差分プライバシーの検証と形式化

2025年2月27日

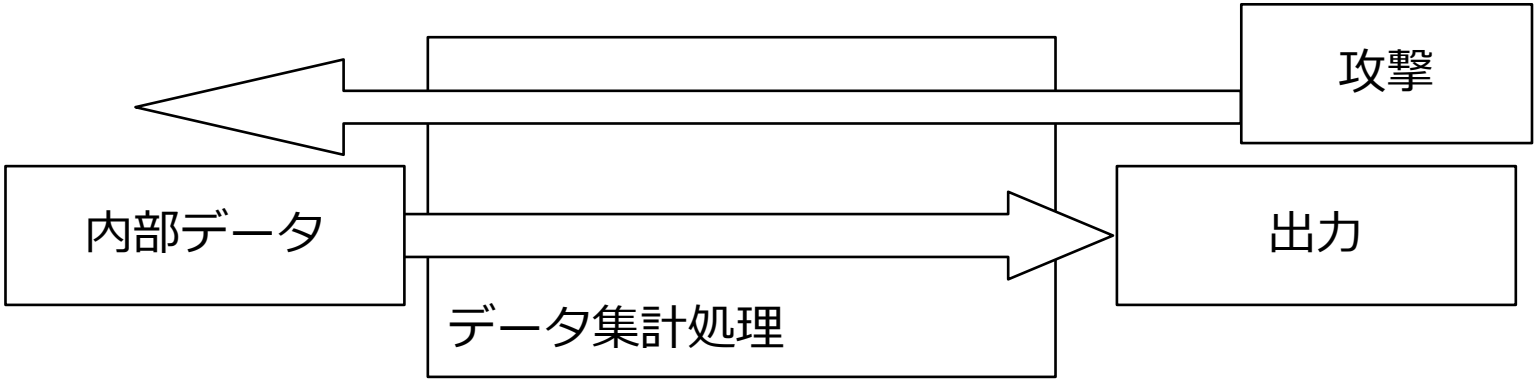
東京科学大学

PL合同セミナー

講演者：佐藤 哲也(東京科学大学)

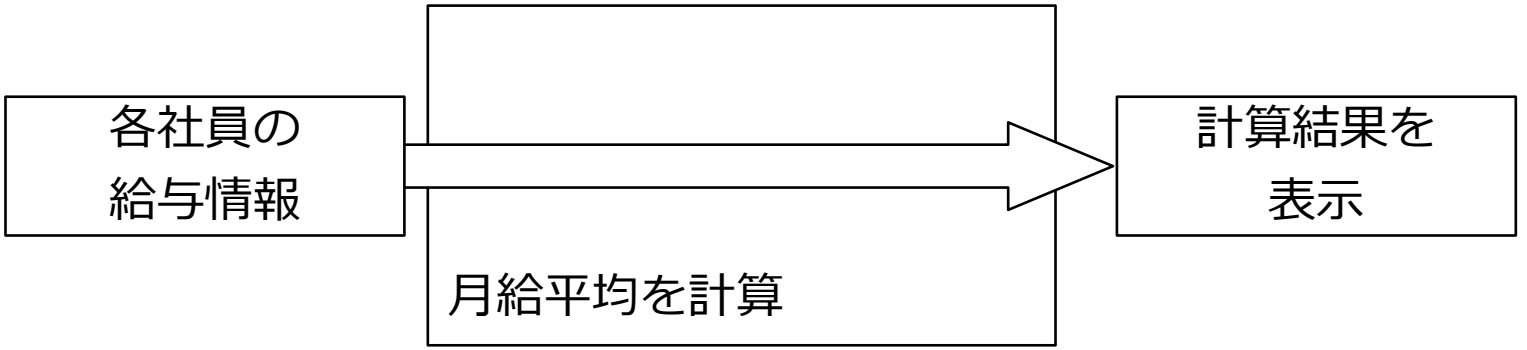
差分プライバシーの背景

- 背景知識攻撃



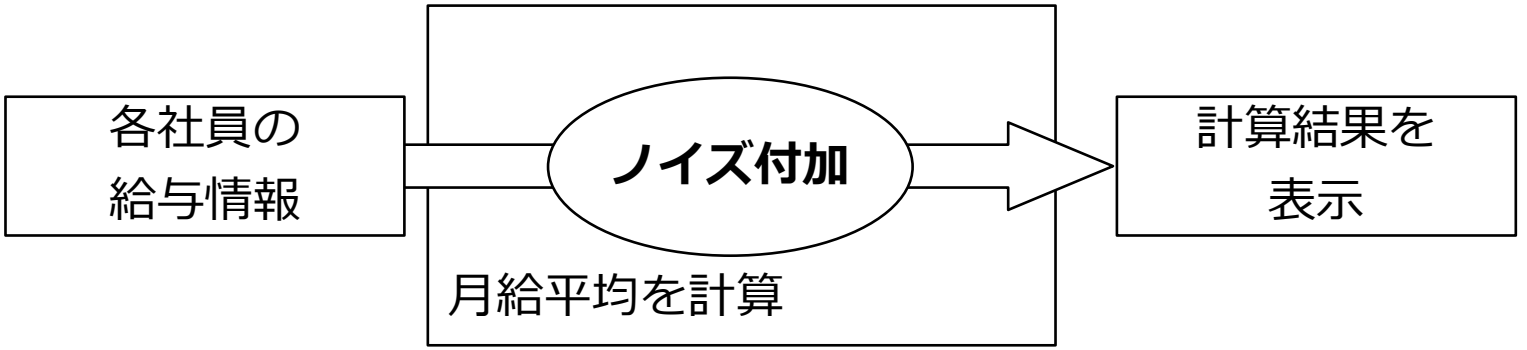
  - 内部データ自体は保護されていたとしても、
  - 十分な背景知識があると、データベースの出力から
    内部データを(統計的に)逆算することが可能となる。

背景知識からの情報漏洩



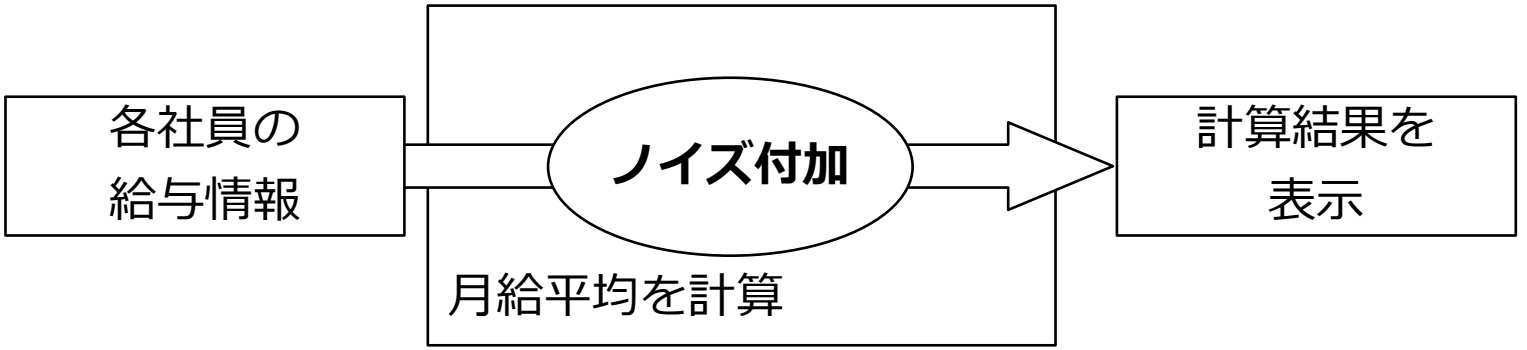
- 以下のようなシナリオを考えてみましょう：
  1. 999人が働く会社があります。
     この会社は月給平均を公開しており、5000ドルと算出されています。
  2. 今月、1000人目の社員――名前はビル――がDBに追加されました。
     月給平均は(敏感に)変化し、5001ドルと算出されています。
  3. 公開しているのは月給平均だけです。
     しかし、上記の状況の変化を知っている(背景知識)皆さんには、
     ビルの月給が6000ドルであると、計算できてしまいます(情報漏洩)。

ノイズ付加による保護



- こういう”筒抜け”の状況を防ぐ方法の一つに
  計算結果(or データ収集時)にノイズを付加する手法がある。
  月給平均にノイズを付加すると、先ほどのシナリオがどうなるか：
  1. 社員999人　→5002ドル(ノイズ：+2ドル)
  2. 社員1000人(1000人目はビル)　→5000ドル(ノイズ：±0ドル)
- このように、ビルの月給を計算するのは統計的に困難となる。
  - 実際には、適切なノイズを付加すれば、という条件がつく。
  - ノイズを加えるので当然出力の精度が落ちる。

差分プライバシーとその問題意識



- 差分プライバシーは出力にノイズを加えて匿名性を保障する手法。
  - ちゃんと匿名性が保障できるかは微妙な議論が必要。
    - ノイズが不適切だと、プライバシーが保障できない。
    - データベースの設定にも依存
      - ビル(月給6000ドル)ではなく、ビル(月給3億ドル)だとすると多少のノイズでは秘匿しきれない。
  - 匿名性と精度はトレードオフの関係にある。
    - 匿名性を保障しつつ、なるべくノイズは小さくしたい。

差分プライバシーの応用例

- QuickTypeの候補出力（タイプした文字を端末から収集・集計）
  https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

- GoogleMapの混雑状況の表示
  https://developers.googleblog.com/en/enabling-developers-and-organizations-to-use-differential-privacy/

- アメリカ国勢調査 (2020年～)
  https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/differential-privacy.html

差分プライバシー

(Differential Privacy)

[Dwork+, TCC 2006] [Dwork+,EUROCRYPT 2006]

- 定義
  - ランダム化されたメカニズム M: Datasets → Prob(Y) が **(ε,δ)-差分プライバシー(DP)**を満たすとは、
    - “隣接する” データセット **D1～D2** について以下が成立：

$$\forall S \subseteq Y.$$
$$\Pr[M(D_1) \in S] \leq \exp(\varepsilon) \Pr[M(D_2) \in S] + \delta$$

  - 直観：確率δの場合を除き、確率比がεで押さえられる
    ((ε,δ)=(0,0) のとき、確率分布は一致する)
  - 感覚的には、ε=0.01～10、δ<<0.01

“隣接する” データセット

- 差分プライバシーは、内部データが更新されたことを秘匿するもので、内部データの構造に強く依存した概念である。

- 直観的には「データ列の1か所だけ異なる」関係にある事を指す。
  - よく使うのは「距離が１以下の関係」や「リストに１成分を挿入」

- 教科書的な定式化：
  - データセットは自然数の配列(ヒストグラム) $D \in \mathbb{N}^n$
    - 各成分 D[i] は i 番目の項目に対応するデータの個数。
  - データセットの”隣接関係”
    - １項目しか違いがないということの数学的表現。

$$D \sim D' \iff \|D - D'\|_1 \leq 1$$

DPの仮説検定的特徴づけ

[Kariouz+, ICML 2015]

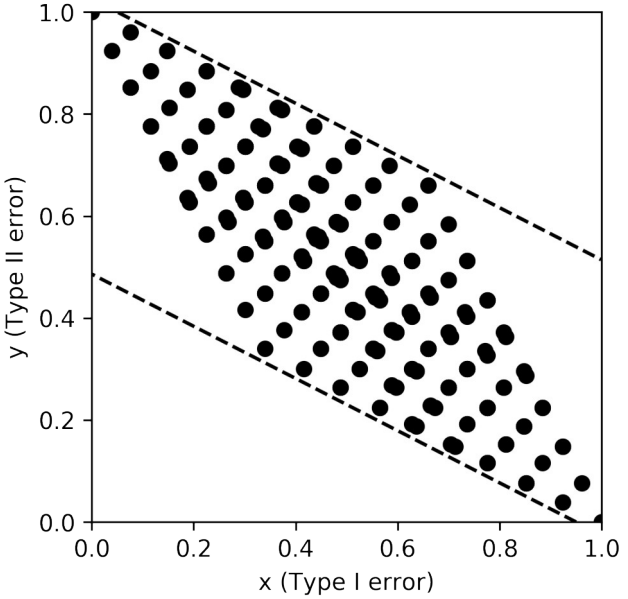
- ランダム化されたメカニズム M: Datasets → Prob(Y) が (ε,δ)-差分プライバシー(DP)を満たすことは以下と同値：
  - 隣接する内部データD1～D2 について

$$\forall S \subseteq Y.\ (\Pr[M(D_1) \in S], \Pr[M(D_2) \notin S]) \in R(\varepsilon, \delta)$$

Rejection　Type I error　Type II error　privacy region

$$R(\varepsilon, \delta) = \{\ (s,t) \mid s + e^{\varepsilon} \cdot t \geq 1 - \delta,\quad t + e^{\varepsilon} \cdot s \geq 1 - \delta\ \}$$

Sは、内部データがD1かD2かのどちらかを判断する仮説検定手法と等価。
　　仮説検定の棄却域を検定統計量で
　　引き戻した逆像に対応する。
帰無仮説…Mの内部データはD1
対立仮説…Mの内部データはD2



Renyi 差分プライバシー

[Mironov, CSF2017]

- Renyiダイバージェンスを用いた定義、機械学習のDPによく使われる。
- ランダム化されたメカニズム M: Datasets → Prob(Y) が
  (α,ρ)-Renyi 差分プライバシー (RDP) を満たすとは、
  - 隣接するデータセット D1 ～ D2 について

$$D_Y^{\alpha}(M(D_1)||M(D_2)) \leq \rho$$

  - α-Renyi ダイバージェンス (α>1) ：

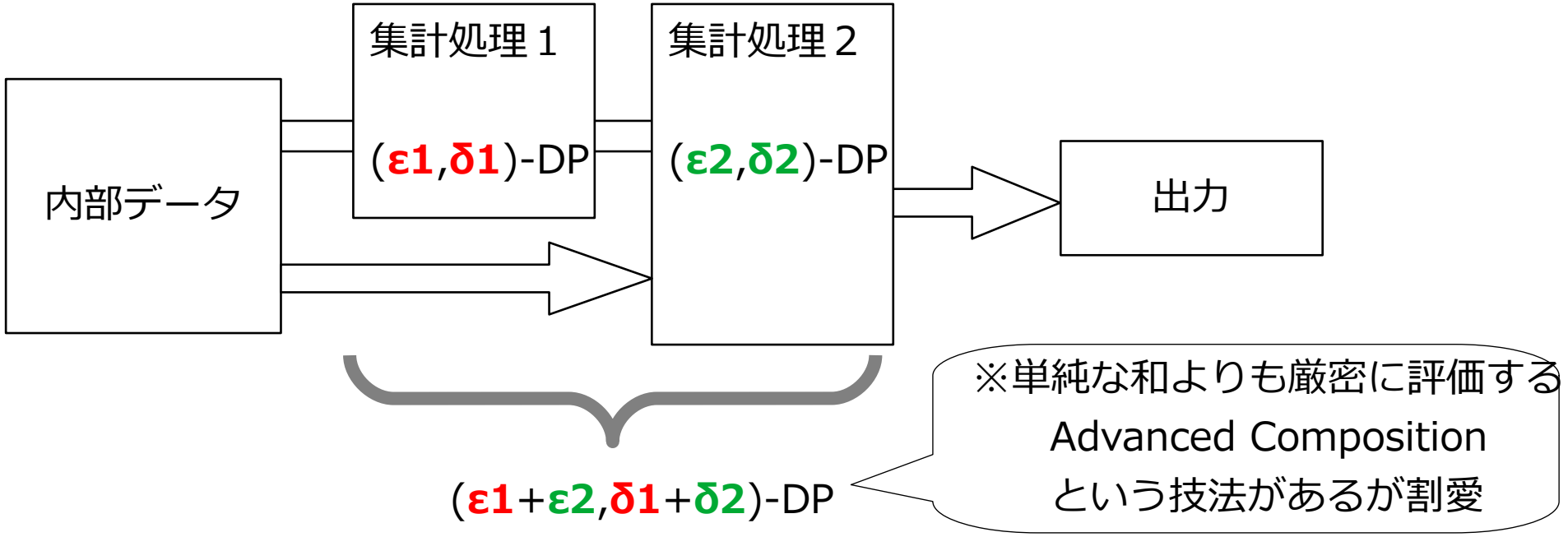
$$D_Y^{\alpha}(\mu_1, \mu_2) = \frac{1}{\alpha - 1} \log \int_Y \mu_1(y)^{\alpha} \mu_2(y)^{1-\alpha} dy$$

ふつうのDPと同様に、
RDPも後述の合成性を持つ。

差分プライバシーの合成性

(Composition theorem)

- 組み合わせたデータベースの差分プライバシーは、
  各ブロックごとに分割して評価できる。

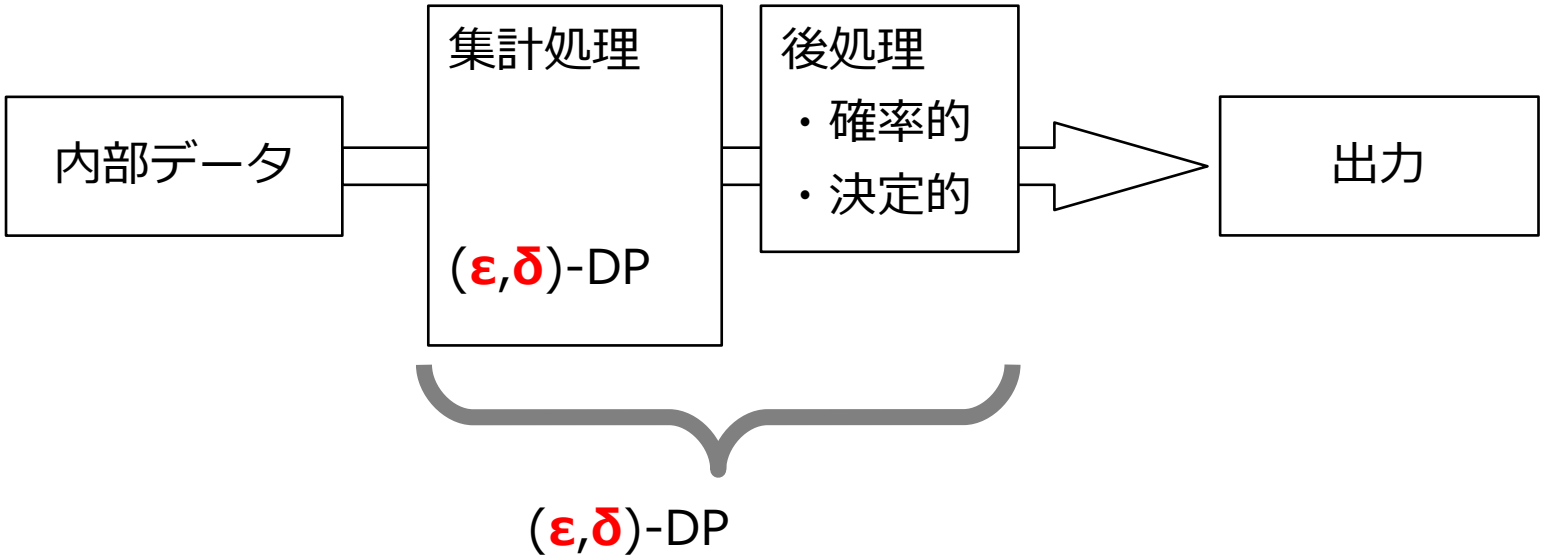


$$M_1{:}(\varepsilon_1,\delta_1)\text{-DP and } \forall y \in Y_1.\ M_2(-,y){:}(\varepsilon_1,\delta_1)\text{-DP}$$
$$\Longrightarrow \{y_1 \leftarrow M_1(x); y_2 \leftarrow M_2(x,y_1); \mathrm{return}(y_1,y_2)\}{:}(\varepsilon_1+\varepsilon_2,\delta_1+\delta_2)\text{-DP}$$

後処理に対する安定性

(Postprocessing)

- 後処理(決定的でも確率的でも)を加えても、差分プライバシーは変化しない。



  - 合成性から示す
    - 後処理部分は内部データを参照しないので、(0,0)-DP。

Randomized Response

- ビットを一定確率で反転。

$$\mathtt{RR}_\varepsilon : \{\top, \bot\} \to \mathrm{Prob}\{\top, \bot\}$$

$$\mathtt{RR}_\varepsilon(b) = \begin{cases} b & \text{with probability } \frac{e^\varepsilon}{e^\varepsilon+1} \\ \neg b & \text{with probability } \frac{1}{e^\varepsilon+1} \end{cases}$$

(ε,0)-DPを満たす。

- 応用：QuickTypeの候補出力
  https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf
  → 収集したデータの各ビットにrandomized responseを適用。

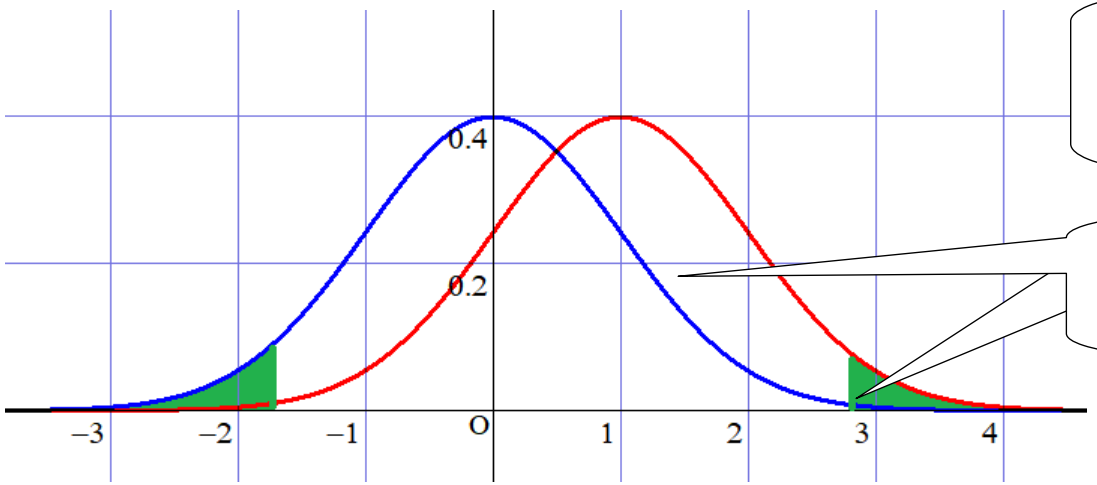
正規分布によるノイズ付加

- 平均0、分散 σ^2 の正規分布からなるノイズを加算

$\mathtt{Gauss}_\sigma : \mathbb{R} \to \mathrm{Prob}(\mathbb{R})$

$\mathtt{Gauss}_\sigma(x)$ は平均 x 分散 σ^2 の正規分布となる

  - 隣接関係 |x-y|≦1 に対して、**(ε,δ(ε))-DP** が言える。
    (δ(ε)の計算は複雑なので割愛)、RDPだと、**(α,α/2σ^2)-RDP**。



確率比は遠方で∞に発散、
誤差 0<δ だけ端をカット

残った中心部分の確率比e^{ε}

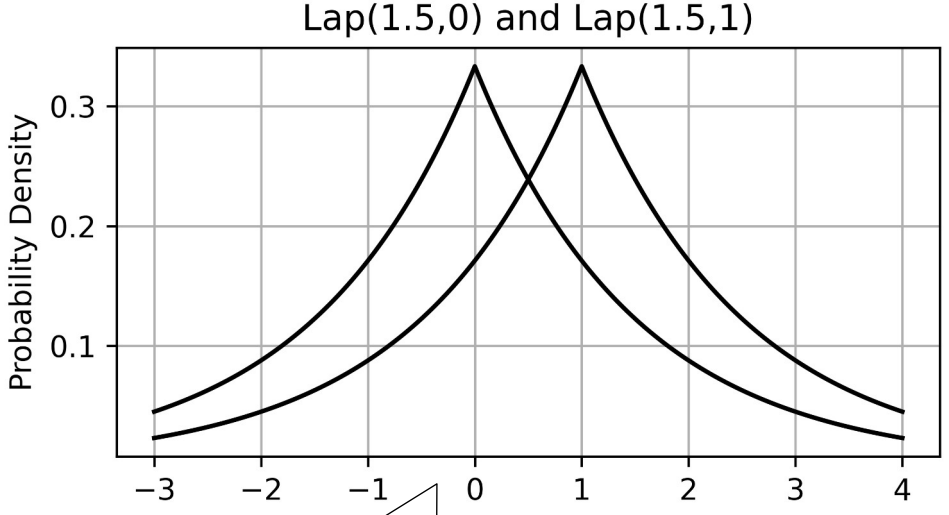
ラプラス分布によるノイズ付加

- 平均0、尺度 b のラプラス分布からなるノイズを加算

$\mathrm{Lap}_b : \mathbb{R} \to \mathrm{Prob}(\mathbb{R})$

$\mathrm{Lap}_b(x)$ は平均 x 尺度 b のラプラス分布となる

  - 隣接関係 |x-y|≦1 に対して、(1/b,0)-DPが得られる



x=0, y=1　という２つの数値に尺度1.5の
ラプラス分布によるノイズを加算した結果

密度関数

$$\Pr[\mathrm{Lap}_b(x) = z] = \frac{1}{2b}\exp(-\frac{|x-z|}{b})$$

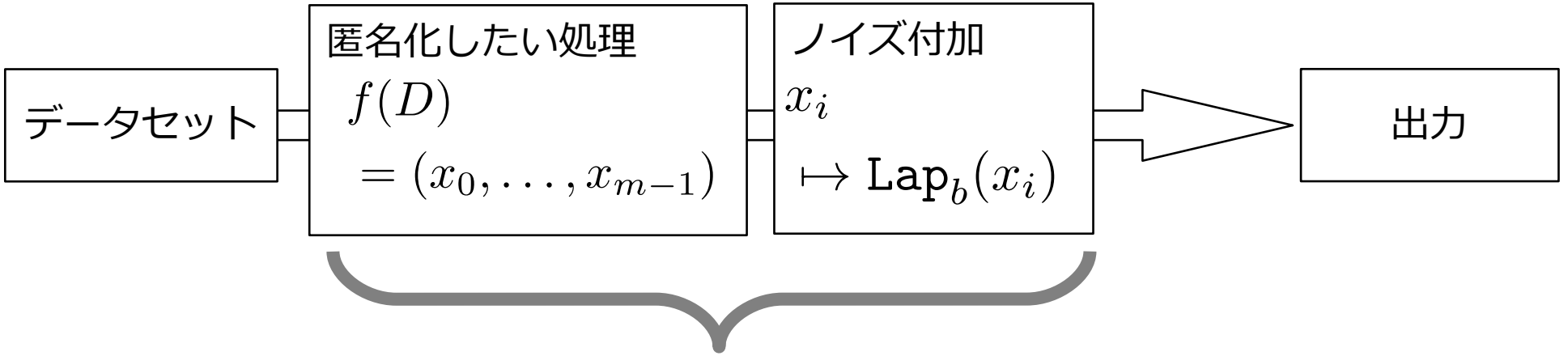
差分プライバシーのための性質

$$|x-y| \le r$$

$$\Longrightarrow \Pr[\mathrm{Lap}_b(x) = z] \le \exp(r/b)\Pr[\mathrm{Lap}_b(y) = z]$$

Laplace Mechanism

- 関数 f のm次元の出力の各成分に、
  ラプラス分布によるノイズを独立に加算する。



f の感度(sensitivity)

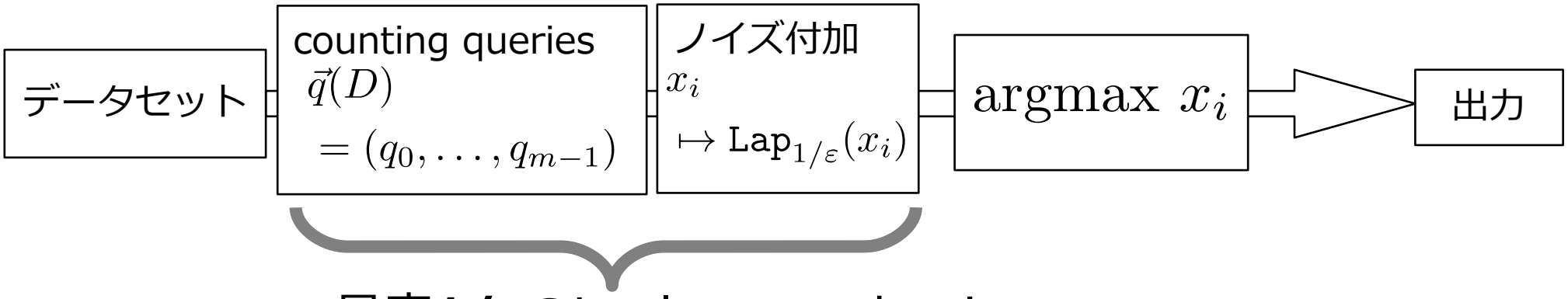
$$\Delta f = \sup\{\|f(D_1) - f(D_2)\|_1 \mid D_1 \sim D_2\}$$

と、ラプラス分布の尺度bに対し、
メカニズム全体は (**Δf/b**,0)-DP。

Report Noisy Max Mechanism

(Dwork and Roth, "The Algorithmic Foundations of Differential Privacy")

- **m個のcounting queries**に関するLaplace mechanismを適用、得られた組の最大値が何番目かを返す。



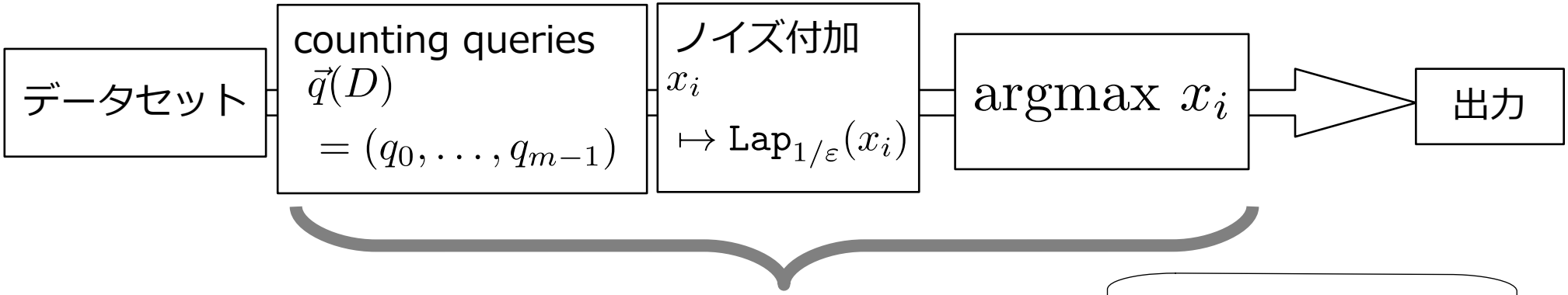
尺度**1/ε**のLaplace mechanism

- 各counting queryはデータセットDの内、特定のクラスに属するデータがいくつあるかを返す。
  - 感度は $\Delta q_i = 1$ 、m個の組では $\Delta \vec{q} = m$

素朴には(mε,0)-DP

RNMの差分プライバシー

- 差分プライバシー



  - 証明は確率分布 $\Pr[\mathrm{RNM}_{\vec{q},m,\varepsilon}(D) = i]$ を評価して行い、argmax 処理や、counting queties 特有の性質を使う。

形式的検証の必要性

- 差分プライバシーはプログラムの小さな変更で崩れ得る。
  - sparse-vector techniqueの研究事例の比較 [Lyu+, VLDB 2017]

差分プライバシーを満たす
Above thresholdメカニズム

Algorithm 2 SVT in Dwork and Roth 2014 [8].

Input: D, Q, Δ, T, c.
1: $\epsilon_1 = \epsilon/2, \quad \rho = \mathsf{Lap}\,(c\Delta/\epsilon_1)$
2: $\epsilon_2 = \epsilon - \epsilon_1, \quad \text{count} = 0$
3: **for** each query $q_i \in Q$ **do**
4: $\quad \nu_i = \mathsf{Lap}\,(2c\Delta/\epsilon_1)$
5: $\quad$ **if** $q_i(D) + \nu_i \geq T + \rho$ **then**
6: $\quad\quad$ Output $a_i = \top$, $\rho = \mathsf{Lap}\,(c\Delta/\epsilon_2)$
7: $\quad\quad$ count = count + 1, **Abort** if count $\geq c$.
8: $\quad$ **else**
9: $\quad\quad$ Output $a_i = \bot$

厳密で正確な検証を行う
「ツール」が必要

実はいかなる差分プライバシーも
満たさないことが**後から**わかった例

Algorithm 3 SVT in Roth's 2011 Lecture Notes [17].

Input: D, Q, Δ, T, c.
1: $\epsilon_1 = \epsilon/2, \quad \rho = \mathsf{Lap}\,(\Delta/\epsilon_1)$,
2: $\epsilon_2 = \epsilon - \epsilon_1, \quad \text{count} = 0$
3: **for** each query $q_i \in Q$ **do**
4: $\quad \nu_i = \mathsf{Lap}\,(c\Delta/\epsilon_2)$
5: $\quad$ **if** $q_i(D) + \nu_i \geq T + \rho$ **then**
6: $\quad\quad$ Output $a_i = q_i(D) + \nu_i$
7: $\quad\quad$ count = count + 1, **Abort** if count $\geq c$.
8: $\quad$ **else**
9: $\quad\quad$ Output $a_i = \bot$

Algorithm 5 SVT in Stoddard et al. 2014 [20].

Input: D, Q, Δ, T.
1: $\epsilon_1 = \epsilon/2, \quad \rho = \mathsf{Lap}\,(\Delta/\epsilon_1)$
2: $\epsilon_2 = \epsilon - \epsilon_1$
3: **for** each query $q_i \in Q$ **do**
4: $\quad \nu_i = 0$
5: $\quad$ **if** $q_i(D) + \nu_i \geq T + \rho$ **then**
6: $\quad\quad$ Output $a_i = \top$
7:
8: $\quad$ **else**
9: $\quad\quad$ Output $a_i = \bot$

関係ホーア論理による検証

- apRHL [Barthe+,POPL2012]
  - 次数付きのjudgment

$$\{\Phi\}\ c_1 \sim^{(\varepsilon,\delta)} c_2\ \{\Psi\}$$

2つの確率的プログラム

$$c_1, c_2 : M \to \mathrm{prob}(M)$$

差分プライバシー
パラメータ

メモリM上の**二項関係**
Φ…事前条件
Ψ…事後条件

  - 差分プライバシーを記述できる

これをゴールとして、
apRHL上で証明を行う

$$\{\mathtt{Adj}\}c \sim^{\varepsilon,\delta} c\{x\langle 1\rangle = x\langle 2\rangle\}$$

データセット上の隣接

出力変数x上のequality

関係ホーア論理の性質

- 以下のような推論規則が健全(ほかにもone-sided ruleなどがある) :

$$\{\Phi\}\texttt{skip} \sim^{(0,0)} \texttt{skip}\{\Phi\}$$

$$\frac{\{\Phi\}c_1 \sim^{(\varepsilon,\delta)} c_2\{\Phi'\} \qquad \{\Phi'\}d_1 \sim^{(\varepsilon',\delta')} d_2\{\Psi\}}{\{\Phi\}c_1; d_1 \sim^{(\varepsilon+\varepsilon',\delta+\delta')} c_2; d_2\{\Psi\}}$$

$$\frac{\Phi' \subseteq \Phi \quad \{\Phi\}c_1 \sim^{(\varepsilon,\delta)} c_2\{\Psi\} \quad \Psi \subseteq \Psi' \quad \varepsilon \leq \varepsilon' \quad \delta \leq \delta'}{\{\Phi'\}c_1 \sim^{(\varepsilon',\delta')} c_2\{\Psi'\}}$$

- 差分プライバシーを記述できる(x : 出力用の変数) :
  - $\{\texttt{Adj}\}c \sim^{\varepsilon,\delta} c\{x\langle 1\rangle = x\langle 2\rangle\}$ の正当性と「cが(ε,δ)-DP」が同値。

関係持ち上げ

- こうした関係ホーア論理では、次数付きjudgment

$$\{\Phi\}\ c_1 \sim^{(\varepsilon,\delta)} c_2\ \{\Psi\} \quad c_1, c_2 \colon M \to \mathrm{prob}(M)$$

  - の正当性を、適当な条件を満たす関係持ち上げで解釈する。

$$(m_1, m_2) \in \Phi \implies (c_1(m_1), c_2(m_2)) \in \Psi^{\sharp(\varepsilon,\delta)}$$

$$(c_1, c_2) \colon \Phi \dot{\to} \Psi^{\sharp(\varepsilon,\delta)}$$

  - plainな二項関係を確率分布上に拡張

$$\text{For } \Psi \subseteq M \times M, \quad \Psi^{\sharp(\varepsilon,\delta)} \subseteq \mathrm{prob}(M) \times \mathrm{prob}(M)$$

  - 差分プライバシーを復元する。

$$\mathrm{Eq}_X^{\sharp(\varepsilon,\delta)} = \{(\mu_1, \mu_2) | \forall S \subseteq X.\ \mu_1(S) \leq e^{\varepsilon}\mu_2(S) + \delta\}$$

  - 確率分布モナドの構造と整合的である(graded liftingを成す)

- 確率分布モナドの構造と整合的である (graded liftingを成す)

$$(\eta_M, \eta_M) \colon \Phi \dot{\to} \Psi^{\sharp(0,0)}$$

$$\frac{(c_1, c_2) \colon \Phi \dot{\to} \Psi^{\sharp(\varepsilon,\delta)}}{(c_1^\sharp, c_2^\sharp) \colon \Phi^{\sharp(\varepsilon',\delta')} \dot{\to} \Psi^{\sharp(\varepsilon'+\varepsilon,\delta'+\delta)}}$$

$$\frac{(c_1, c_2) \colon \Phi \dot{\to} \Psi^{\sharp(\varepsilon,\delta)}, \varepsilon \leq \varepsilon', \delta \leq \delta'}{(c_1, c_2) \colon \Phi \dot{\to} \Psi^{\sharp(\varepsilon',\delta')}}$$

関係持ち上げの構成と拡張

- こういった関係持ち上げを一般的にどう作るか？
  - apRHL [Barthe+,POPL2012] での構成(離散分布上)

$$(R \subseteq Y_1 \times Y_2)^{\sharp(\varepsilon,\delta)}$$

$$= \left\{ (\mu_1, \mu_2) \;\middle|\; \begin{array}{l} \exists \nu_L, \nu_R \in \mathrm{Prob}(R). \mu_1 = \pi_1[\nu_L], \mu_2 = \pi_2[\nu_R] \\ \\ \forall S \subseteq R.\ \nu_L(S) \leq \exp(\varepsilon)\nu_R(S) + \delta \end{array} \right\}$$

  - apRHLの連続化を試みて得たもの [Sato, MFPS2016]

$$(R \subseteq Y_1 \times Y_2)^{\top\top(\varepsilon,\delta)}$$

$$= \{(\mu_1, \mu_2) | \forall A \subseteq Y_1. \mu_1(A) \leq \exp(\varepsilon)\mu_2(R[A]) + \delta\}$$

  - [Sato+,LICS2019] RDPの検証に対応。
    二項関係をSpanに拡張し、持ち上げも合わせて拡張（構成は割愛）
  - [Sato & Katsumata, MSCS 2023] 連続で高階な確率的プログラム(の定式化)[Heunen+,LICS2017]にも対応。

ダイバージェンスによる特徴づけ

[Barthe & Olmedo, ICALP 2013]

- 差分プライバシーの「ダイバージェンスによる定式化」
  - 実は、apRHLのような関係ホーア論理の本質的なデータ

- M: Datasets → Prob(Y) が(ε,δ)-DPを満たす

  ⇔ 隣接するデータD1～D2について

$$\forall S \subseteq Y.\quad \Pr[M(D_1) \in S] \leq \exp(\varepsilon) \Pr[M(D_2) \in S] + \delta$$

  ⇔ 隣接するデータD1～D2について

$$\underbrace{\sup_{S \subseteq Y} \left(\Pr[M(D_1) \in S] - \exp(\varepsilon) \Pr[M(D_2) \in S]\right)}_{\Delta^{\varepsilon}(M(D_1) \| M(D_2))} \leq \delta$$

ダイバージェンス $\triangle^{\varepsilon}$ の基本的性質

- (Eq-単位)反射性

$$\Delta_X^0(\mu,\mu) = 0$$

- 単調性

$$\varepsilon \geq \varepsilon_2 \implies \Delta_X^{\varepsilon}(\mu,\nu) \leq \Delta_X^{\varepsilon_2}(\mu,\nu)$$

- (Eq-)合成性

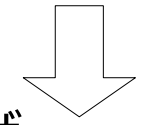
$$\Delta_X^{\varepsilon}(\mu_1,\mu_2) \leq \delta \text{ and } \forall x \in X. \Delta_Y^{\varepsilon_2}(f(x),g(x)) \leq \delta_2$$
$$\implies \Delta_X^{\varepsilon+\varepsilon_2}(f^{\sharp}\mu_1, g^{\sharp}\mu_2) \leq \delta + \delta_2$$

Renyiダイバージェンスも同様の性質を満たす。
一般のモナドと順序付きモノイドでこの議論はできる。

一般化された関係持ち上げ

[Sato & Katsumata, MSCS 2023]

- 任意の直積を持つ圏C、C上の強モナドT、
  順序付きモノイドM、順序付きモノイドが乗った完備半順序Q
- 単調性を持つ関数族 $\Delta = \{\Delta_I^m \colon |TI| \times |TI| \to Q\}_{m \in M, I \in \mathbb{C}}$



- Tの(M×Q)-次数付き関係持ち上げ

$$\tilde{\Delta}(\varepsilon, \delta) I = \{(\nu_1, \nu_2) | \Delta_I^\varepsilon(\nu_1, \nu_2) \leq \delta\}$$

$$T^{[\Delta]}(\varepsilon, \delta)\Phi = \left\{ (\mu_1, \mu_2) \middle| \begin{array}{l} \forall I, \varepsilon', \delta', (k, l) \colon \Phi \dot{\to} \tilde{\Delta}(\varepsilon', \delta') I. \\ (k^\sharp \mu_1, l^\sharp \mu_2) \in \tilde{\Delta}(\varepsilon + \varepsilon', \delta + \delta') I \end{array} \right\}$$

- Δが、(単位-)反射性・合成性を持つ(モナド上のダイバージェンス) ことと
  $T^{[\Delta]}(\varepsilon, \delta)\mathrm{Eq}_X = \{(\mu_1, \mu_2) | \Delta^\varepsilon(\mu_1, \mu_2) \leq \delta\}$ が同値！
  (単位-)反射性は⊆、合成性は⊇と対応。

余談

- [Sato, MFPS2016]で与えた関係持ち上げ

$$(R \subseteq Y_1 \times Y_2)^{\top\top(\varepsilon,\delta)}$$
$$= \{(\mu_1, \mu_2) | \forall A \subseteq Y_1 . \mu_1(A) \leq \exp(\varepsilon)\mu_2(R[A]) + \delta\}$$

はΔをDPのダイバージェンス Δ^{ε} でとると $T^{[\Delta]}(\varepsilon, \delta)R$ に等しい。

$$(R \subseteq Y_1 \times Y_2)^{\top\top(\varepsilon,\delta)}$$
$$= \left\{ (\mu_1, \mu_2) \middle| \begin{array}{c} \forall \varepsilon', \delta', (k, l) \colon R \dot{\to} \tilde{\Delta}(\varepsilon', \delta')2. \\ (k^\sharp \mu_1, l^\sharp \mu_2) \in \tilde{\Delta}(\varepsilon + \varepsilon', \delta + \delta')2 \end{array} \right\}$$
$$=?= \left\{ (\mu_1, \mu_2) \middle| \begin{array}{c} \forall I, \varepsilon', \delta', (k, l) \colon R \dot{\to} \tilde{\Delta}(\varepsilon', \delta')I. \\ (k^\sharp \mu_1, l^\sharp \mu_2) \in \tilde{\Delta}(\varepsilon + \varepsilon', \delta + \delta')I \end{array} \right\}$$
$$= T^{[\Delta]}(\varepsilon, \delta)R$$

RDPの仮説検定的特徴づけ

[Balle+,AISTATS2020]

- ダイバージェンスΔのk-cut

$$\overline{\Delta}_X^k(\mu_1,\mu_2) = \sup_{f:\, X\to \mathrm{Prob}(I)} \Delta_I(f^\sharp\mu_1, f^\sharp\mu_2) \quad (|I|=k)$$

  - Δのk-cut が元のΔであるときk-generatedと呼ぶ。

  - DPの仮説検定的特徴づけ [Kariouz+, ICML 2015] は
    Δ^εが2-generatedであることと等価である。
    また、さっきの余談で述べた等式もこの性質を使って証明できる。
  - Renyi ダイバージェンスは2-generatedではない(Nat-generated)ため、
    等価な仮説検定的特徴づけは存在しない。
    が、2-cutを使ってRDP→DPの良い変換ができる。

RDP→DPの公式

[Balle+,AISTATS2020]

$$\Delta^{\varepsilon} \leq D^{\alpha} \iff \Delta^{\varepsilon} \leq \overline{D^{\alpha}}^{2}$$

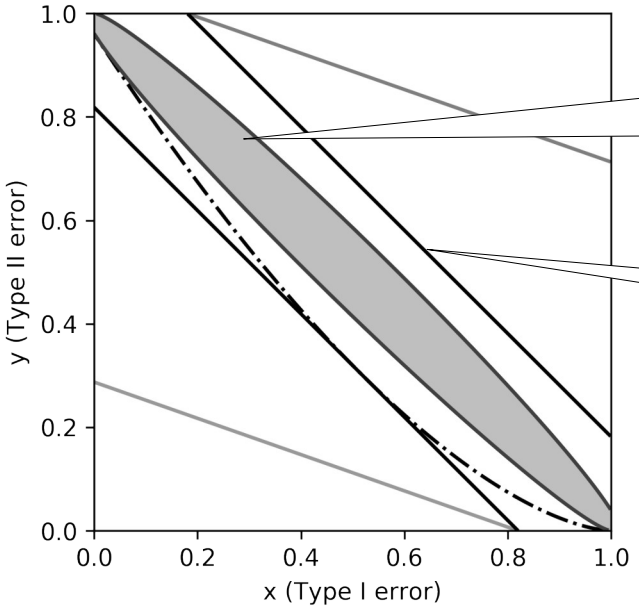
2-cutの最大性

$$\iff \Delta^{\varepsilon}_{\{\mathbf{a},\mathbf{r}\}} \leq \overline{D^{\alpha}}^{2}_{\{\mathbf{a},\mathbf{r}\}}$$

DPが2-gen.

$$\iff R^{D^{\alpha}} \supseteq R^{\Delta^{\varepsilon}}$$

2-cutの仮説検定による等価な特徴づけ



$R^{D^{\alpha}}(\rho)$ (α,ρ)-RDPのプライバシー領域

境界は、$(1-x)^{\alpha} y^{1-\alpha} + x^{\alpha}(1-y)^{1-\alpha} = e^{\rho(\alpha-1)}$

$R^{\Delta^{\varepsilon}}(\delta)$ 本研究で得たcovering

$$(\alpha,\rho)\text{-RDP} \implies \left(\rho + \log\frac{\alpha-1}{\alpha} - \frac{\log\delta + \log\alpha}{\alpha-1}, \delta\right)\text{-DP}$$

Isabelle/HOLによる形式的検証

- 定理証明支援系Isabelle/HOLで差分プライバシーの形式化を進める。
  - 現在の進捗状況 [Sato & Minamide, CPP2025]
    - (標準的な)差分プライバシーの形式化
      - 差分プライバシーの定義をIsabelle/HOLで記述
      - 合成性などの性質の形式的証明をあたえた
        - 差分プライバシーのための統計的ダイバージェンス
    - Laplace mechanismの形式化
      - ラプラス分布の実装(正規分布は既にあった)
      - 差分プライバシーの形式的証明をあたえる
    - Report noisy max mechanismの形式化
      - 差分プライバシーを示す(1500行程度)

AFP(Archive of Formal Proofs ; 公式レポジトリ)エントリ：

- https://www.isa-afp.org/entries/Differential_Privacy.html

動機

- **連続的な確率分布**を念頭に置いて差分プライバシーの形式化を行った。

- 連続的な確率分布でやる動機：
  - 離散的な確率分布より一般的な状況を考えている。
  - いくつかの研究では、離散的なアルゴリズムの差分プライバシー
    であっても、連続的な確率分布を経由することがある。
    - ラプラスメカニズムの浮動小数点数での実装 [Mironov, CCS2012]
      連続的なラプラスメカニズムを ”ideal mechanism” と呼び、
      証明内で呼び出している。
    - 差分プライバシーの拡張の一つである f-DP/Gaussian-DP [Dong+,2019] の
      合成性を示す際、単位区間上の分布(連続的分布)を経由する。
  - 差分プライバシーの研究自体が連続的なセッティングで書かれている事が多い。
    （ありていに言うと、実数上の積分がよく出てくる）

関連研究

- 離散的な状況でのDPの形式化は他の証明支援系で行われている。
  (Isabelle/HOLでは無かった)
  - Coq (Certipriv)[Barthe+, TOPLAS2013]
    - DPの離散版の形式化
    - apRHLの実装
    - Gaussian mechanismの離散版
  - Lean (SampCert) [Tristan+, 2024-ongoing]
    - DP, RDPの離散版の形式化（連続版への拡張を視野に入れてそう）
    - Discrete Laplace mechanism / Discrete Gaussian mechanism
    - Pythonコード生成
    - 論文(プレプリント)は10名以上と、大型のプロジェクト。

Isabelle/HOL

- 公式サイト：https://isabelle.in.tum.de/
- 証明支援系(proof assistant)の一つ。
  - 読みやすい証明用言語Isarを使える。
  - 測度論や確率論のライブラリが充実。
  - 自動証明機能Sledgehammerもある。

```
primrec rev :: "'a list ⇒ 'a list" where
"rev [] = []" |
"rev (x # xs) = rev xs @ [x]"
```

Isabelle/HOLでの
リストの反転と結合の定義

```
primrec append :: "'a list ⇒ 'a list ⇒ 'a list" (infixr "@" 65) where
append_Nil: "[] @ ys = ys" |
append_Cons: "(x#xs) @ ys = x # xs @ ys"
```

Isabelle/HOL上で証明された
リストの反転と結合の性質

```
lemma rev_append [simp]: "rev (xs @ ys) = rev ys @ rev xs"
by (induct xs) auto

lemma rev_rev_ident [simp]: "rev (rev xs) = xs"
by (induct xs) auto
```

Isabelle/HOLでの確率論

- 標準ライブラリにある以下の構成を使う：
  - 測度空間の型　(可測集合と測度を兼ねた型)　(X, Σ_X, μ)
    - 台集合
```
"space"
  :: "'a measure ⇒ 'a set"
```
X
    - 完全加法族
```
"sets"
  :: "'a measure ⇒ 'a set set"
```
Σ_X
    - 測度の評価関数
```
"Sigma_Algebra.measure"
  :: "'a measure ⇒ 'a set ⇒ real"
```
$\mu(A) \quad A \in \Sigma_X$
    - 可測関数
```
"(→M)"
  :: "'a measure ⇒ 'b measure ⇒ ('a ⇒ 'b) set"
```
  - 確率分布モナド(Giryモナド)
```
"prob_algebra M"
  :: "'a measure measure"
```
    - bind
```
"(≫=)"
  :: "'a measure ⇒ ('a ⇒ 'b measure) ⇒ 'b measure"
```
    - return
```
"return"
  :: "'a measure ⇒ 'a ⇒ 'a measure"
```
  - Radon-Nikodym 微分 (密度関数)
  - ルベーグ測度

Isabelle/HOLでの確率的プログラム

- 確率的プログラムを可測関数(Isabelle/HOLの項)として扱う。
  - 実行可能な確率的プログラムも取り扱い可能だが、触れない。

- 例えば、

$$M : \mathbb{R} \to \mathrm{Prob}(\mathbb{R})$$

という可測関数は、

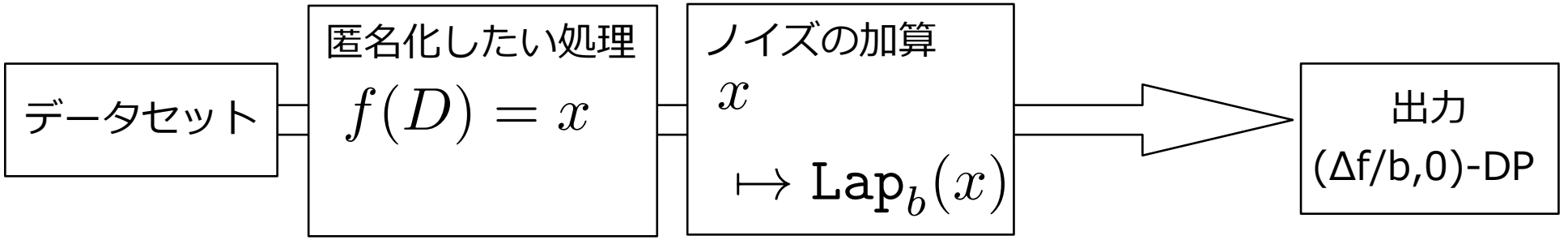
```
"M ∈ X →M prob algebra borel"
```

という条件を満たす(普通の)関数として扱う

  - prob_algebra は確率分布全体の可測空間を構成(Giryモナド)
  - borel は実直線+ボレル代数

Isabelle/HOLでの確率的プログラム

- 1次元のラプラスメカニズムの場合



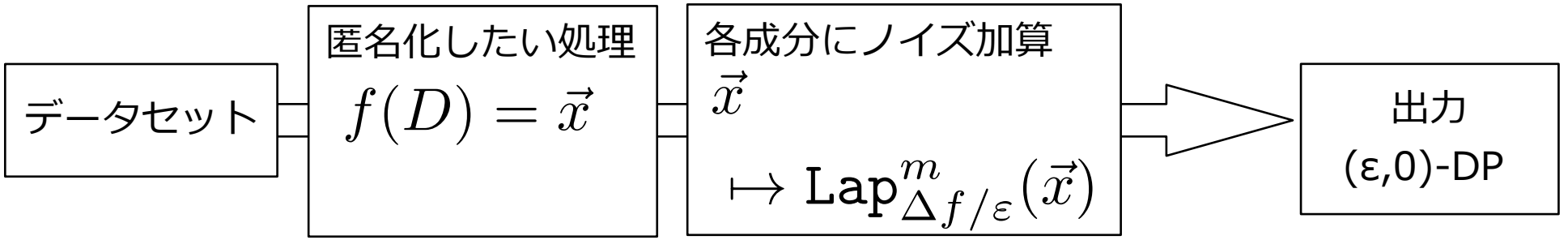

```
"LapMech_1dim ε x = Lap_dist ((real_of_ereal sensitivity) / ε) (f x)"

lemma Lap_dist_def2:
  shows "(Lap_dist b x) = do{r ← Lap_dist0 b; return borel (x + r)}"
```

平均 0 、尺度 b のラプラス分布

それを加算(return式；決定的)

Isabelle/HOLでの確率的プログラム

- m次元で、ノイズ量をsensitivityで決める場合



- f の感度

```
definition sensitivity:: ereal where
  "sensitivity = Sup{ ereal ( ∑ i∈{1..m}. ¦ nth (f x) (i-1) - nth (f y) (i-1) ¦)
                      | x y::'a. x ∈ space X ∧ y ∈ space X ∧ (x,y) ∈ adj}"
```

- ラプラスメカニズム全体

```
"LapMech_list ε x = Lap_dist_list ((real_of_ereal sensitivity) / ε) (f x)"
```

- 各成分にノイズを加算（簡単のため組ではなくリストで形式化する）

```
primrec Lap_dist_list :: "real list ⇒ (real list) measure" where
  "Lap_dist_list [] = return (listM borel) []"
  "Lap_dist_list (x # xs) = do{x1 ← (Lap_dist b x);
     x2 ← (Lap_dist_list xs); return (listM borel) (x1 # x2)}"
```

ラプラス分布の形式化

- 密度関数・累積分布関数の形式化

```
definition laplace_density :: "real ⇒ real ⇒ real ⇒ real" where
  "laplace_density l m x = (if l > 0 then (exp(-¦ x - m¦ / l) / (2* l)) else 0)"

definition laplace_CDF :: "real ⇒ real ⇒ real ⇒ real" where
  "laplace_CDF l m x = (if l > 0
     then (if x < m then (exp((x - m) / l) / 2) else (1 - exp(-(x - m) / l) / 2)) else 0)"
```

  - 密度関数の積分が累積分布関数であることをしめす。

- 尺度b、平均μのラプラス分布

```
definition Lap_dist :: "real ⇒ real ⇒ real measure" where
  "Lap_dist b μ =
    (if b ≤ 0 then return borel μ
      else density lborel (laplace_density b μ))"

lemma Lap_dist_def2:
  shows "(Lap_dist b x) = do{r ← Lap_dist0 b; return borel (x + r)}"
```

平均0のものを平行移動して書ける

データセットの形式化

- 組の代わりにリストを使う。$D \in \mathbb{N}^n$ は長さnのリストと思う。
  - undefinednessを回避したいから。
- リストのL1距離を形式化する。

$$D \sim D' \iff \|D - D'\|_1 \leq 1$$

```
locale results_AFDP =
  fixes n ::nat (* lengt
begin

interpretation L1_norm_list "(UNIV::nat set)" "(λ x y. ¦int x - int y¦)" n
  by(unfold_locales,auto)

definition sp_Dataset :: "nat list measure"where
  "sp_Dataset ≡ restrict_space (listM (count_space UNIV)) space_L1_norm_list"

definition adj_L1_norm :: "(nat list × nat list) set"where
  "adj_L1_norm ≡ {(xs,ys)| xs ys. xs ∈ space sp_Dataset ∧ ys ∈ space sp_Dataset
                            ∧ dist_L1_norm_list xs ys ≤ 1}"
```

```
definition space_L1_norm_list ::"('a list) set" where
  "space_L1_norm_list = {xs. xs∈ lists M ∧ length xs = n}"
```

付属のlocale Metric_space　を応用。

差分プライバシーの形式化

- データセットの隣接関係を抽象化して形式化する。
  - ランダム化されたメカニズム M: X → Prob(Y) が
    (ε,δ)-DP であるとは “隣接する”データセット D～D’ に対して,

$$\forall S \subseteq Y.\ \Pr[M(D) \in S] \leq \exp(\varepsilon) \Pr[M(D') \in S] + \delta$$

  - 隣接関係は対称的であることが暗に仮定されている。それを外す。

- 不等式を形式化→差分プライバシー条件の形式化を行う。

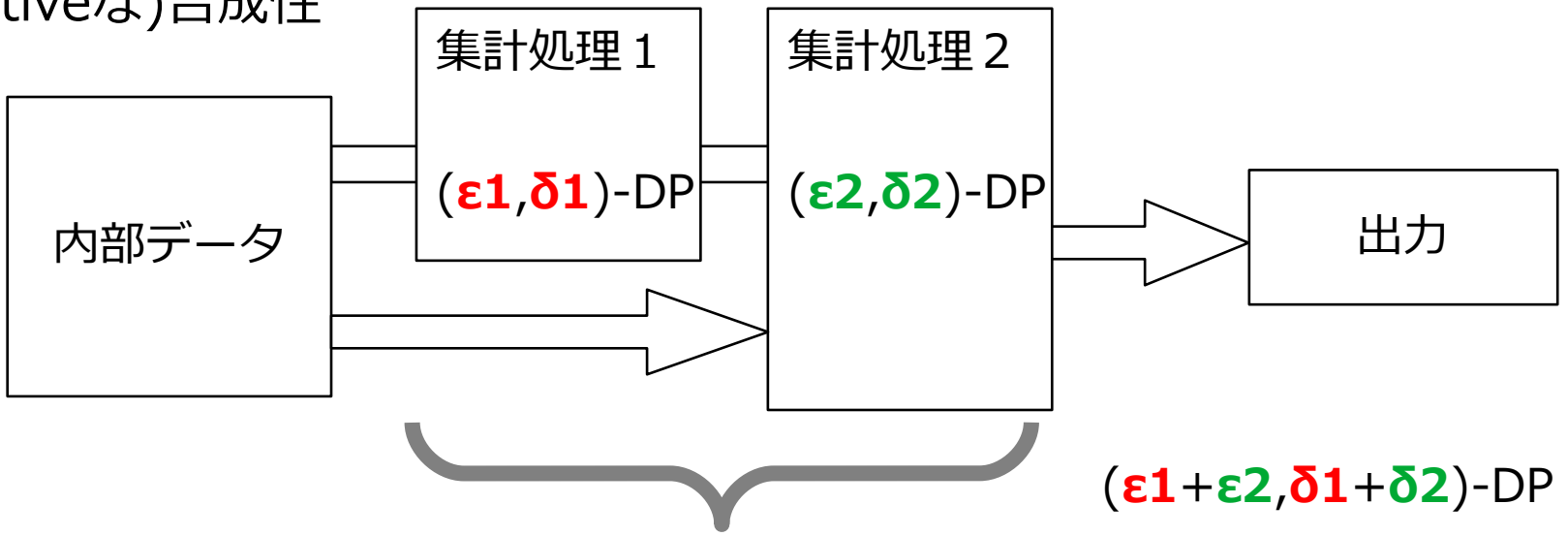
```
definition DP_inequality:: "'a measure ⇒ 'a measure ⇒ real ⇒ real ⇒ bool" where
  "DP_inequality M N ε δ ≡ (∀ A ∈ sets M. measure M A ≤ (exp ε) * measure N A + δ)"

definition differential_privacy :: "('a ⇒ 'b measure) ⇒ ('a rel) ⇒ real ⇒ real ⇒ bool "
  where
  "differential_privacy M adj ε δ ≡
  ∀(d1,d2)∈adj. DP_inequality (M d1) (M d2) ε δ ∧ DP_inequality (M d2) (M d1) ε δ"
```

隣接関係の対称性を外したことで２つに分裂

差分プライバシーの合成性

(Composition theorem)

- (adaptiveな)合成性

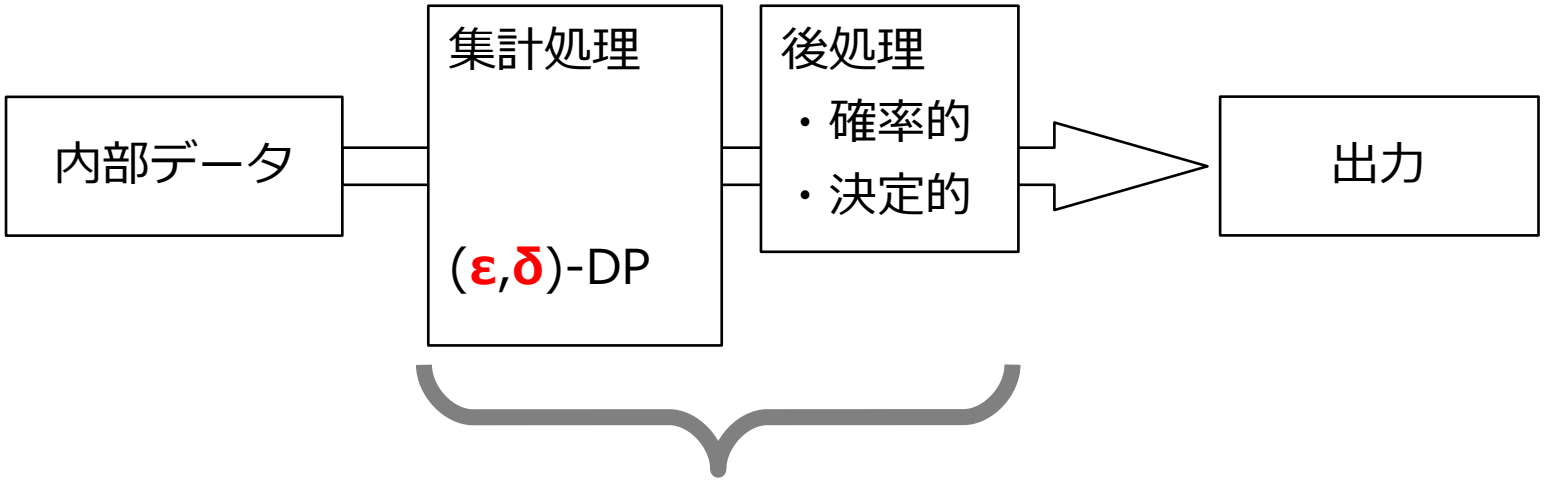



```
proposition differential_privacy_composition_adaptive:
  assumes "ε ≥ 0"
    and "ε' ≥ 0"
    and M: "M ∈ X →M (prob_algebra Y)"
    and DPM: "differential_privacy M adj ε δ"
    and N: "N ∈ (X ⨂M Y) →M (prob_algebra Z)"
    and DPN: "∀ y ∈ space Y. differential_privacy (λ x. N (x,y)) adj ε' δ'"
    and "adj ⊆ (space X) × (space X)"
  shows "differential_privacy (λx. do{y ← M x; N (x, y)}) adj (ε + ε') (δ + δ')"
```

後処理に対する安定性

(Postprocessing)

- 後処理(決定的でも確率的でも)を加えても、差分プライバシーは変化しない。




```
proposition differential_privacy_postprocessing:
  assumes "ε ≥ 0"
    and "differential_privacy M adj ε δ"
    and M: "M ∈ X →M (prob_algebra R)"
    and f: "f ∈ R →M (prob_algebra R')" (*probabilistic post-process*)
    and "adj ⊆ (space X) × (space X)"
  shows "differential_privacy (λx. do{y ← M x; f y}) adj ε δ"
```

ダイバージェンス Δ^{ε} の形式化

- ダイバージェンスを先に形式化する。
  - 定義　(※確率測度であることはまだ言ってない)

```
definition DP_divergence :: "'a measure ⇒ 'a measure ⇒ real ⇒ ereal " where
  "DP_divergence M N ε
    = Sup {ereal(measure M A - (exp ε) * measure N A) | A::'a set. A ∈ (sets M)}"
```

  - 反射性・単調性・合成性（※定理側で確率測度であることを言う）

```
lemma DP_divergence_monotonicity:
  assumes M: "M ∈ space (prob_algebra L)"
    and N: "N ∈ space (prob_algebra L)"
    and "ε1 ≤ ε2"
  shows "DP_divergence M N ε2 < DP_divergence M N ε1"
```

```
lemma DP_divergence_reflexivity:
  shows "DP_divergence M M 0 = 0"
```

```
proposition DP_divergence_composability:
  assumes M: "M ∈ space (prob_algebra L)"
    and N: "N ∈ space (prob_algebra L)"
    and f: "f ∈ L →M prob_algebra K"
    and g: "g ∈ L →M prob_algebra K"
    and div1: "DP_divergence M N ε1 ≤ (δ1 :: real)"
    and div2: "∀x ∈ (space L). DP_divergence (f x) (g x) ε2 ≤ (δ2 :: real)"
    and "0 ≤ ε1" and "0 ≤ ε2"
  shows "DP_divergence (M ≫= f) (N ≫= g) (ε1 + ε2) ≤ δ1 + δ2"
```

合成性の証明（スケッチ）

$$\Pr[\mu \gg\!= f \in S] - \exp(\varepsilon_1 + \varepsilon_2) \Pr[\nu \gg\!= g \in S]$$

$$= \int f(x)(S) d\mu - \exp(\varepsilon) \int g(x)(S) d\nu(x)$$

$$= \int f(x)(S) \cdot \frac{d\mu}{d\pi}(x)\ d\pi - \exp(\varepsilon_1 + \varepsilon_2) \int g(x)(S) \cdot \frac{d\nu}{d\pi}(x)\ d\pi(x)$$

$$= \int f(x)(S) \cdot \frac{d\mu}{d\pi}(x) - \exp(\varepsilon_1 + \varepsilon_2) g(x)(S) \cdot \frac{d\nu}{d\pi}(x)\ d\pi(x)$$

$$\leq \int (\max(0, f(x)(S) - \delta_2) + \delta_2) \cdot \frac{d\mu}{d\pi}(x)$$

$$- \exp(\varepsilon_1) \min(1, \exp(\varepsilon_2) g(x)(S)) \cdot \frac{d\nu}{d\pi}(x)\ d\pi(x)$$

$$= \int \max(0, f(x)(S) - \delta_2) \cdot \frac{d\mu}{d\pi}(x)$$

$$- \exp(\varepsilon_1) \min(1, \exp(\varepsilon_2) g(x)(S)) \cdot \frac{d\nu}{d\pi}(x)\ d\pi(x) + \int \delta_2 \frac{d\mu}{d\pi}(x)\ d\pi$$

$$\leq \int_B \left(\frac{d\mu}{d\pi}(x) - \exp(\varepsilon_1) \frac{d\nu}{d\pi}(x) \right) \cdot \min(1, \exp(\varepsilon_2) \cdot g(x)(S))\ d\pi + \int \delta_2 \frac{d\mu}{d\pi}(x)\ d\pi(x)$$

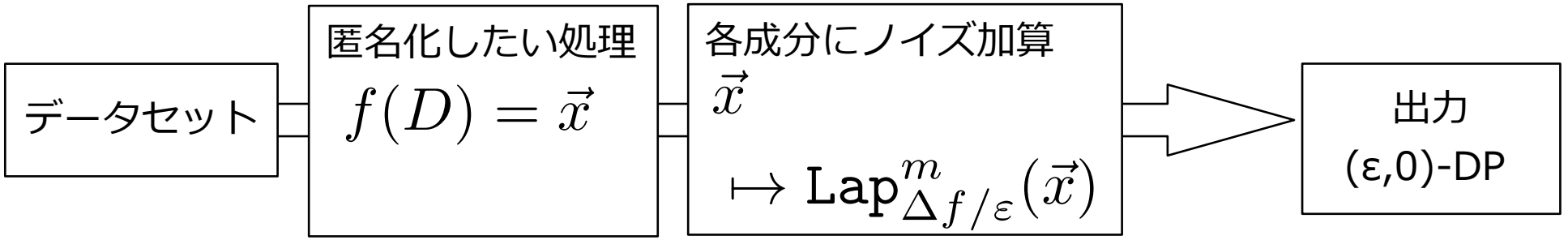
$$\leq \delta_1 + \delta_2$$

Giryモナドのbindを積分に展開

共通の測度 π を考え
密度関数を取る変形
(Radon-Nikodym)

移項や不等式変形が
たくさんある。

ラプラスメカニズムのDP



- 構成

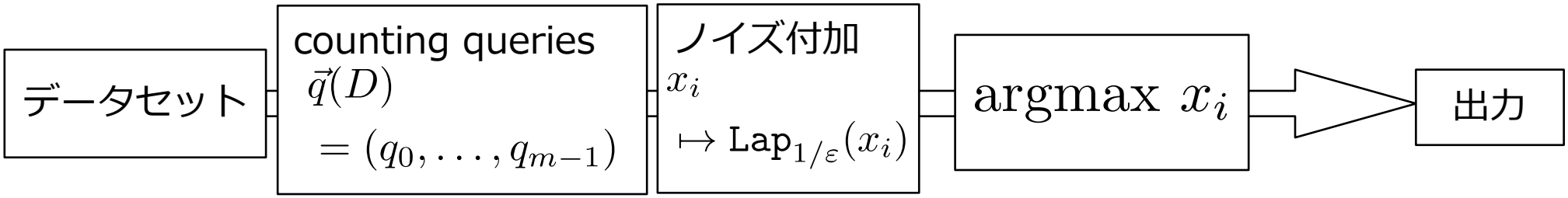
```
definition sensitivity:: ereal where
  "sensitivity = Sup{ ereal ( ∑ i∈{1..m}. ¦ nth (f x) (i-1) - nth (f y) (i-1) ¦)
                      | x y::'a. x ∈ space X ∧ y ∈ space X ∧ (x,y) ∈ adj}"

"LapMech_list ε x = Lap_dist_list ((real_of_ereal sensitivity) / ε) (f x)"
```

- 差分プライバシーの証明は、ラプラス分布の性質と合成性を使う。

```
proposition differential_privacy_LapMech_list:
  assumes pose: "ε > 0"
    and "sensitivity > 0"
    and "sensitivity < ∞"
  shows "differential_privacy (LapMech_list ε) adj ε 0"
```

Report Noisy Max メカニズムのDP

- Counting queryと argmaxを部品として形式化して、全体を形式化する。




```
definition RNM_counting :: "real ⇒ nat list ⇒ nat measure" where
  "RNM_counting ε x = do {
 y ← Lap_dist_list (1 / ε) (counting_query x);
 return (count_space UNIV) (argmax_list y)
 }"

theorem differential_privacy_LapMech_RNM_AFDP:
  assumes pose: "(ε::real) > 0"
  shows " differential_privacy (RNM_counting ε) adj_L1_norm ε 0"
```

ほぼゴリ押し(1500行)なので、
詳細は論文かAFPを見てください。

これからの課題

- 差分プライバシーを満たすアルゴリズムは数多く存在し
  形式化をやっていくことは考えているが、パワーが足りなそう。

- 差分プライバシーから少し離れたトピックを形式化し、
  差分プライバシーの結果を応用としてつなげることはできそう。
  - f-divergence[Csiszár, 1963][Morimoto, J. Phys. Soc. Jpn. 1963]

$$\Delta^f(\mu_1, \mu_2) = \int \mu_2(x) f\left(\frac{\mu_1(x)}{\mu_2(x)}\right) dx$$

fは特定の凸関数

    - 様々な特性を持つ。RDPや様々な統計的尺度の形式化につながる。
  - Blackwell informativeness theorem

$$T_X(\mu_1, \mu_2)(\alpha) = \inf\{\mathbb{E}_{\mu_2}[\neg\phi] | \phi \colon X \to \mathrm{prob}(2), \mathbb{E}_{\mu_1}[\phi] \leq \alpha\}$$

$$T_X(\mu_1, \mu_2) \geq T_Y(\nu_1, \nu_2) \iff \exists h \colon Y \to \mathrm{prob}(X). h^\sharp \nu_i = \mu_i$$

    - [Blackwell, 1954] がもとであるが、証明のギャップ埋めや一般化の論文がいくつも出ていて、昨年2024年にも出ていた。形式化は難しそう！
    - f-DP/Gaussian-DP [Dong+,2019] の理論的基礎。

ご清聴ありがとうございました！