Towards Turn-Key Differential Privacy

Adventures in Function Approximation, Empirical Process Theory and Open-Source Software

Ben Rubinstein

joint with Francesco Aldà

July 25, 2017

School of Computing & Information Systems
The University of Melbourne

One More Time With Feeling: Why Protect Privacy?

WIRED SUBSCRIBE » SECTIONS » BLOGS » REVIEWS » VIDEO » HOW-TO

Sign In | RSS Feeds

THREAT LEVEL

PRIVACY, CRIME AND SECURITY ONLINE

NetFlix Cancels Recommendation Contest After Privacy Lawsuit

By Ryan Singel | March 12, 2010 | 2:48 pm | Categories: privacy

Netflix is canceling its second $1 million Netflix Prize to settle a legal challenge that it breached customer privacy as part of the first contest's race for a better movie-recommendation engine.

Friday's announcement came five months after Netflix had announced a successor to its algorithm-improvement contest. The company at the time said it intended to expand the amount of information it gave to researchers in hopes that its recommendation system — a key part of Netflix's customer retention strategy — would get even better. That was then followed with a

https://www.wired.com/2010/03/netflix-cancels-contest/

ABC NEWS LOCATION: Melbourne, Vic Change

Just In Australia World Business Sport Science Arts Analysis Fact Che

Print Email Facebook Twitter More

Medicare dataset pulled after academics find breach of doctor details possible

By political reporter Stephanie Anderson

Updated 29 Sep 2016, 2:51pm

The Health Department has removed Medicare data from its website amid an investigation into whether personal information has been compromised.

Australian Privacy Commissioner Timothy Pilgrim has launched an investigation after academics found it was possible to figure out some service provider ID numbers in the Medicare Benefits Schedule and Pharmaceutical Benefits Schedule datasets, published on August 1.

The University of Melbourne academics said they notified the department of the issue on September 12, adding that the data was then "immediately removed".

In a joint report, Drs Chris Culnane, Benjamin Rubinstein and Vanessa Teague described the

PHOTO: The Health Department says no patient information has been compromised. (ABC News)

RELATED STORY: Yahoo breach puts focus on Australian consumer hacking protections

RELATED STORY: Thousands of Australian computer log-ins up for sale on dark web

http://www.abc.net.au/news/2016-09-29/medicare-pbs-dataset-pulled-over-encryption-concerns/7888686

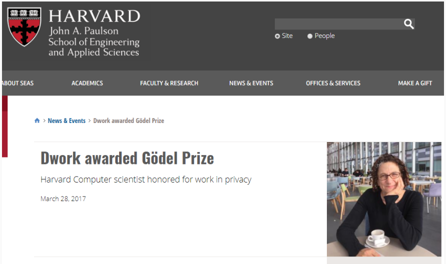
Regulatory & ethical obligations; customer confidence; . . .profits!!

DP Successes (If Privacy Doesn't Inspire You)

Recent deployments

- Google: RAPPOR, Google Chrome
- Apple: iOS 10.x
- Uber: SQL Elastic Sensitivity
- U.S. Census Bureau: OnTheMap
- Transport for NSW: Opal Data Release
- etc.

Active world-leading groups: Harvard, Stanford, Berkeley, CMU, Weizmann, UCL, Oxford, USC, UCSD, UPenn, Caltech, Cornell, Duke, Disney Research, Google Research, Microsoft Research, etc.



Talk Outline

1. Intro to differential privacy
2. The Bernstein mechanism: Private function release
3. The sensitivity sampler: Automating privatisation
4. The `diffpriv` package

Introduction to Differential Privacy

What's DP For?

Release aggregate information on a dataset, but protect individuals.

What's DP For?

Release aggregate information on a dataset, but protect individuals.

Parties: Trusted data curator; **Untrusted receipient**

Variations exist *e.g.*, decentralised curator

Example **target analyses** to privatise

- A function of data: A statistic!
- Probabilistic model fitting with MLE: Estimation procedure
- Deep neural network training: A learner
- KD tree construction: Spatial data analysis

What's DP For?

Release aggregate information on a dataset, but protect individuals.

Parties: Trusted data curator; **Untrusted receipient**

Variations exist *e.g.*, decentralised curator

Example **target analyses** to privatise

- A function of data: A statistic!
- Probabilistic model fitting with MLE: Estimation procedure
- Deep neural network training: A learner
- KD tree construction: Spatial data analysis

In general, privacy/utility must be in tension. *Lower bounds later.*

Records, Databases, Target Functions, Mechanisms

A database D is a sequence of n records from domain set $\mathcal{D}$.

A target function for privatisation $f : \mathcal{D}^n \to \mathcal{B}$ a response set

Example: Sample Mean

Consider releasing the average of scalars, *e.g.*, test scores

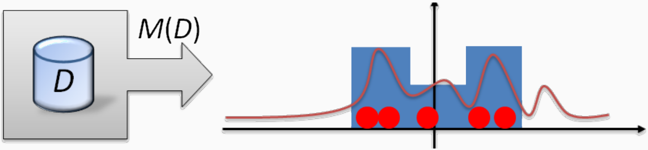
$\mathcal{D} = \mathcal{B} = \mathbb{R}$ and $f(D) = \frac{1}{n}\sum_{i=1}^{n} D_i$

```
> D <- rnorm(1000) # 1000 standard normal samples
> f <- mean
> f(D)
[1] 0.03339015
```

Records, Databases, Target Functions, Mechanisms (cont.)

A mechanism $\mathcal{M}$ maps D to a random response in $\mathcal{B}$.

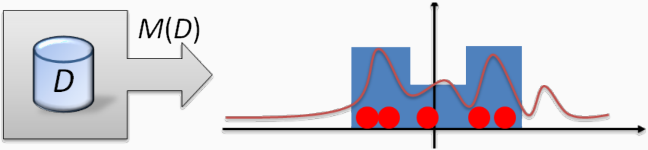
Response distribution: $\Pr(\mathcal{M}(D) \in B)$ for $B \subset \mathcal{B}$.



Records, Databases, Target Functions, Mechanisms (cont.)

A mechanism $\mathcal{M}$ maps D to a random response in $\mathcal{B}$.

Response distribution: $\Pr(\mathcal{M}(D) \in B)$ for $B \subset \mathcal{B}$.



Example: Blood Type

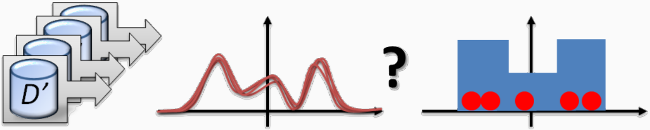
Everyone in D have same blood type? $f(D) = 1[D_1 = \ldots = D_n]$.

$\mathcal{M}(D) \sim Bernoulli(0.5)$ $\quad$ $\mathcal{M}(D) = \begin{cases} f(D)\ , & w.p.\ 0.9\ , \\ 1 - f(D) & w.p.\ 0.1 \end{cases}$

Utility measures (high probability) proximity of $\mathcal{M}(D), f(D)$

Defining Differential Privacy

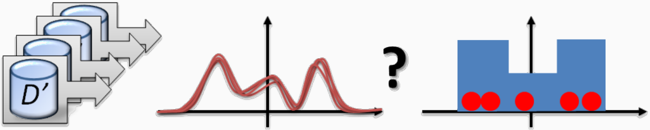
Intuition: Response indistinguishable on changing any one record



Databases D, D' are called neighbouring if they differ on one record

Defining Differential Privacy

Intuition: Response indistinguishable on changing any one record



Databases D, D' are called neighbouring if they differ on one record

$\mathcal{M}$ is ϵ-Differentially Private

If for all neighbouring $D, D' \in \mathcal{D}^n$, for all $B \subset \mathcal{B}$, we have that $\Pr(\mathcal{M}(D) \in B) \leq \exp(\epsilon) \cdot \Pr(\mathcal{M}(D') \in B)$. Where $\epsilon > 0$.

That is $\log\left(\frac{\Pr(\mathcal{M}(D)\in B)}{\Pr(\mathcal{M}(D')\in B)}\right) \leq \epsilon$: Smaller $\epsilon > 0$, more privacy.

Semantic privacy with strong threat model; *worst-case on DBs.*

Example: Numeric Releases with the Laplace Mechanism

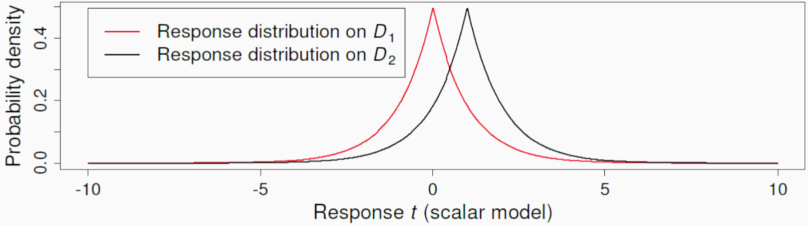
Consider target $f : \mathcal{D} \to \mathbb{R}^d$

e.g., a covariance matrix, regression coefficients, classifier weights

Smooth the target by adding zero-mean Laplace noise to output.

Laplace Mechanism

Given parameters $\Delta, \epsilon > 0$, release $\mathcal{M}(D) \sim f(D) + Lap(\Delta/\epsilon)$.



Example: Hello World – Sample Mean of $D_i \in [0, 1]$

Global Sensitivity

Many generic mechanisms like Laplace operate by smoothing f.
Less smoothing needed for already-smooth f; How to measure?

Consider target $f : \mathcal{D} \to \mathcal{B}$ with normed response space $\mathcal{B}$.

Global Sensitivity

Many generic mechanisms like Laplace operate by smoothing f.
Less smoothing needed for already-smooth f; How to measure?

Consider target $f : \mathcal{D} \to \mathcal{B}$ with normed response space $\mathcal{B}$.

Global sensitivity

$\Delta(f) = \max_{D,D'} \|f(D) - f(D')\|_{\mathcal{B}}$ over neighbouring DBs in $\mathcal{D}^n$.

A type of Lipschitz condition. (Weakest form of smoothness.)

Example: Sample Mean

Take $f(D) = \frac{1}{n}\sum_{i=1}^{n} D_i$ in $\mathcal{B} = \mathbb{R}$, with absolute as norm.
If $D_i \in [0, 1]$ then $\Delta(f) = 1/n$.

Privacy of the Laplace Mechanism

Recall

- $\Delta(f) = \max_{D,D'} \|f(D) - f(D')\|_{\mathcal{B}}$ over neighbouring DBs.
- $\mathcal{M}(D) \sim f(D) + Lap(\Delta/\epsilon)$.

Theorem: Laplace Mechanism Privacy

If Δ is L_1-gobal sensitivity of f, then $\mathcal{M}$ is ϵ-DP.

Why L_1? *multivariate Laplace has density exponential in L_1.*

Privacy of the Laplace Mechanism

Recall

- $\Delta(f) = \max_{D,D'} \|f(D) - f(D')\|_{\mathcal{B}}$ over neighbouring DBs.
- $\mathcal{M}(D) \sim f(D) + Lap(\Delta/\epsilon)$.

Theorem: Laplace Mechanism Privacy

If Δ is L_1-gobal sensitivity of f, then $\mathcal{M}$ is ϵ-DP.

Why L_1? *multivariate Laplace has density exponential in L_1.*

More privacy (smaller ϵ), the more noise needed, lower utility.
The smoother the target (low Δ), the less smoothing needed.

Notes

- Generic mechanisms like Laplace have driven DP's ascent
- Another driver: A calculus of composition
- Many applications explored in telecom, health, web, etc.
- Utility bounds exist for simpler mechanisms: Guide choices
- Empirical investigations: some mechanisms work, some don't
- Lower bounds illustrate impossibility results

The Bernstein Mechanism:
Private Function Release – AAAI'17

Bernstein vs. Laplace Mechanisms

Problem: What about releasing a function? A trained classifier?

	Laplace Mechanism	**Bernstein Mechanism**
Operation		
Response space $\mathcal{B}$	$\mathbb{R}^d$	functions: $[0,1]^d \to \mathbb{R}$
Perturbation	output	output
Privacy		
Requires access to	$f(D)$, $\Delta(f)$	$f(D)$, $\Delta(f)$
Sensitivity norm	L_1	L_1 of $f(\cdot)$ evaluated on lattice
Privacy guarantee	ϵ-DP	ϵ-DP
Utility		
Conditions	-	Smooth $f(\cdot)$

Bernstein Mechanism: Sketch

Goal: Privately release function g returned by $f : \mathcal{D}^n \to \mathbb{R}^{[0,1]^d}$

Parameters: degree k, sensitivity Δ, privacy $\epsilon > 0$

Bernstein Mechanism: Sketch

Goal: Privately release function g returned by $f : \mathcal{D}^n \rightarrow \mathbb{R}^{[0,1]^d}$
Parameters: degree k, sensitivity Δ, privacy $\epsilon > 0$

1. Function $g \longleftarrow$ Evaluate $f(D)$

Bernstein Mechanism: Sketch

Goal: Privately release function g returned by $f : \mathcal{D}^n \to \mathbb{R}^{[0,1]^d}$
Parameters: degree k, sensitivity Δ, privacy $\epsilon > 0$

1. Function $g \longleftarrow$ Evaluate $f(D)$
2. Coefficients $\mathbf{c} \longleftarrow$ Approximate g on a grid over $[0,1]^d$

Bernstein Mechanism: Sketch

Goal: Privately release function g returned by $f : \mathcal{D}^n \to \mathbb{R}^{[0,1]^d}$
Parameters: degree k, sensitivity Δ, privacy $\epsilon > 0$

1. Function $g \longleftarrow$ Evaluate $f(D)$
2. Coefficients $\mathbf{c} \longleftarrow$ Approximate g on a grid over $[0,1]^d$
3. Coefficients $\tilde{\mathbf{c}} \longleftarrow$ perturb $\mathbf{c}$ by Laplace mechanism

Bernstein Mechanism: Sketch

Goal: Privately release function g returned by $f : \mathcal{D}^n \rightarrow \mathbb{R}^{[0,1]^d}$
Parameters: degree k, sensitivity Δ, privacy $\epsilon > 0$

1. Function $g \longleftarrow$ Evaluate $f(D)$
2. Coefficients $\mathbf{c} \longleftarrow$ Approximate g on a grid over $[0, 1]^d$
3. Coefficients $\tilde{\mathbf{c}} \longleftarrow$ perturb $\mathbf{c}$ by Laplace mechanism
4. Release coefficients $\tilde{\mathbf{c}}$

Bernstein Mechanism: Sketch

Goal: Privately release function g returned by $f : \mathcal{D}^n \to \mathbb{R}^{[0,1]^d}$
Parameters: degree k, sensitivity Δ, privacy $\epsilon > 0$

1. Function $g \longleftarrow$ Evaluate $f(D)$
2. Coefficients $\mathbf{c} \longleftarrow$ Approximate g on a grid over $[0,1]^d$
3. Coefficients $\tilde{\mathbf{c}} \longleftarrow$ perturb $\mathbf{c}$ by Laplace mechanism
4. Release coefficients $\tilde{\mathbf{c}}$

Reconstruct release function

4. $\tilde{g} \longleftarrow$ perturbed coefficients $\tilde{\mathbf{c}}$, dot, public basis functions

Aside: Bernstein Function Approximation

Goal: Approximate $g : [0, 1] \to \mathbb{R}$ by smooth polynomial

Aside: Bernstein Function Approximation

Goal: Approximate $g : [0, 1] \to \mathbb{R}$ by smooth polynomial

Degree-k basis $b_{\nu,k}(x) = \binom{k}{\nu} x^{\nu}(1 - x)^{k-\nu}$ for $\nu \in \{0, \dots, k\}$

Aside: Bernstein Function Approximation

Goal: Approximate $g : [0, 1] \to \mathbb{R}$ by smooth polynomial

Degree-k basis $b_{\nu,k}(x) = \binom{k}{\nu} x^{\nu}(1 - x)^{k-\nu}$ for $\nu \in \{0, \ldots, k\}$

Coefficients $\mathbf{c}$: evaluations on grid $g(0/k), g(1/k), \ldots, g(k/k)$

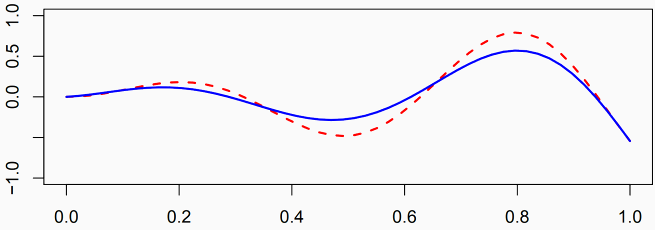
Aside: Bernstein Function Approximation

Goal: Approximate $g : [0,1] \to \mathbb{R}$ by smooth polynomial

Degree-k basis $b_{\nu,k}(x) = \binom{k}{\nu} x^{\nu}(1-x)^{k-\nu}$ for $\nu \in \{0, \ldots, k\}$

Coefficients **c**: evaluations on grid $g(0/k), g(1/k), \ldots, g(k/k)$

Bernstein operator: $g(x) \approx \sum_{\nu=0}^{k} g(\nu/k) b_{\nu,k}(x)$



$k = 50$

Bernstein Utility

Utility: $\leq \alpha$ error whp $\geq 1 - \beta$

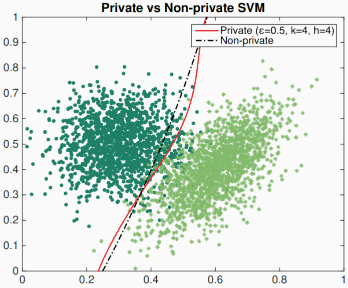
1. $(2h, T)$-smooth target:

   $$\alpha = O\left(\frac{\Delta}{\epsilon} \log \frac{1}{\beta}\right)^{\frac{h}{d+h}}$$

2. (γ, L)-Hölder continuous:

   $$\alpha = O\left(\frac{\Delta}{\epsilon} \log \frac{1}{\beta}\right)^{\frac{\gamma}{2d+\gamma}}$$

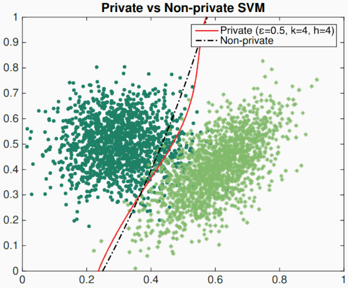
3. Linear target: $\alpha = O\left(\frac{\Delta}{\epsilon} \log \frac{1}{\beta}\right)$



Bernstein Utility

Utility: $\leq \alpha$ error whp $\geq 1-\beta$

1. $(2h, T)$-smooth target:
   $\alpha = O\left(\frac{\Delta}{\epsilon}\log\frac{1}{\beta}\right)^{\frac{h}{d+h}}$
2. (γ, L)-Hölder continuous:
   $\alpha = O\left(\frac{\Delta}{\epsilon}\log\frac{1}{\beta}\right)^{\frac{\gamma}{2d+\gamma}}$
3. Linear target: $\alpha = O\left(\frac{\Delta}{\epsilon}\log\frac{1}{\beta}\right)$



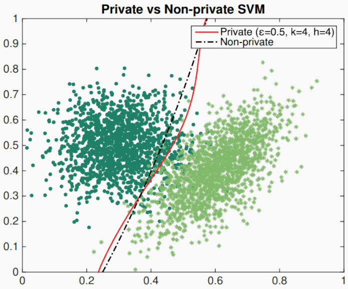
Proschan'65: Concentration of convex comb of iid log-concave rv

Weierstrass Theorem: uniform approximation

Bernstein Utility

Utility: $\leq \alpha$ error whp $\geq 1-\beta$

1. $(2h, T)$-smooth target:
   $$\alpha = O\left(\frac{\Delta}{\epsilon}\log\frac{1}{\beta}\right)^{\frac{h}{d+h}}$$
2. (γ, L)-Hölder continuous:
   $$\alpha = O\left(\frac{\Delta}{\epsilon}\log\frac{1}{\beta}\right)^{\frac{\gamma}{2d+\gamma}}$$
3. Linear target: $\alpha = O\left(\frac{\Delta}{\epsilon}\log\frac{1}{\beta}\right)$



Proschan'65: Concentration of convex comb of iid log-concave rv

Weierstrass Theorem: uniform approximation

Lower bound: There exists a target s.t. all ϵ-DP mechanisms introduce $\geq \Omega(\Delta/\epsilon)$ error with probability going to 1

The Sensitivity Sampler: Automating Privatisation – ICML'17

“Just bound sensitivity” he said, “It will be great” he said.

"Just bound sensitivity" he said, "It will be great" he said.

Bound sensitivity for releasing SVM classifier (Rubinstein et al. 12)

75

for all $\mathbf{x} \in \mathbb{R}^d$. *For each database* $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, *define*

$$\mathbf{w}_S \in \arg\min_{\mathbf{w}\in\mathbb{R}^F} \frac{C}{n}\sum_{i=1}^n \ell\left(y_i, f_{\mathbf{w}}(\mathbf{x}_i)\right) + \frac{1}{2}\|\mathbf{w}\|_2^2 .$$

Then for every pair of neighboring databases D, D' *of* n *entries, we have* $\|\mathbf{w}_D - \mathbf{w}_{D'}\|_2 \le 4LC\kappa/n$, *and* $\|\mathbf{w}_D - \mathbf{w}_{D'}\|_1 \le 4LC\kappa\sqrt{F}/n$.

Proof. The argument closely follows the proof of the SVM's uniform stability (Schölkopf and Smola, 2001, Theorem 12.4). For convenience we define for any training set S

$$\begin{aligned} R_{\text{reg}}(\mathbf{w}, S) &= \frac{C}{n}\sum_{i=1}^n \ell\left(y_i, f_{\mathbf{w}}(\mathbf{x}_i)\right) + \frac{1}{2}\|\mathbf{w}\|_2^2 \\ R_{\text{emp}}(\mathbf{w}, S) &= \frac{1}{n}\sum_{i=1}^n \ell(y_i, f_{\mathbf{w}}(\mathbf{x}_i)). \end{aligned}$$

Then the first-order necessary KKT conditions imply

$$\mathbf{0} \in \partial_{\mathbf{w}} R_{\text{reg}}(\mathbf{w}_D, D) = C\partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_D, D) + \mathbf{w}_D\,, \tag{2}$$

$$\mathbf{0} \in \partial_{\mathbf{w}} R_{\text{reg}}(\mathbf{w}_{D'}, D') = C\partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_{D'}, D') + \mathbf{w}_{D'} \tag{3}$$

where $\partial_{\mathbf{w}}$ is the subdifferential operator wrt $\mathbf{w}$. Define the auxiliary risk function

$$\tilde{R}(\mathbf{w}) = C\left\langle \partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_D, D) - \partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_{D'}, D'), \mathbf{w} - \mathbf{w}_{D'} \right\rangle + \frac{1}{2}\|\mathbf{w} - \mathbf{w}_{D'}\|_2^2 .$$

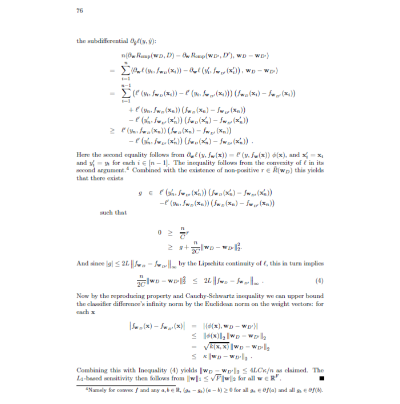
Note that $\tilde{R}(\cdot)$ maps to sets of reals. It is easy to see that $\tilde{R}(\mathbf{w})$ is strictly convex in $\mathbf{w}$. Substituting $\mathbf{w}_{D'}$ into $\tilde{R}(\mathbf{w})$ yields

$$\begin{aligned} \tilde{R}(\mathbf{w}_{D'}) &= C\left\langle \partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_D, D) - \partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_{D'}, D'), 0 \right\rangle + \frac{1}{2}\|0\|_2^2 \\ &= \{0\}. \end{aligned}$$

And by Equation (3)

$$\begin{aligned} C\partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_D, D) + \mathbf{w} &\in C\partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_D, D) - C\partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_{D'}, D') + \mathbf{w} - \mathbf{w}_{D'} \\ &= \partial_{\mathbf{w}} \tilde{R}(\mathbf{w})\,, \end{aligned}$$

which combined with Equation (2) implies $\mathbf{0} \in \partial_{\mathbf{w}}\tilde{R}(\mathbf{w}_D)$, so that $\tilde{R}(\mathbf{w})$ is minimized at $\mathbf{w}_D$. Thus there exists some non-positive $r \in \tilde{R}(\mathbf{w}_D)$. Next simplify the first term of $\tilde{R}(\mathbf{w}_D)$, scaled by n/C for notational convenience. In what follows we denote by $\ell'(y, \hat{y})$



76

the subdifferential $\partial_{\hat{y}}\ell(y, \hat{y})$:

$$\begin{aligned} & n\left\langle \partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_D, D) - \partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_{D'}, D'), \mathbf{w}_D - \mathbf{w}_{D'} \right\rangle \\ = & \sum_{i=1}^n \left\langle \partial_{\mathbf{w}}\ell\left(y_i, f_{\mathbf{w}_D}(\mathbf{x}_i)\right) - \partial_{\mathbf{w}}\ell\left(y_i', f_{\mathbf{w}_{D'}}(\mathbf{x}_i')\right), \mathbf{w}_D - \mathbf{w}_{D'} \right\rangle \\ = & \sum_{i=1}^{n-1} \left(\ell'\left(y_i, f_{\mathbf{w}_D}(\mathbf{x}_i)\right) - \ell'\left(y_i, f_{\mathbf{w}_{D'}}(\mathbf{x}_i)\right)\right)\left(f_{\mathbf{w}_D}(\mathbf{x}_i) - f_{\mathbf{w}_{D'}}(\mathbf{x}_i)\right) \\ & \quad + \ell'\left(y_n, f_{\mathbf{w}_D}(\mathbf{x}_n)\right)\left(f_{\mathbf{w}_D}(\mathbf{x}_n) - f_{\mathbf{w}_{D'}}(\mathbf{x}_n)\right) \\ & \quad - \ell'\left(y_n', f_{\mathbf{w}_{D'}}(\mathbf{x}_n')\right)\left(f_{\mathbf{w}_D}(\mathbf{x}_n') - f_{\mathbf{w}_{D'}}(\mathbf{x}_n')\right) \\ \ge & \; \ell'\left(y_n, f_{\mathbf{w}_D}(\mathbf{x}_n)\right)\left(f_{\mathbf{w}_D}(\mathbf{x}_n) - f_{\mathbf{w}_{D'}}(\mathbf{x}_n)\right) \\ & \quad - \ell'\left(y_n', f_{\mathbf{w}_{D'}}(\mathbf{x}_n')\right)\left(f_{\mathbf{w}_D}(\mathbf{x}_n') - f_{\mathbf{w}_{D'}}(\mathbf{x}_n')\right) . \end{aligned}$$

Here the second equality follows from $\partial_{\mathbf{w}}\ell\left(y, f_{\mathbf{w}}(\mathbf{x})\right) = \ell'\left(y, f_{\mathbf{w}}(\mathbf{x})\right)\phi(\mathbf{x})$, and $\mathbf{x}_i' = \mathbf{x}_i$ and $y_i' = y_i$ for each $i \in [n-1]$. The inequality follows from the convexity of ℓ in its second argument.[4] Combined with the existence of non-positive $r \in \tilde{R}(\mathbf{w}_D)$ this yields that there exists

$$\begin{aligned} g \in & \; \ell'\left(y_n', f_{\mathbf{w}_{D'}}(\mathbf{x}_n')\right)\left(f_{\mathbf{w}_D}(\mathbf{x}_n') - f_{\mathbf{w}_{D'}}(\mathbf{x}_n')\right) \\ & -\ell'\left(y_n, f_{\mathbf{w}_D}(\mathbf{x}_n)\right)\left(f_{\mathbf{w}_D}(\mathbf{x}_n) - f_{\mathbf{w}_{D'}}(\mathbf{x}_n)\right) \end{aligned}$$

such that

$$\begin{aligned} 0 &\ge \frac{n}{C} r \\ &\ge g + \frac{n}{2C}\|\mathbf{w}_D - \mathbf{w}_{D'}\|_2^2 . \end{aligned}$$

And since $|g| \le 2L\left\|f_{\mathbf{w}_D} - f_{\mathbf{w}_{D'}}\right\|_\infty$ by the Lipschitz continuity of ℓ, this in turn implies

$$\frac{n}{2C}\|\mathbf{w}_D - \mathbf{w}_{D'}\|_2^2 \le 2L\left\|f_{\mathbf{w}_D} - f_{\mathbf{w}_{D'}}\right\|_\infty . \tag{4}$$

Now by the reproducing property and Cauchy-Schwartz inequality we can upper bound the classifier difference's infinity norm by the Euclidean norm on the weight vectors: for each $\mathbf{x}$

$$\begin{aligned} \left|f_{\mathbf{w}_D}(\mathbf{x}) - f_{\mathbf{w}_{D'}}(\mathbf{x})\right| &= \left|\left\langle \phi(\mathbf{x}), \mathbf{w}_D - \mathbf{w}_{D'} \right\rangle\right| \\ &\le \|\phi(\mathbf{x})\|_2 \|\mathbf{w}_D - \mathbf{w}_{D'}\|_2 \\ &= \sqrt{k(\mathbf{x}, \mathbf{x})}\,\|\mathbf{w}_D - \mathbf{w}_{D'}\|_2 \\ &\le \kappa \|\mathbf{w}_D - \mathbf{w}_{D'}\|_2 . \end{aligned}$$

Combining this with Inequality (4) yields $\|\mathbf{w}_D - \mathbf{w}_{D'}\|_2 \le 4LC\kappa/n$ as claimed. The L_1-based sensitivity then follows from $\|\mathbf{w}\|_1 \le \sqrt{F}\|\mathbf{w}\|_2$ for all $\mathbf{w} \in \mathbb{R}^F$. ■

[4] Namely for convex f and any $a, b \in \mathbb{R}$, $(g_a - g_b)(a - b) \ge 0$ for all $g_a \in \partial f(a)$ and all $g_b \in \partial f(b)$.

“Just bound sensitivity” he said, “It will be great” he said.

Bound sensitivity for releasing SVM classifier (Rubinstein et al. 12)

75

for all $\mathbf{x} \in \mathbb{R}^d$*. For each database* $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$*, define*

$$\mathbf{w}_S \in \arg\min_{\mathbf{w}\in\mathbb{R}^F} \frac{C}{n}\sum_{i=1}^n \ell\left(y_i, f_{\mathbf{w}}(\mathbf{x}_i)\right) + \frac{1}{2}\|\mathbf{w}\|_2^2 .$$

Then for every pair of neighboring databases D, D' *of* n *entries, we have* $\|\mathbf{w}_D - \mathbf{w}_{D'}\|_2 \le 4LC\kappa/n$*, and* $\|\mathbf{w}_D - \mathbf{w}_{D'}\|_1 \le 4LC\kappa\sqrt{F}/n$*.*

Proof. The argument closely follows the proof of the SVM's uniform stability (Schölkopf and Smola, 2001, Theorem 12.4). For convenience we define for any training set S

$$\begin{aligned} R_{\text{reg}}(\mathbf{w}, S) &= \frac{C}{n}\sum_{i=1}^n \ell\left(y_i, f_{\mathbf{w}}(\mathbf{x}_i)\right) + \frac{1}{2}\|\mathbf{w}\|_2^2 \\ R_{\text{emp}}(\mathbf{w}, S) &= \frac{1}{n}\sum_{i=1}^n \ell(y_i, f_{\mathbf{w}}(\mathbf{x}_i)). \end{aligned}$$

Then the first-order necessary KKT conditions imply

$$\mathbf{0} \in \partial_{\mathbf{w}} R_{\text{reg}}(\mathbf{w}_D, D) = C\partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_D, D) + \mathbf{w}_D\,, \quad (2)$$

$$\mathbf{0} \in \partial_{\mathbf{w}} R_{\text{reg}}(\mathbf{w}_{D'}, D') = C\partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_{D'}, D') + \mathbf{w}_{D'} \quad (3)$$

where $\partial_{\mathbf{w}}$ is the subdifferential operator wrt $\mathbf{w}$. Define the auxiliary risk function

$$\tilde{R}(\mathbf{w}) = C\langle \partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_D, D) - \partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_{D'}, D'), \mathbf{w} - \mathbf{w}_{D'}\rangle + \frac{1}{2}\|\mathbf{w} - \mathbf{w}_{D'}\|_2^2 .$$

Note that $\tilde{R}(\cdot)$ maps to sets of reals. It is easy to see that $\tilde{R}(\mathbf{w})$ is strictly convex in $\mathbf{w}$. Substituting $\mathbf{w}_{D'}$ into $\tilde{R}(\mathbf{w})$ yields

$$\begin{aligned} \tilde{R}(\mathbf{w}_{D'}) &= C\left\langle \partial_{\mathbf{w}} R_{\text{emp}}\left(\mathbf{w}_D, D\right) - \partial_{\mathbf{w}} R_{\text{emp}}\left(\mathbf{w}_{D'}, D'\right), 0\right\rangle + \frac{1}{2}\|0\|_2^2 \\ &= \{0\}. \end{aligned}$$

And by Equation (3)

$$\begin{aligned} C\partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_D, D) + \mathbf{w} &\in C\partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_D, D) - C\partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_{D'}, D') + \mathbf{w} - \mathbf{w}_{D'} \\ &= \partial_{\mathbf{w}} \tilde{R}(\mathbf{w})\,, \end{aligned}$$

which combined with Equation (2) implies $\mathbf{0} \in \partial_{\mathbf{w}} \tilde{R}(\mathbf{w}_D)$, so that $\tilde{R}(\mathbf{w})$ is minimized at $\mathbf{w}_D$. Thus there exists some non-positive $r \in \tilde{R}(\mathbf{w}_D)$. Next simplify the first term of $\tilde{R}(\mathbf{w}_D)$, scaled by n/C for notational convenience. In what follows we denote by $\ell'(y, \hat{y})$

76

the subdifferential $\partial_{\hat{y}}\ell(y, \hat{y})$:

$$\begin{aligned} & n\langle \partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_D, D) - \partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}_{D'}, D'), \mathbf{w}_D - \mathbf{w}_{D'}\rangle \\ = & \sum_{i=1}^n \left\langle \partial_{\mathbf{w}}\ell\left(y_i, f_{\mathbf{w}_D}(\mathbf{x}_i)\right) - \partial_{\mathbf{w}}\ell\left(y_i', f_{\mathbf{w}_{D'}}(\mathbf{x}_i')\right), \mathbf{w}_D - \mathbf{w}_{D'}\right\rangle \\ = & \sum_{i=1}^{n-1} \left(\ell'\left(y_i, f_{\mathbf{w}_D}(\mathbf{x}_i)\right) - \ell'\left(y_i, f_{\mathbf{w}_{D'}}(\mathbf{x}_i)\right)\right)\left(f_{\mathbf{w}_D}(\mathbf{x}_i) - f_{\mathbf{w}_{D'}}(\mathbf{x}_i)\right) \\ & + \ell'\left(y_n, f_{\mathbf{w}_D}(\mathbf{x}_n)\right)\left(f_{\mathbf{w}_D}(\mathbf{x}_n) - f_{\mathbf{w}_{D'}}(\mathbf{x}_n)\right) \\ & - \ell'\left(y_n', f_{\mathbf{w}_{D'}}(\mathbf{x}_n')\right)\left(f_{\mathbf{w}_D}(\mathbf{x}_n') - f_{\mathbf{w}_{D'}}(\mathbf{x}_n')\right) \\ \ge & \ \ell'\left(y_n, f_{\mathbf{w}_D}(\mathbf{x}_n)\right)\left(f_{\mathbf{w}_D}(\mathbf{x}_n) - f_{\mathbf{w}_{D'}}(\mathbf{x}_n)\right) \\ & - \ell'\left(y_n', f_{\mathbf{w}_{D'}}(\mathbf{x}_n')\right)\left(f_{\mathbf{w}_D}(\mathbf{x}_n') - f_{\mathbf{w}_{D'}}(\mathbf{x}_n')\right)\,. \end{aligned}$$

Here the second equality follows from $\partial_{\mathbf{w}}\ell\left(y, f_{\mathbf{w}}(\mathbf{x})\right) = \ell'\left(y, f_{\mathbf{w}}(\mathbf{x})\right)\phi(\mathbf{x})$, and $\mathbf{x}_i' = \mathbf{x}_i$ and $y_i' = y_i$ for each $i \in [n-1]$. The inequality follows from the convexity of ℓ in its second argument.[4] Combined with the existence of non-positive $r \in \tilde{R}(\mathbf{w}_D)$ this yields that there exists

$$\begin{aligned} g \in & \ \ell'\left(y_n', f_{\mathbf{w}_{D'}}(\mathbf{x}_n')\right)\left(f_{\mathbf{w}_D}(\mathbf{x}_n') - f_{\mathbf{w}_{D'}}(\mathbf{x}_n')\right) \\ & -\ell'\left(y_n, f_{\mathbf{w}_D}(\mathbf{x}_n)\right)\left(f_{\mathbf{w}_D}(\mathbf{x}_n) - f_{\mathbf{w}_{D'}}(\mathbf{x}_n)\right) \end{aligned}$$

such that

$$\begin{aligned} 0 &\ge \frac{n}{C}r \\ &\ge g + \frac{n}{2C}\|\mathbf{w}_D - \mathbf{w}_{D'}\|_2^2 . \end{aligned}$$

And since $|g| \le 2L\left\|f_{\mathbf{w}_D} - f_{\mathbf{w}_{D'}}\right\|_\infty$ by the Lipschitz continuity of ℓ, this in turn implies

$$\frac{n}{2C}\|\mathbf{w}_D - \mathbf{w}_{D'}\|_2^2 \le 2L\left\|f_{\mathbf{w}_D} - f_{\mathbf{w}_{D'}}\right\|_\infty . \quad (4)$$

Now by the reproducing property and Cauchy-Schwartz inequality we can upper bound the classifier difference's infinity norm by the Euclidean norm on the weight vectors: for each $\mathbf{x}$

$$\begin{aligned} \left|f_{\mathbf{w}_D}(\mathbf{x}) - f_{\mathbf{w}_{D'}}(\mathbf{x})\right| &= |\langle \phi(\mathbf{x}), \mathbf{w}_D - \mathbf{w}_{D'}\rangle| \\ &\le \|\phi(\mathbf{x})\|_2 \|\mathbf{w}_D - \mathbf{w}_{D'}\|_2 \\ &= \sqrt{k(\mathbf{x}, \mathbf{x})}\,\|\mathbf{w}_D - \mathbf{w}_{D'}\|_2 \\ &\le \kappa\,\|\mathbf{w}_D - \mathbf{w}_{D'}\|_2 \,. \end{aligned}$$

Combining this with Inequality (4) yields $\|\mathbf{w}_D - \mathbf{w}_{D'}\|_2 \le 4LC\kappa/n$ as claimed. The L_1-based sensitivity then follows from $\|\mathbf{w}\|_1 \le \sqrt{F}\|\mathbf{w}\|_2$ for all $\mathbf{w} \in \mathbb{R}^F$. ■

[4] Namely for convex f and any $a, b \in \mathbb{R}$, $(g_a - g_b)(a - b) \ge 0$ for all $g_a \in \partial f(a)$ and all $g_b \in \partial f(b)$.

Simple? Subdifferentials, algorithmic stability, convex auxiliary risk

“Laws of Mathematics are Very Commendable but...”

“Laws of Mathematics are Very Commendable but...”

Apply generic mechanisms without bounding sensitivity?

Existing work: Restrict targets until sensitivity can be ‘composed’ *e.g.*, recent Uber/Berkeley Elastic Sensitivity system.

This work: Permit *any* target, but won’t bound target sensitivity over all DB pairs. Instead sensitivity over all reasonable DBs.

Key ideas

- High-prob bound on sensitivity $\Rightarrow$ Mechanisms probably DP
- Sampling, Emp process theory $\Rightarrow$ High-prob sensitivity bound

Idea 1: Sensitivity-Induced Privacy

Mechanism $\mathcal{M}$ (on target f) is sensitivity-induced private

If for neighbouring D, D': $\|f(D) - f(D')\|_{\mathcal{B}} \leq \Delta$ implies
$\forall B \subset \mathcal{B}, \Pr(\mathcal{M}_\Delta(D) \in B) \leq \exp(\epsilon) \cdot \Pr(\mathcal{M}_\Delta(D') \in B)$

Many mechanisms! Laplace, Gaussian, exponential, Bernstein

Connecting the dots:

- Choose a 'natural' distribution P on $\mathcal{D}$
- $\Pr(\mathcal{M}_\Delta$ being ϵ-DP on $D, D') \geq \Pr(\|f(D) - f(D')\|_{\mathcal{B}} \leq \Delta)$
- (γ, ϵ)-random DP (Hall et al. 2012):
  $\Pr(\mathcal{M}_\Delta$ being ϵ-DP on $D, D') \geq 1 - \gamma$
  Intuition: DP on most databases, ignore the pathological.

Idea 2: Sample and Estimate $\Pr\left(\|f(D) - f(D')\|_{\mathcal{B}} \leq \Delta\right)$

Define $G = \|f(D) - f(D')\|_{\mathcal{B}}$ from neighbouring $D, D' \sim P^n$

- CDF of G is $\Pr\left(\|f(D) - f(D')\|_{\mathcal{B}} \leq \Delta\right)$
- Idea 1: $\mathcal{M}_\Delta$ is RDP with confidence $1 - \gamma = CDF(\Delta)$
- Compute then invert $\Delta = CDF^{-1}(1 - \gamma)$? ...*groan*

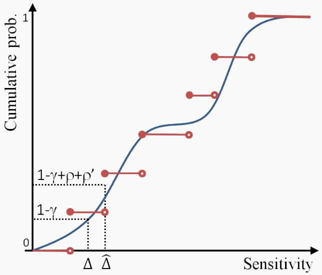
Idea 2: Sample and Estimate $\Pr(\|f(D) - f(D')\|_{\mathcal{B}} \leq \Delta)$

Define $G = \|f(D) - f(D')\|_{\mathcal{B}}$ from neighbouring $D, D' \sim P^n$

- CDF of G is $\Pr(\|f(D) - f(D')\|_{\mathcal{B}} \leq \Delta)$
- Idea 1: $\mathcal{M}_\Delta$ is RDP with confidence $1 - \gamma = CDF(\Delta)$
- Compute then invert $\Delta = CDF^{-1}(1 - \gamma)$? ...*groan*

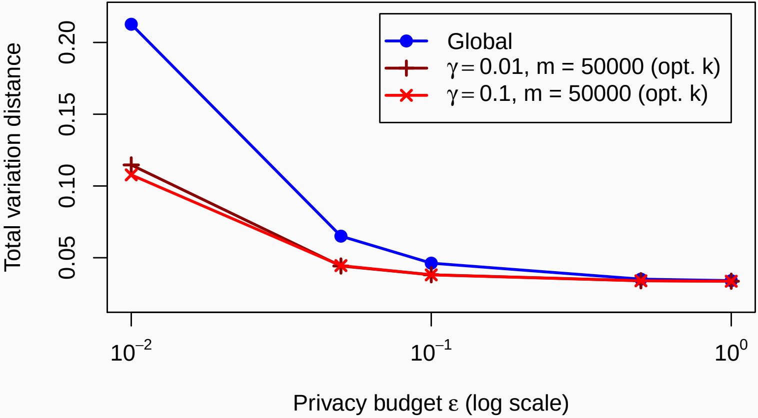
Algorithm: Sensitivity-sampler

1. Sample target: $G_1, \ldots, G_m \sim G$
2. Empirical CDF: $\frac{1}{m} \sum_{i=1}^{m} 1[G_i \leq \Delta]$
3. Dvoretsky-Kiefer-Wolfowitz: ECDF ρ' close to CDF, whp $1 - \rho$
4. $\Delta = ECDF^{-1}(1 - \gamma + \rho + \rho')$



Example: Priestly-Chao Kernel Regression

Density Estimation: Utility vs Privacy



Synthetic $n = 5000$ (1000 repeats); Bernstein with $k = 10, h = 3$

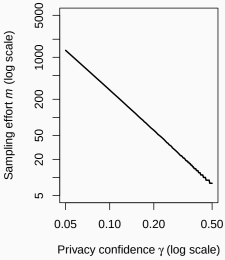
When resource constrained, can strike 'optimal' trade-offs:

Table 1. Optimal ρ operating points for budgeted resources—γ or m—minimising m, γ or k; proved in (Rubinstein & Aldà, 2017).

Budgeted	Optimise	ρ	γ	m	k
$\gamma \in (0,1)$	m	$\exp\left(W_{-1}\left(-\frac{\gamma}{2\sqrt{e}}\right)+\frac{1}{2}\right)$	$\bullet$	$\left\lceil \frac{\log\left(\frac{1}{\rho}\right)}{2(\gamma-\rho)^2} \right\rceil$	$\left\lceil m\left(1-\gamma+\rho+\sqrt{\frac{\log\left(\frac{1}{\rho}\right)}{2m}}\right)\right\rceil$
$m \in \mathbb{N}, \gamma$	k	$\exp\left(\frac{1}{2}W_{-1}\left(-\frac{1}{4m}\right)\right)$	$\geq \rho+\sqrt{\frac{\log\left(\frac{1}{\rho}\right)}{2m}}$	$\bullet$	$\left\lceil m\left(1-\gamma+\rho+\sqrt{\frac{\log\left(\frac{1}{\rho}\right)}{2m}}\right)\right\rceil$
$m \in \mathbb{N}$	γ	$\exp\left(\frac{1}{2}W_{-1}\left(-\frac{1}{4m}\right)\right)$	$\rho+\sqrt{\frac{\log\left(\frac{1}{\rho}\right)}{2m}}$	$\bullet$	m

Estimate sensitivity offline & in parallel

- m up, then RDP confidence $1-\gamma$ up



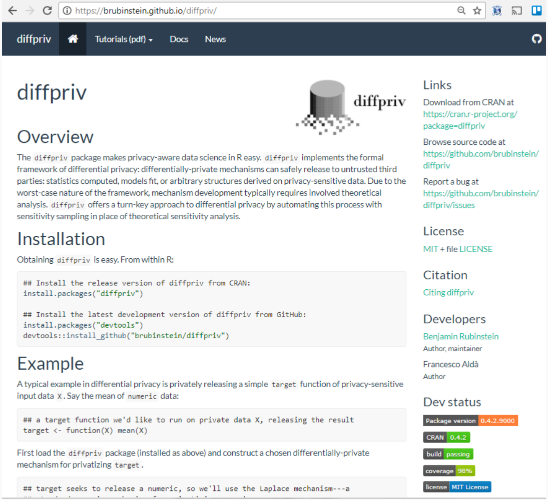
Distribution P on records:

- Non-informative *e.g.*, uniform, Gaussian
- A (public) Bayesian prior
- Density fit privately to data

The `diffpriv` Package

`diffpriv` on CRAN and GitHub

Google: 'diffpriv R'



Open-source R

'Official' on CRAN with rigorous submission process

`roxygen2` docs

Tutorial vignettes

98% Codecov

Travis CI

```
install.packages("diffpriv")
```

Architecture Highlights

`DPMech`: `VIRTUAL S4` class for sensitivity-induced mechanisms

1. Slot `target`: The non-private target function f
2. Slot `sensitivity`: Sensitivity of f to calibrate mechanism
3. `releaseResponse()`: Sample from response distribution
4. `sensitivityNorm()`: $\Delta_f(D_1, D_2) = \|f(D_1) - f(D_2)\|_{\mathcal{B}}$
5. `sensitivitySampler()`: Probes #4 to fill #2

Architecture Highlights

`DPMech`: `VIRTUAL S4` class for sensitivity-induced mechanisms

1. Slot `target`: The non-private target function f
2. Slot `sensitivity`: Sensitivity of f to calibrate mechanism
3. `releaseResponse()`: Sample from response distribution
4. `sensitivityNorm()`: $\Delta_f(D_1, D_2) = \|f(D_1) - f(D_2)\|_{\mathcal{B}}$
5. `sensitivitySampler()`: Probes #4 to fill #2

Included generic mechanisms, all subclass `DPMech`

- `DPMechLaplace`, `DPMechGaussian`: numeric release
- `DPMechExponential`: private optimisation
- `DPMechBernstein`: function release

Conclusions

Differential privacy

- Semantic privacy; practical in many ways; complements cryto
- Many deep connections between TCS, Stats/Learning, S&P

AAAI'17 Bernstein mechanism for private function release

ICML'17 Sensitivity sampler for automated RDP privatisation

`diffpriv` open-source R package implements these and more

Thankyou!

`http://bipr.net`

Narayanan & Shmatikov (2008) on *k*-Anonymity

"Sanitization techniques from *k*-anonymity literature... do not provide meaningful privacy guarantees"

"A popular approach to micro-data privacy is *k*-anonymity... This does not guarantee any privacy, because the values of sensitive attributes associated with a given quasi-identifier may not be sufficiently diverse [20, 21] or the adversary may know more than just the quasi-identifiers [20]. Furthermore... completely fails on high-dimensional datasets [2], such as the Netflix Prize dataset..."

Iterated Bernstein Operator

Order h, degree k

Bernstein operator:

$B_k(g; x) = \sum_{\nu=0}^{k} g(\nu/k) b_{\nu,k}(x)$

Iterated Bernstein operator:

$B_k^{(h)} = \sum_{i=1}^{h} (-1)^{i-1} B_k^i$ where $B_k^i = B_k \circ B_k^{i-1}$

Multivariate:

Evaluate g over lattice, Basis polynomials become products

Comparing DP Relaxations

ϵ-differential privacy

- Worst case on databases, Worst case on responses

(ϵ, δ)-differential privacy

- Worst case on databases, Protection for likely responses

(ϵ, γ)-random differential privacy

- Protection for likely databases, Worst case on responses