差分隐私

Xiaoxu Lin
Jinan University
xiaoxulin@stu2019.jnu.edu.cn

Abstract

简单介绍一下差分隐私吧

1 Introduction

在讨论关于差分隐私的具体技术之前，我们首先应该搞清楚使用差分隐私技术的动机。所以本节的核心是一个问题：为什么要使用差分隐私？

1.1 其他隐私保护技术

相对于差分隐私，人们可能更熟悉其他针对数据集的隐私保护技术，比如 K-anonymity[1], L-diversity[2] 以及 T-closeness[3]。人们通过这些技术手段来产生一个新的数据集，隐私保护 (PP) 数据集, 来保护用户的隐私信息。前面所提到三种隐私保护技术首先都将用户标识信息去除，比如姓名、身份证号，留下准标识符 (quasi-identifier)，比如性别、年龄、国籍，以及对应的敏感信息，比如疾病。含有相同准标识符的数据项集合叫做等价类 (equivalence class)。K-anonymity 要求数据集中的每一个数据项至少有另外 k-1 个数据项的准标识符与它相同，也就是说每个等价类的阶至少为 k，如图 1 所示。但是如果一个等价类拥有相同的敏感信息时，k -anonymity 就不发挥作用了，这又被称为同质攻击 (homogeneity attack) 就比如图 1 左的第 9、10、11 项都患有癌症。

L-diversity 就是为了解决上述情况，L-diversity 要求每个等价类的敏感信息至少有 L 个不同的值 (Distinction L-Diversity)。我们再考虑一种情况，一个等价类里面有 100 个数据项且只有两个不同的敏感信息值 (L=2)，其中有 99 个人有着相同的敏感信息值，在这种情况下虽然攻击者并不能直接锁定目标数据项的敏感信息值，但是能大概率 (99%) 地断言目标数据项的敏感信息值。另外，L-diversity 并不能区分敏感信息值之间的语义关系，如图 2 所示，这个表显然符合 3-diversity，但是如果知道了某个人属于第一个等价类，那么就可以推断出他工资低且患有胃部相关疾病。

T-closeness 要求每个等价类里面敏感信息分布与整个数据集敏感信息分布之差不超过阈值 T。这种分布之间的差异通过推土机距离 EMD(Earth Mover's Distance)，也叫做

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

图 1: 4-anonymity(左)，6-anonymity(右)

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

图 2: 3-diversity

瓦瑟斯坦距离 (Wasserstein distance)，来表示。用一种通俗的语言来说，EMD 是一个运输问题的最优解，想象现在有 m 个土堆 $P = p_1, p_2, .., p_m$，我们需要将土堆 P 变成土堆 $Q = (q_1, q_2, ..., q_m)$，那么运输问题的最优解就是如何做最少的功完成土堆的转换。那么用数学语言描述则为

在约束

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m, 1 \leq j \leq m \qquad (c_1)$$

$$p_i - \sum_{j=1}^{m} f_{ij} + \sum_{j=1}^{m} f_{ji} = q_i \qquad 1 \leq i \leq m \qquad (c_2)$$

$$\sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1 \qquad (c_3)$$

对如下目标进行优化

$$\min WORK(P, Q, F) = \sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij} f_{ij}$$

d_{ij} 可以当作是土堆间的距离，f_{ij} 可以看成是土堆间需要移动的质量。(说实话，我没明白约束 c3 的意义)。我们知道度量两个分布之间差异的方法有很多，比如 KL 散度？而选用 EMD 来度量分布间的差异的原因是它能度量敏感信息值之间的语义差异，EMD 中会给离群值赋予更高的权重 (距离更远)。但是 EMD 并不能度量敏感信息 value 与从其获得的信息增益之间的关系，比如 (0.01,0.99) 与 (0.11,0.89) 之间的 EMD 是 0.1，(0.4,0.6) 与 (0.5,0.5) 之间的 EMD 也是 0.1，但是前者的变化率是 1000%，而后者只是 25%。

可以看到，以上所提到的隐私保护技术的局限性在于它们都只能有效应地应对特定的攻击而且依赖对攻击者有一些特殊背景知识的假设 [4]，但是枚举攻击者所有可能拥有的背景知识是不可能。

1.2 重构攻击

你可能会想，如果不公开 (匿名化) 数据集本身，公开的是关于元数据的统计量 (statistics)，是不是就不会导致隐私泄露了呢？然而即使仅仅公布统计量，也无法阻止个人用户信息的泄露。所谓的重构攻击 (reconstruction attack) 就是利用统计数据中的冗余来对数据库进行重构。下面通过一个简单的例子 [5] 来展现如何进行重构攻击。

表一是一张虚构的人口统计表，它每一行统计了不同群体的个数 (Count)、年龄中位数 (Median) 以及年龄平均数 (Mean)。其中 (D) 表示该群体的个体数小于三，不予展示统计量。让我们考虑统计量 2B，表示这里有三个男人，他们年龄的中位数是 30，平均数是 44。现在我们来枚举一下这三位男性的年龄组合，我们用 A,B,C 来代表这三个男人的年龄。首先，年龄是有一个范围的，不妨假设他们三个的年龄在 0 到 125 之间，也就是 $0 \le A, B, C \le 125$，根据可重复组合公式

$$\binom{126+3-1}{3} \approx 300k$$

可得现在有约 300k 个组合，然后我们进一步利用中位数为 30 和平均数 44 作为约束条件，只剩下表 2 所列出的 30 种可能性。我们仅仅只用了 2B 统计量就可以将范围锁定在 30 个组合之内，这足以展现重构攻击的威力。另外，上述约束问题 (线性规划) 的求解并不是 NP-Hard 的，反之现已有高效解决器工具 (SAT)。

在认识了重构攻击之后，我们知道了即使是群体的统计信息也会导致个体信息的泄露。一个自然的解决办法就是公布不准确的信息，这种不准确可以通过在信息中添加噪声来实现。那么，现在有一个问题，为了保护隐私不被泄露，至少要引入多少不准确性呢？为了讨论这个问题，首先引入一个简单的模型。在该模型中，数据库被抽象成一个 n 位

表 1: 一张虚构的人口统计表

		Age		
Statistic	**Group**	**Count**	**Median**	**Mean**
1A	Total Population	7	30	38
2A	Female	4	30	33.5
2B	Male	3	30	44
2C	Black or African American	4	51	48.5
2D	White	3	24	24
3A	Single Adults	(D)	(D)	(D)
3B	Married Adults	4	51	54
4A	Black or African American Female	3	36	36.7
4B	Black or African American Male	(D)	(D)	(D)
4C	White Male	(D)	(D)	(D)
4D	White Female	(D)	(D)	(D)
5A	Persons Under 5 Years	(D)	(D)	(D)
5B	Persons Under 18 Years	(D)	(D)	(D)
5C	Persons 64 Years or Over	(D)	(D)	(D)

表 2: Remaining Possible Results

A	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
1	30	101	11	30	91	21	30	81
2	30	100	12	30	90	22	30	80
3	30	99	13	30	89	23	30	79
4	30	98	14	30	88	24	30	78
5	30	97	15	30	87	25	30	77
6	30	96	16	30	86	26	30	76
7	30	95	17	30	85	27	30	75
8	30	94	18	30	84	28	30	74
9	30	93	19	30	83	29	30	73
10	30	92	20	30	82	30	30	72

表 3:

Name	Postal Code	Date of Birth	Sex	Has Disease?
Alice	K8V7R6	5/2/1984	F	1
Bob	V5K5J9	2/8/2001	M	0
Charlie	V1C7J	10/10/1954	M	1
David	R4K5T1	4/4/1944	M	0
Eve	G7N8Y3	1/1/1980	F	1

的 Boolean 向量，对于数据库的询问类型是子集询问 (subset query)。那么一个子集询问也可以用向量来表示，询问结果则为两个向量的内积。为了读者方便理解，接下来我举一个例子，请看表 3。

在表 3 中，敏感信息 (属性) 是"Has Disease?" 这一项，它的取值空间只有两个值 {0,1}，所以我们可以用一个 n 维 Boolean 向量来表示数据库。而询问的格式是" 在给出的集合中有多少人'Has Disease? $= 1$'"。同样，可以用一个 n 位 Boolean 向量来代表一个询问，把在询问集合内的对象的值置 1，反之置 0。表 3 中的数据库向量为 $A = \{1, 0, 1, 0, 1\}$，假设我们想要询问 Alice，Bob 和 Charlie 三人中'Has Disease?=1" 的人的数量，那么询问向量为 $Q = \{1, 1, 1, 0, 0\}$。这样，询问结果可以表示为 $A \cdot Q = \sum_{i=1}^{5} a_i q_i = 2$

现在接着讨论至少需要添加多少噪声这个问题。因为讨论的是噪声量级的下界，所以应该设置一个较为宽松的目标，也就是仅仅只防止最坏情况的出现，防止出现隐私灾难 (privacy catastrophe)。现对隐私" 灾难 " 进行一个确切的定义

Definition 1 (blatantly non-private)**.** *如果敌手能利用算法 R 来构造数据库 $c \in \{0,1\}^n$ 且 c 与真实数据库 d 至多有 $o(n)$ 个数据项不同，即 $\|c - d\|_1 = o(n)$, 则称算法 R 是 blatantly non-private 的*

在定义了隐私" 灾难 " 之后，接下来具体讨论噪声的量级与隐私泄露量级之间的关系。

Theorem 1. *如果一个机制 M 添加的噪声量级是 E, 那么存在敌手能重构数据库且至多只有 $4E$ 个数据项不同*

现给出 Theorem 1 的构造性证明，也就是直接构造出一个符合 Theorem 1 的攻击方法

Proof. 敌手的攻击算法分为两个阶段:

1. 构造 2^n 个询问集 S，获得询问 $M(S)$

2. 对于每个候选数据集 $c \in \{0,1\}^n$, 如果存在 S，使得 $\left|\sum_{i \in S} c_i - M(S)\right| > E$，则除去。否则，输出 c 并停止。

用 d 表示真实数据集，现将 d 分成两个不相交的子集 I_0, I_1，其中 $I_0 = \{i | d_i = 0\}, I_1 = \{i | d_i = 1\}$ 对于攻击算法输出的 c，有 $\left|\sum_{i \in I_0} c_i - M(I_0)\right| \leq E$ 且 $\left|\sum_{i \in I_1} c_i - M(I_1)\right| \leq E$

根据 Theorem 1 的假设，机制 M 添加噪声严格小于 E，那么 $\left|\sum_{i\in I_0} d_i - M(I_0)\right| \leq E$ 且 $\left|\sum_{i\in I_1} d_i - M(I_1)\right| \leq E$ 根据三角不等式可得，$\left|\sum_{i\in I_0} d_i - \sum_{i\in I_0} c_i\right| \leq 2E$ 且 $\left|\sum_{i\in I_1} d_i - \sum_{i\in I_1} c_i\right| \leq 2E$

因为 I_0, I_1 是真实数据集 d 的两个不相交子集且 $I_0 \cap I_1 = d$，所以两部分的误差可以直接相加，得 $\|c - d\|_1 \leq 4E$

□

根据 Theorem 1 可以得出一个简单的结论，如果一个机制 M 所添加的噪声量级为 $\frac{n}{401}$，那么敌手重构出的数据库至多只有 $\frac{n}{100.25}$ 个数据项不同，也就是至少有 $n - \frac{n}{100.25} \approx 0.99 * n$ 个数据项是相同的，实现了 99% 的重构率。

可以看到，根据 Theorem 1 给出的结论是十分可怕的，但是证明中使用的方法需要 2^n 次询问，这显然是不实际的，所以是否有高效可实现的攻击方法呢？

Theorem 2 ([6]). *如果攻击者被允许提出 $O(n)$ 次询问且机制 M 添加的噪声量级为 $E = O(\alpha\sqrt{n})$，那么存在敌手可以重构数据库且至多只有 $O(\alpha^2)$ 个数据项项不同。*

Theorem 2 的证明更为复杂、难以理解，所以接下来以一个较为直观的角度去解释它，其核心思想就是候选数据库得到的询问结果会与真实数据库得到的询问结果大概率相差很多，这种差异无法通过 $O(\sqrt{n})$ 量级的噪声来掩盖，所以攻击者一旦看到返回的询问结果与当前手中候选 (构造) 数据库得到的询问结果相差很多，那么就可以认定当前手中的候选数据库与真实数据库相差很多，从而排除掉手中的候选数据库。更进一步来说，因为可以进行 $O(n)$ 次询问，所以攻击者可以 $O(n)$ 进行次比对，取 union_bound 后，可以认为留下的候选数据库一定与真实数据库相差不会太大。

如果想看原版 Theorem 2 以及相应证明，可以参考 [7] 的章节 8.1。顺便提一嘴，[7] 可以说是差分隐私的教科书，如果想有入门的同学可以仔细看下。事实上，满足 Theorem 2 且高效的攻击方法是存在的，甚至存在一种高效攻击 [8] 能容许所有回答中有 0.239 比例带有不受约束的大噪声。

总结一下，第一种攻击方法需要 2^n 次询问来打破 $O(n)$ 噪声的防御，第二种攻击方式需要 $O(n)$ 次询问来打破 $O(\sqrt{n})$ 噪声的防御。在某种意义上来说，第二种攻击是" 紧凑的"，在理论上是最优的攻击手段。而差分隐私恰好允许在 $O(\sqrt{n})$ 的噪声量级下容许敌手进行 $O(n)$ 次的询问，也就是说差分隐私定义中的噪声是恰好能防御第二种攻击方式的最少噪声 (对于这一点我无法做更多的解释，因为没搞懂，请参考 [7] 的章节 8.2)。

2 差分隐私

2.1 差分隐私的定义

Definition 2 (Differential Privacy). *如果一个随机算法 $M : \mathcal{X}^n \to \mathcal{Y}$，对于所有相邻数据集 $X, X' \in \mathcal{X}^n$ 和所有 $T \subseteq \mathcal{Y}$ 满足*

$$Pr[M(X) \in T] \leq e^{\epsilon} Pr[M(X') \in T]$$

则称 M 是满足 $\epsilon - DP$ 的

直观来看，差分隐私衡量的是两个输出概率分布的相似性，ϵ 越小，两个分布的相似性就越高，反之两个分布的相似性越低，当 $\epsilon = 0$ 时，算法 M 对于任意数据集都具有相同的输出分布，也就是说该算法的输出与所依赖的数据集无关，是完全随机的。由于输入项的相邻性是对称的，所以差分隐私也具有对称性。

接来下将从假设检验的角度解释差分隐私 [9]，可以理解从敌手的角度来看待。现有像个假设 H_0 和 H_1

$$H_0 : \text{算法所依赖的数据集是 } X$$

$$H_1 : \text{算法所依赖的数据集是 } X'$$

设假阳率 (false positive) 为 p，也就是当 H_0 为真时选择了 H_1；设假阴率 (false negative) 为 q，也就是也就是当 H_1 为真时选择了 H_0，那么

$$\epsilon - DP \Rightarrow \begin{cases} e^{\epsilon}p + q \geq 1 \\ p + e^{\epsilon}q \geq 1 \end{cases}$$

从敌手的角度来看，肯定是希望 p 和 q 都同时小，但是 ϵ 限制了 p 和 q 的下界，当 ϵ 较小时，比如 $\epsilon = 0$ 时，$p + q \geq 1$，和随机猜并没有什么差别；而当 ϵ 较大时，由于 e^{ϵ} 的加权，p 和 q 都可以相对较小一些。

2.2 差分隐私的性质

Proposition 1 (Post-Processing). $M : \mathcal{X}^n \to \mathcal{Y}$ 是一个 $\epsilon - DP$ 的随机算法。那么对于任意随机映射 $f : \mathcal{Y} \to \mathcal{Y}'$，$f \odot M : \mathcal{X}^n \to \mathcal{Y}'$ 也是 $\epsilon - DP$ 的

Proof. 这里我们只需要对一个确定性函数 $f : \mathcal{Y} \to \mathcal{Y}'$ 进行证明就可以了。因为任意一个随机映射都可以分解称一组确定性函数的凸组合 (convex combination)。

首先给定一对相邻数据集 X, X', 对于任意 $S \subseteq \mathcal{Y}'$, 存在子集 $T = \{y \in \mathcal{Y} | f(y) \in S\} \subseteq \mathcal{Y}$

$$\begin{aligned} Pr[f(M(x)) \in S] &= Pr[M(x) \in T] \\ &\leq Pr[M(x') \in T] \\ &= Pr[f(M(x')) \in S] \end{aligned}$$

□

个人觉得这个性质非常重要，有种一劳永逸的感觉，但是维持 Post-Processing 性质的前提是不知道有关数据库的额外信息，也就是后续操作不引入在进行差分隐私操作前的信息。比如，我有一个 $\epsilon = 0$ 差分隐私机制 M，无论输入时什么，输出都是 1。现有两个输入，X=1 和 Y=100，显然是无法通过 M(X) 和 M(Y) 来猜测输入时 X 还是 Y，但是可以通过 f(M(X))=M(X)+X=2 和 f(M(Y))=M(Y)+Y=101 来进行猜测。

Proposition 2 (Basic Composition). 假设 $M = (M_1, M_2, \ldots, M_k)$ 是由 k 个 $\epsilon - DP$ 算法组成的序列，且本质上可以自适应的序列选择，那么 M 是 $k\epsilon - DP$ 的

Proof.

$$\begin{aligned}\frac{Pr[M(x) = y]}{Pr[M(x') = y]} &= \prod_{i=1}^{k} \frac{Pr[M_i(x) = y_i | (M_1(x), M_2(x), \ldots, M_{i-1}(x) = (y_1, y_2, \ldots, y_{i-1}))]}{Pr[M_i(x') = y_i | (M_1(x'), M_2(x'), \ldots, M_{i-1}(x') = (y_1, y_2, \ldots, y_{i-1}))]} \\ &= \prod_{i=1}^{k} e^{\epsilon} \\ &= e^{k\epsilon}\end{aligned}$$

□

自适应的序列选择的意思是 M_i 的输出取决于 $M_{i-1}, \ldots, M_1$ 的输出，这一点也在证明中得以体现。通过这条性质可以看到隐私参数随着序列长度线性增长，这也意味着如果约束整个序列的隐私参数为 ϵ，那么序列中的每个算法 M_i 的隐私参数应该设置为 $\epsilon' = \frac{\epsilon}{k}$。序列长度的增加伴随着 ϵ' 的减小，算法的随机性增强，实用性就会随着序列长度线性降低，所以需要一定的技术手段来减缓实用性衰退的速度。

Proposition 3 (Group Privacy). $M : \mathcal{X}^n \to \mathcal{Y}$ 是一个 $\epsilon - DP$ 的随机算法。那么对于所有 $|x - y|_1 \leq k$ 和所有 $S \subseteq Range(M)$ 有

$$Pr[M(x) \in S] \leq e^{k\epsilon} Pr[M(y) \in S]$$

Proof. 我们可以把 x,y 抽象为一张图上的两个点，这两个点的距离 $d(x, y) = k$，除了 x,y 一定还存在 k-1 个点来组成最短路径, 不妨设这条最短路径为 $(x, z_1, z_2, \ldots, z_k - 1, y)$，其中 $d(x, z_1) = d_(z_i, z_{i+1}) = d(z_{k-1}, y) = 1$，那么

$$\begin{aligned}Pr[M(x) \in S] &\leq e^{\epsilon} Pr[M(z_1) \in T] \\ &\leq e^{2\epsilon} Pr[M(z_2) \in T] \\ &\vdots \\ &\leq e^{(k-1)\epsilon} Pr[M(z_{k-1}) \in T] \\ &\leq e^{k\epsilon} Pr[M(y) \in S]\end{aligned}$$

□

关于 Group Privacy 的用处倒是没那么多遇见过，在我的理解中，Group Privacy 性质表明了差分隐私不仅仅只保护相邻输入项，也保护不相邻项，保护程度随着两个项的距离递减。

3 差分隐私的变种

前面所介绍的差分隐私可以称其为 PureDP，它并不允许任何形式、任何量级 (哪怕一点点) 的隐私泄露，这样会大大降低算法的实用性。所以，人们引入差分隐私的变种

(varient DP)，也可以理解为 PureDP 的松弛版本，来松弛隐私保证，用少量的隐私损失 (风险) 换取实用性的巨大提升，达到更好的 privacy-utility trade-off。

3.1 $(\epsilon, \delta) - DP$

Definition 3 ($(\epsilon, \delta) - DP$)**.** *如果一个随机算法 $M : \mathcal{X}^n \to \mathcal{Y}$，对于所有相邻数据集 $X, X' \in \mathcal{X}^n$ 和所有 $T \subseteq \mathcal{Y}$ 满足*

$$Pr[M(X) \in T] \leq e^{\epsilon} Pr[M(X') \in T] + \delta$$

则称 M 是满足 $(\epsilon, \delta) - DP$ 的

值得一提的是$(\epsilon, \delta) - DP$ 中的 δ 参数应该是 cryptographically negligible 的。用人话说就是 $\delta = negl(n)$，即对于所有正整数 c，都有 $\delta < \frac{1}{n^c}$。一般来说，会将 c 设置成 $\frac{1}{n^{1.1}}$ 举一个简单的例子来说明设置 $\delta < \frac{1}{n^c}$ 的原因。假设现有一个算法，它遍历 n 个点，对于每个点有 δ 的概率输出该点的值 (认为是敏感信息)，否则输出"$\perp$"。不难证明上述算法是 $(\epsilon, \delta) - DP$ 的, 其至少输出一个点的值的概率为 $1 - (1 - \delta)^n$，当 δ 较小时，经过泰勒展开的值约为 δn。如果不满足 $\delta < \frac{1}{n}$，那么这个算法就会有非平凡 (non-trivial) 的概率输出一个点的值。

那么 $(\epsilon, \delta) - DP$ 的松驰性体现在那里呢？接下来将做一个直观的解释，并不做完整严谨的证明。首先引入隐私损失随机变量

Definition 4 (Privacy Loss Random Variable)**.** *设 Y 和 Z 是两个随机变量，则它们的隐私损失随机变量为 $\mathcal{L}_{Y||Z} = \ln(\frac{Pr[Y=t]}{Pr[Z=t]})$，其中 $t \sim Y$。如果 Y 和 Z 的支集 (support) 不相等，那么将它们的隐私损失损失随机变量视为未定义的*

一个随机变量的支集定义为所有概率不为 0 的事件的并集。其实与其在 Y 和 Z 支集不相等的情况下认为隐私损失随机变量未定义，不如认为隐私损失随机变量是无穷的，因为支集不相等意味着存在着观察值 t，这个 t 只可能被 Y 或者 Z 之一输出，如果敌手掌握了这个观察值 t，那么就可以轻易判断依赖数据集是 Y 还是 Z。

第一次看到隐私损失随机变量的定义时，我想到了 KL(Kullback-Leibler) 散度，也叫相对熵，因为它们两个太像了。通过它们两个的定义不难发现，隐私损失随机变量的期望就是分布 Y 和 Z 的 KL 散度。

那么对于 $(\epsilon, \delta) - DP$ 一个直观的理解就是

$$(\epsilon, \delta) - DP \rightleftharpoons Pr[\mathcal{L}_{Y||Z} \geq \epsilon] \leq \delta$$

上述结论的充分性很容易证明，但是必要性的证明比较复杂，有兴趣的请参考 [7] 章节 3.5.1。对比 $\epsilon - DP$

$$\epsilon - DP \rightleftharpoons Pr[\mathcal{L}_{Y||Z} \geq \epsilon] = 0$$

可以看到，$(\epsilon, \delta) - DP$ 允许隐损失随机变量有 δ 的概率超出 ϵ，而 $\epsilon - DP$ 不允许。

通过以上对 $(\epsilon, \delta) - DP$ 的描述，不难发现 $(\epsilon, \delta) - DP$ 有一个严重的缺陷，就是 $(\epsilon, \delta) - DP$ 只约束了隐私损失随机变量大于 ϵ 的概率，但是并没有明确定义隐私损失的程度，也就

是说

$$Pr[\mathcal{L}_{Y||Z} = \infty] = \delta$$

和

$$Pr[\mathcal{L}_{Y||Z} = 1.1\epsilon] = \delta$$

在 $(\epsilon, \delta) - DP$ 定义下是无法区分的。

另外，$(\epsilon, \delta) - DP$ 相较于 $\epsilon - DP$ 的优点在于它提供了更好的 privacy-utility trade-off，体现在如下定理

Theorem 3 (Advanced Composition[10]). *对于所有 $\epsilon, \delta, \delta' > 0$，让 $M = (M_1, M_2, \ldots, M_k)$ 是由 k 个 $(\epsilon, \delta) - DP$ 算法组成的序列，其中 M_i 可以自适应的序列选择。那么序列 M 是 $(\bar{\epsilon}, \bar{\delta}) - DP$，其中*

$$\bar{\epsilon} = \epsilon\sqrt{2k\log\frac{1}{\delta'}} + k\epsilon\frac{e^\epsilon - 1}{e^\epsilon + 1} \qquad \bar{\delta} = k\delta + \delta'$$

一个简单的理解就是 $(\epsilon, \delta) - DP$ 的隐私衰退和序列长度成亚线性 (sublinear) 关系，即 $\bar{\epsilon} = \epsilon\sqrt{k}$。对比 $\epsilon - DP$ 是一个巨大的提升。

接下来就尝试一下去证明 Theorem 3，当证明复杂定理时，一个思路是构造一个简单的模型，并证明这个模型不失一般性 (generlization)，即对于任意情况都可以规约 (reduce) 到这个简单模型，那么这时只需要在这个模型上证明定理即可。

Proof.
1、构造一个简单模型。为了证明 Theorem 3，需要构造符合 $(\epsilon, \delta) - DP$ 的两个分布 U,V。即构造出构造出两个分布 U、V，满足 $Pr[\mathcal{L}_{U||V} \geq \epsilon] \leq \delta$ 且 $Pr[\mathcal{L}_{V||U} \geq \epsilon] \leq \delta$。

t	Pr[U=t]	Pr[V=t]
0	$\frac{e^\epsilon(1-\delta)}{\epsilon+1}$	$\frac{(1-\delta)}{\epsilon+1}$
1	$\frac{(1-\delta)}{\epsilon+1}$	$\frac{e^\epsilon(1-\delta)}{\epsilon+1}$
'I am U'	δ	0
'I am V'	0	δ

通过上述概率分布列不难看出 $Pr[\mathcal{L}_{U||V} \geq \epsilon] \leq \delta$ 且 $Pr[\mathcal{L}_{V||U} \geq \epsilon] \leq \delta$。

2、证明规约的可行性。

Lemma 1. *对于任意两个满足 $Pr[\mathcal{L}_{A||B} \geq \epsilon] \leq \delta$ 且 $Pr[\mathcal{L}_{B||A} \geq \epsilon] \leq \delta$ 的随机变量 A、B，存在随即映射 F 使得 $F(U)$ $A, F(U)$ B*

Proof. 给出构造性证明，映射可以遵循如下方式：分别将输出'I am U',0,1,'I am V' 对应到 $L(x) > e^\epsilon, 0 < L(x) \leq e^\epsilon, e^{-\epsilon} <= L(x) \leq 0, L(x) < e^{-\epsilon}$，其中 $L(x) = \frac{Pr[A=x]}{Pr[B=x]}$ 参考图 3 进行理解 □

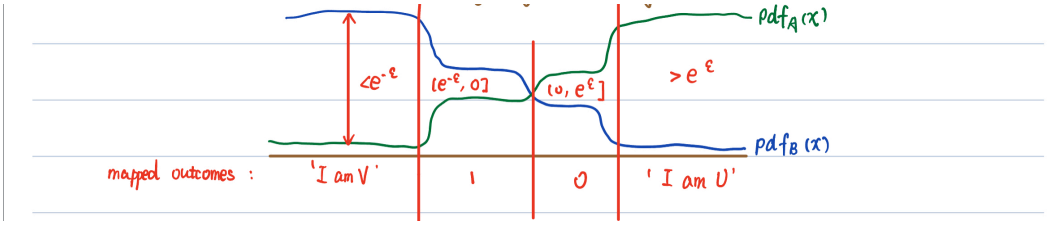

图 3: Lemma 1 示意图

那么根据差分隐私的 Post-Processing 性质，只需要证明分布 U,V 满足 Advanced Composition 定理即可。

Corollary 1. 对于任意由 k 个 $(\epsilon, \delta) - DP$ 算法组成的序列 M，存在随机映射 $F*$，满足如下性质：

对于一对相邻数据集下 x,x'

$M(x) \sim F^*(U_1, U_2, \ldots, U_k)$，其中 $U_1, U_2, \ldots, U_k$ 关于 U 独立同分布

$M(x) \sim F^*(V_1, V_2, \ldots, V_k)$，其中 $V_1, V_2, \ldots, V_k$ 关于 V 独立同分布

3、从 U,V 着手证明 Advanced Composition 定理

设事件 $E_1 = \{z = (z_1, z_2, \ldots, z_k) | \exists z_j (i \in [k]), z_j =' I \quad am \quad U'\}$

那么 $Pr_{U_1, U_2, \ldots, U_k}[E_1] = 1 - (1-\delta)^k \approx k\delta$(从分布 U 中采样 k 次，出现隐私泄露的概率)

对于任意 $z \in \{0,1\}^k$，计算当观察值为 z 时，U,V 的隐私损失变量的取值

$$\begin{aligned}
\ln \frac{Pr[U=z]}{Pr[V=z]} &= \ln \prod_{i=1}^{k} \frac{Pr[U_i = z_i]}{Pr[V_i = z_i]} \\
&= \sum_{i=1}^{k} \ln \frac{(1-\delta) exp(\epsilon(1-z_i))/(exp(\epsilon)+1)}{(1-\delta) exp(\epsilon z_i)/(exp(\epsilon)+1)} \\
&= \sum_{i=1}^{k} \epsilon(1-2z_i)
\end{aligned}$$

计算当不发生隐私灾难时 (后面有解释)，隐私损失随机变量的期望，即

$$\begin{aligned}
E_{z \sim (U_1, U_2, \ldots, U_k)} \left[\ln \frac{Pr[U=z]}{Pr[V=z]} | \bar{E_1} \right] &= \sum_{i=1}^{k} (\epsilon * Pr[z_i = 0 | \bar{E_1}] - \epsilon * Pr[z_i = 1 | \bar{E_1}]) \\
&= \sum_{i=1}^{k} (\epsilon * \frac{e^\epsilon}{e^\epsilon + 1} - \epsilon * \frac{1}{e^\epsilon + 1}) \\
&= k\epsilon \frac{e^\epsilon - 1}{e^\epsilon + 1}
\end{aligned}$$

设事件 E_2 表示隐私损失随机变量偏离期望值大于 $t\sqrt{k}\epsilon$ 的情况，象征着离群概率。

$$E_2 = \{z | \ln \frac{Pr[U=z]}{Pr[V=z]} > k\epsilon \frac{e^\epsilon - 1}{e^\epsilon + 1} + t\sqrt{k}\epsilon\}$$

仔细想想，其实 $(\epsilon, \delta) - DP$ 约束的就是隐私损失随机变量的离群概率。那么通过什么工具可以获得离群概率的上界呢？集中不等式 (Concentration Inequity)!

这里选取切诺夫不等式 (Chernoff inequity)，又叫切诺夫界 (Chernoff bound) 来估计离群概率上界。一般切诺夫界用于 n 个独立同分布的伯努利随机变量和，但是它也有非伯努利分布版本

Theorem 4 (Chernoff Bound(non-Bernoulli version))**.** 令 $X_1, X_2, \ldots, X_n$ 为 n 个随机变量，其中 $a \le X_i \le b$。令 $X = \sum_{i=1}^n X^i$,$u = E(X)$，则

$$Pr[X \ge u + \tau] \le e^{-\frac{2\tau^2}{n(b-a)^2}}$$

$$Pr[X \le u - \tau] \le e^{-\frac{\tau^2}{n(b-a)^2}}$$

有兴趣的可以证明一下，大致思路是求出矩生成函数 (MGF) 的上界，然后代入马尔可夫不等式 (Markov's Inequity)。

话接正题，对于一个 k 随机序列的隐私损失随机变量本身就是 k 个随机变量的和，即

$$\mathcal{L}_{U||V} = \ln \frac{Pr[U=z]}{Pr[V=z]} = \sum_{i=1}^{k} \epsilon(1 - 2z_i)$$

不难看出,

$$-\epsilon \le \epsilon(1 - 2z_i) \le \epsilon$$

在不发生隐私损失的情况下，其期望为

$$E_{z\sim(U_1,U_2,\ldots,U_k)} \left[\ln \frac{Pr[U=z]}{Pr[V=z]} \mid \bar{E_1} \right] = k\epsilon \frac{e^\epsilon - 1}{e^\epsilon + 1}$$

有了上述条件，就能通过 Chernoff Bound 求出，当允许发生隐私损失的概率上界为 δ' 时，随机序列隐私损失界的位置，也就是 $\bar{\epsilon}$ 的值。

将上述相应的统计量带入 Chernoff Bound，得

$$\begin{aligned} Pr\left[E_2 | \bar{E_1}\right] &= Pr\left[\ln \frac{Pr[U=z]}{Pr[V=z]} > k\epsilon \frac{e^\epsilon - 1}{e^\epsilon + 1} + t\sqrt{k}\epsilon \right] \\ &\le \exp(-\frac{2t^2 k\epsilon^2}{k4\epsilon^2}) \\ &= \exp(-\frac{t^2}{2}) \end{aligned}$$

$Pr[E_2|\bar{E_1}]$ 表示在没有发生隐私”灾难“时，隐私损失随机变量得离群概率，根据 $(\epsilon, \delta) - DP$ 的定义，将其上界设为 δ'，即令

$$\exp(-\frac{2t^2 k\epsilon^2}{k4\epsilon^2}) = \delta'$$

得,

$$t = \sqrt{-2\ln\bar{\delta}} = \sqrt{2\ln\frac{1}{\delta'}}$$

整理即得

$$Pr[E_2|\bar{E_1}] = Pr\left[\ln\frac{Pr[U=z]}{Pr[V=z]} > k\epsilon\frac{e^\epsilon-1}{e^\epsilon+1} + \epsilon\sqrt{2k\ln\frac{1}{\delta'}}\right] \leq \delta'$$

现在开始进行整理

令

$$\bar{\epsilon} = k\epsilon\frac{e^\epsilon-1}{e^\epsilon+1} + \epsilon\sqrt{2k\ln\frac{1}{\delta'}}$$

得

$$Pr[U = Z \cap \bar{E_1} \cap \bar{E_2}] \leq e^{\bar{\epsilon}} Pr[V = Z \cap \bar{E_1} \cap \bar{E_2}] \leq e^{\bar{\epsilon}} Pr[V = Z] \tag{1}$$

$$Pr[U = Z \cap E_1] \leq Pr[E_1] \leq k\delta \tag{2}$$

$$Pr[U = Z \cap \bar{E_1} \cap E_2] \leq Pr[E_2 \cap \bar{E_1}] = Pr[E_2|\bar{E_1}] * Pr[\bar{E_1}] \leq Pr[E_2|\bar{E_1}] \leq \delta' \tag{3}$$

那么根据全概率公式，可得

$$\begin{aligned} Pr[U = Z] &= Pr[U = Z \cap \bar{E_1} \cap \bar{E_2}] + Pr[U = Z \cap E_1] + Pr[U = Z \cap \bar{E_1} \cap E_2] \\ &\leq e^{\bar{\epsilon}} Pr[V = Z] + k\delta + \delta' \end{aligned}$$

□

再一次捋顺 Advanced Composition 证明，对 $(\epsilon, \delta) - DP$ 有了新的理解，观察证明最后的全概率公式，不难发现 $(\epsilon, \delta) - DP$ 的定义包含三个部分，分别是

1、隐私损失 bounded 住的情况，对应公式 (1)，对于该情况，需要通过参数 ϵ 来指明 upperbound;

2、发生隐私" 灾难" 的，也就是隐私完全暴露的情况，对应公式 (2)，只需指明概率上界即可;

3、没有发生隐私" 灾难"，但是隐私损失超过了参数 ϵ 指明的 upperbound，对应公式 (3)，这时也是只需指明概率上界。

通过观察不等式 (1)、(2)、(3) 的放缩，不难发现即使是 Advanced Composition 也不是最紧 (tight) 的，也就是说在理论上对于一个 k 随机序列 (k-fold)，是可能达到比 Advanced Composition 更小的隐私损失的。

4 Conclusion

总的来说，这篇文章写的很匆忙，也有许多需要修改的地方。在我的设想中，Introduction 应该写明白差分隐私的优越性，虽然说开头提到的三种隐私保护技术的优缺点还是讲的比较明白，且对于重构攻击的讲解应该也是比较简单易懂的，但是最后没有说明白根据差分隐私定义来添加的噪声是 match 抵御重构攻击所需噪声的 lowerbound 的。

对于差分隐私的定义做了介绍，以及从假设检验的角度来解释差分隐私。其实从假设检验角度来解释并不利于直观易解，而是 operational，换句话说利于系统机制的设计。个人觉得从散度 (divergence) 的角度理解是最易懂的。介绍差分隐私的定义同时，本文也讨论了差分隐私几个重要的性质，这几个性质可以说是差分隐私的核心。

由于 PureDP 的实用性太差了，现在更多用的是差分隐私的变种 (松弛版本)。本文所介绍的 $(\epsilon, \delta) - DP$ 应该是最早提出也是比较常用的差分隐私变种。本文通过隐私损失随机变量直观地解释了 $(\epsilon, \delta) - DP$ 的松弛性，并介绍了 Advanced Composition，量化解释了 $(\epsilon, \delta) - DP$ 的松弛程度。本文给出的 Advanced Composition 的证明虽然说十分繁琐，但是是比较容易理解的，可以对比 [7] 的章节 3.5 的证明。证明的方法其实也值得学习，建模以及规约有利于将证明化繁为简。另外，理解整个证明的思路有助于加深对 $(\epsilon, \delta) - DP$ 的理解。

本文也稍微提到了 $(\epsilon, \delta) - DP$ 的局限性之一，只定义了隐私损失的概率，并不能区分隐私损失的程度。另一局限性是在实际使用 Advanced Composition 时，会涉及到三个参数 $\epsilon, \delta, \delta'$，出于隐私性能的原因，需要进行参数的搜索，从而导致参数的组合爆炸 (combinational explosion)。为了解决上述两个局限性以及提高 privacy-utility 的 trade-ff，提出了基于 Ŕenyi divergence 的 Ŕenyi-DP(RDP)[11]。当然，差分隐私的变种并不止这些。比如，相比于之前基于 divergence 定义的 DP，Gaussian Differential Privacy[12] 基于假设检验来定义；相比于一般 DP 是 information theoretic(只要敌手没有额外信息，无论有多少算力都没办法破解)，Computational Differential Privacy[13] 针对于有限算力的敌手；针对没有可信服务器的分布式环境下的 LocalDP(LDP)[14]，LDP 是记录级细粒度的差分隐私 (传统差分隐私是定义在相邻数据集下)，隐私保护程度之高导致了其实用性十分的差，现有通过密码技术 (混洗模型 [15]，一种匿名信道技术) 来提升 LDP 性能的，LDP 和联邦学习 (federated learning) 的关系比较紧密。

本文对于差分隐私只讲了定义层面，对于新手来说对于如果应用差分隐私应该还是一头雾水。限于时间原因并没有将差分隐私的 building block。目前主流实现差分隐私机制是基于噪声的机制，比如高斯机制、拉普拉斯机制、离散高斯机制 [16]，也有其他不加噪声的机制，比如随机响应 (Randomized Response)、指数机制等等。

本文本来是想介绍差分隐私深度学习，如何将差分隐私与深度学习算法结合以及之中存在的问题，可惜 DDL 到了。

References

[1] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.

[3] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

[4] Jiang, Honglu, et al. "Applications of differential privacy in social network analysis: a survey." *IEEE Transactions on Knowledge and Data Engineering* (2021).

[5] Garfinkel, Simson, John M. Abowd, and Christian Martindale. "Understanding database reconstruction attacks on public data." *Communications of the ACM* 62.3 (2019): 46-53.

[6] Dinur, Irit, and Kobbi Nissim. "Revealing information while preserving privacy." *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 2003

[7] Dwork, Cynthia, and Aaron Roth. "The algorithmic foundations of differential privacy." *Found. Trends Theor . Comput. Sci.* 9.3-4 (2014): 211-407.

[8] Dwork, Cynthia, Frank McSherry, and Kunal Talwar. "The price of privacy and the limits of LP decoding." *Proceedings of the thirty-ninth annual ACM Symposium on Theory of Computing*. 2007.

[9] Wasserman, Larry, and Shuheng Zhou. "A statistical framework for differential privacy." *Journal of the American Statistical Association* 105.489 (2010): 375-389.

[10] Dwork, Cynthia, Guy N. Rothblum, and Salil Vadhan. "Boosting and differential privacy." *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, 2010.

[11] Mironov, Ilya. "Rényi differential privacy." *2017 IEEE 30th computer security foundations symposium (CSF)*. IEEE, 2017.

[12] Dong, Jinshuo, Aaron Roth, and Weijie J. Su. "Gaussian differential privacy." *arXiv preprint arXiv*:1905.02383 (2019).

[13] Mironov, Ilya, et al. "Computational differential privacy." *Annual International Cryptology Conference*. Springer, Berlin, Heidelberg, 2009.

[14] Duchi, John C., Michael I. Jordan, and Martin J. Wainwright. "Local privacy and statistical minimax rates." *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE, 2013.

[15] Cheu, Albert, et al. "Distributed differential privacy via shuffling." *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, Cham, 2019.

[16] Canonne, Clément L., Gautam Kamath, and Thomas Steinke. "The discrete gaussian for differential privacy." *Advances in Neural Information Processing Systems* 33 (2020): 15676-15688.