Arbitrages statistiques dans l'apprentissage automatique confidentiel

Soutenance de stage

Alexi Canesse sous la supervision d'Aurélien Garivier, Professeur,
UMPA et École Normale Supérieure de Lyon

Stage de recherche effectué à l'UMPA dans le cadre de la
L3 informatique fondamental de l'ÉNS de Lyon

30 août 2022

Table des matières

Background essentiel sur la *differential privacy* (1/4)

La *differential privacy* [Dwo+06] quantifie la perte de confidentialité subit par un individu en participant à une étude.

Background essentiel sur la *differential privacy* (1/4)

La *differential privacy* [Dwo+06] quantifie la perte de confidentialité subit par un individu en participant à une étude.

Définition (Jeu de donnés voisins)

On dit que deux jeux de donnés x et y sont voisins et on note $\mathrm{d}_{\mathrm{Ham}}(x, y) \leq 1$ s'ils diffèrent sur au plus une entrée *ie* la distance de Hamming qui les sépare et majorée par 1.

Background essentiel sur la *differential privacy* (1/4)

La *differential privacy* [Dwo+06] quantifie la perte de confidentialité subit par un individu en participant à une étude.

Définition (Jeu de donnés voisins)

On dit que deux jeux de donnés x et y sont voisins et on note $\mathrm{d}_{\mathrm{Ham}}(x, y) \leq 1$ s'ils diffèrent sur au plus une entrée *ie* la distance de Hamming qui les sépare et majorée par 1.

Définition (Differential privacy)

On dit qu'un mécanisme aléatoire $\mathcal{M} : \mathcal{X}^{(\mathbb{N})} \to \mathcal{T}$ est ε-***differentially private*** si pour tout $\mathcal{S} \subset \mathcal{T}$ mesurable,

$$\forall x, y \in \mathcal{X}^{(\mathbb{N})} \quad \mathrm{d}_{\mathrm{Ham}}(x, y) \leq 1 \quad \Rightarrow \quad \mathbb{P}(\mathcal{M}(x) \in \mathcal{S}) \leq \exp(\varepsilon)\mathbb{P}(\mathcal{M}(y) \in \mathcal{S}).$$

a. La notation $\mathcal{X}^{(\mathbb{N})}$ désigne, de manière usuelle, l'ensemble des suites finies de $\mathcal{X}$.

Background essentiel sur la *differential privacy* (2/4)

Définition (Sensibilitée d'une requête)

Soit $\mathcal{X}$ un ensemble, $(\mathcal{T}, \mathcal{N})$ un espace mesuré et $f : \mathcal{X}^{(\mathbb{N})} \to \mathcal{T}$ une requête. On appel **sensibilité de** f la grandeur Δf que l'on définit de la manière suivante :

$$\Delta f = \sup_{x,y \in \mathcal{X}^{(\mathbb{N})}} \{\mathcal{N}(f(x), f(y)) \mid \mathrm{d}_{\mathrm{Ham}}(x, y) \leq 1\}.$$

Background essentiel sur la *differential privacy* (2/4)

Définition (Sensibilitée d'une requête)

Soit $\mathcal{X}$ un ensemble, $(\mathcal{T}, \mathcal{N})$ un espace mesuré et $f : \mathcal{X}^{(\mathbb{N})} \to \mathcal{T}$ une requête. On appel **sensibilité de** f la grandeur Δf que l'on définit de la manière suivante :

$$\Delta f = \sup_{x,y \in \mathcal{X}^{(\mathbb{N})}} \{\mathcal{N}(f(x), f(y)) \mid \mathrm{d}_{\mathrm{Ham}}(x, y) \leq 1\}.$$

Définition (Mécanisme de Laplace [Dwo+06])

Soit $\mathcal{X}$ un ensemble de base, $\varepsilon \in \mathbb{R}_+^\star$, $n \in \mathbb{N}$ et $f : \mathcal{X}^{(\mathbb{N})} \to \mathbb{R}$ une requête. Notons Lap la fonction qui retourne une variable aléatoire suivant la loi de Laplace dont la paramètre est l'argument. On appel **mécanisme de Laplace** la fonction

$$\mathcal{M}_{f,\varepsilon} : \left\{ \begin{array}{ccc} \mathcal{X}^{(\mathbb{N})} & \to & \mathbb{R} \\ x & \mapsto & f(x) + \mathrm{Lap}\left(\Delta f / \varepsilon\right) \end{array} \right.$$

a. Il s'agit de la loi de densité $x \mapsto 1/(2b) \exp(-|x|/b)$ où b est le paramètre.

Background essentiel sur la *differential privacy* (3/4)

Définition (Mécanisme de Laplace [Dwo+06])

Soit $\mathcal{X}$ un ensemble de base, $\varepsilon \in \mathbb{R}_+^\star$, $n \in \mathbb{N}$ et $f : \mathcal{X}^{(\mathbb{N})} \to \mathbb{R}$ une requête. Notons Lap la fonction qui retourne une variable aléatoire suivant la loi de Laplace dont la paramètre est l'argument. On appel **mécanisme de Laplace** la fonction

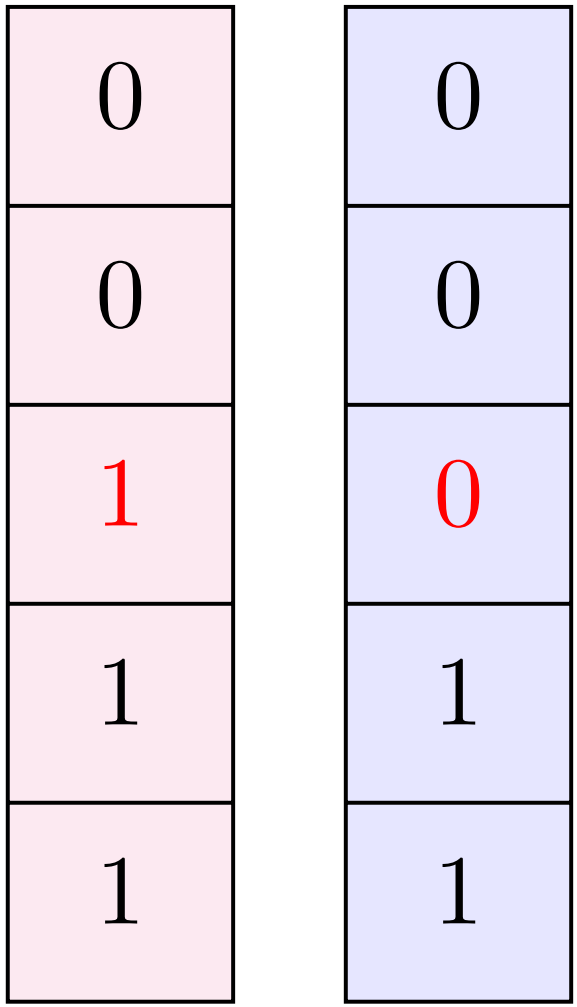
$$\mathcal{M}_{f,\varepsilon} : \left\{ \begin{array}{ccc} \mathcal{X}^{(\mathbb{N})} & \to & \mathbb{R} \\ x & \mapsto & f(x) + \mathrm{Lap}\left(\Delta f/\varepsilon\right) \end{array} \right.$$

Théorème

Le mécanisme de Laplace *est differentially private.*

L'échec de la méthode naïve

Soit $x \in [0,1]^n$. On a $\Delta\text{méd} = 1$.



L'échec de la méthode naïve

Soit $x \in [0,1]^n$. On a $\Delta\text{méd} = 1$.

Or,

$$|\text{méd}(x) - \mathcal{M}_{\text{méd},\varepsilon}| = |\text{Lap}(1/\varepsilon)|.$$

L'échec de la méthode naïve

Soit $x \in [0,1]^n$. On a $\Delta\text{méd} = 1$.

Or,

$$|\text{méd}(x) - \mathcal{M}_{\text{méd},\varepsilon}| = |\text{Lap}(1/\varepsilon)|.$$

D'où,

$$\mathbb{E}\left(|\text{méd}(x) - \mathcal{M}_{f,\varepsilon}|\right) = \int_{\mathbb{R}} \mathbb{P}\left(|\text{Lap}(1/\varepsilon)| > t\right) \mathrm{d}t = \int_{\mathbb{R}_+} e^{-\varepsilon t} \mathrm{d}t = \frac{1}{\varepsilon}.$$

L'échec de la méthode naïve

Soit $x \in [0,1]^n$. On a $\Delta\text{méd} = 1$.

Or,

$$|\text{méd}(x) - \mathcal{M}_{\text{méd},\varepsilon}| = |\text{Lap}(1/\varepsilon)|.$$

D'où,

$$\mathbb{E}\left(|\text{méd}(x) - \mathcal{M}_{f,\varepsilon}|\right) = \int_{\mathbb{R}} \mathbb{P}\left(|\text{Lap}(1/\varepsilon)| > t\right) \mathrm{d}t = \int_{\mathbb{R}_+} e^{-\varepsilon t} \mathrm{d}t = \frac{1}{\varepsilon}.$$

Néanmoins nous avons forcement $\text{méd}(x) \in [0,1]$. **Cette méthode ne peut donc pas convenir**.

Table des matières

AboveThreshold [Dwo+06]

```
1  AboveThreshold(database, queries, threshold, epsilon){
2    Assert("les requêtes sont toutes de sensibilité 1");
3    result = 0;
4    noisyThreshold = threshold + Lap(2/epsilon);
5    for(querie in queries){
6        nu = Lap(4/epsilon);
7        if(querie(D) + nu > noisyThreshold)
8            return result;
9        else
10           ++result;
11   }
12   return -1; /* Aucune requête n'a dépassé le seuil */
13 }
```

Théorème

`AboveThreshold` est ε-differentially private.

Présentation de la méthode des histogrammes

```
1  HistogramMethod(database, epsilon, a, b){
2    steps = 1.5*n/log(n);
3    epsilon /= 9; /* composition theorem */
4    result = {};
5    for(d in {1 ... 9}){ /* which decile */
6        T = d*card(database)/10;
7        for(i in {1 ... steps}){
8            fi = x -> card({element in x | element < i*(b-a)
9                                                    /steps});
10           queries.push_back(fi);
11       }
12       T = d*card(database)/10;
13       result.push_back(AboveThreshold(database, queries, T,
14                                       epsilon)*(b-a)/steps});
15   }
16   return result;
17 }
```

Analyse de précision - le cas de la loi uniforme standard (2/4)

Corollaire ((im)Précision moyenne de `HistogramMethod`)

Soit X un ensemble de n (tel que $8\log(3n\sqrt{n}))/\varepsilon) \leq n/10$) variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme standard. Soit $i \in [\![1, 9]\!]$. Notons $(d_i^l)_i$ les déciles de la loi. Posons A la variable aléatoire `HistogramMethod(X, epsilon, 0, 1)`, $\alpha = 8\log(3n\sqrt{n})/\varepsilon)$. *On a*

$$\begin{aligned}
\mathbb{E}\left(|A_i - d_i^l|\right) \leq & \int_0^{d_i^l} \left(1 - I_{d_i^l + t}(\lceil in/10 + \alpha \rceil, n - \lceil in/10 + \alpha \rceil + 1)\right) \mathrm{d}t \\
& + \int_0^{d_i^l} I_{d_i^l - t}(\lfloor in/10 - \alpha \rfloor, n - \lfloor in/10 - \alpha \rfloor + 1)\mathrm{d}t + \frac{1}{\sqrt{n}\log n} \\
& + I_{d_i^l - 0.1}(\lfloor in/10 - \alpha \rfloor, n - \lfloor in/10 - \alpha \rfloor + 1) \\
& + I_{1 - d_i^l - 0.1}(n - \lceil in/10 + \alpha \rceil + 1, \lceil in/10 + \alpha \rceil) \\
& + \frac{2\log n}{3n} + \frac{d_i^l}{\sqrt{n}\log n}.
\end{aligned}$$

Analyse de précision - le cas de la loi uniforme standard (3/4)

Théorème ((im)Précision moyenne de `HistogramMethod`)

Soit X un ensemble de n (tel que $0 \leq 8\log(3n\sqrt{n})/\varepsilon) \leq n/10$) variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur [0,1]. Soit $i \in [\![1, 9]\!]$ et $k \in \mathbb{N}$. Notons $(d_i)_i$ les décile de la loi. Posons A la variable aléatoire `HistogramMethod(X, epsilon, 0, 1)` *et $\alpha = 8\log(3n\sqrt{n})/\varepsilon)$. On a,*

$$\mathbb{E}\left(|A_i - d_i^l|\right) \leq 2\sqrt{\frac{\pi}{2n}} + \frac{d_i^l + 1}{\sqrt{n}\log n} + \frac{\log n}{n}\left(\frac{2}{3} + \frac{16}{\varepsilon}\log(3)\right) + 2\exp\left(-2n\left(0.1 - \frac{\alpha}{n}\right)^2\right).$$

Table des matières

Présentation du mécanisme (1/2)

Définition (Longueur)

Soit $x \in \mathcal{X}^{(\mathbb{N})}$, $f\colon \mathcal{X}^{(\mathbb{N})} \to \mathcal{T}$ et $t \in \mathcal{T}$. La longueur est le nombre minimum de valeurs à modifier dans x pour obtenir x' tel que $f(x') = t$. On a

$$\mathrm{len}_f(x, t) := \inf_{x' \in \mathcal{X}^{(\mathbb{N})}} \{||x - x'||_1 \mid f(x') = t\}.$$

Définition (Mécanisme de sensibilité inverse [AD20])

Soit $f\colon \mathcal{X}^{(\mathbb{N})} \to \mathcal{T}$ et $\varepsilon \in \mathbb{R}_+$. Pour une mesure μ sur $\mathcal{T}$, on définit le mécanisme aléatoire $\mathcal{M}(x)$ par sa fonction de densité

$$t \mapsto \frac{\exp(-\mathrm{len}_f(x, t)\varepsilon/2)}{\int_{\mathcal{T}} \exp(-\mathrm{len}_f(x, s)\varepsilon/2)\mathrm{d}\mu(s)}.$$

Présentation du mécanisme (2/2)

Définition (Longueur lisse)

Soit $x \in \mathcal{X}^{(\mathbb{N})}$, $f : \mathcal{X}^{(\mathbb{N})} \to \mathcal{T}$ et $\rho \in \mathbb{R}_+$. Si $\mathcal{N}$ est une norme sur $\mathcal{T}$,

$$\mathrm{len}_f^\rho : \begin{cases} \mathcal{T} & \to & \mathbb{N} \\ t & \mapsto & \inf_{s \in \mathcal{T}, \mathcal{N}(s,t) \leq \rho} \{\mathrm{len}_f(x, s)\} \end{cases}.$$

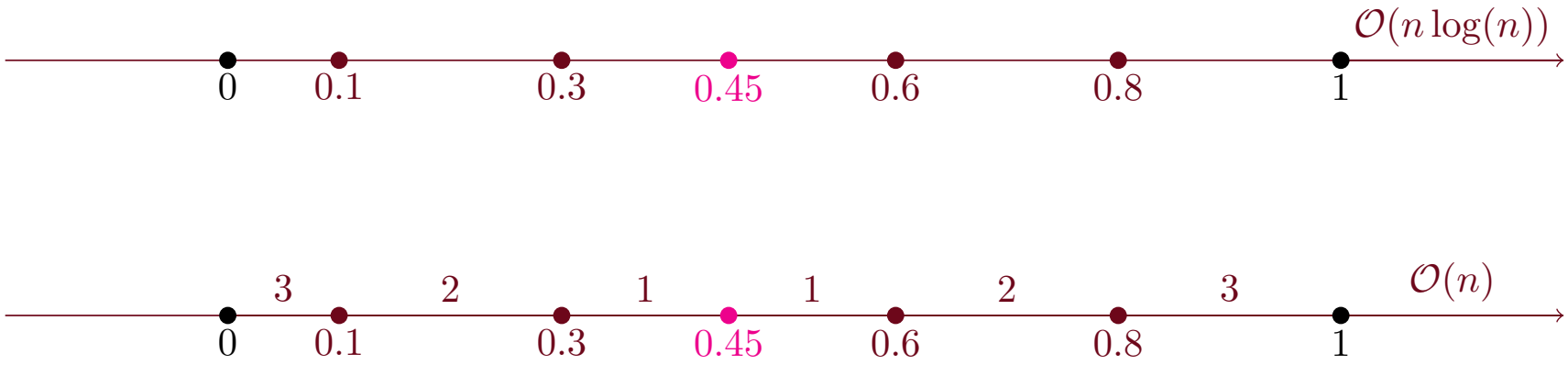
Définition (Mécanisme de sensibilité inverse ρ-lisse [AD20])

Soit $f : \mathcal{X}^{(\mathbb{N})} \to \mathcal{T}$ et $\rho, \varepsilon \in \mathbb{R}_+$. Pour une mesure μ sur $\mathcal{T}$, on définit le mécanisme aléatoire $\mathcal{M}_{\mathrm{cont}}(x)$ par sa fonction de densité

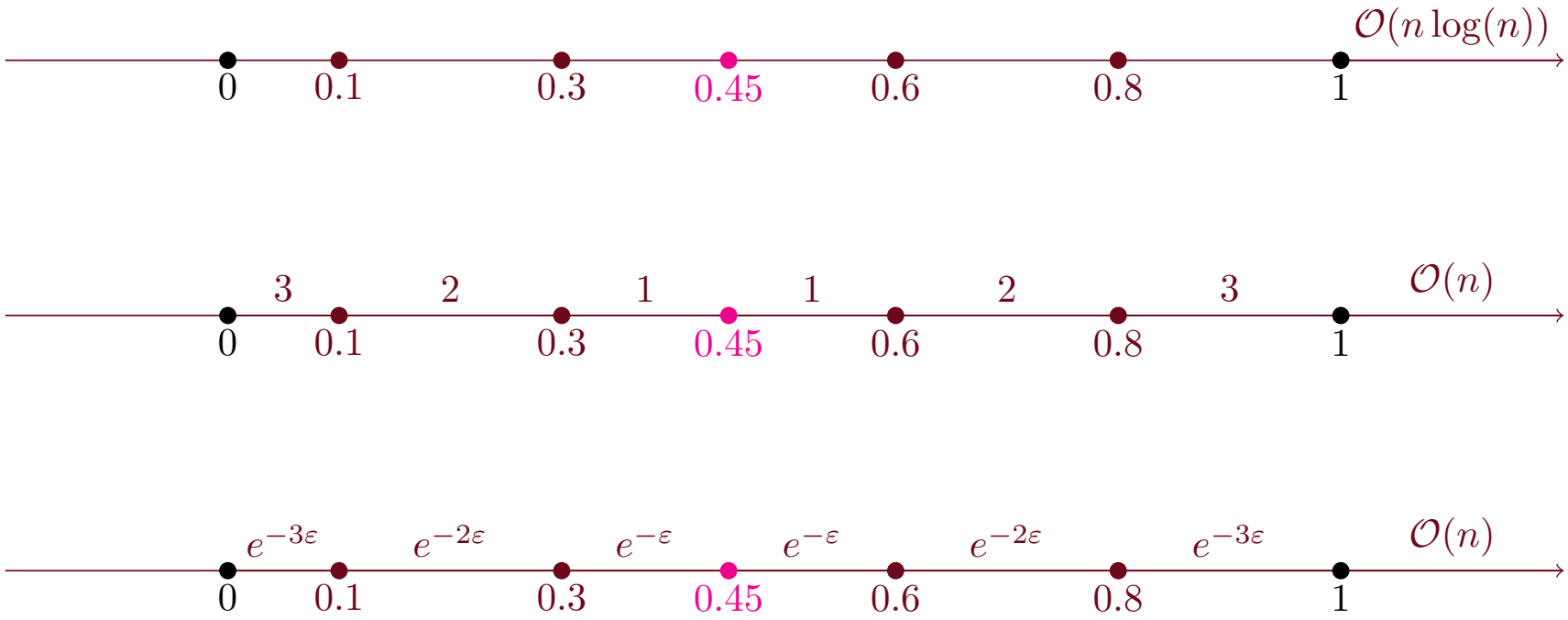
$$t \mapsto \frac{\exp(-\mathrm{len}_f^\rho(x, t)\varepsilon/2)}{\int_{\mathcal{T}} \exp(-\mathrm{len}_f^\rho(x, s)\varepsilon/2)\mathrm{d}\mu(s)}.$$

Analyse de complexité

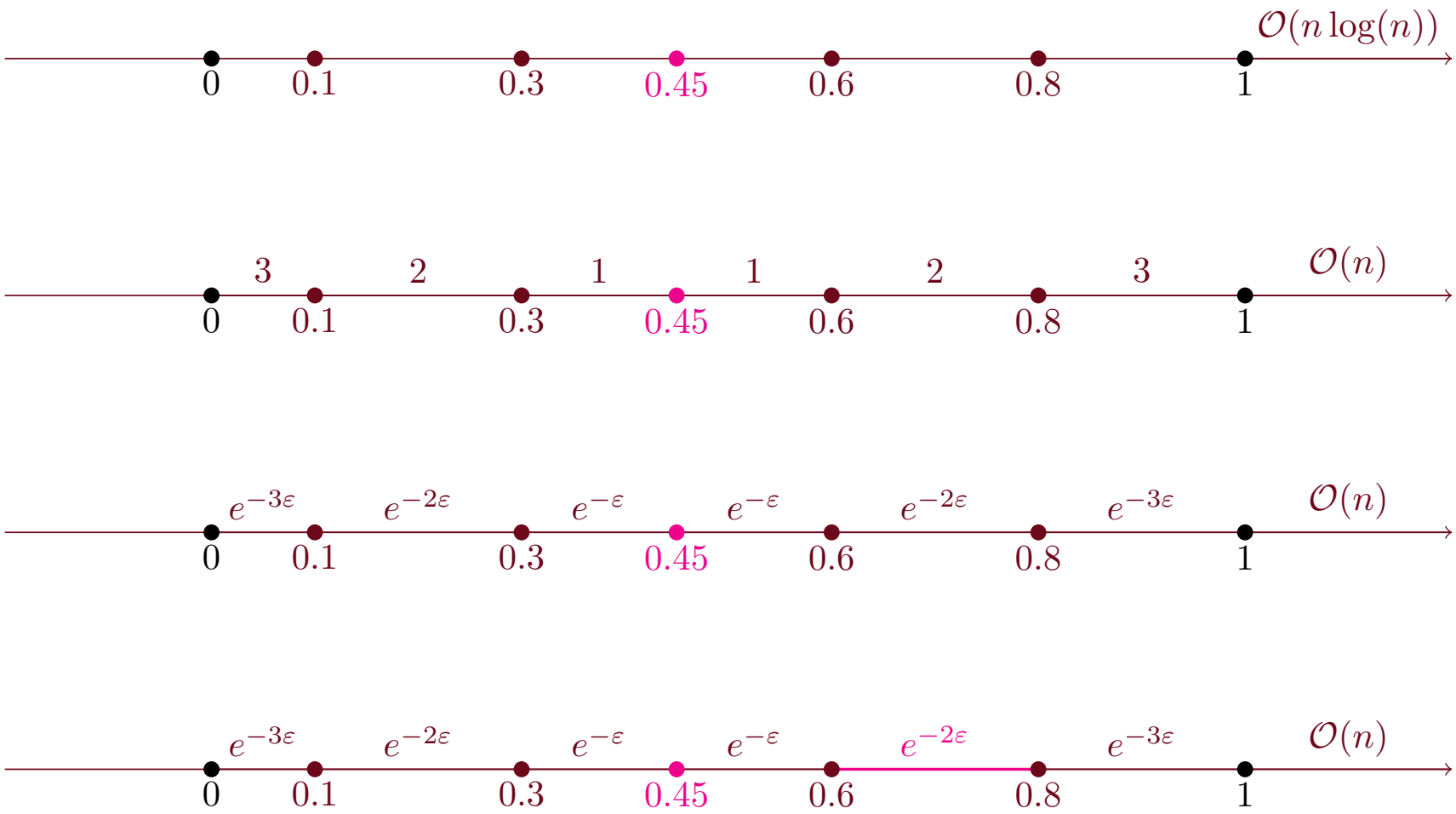
Analyse de complexité



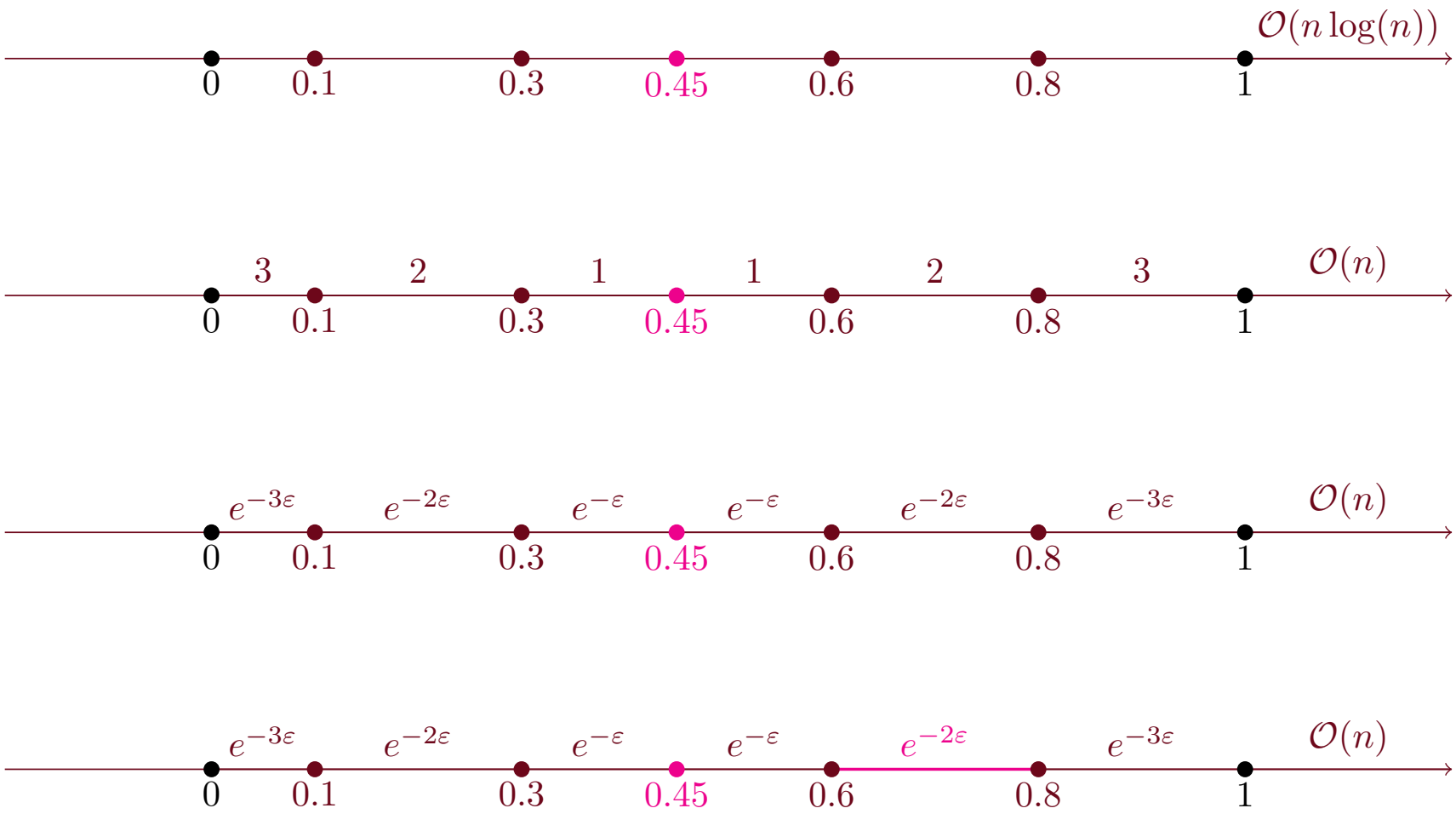
Analyse de complexité



Analyse de complexité



Analyse de complexité



La complexité de l'algorithme est donc au pire en $\mathcal{O}(n\log(n))$.

Précision pour l'estimation de déciles

Théorème

Soit $\rho \in]0, 1 - 4/\sqrt{n}]$ et $X \in [0,1]^n$ dont les éléments sont obtenues à partir de la loi uniforme standard. On note $(d_i^l)_i$ les déciles de la loi. Notons alors enfin $\mathcal{M}_{cont}$ le mécanisme de sensibilité inverse ρ-lisse.

$$\mathbb{E}(|\mathcal{M}_{cont,i} - d_i^l|) \leq \rho + \frac{4}{\sqrt{n}} + \frac{4}{n\varepsilon\rho} \exp\left(-\frac{\sqrt{n}\varepsilon}{2}\right) + \frac{16}{\sqrt{n}} \exp\left(-\frac{\sqrt{n}}{4}\right).$$

Table des matières

Résultats expérimentaux

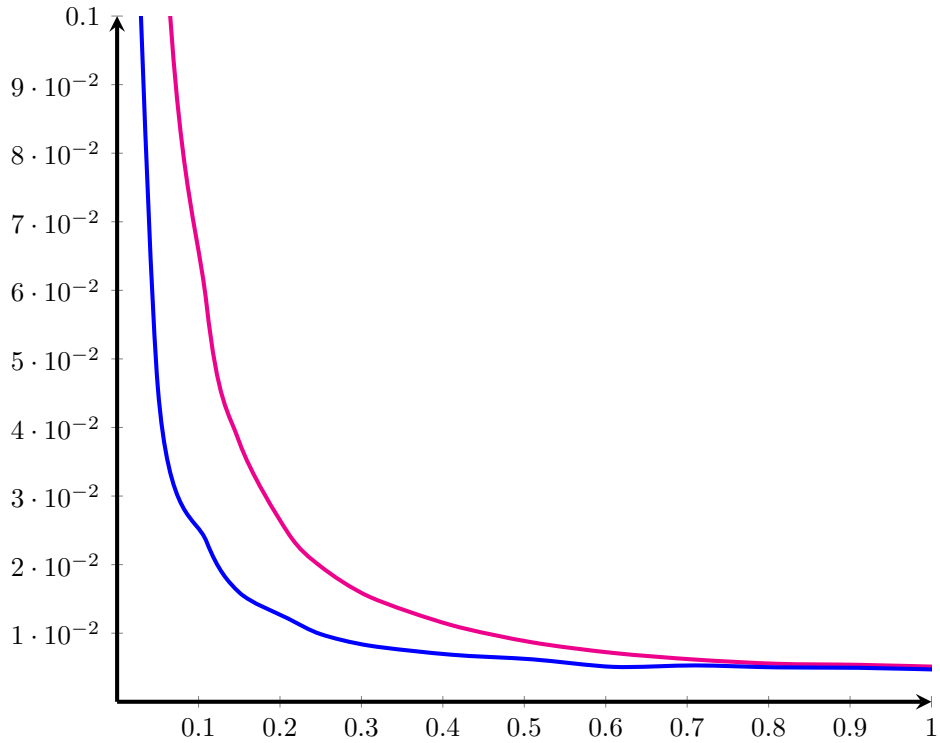


Figure – Écart-quadratique moyen sur le calcul des déciles en fonction de ε pour $n = 10\ 000$. La méthode des histogrammes est en magenta et le mécanisme de sensibilité inverse est en bleu.

Résultats expérimentaux

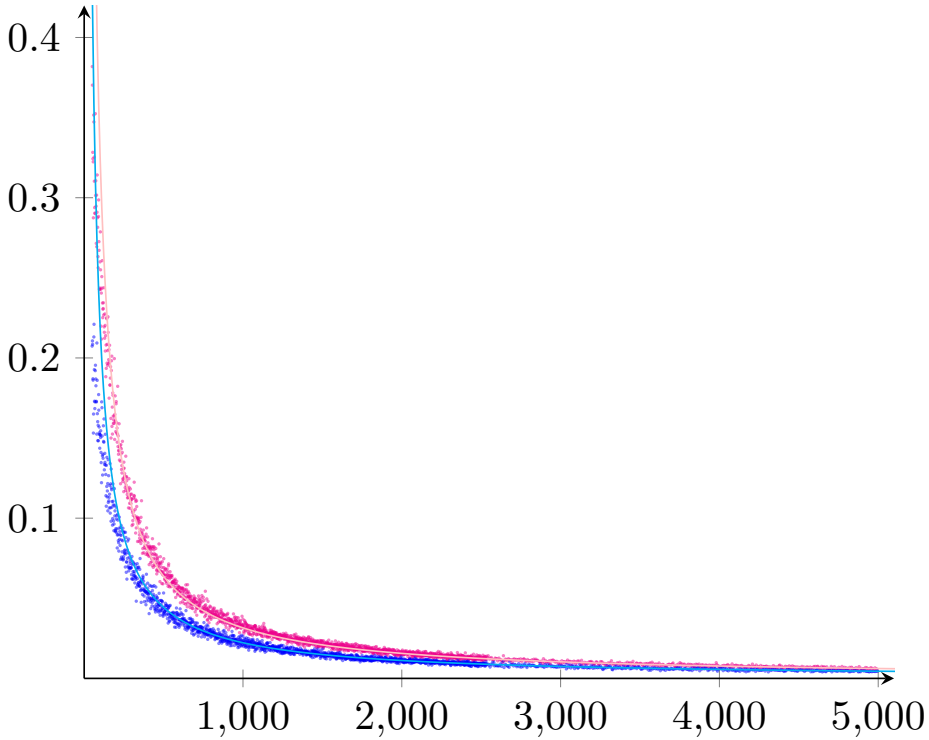


Figure – Écart-quadratique moyen sur le calcul des déciles en fonction de n (la taille de l'échantillon) avec $\varepsilon = 1$. La méthode des histogrammes est en magenta et le mécanisme de sensibilité inverse est en bleu.

Résultats expérimentaux

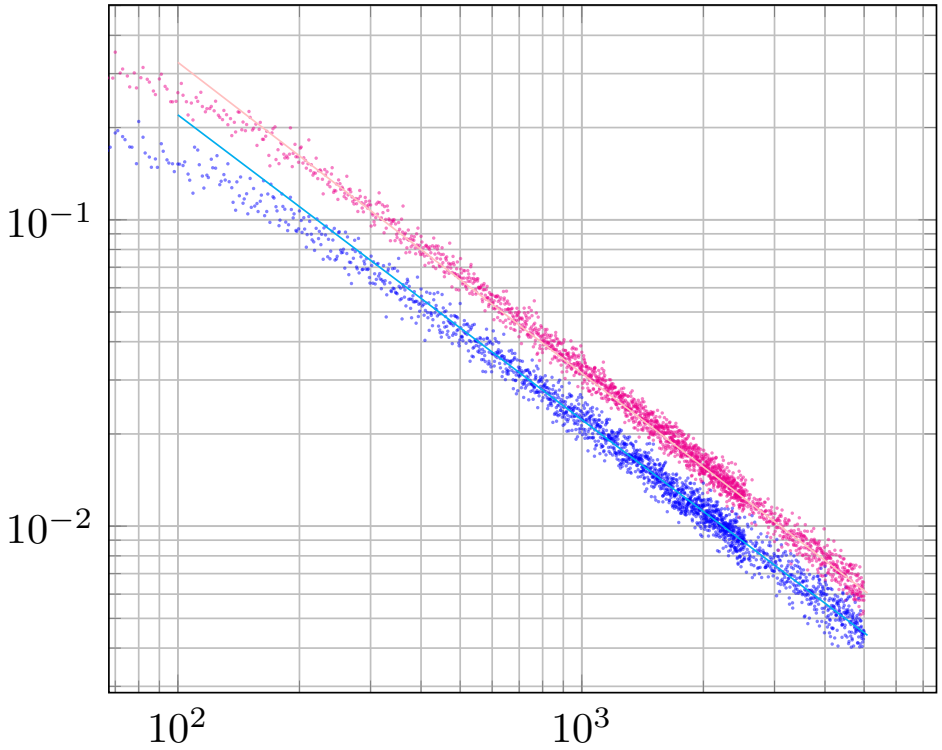


Figure – Écart-quadratique moyen sur le calcul des déciles en fonction de n (la taille de l'échantillon) avec $\varepsilon = 1$. La méthode des histogrammes est en magenta et le mécanisme de sensibilité inverse est en bleu.

Données réelles

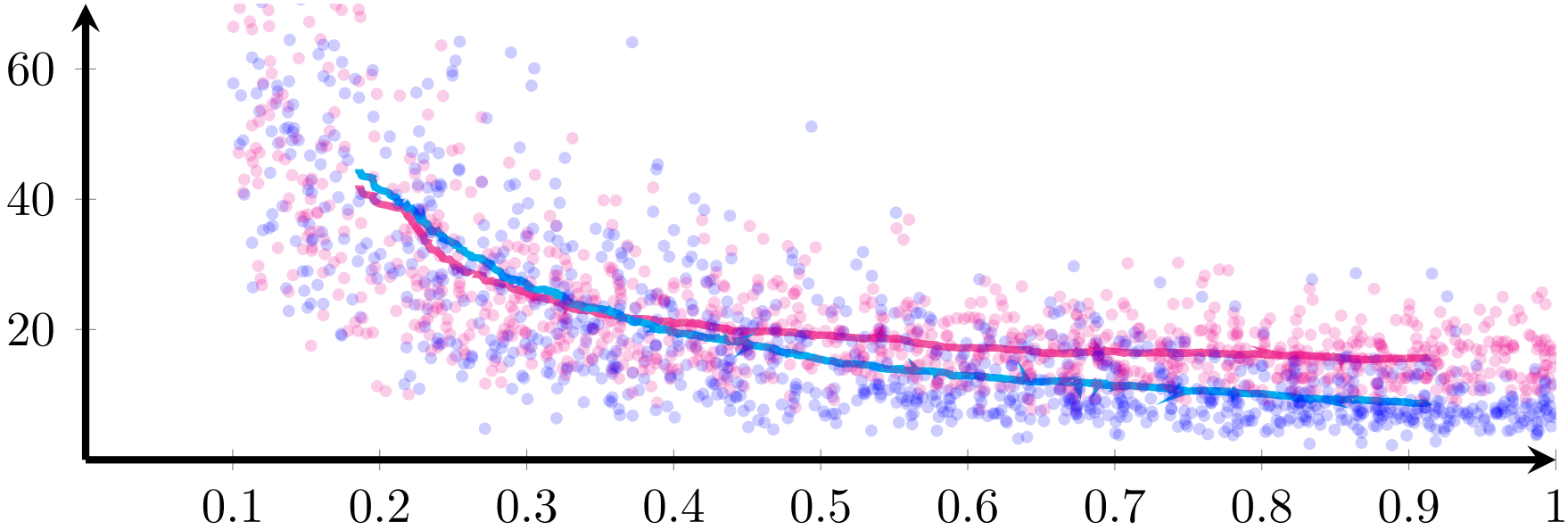


Figure – Écart-quadratique sur le calcul des déciles en fonction de ε. La méthode des histogrammes est en magenta et le mécanisme de sensibilité inverse est en bleu.

Table des matières

Conclusion


WE PUT A BANNER ON OUR WEBSITE SAYING, "WE RESPECT YOUR PRIVACY"
AND THEN PROCEEDED TO TELL VISITORS HOW WE SELL THEIR PERSONAL DATA
imgflip.com


YOU CAN'T HAVE A DATA BREACH
IF YOU DON'T SECURE YOUR DATA

Table des matières

Référence I

[AD20] Hilal Asi et John C. Duchi. « Near Instance-Optimality in Differential Privacy ». In : *ArXiv* abs/2005.10630 (mai 2020). URL : https://arxiv.org/pdf/2005.10630.pdf.

[Aux+19] Brooke Auxier et al. *Americans and Privacy : Concerned, Confused and Feeling Lack of Control Over Their Personal Information*. 15 nov. 2019. URL : https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2019/11/Pew-Research-Center_PI_2019.11.15_Privacy_FINAL.pdf (visité le 20/07/2022).

[Dwo+06] Cynthia Dwork et al. « Calibrating Noise to Sensitivity in Private Data Analysis ». In : *Theory of Cryptography*. Sous la dir. de Shai Halevi et Tal Rabin. Berlin, Heidelberg : Springer Berlin Heidelberg, 2006, p. 265-284. ISBN : 978-3-540-32732-5.

[Mon+13] Yves-Alexandre de Montjoye et al. « Unique in the Crowd : The privacy bounds of human mobility ». In : *Nature* 3 (mars 2013). URL : https://doi.org/10.1038/srep01376.

[PE16] European Parliament et Council of the European Union. *Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive)*. 14 avr. 2016. URL : https://gdpr-info.eu/recitals/no-26/ (visité le 20/07/2022).

Référence II

[Swe00] Latanya Sweeney. « Simple Demographics Often Identify People Uniquely ». In : (jan. 2000). DOI : 10.1184/R1/6625769.v1. URL : https://dataprivacylab.org/projects/identifiability/paper1.pdf (visité le 20/07/2022).

De l'importance de respecter la confidentialité

"79% of adults assert they are very or somewhat concerned about how companies are using the data they collect about them, while 64% say they have the same level of concern about government data collection"

"a majority think the potential risks of data collection outweigh the benefits"

Brooke Auxier et al. *Americans and Privacy : Concerned, Confused and Feeling Lack of Control Over Their Personal Information*. 15 nov. 2019. URL : https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2019/11/Pew-Research-Center_PI_2019.11.15_Privacy_FINAL.pdf (visité le 20/07/2022)

Anonymiser les données n'est pas suffisant

Anonymiser les données n'est pas suffisant

European Parliament et Council of the European Union. *Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive)*. 14 avr. 2016. URL : https://gdpr-info.eu/recitals/no-26/ (visité le 20/07/2022)

Anonymiser les données n'est pas suffisant

European Parliament et Council of the European Union. *Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive)*. 14 avr. 2016. URL : https://gdpr-info.eu/recitals/no-26/ (visité le 20/07/2022)

Yves-Alexandre de Montjoye et al. « Unique in the Crowd : The privacy bounds of human mobility ». In : *Nature* 3 (mars 2013). URL : https://doi.org/10.1038/srep01376

Anonymiser les données n'est pas suffisant

European Parliament et Council of the European Union. *Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive)*. 14 avr. 2016. URL : https://gdpr-info.eu/recitals/no-26/ (visité le 20/07/2022)

Yves-Alexandre de Montjoye et al. « Unique in the Crowd : The privacy bounds of human mobility ». In : *Nature* 3 (mars 2013). URL : https://doi.org/10.1038/srep01376

Latanya Sweeney. « Simple Demographics Often Identify People Uniquely ». In : (jan. 2000). DOI : 10.1184/R1/6625769.v1. URL : https://dataprivacylab.org/projects/identifiability/paper1.pdf (visité le 20/07/2022)

Théorème de composition

Théorème (Théorème de composition (simple) [Dwo+06])

Soit $\mathcal{X}$ un ensemble de base, $n \in \mathbb{N}$ un nombre de mécanismes, $(\mathcal{T}_i)_{i \leq n}$ des ensembles d'arrivé et $(\mathcal{M}_i : \mathcal{X}^{(\mathbb{N})} \to \mathcal{T}_i)_{i \leq n}$ des mécanismes mutuellement indépendants respectivement ε_i-differentially private. L'exécution des n mécanismes est $\left(\sum_{i=1}^{n} \varepsilon_i\right)$-differentially private.

Théorème de composition

Théorème (Théorème de composition (simple) [Dwo+06])

Soit $\mathcal{X}$ un ensemble de base, $n \in \mathbb{N}$ un nombre de mécanismes, $(\mathcal{T}_i)_{i \leq n}$ des ensembles d'arrivé et $(\mathcal{M}_i : \mathcal{X}^{(\mathbb{N})} \to \mathcal{T}_i)_{i \leq n}$ des mécanismes mutuellement indépendants respectivement ε_i-differentially private. L'exécution des n mécanismes est $\left(\sum_{i=1}^{n} \varepsilon_i\right)$-differentially private.

Démonstration.

Considérons $(\mathcal{S}_i)_{i \leq n} \subset \prod_{i=1}^{n} \mathcal{T}_i$ et $x, x' \in \mathcal{X}^{(\mathbb{N})}$ tel que $\mathrm{d}_{\mathrm{Ham}(x,x') \leq 1}$.

$$\begin{aligned}
\mathbb{P}\left(\mathcal{M}_1(x) \in \mathcal{S}_1 \ \wedge \ \mathcal{M}_2(x) \in \mathcal{S}_2 \ \wedge \dots\right) &\overset{\text{indé.}}{=} \prod_{i=1}^{n} \mathbb{P}\left(\mathcal{M}_i(x) \in \mathcal{S}_i\right) \\
&\overset{\text{DP}}{\leq} \prod_{i=1}^{n} e^{\varepsilon_i} \mathbb{P}\left(\mathcal{M}_i(x') \in \mathcal{S}_i\right) \\
&\overset{\text{indé.}}{=} e^{\sum_{i=1}^{n} \varepsilon_i} \mathbb{P}\left(\mathcal{M}_1(x') \in \mathcal{S}_1 \ \wedge \dots\right)
\end{aligned}$$

□

Analyse de précision - le cas de la loi uniforme standard (1/5)

Définition (Fonction Beta incomplète (régularisée))

On appel respectivement fonction beta incomplète et fonction beta incomplète régularisée les fonctions

$$\mathrm{B} : \begin{cases} [0,1] \times (\mathbb{R}_+^\star)^2 & \to & \mathbb{R}_+ \\ (x, \alpha, \beta) & \mapsto & \displaystyle\int_0^x t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t \end{cases}$$

$$I_\bullet : \begin{cases} (\mathbb{R}_+^\star)^2 & \to & \mathbb{R}_+ \\ (\alpha, \beta) & \mapsto & \dfrac{B(\bullet, \alpha, \beta)}{B(1, \alpha, \beta)} \end{cases} .$$

Analyse de précision - le cas de la loi uniforme standard (1/5)

Définition (Fonction Beta incomplète (régularisée))

On appel respectivement fonction beta incomplète et fonction beta incomplète régularisée les fonctions

$$\mathrm{B} : (x, \alpha, \beta) \mapsto \int_0^x t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t \quad \text{et} \quad I_\bullet : (\alpha, \beta) \mapsto \frac{B(\bullet, \alpha, \beta)}{B(1, \alpha, \beta)}.$$

Définition (Loi beta)

On appel loi beta de paramètre $(\alpha, \beta) \in \mathbb{R}_+^\star$ la loi de densité

$$f_{\alpha,\beta} : [0,1] \ni x \mapsto \frac{x^{\alpha-1}(1-x)^{\beta}-1}{B(1, \alpha, \beta)}$$

Analyse de précision - le cas de la loi uniforme standard (1/4)

Définition (Statistique d'ordre)

Soit X un échantillon statistique de cardinal $n \in \mathbb{N}$. Pour tout $k \in [\![1, n]\!]$ on note $X_{(i)}$ et on appel **statistique d'ordre** de rang k la k-ème plus petite valeur de l'échantillon.

Théorème (Loi des statistiques d'ordre d'un échantillon issue de $\mathcal{U}(0,1)$)

Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur [0,1] et $k \in [\![1, n]\!]$. La k-ème statistique d'ordre de X, $X_{(k)}$, est distribuée suivant la loi beta de paramètre $(k, n-k+1)$.

Précision pour l'estimation de déciles

Théorème (Ecart avec les déciles empiriques)

Soit $\gamma \in \mathbb{R}_+^\star$, $u \in [0, \gamma/4]$, $\rho \in \mathbb{R}_+$ et $X \in [0,R]^n$ dont les éléments sont obtenues à partir d'une loi P de densité π_P continue au voisinage de ses déciles. On pose $p_{min,i} = \inf_{t \in [d_i^l - 2\gamma, d_i^l + 2\gamma]} \pi_P(t)$. On note $(d_i)_i$ les déciles empirique de X et $(d_i^l)_i$ les déciles de la loi. Notons alors enfin $\mathcal{M}_{cont}$ le mécanisme de sensibilité inverse ρ-lisse. On a

$$\mathbb{P}\left(|\mathcal{M}_{cont,i} - d_i| > 2u + \rho\right) \leq \frac{R}{2\rho} \exp\left(-\frac{np_{min,i}u\varepsilon}{4}\right) + 4\exp\left(-\frac{n\gamma^2 p_{min,i}^2}{8}\right) + \frac{2\gamma}{u}\exp\left(-\frac{np_{min,i}u}{8}\right).$$

Précision pour l'estimation de déciles

Théorème (Ecart avec les déciles théoriques)

Soit $\gamma \in \mathbb{R}_+^\star$, $u \in [0, \gamma/4]$, $\rho \in \mathbb{R}_+$ et $X \in [0, R]^n$ dont les éléments sont obtenues à partir d'une loi P de densité π_P continue au voisinage de ses déciles. On pose $p_{min,i} = \inf_{t \in [d_i^l - 2\gamma, d_i^l + 2\gamma]} \pi_P(t)$. On note $(d_i^l)_i$ les déciles de la loi. Notons alors enfin $\mathcal{M}_{cont}$ le mécanisme de sensibilité inverse ρ-lisse. On a

$$\mathbb{P}\left(|\mathcal{M}_{cont,i} - d_i^l| > 2u + \rho\right) \leq \frac{R}{2\rho} \exp\left(-nup_{min,i}\varepsilon/4\right) + \frac{2\gamma}{u} \exp\left(-\frac{1}{8} nup_{min,i}\right).$$

Précision pour l'estimation de déciles

Théorème (Ecart avec les déciles théoriques)

Soit $\gamma \in \mathbb{R}_+^\star$, $u \in [0, \gamma/4]$, $\rho \in \mathbb{R}_+$ et $X \in [0,R]^n$ dont les éléments sont obtenues à partir d'une loi P de densité π_P continue au voisinage de ses déciles. On pose $p_{min,i} = \inf_{t \in [d_i^l - 2\gamma, d_i^l + 2\gamma]} \pi_P(t)$. On note $(d_i^l)_i$ les déciles de la loi. Notons alors enfin $\mathcal{M}_{cont}$ le mécanisme de sensibilité inverse ρ-lisse. On a

$$\mathbb{P}\left(|\mathcal{M}_{cont,i} - d_i^l| > 2u + \rho\right) \leq \frac{R}{2\rho} \exp\left(-nup_{min,i}\varepsilon/4\right) + \frac{2\gamma}{u} \exp\left(-\frac{1}{8} nup_{min,i}\right).$$

Théorème

Soit $\rho \in]0, 1 - 4/\sqrt{n}]$ et $X \in [0,1]^n$ dont les éléments sont obtenues à partir de la loi uniforme standard. On note $(d_i^l)_i$ les déciles de la loi. Notons alors enfin $\mathcal{M}_{cont}$ le mécanisme de sensibilité inverse ρ-lisse.

$$\mathbb{E}(|\mathcal{M}_{cont,i} - d_i^l|) \leq \rho + \frac{4}{\sqrt{n}} + \frac{4}{n\varepsilon\rho} \exp\left(-\frac{\sqrt{n}\varepsilon}{2}\right) + \frac{16}{\sqrt{n}} \exp\left(-\frac{\sqrt{n}}{4}\right).$$

Comparaison des bornes obtenues

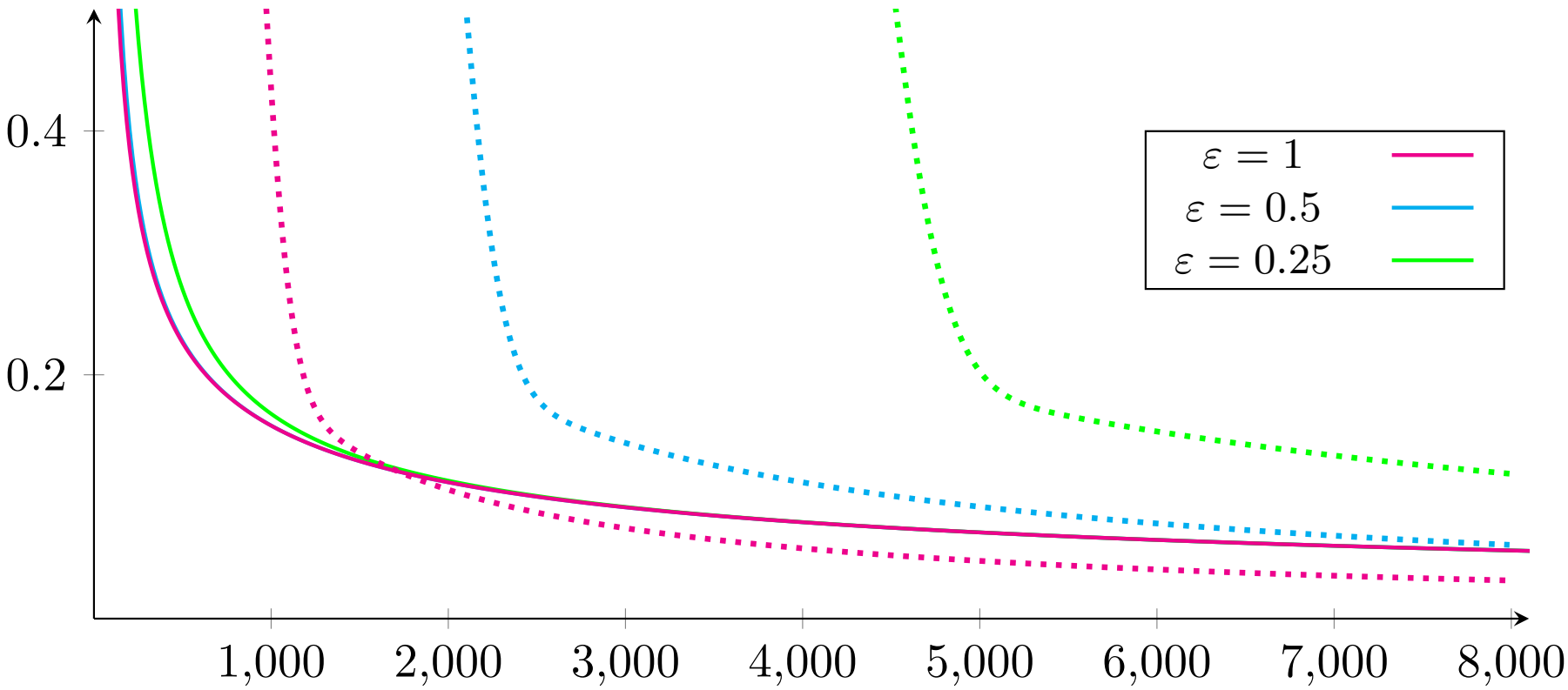


Figure – Graphe des bornes obtenues [21, 8] en fonction de n pour $\rho = 1/\sqrt{n}$. Le mécanisme exponentiel est en ligne continues et la méthode des histogrammes est en lignes pointillées.