Label Differential Privacy におけるラベル情報漏洩の要因分析

関口 ひなた　(指導教員：小口 正人)

1　はじめに

機密データの共有は重要だが，プライバシ侵害のリスクを伴うため，信頼性の高いプライバシ保護技術が求められる．差分プライバシ（Differential Privacy: DP）は，データにノイズを加えることでプライバシを保証する手法であり，実用化が進んでいる．しかし，ノイズ付加によるデータ有用性の低下が課題となっている．ラベル差分プライバシ（Label DP）は，特徴量が非センシティブでラベル情報がセンシティブである場合に，ラベル情報にのみノイズを加えることで学習性能を維持しつつ，ラベルのプライバシを保護する手法である [3]．ただし，ラベルに関連する特徴量データからの情報漏洩には対応できない．

本研究では，特徴量公開によるプライバシリスクを評価し，「予期せぬ漏洩」を定義して要因を解明する．

2　準備

本研究では次の前提条件を基に議論を進める：まず，$\mathcal{X}$ は特徴量空間を，$\mathcal{Y}$ はバイナリのラベル空間を表す．また，入力特徴量とラベルの相関を表す $\mathcal{X} \times \mathcal{Y}$ 上の結合分布 $\mathcal{P}$ が存在すると仮定する．さらに，$\mathcal{X}$ 上の周辺分布を $\mathcal{P}(\mathcal{X})$，x を与えたときの y の条件付き分布を $\mathcal{P}(\mathcal{Y}|x)$ と表記する．最後に，$\mathcal{D} \coloneqq (\mathcal{X} \times \mathcal{Y})^n$ はデータセットのドメインを表す．

2.1　プライバシ保護の定義

差分プライバシ（Differential Privacy: DP）は，統計値の公開におけるプライバシ漏洩を定量化する数学的な定義である．

定義 2.1.1 (ε-ラベル差分プライバシ)**.** *[3]* メカニズム $\mathcal{M}$ が，最大で *1* つのラベルのみ異なるデータセット D, D' と出力サブセット Z に対して $\frac{\mathbb{P}[\mathcal{M}(D)\in Z]}{\mathbb{P}[\mathcal{M}(D')\in Z]} \leq e^{\varepsilon}.$ を満たす場合，*ε-Label DP* を満たす．

定義 2.1.2 (ε-部分センシティブ差分プライバシ)**.** *[1]* 特徴量空間が，センシティブな特徴量空間と非センシティブな特徴量空間で構成されているとする（$\mathcal{X} = \mathcal{X}_{priv} \times \mathcal{X}_{pub}$）．このとき，最大で *1* つのラベルとセンシティブな特徴量が異なるデータセット D, D' と出力サブセット Z に対して，$\frac{\mathbb{P}[\mathcal{M}(D)\in Z]}{\mathbb{P}[\mathcal{M}(D')\in Z]} \leq e^{\varepsilon}.$ を満たす場合，*ε-semi sensitive DP* を満たす．

プライバシ保護手法．本研究では，プライバシ保護手法として ε-DP を満たすラプラス機構と，ε-Label DP を満たすランダム応答機構（RR）を取り上げる．ラプラス機構 [2] は，元のデータにラプラス分布に従うノイズを加えることで，データの改変を行う．RR [3] は，特徴量には処理を加えず，真のラベルを一定の確率で別のラベルに置き換える操作を行う．この手法では，攻撃者が元のラベルを推測する確率を制限することが可能であり，ε によってプライバシ保護の強度を調整できる．数式では，$\mathbb{P}(\tilde{y}_i = y_i) = 1 - \pi, \quad \mathbb{P}(\tilde{y}_i \neq y_i) = \pi,$ と表現される．ここで，反転確率は $\pi = \frac{1}{1+e^{\varepsilon}}$ である．

2.2　ラベル推論攻撃

攻撃の定式化．本研究では，攻撃性能期待値を評価するために，先行研究 [4] に基づき，以下のように定義する．この定式化に基づき，攻撃成功確率の評価を行う枠組みを構築する．

定義 2.2.1 (個別攻撃性能期待値 [4])**.** 入力データ x_i における個別攻撃性能期待値（Individual Expected Attack Utility）は次のように定義される：

$$\mathrm{IEAU}_i(\mathcal{A}, \mathcal{M}, \mathcal{P}, x_i) = \mathop{\mathbb{P}}_{y_i\sim\mathcal{P}(\mathcal{Y}|x_i)(\mathbf{X}^{(-i)},\mathbf{y}^{(-i)})\sim\mathcal{D}^{m-1},\ coins\ of\ \mathcal{M}} \Big(\mathcal{A}(\mathbf{X}, \mathcal{M}(\mathbf{X}, \mathbf{y}))_i = y_i \,\Big|\, x_i\Big)$$

ここで，$\mathbf{X}^{(-i)}$ は i 番目の要素 x_i を除いた $\mathbf{X}$ を表し，同様に $\mathbf{y}^{(-i)}$ は $\mathbf{y}$ から i 番目の要素を除いたものである．$\mathrm{IEAU}_i(\mathcal{A}, \mathcal{M}, \mathcal{P}, x_i)$ は，x_i に関連付けられたラベル y_i が条件付き分布 $\mathcal{P}(\mathcal{Y}|x_i)$ からサンプリングされる場合における，特定のデータ x_i に対する（期待される）攻撃効用を強調するものである．

攻撃者の定式化．本研究では，ε-Label DP が匿名化ラベルのみを制御し，攻撃者 $\mathcal{A}_{naive}$ を想定している現状の課題を明確化し，攻撃成功確率を評価する枠組みを構築する．攻撃者は 3 種類に分類される．1 つ目は，条件付き確率 $\mathbb{P}(y \mid x)$ と特徴量 $\mathbf{X}$ に基づいて攻撃を行う $\mathcal{A}_{uninformed}$ である．2 つ目は，匿名化ラベル $\tilde{\mathbf{y}}$ と反転確率 π に基づいて攻撃を行う $\mathcal{A}_{naive}$ である．3 つ目は，$\mathbb{P}(y \mid x)$，$\mathbf{X}$，π，およびメカニズム出力 $(\mathbf{X}, \tilde{\mathbf{y}})$ に基づいて最適な攻撃を行う $\mathcal{A}_{informed}$ である [4]．

3　ラベル差分プライバシの限界

ラベル差分プライバシ（Label DP）は，ラベルの匿名性を保証しながら特徴量を公開し，モデル精度を向上させる手法である．しかし，公開された特徴量とラベルに強い相関がある場合，特徴量からラベルが推測されるリスクがある．例えば，血圧が特徴量でラベルが高血圧症である場合，匿名化したラベルは意味を持たない．このような漏洩は，Label DP の保証範囲を超える問題である．本研究では，特徴量によるラベル情報の漏洩を「予期せぬ漏洩」と定義し，これを定式化する．

定義 3.0.1. i 番目のデータポイント x_i が固定されたとき，個別予期せぬ漏洩（Individual UnExpected attack Advantage ）は次のように定義される*:*

$$\mathrm{IUEAdv}(\mathcal{M}, \mathcal{D}, x_i) = \sup_{\mathcal{A}_{informed}} \mathrm{IEAU}(\mathcal{A}_{informed}, \mathcal{M}, \mathcal{D}, x_i) - \mathrm{IEAU}(\mathcal{A}_{naive}, \mathcal{M}, \mathcal{D}, x_i)$$

命題 3.0.2. ランダム応答機構を適用するメカニズム *[3]*$\mathcal{M}_{RR} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \tilde{\mathcal{Y}}$ を用いた場合，予期せぬ漏洩 IUEAdv の最大値は ε*,* および事前確率 $\mathbb{P}(y_i)$ と

事後確率 $\mathbb{P}(y_i \mid x_i)$ に依存する：

$$\mathrm{IUEAdv}(\mathcal{M}_{RR}, \mathcal{D}, x_i) \leq \frac{1}{1+\frac{\mathbb{P}(1-y_i|x_i)}{\mathbb{P}(y_i|x_i)}e^{-\varepsilon}} - \frac{1}{1+\frac{\mathbb{P}(1-y_i)}{\mathbb{P}(y_i)}e^{-\varepsilon}}$$

系 3.0.3. $k = \frac{\frac{\mathbb{P}(y_i|x_i)}{\mathbb{P}(y_i)}}{\frac{\mathbb{P}(1-y_i|x_i)}{\mathbb{P}(1-y_i)}} > 1$ を満たすとき，予期せぬ漏洩は発生しうる．また，予期せぬ漏洩の上限は k に比例して増加する．

4 ε-部分センシティブ差分プライバシ

3 章で述べた課題に対処しつつ，ユーティリティを低下させないようにするための定義として，ε-semi sensitive DP（定義 2.1.2）が提案されている．

この定義を満たすメカニズムとして，先行研究 [1] では，センシティブな特徴量にはラプラス機構を，ラベルには RR をそれぞれ適用し，非センシティブな特徴量にはノイズを加えずに学習を行う手法が提案されている．

命題 4.0.1. あるメカニズム $cM : \mathcal{X}^{pub} \times \mathcal{X}^{priv} \times \mathcal{Y} \mapsto \mathcal{X}^{pub} \times \mathcal{X}^{priv} \times \mathcal{Y}$ について，出力が $\tilde{x}_{priv} = f_{Lap}(x_{priv}), x_{pub}, \tilde{y} = RR(y)$ であったとき，以下を満たす：

$$\frac{\mathbb{P}(y=1 \mid \tilde{x}_{priv}, x_{pub}, \tilde{y})}{\mathbb{P}(y=0 \mid \tilde{x}_{priv}, x_{pub}, \tilde{y})} \leq e^{\varepsilon_1+\varepsilon_2} \cdot \frac{\mathbb{P}(y=1 \mid x_{pub})}{\mathbb{P}(y=0 \mid x_{pub})}$$

ここで，$f_{Lap} : \mathcal{X} \mapsto \mathcal{X}$ は ε'-*DP* を，$RR : \mathcal{Y} \mapsto \mathcal{Y}$ は ε_2-*Label DP* を満たすとし，$\varepsilon_1 = \max\{\varepsilon', \ln \mid \max_{x \in \mathcal{X}_{pub}}\{\frac{\mathbb{P}(x|y=1)}{\mathbb{P}(x|y=0)}\} \mid\}$ とする．

この命題から，センシティブ，非センシティブな特徴量の分布とラベルの関係が強い場合，ラベルのプライバシ保証が劣化する可能性があることが分かる．

5 実験

本章では，ラベル差分プライバシにおける「予期せぬ漏洩」と，条件付き特徴量における ε-ラベル差分プライバシを満たすメカニズムの評価を目的とした実験について説明する．本章では，$k = \frac{\frac{\mathbb{P}(y|\tilde{x})}{\mathbb{P}(y)}}{\frac{\mathbb{P}(y'|\tilde{x})}{\mathbb{P}(y')}}$ とし，これを特徴量とラベルの関係性の強さの指標とする．

シミュレーションデータセット． 特徴量をガウス分布 $\mathcal{N}(e_i, \sigma^2 I_{10})$ からサンプリングし，各次元を $[0,1]$ に正規化した 10 次元データセットを作成する．ラベルは $y \in \{0,1\}$ とし，以下の 2 種類を構築する：

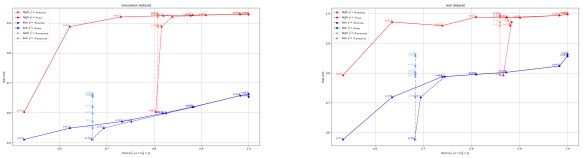
1. $\mathcal{D}_{\mathbf{high}}$: ラベルにより平均ベクトルが大きく異なる．
2. $\mathcal{D}_{\mathbf{low}}$: ラベル間で平均ベクトルの差が小さい．

実データセット． Breast Cancer データセット*から以下の 2 種類を構築する：

1. $\mathcal{D}_{\mathbf{high}}$: 相互情報量が最も高い 10 個の特徴量を使用．
2. $\mathcal{D}_{\mathbf{low}}$: 相互情報量が最も低い 10 個の特徴量を使用．

これらのデータセットを用いて，特徴量とラベルの関係性がプライバシ漏洩に与える影響を検証する．

*https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data



(a) シミュレーション　(b) Breast Cancer

図 1: 予期せぬ漏洩の比較結果．

予期せぬ漏洩の評価． 本章では，ラベル差分プライバシの下で発生する「予期せぬ漏洩」の程度を検証するための実験結果について述べる．グラフ内の赤色の線は $\mathcal{D}_{\text{high}}$ の結果を，青色の線は $\mathcal{D}_{\text{low}}$ の結果を表している．点線は攻撃者 $\mathcal{A}_{\text{informed}}$，実線は攻撃者 $\mathcal{A}_{\text{naive}}$，薄い線は攻撃者 $\mathcal{A}_{\text{uninformed}}$ の結果を示す．横軸は攻撃者の正解率（プライバシ），縦軸は各 ε で訓練したモデルのテスト AUC（ユーティリティ）を表す．図 1(a) はシミュレーションデータ，図 1(b) は実データの結果を示す．実世界では正確な $\mathbb{P}(y_i \mid x_i)$ は不明のため，LLM（GPT 4o）を用いて $\mathbb{P}'(y_i \mid x_i)$ を再現し $\mathcal{A}_{\text{uninformed}}$ の結果とした．両データセットで，命題 3.0.2 の通り，赤い線（k が大きい）は実線と点線の差が大きく，特徴量とラベルの関係性が強いほど予期せぬ漏洩が増加している．

6 結論

本研究では，公開された特徴量によるラベル情報の漏洩を「予期せぬ漏洩」と定義し，従来の ε-ラベル差分プライバシの限界を明らかにした．特に，特徴量とラベルが強く相関する場合，この漏洩が ε のみでは制御困難であることを示した．また，部分的に特徴量を匿名化するメカニズムにおいても，特徴量とラベルの関係に依存してラベル情報が漏洩することを理論的に示した．本研究は，特徴量とラベルの相関を考慮した新たなプライバシ基準の必要性を示している．

参考文献

[1] Chua, L., Cui, Q., Ghazi, B., Harrison, C., Kamath, P., Krichene, W., Kumar, R., Manurangsi, P., Narra, K. G., Sinha, A., et al.: Training differentially private ad prediction models with semi-sensitive features, *arXiv preprint arXiv:2401.15246* (2024).

[2] Dwork, C.: Differential privacy, in *Proceedings of the 33rd international conference on Automata, Languages and Programming-Volume Part II*, pp. 1–12Springer-Verlag (2006).

[3] Ghazi, B., Golowich, N., Kumar, R., Manurangsi, P. and Zhang, C.: Deep Learning with Label Differential Privacy, *Advances in Neural Information Processing Systems*, Vol. 34, (2021).

[4] Robert Istvan Busa-Fekete, C. G. A. M. m. A. S., Travis Dick and Swanberg, M.: Auditing privacy mechanisms via label inference attacks, *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (2024).

A　定理・命題・系の証明集

A.1　命題 3.0.2 の証明

IUEAdv は $\mathbb{P}(y_i \mid x_i, \tilde{y}_i) - \mathbb{P}(y_i \mid \tilde{y}_i)$ として表される.

Proof.

$$
\begin{aligned}
&\text{IUEAdv} \\
&= \mathbb{P}(Y = y_i \mid \mathcal{M}_{RR}(\mathbf{X}, \mathbf{y})_i = (x_i, \tilde{y}_i)) - \mathbb{P}(Y = y_i \mid RR(\mathbf{y})_i = \tilde{y}_i) \\
&= \mathbb{P}(y_i \mid x_i, \tilde{y}_i) - \mathbb{P}(y_i \mid \tilde{y}_i) \\
&= \mathbb{P}(y_i \mid x_i, \tilde{y}_i) - \frac{\mathbb{P}(y_i, \tilde{y}_i)}{\mathbb{P}(\tilde{y}_i)} \\
&= \mathbb{P}(y_i \mid x_i, \tilde{y}_i) - \frac{\mathbb{P}(\tilde{y}_i \mid y_i)\,\mathbb{P}(y_i)}{\mathbb{P}(\tilde{y}_i \mid y_i)\,\mathbb{P}(y_i) + \mathbb{P}(\tilde{y}_i \mid 1 - y_i)\,\mathbb{P}(1 - y_i)} \\
&\leq \mathbb{P}(y_i \mid x_i, \tilde{y}_i) - \frac{\frac{e^\varepsilon}{1+e^\varepsilon}\,\mathbb{P}(y_i)}{\frac{e^\varepsilon}{1+e^\varepsilon}\,\mathbb{P}(y_i) + \frac{1}{1+e^\varepsilon}\,\mathbb{P}(1 - y_i)} \\
&= \mathbb{P}(y_i \mid x_i, \tilde{y}_i) - \frac{e^\varepsilon\,\mathbb{P}(y_i)}{e^\varepsilon\,\mathbb{P}(y_i) + \mathbb{P}(1 - y_i)} \\
&= \mathbb{P}(y_i \mid x_i, \tilde{y}_i) - \frac{1}{1 + \frac{\mathbb{P}(1-y_i)}{\mathbb{P}(y_i)} e^{-\varepsilon}} \\
&\leq \frac{1}{1 + \frac{\mathbb{P}(1-y_i \mid x_i)}{\mathbb{P}(y_i \mid x_i)} e^{-\varepsilon}} - \frac{1}{1 + \frac{\mathbb{P}(1-y_i)}{\mathbb{P}(y_i)} e^{-\varepsilon}} \quad (\because (1))
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}(y_i \mid x_i, \tilde{y}_i) \qquad (1) \\
&= \frac{\mathbb{P}(\tilde{y}_i \mid x_i, y_i)\,\mathbb{P}(y_i \mid x_i)}{\mathbb{P}(\tilde{y}_i \mid x_i)}
\end{aligned}
$$

ここで，$\mathcal{M}_{RR}$ の出力は x_i に依存しないので,

$$
\begin{aligned}
&= \frac{\mathbb{P}(\tilde{y}_i \mid y_i)\,\mathbb{P}(y_i \mid x_i)}{\mathbb{P}(\tilde{y}_i \mid x_i)} \\
&= \frac{\mathbb{P}(\tilde{y}_i \mid y_i)\,\mathbb{P}(y_i \mid x_i)}{\mathbb{P}(\tilde{y}_i \mid y_i)\,\mathbb{P}(y_i \mid x_i) + \mathbb{P}(\tilde{y}_i \mid 1 - y_i)\,\mathbb{P}(1 - y_i \mid x_i)} \\
&\leq \frac{\frac{e^\varepsilon}{1+e^\varepsilon}\,\mathbb{P}(y_i \mid x_i)}{\frac{e^\varepsilon}{1+e^\varepsilon}\,\mathbb{P}(y_i \mid x_i) + \frac{1}{1+e^\varepsilon}\,\mathbb{P}(1 - y_i \mid x_i)} \\
&= \frac{e^\varepsilon\,\mathbb{P}(y_i \mid x_i)}{e^\varepsilon\,\mathbb{P}(y_i \mid x_i) + \mathbb{P}(1 - y_i \mid x_i)} \\
&= \frac{1}{1 + e^{-\varepsilon}\frac{\mathbb{P}(1-y_i \mid x_i)}{\mathbb{P}(y_i \mid x_i)}}
\end{aligned}
$$

□

A.2　系 3.0.3 の証明

Proof.

$$
\begin{aligned}
&\frac{\frac{\mathbb{P}(y_i \mid x_i)}{\mathbb{P}(y_i)}}{\frac{\mathbb{P}(1-y_i \mid x_i)}{\mathbb{P}(1-y_i)}} > 1 \\
&\frac{\mathbb{P}(1 - y_i)}{\mathbb{P}(y_i)} \cdot \frac{\mathbb{P}(y_i \mid x_i)}{\mathbb{P}(1 - y_i \mid x_i)} > 1 \\
&\frac{\mathbb{P}(1 - y_i)}{\mathbb{P}(y_i)} > \frac{\mathbb{P}(1 - y_i \mid x_i)}{\mathbb{P}(y_i \mid x_i)} \\
&\frac{1}{1 + \frac{\mathbb{P}(1-y_i \mid x_i)}{\mathbb{P}(y_i \mid x_i)} e^{-\varepsilon}} > \frac{1}{1 + \frac{\mathbb{P}(1-y_i)}{\mathbb{P}(y_i)} e^{-\varepsilon}}
\end{aligned}
$$

□

A.3　命題 4.0.1 の証明

Proof.

$$
\begin{aligned}
&\mathbb{P}(y = 1 \mid x_{\text{pub}}, \tilde{x}_{\text{priv}}, \tilde{y}) \\
&= \frac{\mathbb{P}(x_{\text{pub}}, \tilde{x}_{\text{priv}}, \tilde{y} \mid y = 1)\,\mathbb{P}(y = 1)}{\sum_{c \in \mathcal{Y}} \mathbb{P}(x_{\text{pub}}, \tilde{x}_{\text{priv}}, \tilde{y} \mid y = c)\,\mathbb{P}(y = c)} \\
&= \frac{\mathbb{P}(\tilde{x}_{\text{priv}}, \tilde{y} \mid x_{\text{pub}}, y = 1)\,\mathbb{P}(x_{\text{pub}} \mid y = 1)\,\mathbb{P}(y = 1)}{\sum_{c \in \mathcal{Y}} \mathbb{P}(\tilde{x}_{\text{priv}}, \tilde{y} \mid x_{\text{pub}}, y = c)\,\mathbb{P}(x_{\text{pub}} \mid y = c)\,\mathbb{P}(y = c)} \\
&= \frac{\mathbb{P}(\tilde{x}_{\text{priv}}, \tilde{y} \mid x_{\text{pub}}, y = 1)\frac{\mathbb{P}(y=1 \mid x_{\text{pub}})\,\mathbb{P}(x_{\text{pub}})}{\mathbb{P}(y=1)}\,\mathbb{P}(y = 1)}{\sum_{c \in \mathcal{Y}} \mathbb{P}(\tilde{x}_{\text{priv}}, \tilde{y} \mid x_{\text{pub}}, y = c)\frac{\mathbb{P}(y=c \mid x_{\text{pub}})\,\mathbb{P}(x_{\text{pub}})}{\mathbb{P}(y=c)}\,\mathbb{P}(y = c)} \\
&= \frac{\mathbb{P}(\tilde{x}_{\text{priv}}, \tilde{y} \mid x_{\text{pub}}, y = 1)\,\mathbb{P}(y = 1 \mid x_{\text{pub}})\,\mathbb{P}(x_{\text{pub}})}{\sum_{c \in \mathcal{Y}} \mathbb{P}(\tilde{x}_{\text{priv}}, \tilde{y} \mid x_{\text{pub}}, y = c)\,\mathbb{P}(y = c \mid x_{\text{pub}})\,\mathbb{P}(x_{\text{pub}})}
\end{aligned}
$$

ここで，$y = 0$ においても同様にして,

$$
\begin{aligned}
&\frac{\mathbb{P}(y = 1 \mid x_{\text{pub}}, \tilde{x}_{\text{priv}}, \tilde{y})}{\mathbb{P}(y = 0 \mid x_{\text{pub}}, \tilde{x}_{\text{priv}}, \tilde{y})} \\
&= \frac{\mathbb{P}(\tilde{x}_{\text{priv}}, \tilde{y} \mid x_{\text{pub}}, y = 1)\,\mathbb{P}(y = 1 \mid x_{\text{pub}})}{\mathbb{P}(\tilde{x}_{\text{priv}}, \tilde{y} \mid x_{\text{pub}}, y = 0)\,\mathbb{P}(y = 0 \mid x_{\text{pub}})}
\end{aligned}
$$

したがって，$\frac{\mathbb{P}(\tilde{x}_{\text{priv}}, \tilde{y} \mid x_{\text{pub}}, y=1)}{\mathbb{P}(\tilde{x}_{\text{priv}}, \tilde{y} \mid x_{\text{pub}}, y=0)}$ のバウンドについて考えれば良い．ここで，$\tilde{x}_{\text{priv}}, \tilde{y}$ は独立に匿名化されることを考えると,

$$
\begin{aligned}
&\frac{\mathbb{P}(\tilde{x}_{\text{priv}}, \tilde{y} \mid x_{\text{pub}}, y = 1)}{\mathbb{P}(\tilde{x}_{\text{priv}}, \tilde{y} \mid x_{\text{pub}}, y = 0)} \\
&= \frac{\mathbb{P}(\tilde{x}_{\text{priv}} \mid x_{\text{pub}}, y = 1)}{\mathbb{P}(\tilde{x}_{\text{priv}} \mid x_{\text{pub}}, y = 0)} \frac{\mathbb{P}(\tilde{y} \mid x_{\text{pub}}, y = 1)}{\mathbb{P}(\tilde{y} \mid x_{\text{pub}}, y = 0)}
\end{aligned}
$$

$\tilde{y} = RR(y)$ より,

$$
\frac{\mathbb{P}(\tilde{y} \mid x_{\text{pub}}, y = 1)}{\mathbb{P}(\tilde{y} \mid x_{\text{pub}}, y = 0)} = \frac{\mathbb{P}(\tilde{y} \mid y = 1)}{\mathbb{P}(\tilde{y} \mid y = 0)} \leq e^{\varepsilon_2}
$$

ここで，特徴量の各次元は独立にサンプリングされていると仮定すると,

$$
\frac{\mathbb{P}(\tilde{x}_{\text{priv}} \mid x_{\text{pub}}, y = 1)}{\mathbb{P}(\tilde{x}_{\text{priv}} \mid x_{\text{pub}}, y = 0)} = \frac{\mathbb{P}(\tilde{x}_{\text{priv}} \mid y = 1)}{\mathbb{P}(\tilde{x}_{\text{priv}} \mid y = 0)}
$$

また，ε_1 の定義より，

$$\mathbb{P}(\tilde{x}_{\text{priv}} \mid y=1) = \int_x \mathbb{P}(\tilde{x}_{\text{priv}} \mid x)\,\mathbb{P}(x \mid y=1)dx$$
$$\leq \int_x \mathbb{P}(\tilde{x}_{\text{priv}} \mid x) e_1^{\varepsilon}\,\mathbb{P}(x \mid y=0)dx$$
$$= \mathbb{P}(\tilde{x}_{\text{priv}} \mid y=0)$$

または，

$$\forall x, x' \in \mathcal{X}_{\text{priv}},$$
$$\frac{\mathbb{P}(\tilde{x}_{\text{priv}} \mid x_{\text{pub}}, y=1)}{\mathbb{P}(\tilde{x}_{\text{priv}} \mid x_{\text{pub}}, y=0)} \leq \max_{x,x' \in \mathcal{X}_{\text{priv}}} \frac{\mathbb{P}(\tilde{x}_{\text{priv}} \mid x)}{\mathbb{P}(\tilde{x}_{\text{priv}} \mid x')} \leq e^{\varepsilon_1}$$

したがって，命題は示された．　□