New Statistical Applications for Differential Privacy

Rob Hall

11/5/2012

Committee: Stephen Fienberg, Larry Wasserman, Alessandro Rinaldo, Adam Smith.

rjhall@cs.cmu.edu

http://www.cs.cmu.edu/~rjhall

Carnegie Mellon University
Statistics Department

MACHINE LEARNING DEPARTMENT

Contributions

- Overall theme: construction of practical methods for privacy-preserving statistical inference.

- Part 1: Private histograms
  - Laplace noise addition spoils histograms when n > p
  - A hard thresholding approach results in greatly increased utility.

- Part 2: Approximate differential privacy in an RKHS
  - Port noise addition techniques to an infinite dimensional function space.

- Part 3: Differential privacy for kernel density estimation
  - Laplace noise spoils the convergence rate in dimension > 3.
  - A new method obtains the correct rate.

Importance of Privacy

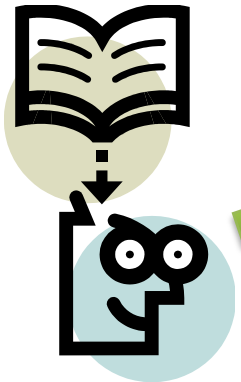
1. Agency has data about individuals.

Patient ID	Tobacco	Age	Weight	Heart Disease
0001	Y	36	170	N
0002	N	26	150	N
0003	N	45	165	Y
…	…	…	…	…

2. They release some statistics (e.g., for researchers to examine).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3957.0     4000.1   0.989   0.3313
UNEM          1133.8      513.1   2.210   0.0358 *
---
```

3. Somebody combines this with some side-information and identifies some individuals in the database.



Embarrassment to agency, individuals, etc. (at very least).

Importance of Privacy

1. Agency has data about individuals.

Patient ID	Tobacco	Age	Weight	Heart Disease
0001	Y	36	170	N
0002	N	26	150	N
0003	N	45	165	Y
…	…	…	…	…

2. They release some statistics (e.g., for researchers to examine).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3957.0     4000.1   0.989   0.3313
UNEM          1133.8      513.1   2.210   0.0358 *
---
```

3. Somebody combines this with some side information

… Embarrassment to individuals in the database. … etc. (at very least).

Goal of "Privacy-Preserving Data Mining" is to **prevent identification**, while **maintaining utility** in the release.

Privacy via Noise Addition

- Consider algorithms which incorporate randomness:

Patient ID	Age	Weight
0001	36	170
0002	26	150
0003	45	165
...	...	...

$D = (X, y)$

Private **Input** data

Some **private** quantity of interest

$$\theta(D) = (X^T X)^{-1} X^T y$$

Output (to world)

$$\tilde{\theta} \sim P_D$$

e.g., $\tilde{\theta} \sim \mathcal{N}(\theta(D), \Sigma)$

- We characterize private methods as sets of probability measures indexed by datasets: $\mathbf{P} = \{P_D : D \in \mathcal{D}\}$

Called a "mechanism."

(α,β)-Differential Privacy

- A mechanism is called (α,β)-Differentially Private when:

$$\forall D \sim D', \ \forall A \in \mathcal{A}: \ P_D(A) \leq e^{\alpha} P_{D'}(A) + \beta$$

A pair of databases which differ by one element ("adjacent").

The measurable sets in the output space.

- Remarks:
  - Symmetric in D and D'.
  - Means that the distribution of the output "doesn't change much" when the input changes by one element.
  - Called "Approximate Differential Privacy" when β > 0.
  - Implications…[WZ]

Methods of Differential Privacy

- Almost all methods to achieve (α,0)-differential privacy boil down to this:

$$dP_D(x) \propto \exp\{-\Delta d(x, \theta(D))\}$$

Metric on the output space

$$\Delta = \alpha \left(\sup_{D \sim D'} d(\theta(D), \theta(D')) \right)^{-1}$$

"global sensitivity"

- Leads to
  - Laplace noise, "gamma" noise (Euclidean space with L1 or L2 norms).
  - Exponential mechanism [MT] (discrete output space).
  - K-norm mechanism [HT] (Euclidean space with a Minkowski norm).

- For β > 0, lighter tailed distributions may be used (e.g., Gaussian).

Part 1: Lower Risk Bounds

- Noise addition leads to a loss in "utility" which we characterize by risk:

$$R(\theta, \mathbf{P}, D) = \int_{\mathcal{Z}} \ell\left(\theta(D), \tilde{\theta}\right) \, dP_D(\tilde{\theta})$$

Depends on the function to release, the mechanism, and the data itself.

The "loss function."

- We characterize the cost of differential privacy by the minimum **worst case** risk available:

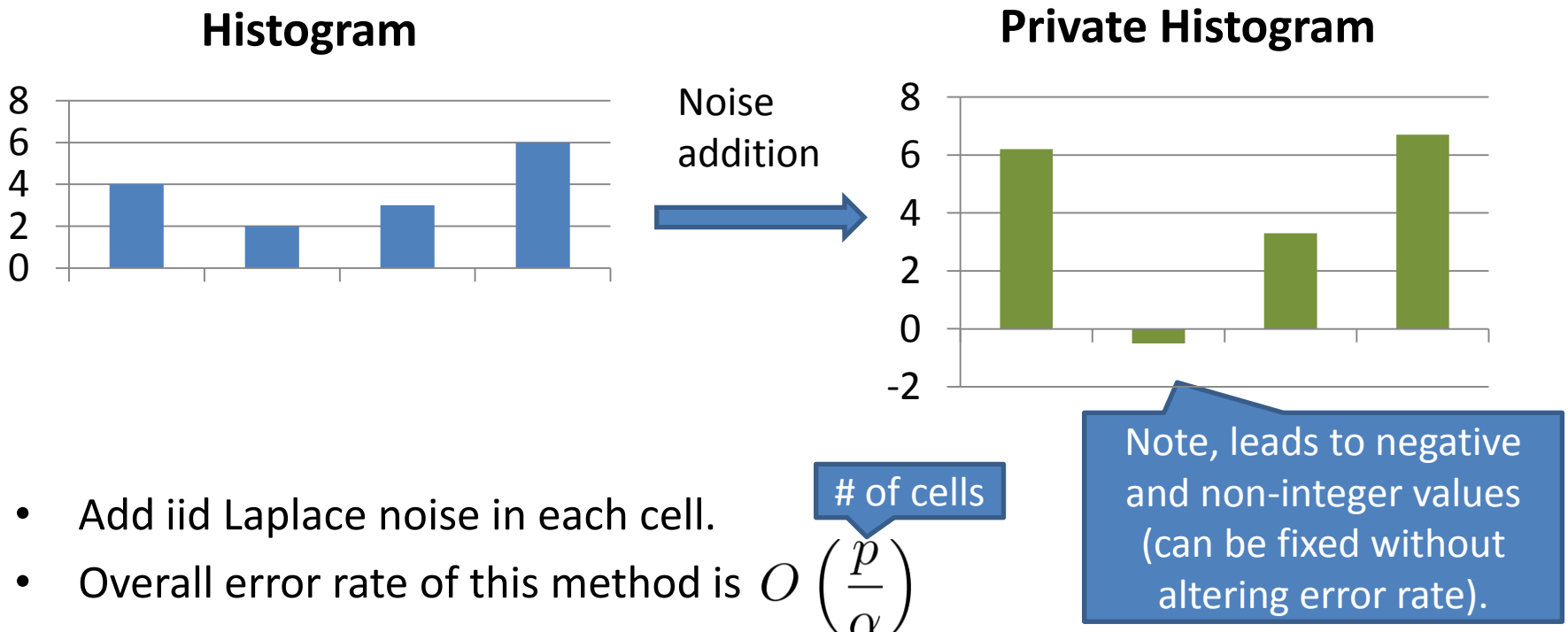
$$R^{\star}_{(\alpha,\beta)}(\theta) = \inf_{\mathbf{P} \in \mathrm{DP}(\alpha,\beta)} \sup_{D \in \mathcal{D}} R(\theta, \mathbf{P}, D)$$

Depends only on the function to release.

The smallest among all methods P which admit differential privacy.

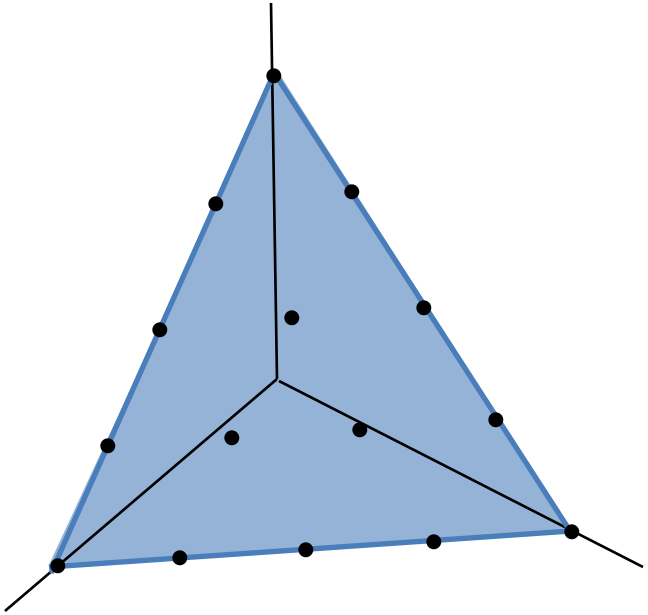
Basic Private Histograms

- Histograms were one of the first statistical estimators to be studied under (α,0)-DP.



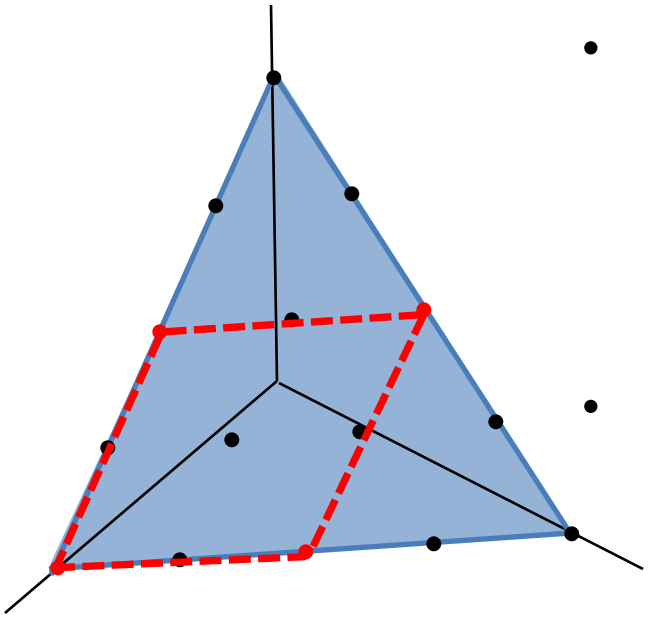
- Add iid Laplace noise in each cell.
- Overall error rate of this method is $O\left(\frac{p}{\alpha}\right)$ (# of cells)
- Is this the minimax *rate*?

Lower Risk Bounds for Histograms



- Regard the set of histograms as lattice points in an appropriately scaled simplex.
- WLoG: consider mechanism to be indexed by histograms rather than datasets.

Lower Risk Bounds for Histograms



- For a hypercube of lattice points having appropriate width:
  - (α,0)-DP ensures the distributions at these points are close (in KL).
  - Fano's inequality leads to lower risk bound.
- Alternative technique due to [HT].
  - Only holds for mechanisms which admit a kind of "continuous extension" (unclear how important this is [De]).

e.g., for l_1 loss

Some universal constant

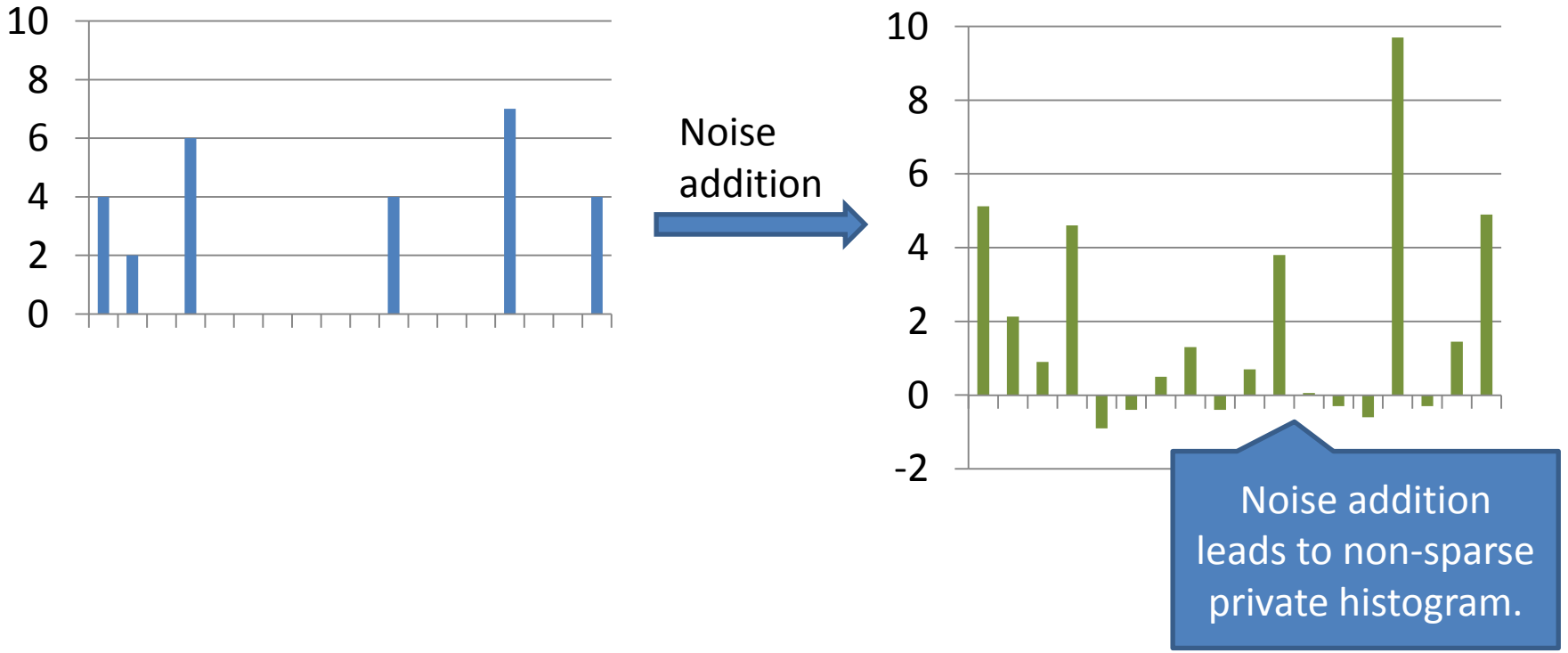
$$R^{\star}_{(\alpha,0)} \geq \frac{c_0 p}{\alpha}$$

Laplace noise addition achieves the optimal rate.

Histograms in High Dimensions

- When $p > n$, noise addition completely swamps the data.

- In high dimensional histograms we encounter sparsity



Histograms in High Dimensions

- Sacrifice minimaxity in order to obtain better performance when the histogram is sparse.

- Suppose that only $q < p$ cells will ever be occupied.

$$\inf_{\mathbf{P} \in \mathrm{DP}(\alpha,0)} \sup_{D \in \mathcal{D}_q} R(\theta, \mathbf{P}, D) > \frac{c_0 q \log(p/q)}{\alpha}$$

Restrict to datasets which produce sparse histograms

Now depends linearly on q, logarithmically on p

- Thus if the number of occupied cells is small relative to n there is still hope.

- Note similarity to sparse normal means problem [J11].

Methods for Sparse Private Histograms

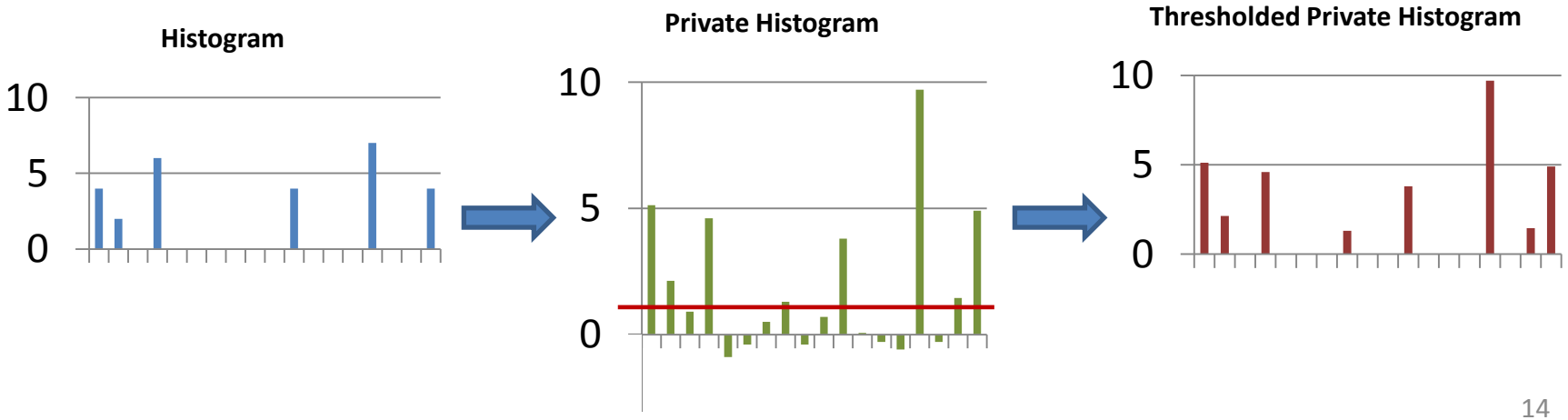
- The Laplace noise technique evidently doesn't achieve this rate (since the risk is the same for all histograms).
- We hard-threshold the resulting histogram [CPST]

Output for cell *i*

$$\tilde{\theta}_i = \begin{cases} 0 & \theta_i(D) + L_i \leq \tau \\ \theta_i(D) + L_i & \text{o/w}, \end{cases}$$

Threshold

True cell value plus Laplace noise.



Low Risk for Thresholded Histograms

- We hard-threshold the resulting histogram [CPST]

Output for cell *i*

$$\tilde{\theta}_i = \begin{cases} 0 & \theta_i(D) + L_i \le \tau \\ \theta_i(D) + L_i & \text{o/w}, \end{cases}$$

Threshold

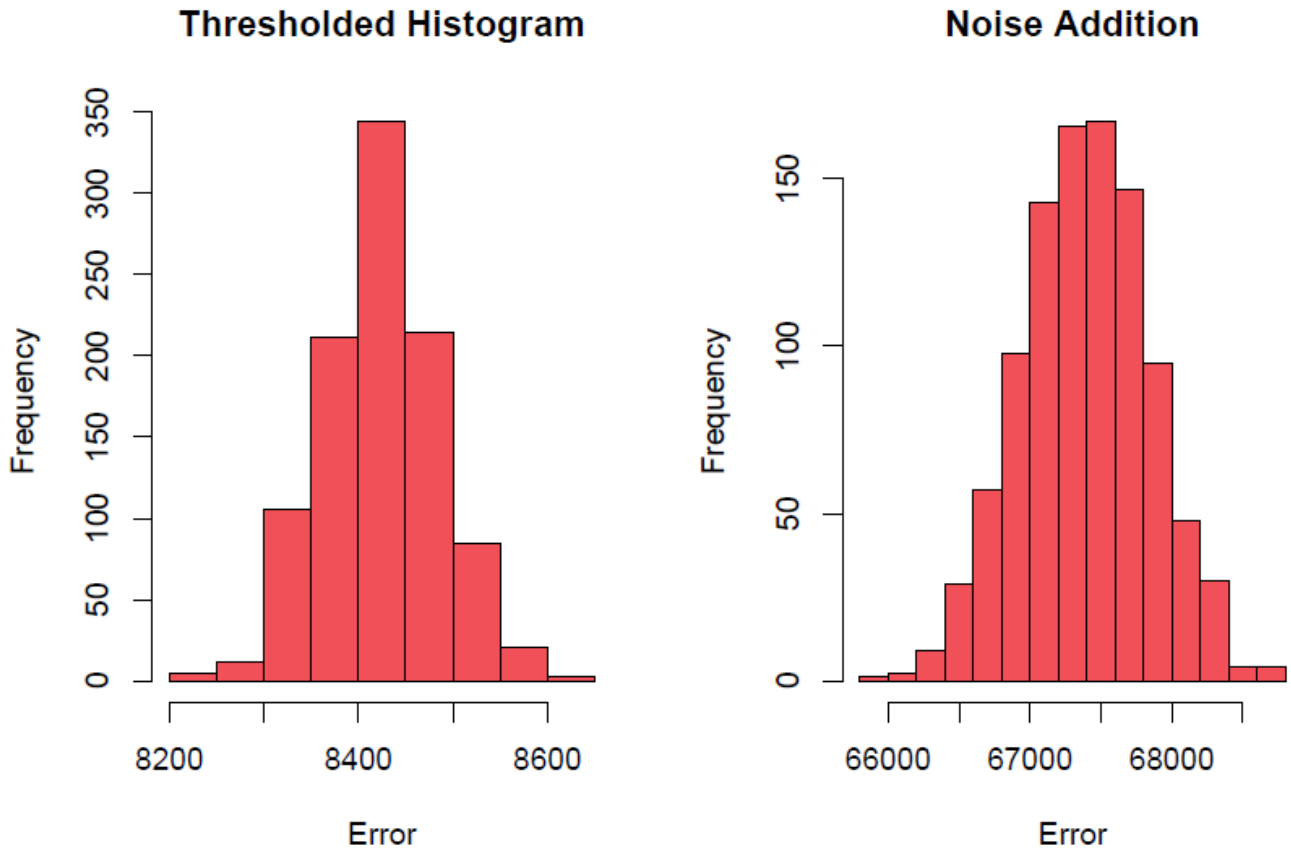
True cell value plus Laplace noise.

- If we choose $\tau = \tau(p, \alpha) = \frac{2}{\alpha} \log p$ then the risk is $O\left(\frac{q \log p}{\alpha}\right)$

- This is close to the "restricted minimax risk" but off by an additive amount of the order $q \log q$.

- To achieve actual minimax rate: know *q* in advance or do FDR [J11].

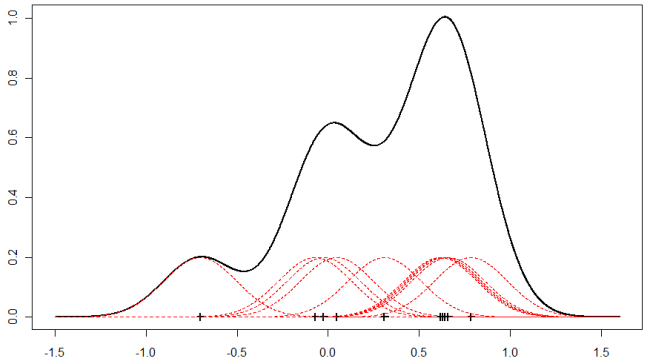
Thresholding Dramatically Improves Utility

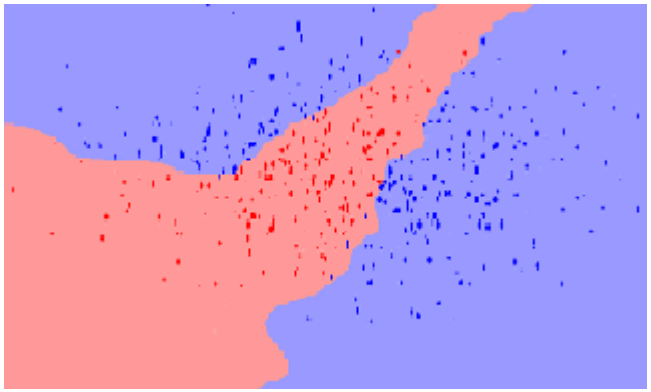


Error distribution over several trials, NLTCS data with $p=2^{16}$, $q=3152$ and $n=21574$.

Part 2: Approximate Privacy in an RKHS

- Sometimes the goal is to release a function





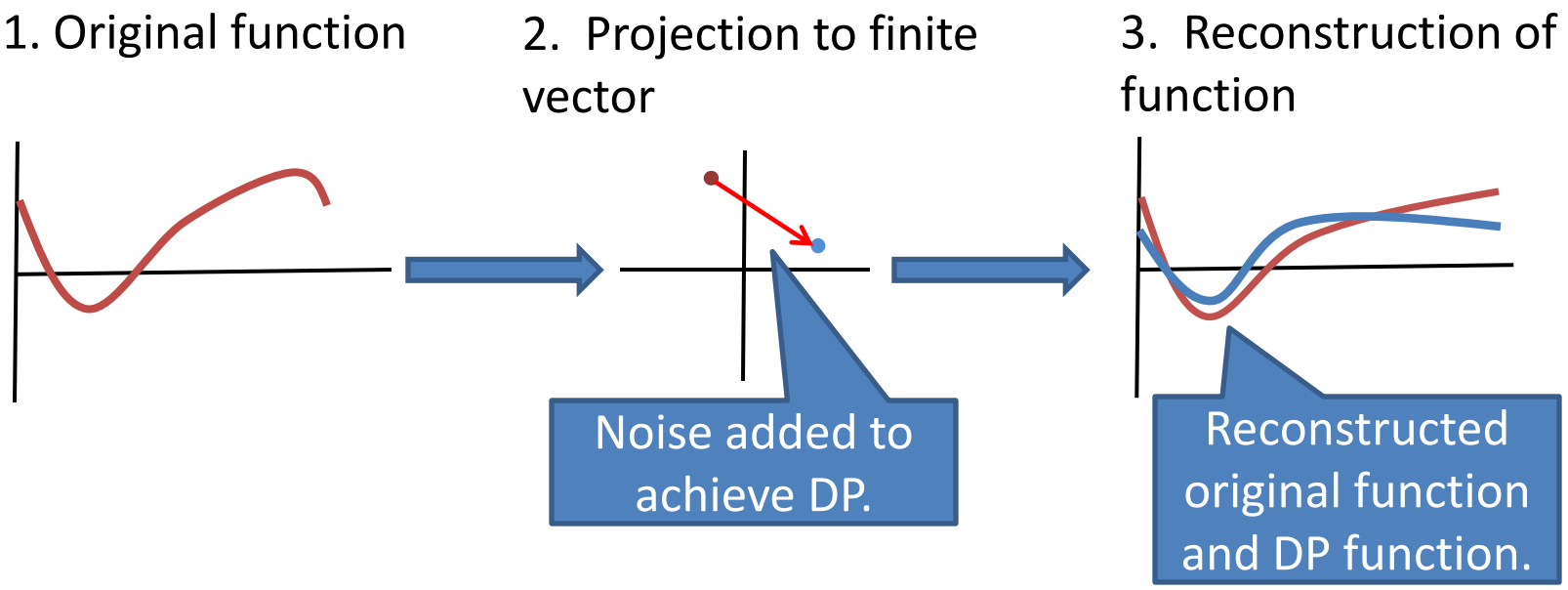
e.g., kernel density estimation, SVM regression functions etc.

- The basic techniques to achieve DP do not work in this infinite dimensional setting (no dominating measure).

- Techniques exist for DP linear SVMs though [CMS,RBHT].

Project, Privatize then Reconstruct

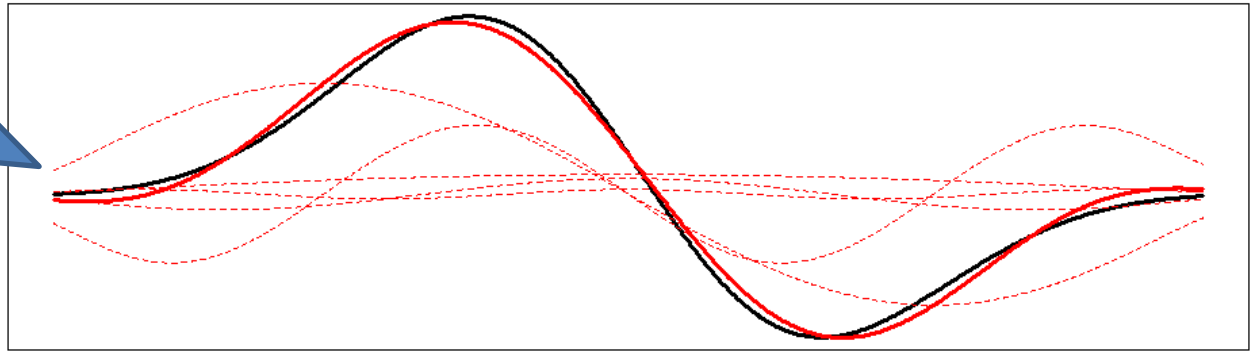
- Idea: reduce to finite m-dimensional vector and apply existing DP techniques to coordinates.



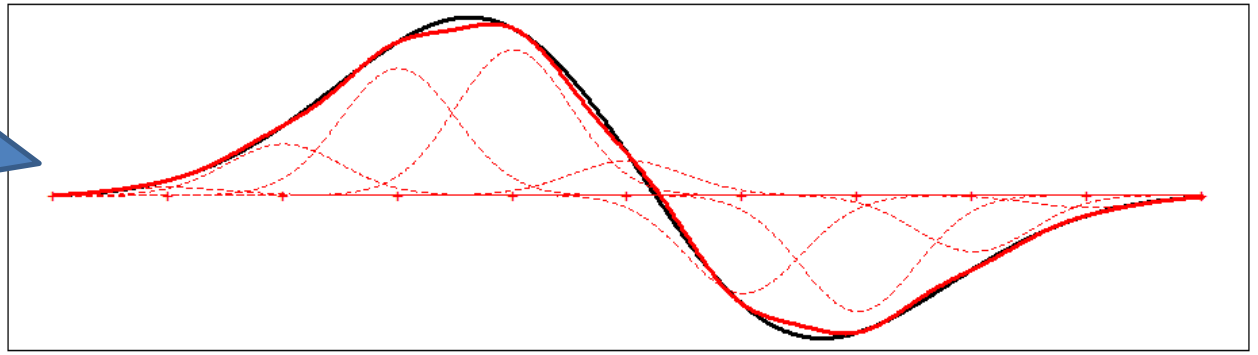
RKHS Projections

- Two ideas for finite representations:

Expansion into orthonormal basis of eigenfunctions.



Approximation by linear combination of kernels



Alternatives: Taylor's series, Fourier series etc.

(α,β)-DP in an RKHS

- Add **Gaussian noise** calibrated to sensitivity.

$$\|\theta(f_D) - \theta(f_{D'})\|_\Sigma \leq \|f_D - f_{D'}\|_{\mathcal{H}(K)}$$

Sensitivity of finite vectors.

RKHS distance between functions.

- Measured in L_2, the error decomposes:

$$\mathbb{E}\left\|\tilde{f}_D - f_D\right\|_2^2 \leq \mathbb{E}\left\|\tilde{f}_D - \hat{f}_D\right\|_2^2 + \sup_D \left\|f_D - \hat{f}_D\right\|_2^2$$

Noise addition

Approximation error

- Allow the dimension to grow to infinity:
  - The first term is **bounded** (easy to see for eigen-representation).
  - The second term **vanishes** (under e.g., seperability).

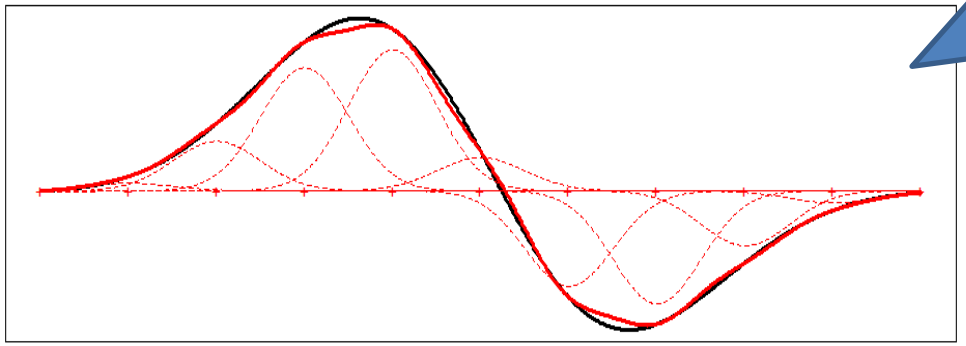
(α,β)-DP in an RKHS

- Both techniques become the same thing: **addition of a GP** with mean zero and covariance function = the reproducing kernel of the RKHS.
  - Eigendecomposition:

$$\tilde{f}_D = \sum_{i \geq i} \left(\lambda_i \langle f_D, \psi_i \rangle_{\mathcal{H}(K)} + \sqrt{\lambda_i} Z_i \right) \psi_i = f_D + \sum_{i \geq 1} \sqrt{\lambda_i} Z_i \psi_i$$

"Coordinates" of *f* | Gaussian noise | Gaussian Process

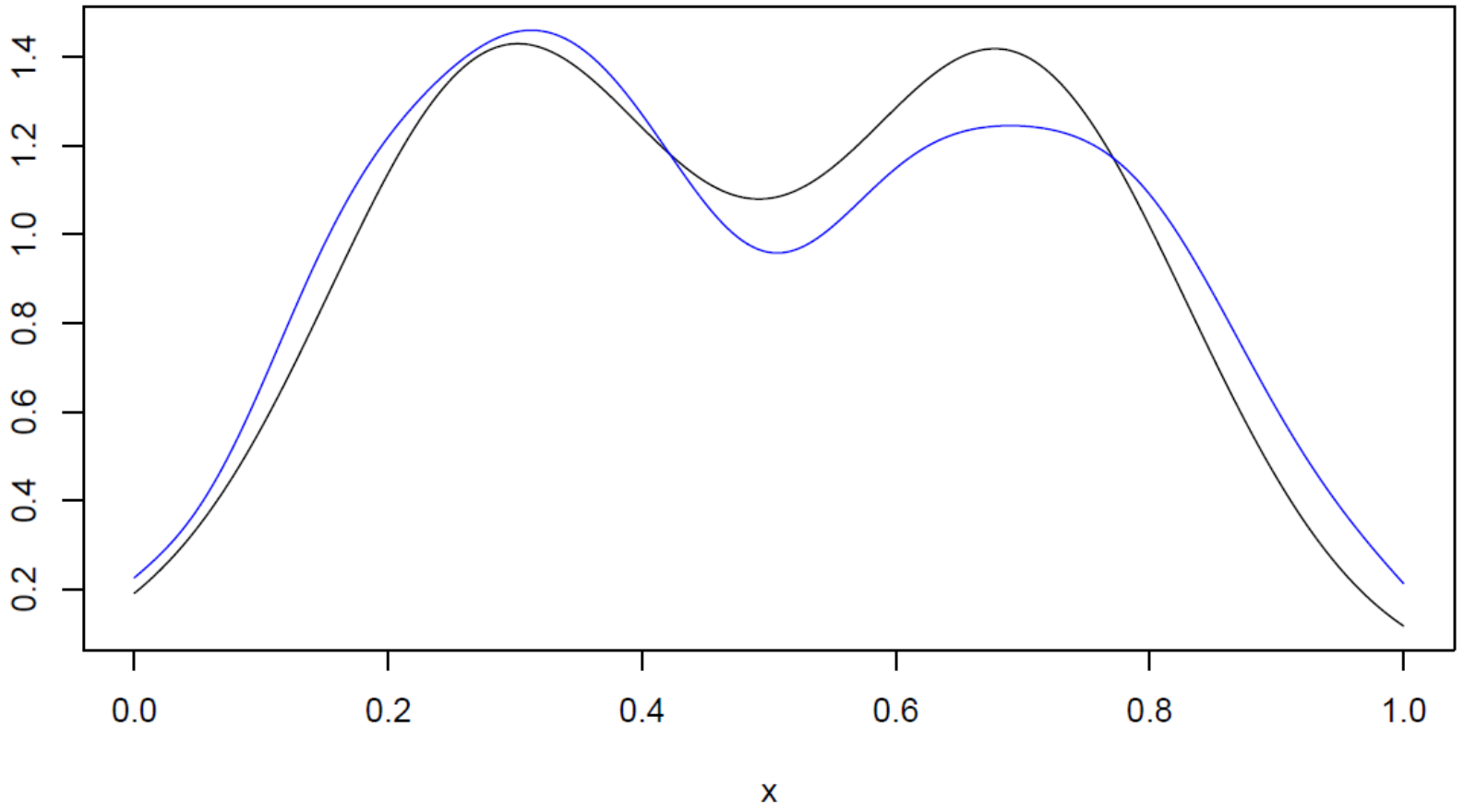
  - Linear combination of kernels:



Equivalent to output of function values + noise at kernel locations.

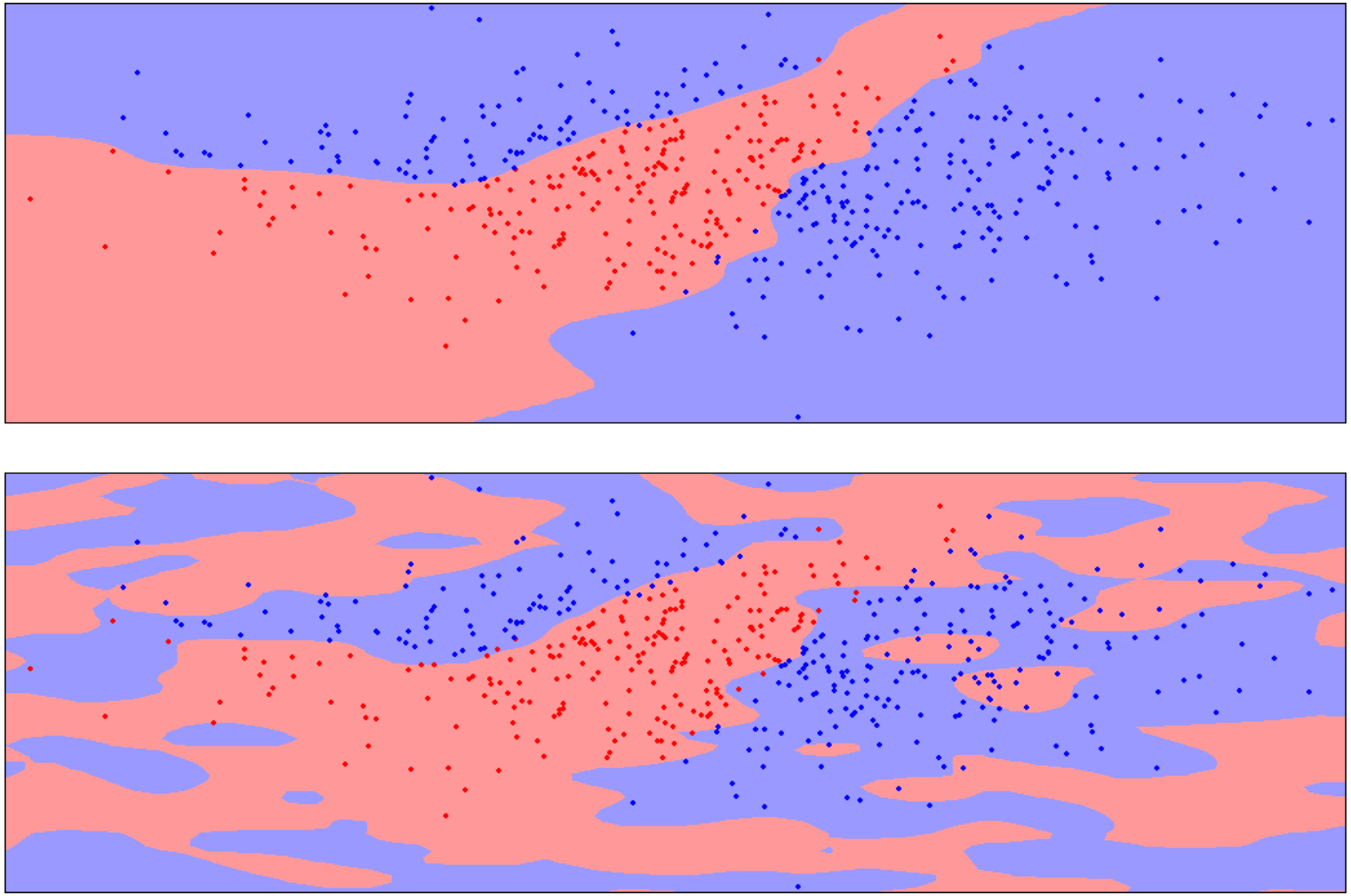
Kolmogorov's extension theorem…

Example 1: (α,β)-DP Kernel Density Estimation

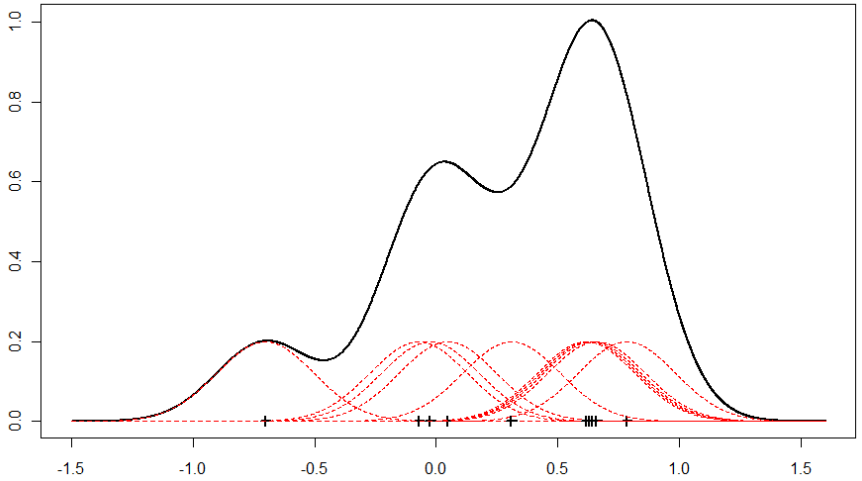


Remarks: released function is smooth, error is of smaller order than sampling error.

Example 2: (α,β)-DP Kernel SVM



Part 3: Differentially Private Kernel Density Estimation



- So far we have an (α,β)-DP method for this task which does not spoil the convergence rate.
- We turn attention to the construction of a (α,0)-DP method.
- Unfortunately cannot simply modify the previous technique.
- [WZ] give a technique which works in one dimension under different regularity conditions.

Facts About Kernel Density Estimation

- Recall the form

$$\hat{f}_X(x) = \frac{1}{nh}\sum_{i=1}^{n}\phi\left(\frac{x - x_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n}\sigma_h(x - x_i)$$

Standard normal density.

Normal density with std. dev. h

- Under appropriate regularity conditions, and when

$$h = h_n = O\left(n^{-1/(4+d)}\right)$$

then this estimate enjoys the minimax convergence rate

$$\mathbb{E}\int_{[0,1]^d}\left(\hat{f}_X(x) - f(x)\right)^2 = c_1 h^4 + \frac{c_2}{nh^d} = O\left(n^{-4/(4+d)}\right)$$

Basic Private Kernel Density Estimation

- We eigendecompose the Gaussian kernel into the Fourier basis

Periodic version of kernel.

$$\tilde{\phi}_\sigma(x,y) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x)\psi_i(y)$$

$$\lambda_1 = 1, \quad \lambda_{2j} = \lambda_{2j+1} = \exp\left\{-2j^2\sigma^2\right\}$$

$$\psi_1(x) = 1,$$
$$\psi_{2j}(x) = \sqrt{2}\cos(2j\pi x),$$
$$\psi_{2j+1}(x) = \sqrt{2}\sin(2j\pi x)$$

- Truncation to the first m terms leads to the modified KDE

Coordinates in Fourier basis

$$\hat{f}_X^m(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_j\psi_j(x_i)\psi_j(x) = \sum_{j=1}^{m}\lambda_j\left(\frac{1}{n}\sum_{i=1}^{n}\psi_j(x_i)\right)\psi_j(x)$$

- Following [WZ] we add Laplace noise to each coordinate…

$$\tilde{f}_X^{m,\alpha}(x) = \sum_{j=1}^{m}\lambda_j\left(\frac{1}{n}\sum_{i=1}^{n}\psi_j(x_i) + L_j\right)\psi_j(x)$$

For d>1 use a tensor product construction

Laplace Noise Spoils the Rate

- The error may be written

$$\mathbb{E}\|\hat{f}_X - \tilde{f}_X^{m,\alpha}\|^2 \leq \mathbb{E}\|\hat{f}_X - \tilde{f}_X^{m}\|^2 + \mathbb{E}\|\tilde{f}_X^{m} - \tilde{f}_X^{m,\alpha}\|^2$$

Error due to truncation

Error due to noise

- The first term is in the minimax rate so long as

$$m = O\left(\frac{\log n}{h}\right) = O\left(n^{1/(4+d)} \log n\right)$$

- But then the second term is in the order

$$O\left(\frac{m^{3d}}{n^2}\right) = O\left(n^{\frac{d-8}{d+4}} (\log n)^{3d}\right)$$

Incorrect rate when d > 3

- Independent Laplace noise adds more error than is necessary.
  - Here the coordinates exhibit dependence.

The “K-Norm” Method Achieves the Minimax Rate

- In one dimension, each kernel function corresponds to a point on the trigonometic moment curve

$$\Psi_m(x) = (\psi_1(x), \ldots, \psi_m(x))$$

and the truncated estimate corresponds to the average of these, which resides in the convex body

$$K_m = \text{conv}\left\{\Psi_m(x), x \in [0,1]\right\}$$

- This is known as the “Caratheodory Orbitope” and is studied for its facial structure in the algebraic geometry literature [SSS].
- Using noise calibrated to the seminorm associated with this body [HT] leads to error in the correct order.
- Efficient sampling is feasible using MCMC (at least when d=1).

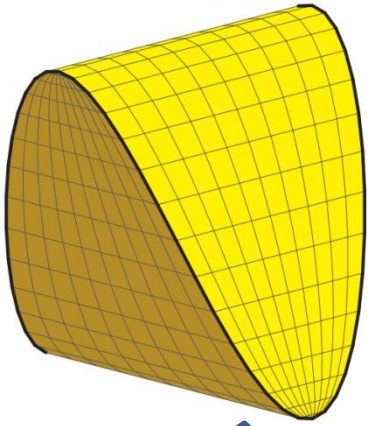


Image from [SSS].

Contributions

- Discrete estimators
  - Study of the utility of private histograms (contingency tables etc).
  - Theoretical demonstration of improved error rates in sparse settings.

- Private functions
  - Theoretically clean method for the release of private functions.
  - Easy to implement in practice (see text).

- Private density estimation
  - A density estimator in the correct minimax rate.
  - A new way of using the K-norm approach of [HT].
  - Allows e.g., the construction of private synthetic data through sampling.

Open Problems

- Discrete estimators
  - Attainment of the true minimax rate in a sparse setting via FDR.
- Private functions
  - What can be done to relax the requirement for approximate DP?
- Private density estimation
  - How can the sampling be done efficiently in high dimensions?
  - What other kinds of functions are amenable to this analysis?

Bibliography

- [CMS] Chaudhuri, K., Monteleoni, C. and Sarwate, A. Differentially private empirical risk minimization. JMLR 2011.
- [CPST] Cormode, G., Procopiuc, C., Srivastava, D. and Tran, T. Differentially private publication of sparse data. CoRR 2011.
- [De] De, A. Lower bounds in differential privacy. CoRR 2011.
- [HT] Hardt, M. and Talwar, K. On the geometry of Differential privacy. STOC 2010.
- [J11] Johnstone, I. Gaussian estimation: Sequence and multiresolution models (textbook draft).
- [MT] McSherry, F. and Talwar, K. Mechanism design via differential privacy. FOCS 2007.
- [RBHT] Rubinstein, B., Bartlett, P., Huang, L. and Taft, N. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. JPC 2012.
- [SSS] Sanyal, R., Sottile, F. and Sturmfels, B. Orbitopes. CoRR 2009.
- [WZ] Wasserman, L. and Zhou, S. A statistical framework for differential privacy. JASA 2010.

Thanks.

Part 4: Relaxed Privacy Definitions

- Consider sparse high-dimensional settings
  - Census data, genetic data etc.
- Utility loss required by DP may render inference impossible.
  - c.f., Charest's thesis.
- Possible solutions:
  - Abandon the inference.
  - Abandon privacy and publish the raw data.
  - Use α-DP with some huge α.
  - Adhere to some weaker notion of privacy.

i.e., abandon hope of protecting against DP adversary.

Random Differential Privacy

- Move from a **worst-case** to an **average case** guarantee:

Usual DP condition

$$\mathbb{P}\left(\forall A \in \mathcal{A},\ P_D(A) \leq e^{\alpha} P_{D'}(A) + \beta\right) \geq 1 - \gamma$$

The (n+1)-fold product measure over elements of D,D'

A new parameter

- Much weaker than DP:
  - Protects a random individual with high probability.
  - Individuals in the tail of P(x) may be exposed.
  - Nevertheless may be useful in certain situations.

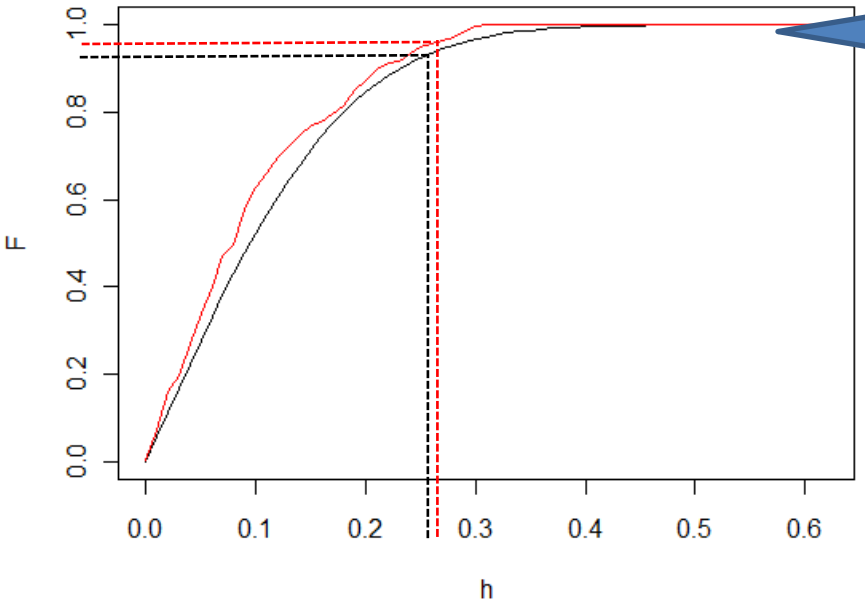
- Composes nicely (DP composition + "union bound").

RDP via Sensitivity Analysis

- Consider functions g for which the sensitivity decomposes as

$$\sup_{D \sim D'} \left| g(D) - g(D') \right| = n^{-1} \sup_{d,d'} h(d,d')$$

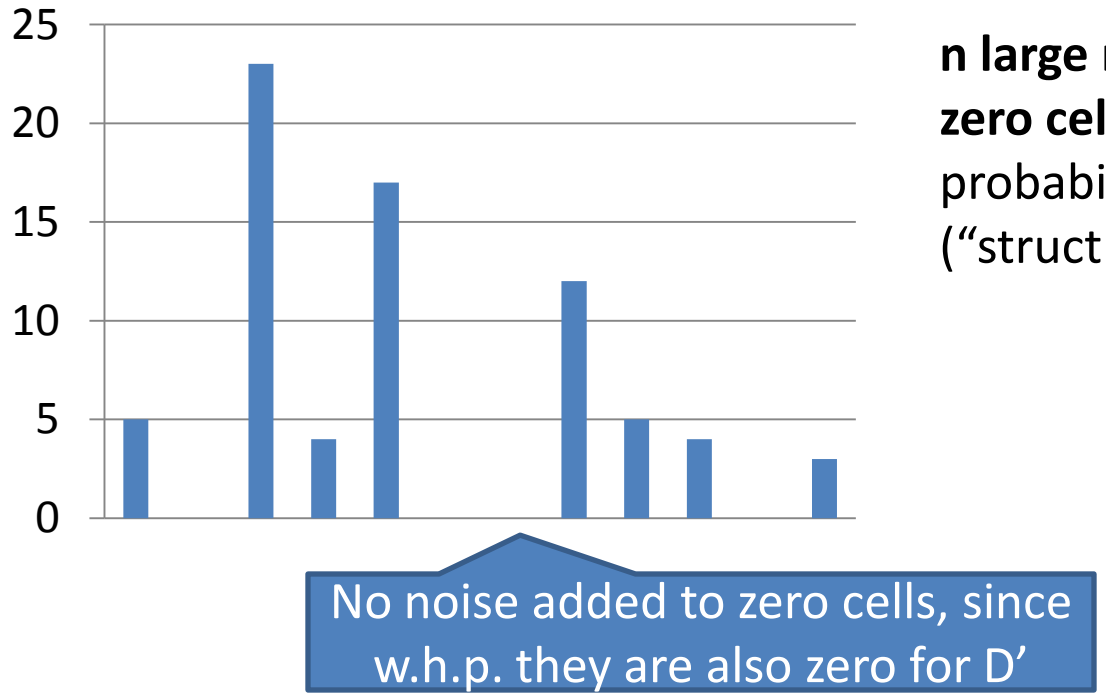
- Taking a sample quantile of $h(x_i,x_j)$ leads to a bound on the sensitivity which holds w.h.p:



Invoke e.g., DKW to get upper bound on requisite quantile.

- "Sensitivity bound" depends on data
- Requires "smooth sensitivity" treatment.
- Smoothness established (w.h.p) by Kiefer's work on U-quantiles.

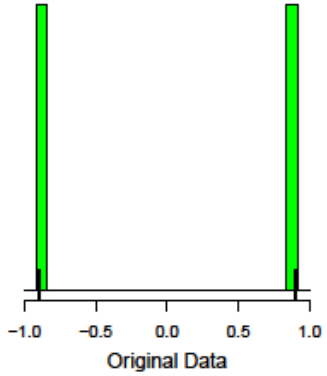
RDP Sparse Histograms

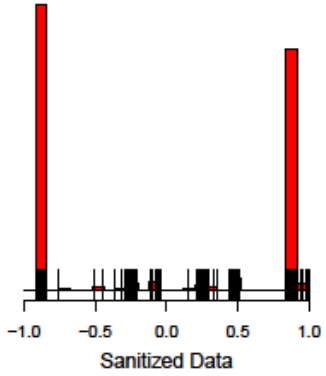


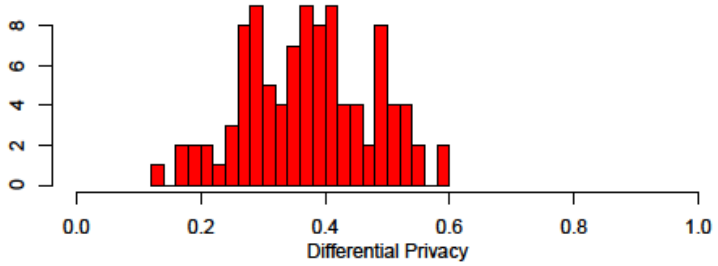
n large relative to number of zero cells: zero cells have low probability under P(x) ("structural zeroes").

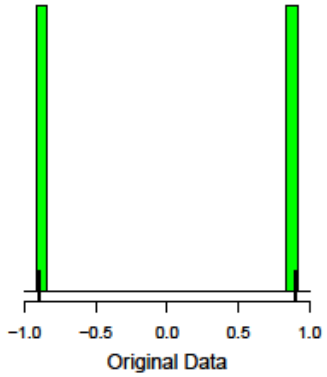
The "adversary" doesn't learn anything from these cells since (from his data alone) he could already conclude they have low probability.

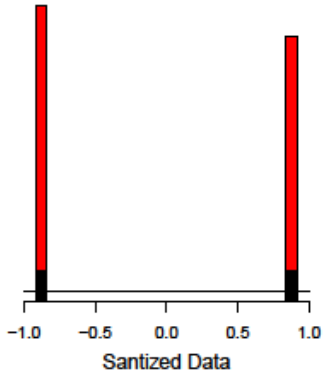
RDP Sparse Histograms

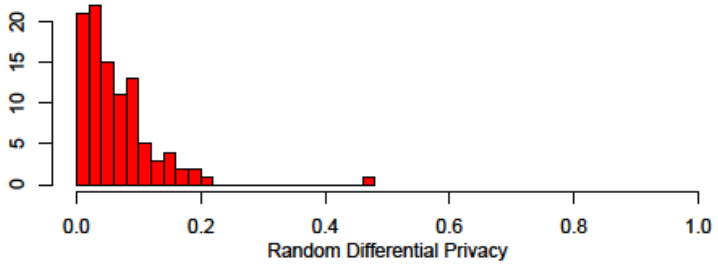












Methods for Sparse Private Histograms

- Just as in the sparse normal means problem, if *q* were known we could take

$$\tau = \frac{2}{\alpha} \log \frac{p}{q}$$

  which leads to hard-thresholding at the minimax rate.

- However since *q* is generally unknown this is not satisfactory

- As in the normal means problem, the solution will be to control the "false discovery rate" (FDR) [cite].
  - Mostly a theoretical exercise since we anticipate q (and hence the reduction in risk) to be small, whereas the analysis is considerably more complex.
  - Thus left as future work.