一般隐私目标向量的额外结构。

另一条研究路线考虑不同代理接收解的不同部分的优化问题。Hsu 等人 [30]研究联合差分隐私下的隐私匹配与分配，而 Hsu 等人 [32] 将这一视角扩展到带有共享资源约束的一大类凸规划。联合隐私保护某一代理的信息，使其不因其他代理所接收的输出而泄露，同时允许该代理自身的分配依赖于其输入。Danassis 等人 [16]在分段局部隐私表述下发展了去中心化资源分配。资源分配隐私也会通过对共享资源的可观察的竞争而产生，Chen 等人 [13]对此进行了研究。这些工作共同凸显了同时明确敏感信息以及观察者对解的访问权限的重要性。

匹配特定结果涉及若干其他隐私模型。Dinitz 等人 [18] 研究了边隐私和节点隐私下的匹配，并明确区分了显式发布的匹配与隐式表示。对于最小权完美匹配，Sealfon [52, 附录 B.2] 给出了拓扑公开、边权重私有且满足 ℓ_1 邻接关系时的算法和下界。Imola 等人 [37] 在这一设定下加强了基于重构的下界。这些结果为优化问题提供了直接先例，尽管边权重的任意变化不同于移动一个私有点所引起的匹配成本协调变化。Anandan 和 Clifton [3]

将安全的二分图匹配计算与数值匹配成本的差分隐私相结合，说明了保护最优值与保护已发布分配之间的区别。

私有最优传输与此密切相关，因为对于两个大小相同且权重相等的点集，未正则化的离散传输存在一个最优置换耦合；其目标函数等于相应最小匹配成本除以点的数量。LeTien 等人[41]将私有投影表示和重构的传输成本用于领域自适应。Wang 和 Goldfeld [63]提出了一种用于熵正则化传输的私有 Sinkhorn 过程，并进一步提出随机化传输成本发布。这些方法将私有几何表示与约束分配联系起来，但它们的隐私模型，以及后一种方法中的正则化目标函数，都不同于度量隐私下采用未平方欧几里得代价的匹配。

A.3 私有几何表示与距离查询 s

私有表示可以保留下游任务所需的信息，而无须准确重构每个原始坐标。Kenthapadi 等人 [38] 将 Johnson–Lindenstrauss 投影与高斯扰动相结合，并校正由此产生的平方距离估计偏差。Huang 和 Roth [33]利用度量结构发布私有距离查询。最近，Backurs 等人 [6] 开发了用于计算与数据集的聚合相似度和距离的私有数据结构。聚合查询答案与点特定分配成本的用途不同，但这些研究共同揭示了几何结构、私有表示和下游效用之间的联系。私有投影表示也已用于欧几里得最小生成树 [69], ，这为几何结构如何助力私有优化提供了另一个实例。

公共参考点提供了另一种结构来源。终端维数缩减能够保留指定终端点到其他点的距离。本研究采用的精确投影加正交范数表示见 Narayanan 和 Nelson [50, 第 1 节1]。因此，维数缩减以及根据私有草图重建距离均有先例可循。我们的投影表示机制在这种保距表示中加入校准噪声，并对由此产生的额外匹配成本进行界定。

经典随机分配研究随机边权重下的最优匹配成本。Linusson 和 Wästlund [43] 证明了独立指数成本方阵的精确期望最优值。在私有优化领域，Hardt 和 Talwar [29] 利用范数球的几何性质构造噪声分布，并确立隐私–效用界。

A.4 位置隐私分配与空间众包

空间众包领域关于敏感位置分配问题的应用研究十分丰富。早期研究利用私有空间分解支持任务分发，同时保护工人群体 [58]。Prorok 和 Kumar [51] 研究了面向按需出行系统的位置隐私车辆分配。基于树的表示支持在线任务分配 [57]；四叉树泛化支持个性化隐私 [45]；基于图的方法则支持在位置扰动后进行在线分配 [62]。Li 等人 [42]研究了由不可信服务器执行的批量任务分配。Du 等人 [19] 将不断变化的隐私选择和补偿纳入任务分配。Zhang 等人 [68] 提出 CANOE，通过对工人偏好的任务进行分组，并收集工人真实位置到扰动后的群组代表点的距离，为基于混淆距离信息支持分配提供了一种密切相关的方法。

这些表述在到达假设、容量、披露和效用目标方面有所不同。例如，Xu 等人 [67] 研究具有已知分布的到达与司机匹配的在线拼车

预算，并考虑匹配数量、利润和取货距离。隐私度量同样重要：Lv 等人的机制 [45] 使用选定层级上编号四叉树节点之间的差异，而非连续欧氏位移。该文献推动了私有分配，并提供了面向具体应用的机制。当前表述将发布最小总距离完美匹配这一优化问题单独分离出来，其中包含公开分配约束和加性效用保证，且不要求隐私位置有界或具有指定的分布。

B 差分隐私与隐私校准

B.1 差分隐私与 ε 的含义

在差分隐私中，两个大小相同的数据集若有一条记录不同，则互为邻居，无论该差异的幅度如何。机制 M_Y 满足 ε-差分隐私，如果对于每一对这样的 X， X' 以及每一个可测事件 A，

$$\Pr[M_Y(X) \in A] \le e^{\varepsilon} \Pr[M_Y(X') \in A].$$

这是差分隐私的替换邻居定义 [20]。将汉明距离定义为

$$d_H(X, X') = \sum_{i=1}^{n} \mathbf{1}_{\{\boldsymbol{x}_i \neq \boldsymbol{x}'_i\}}.$$

用 d_H 替换 D 于（3）可恢复这一概念：邻居的汉明距离为一，并且一次替换一条不同记录可为任意对 [12]。

在我们的欧几里得度量 D 下，转而假设邻居仅在一条记录上相差的距离至多为 1。则 $D(X, X') \le 1$，因此度量隐私蕴含该邻域关系下的 ε-差分隐私。更一般地，选取距离 $r > 0$ 以及容差因子 $c > 1$：对于距离至多为 r 的输入，任何输出事件的概率变化至多应为因子 c。遵循 Andrés 等人 [4], 对度量隐私的解释，令

$$\varepsilon = \frac{\log c}{r}.$$

事实上，只要 $D(X, X') \le r$，

$$\Pr[M_Y(X) \in A] \le e^{\varepsilon r} \Pr[M_Y(X') \in A] = c \Pr[M_Y(X') \in A].$$

对于一条记录的变更， r 是该点的欧几里得位移。

半径 r 以数据的单位度量，无论是物理单位还是潜在表示中的距离；而 ε 具有倒数单位，因此 εr 是无量纲的。例如，从米改为厘米会使数值距离乘以 100，并使 ε 除以 100，从而保持相同的隐私保证。同样，在潜在空间中进行校准取决于其坐标的尺度。

B.2 实验数据集上的校准

设有一名攻击者，知道所有私有数据点及其身份，但不知道两名个体分别对应两个已知点 $\boldsymbol{x}$, $\boldsymbol{x}' \in \mathbb{R}^d$中的哪一个。攻击者认为有两个数据集 X_0, X_1 的可能性相同，它们仅在这两个点互换这一点上有所不同。观察到一次匹配（包括分配给这两名个体的点）后，攻击者猜测所使用的是哪个数据集。观察非隐私匹配时，攻击准确率为 100%；而观察均匀随机匹配时，准确率为 50%。

写出 $r = \|\boldsymbol{x} - \boldsymbol{x}'\|_2$。在该交换下，两个点都会改变，因此 $D(X_0, X_1) = 2r$。根据度量隐私的定义（3），对于每一个匹配 σ （具有正概率），

$$e^{-2\varepsilon r} \le \frac{\Pr[M_Y(X_0) = \sigma]}{\Pr[M_Y(X_1) = \sigma]} \le e^{2\varepsilon r}.$$

在先验概率相等的情况下，该比值也是观察到 σ 后的后验优势比。因此，即使是最优攻击者，其

$$\Pr[\text{correct guess}] \le \frac{e^{2\varepsilon r}}{1 + e^{2\varepsilon r}}. \tag{16}$$

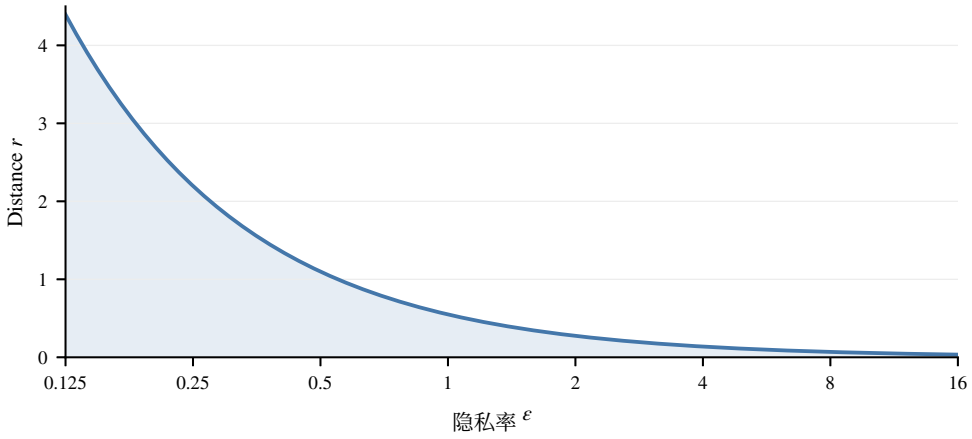

图4：阴影区域满足 $r \le \log 3/(2\varepsilon)$，从而保证在 εD-隐私匹配上的识别准确率至多为 75%。

因此，要将识别准确率限制在最高 75%，而发布前为 50%，只需选择

$$r \le \frac{\log 3}{2\varepsilon}, \qquad \text{equivalently} \qquad \varepsilon \le \frac{\log 3}{2r}, \tag{17}$$

如图4所示。

针对声呐的推断攻击。该假设检验实验汇集了 111 个金属圆柱体观测值和 97 个岩石观测值，从而在 $[0,1]$ 中得到具有 $d = 60$ 个坐标的 $N = 208$ 个数据点。我们使用源数据集中的全部观测值，不再进行进一步标准化。不同点之间欧几里得距离的最小值、中位数和最大值分别为 0.1746、1.7786 和 3.5295。这些描述的是所观测的数据集；总体中可能的数据点也可以位于已观测点之间。

在 $\varepsilon = 16$ 时，εD-隐私匹配保证：对于相距至多 $r = 0.0343$ 的点之间的交换，准确率至多为 75%。该距离低于观测到的最小距离，因此在此攻击模型下 $\varepsilon = 16$ 较大。若要在观测距离的中位数处获得同样的保证，则需要 $\varepsilon \le 0.309$。应当结合攻击者的广泛知识来看待这一弱点：候选点、所有其他个体的数据以及隐私机制本身均已已知，仅剩下双向身份歧义。在攻击者侧信息较少、或仅观测到检验统计量的应用中，这种推断将困难得多。因此，面对信息较少的对手，$\varepsilon = 16$ 仍可作为一种有用的选择，既能限制其区分相似数据点的能力，又能为双样本检验保留证据。

C 结果的证明

C.1 匹配稳定性

引理 7（匹配稳定性）。令 $\boldsymbol{C} \in \mathbb{R}^{n\times n}$，并令 $\boldsymbol{C} = \boldsymbol{C} + \boldsymbol{E}$。假设 $\widehat{\sigma}$ 使 $\boldsymbol{C}$的匹配成本最小。记 $u(\sigma) = \sum_i C_{i,\,\sigma(i)}$ 且 $\mathrm{OPT} = \min_{\sigma\in S_n} u(\sigma)$，则有

$$0 \le \mathrm{err}(\widehat{\sigma}) \le \sum_{i=1}^{n} \big(\max_j E_{ij} - \min_j E_{ij}\big). \tag{18}$$

使用列范围之和的相应界也成立。

^证明。 令 $\sigma^\star$ 使真实成本最小。根据 $\boldsymbol{C}$ 的最优性，有

$$u(\widehat{\sigma}) + \sum_i E_{i,\widehat{\sigma}(i)} \le u(\sigma^\star) + \sum_i E_{i,\sigma^\star(i)}.$$

因此，

$$\mathrm{err}(\widehat{\sigma}) \le \sum_i \big(E_{i,\sigma^\star(i)} - E_{i,\widehat{\sigma}(i)}\big) \le \sum_i \big(\max_j E_{ij} - \min_j E_{ij}\big).$$

这两个排列也都恰好从每一列中选取一个条目，因此同一差值的上界为各列范围之和。根据 OPT 的定义可知其非负。 □

特别地，行偏移或列偏移会给每个可行匹配成本都加上相同的常数，并且不会改变任何最优解或超额成本。

C.2 原始点

定理 *2*的证明。 拉普拉斯机制为每个报告提供局部 ε-度量隐私。并行组合与后处理为输出匹配提供 εD-隐私。

对于效用，成本误差为

$$E_{ij} = \|\boldsymbol{x}_i + \mathbf{Z}_i - \boldsymbol{y}_j\|_2 - \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2.$$

反三角不等式给出 $|E_{ij}| \le \|\mathbf{Z}_i\|_2$，因此第 i 行中的误差范围至多为 $2\|\mathbf{Z}_i\|_2$。因此，引理7 表明，对于每个噪声实现，

$$\mathrm{err}(\widehat{\sigma}) \le 2\sum_{i=1}^n \|\mathbf{Z}_i\|_2.$$

由于 $\mathbb{E}\|\mathbf{Z}_i\|_2 = d/\varepsilon$，取期望可得

$$\mathbb{E}[\mathrm{err}(\widehat{\sigma})] \le 2\sum_{i=1}^n \mathbb{E}\|\mathbf{Z}_i\|_2 = \frac{2nd}{\varepsilon}.$$ □

C.3 距离向量

定理 *3* 的证明。 对于每个公开候选点 $\boldsymbol{y}_j$，反三角不等式给出

$$\big|\|\boldsymbol{x} - \boldsymbol{y}_j\|_2 - \|\boldsymbol{x}' - \boldsymbol{y}_j\|_2\big| \le \|\boldsymbol{x} - \boldsymbol{x}'\|_2.$$

因此， Φ_Y 是 $n^{1/p}$-利普希茨映射到 ℓ_p:

$$\|\Phi_Y(\boldsymbol{x}) - \Phi_Y(\boldsymbol{x}')\|_p \le n^{1/p}\|\boldsymbol{x} - \boldsymbol{x}'\|_2.$$

因此，所选噪声可提供局部 ε-度量隐私。并行组合和后处理为输出匹配提供 εD-隐私。

设 $Z_{ij} = [\mathbf{Z}_i]_j$。第 i 行的成本误差恰好为 $\mathbf{Z}_i$，因此引理 7 给出

$$\mathrm{err}(\widehat{\sigma}) \le \sum_{i=1}^n \big(\max_j Z_{ij} - \min_j Z_{ij}\big) \le 2\sum_{i=1}^n \|\mathbf{Z}_i\|_\infty \le 2\sum_{i=1}^n \|\mathbf{Z}_i\|_p.$$

The ℓ_p 噪声幅度满足

$$\|\mathbf{Z}_i\|_p \sim \mathrm{Gamma}\Big(n, \mathrm{rate}\ \frac{\varepsilon}{n^{1/p}}\Big), \qquad \mathbb{E}\|\mathbf{Z}_i\|_p = \frac{n^{1+1/p}}{\varepsilon}.$$

由于噪声向量同分布，取期望可得

$$\mathbb{E}[\mathrm{err}(\widehat{\sigma})] \le 2n\,\mathbb{E}\|\mathbf{Z}_1\|_p = \frac{2n^{2+1/p}}{\varepsilon}.$$

由于上界在 $p=\infty$, $\mathbb{E}\|\mathbf{Z}_1\|_\infty = n/\varepsilon$ 处取得最小值，从而得到 $2n^2/\varepsilon$。 □

C.4 投影表示

定理证明*4*. 当 $s=d$，这是原点机制。对于 $s<d$，正交性与反三角不等式给出

$$\begin{aligned}\|f_Y(\boldsymbol{x}) - f_Y(\boldsymbol{x}')\|_2^2 &= \|P_W(\boldsymbol{x}-\boldsymbol{x}')\|_2^2 + \big(\|P_{W^\perp}\boldsymbol{x}\|_2 - \|P_{W^\perp}\boldsymbol{x}'\|_2\big)^2 \\ &\le \|P_W(\boldsymbol{x}-\boldsymbol{x}')\|_2^2 + \|P_{W^\perp}(\boldsymbol{x}-\boldsymbol{x}')\|_2^2 = \|\boldsymbol{x}-\boldsymbol{x}'\|_2^2.\end{aligned}$$

因此，添加拉普拉斯噪声满足局部 ε-度量隐私，而组合与后处理使匹配满足 εD-隐私。

距离保持与反三角不等式给出

$$|E_{ij}| = |\widehat{C}_{ij} - C_{ij}| \le \|\mathbf{Z}_i\|_2.$$

应用引理 7 并取期望可得

$$\mathbb{E}[\mathrm{err}(\widehat{\sigma})] \le 2\sum_{i=1}^{n}\mathbb{E}\|\mathbf{Z}_i\|_2 = \frac{2ns}{\varepsilon}.$$

最后， $\boldsymbol{y}_1 = 0$ 给出 $r \le n-1$，因此 $s = \min\{r+1, d\} \le \min\{n, d\}$。 □

C.5 Gumbel 匹配

我们证明在行内成本变化满足某个条件时 $n \times n$ 匹配的界限。这涵盖欧几里得代价和平方欧几里得代价。

定理 8。 假设每个代价均具有 $C_{ij} = C_{ij}(\boldsymbol{x}_i)$ 的形式，并且对于某个 $L>0$，满足

$$\big|\big(C_{ij}(\boldsymbol{x}) - C_{ij}(\boldsymbol{x}')\big) - \big(C_{ik}(\boldsymbol{x}) - C_{ik}(\boldsymbol{x}')\big)\big| \le L\|\boldsymbol{x}-\boldsymbol{x}'\|_2 \tag{19}$$

对于所有 i、j、$k \in [n]$ 以及 $\boldsymbol{x}$、$\boldsymbol{x}' \in \mathbb{R}^d$。对于相互独立的 $G_{ij} \sim \mathrm{Gumbel}(0、1)$，该匹配

$$\widehat{\sigma} \in \arg\max_{\sigma \in S_n} \sum_{i=1}^{n}\left(-C_{i,\sigma(i)} + \frac{L}{\varepsilon}G_{i,\sigma(i)}\right)$$

满足 εD-隐私且

$$\mathbb{E}[\mathrm{err}(\widehat{\sigma})] \le \frac{nL\log n}{\varepsilon}.$$

证明。设 $\gamma = L/\varepsilon$。首先假设 X, X' 仅在第 i 行不同。对该行之外的所有噪声取条件，记为 G_{-i}。对每个 $j \in [n]$，令

$$B_i(j) = \max_{\sigma \in S_n : \sigma(i)=j} \sum_{k \neq i} \bigl(-C_{k,\sigma(k)} + \gamma G_{k,\sigma(k)}\bigr),$$

并令 $\sigma^{(j)}$ 表示相应的最大化排列，并对任何平局采用固定规则进行处理。$B_i(j)$ 和 $\sigma^{(j)}$ 仅取决于不等于 i 和 G_{-i} 的行，因此对于 X 和 X'，它们相同。

令 $J = \widehat{\sigma}(i)$。则 $\widehat{\sigma} = \sigma^{(J)}$，且

$$J = \arg\max_{j \in [n]} \bigl(-C_{ij}(\boldsymbol{x}_i) + B_i(j) + \gamma G_{ij}\bigr).$$

定义

$$p_j(\boldsymbol{x}_i) = \Pr[J = j \mid G_{-i}].$$

Gumbel-max 技巧 [46] 给出条件概率

$$p_j(\boldsymbol{x}_i) = \frac{\exp\bigl((-C_{ij}(\boldsymbol{x}_i) + B_i(j))/\gamma\bigr)}{\sum_{k=1}^{n} \exp\bigl((-C_{ik}(\boldsymbol{x}_i) + B_i(k))/\gamma\bigr)}.$$

令 $\delta_j = C_{ij}(\boldsymbol{x}'_i) - C_{ij}(\boldsymbol{x}_i)$。条件（19）将它们的取值范围限制在 $L\|\boldsymbol{x}_i - \boldsymbol{x}'_i\|_2$。因此

$$\frac{p_j(\boldsymbol{x}_i)}{p_j(\boldsymbol{x}'_i)} = \frac{e^{\delta_j/\gamma}}{\sum_{k=1}^{n} p_k(\boldsymbol{x}'_i) e^{\delta_k/\gamma}} \leq \exp\left(\frac{\max_k \delta_k - \min_k \delta_k}{\gamma}\right) \leq e^{\varepsilon \|\boldsymbol{x}_i - \boldsymbol{x}'_i\|_2}.$$

对于任意事件 $A \subseteq S_n$，输出 $\sigma^{(J)}$ 属于 A，这意味着

$$\begin{aligned}
\Pr[M_Y(X) \in A \mid G_{-i}] &= \sum_{j:\, \sigma^{(j)} \in A} p_j(\boldsymbol{x}_i) \\
&\leq e^{\varepsilon \|\boldsymbol{x}_i - \boldsymbol{x}'_i\|_2} \sum_{j:\, \sigma^{(j)} \in A} p_j(\boldsymbol{x}'_i) \\
&= e^{\varepsilon \|\boldsymbol{x}_i - \boldsymbol{x}'_i\|_2} \Pr[M_Y(X') \in A \mid G_{-i}].
\end{aligned}$$

由于 G_{-i} 与 X, X' 相互独立，对两边关于它积分可得

$$\Pr[M_Y(X) \in A] \leq e^{\varepsilon \|\boldsymbol{x}_i - \boldsymbol{x}'_i\|_2} \Pr[M_Y(X') \in A].$$

对于任意 X 和 X'，逐条替换其记录可得

$$\Pr[M_Y(X) \in A] \leq \exp\left(\varepsilon \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{x}'_i\|_2\right) \Pr[M_Y(X') \in A] = e^{\varepsilon D(X,X')} \Pr[M_Y(X') \in A],$$

从而确立了 εD-隐私。

对于效用，固定一个与噪声无关的最小代价匹配 $\sigma^\star$。扰动后分数的最优性给出

$$\mathrm{err}(\widehat{\sigma}) \leq \gamma \sum_{i=1}^{n} \bigl(G_{i,\widehat{\sigma}(i)} - G_{i,\sigma^\star(i)}\bigr).$$

标准的 Gumbel 累积分布函数表明，$\max_{j \in [n]} G_{ij}$ 与 $G_{i1} + \log n$ 具有相同的分布。因此，

$$\mathbb{E}[\mathrm{err}(\widehat{\sigma})] \leq \gamma \sum_{i=1}^{n} \left(\mathbb{E} \max_{j \in [n]} G_{ij} - \mathbb{E} G_{i,\sigma^\star(i)}\right) = \gamma n \log n = \frac{Ln \log n}{\varepsilon}. \qquad \square$$

定理 5的证明。 对于欧几里得代价，反三角不等式给出 s

$$\big|\|\boldsymbol{x}-\boldsymbol{y}_j\|_2-\|\boldsymbol{x}'-\boldsymbol{y}_j\|_2\big|\le\|\boldsymbol{x}-\boldsymbol{x}'\|_2.$$

因此（19）在 $L=2$ 时成立，定理 8 证明了这些结论。 □

C.6 隐私下界

定理 6 的证明。在欧几里得单位球中选取点 $\boldsymbol{v}_1,\ \ldots,\ \boldsymbol{v}_k$，使其相互距离大于 $1/2$，其中

$$k=\min\{n,2^d\},\qquad h=\frac{\log k}{6\varepsilon},\qquad G=\lfloor n/k\rfloor,\qquad N=Gk.$$

这样的点确实存在：分离度大于 $1/2$ 的一个极大集合，其半径为$1/2$的球覆盖单位球，因此由体积比较可知，至少有 2^d 个点。由于 $k\ge 2$ 和 $n-N<k\le N$，我们有$N\ge n/2$。

将前 N个索引划分为 G组 $B_1,\ \ldots,\ B_G$，每组大小为 k。令 $\boldsymbol{e}_1$表示第一个坐标向量，并设 $\boldsymbol{c}_g=10hg\boldsymbol{e}_1$。对于 $a=1,\ \ldots,\ k$，将索引为 $j=\ (g-1)\ k+a$的公共点放置在

$$\boldsymbol{y}_j=\boldsymbol{c}_g+h\boldsymbol{v}_a+\boldsymbol{u}_j,\qquad\|\boldsymbol{u}_j\|_2\le h/16.$$

将每个剩余的公共点放在单独的位置 $10h\ (G+a)\ \boldsymbol{e}_1+\boldsymbol{u}_{N+a}$，其中 $a=1,\ \ldots,\ n-N$，并对其扰动采用相同的上界。每个完整组的直径至多为 $17h/8$。同一组内不同的点之间的距离大于 $h/2-2h/16=3h/8$。不同组中的点（包括剩余的单个点）之间的距离至少为 $10h-2\ (17h/16)$ $=63h/8$。因此，所有不同的公共点之间的距离至少为 $3h/8$。

固定偏移量 $\boldsymbol{b}_1,\ldots,\boldsymbol{b}_n$，其中 $\|\boldsymbol{b}_i\|_2\le h/16$。令 Π 为每个完整组内的均匀随机置换，各组之间相互独立，并固定其余下标。对于每个可能的排列π，按如下方式定义一个私有输入

$$\boldsymbol{x}_i^{\pi}=\boldsymbol{y}_{\pi(i)}+\boldsymbol{b}_i,\qquad i=1,\ldots,n.$$

在该族中，所有公共扰动和私有偏移量保持固定。植入匹配对象 $\pi(i)$的代价为 $\|\boldsymbol{b}_i\|_2\le h/16$，而其他每个匹配对象的代价至少为 $3h/8-h/16=5h/16$。因此 π 是唯一的最优匹配，并且该输入上的每个匹配 σ 都满足

$$\mathrm{err}(\sigma)\ge\frac{h}{4}\sum_{i=1}^{n}\mathbf{1}_{\{\sigma(i)\ne\pi(i)\}}.\tag{20}$$

固定任意 εD-隐私匹配机制 M_Y，并记 $\widehat{\sigma}=M_Y(X^{\Pi})$。考虑完整组 B_g中的一行 i 。对于每个 $a\in B_g$，记$\Pr_a$ 为在 $\Pi(i)=a$条件下的概率，并对其他排列选择和该机制取平均。对于不同的 a, $b\in B_g$，交换目标标签 a,b 是满足 $\pi(i)=a$ 的排列与满足 $\pi(i)=b$的排列之间的一个双射。成对的私有输入恰好在两行上不同，且它们的固定偏移量相互抵消，从而得到

$$D(X^{\pi},X^{\pi'})=2\|\boldsymbol{y}_a-\boldsymbol{y}_b\|_2\le 17h/4.$$

对这些成对输入求度量隐私不等式的平均，对于每一个关于$\widehat{\sigma}(i)$的事件 A，可得

$$\Pr_a(A)\le\beta\Pr_b(A),\qquad\beta=\exp(17\varepsilon h/4)\le k.$$

特别地，对于每个 $a \in B_g$，

$$\sum_{b \in B_g} \Pr_b\{\widehat{\sigma}(i) = a\} \geq \left(1 + \frac{k-1}{\beta}\right) \Pr_a\{\widehat{\sigma}(i) = a\}.$$

对 $a \in B_g$ 求和，并利用可能被分配的标签的不相交性，可得

$$\left(1 + \frac{k-1}{\beta}\right) \sum_{a \in B_g} \Pr_a\{\widehat{\sigma}(i) = a\} \leq \sum_{b \in B_g} \sum_{a \in B_g} \Pr_b\{\widehat{\sigma}(i) = a\} \leq k.$$

由于 $\Pi(i)$在 B_g上是均匀的，因此

$$\Pr\{\widehat{\sigma}(i) \neq \Pi(i)\} \geq \frac{k-1}{\beta + k - 1} \geq \frac{k-1}{2k-1} \geq \frac{1}{3}.$$

组外的分配也算作不正确，因此该论证并不限制机制对伙伴的选择。

对（20）取期望，并对 N 个分组行求和，可得

$$\mathbb{E}_{\Pi, M_Y}[\mathrm{err}(\widehat{\sigma})] \geq \frac{hN}{12} \geq \frac{nh}{24} = \frac{n \log k}{144\varepsilon}.$$

最后，

$$\log k = \min\{\log n, d \log 2\} \geq (\log 2) \min\{\log n, d\}, \qquad \frac{\log 2}{144} > \frac{1}{256}.$$

因此，有限族中至少有一个私有输入的期望超额成本不低于所声称的界。公共构造是在选择机制之前固定的。

对于所有允许的公共扰动和私有偏移量，该论证均成立。特别是，它们可以在开球内变化而不影响上界；此构造不要求私有点与公共点重合，也不要求距离具有精确对称性。

□

D 推广到其他设置

D.1 平方欧几里得代价的上界

平方目标函数使我们能够舍弃代价行内的常数项。这简化了投影表示，也使我们能够在不对 X 设定上界的情况下校准距离向量噪声。令

$$\boldsymbol{c} = \frac{1}{n} \sum_j \boldsymbol{y}_j, \qquad W = \mathrm{span}\{\boldsymbol{y}_j - \boldsymbol{c} : j \in [n]\}, \qquad r = \dim W.$$

$\mathrm{diam}(Y) = \max_{j,\ k} \|\boldsymbol{y}_j - \boldsymbol{y}_k\|_2$。

原始点。 发布 $\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i + \mathbf{Z}_i$，并添加与定理 d 中相同的 2 维径向拉普拉斯噪声，然后最小化 $\sum_i \|\widetilde{\boldsymbol{x}}_i - \boldsymbol{y}_{\sigma(i)}\|_2^2$。

投影表示。 设 $\boldsymbol{B} \in \mathbb{R}^{d \times r}$ 的正交归一列张成 W，并记$\boldsymbol{u}_i = \boldsymbol{B}^\top(\boldsymbol{x}_i - \boldsymbol{c})$ 和 $\boldsymbol{v}_j = \boldsymbol{B}^\top(\boldsymbol{y}_j - \boldsymbol{c})$。平方代价可分解为

$$\|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2^2 = \|\boldsymbol{u}_i - \boldsymbol{v}_j\|_2^2 + \|P_{W^\perp}(\boldsymbol{x}_i - \boldsymbol{c})\|_2^2.$$

最后一项是行常数，不会影响最优指派。因此，投影表示在 $\mathbb{R}^r$ 中发布带有径向拉普拉斯噪声的 $\widetilde{\boldsymbol{u}}_i = \boldsymbol{u}_i + \mathbf{Z}_i$，并最小化 $\sum_i \|\widetilde{\boldsymbol{u}}_i - \boldsymbol{v}_{\sigma(i)}\|_2^2$。

距离向量。使用中心化平方代价表示

$$[h_Y(\boldsymbol{x})]_j = \|\boldsymbol{x}-\boldsymbol{y}_j\|_2^2 - \|\boldsymbol{x}-\boldsymbol{c}\|_2^2 = \|\boldsymbol{y}_j-\boldsymbol{c}\|_2^2 - 2\langle \boldsymbol{x}-\boldsymbol{c}, \boldsymbol{y}_j-\boldsymbol{c}\rangle. \tag{21}$$

从每一行的所有元素中减去相同的项，不会改变匹配。每条记录释放 $h_Y(\boldsymbol{x}_i)+\mathbf{Z}_i$，其密度为

$$p_{\mathbf{Z}_i}(\boldsymbol{z}) \propto \exp\left(-\frac{\varepsilon\|\boldsymbol{z}\|_\infty}{2\max_j \|\boldsymbol{y}_j-\boldsymbol{c}\|_2}\right).$$

中心化使噪声尺度由 $\max_j \|\boldsymbol{y}_j-\boldsymbol{c}\|_2$ 决定，该值可能小于 $\mathrm{diam}(Y)$。

Gumbel。 抽取相互独立的标准 Gumbel 随机变量，并返回

$$\widehat{\sigma} \in \arg\max_{\sigma\in S_n} \sum_i \left(-\|\boldsymbol{x}_i-\boldsymbol{y}_{\sigma(i)}\|_2^2 + \frac{2\,\mathrm{diam}(Y)}{\varepsilon} G_{i,\sigma(i)}\right). \tag{22}$$

定理 8 适用，其中 $L=2\mathrm{diam}(Y)$，因为

$$\begin{aligned}&\left|\left(\|\boldsymbol{x}-\boldsymbol{y}_j\|_2^2 - \|\boldsymbol{x}'-\boldsymbol{y}_j\|_2^2\right) - \left(\|\boldsymbol{x}-\boldsymbol{y}_k\|_2^2 - \|\boldsymbol{x}'-\boldsymbol{y}_k\|_2^2\right)\right| \\ &\quad = 2|\langle \boldsymbol{x}-\boldsymbol{x}', \boldsymbol{y}_j-\boldsymbol{y}_k\rangle| \le 2\,\mathrm{diam}(Y)\|\boldsymbol{x}-\boldsymbol{x}'\|_2.\end{aligned}$$

命题 9。 对于 $n\ge 2$ 和 $\mathrm{diam}(Y)>0$，这四种机制满足 εD-隐私，并且在平方欧几里得代价下具有以下期望超额代价上界，其中 $a_n(s)=\min\{s, 2\sqrt{s\log n}\}$：

机制	$\mathbb{E}[\mathrm{err}(\widehat{\sigma})]$的上界
原始点	$2n$直径$(Y)a_n(d)/\varepsilon$
投影表示	$2n$直径$(Y)a_n(r)/\varepsilon$
距离向量	$4n^2$最大值$_j\|\boldsymbol{y}_j-\boldsymbol{c}\|_2/\varepsilon$
Gumbel	$2n$ 直径(Y) $\log\ n/\varepsilon$

证明。 投影是 1-利普希茨的，并且 h_Y 是 2 倍最大值$_j\|\boldsymbol{y}_j-\boldsymbol{c}\|_2$-利普希茨的（相对于无穷范数），因此其拉普拉斯报告满足隐私要求。对于原始点，展开扰动后的平方代价后，误差为 $-2\langle \mathbf{Z}_i,\ \boldsymbol{y}_j\rangle$ 去掉行常数后。引理 7 将超额代价界定为各行振荡之和

$$2\max_{j,k}\langle \mathbf{Z}_i, \boldsymbol{y}_j-\boldsymbol{y}_k\rangle.$$

回顾一下， $\mathbb{R}^s$ 中的径向拉普拉斯噪声形式为 $\mathbf{Z}=R\boldsymbol{\Theta}$，其中 $\mathbb{E}R = s/\varepsilon$ ，并有一个独立的均匀方向 $\boldsymbol{\Theta}\in\mathbb{S}^{s-1}$。对于单位方向，与候选差值的最大内积不超过直径$(Y)$。球面矩公式还给出

$$\mathbb{E}\exp(t\langle\boldsymbol{\Theta},\boldsymbol{v}\rangle) \le \exp\left(\frac{t^2\|\boldsymbol{v}\|_2^2}{2s}\right).$$

将其应用于至多 n^2 个范数至多为 $\mathrm{diam}(Y)$ 的差分，然后优化指数界，可得

$$\mathbb{E}\max_{j,k}\langle\boldsymbol{\Theta}, \boldsymbol{y}_j-\boldsymbol{y}_k\rangle \le \mathrm{diam}(Y)\min\left\{1, 2\sqrt{\frac{\log n}{s}}\right\}.$$

乘以 $2\mathbb{E}R$ 并对各行求和，得到带有 $s = d$的原始点界。用 $s = r$ 和 $\boldsymbol{v}_j$ 代替 $\boldsymbol{y}_j$，同样的论证给出投影表示界，因为投影保持 Y中各点之间的距离。对于距离向量，每个噪声行的期望无穷范数为 $2n\max_j \|\boldsymbol{y}_j - \boldsymbol{c}\|_2/\varepsilon$；其振荡至多为该范数的两倍。对稳定性界求和得到 $4n^2\max_j \|\boldsymbol{y}_j - \boldsymbol{c}\|_2/\varepsilon$。Gumbel 保证由定理 8 在$L = 2\mathrm{diam}(Y)$下得出。

□

D.2 两个数据集均为私有时的匹配

假设两个数据集均为私有，但其大小和索引仍为公开信息。将隐私度量扩展为

$$D_{\mathrm{both}}\big((X,Y),(X',Y')\big) = D(X,X') + D(Y,Y'). \tag{23}$$

我们要求发布的匹配 $M(X,\ Y)$ 满足

$$\Pr[M(X,Y) \in A] \le e^{\varepsilon D_{\mathrm{both}}((X,Y),(X',Y'))} \Pr[M(X',Y') \in A].$$

下文的构造给出了表 2 中的期望超额成本界。

表 2：两个数据集均为私有时，在 $\mathbb{E}[\mathrm{err}(\widehat{\sigma})]$ 下的上界（23）。破折号表示投影构造无法直接扩展。

机制	欧几里得代价
原始点	$4nd/\varepsilon$
投影表示	—
距离向量	$4n^2/\varepsilon$
Gumbel	$2n\log\ n/\varepsilon$

原始点。 分别对两个数据集中的每个点添加 d维、速率为 ε的径向拉普拉斯噪声，并根据报告计算匹配。它们的联合发布满足$\varepsilon D_{\mathrm{both}}$-隐私。对于欧几里得代价，反三角不等式给出

$$\mathrm{err}(\widehat{\sigma}) \le 2\sum_i \|\widetilde{\boldsymbol{x}}_i - \boldsymbol{x}_i\|_2 + 2\sum_j \|\widetilde{\boldsymbol{y}}_j - \boldsymbol{y}_j\|_2,$$

其期望至多为 $4nd/\varepsilon$。

投影表示。 (8) 中的表示依赖于 Y 的精确仿射张成空间。当 Y 是私有的时，该空间也需要受到保护，因此现有机制及其保证并不会自动适用。

距离向量。 中心化机制可以通过向代价矩阵的行和列添加独立噪声来保护两个数据集 $\boldsymbol{C}$ （1）中，分别为 $\boldsymbol{Z}^{\mathrm{row}}$ 的行和 $\boldsymbol{Z}^{\mathrm{col}}$ 的列生成相互独立的噪声向量，每个向量的密度为

$$p(\boldsymbol{z}) \propto \exp(-\varepsilon\|\boldsymbol{z}\|_\infty), \qquad \boldsymbol{z} \in \mathbb{R}^n.$$

返回 $\boldsymbol{C} + \boldsymbol{Z}^{\mathrm{row}} + \boldsymbol{Z}^{\mathrm{col}}$的最小代价匹配。对 $\boldsymbol{x}_i$ 的改变仅影响行 i，其无穷范数变化至多为 $\|\boldsymbol{x}_i - \boldsymbol{x}_i'\|_2$。在以列噪声为条件时，行噪声

因此给出所需的隐私因子。将行与列互换后，同一论证可保护对 $\boldsymbol{y}_j$的改变。逐一替换这些点即可确立 $\varepsilon D_{\text{both}}$-隐私。引理7中的稳定性论证将超额代价界定为 $\boldsymbol{Z}^{\text{row}}$ 的行振荡与 $\boldsymbol{Z}^{\text{col}}$的列振荡之和。每个噪声向量的期望无穷范数均为 n/ε，从而得到 $\mathbb{E}[\text{err}(\widehat{\sigma})] \leq 4n^2/\varepsilon$。

Gumbel。 对于欧几里得代价，该机制无需修改。定理 8 保护了对 X 的更改。转置代价矩阵和噪声矩阵后，对 Y 的更改也可用同样的证明：固定发生变化的列以外的噪声，并对其可能的匹配对象应用 Gumbel-max 论证。先替换行、再替换列，即可证明 $\varepsilon D_{\text{both}}$-隐私，而效用界仍为$2n \log n/\varepsilon$。

表 3：采样前的数据集规模、环境维度和原始来源。所有隐私扫描均使用每侧 64 个点。

数据集	可用记录	d	来源
噪声环	1,050	2	Graves和Pedrycz [28]
圆	4,000	2	Truszczyński [59]
链环	1,000	3	Ultsch [61]
Iris	150	4	费希尔 [23]
纸币	1,372	4	Lohweg [44]
大肠杆菌	336	7	Nakai [49]
酵母	1,484	8	Nakai [48]
玻璃	214	9	German [26]
Wine	178	13	Aeberhard 和 Forina [1]
Statlog（图像分割）	2,310	19	UCI机器学习仓库 [60]
WDBC	569	30	Wolberg 等人 [64]
电离层	351	34	Sigillito等人 [56]
声呐	208	60	Sejnowski和Gorman [53]
Worms-64	105,000	64	Sieranoja和Fränti [55]
MNIST	10,000	784	LeCun 等人 [40]
Fashion-MNIST	10,000	784	Xiao 等人 [65]

E 数值实验的详细信息

E.1 数据集与特征处理

表 3 列出了十六个数据集。聚类套件固定为 1.2.0 版。对于 MNIST 和 Fashion-MNIST，我们使用该套件中的 10,000 张图像测试子集。鸢尾花使用该套件所附的 R 数据集副本。前三种几何数据和 Worms-64 是合成数据；其余数据集包含真实测量数据或图像。这些来源为匹配提供了不同的维度、形状和特征分布，且不受其原始聚类或分类标签的影响。

对于每个母集合，我们在划分公开/私有数据之前，对特征变换拟合一次。变化的数值特征 z_j 会根据合并母集合的经验均值和标准差映射为 $(z_j - \mu_j)/\sigma_j$。常量特征会中心化为零，并保留其坐标。固定来源值为 a_j，b_j 的二元特征会映射为 $(z_j - a_j)/(b_j - a_j)$。因此，其权重与抽样集合中该指示变量的频率无关。

葡萄酒、WDBC、玻璃、声呐和纸币数据集的所有特征均采用数值标准化。Stat-log 对变化的数值特征采用该方法，并保留其常量坐标。大肠杆菌数据集对五个数值评分进行标准化，并将其两个二元特征映射为 $\{0, 1\}$。酵母数据集对七个数值评分进行标准化，并将其二元指示变量映射为 $\{0, 1\}$。电离层数据集保留第一个二元坐标和第二个零坐标，并对第三至第三十四个坐标进行标准化。鸢尾花数据集、合成数据集以及两个图像数据集在采用统一的 RMS 变换前，均保留其相对坐标尺度。

例如，来自大肠杆菌的一个样本可以示意性地表示为（c_1，c_2，b_1，b_2，c_3，c_4，c_5），其中数值评分为 c_j，二元指示变量为 b_j。经过特征处理后，对应的向量为

$$\left(\frac{c_1 - \mu_1}{\sigma_1}, \frac{c_2 - \mu_2}{\sigma_2}, b_1, b_2, \frac{c_3 - \mu_3}{\sigma_3}, \frac{c_4 - \mu_4}{\sigma_4}, \frac{c_5 - \mu_5}{\sigma_5} \right), \qquad b_1, b_2 \in \{0, 1\}.$$

随后，将通用的中心化和 RMS 重缩放应用于整个向量，包括二进制坐标。

E.2 距离尺度、投影链与嵌套样本

For 实际实例中的 $2n$ 汇总后的特征处理向量 $\boldsymbol{v}_1$、……、 $\boldsymbol{v}_{2n}$ 定义为

$$\bar{\boldsymbol{v}} = \frac{1}{2n}\sum_{i=1}^{2n}\boldsymbol{v}_i, \qquad a^2 = \frac{1}{2n}\sum_{i=1}^{2n}\|\boldsymbol{v}_i - \bar{\boldsymbol{v}}\|_2^2, \qquad \boldsymbol{z}_i = \frac{\boldsymbol{v}_i - \bar{\boldsymbol{v}}}{a}. \tag{24}$$

由此得到的汇集点云均值为零，均方根半径为一。特征权重、中心和半径均视为离线基准构建中的固定部分。机制的隐私保证是在这些归一化输入上评估的。

如附录 B.1所述，单记录位移下的隐私比较 ρ由 $\exp(\varepsilon\rho)$ 控制。具体而言，因子c 隐私保证可扩展至归一化距离 $\rho_c = \log(c)/\varepsilon$。在 $\varepsilon = 8$处，二倍距离约为 0.0866；在 $\varepsilon = 16$处，约为 0.0433。公式（24）以共同的均方根半径为单位表示这些距离。对于固定的特征变换和半径 a，特征空间位移 δ 变为 δ/a，对应的隐私因子为 $\exp(\varepsilon\delta/a)$。因此，在均方根归一化后保持 ε 不变，就能使保护尺度相对于该半径保持固定。这种校准方式说明了为何要在两组扫描中控制半径。

Fashion-MNIST 投影。每个父样本集包含 1,600 张互不相同的全维图像。将其中心化并按均方根归一化后的数据矩阵记为 $\boldsymbol{Z} \in \mathbb{R}^{1600\times 784}$。对于五条相互独立的链，分别抽取一个 Haar分布正交矩阵 $\boldsymbol{O} \in \mathbb{R}^{784\times 784}$。我们通过高斯 QR 分解得到该矩阵，并令三角因子的对角线为正，以此固定符号 [47]。令 $\boldsymbol{O}_d$ 包含其前d 列。在每个目标维度下，设

$$\begin{aligned} \boldsymbol{V}_d &= \boldsymbol{Z}\boldsymbol{O}_d, \qquad \bar{\boldsymbol{v}}_d = \frac{1}{1600}\boldsymbol{V}_d^\top \mathbf{1}, \\ a_d^2 &= \frac{1}{1600}\sum_i \|(\boldsymbol{V}_d)_{i,:}^\top - \bar{\boldsymbol{v}}_d\|_2^2, \qquad \boldsymbol{Z}_d = \frac{\boldsymbol{V}_d - \mathbf{1}\bar{\boldsymbol{v}}_d^\top}{a_d}. \end{aligned} \tag{25}$$

这里，$(\boldsymbol{V}_d)_{i,:}$ 是投影数据矩阵的第 i 行。每个目标维度在五个固定划分中都保留相同的 800 条公开记录和 800 条私有记录。

这些前缀 $\boldsymbol{O}_d$ 构成嵌套均匀随机子空间。它们的联合子空间分布与从 784 维开始依次进行均匀正交降维所得的分布一致，区别仅在于每个保留空间内的正交坐标变换。所有四种匹配机制在此类坐标旋转下的分布均保持不变。对于固定行向量 $\boldsymbol{v}$,

$$\mathbb{E}_{\boldsymbol{O}}\|\boldsymbol{v}\boldsymbol{O}_d\|_2^2 = \frac{d}{784}\|\boldsymbol{v}\|_2^2.$$

乘以$\sqrt{784/d}$ 在期望意义下会保持平方长度不变。用（25）中的实际汇总均方根半径进行除法，会抵消这个标量，并固定每个实际生成的投影点云的半径。

这也解释了其与迭代重缩放的等价性。若从 d 到 d' 的约化由带有 $\boldsymbol{O}_d\boldsymbol{A} = \boldsymbol{O}_{d'}$的标准正交矩阵 $\boldsymbol{A}$ 表示，则当前中心化数据投影到$(\boldsymbol{Z}\boldsymbol{O}_{d'} - \mathbf{1}\bar{\boldsymbol{v}}_{d'}^\top)/a_d$。它们的新均方根半径为 $a_{d'}/a_d$，因此下一次归一化恰好给出 $\boldsymbol{Z}_{d'}$。中间的标量归一化因而在代数上相互抵消。该实验在相同的均方根半径下比较一族投影匹配几何，并对五条独立链取平均，以计入保留子空间的选择。

目标维度为 784、604、404、304、254，从 204 到 4 每隔 5 取一个维度，以及从 1 到 12 的所有整数：共五十六个不同的值。图2(a) 展示了截至 104 的部分。表11 列出了所有测量维度。

纸币嵌套样本。每个父样本集包含 400 条不同的记录。我们在该父样本集上拟合一次四个特征的均值和标准差。五个分区各自排列 200 条公开记录和 200 条私有记录；对于目标 n，我们从每一侧各取前 n 条记录。因此，当 n 增大时，较小样本仍保留其成员。对于每个选出的合并样本，我们使用（24）重新计算 $\bar{\boldsymbol{v}}_n$ 和a_n ，并保持特征权重不变。随后，同一个数值 $\varepsilon = 8$ 便适用于整个均方根半径为 1 的范围。样本量的变化会影响可能的配对集合以及实际形成的几何结构，同时其总体距离尺度受到控制。

E.3 机制校准与重复运行

对于欧几里得代价，原始点添加 d维径向拉普拉斯噪声，其密度正比于 $\exp(-\varepsilon\|\boldsymbol{z}\|_2)$。投影表示在 $s = \min\{r+1, d\}$个坐标上添加相同类型的噪声，其中 r 是公开点的仿射维数。距离向量对每个代价行添加无穷范数拉普拉斯噪声，其密度正比于 $\exp(-\varepsilon\|\boldsymbol{z}\|_\infty)$。Gumbel 向负边代价添加尺度为 $2/\varepsilon$ 的独立标准 Gumbel 随机变量，并最大化扰动后的分配得分。

对于平方代价，原始点保留其 d维报告。投影表示使用公共仿射张成空间：给定公开正交基 $\boldsymbol{B} \in \mathbb{R}^{d\times r}$ 和公开中心 $\boldsymbol{c}$，它报告 $\boldsymbol{B}^\top(\boldsymbol{x}_i - \boldsymbol{c}) + \mathbf{Z}_i$，并添加 r维径向拉普拉斯噪声。垂直方向的平方范数在每个代价行内都是常数，因此分配仅使用这些投影坐标。距离向量扰动

$$h_j(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{y}_j\|_2^2 - \|\boldsymbol{x} - \boldsymbol{c}\|_2^2 = \|\boldsymbol{y}_j - \boldsymbol{c}\|_2^2 - 2\langle \boldsymbol{x} - \boldsymbol{c}, \boldsymbol{y}_j - \boldsymbol{c}\rangle, \qquad \boldsymbol{c} = \frac{1}{n}\sum_j \boldsymbol{y}_j,$$

使用密度正比于 $\exp[-\varepsilon\|\boldsymbol{z}\|_\infty/(2\max_j \|\boldsymbol{y}_j - \boldsymbol{c}\|_2)]$。Gumbel 的平方代价噪声尺度为 $2\mathrm{diam}(Y)/\varepsilon$。这些校准使两个目标函数都在相同的逐记录欧几里得度量下满足隐私保证；平方代价的界见附录 D.1。

我们根据中心化公开矩阵确定公开仿射维数，保留大于 $\max\{n, d\}\sigma_{\max}$ 倍机器精度的奇异值，其中 $\sigma_{\max}$ 是该矩阵的最大奇异值。

每当投影后的报告与原始点报告相同时，我们就共享完整的随机实现，包括其分配。因此，这些重合的曲线表示同一个随机化机制。对于每种固定几何结构，单位噪声抽样会在不同隐私预算间共享，并按 ε缩放；除此之外，各机制使用独立的随机通道。每项所述隐私保证均针对在指定预算下的一次执行。

这二十个父集合是从有限源数据集中独立抽取的，不同父集合可能包含部分相同的记录。在每个父集合内，所有选取的记录均互不相同，且每种公开/私有划分都互不重叠。五十次运行用于估计给定该划分时的机制期望。我们先对这些运行、划分和投影链取平均，再计算二十个父集合均值及其 Student-t 区间。报告的区间均为逐点区间，未针对不同设置之间的比较进行调整。

E.4 数值

以下各表列出了每个目标下、三个代表性预算对应的全部十六个数据集，以及两次专门扫描中测量的每个维度和样本量。每个机制条目均报告均值及其逐点95%置信区间的半宽度。除纸币表报告每对超额成本外，其余均使用总超额成本。

表 4：纸币尺寸扫描，$d = 4$, $\varepsilon = 8$。机制条目为每对超额成本均值 ±逐点95%置信区间半宽。RAND 为 (RAND $-$ OPT)$/n$；$B/n = 2\min\{4, \log\ n\}/8$。

n	Gumbel 原始点	投影	表示 距离向量 RAND		B/n
20	0.162 ± 0.003	0.091 ± 0.003	0.091 ± 0.003	0.359 ± 0.008	0.696 0.749
40	0.205 ± 0.003	0.119 ± 0.003	0.119 ± 0.003	0.484 ± 0.010	0.822 0.922
60	0.232 ± 0.002	0.137 ± 0.002	0.137 ± 0.002	0.547 ± 0.006	0.883 1.000
80	0.252 ± 0.002	0.149 ± 0.002	0.149 ± 0.002	0.588 ± 0.006	0.924 1.000
100	0.265 ± 0.002	0.158 ± 0.002	0.158 ± 0.002	0.617 ± 0.004	0.953 1.000
120	0.277 ± 0.003	0.166 ± 0.002	0.166 ± 0.002	0.640 ± 0.005	0.976 1.000
140	0.285 ± 0.002	0.172 ± 0.002	0.172 ± 0.002	0.655 ± 0.005	0.992 1.000
160	0.294 ± 0.002	0.179 ± 0.001	0.179 ± 0.001	0.670 ± 0.004	1.007 1.000
180	0.301 ± 0.002	0.183 ± 0.001	0.183 ± 0.001	0.683 ± 0.004	1.019 1.000
200	0.307 ± 0.002	0.188 ± 0.001	0.188 ± 0.001	0.693 ± 0.004	1.030 1.000

表5：$\varepsilon = 4$, $n = 64$处的欧几里得超额匹配代价。各项为均值 ± 逐点95%置信区间半宽，基于二十个父集合计算。RAND表示RAND −OPT。

数据集	Gumbel 原始点	投影	表示 距离向量 RAND		
噪声环	25.534 ± 0.517	13.661 ± 0.337	13.661 ± 0.337	44.320 ± 0.752	58.825
圆	27.534 ± 0.178	14.271 ± 0.155	14.271 ± 0.155	46.525 ± 0.260	65.009
链环	29.642 ± 0.349	18.220 ± 0.314	18.220 ± 0.314	47.962 ± 0.500	64.143
Iris	18.372 ± 0.347	14.700 ± 0.360	14.700 ± 0.360	33.181 ± 0.686	58.935
纸币	29.061 ± 0.450	20.038 ± 0.313	20.038 ± 0.313	44.162 ± 0.443	57.155
大肠杆菌	29.831 ± 0.313	24.482 ± 0.275	24.296 ± 0.328	41.697 ± 0.367	52.776
酵母	29.219 ± 0.402	24.193 ± 0.281	23.979 ± 0.305	36.331 ± 0.615	41.638
玻璃	26.717 ± 0.334	25.480 ± 0.372	25.480 ± 0.372	37.253 ± 0.695	43.824
Wine	27.969 ± 0.283	27.355 ± 0.298	27.355 ± 0.298	35.426 ± 0.297	42.668
Statlog	28.433 ± 0.513	33.398 ± 0.622	33.335 ± 0.657	39.350 ± 0.731	48.201
WDBC	25.814 ± 0.269	29.746 ± 0.319	29.746 ± 0.319	33.190 ± 0.309	40.129
电离层	25.661 ± 0.323	30.743 ± 0.490	30.743 ± 0.490	32.017 ± 0.464	37.223
声呐	24.745 ± 0.245	27.940 ± 0.371	27.940 ± 0.371	28.412 ± 0.350	31.600
Worms-64	14.689 ± 0.170	15.853 ± 0.176	15.853 ± 0.176	15.572 ± 0.182	16.423
MNIST	21.126 ± 0.311	23.480 ± 0.394	22.319 ± 0.366	22.619 ± 0.361	23.943
Fashion-MNIST	25.312 ± 0.360	33.661 ± 0.621	30.369 ± 0.562	30.374 ± 0.532	34.937

表6：$\varepsilon = 8$, $n = 64$处的欧几里得超额匹配代价。表中数值为二十个父集合的均值 ± 逐点95%置信区间半宽。RAND表示RAND −OPT。

数据集	Gumbel 原始点	投影	表示 距离向量 RAND		
噪声环	10.757 ± 0.179	5.273 ± 0.103	5.273 ± 0.103	34.399 ± 0.654	58.825
圆	11.942 ± 0.101	5.359 ± 0.068	5.359 ± 0.068	35.149 ± 0.248	65.009
链环	12.191 ± 0.124	6.568 ± 0.097	6.568 ± 0.097	37.409 ± 0.441	64.143
Iris	10.927 ± 0.120	7.588 ± 0.140	7.588 ± 0.140	22.245 ± 0.455	58.935
纸币	14.909 ± 0.274	8.885 ± 0.141	8.885 ± 0.141	35.375 ± 0.407	57.155
大肠杆菌	16.362 ± 0.161	12.025 ± 0.156	11.902 ± 0.194	33.942 ± 0.349	52.776
酵母	18.259 ± 0.193	13.565 ± 0.201	13.366 ± 0.211	31.768 ± 0.486	41.638
玻璃	15.002 ± 0.189	13.981 ± 0.175	13.981 ± 0.175	31.604 ± 0.532	43.824
Wine	17.265 ± 0.159	16.434 ± 0.158	16.434 ± 0.158	29.867 ± 0.295	42.668
Statlog	15.682 ± 0.220	21.052 ± 0.398	21.015 ± 0.424	32.750 ± 0.662	48.201
WDBC	16.477 ± 0.185	21.256 ± 0.222	21.256 ± 0.222	27.949 ± 0.259	40.129
电离层	15.834 ± 0.216	23.912 ± 0.308	23.912 ± 0.308	27.659 ± 0.402	37.223
声呐	17.628 ± 0.135	24.114 ± 0.305	24.114 ± 0.305	25.539 ± 0.289	31.600
Worms-64	12.656 ± 0.167	15.231 ± 0.179	15.231 ± 0.179	14.774 ± 0.181	16.423
MNIST	17.499 ± 0.212	23.018 ± 0.379	20.484 ± 0.317	21.307 ± 0.327	23.943
Fashion-MNIST	16.724 ± 0.190	32.351 ± 0.597	25.625 ± 0.432	26.492 ± 0.429	34.937

表7： $\varepsilon = 16$、 $n = 64$处的欧几里得超额匹配代价。各项为二十个父集合上的均值 ± 逐点95%置信区间半宽。RAND表示RAND −OPT。

数据集	Gumbel 原始点		投影 表示 距离向量	RAND	
噪声环	4.483 ± 0.075	1.994 ± 0.047	1.994 ± 0.047	23.424 ± 0.479	58.825
圆	4.182 ± 0.055	1.786 ± 0.038	1.786 ± 0.038	23.769 ± 0.220	65.009
链环	4.109 ± 0.047	2.161 ± 0.028	2.161 ± 0.028	26.132 ± 0.411	64.143
Iris	6.242 ± 0.047	3.666 ± 0.067	3.666 ± 0.067	15.044 ± 0.263	58.935
纸币	5.886 ± 0.078	3.081 ± 0.051	3.081 ± 0.051	25.668 ± 0.400	57.155
大肠杆菌	6.968 ± 0.078	4.658 ± 0.054	4.598 ± 0.077	25.172 ± 0.328	52.776
酵母	7.864 ± 0.131	5.521 ± 0.108	5.413 ± 0.125	25.437 ± 0.362	41.638
玻璃	6.768 ± 0.117	5.998 ± 0.109	5.998 ± 0.109	24.011 ± 0.368	43.824
Wine	7.849 ± 0.062	7.255 ± 0.058	7.255 ± 0.058	23.007 ± 0.271	42.668
Statlog	6.519 ± 0.138	9.578 ± 0.132	9.557 ± 0.142	24.846 ± 0.568	48.201
WDBC	8.088 ± 0.066	11.832 ± 0.104	11.832 ± 0.104	21.834 ± 0.202	40.129
电离层	7.241 ± 0.097	13.861 ± 0.146	13.861 ± 0.146	21.803 ± 0.346	37.223
声呐	8.675 ± 0.113	17.107 ± 0.195	17.107 ± 0.195	21.343 ± 0.216	31.600
Worms-64	9.064 ± 0.139	13.742 ± 0.173	13.742 ± 0.173	13.352 ± 0.180	16.423
MNIST	10.201 ± 0.115	22.047 ± 0.357	16.465 ± 0.213	18.990 ± 0.267	23.943
Fashion-MNIST	7.814 ± 0.100	29.713 ± 0.541	17.215 ± 0.259	21.080 ± 0.320	34.937

表 8：在 $\varepsilon = 4$, $n = 64$ 处的平方欧几里得超额匹配代价。表中数值为二十个父集合的均值 ± 逐点95%置信区间半宽。RAND 表示 RAND − OPT。

数据集	Gumbel 原始点		投影 表示 距离向量	RAND	
噪声环	59.764 ± 1.953	15.084 ± 0.228	15.084 ± 0.228	87.897 ± 1.756	115.118
圆	50.292 ± 0.638	14.914 ± 0.192	14.914 ± 0.192	81.599 ± 0.615	120.965
链环	55.844 ± 0.741	22.073 ± 0.294	22.073 ± 0.294	83.648 ± 0.957	118.613
Iris	35.820 ± 0.508	16.634 ± 0.382	16.634 ± 0.382	67.329 ± 0.717	120.007
纸币	60.454 ± 1.069	27.045 ± 0.482	27.045 ± 0.482	91.231 ± 1.102	115.218
大肠杆菌	61.596 ± 0.699	41.045 ± 0.411	37.895 ± 0.628	87.604 ± 0.571	108.964
酵母	63.979 ± 1.013	43.410 ± 0.756	40.992 ± 0.715	80.105 ± 1.993	87.807
玻璃	62.465 ± 0.970	43.797 ± 0.585	43.797 ± 0.585	83.081 ± 1.641	90.674
Wine	58.688 ± 0.975	55.376 ± 0.404	55.376 ± 0.404	78.604 ± 0.637	94.353
Statlog	58.330 ± 1.754	59.246 ± 1.472	58.289 ± 1.472	84.449 ± 2.020	95.863
WDBC	57.725 ± 0.916	62.689 ± 0.603	62.689 ± 0.603	79.612 ± 0.792	93.257
电离层	53.854 ± 0.727	58.583 ± 0.828	58.261 ± 0.829	66.316 ± 1.062	74.116
声呐	53.709 ± 0.419	63.241 ± 0.626	63.241 ± 0.626	66.900 ± 0.599	73.229
Worms-64	28.295 ± 0.474	28.691 ± 0.484	28.613 ± 0.457	29.777 ± 0.480	30.584
MNIST	47.499 ± 0.546	56.404 ± 0.704	53.069 ± 0.657	54.231 ± 0.663	57.572
Fashion-MNIST	52.339 ± 0.518	77.530 ± 1.195	68.736 ± 0.993	69.436 ± 0.866	80.839

表9： $\varepsilon = 8$、 $n = 64$处的平方欧几里得超额匹配代价。条目为二十个父集合上的均值 ± 逐点95%置信区间半宽。RAND表示RAND − OPT。

数据集	Gumbel 原始点		投影 表示	距离向量	RAND
噪声环	34.186 ± 1.955	4.331 ± 0.101	4.331 ± 0.101	70.761 ± 1.915	115.118
圆	26.482 ± 0.502	4.245 ± 0.076	4.245 ± 0.076	59.780 ± 0.613	120.965
链环	31.528 ± 0.493	6.246 ± 0.074	6.246 ± 0.074	64.207 ± 0.932	118.613
Iris	18.960 ± 0.229	6.069 ± 0.121	6.069 ± 0.121	42.697 ± 0.589	120.007
纸币	37.000 ± 0.925	9.430 ± 0.210	9.430 ± 0.210	74.550 ± 1.366	115.218
大肠杆菌	36.535 ± 0.620	16.707 ± 0.203	14.919 ± 0.311	71.737 ± 0.677	108.964
酵母	46.604 ± 1.413	21.019 ± 0.354	19.303 ± 0.387	73.027 ± 1.509	87.807
玻璃	42.019 ± 0.780	19.799 ± 0.423	19.799 ± 0.423	75.813 ± 1.477	90.674
Wine	36.907 ± 0.775	30.284 ± 0.265	30.284 ± 0.265	66.372 ± 0.659	94.353
Statlog	37.526 ± 1.412	32.102 ± 0.845	31.137 ± 0.771	74.668 ± 2.121	95.863
WDBC	38.998 ± 0.933	40.248 ± 0.558	40.248 ± 0.558	69.153 ± 0.833	93.257
电离层	36.477 ± 0.538	43.315 ± 0.563	42.796 ± 0.544	59.296 ± 0.914	74.116
声呐	37.549 ± 0.427	53.101 ± 0.444	53.101 ± 0.444	61.179 ± 0.500	73.229
Worms-64	25.775 ± 0.474	26.799 ± 0.465	26.722 ± 0.424	28.979 ± 0.490	30.584
MNIST	35.823 ± 0.377	55.164 ± 0.703	48.262 ± 0.577	50.883 ± 0.598	57.572
Fashion-MNIST	33.151 ± 0.562	74.128 ± 1.098	56.452 ± 0.670	60.051 ± 0.682	80.839

表10： $\varepsilon = 16$、 $n = 64$处的平方欧几里得超额匹配代价。条目为二十个父集合上的均值 ± 逐点95%置信区间半宽。RAND表示RAND − OPT。

数据集	Gumbel 原始点		投影 表示	距离向量	RAND
噪声环	15.298 ± 1.107	1.315 ± 0.050	1.315 ± 0.050	52.260 ± 1.837	115.118
圆	12.953 ± 0.288	1.116 ± 0.021	1.116 ± 0.021	39.429 ± 0.555	120.965
链环	14.216 ± 0.229	1.650 ± 0.034	1.650 ± 0.034	45.803 ± 0.743	118.613
Iris	11.434 ± 0.168	2.354 ± 0.055	2.354 ± 0.055	25.079 ± 0.348	120.007
纸币	19.961 ± 0.591	2.732 ± 0.050	2.732 ± 0.050	55.296 ± 1.254	115.218
大肠杆菌	18.809 ± 0.351	5.384 ± 0.079	4.736 ± 0.110	53.107 ± 0.628	108.964
酵母	27.891 ± 1.467	7.472 ± 0.151	6.738 ± 0.188	62.284 ± 1.412	87.807
玻璃	23.282 ± 0.600	6.787 ± 0.255	6.787 ± 0.255	63.690 ± 1.053	90.674
Wine	19.076 ± 0.444	11.831 ± 0.141	11.831 ± 0.141	50.907 ± 0.617	94.353
Statlog	20.573 ± 1.087	11.736 ± 0.252	11.248 ± 0.245	60.454 ± 1.972	95.863
WDBC	23.192 ± 0.711	19.131 ± 0.297	19.131 ± 0.297	55.421 ± 0.998	93.257
电离层	18.465 ± 0.456	22.610 ± 0.312	22.115 ± 0.295	48.746 ± 0.715	74.116
声呐	20.692 ± 0.324	35.405 ± 0.264	35.405 ± 0.264	52.167 ± 0.441	73.229
Worms-64	20.569 ± 0.540	23.107 ± 0.397	22.959 ± 0.385	27.487 ± 0.489	30.584
MNIST	18.143 ± 0.299	52.628 ± 0.620	38.077 ± 0.332	45.104 ± 0.556	57.572
Fashion-MNIST	16.168 ± 0.334	67.118 ± 0.942	35.844 ± 0.448	47.529 ± 0.537	80.839

表 11：Fashion-MNIST 维度扫描，$n = 800$，$\varepsilon = 16$。机制条目为总超额成本均值 ± 逐点95%置信区间半宽。RAND 为 RAND $-$ OPT；$B = 2n\min\{d, \log n\}/\varepsilon$。

d	Gumbel	原始点	投影表示	距离向量	RAND	B
1	65.556 ± 1.052	33.914 ± 0.265	33.914 ± 0.265	296.420 ± 2.252	851.812	100.000
2	127.698 ± 0.602	52.767 ± 0.460	52.767 ± 0.460	463.815 ± 4.463	903.867	200.000
3	155.818 ± 0.409	65.883 ± 0.507	65.883 ± 0.507	508.928 ± 5.984	879.018	300.000
4	169.041 ± 0.355	76.365 ± 0.514	76.365 ± 0.514	512.379 ± 5.700	840.827	400.000
5	176.880 ± 0.396	86.390 ± 0.515	86.390 ± 0.515	505.198 ± 5.196	805.214	500.000
6	181.618 ± 0.451	95.989 ± 0.472	95.989 ± 0.472	495.192 ± 4.711	776.106	600.000
7	184.459 ± 0.508	104.865 ± 0.457	104.865 ± 0.457	487.454 ± 4.376	752.415	668.461
8	186.110 ± 0.453	113.312 ± 0.347	113.312 ± 0.347	480.688 ± 3.694	733.505	668.461
9	187.310 ± 0.511	121.565 ± 0.309	121.565 ± 0.309	473.888 ± 3.502	718.097	668.461
10	188.057 ± 0.616	129.248 ± 0.283	129.248 ± 0.283	468.736 ± 2.996	705.000	668.461
11	188.419 ± 0.692	136.556 ± 0.276	136.556 ± 0.276	464.097 ± 3.111	694.101	668.461
12	188.441 ± 0.633	143.597 ± 0.207	143.597 ± 0.207	459.293 ± 2.920	685.759	668.461
14	188.353 ± 0.816	156.747 ± 0.350	156.747 ± 0.350	453.078 ± 2.687	671.547	668.461
19	188.143 ± 0.740	185.864 ± 0.497	185.864 ± 0.497	442.387 ± 2.290	648.225	668.461
24	187.388 ± 0.745	210.512 ± 0.611	210.512 ± 0.611	436.357 ± 1.901	634.520	668.461
29	186.591 ± 0.832	231.696 ± 0.848	231.696 ± 0.848	432.399 ± 1.916	626.210	668.461
34	185.913 ± 0.870	250.188 ± 0.922	250.188 ± 0.922	429.403 ± 1.889	620.069	668.461
39	185.590 ± 0.766	266.560 ± 1.196	266.560 ± 1.196	427.358 ± 2.095	615.051	668.461
44	185.064 ± 0.732	280.933 ± 1.241	280.933 ± 1.241	425.563 ± 1.995	612.022	668.461
49	184.739 ± 0.751	293.727 ± 1.225	293.727 ± 1.225	424.060 ± 1.908	609.377	668.461
54	184.547 ± 0.763	305.562 ± 1.425	305.562 ± 1.425	423.357 ± 1.981	606.543	668.461
59	184.279 ± 0.825	315.995 ± 1.479	315.995 ± 1.479	422.279 ± 1.983	604.979	668.461
64	184.048 ± 0.794	325.447 ± 1.478	325.447 ± 1.478	421.351 ± 1.985	603.348	668.461
69	183.764 ± 0.752	334.038 ± 1.466	334.038 ± 1.466	420.633 ± 1.927	602.115	668.461
74	183.635 ± 0.782	342.051 ± 1.604	342.051 ± 1.604	420.124 ± 1.936	600.603	668.461
79	183.492 ± 0.742	349.199 ± 1.596	349.199 ± 1.596	419.541 ± 1.872	599.486	668.461
84	183.524 ± 0.744	355.947 ± 1.549	355.947 ± 1.549	419.042 ± 1.739	598.232	668.461
89	183.381 ± 0.696	362.184 ± 1.477	362.184 ± 1.477	418.421 ± 1.707	597.376	668.461
94	183.283 ± 0.681	367.951 ± 1.441	367.951 ± 1.441	418.242 ± 1.622	596.566	668.461
99	183.187 ± 0.610	373.384 ± 1.471	373.384 ± 1.471	417.901 ± 1.630	595.939	668.461
104	183.079 ± 0.591	378.428 ± 1.496	378.428 ± 1.496	417.639 ± 1.578	595.442	668.461
109	182.920 ± 0.598	383.089 ± 1.527	383.089 ± 1.527	417.264 ± 1.613	595.207	668.461
114	182.786 ± 0.574	387.459 ± 1.458	387.459 ± 1.458	417.006 ± 1.619	594.677	668.461
119	182.802 ± 0.541	391.677 ± 1.452	391.677 ± 1.452	416.616 ± 1.579	594.234	668.461
124	182.602 ± 0.565	395.535 ± 1.494	395.535 ± 1.494	416.539 ± 1.561	594.072	668.461
129	182.522 ± 0.508	399.354 ± 1.425	399.354 ± 1.425	416.342 ± 1.539	593.683	668.461
134	182.535 ± 0.538	402.976 ± 1.475	402.976 ± 1.475	416.251 ± 1.527	593.084	668.461
139	182.512 ± 0.587	406.063 ± 1.485	406.063 ± 1.485	415.989 ± 1.519	592.625	668.461
144	182.478 ± 0.604	409.245 ± 1.507	409.245 ± 1.507	415.778 ± 1.509	592.258	668.461
149	182.440 ± 0.558	412.341 ± 1.523	412.341 ± 1.523	415.706 ± 1.489	591.982	668.461
154	182.393 ± 0.599	415.240 ± 1.489	415.240 ± 1.489	415.536 ± 1.496	591.722	668.461
159	182.371 ± 0.574	417.923 ± 1.531	417.923 ± 1.531	415.415 ± 1.490	591.492	668.461
164	182.332 ± 0.590	420.528 ± 1.565	420.528 ± 1.565	415.221 ± 1.520	591.338	668.461
169	182.310 ± 0.591	423.090 ± 1.553	423.090 ± 1.553	415.236 ± 1.500	591.146	668.461
174	182.228 ± 0.594	425.420 ± 1.574	425.420 ± 1.574	415.100 ± 1.482	590.818	668.461
179	182.198 ± 0.569	427.758 ± 1.593	427.758 ± 1.593	415.075 ± 1.497	590.577	668.461
184	182.156 ± 0.548	430.035 ± 1.586	430.035 ± 1.586	414.976 ± 1.473	590.417	668.461
189	182.107 ± 0.546	432.279 ± 1.565	432.279 ± 1.565	414.917 ± 1.497	590.467	668.461
194	181.998 ± 0.553	434.260 ± 1.623	434.260 ± 1.623	414.853 ± 1.574	590.349	668.461
199	182.029 ± 0.554	436.081 ± 1.634	436.081 ± 1.634	414.670 ± 1.537	590.111	668.461
204	181.981 ± 0.581	438.166 ± 1.619	438.166 ± 1.619	414.668 ± 1.507	589.868	668.461
254	181.868 ± 0.601	453.813 ± 1.651	453.813 ± 1.651	414.083 ± 1.457	588.641	668.461
304	181.773 ± 0.577	465.543 ± 1.652	465.543 ± 1.652	413.482 ± 1.483	587.824	668.461
404	181.573 ± 0.549	481.966 ± 1.629	481.966 ± 1.629	413.179 ± 1.458	586.724	668.461
604	181.446 ± 0.565	501.005 ± 1.696	501.005 ± 1.696	412.602 ± 1.480	585.394	668.461
784	181.312 ± 0.562	511.643 ± 1.637	511.655 ± 1.643	412.208 ± 1.400	584.989	668.461

E.5 隐私预算扫描

以下各图展示了全部 16 个基准测试的完整隐私预算扫描，每侧 $n = 64$。每个面板使用各自的纵轴刻度；灰色虚线表示对应数据集和目标的随机基准。附录 E.4 报告了代表性预算下的均值和置信区间。

E.5.1 欧几里得代价

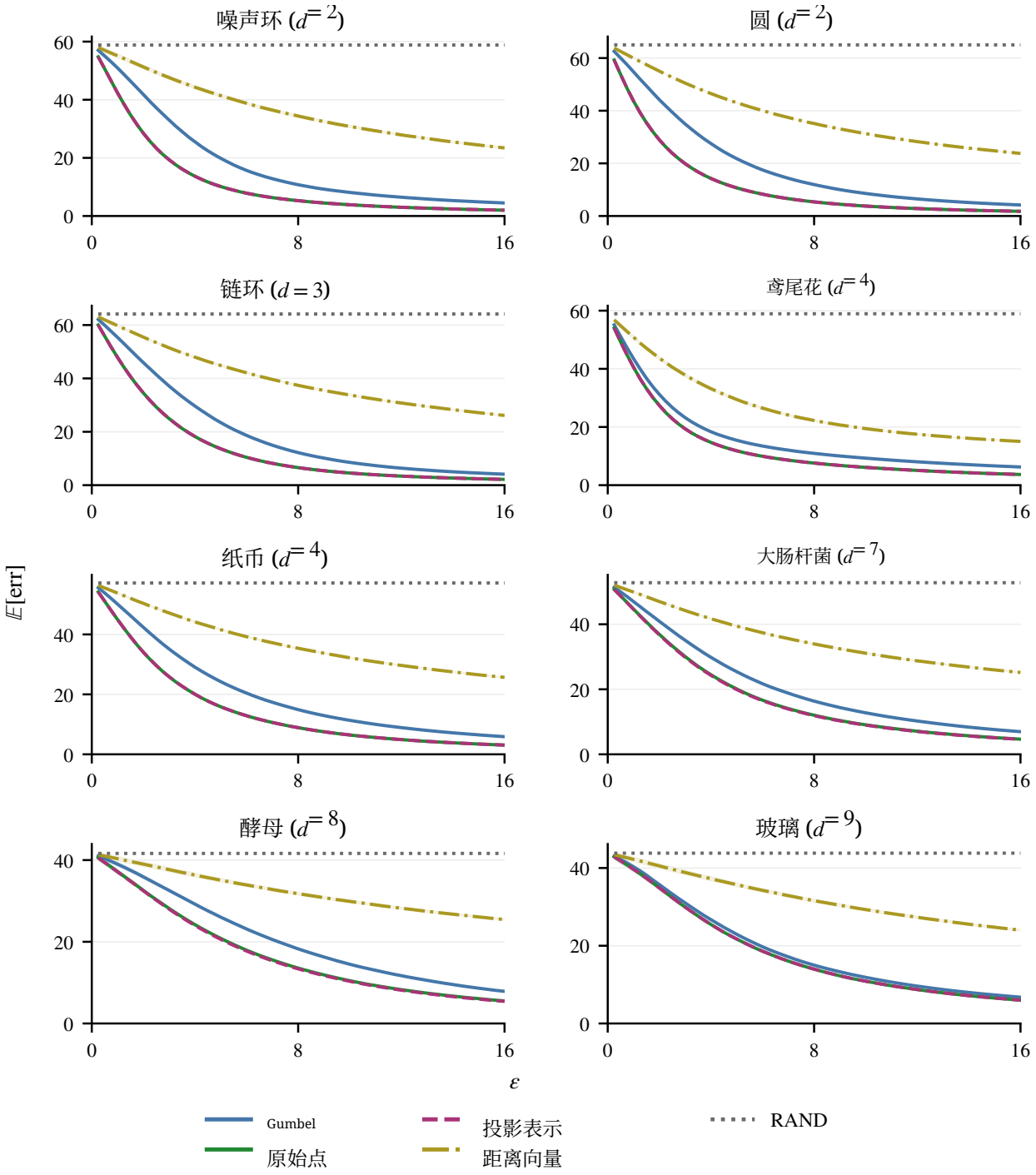

图5：八个最低维度基准测试在不同隐私预算下的欧几里得超额匹配代价。

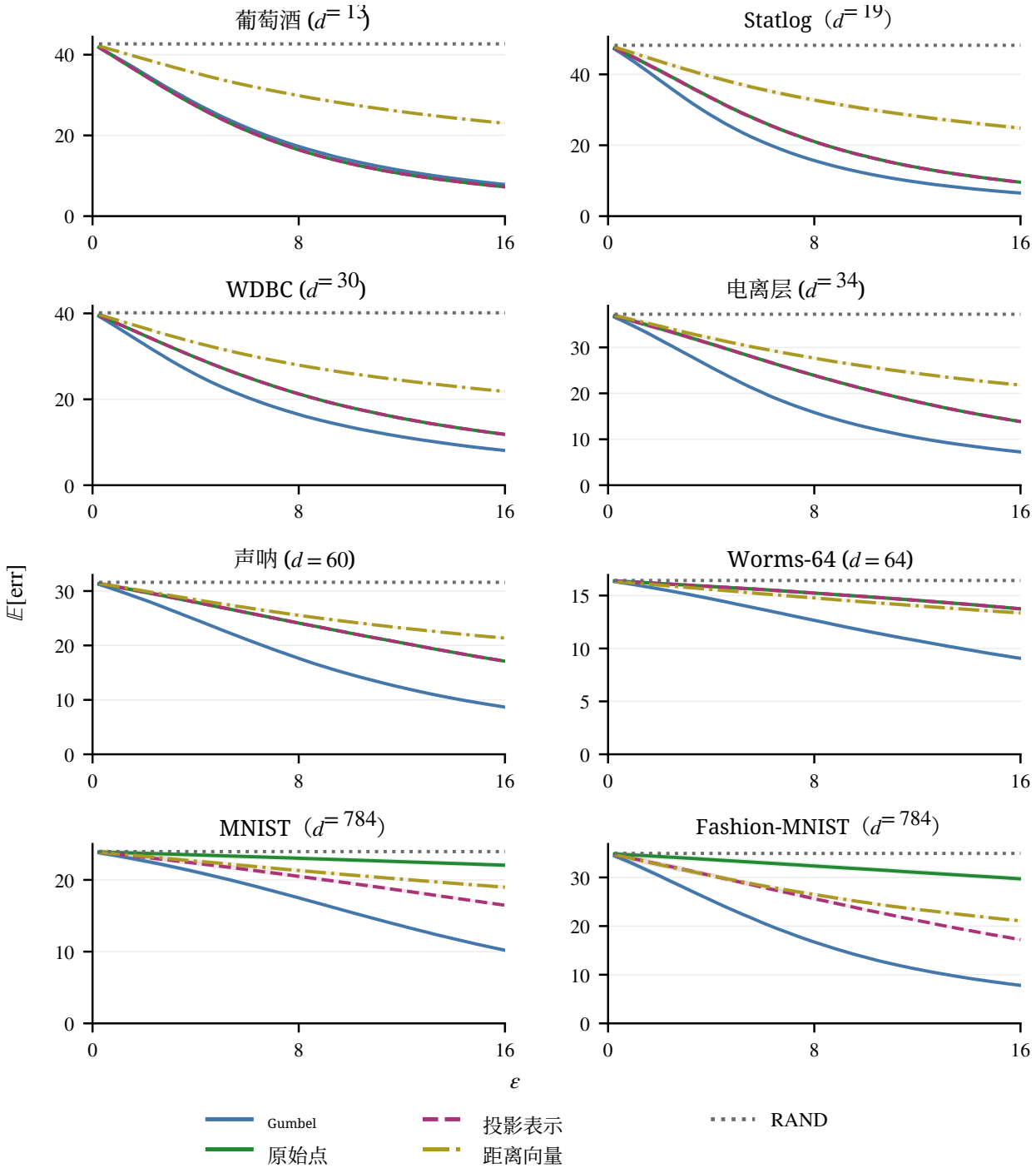

图6：Euclidea　其余八个基准测试在不同隐私预算下的额外匹配成本 .

E.5.2 平方欧几里得代价

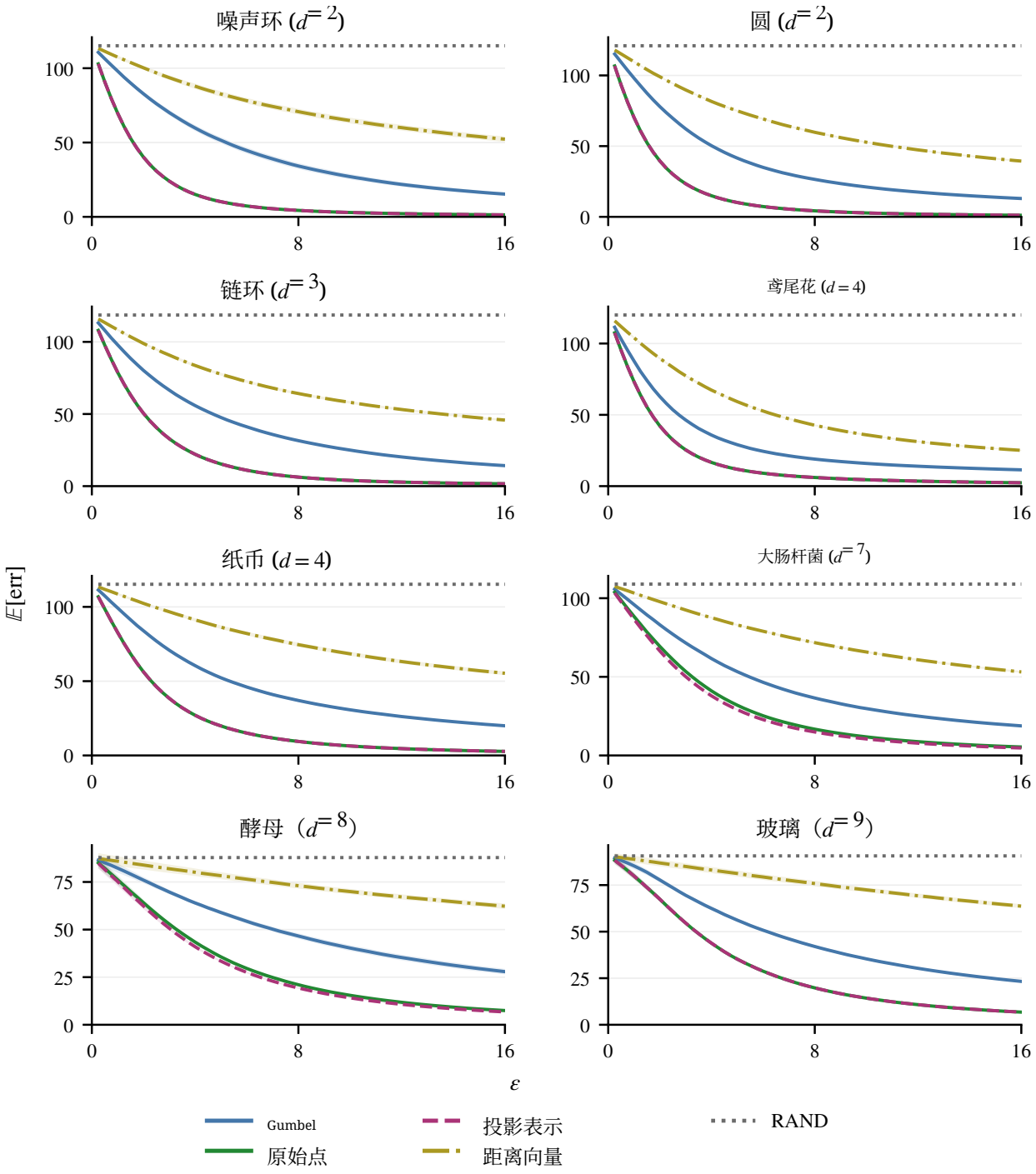

图 7：八个最低维度基准测试在不同隐私预算下的平方欧几里得超额匹配代价。

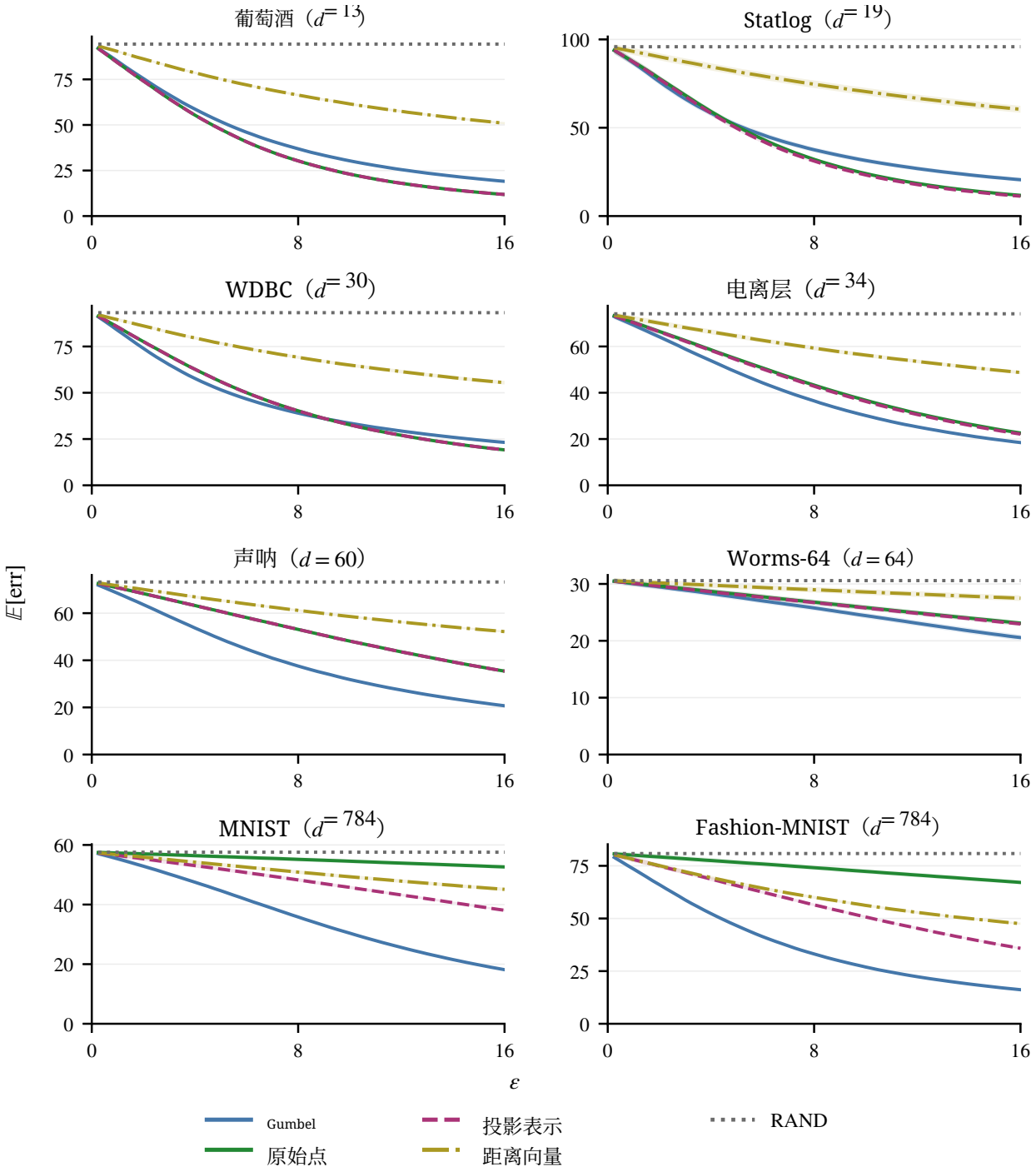

图 8：其余八个基准测试在不同隐私预算下的平方欧几里得超额匹配代价。

F 私有秩能量实验的细节

F.1 参考网格与零分布

该网格由编号为 1 至 208 的未打乱 Halton 点组成，并以前 60 个素数作为坐标基。它独立于测量值与标签。其仿射秩为 47。匹配既不使用队列标签，也不使用依赖标签的约束，且随机平局裁决在分布上对称地对待各行。

在 H_0 下，参考秩在 Y 的所有排列上均匀分布。我们通过蒙特卡罗模拟来近似所得的零分布。由于 T 在队列内排列下保持不变，每次抽取只需为第一队列均匀选取一个包含 a 个参考点的子集，其余 b 个则分配给第二队列。利用 $J = 9{,}999$ 个模拟统计量$T^{(1)}, \dots, T^{(J)}$，我们计算

$$\widehat{p} = \frac{1 + \sum_{j=1}^{J} \mathbf{1}\{T^{(j)} \geq T(\widehat{\sigma})\}}{J + 1}. \tag{26}$$

在整个计算过程中，观测到的私有分配保持固定。由于每一次分配都将所有参考点各使用一次，这些模拟的原假设统计量可用于每一种方法和每一次运行。

图中的置信区间使用 Student-t 分布，根据每种方法和 ε各自的 1,000 次独立运行计算得出，同时固定模拟的原假设样本。

在每个固定的 ε 下，各次运行彼此独立。为比较不同预算，同一组底层噪声抽样会按比例缩放；这些比较不会将独立运行的隐私组合起来。隐私保证仅适用于每次单独执行。