New Paradigms and Optimality Guarantees in Statistical Learning and Estimation

Yu-Xiang Wang

In partial fulfillment of

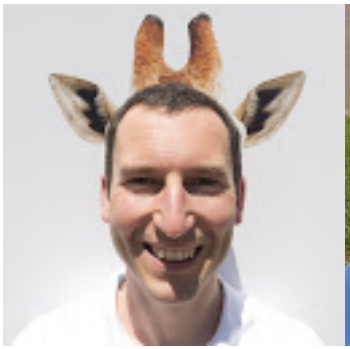
PhD in Statistics and Machine Learning

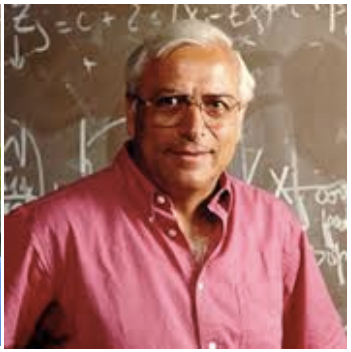
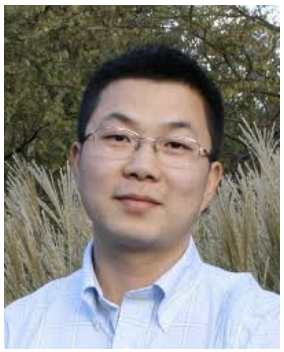
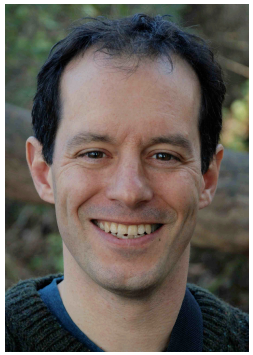
Thesis committee:

Ryan J. Tibshirani (Chair), Jing Lei, Alex J. Smola,

Adam D. Smith (Boston Univ.) , Steve E. Fienberg

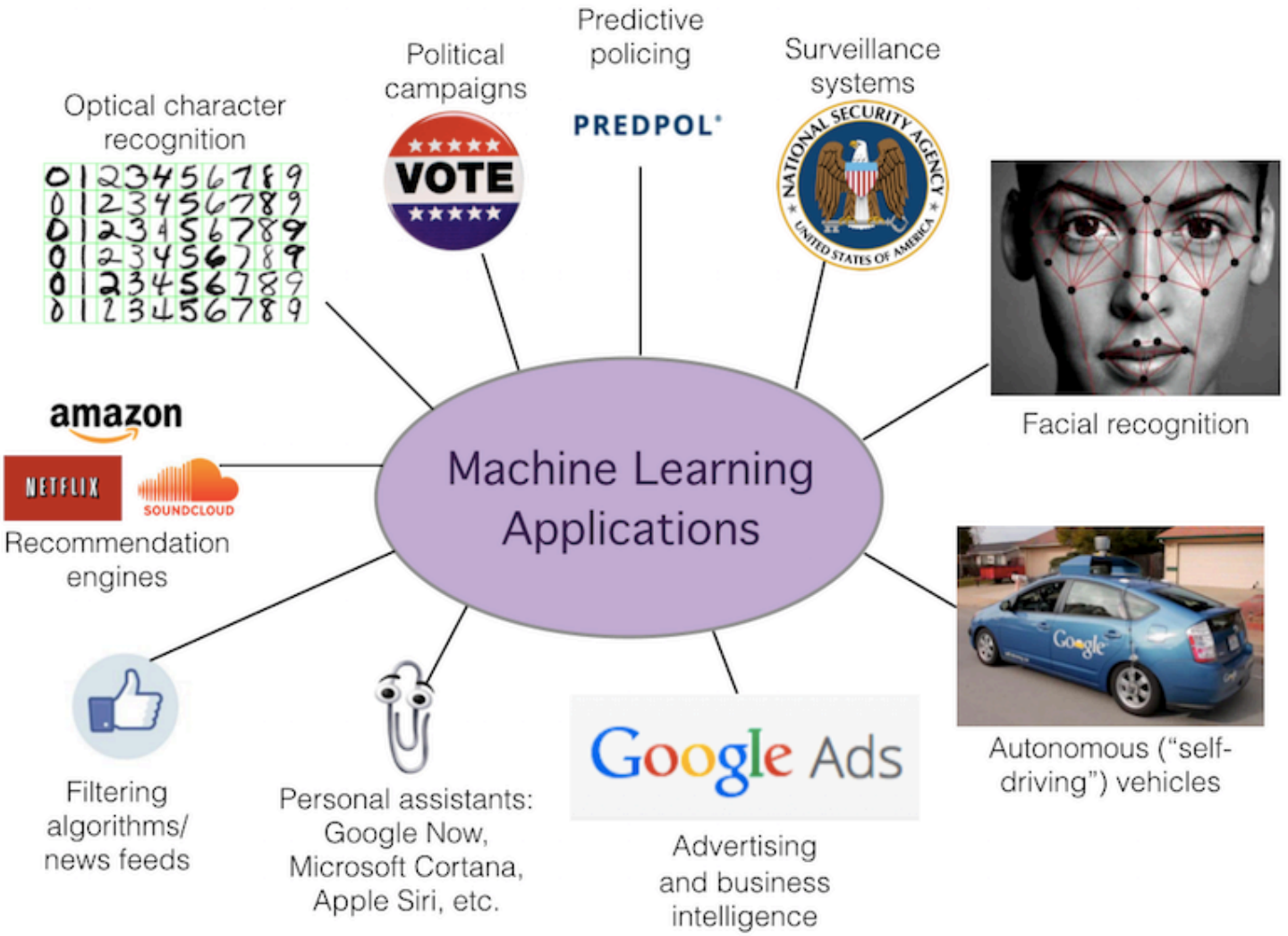
Acknowledgments



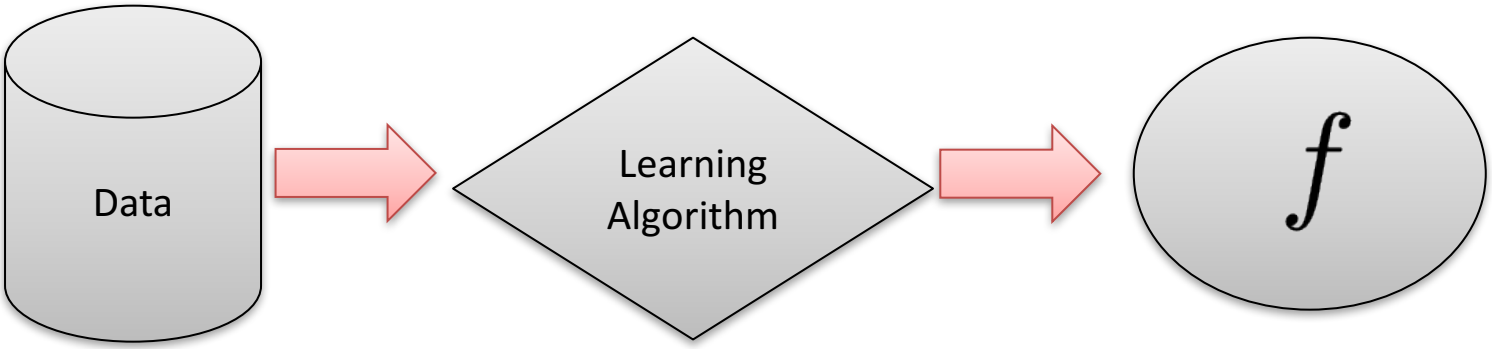







The empirical success of ML



New challenges that arises



Feature-label pairs
Unlabeled features
Feature points

Support Vector Machine
K-means clustering
Kernel density estimation

Classifier
K cluster centers
Estimated density function

Data challenge
- Data owner won't share unless **privacy** is promised.

Modeling challenge
- How general
- How specific
- What's the tradeoff?

Implicit overfitting:
- How do we ensure the learned outcome is statistically valid
- Even after multiple rounds of selections

Outline

- Part I: Privacy (25 min)
- Part II: Trend filtering (25 min)
- Part III: Sequential selective estimation (10 min)

Part I: Privacy

Lessons from privacy breaches

- Need methods with provably privacy guarantee!
- Need new ways to interact with datasets.

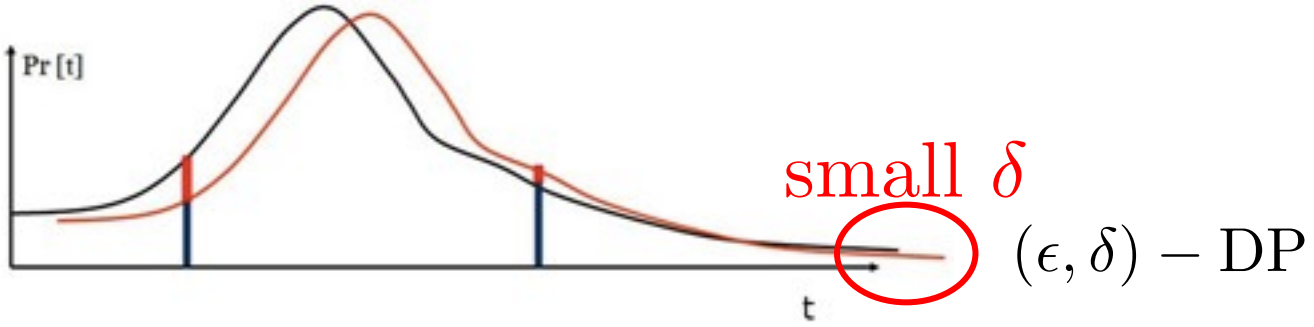
History of privacy technologies

- Statistical disclosure control
  - [Duncan et. al., Hundepool et. al., since 1970s]
- k-anonymity, l-divergence, t-closeness
  - [Sweeny, Machanavajjhala et. al., Li et. al., 2002-2007]
- Differential privacy
  - [Dwork, McSherry, Nissim, Smith, 2006++]
  - Gödel Prize 2017

Formal definition of DP

- Let Z,Z' be any two datasets that differ only by one row, and A is a randomized algorithm. We say A is ε-DP if for all output h

$$\sup_{Z,Z':d(Z,Z')\leq 1} \sup_{h\in\mathcal{H}} \log \frac{p_{h\sim\mathcal{A}(Z)}(h)}{p_{h\sim\mathcal{A}(Z')}(h)} \leq \epsilon$$



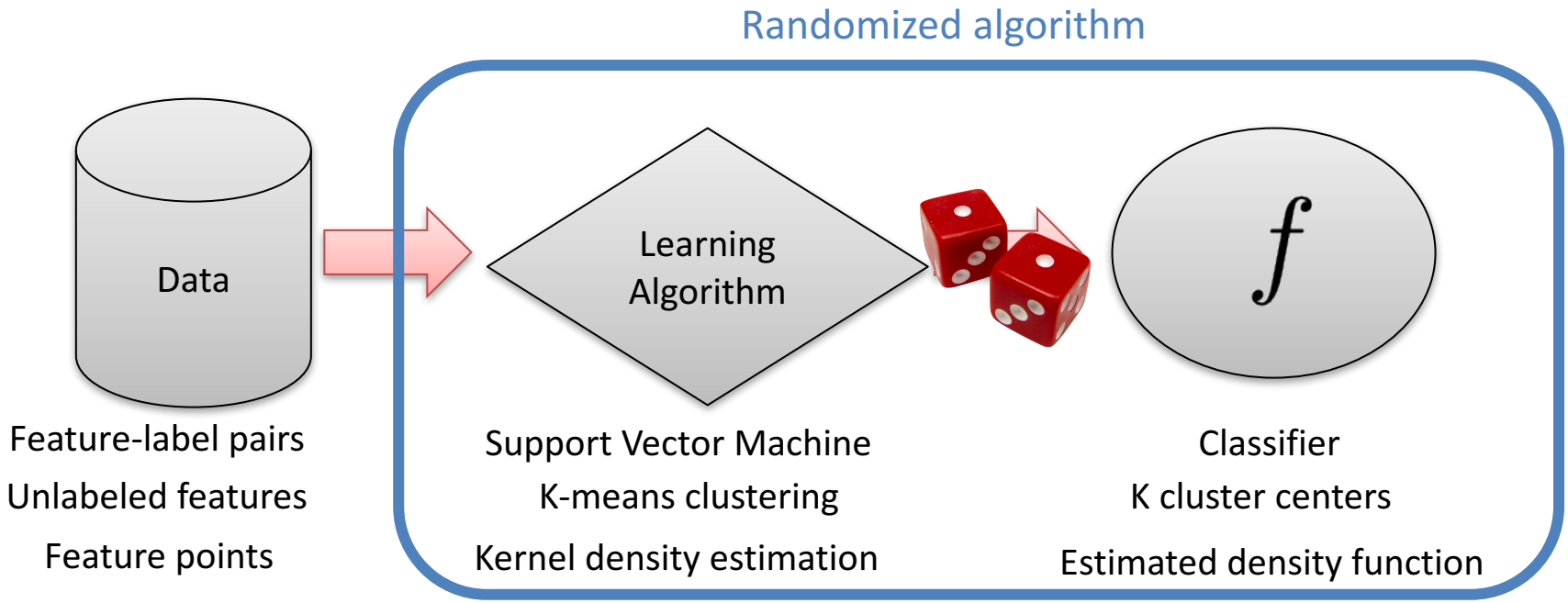
Example: Who voted for Trump?

- How many people in this room voted for Trump?

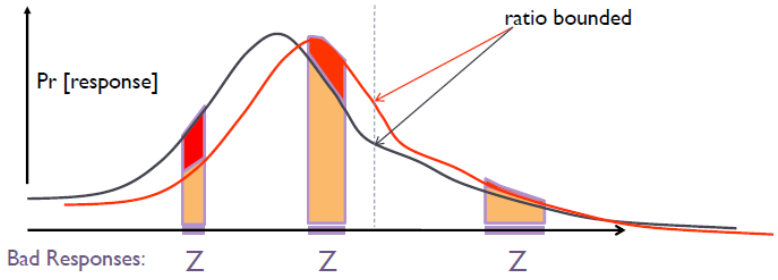
- Let's say the answer is 8.
  - If I know the political view of everybody, except Ryan.
  - Then I can easily infer his choice.

- DP releases: 8 + noise.

Differentially private machine learning

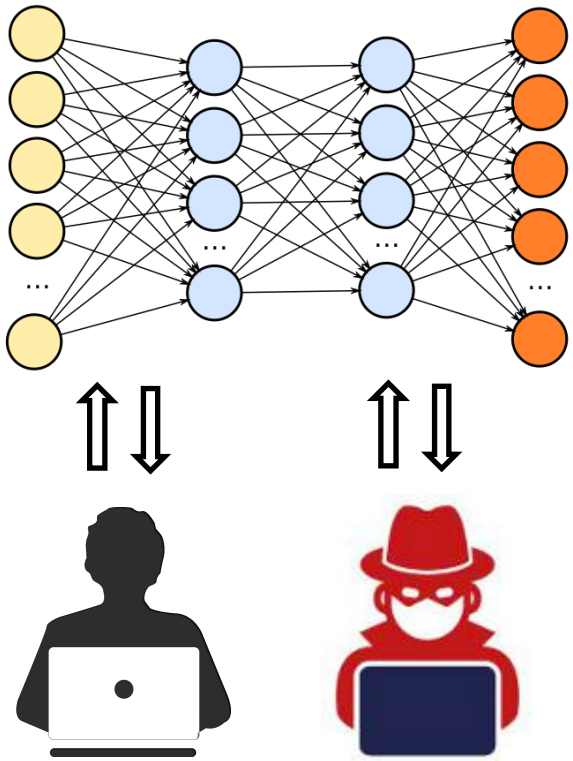


$$\log \frac{\mathbb{P}(f \in S|Data)}{\mathbb{P}(f \in S|Data')} \leq \epsilon$$



[Blum et. al. , Kasiviswanathan et. al., Chaudhuri et. al., Duchi et. al., Kifer et. al., Bassily et. al., 2008 onwards]

Example: Recommendation System



- If A is private, prediction is “post-processing”.

Contribution

- What I talked about at my proposal
  - Learnability under differential privacy [W.,Lei, Fienberg, JMLR'16]
  - Generic / scalable algorithm for DP-learning: OPS [W.,Fienberg, Smola, ICML'15]
  - On-Average KL-privacy [W.,Lei, Fienberg, PSD'16]

- Questions I received:
  - On-Average KL private is not quite private.
  - OPS might not be the best algorithm for linear regression.

- Today:
  - Per-instance DP and pDP for all.
  - AdaOPS algorithm for private linear regression

The need to weaken DP

- In theory
  - A lot of simple problems are not privately learnable.(W., Lei, Fienberg, JMLR'16)
- In practice
  - Poor utility due to too much noise. e.g., Contingency Table (Fienberg et. al. 2010), GWAS data (Yu et. al., PSD'14).
  - Hard to use. Need a lot of tricks/hacks to work. e.g., "clipping" "rescaling" as in the Netflix data. (Liu, W., Smola, RecSys'15)

On-Average KL-Privacy

- Differential Privacy:

Max-Divergence

$$\sup_{Z,Z':d(Z,Z')\leq 1} \sup_{h\in\mathcal{H}} \log \frac{p_{h\sim\mathcal{A}(Z)}(h)}{p_{h\sim\mathcal{A}(Z')}(h)} \leq \epsilon$$

- On-Average KL-Privacy:

$$\mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\log \frac{p_{h\sim\mathcal{A}(Z)}(h)}{p_{h\sim\mathcal{A}([Z_{-1},z])}(h)}\right] \leq \epsilon.$$

KL-Divergence

Per-instance DP

- **Definition**: A is ε-pDP on (Z,z) if

$$\sup_{Z,Z':d(Z,Z')\leq 1} \sup_{h\in\mathcal{H}} \log \frac{p_{h\sim\mathcal{A}(Z)}(h)}{p_{h\sim\mathcal{A}(Z')}(h)} \leq \epsilon$$

  - a strict generalization
  - Measures the privacy loss a specific person z suffers from running A on a specific data set Z.

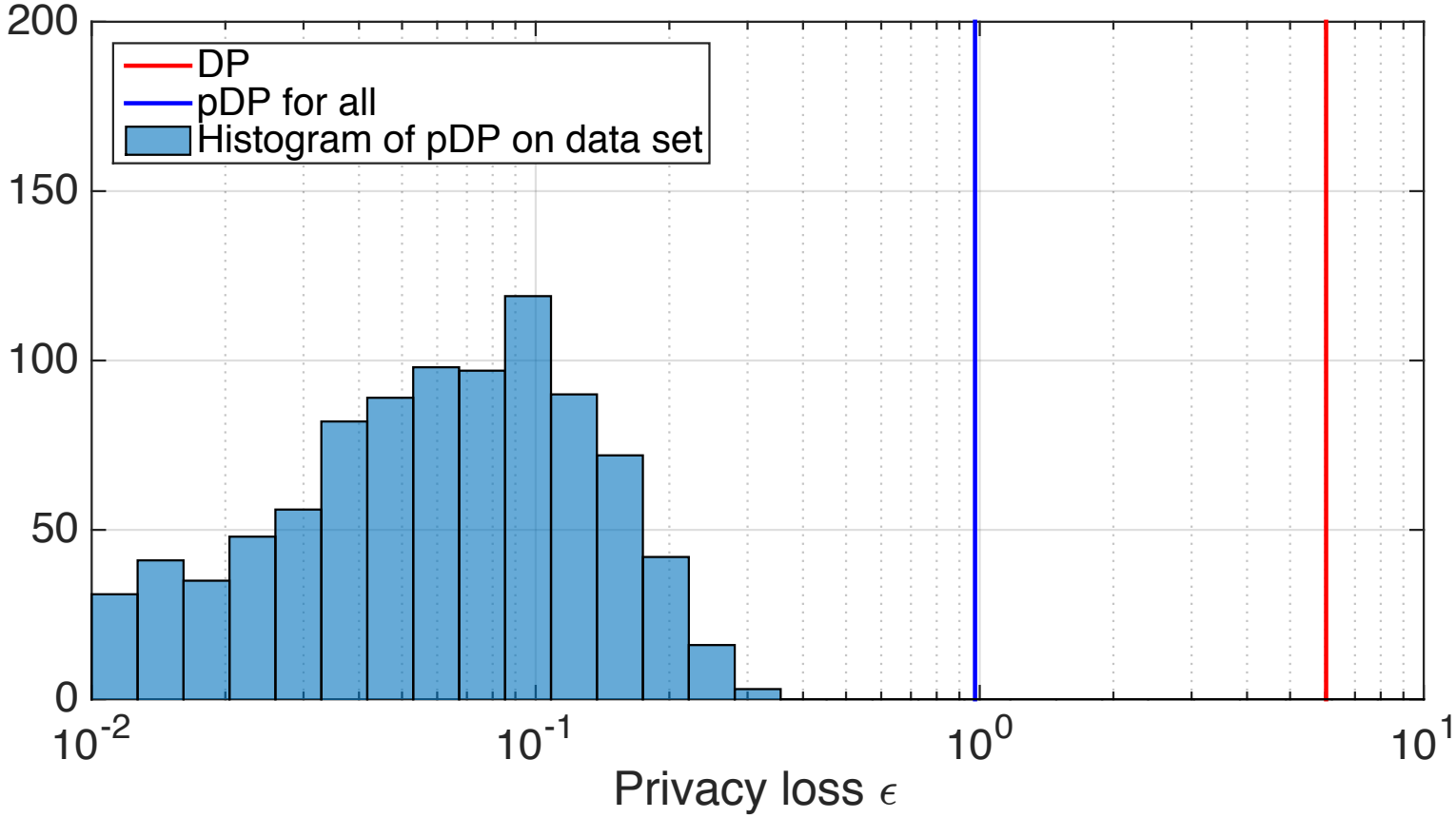
Implicit adversary models

- DP:
  - Adversary choose both data set Z and target z.
- pDP for all:
  - Adversary is given a fixed Z, but can choose a target z
- pDP:
  - Adversary is given data set Z and target z

pDP vs. DP: an illustration

Generate data set by linear Gaussian model. Fix the algorithm below.

$$\tilde{\theta} \sim N((X^T X + I)^{-1} X\mathbf{y}, \sigma^2 I), \quad \sigma = 4$$



Per-instance sensitivity

- The per instance sensitivity of function f

$$\Delta_{\|\cdot\|_*}(f, Z, z) = \|f(Z) - f([Z, z])\|_*$$

- Global sensitivity : max over (Z,z)
- Local sensitivity: fix Z, max over z

Stability of stationary points

- Let f be an optimization query:
  - Find me a stationary point of the loss function

$$f(Z) \in \{\theta | \nabla \mathcal{L}_Z(\theta) = 0\}$$

Lemma: Critical points of $\mathcal{L}_Z$ and $\mathcal{L}_{[Z,z]} = \mathcal{L}_Z + \ell_z$ obey that

$$\hat{\theta}' - \hat{\theta} = \left[\int_{\hat{\theta}}^{\hat{\theta}'} \nabla^2 \mathcal{L}_Z(t) dt \right]^{-1} \nabla \ell_z(\hat{\theta}')$$

Per-instance sensitivity of linear regression coefficients

- per-instance sensitivity in A-norm is

$$|y - x\hat{\theta}'|\sqrt{x^T(X^TX)^{-1}A(X^TX)^{-1}x}$$

Residual/prediction error

Statistical leverage score, when A ≈ X^TX

- Multivariate Gaussian mechanism for pDP.

Which ``A'' to use for Multivariate Gaussian noise adding?

- Standard choice:
  - A $\propto$ Identity $\Leftrightarrow$ Output Pert. [CMS-2013]
- Democratic choice:
  - A $\propto$ (X^TX)^2 $\Leftrightarrow$ Obj Pert. [CMS-2013]
- ``Fisher'' choice:
  - A $\propto$ X^TX $\Leftrightarrow$ OPS

Refined statistical analysis of OPS for linear regression

- Previous analysis [W. Fienberg, Smola, 2015]
  - $(1 + 4B/\varepsilon)$-efficiency and ε-DP
  - Restrict domain s.t. loss function < B

- Direct analysis using pDP:

$$1 + O\left(\frac{d\log(1/\delta)}{n\epsilon^2}\right)$$ and (ε,δ)-pDP for all unit x

No domain restriction needed!

Faster rate, better dimension-dependence than [Smith, 2008] and [Dwork & Smith, 2009], who first obtain such 1+ o(1) statistical efficiency.

Regret of OPS in agnostic setting

- Let $F(\theta) = 0.5\|\mathbf{y} - X\theta\|^2$

- OPS on regularized objective $F(\theta) + \frac{\lambda}{2}\|\theta\|_2^2$

$$F(\tilde{\theta}) - F(\theta^*) \leq \frac{d\log(d/\delta)\log(2/\delta)}{[\lambda + \lambda_{\min}(X^T X)]\epsilon^2} + \lambda\|\theta^*\|_2^2$$

With probability 1-δ

Matches both lower bounds in [Bassily et. al., 14].

High probability bound. Run time does not depend on ε.
Works in unbounded domain. highly practical.

Two expected complaints

- Linear regression is a bit restrictive.
  - All can be generalized!
  - Two ongoing work:
    - (1) Self-concordant GLM
    - (2) Morse-Smale stable nonconvex optimization
- pDP for all ≠ DP (cannot calibrate noise to ε)
  - Next slide!
  - pDP => DP with Propose-Test-Release [Dwork & Lei, 2009]

AdaOPS for Linear Regression

1. DP-release of $\bar{\lambda} > \lambda_{\min}(XX^T)$ **1-Stable by Weyl's lemma**

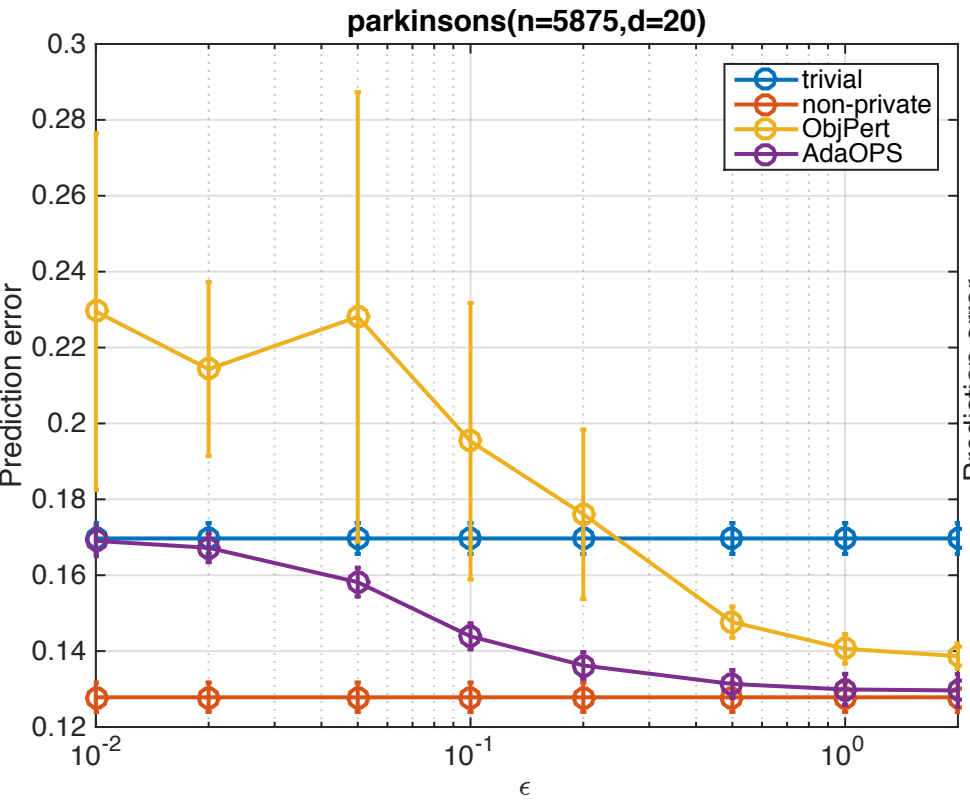
2. DP-release of $\bar{B} > \|\theta^*\|_2$ **1-Stable after log(1+ .) transform**

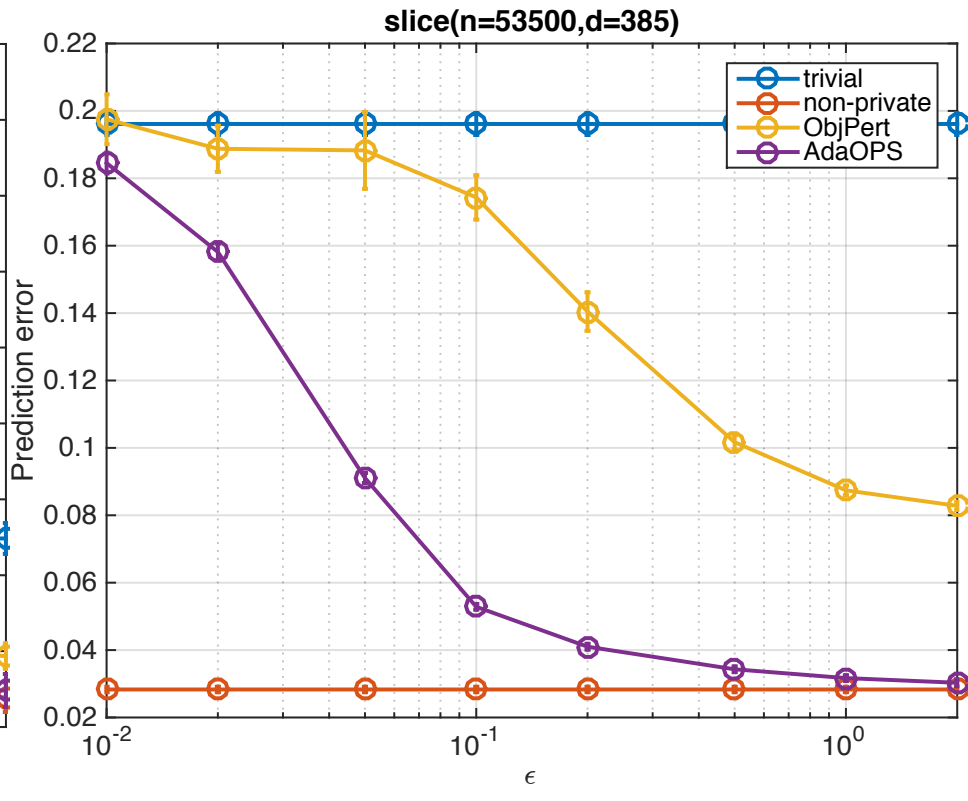
3. Choose γ, λ appropriately using the remaining balance of ϵ, δ

Regularization plays a more important role than noise

1. Output: $\tilde{\theta} \sim N(\theta^*, \gamma^{-1}(XX^T + \lambda I)^{-1})$

AdaOPS on real data sets





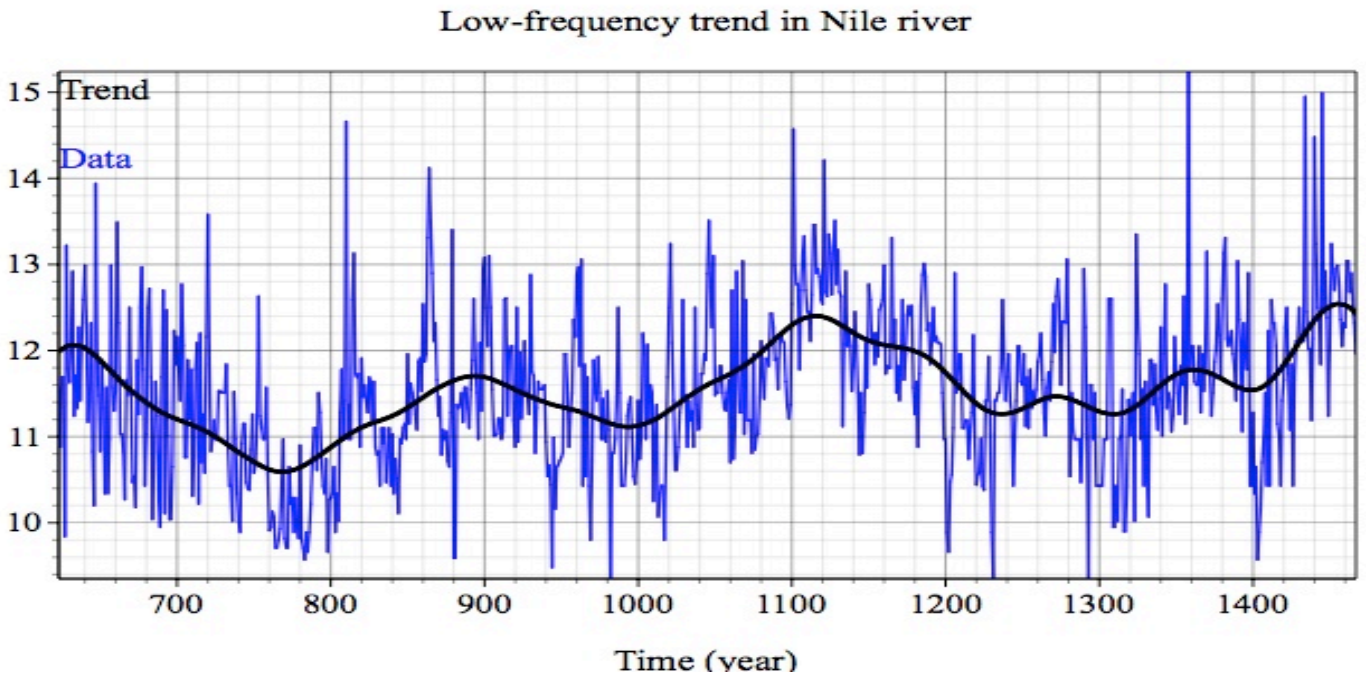
- There are 34 others data sets with results that look just like these.

- Other methods compared to:
NoisySGD,sufficient Statistics perturbation, and output perturbation.
None of them is stronger than ObjPert in this case.

Summary of Part I

- pDP provides more comprehensive summary of the privacy effect of randomization.
- pDP can be used as a tool to design data-adaptive DP algorithm. (complemetary to PTR and smooth sensitivity)
- AdaOPS is quite promising for practical DP learning.
- Future work:
  - Privately release pDP losses.
  - pDP and the economics of data collection

Part II Trend filtering



Nonparametric regression

- 50+ years of associated literature

  [Nadaraya, Watson, 1964]

  – Kernels, splines, local polynomials

  – Gaussian processes and RKHS

  – CART, neural networks

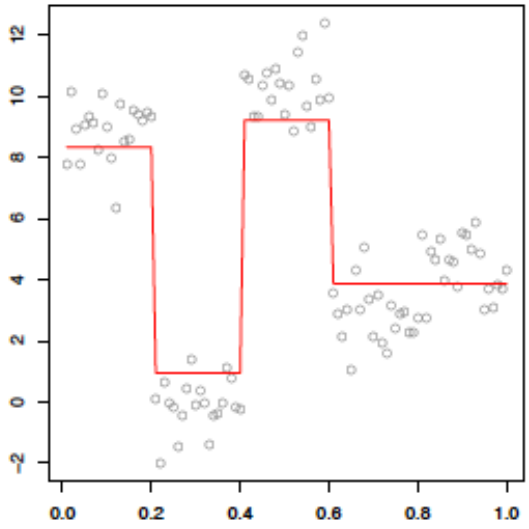
- Also known as smoothing, signal denoising /filtering in signal processing & control.

Adapting to local smoothness

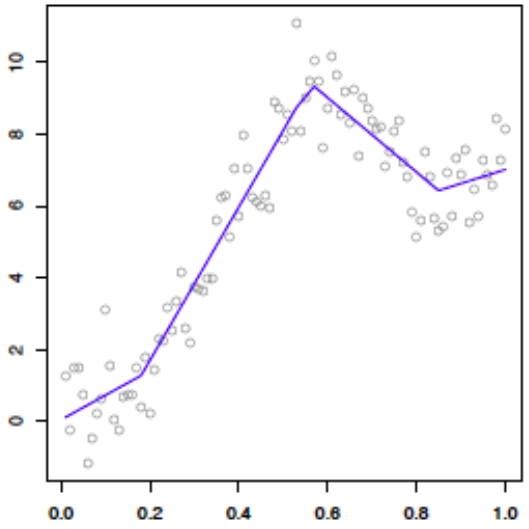
- Some parts smooth, other parts wiggly.
  - Wavelets [Donoho&Johnston,1998], adaptive kernel [Lepski,1999], adaptive splines [Mammen&Van De Geer,2001]
  - a.k.a, multiscale, multi-resolution compression, used in JPEG2000.
  - New comer: Trend filtering! [Steidl,2006; Kim et. al. 2009, Tibshirani, 2013; W.,Smola, Tibshirani, 2014]

Univariate trend filtering

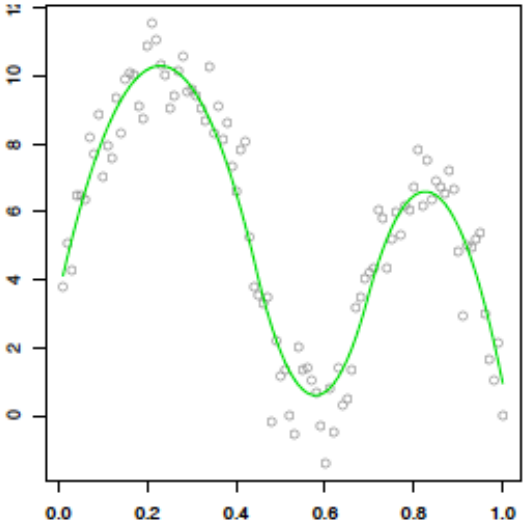
$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2}\|y - \beta\|_2^2 + \lambda \|D^{(k+1)}\beta\|_1$$



Constant, $k = 0$
(Fused lasso)

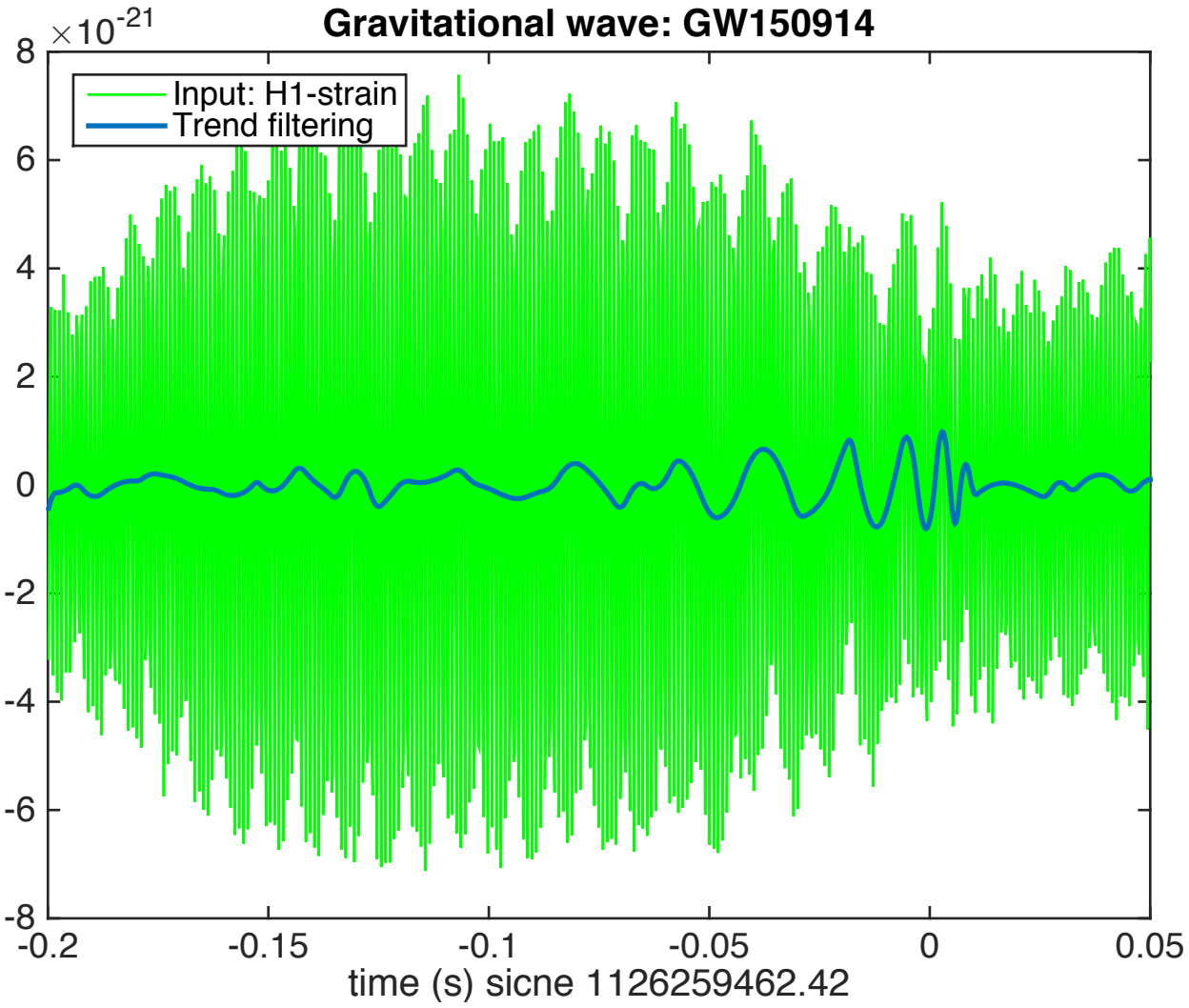


Linear, $k = 1$

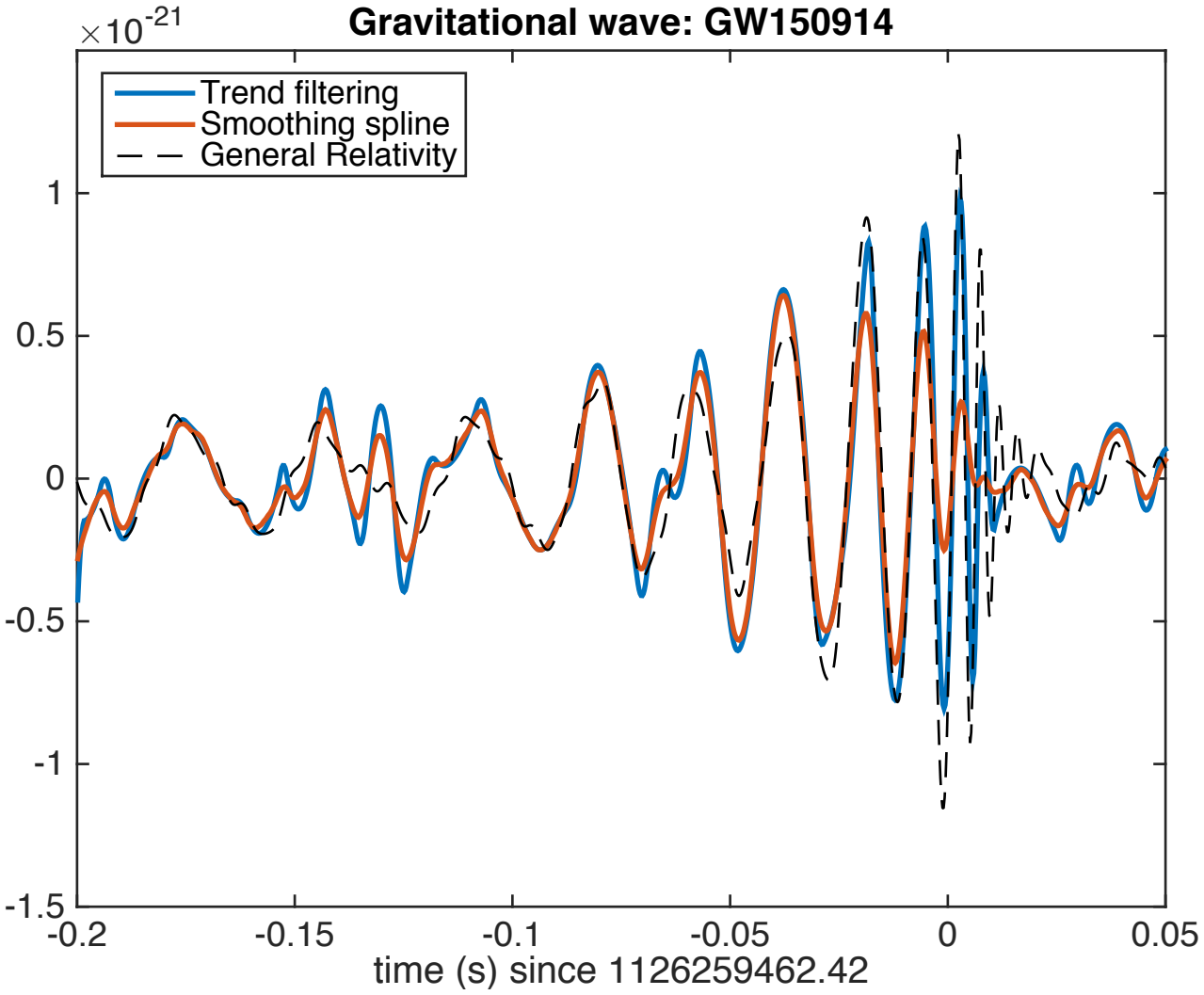


Quadratic, $k = 2$

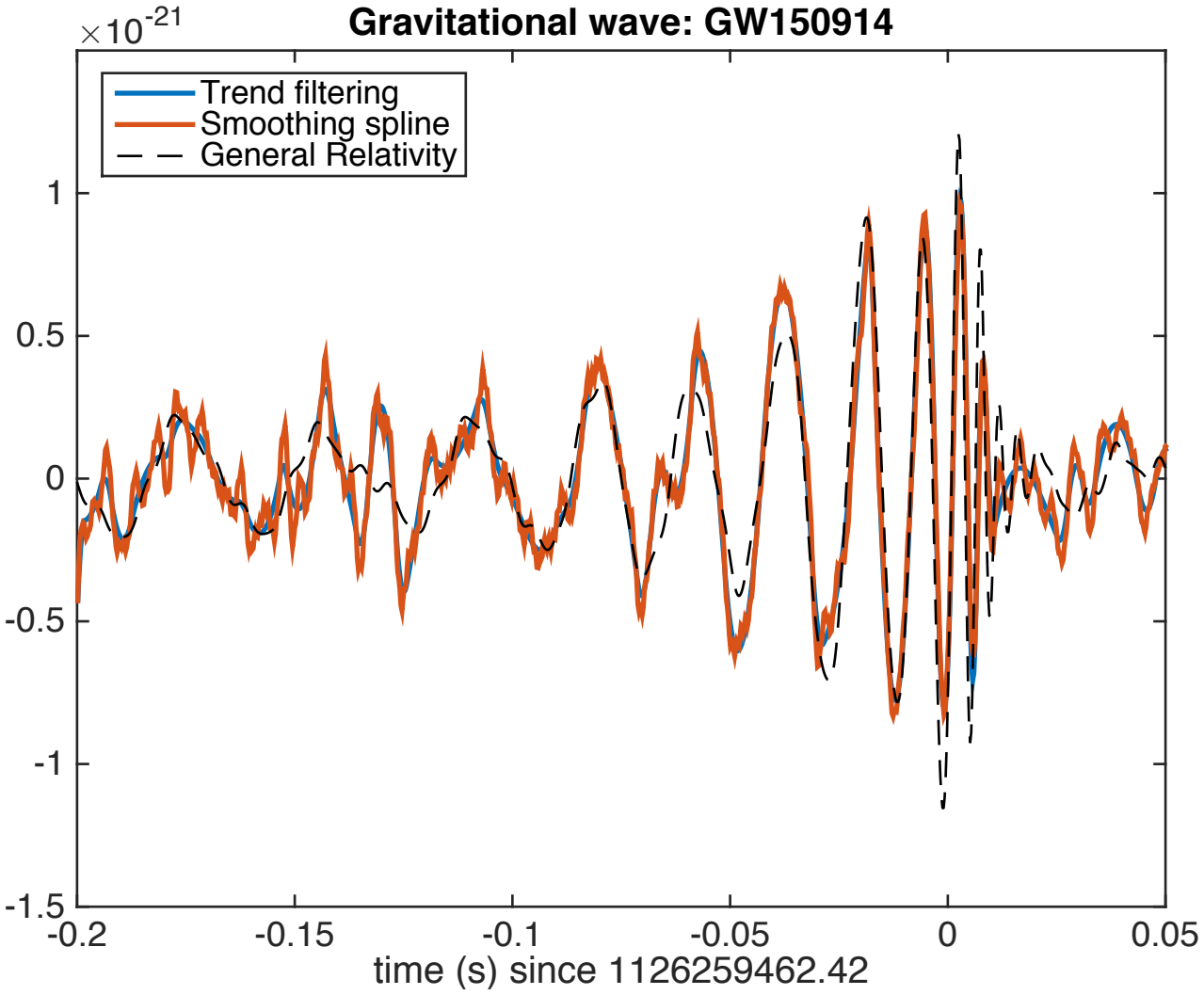
A BIG Example: merger of two black holes



A BIG Example: merger of two black holes



A BIG Example: merger of two black holes



Theory behind trend filtering

- Observations:

$$y_i = f_0(x_i) + \epsilon_i, \quad i = 1, \ldots n$$

- TV-class:

$$\mathcal{F}_k = \{f : \mathrm{TV}(f^{(k)}) \leq C\}$$

- Error rate:

$$O_{\mathbb{P}}(n^{-(2k+2)/(2k+3)})$$

- Drawback: only in for univariate functions.

Contributions

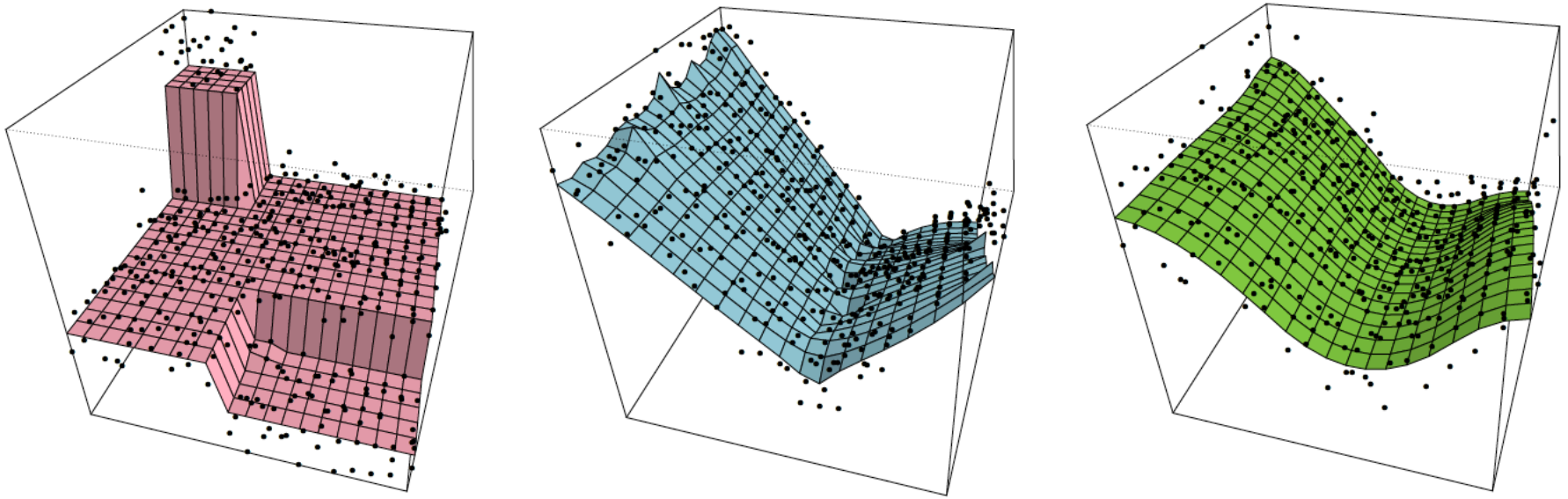
- Trend Filtering on Graphs [W., Sharpnack, Smola, Tibshirani, 2014]
  - Method, properties, algorithm, applications
  - Error bounds depends on graph-theoretic quantities

- Minimax theory on d-dim lattice grids.
  - For k=0, [Sadhanala, W., Tibshirani, 2016]
  - For k > 0, d=2 [Sadhanala, W., Shaprnack, Tisbshirani, 2017]

Trend filtering on graphs

$$\min_{\theta \in \mathbb{R}^n} \|y - \theta\|_2^2 + \lambda \|\Delta^{(k+1)} \theta\|_1$$

$$\Delta^{(1)} = D, \;\; \Delta^{(2)} = L, \;\; \Delta^{(3)} = DL, \;\; \Delta^{(4)} = L^2, \;\; \ldots$$



Example: TV-denoising

Noisy image

Laplacian smoothing

TV denoising

$$\hat{\theta}^{\mathrm{LS}} = \arg\min_{\theta} \|\theta - y\|^2 + \lambda \|D\theta\|_2^2 = \underbrace{(\lambda D^T D + I)^{-1} y}_{\text{a linear smoother}}$$

$$\hat{\theta}^{\mathrm{TV}} = \arg\min_{\theta} \|\theta - y\|^2 + \lambda \|D\theta\|_1 \text{ — not a linear smoother}$$

Minimax theory of GTF on grids

An **estimator** $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^n$ that takes in $\theta_0 +$ i.i.d. Gaussian noise and produces an estimator.

Mean square error:

$$\mathrm{MSE}(\hat{\theta}, \theta_0) = \frac{1}{n}\|\hat{\theta} - \theta_0\|_2^2.$$

Minimax risk:

$$R(\mathcal{K}) = \min_{\hat{\theta}} \max_{\theta_0 \in \mathcal{K}} \mathbb{E}\big[\mathrm{MSE}(\hat{\theta}, \theta_0)\big].$$

Minimax linear risk:

$$R_L(\mathcal{K}) = \min_{\hat{\theta}\ \text{linear}} \max_{\theta_0 \in \mathcal{K}} \mathbb{E}\big[\mathrm{MSE}(\hat{\theta}, \theta_0)\big],$$

Discrete Sobolev and TV-Class

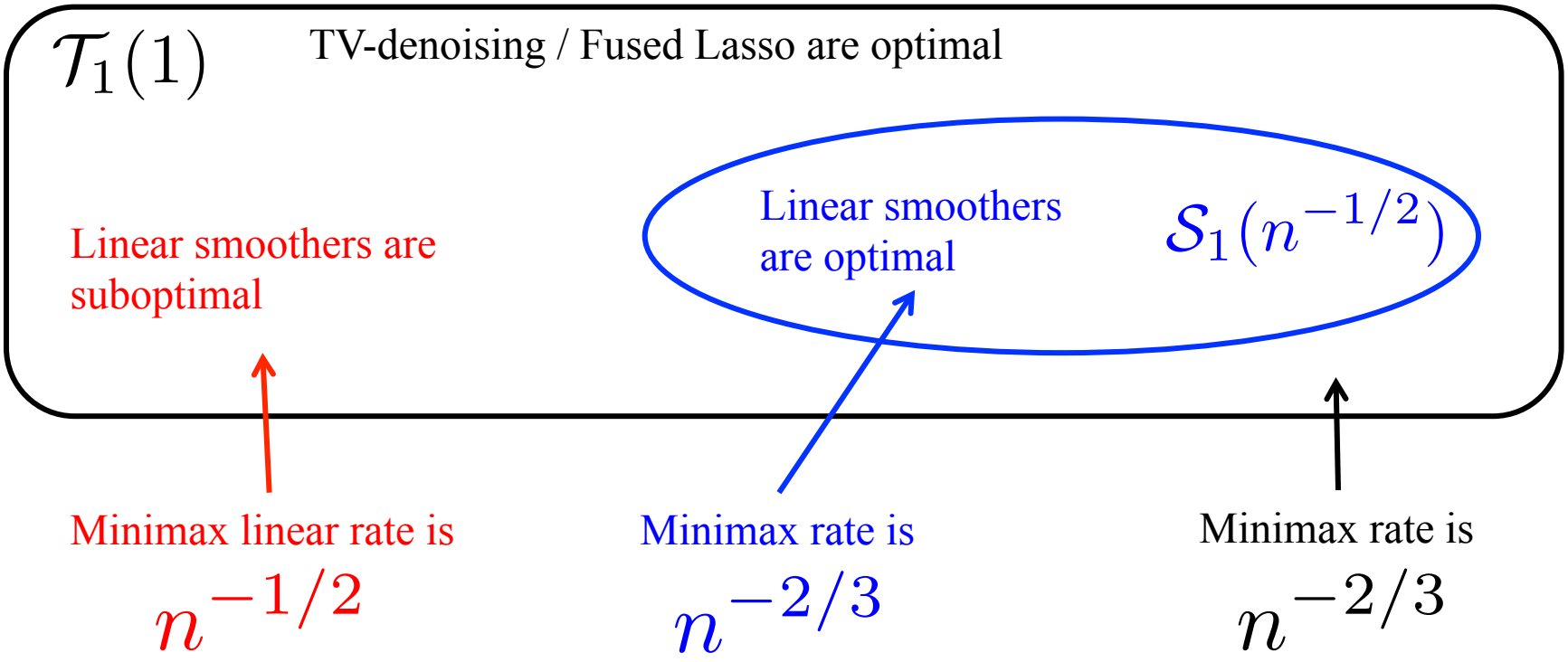
Define "function" classes

$$\text{TV Classes: } \mathcal{T}_d(C_n) = \{\theta : \|D\theta\|_1 \leq C_n\},$$

$$\text{Sobolev Classes: } \mathcal{S}_d(C_n') = \{\theta : \|D\theta\|_2 \leq C_n'\},$$

Where D is the incidence matrix for the d-dimensional grid graph with a total of n vertices.

Summary of known results in 1D



[Donoho, Liu, MacGibbon, 1994; Johnstone and Donoho, 1998]

Curse of dimensionality

- As d gets larger, on Sobolev space
- Classic nonparametric regression theory gives

$$n^{-\frac{2k}{2k+d}}$$

- We should expect the rate to get worse as d increase.

A surprising upper bound

[Hutter and Rigollet, 2016]

Theorem (Hütter and Rigollet, 2016): Total variation denoising estimator obeys

$$\mathrm{MSE}(\hat{\theta}^{\mathrm{TV}}, \theta_0) = O_{\mathbb{P}}\left(\frac{C_n \log n}{n}\right) \text{ for } d = 2,$$

$$\mathrm{MSE}(\hat{\theta}^{\mathrm{TV}}, \theta_0) = O_{\mathbb{P}}\left(\frac{C_n \sqrt{\log n}}{n}\right) \text{ for } d \geq 3,$$

Is this too good to be true?

Where did the curse-of-dimensionality go?

An even more surprising upper bound

Lemma (Sadhanala, W. and Tibshirani, 2016): A trivial estimator $\hat{\theta}^{\text{mean}}$ that outputs $\bar{y}\mathbb{1}$ obeys

$$\sup_{\theta_0 \in \mathcal{F}(C_n)} \mathbb{E}[\text{MSE}(\hat{\theta}^{\text{mean}}, \theta_0)] = O\left(\frac{\sigma^2 + C_n^2 \log n}{n}\right)$$

- This is a linear smoother!

Matching lower bounds for both surprising upper bounds

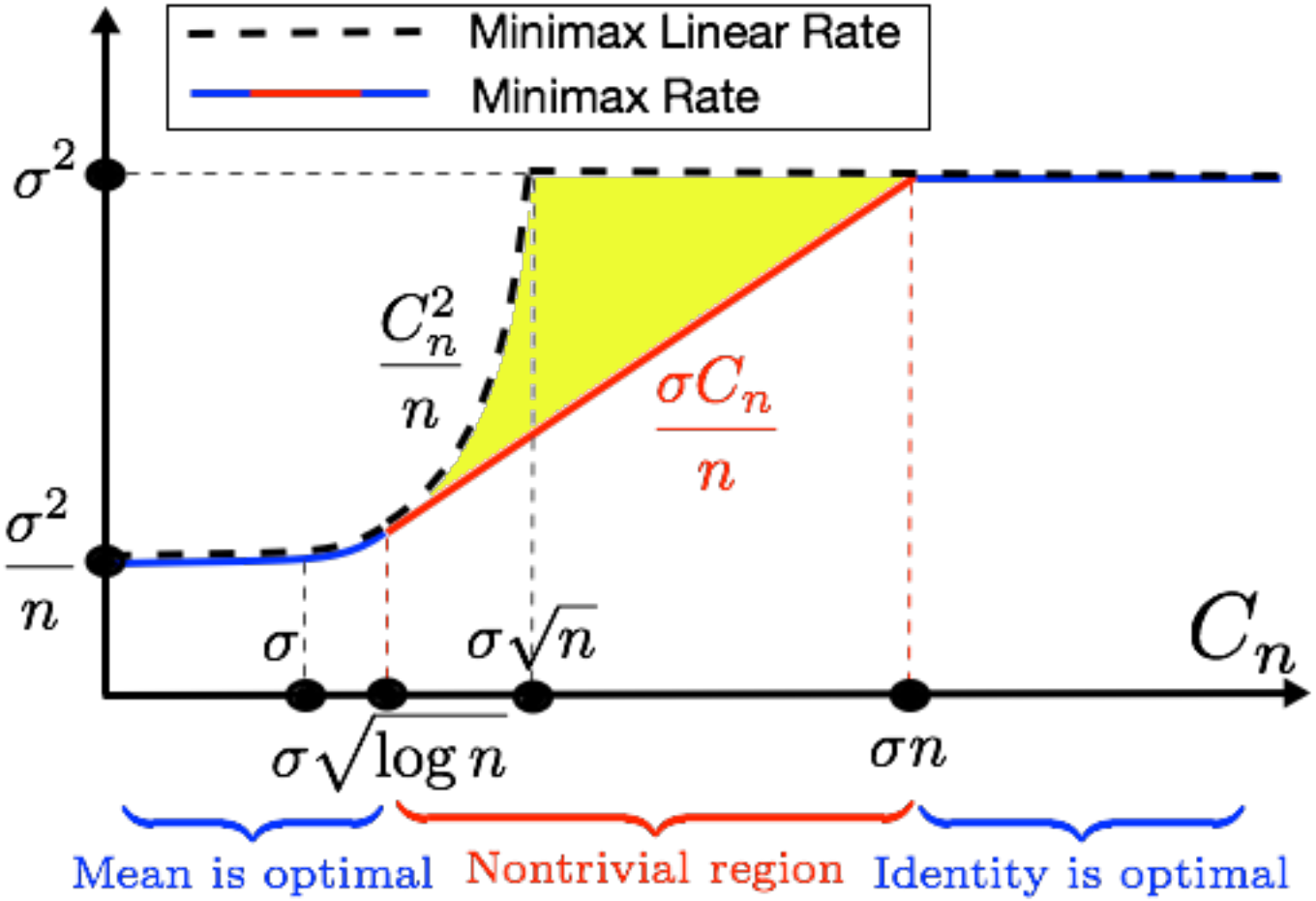
Theorem (Sadhanala, W. and Tibshirani, 2016): For constant d, and nontrivial region of C_n:

$$R(\mathcal{T}_d(C_n)) \asymp \frac{\sigma^2 + \sigma C_n}{n}.$$

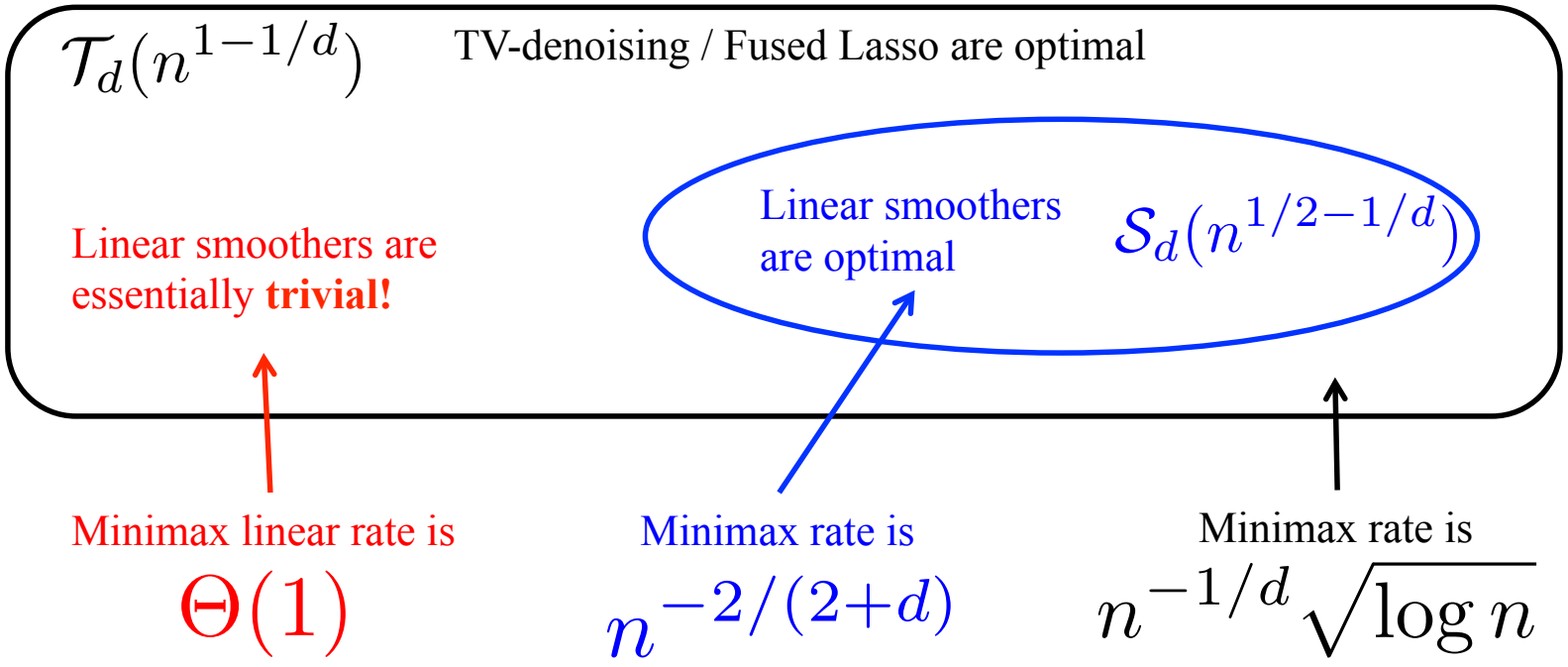
$$R_L(\mathcal{T}_d(C_n)) \asymp \frac{\sigma^2 + C_n^2}{n}.$$

- TV-denoising is optimal.
- No linear smoother can outperform the mean estimator.

How do we make sense of all these?



Rates under canonical scaling



- **Constant minimax linear rate!**
- **Starting d>2, we do not get local adaptivity for free!**

A few notes about proof techniques

- Upper bounds:
  - d-dim grid's Laplacian matrix is structured and can be diagonalized by DCT
  - Prove that D is incoherent
  - Careful calculation and thresholding the spectrum

- Lower bounds:
  - Embedding L1 ball inside a TV-ball
  - Gaussian model selection (Berge and Massart, 2001)
  - Linear smoother lower bounds: use and quadratically convex hull technique (Donoho, Liu, MacGibbon, 1990)

Summary of Part II

	d=1	d=2	d>2
k=0	$n^{-2/3}$	$n^{-2/4}$	$n^{-\frac{1}{d}}$
k=1	$n^{-4/5}$	$n^{-4/6}$	?
k>1	$n^{-\frac{2k+2}{2k+3}}$	$n^{-\frac{2k+2}{2k+4}}$	?

(Sadhanala, W., Tibshirani, 2016)

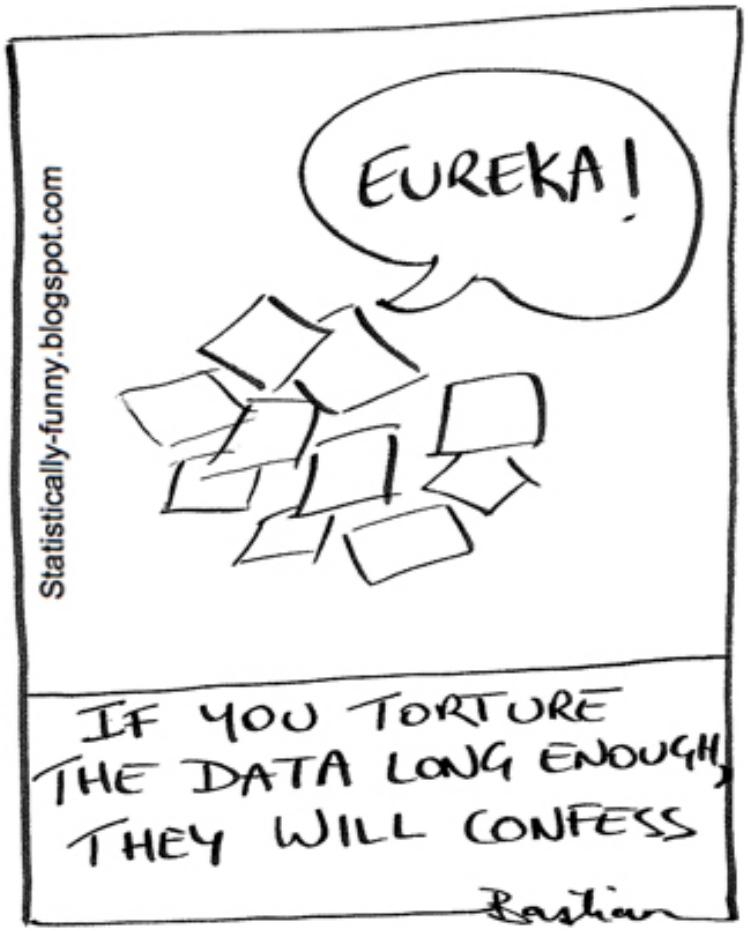
Special:
locally adaptivity
comes with
a statistical cost

Univariate TF
(Tibshirani, 2014)
(Mammen,Van De Geer, 2001)

(Sadhanala, W., Tibshirani, 2017)

Part III sequential learning

Data analysis is conditional/adaptive



- "All inferences are conditional inferences."
  - Jonathan Taylor (via Ryan)
- "Why most published research findings are false?"
  - John Ioannidis, 2005
- "A garden of forking paths"
  - Gelman and Loken, 2013

Related work

- ADA via Differential privacy (DFHPRR15, BNSSSU15, etc…)
  - Similar setting. DP is unnecessarily strong for the purpose. Need low-sensitivity.
  - We work with conditional expectations directly.

- Lower bounds via finger printing codes (Hardt, Ullman, Steinke, etc)
  - Uniform prior => discrete distribution of atoms (fingerprinting coe).
  - Reconstruction attack of iid data

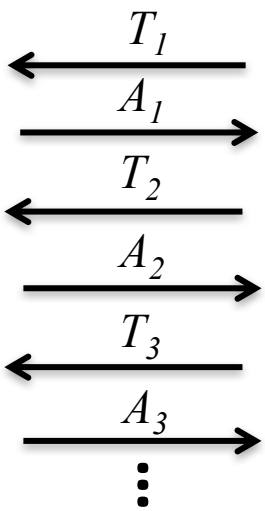
- Bayesian Adaptive Data Analysis (Elder, 2017)
  - The player and adversary share the same prior. Information symmetry.

- Post-selection inference (Taylor, Tibshirani, Fithian, Lee, etc.)
  - The focus is to have correct confidence interval, despite selection bias.
  - Fixed procedure, lasso-like. Not adaptive.
  - We prevent finding significantly biased statistics in the first place.

Gaussian adaptive data analysis

$\phi_T \sim N(\mu_T, \Sigma)$

$T_1, \ldots, T_k \in \mathcal{T}$



Player

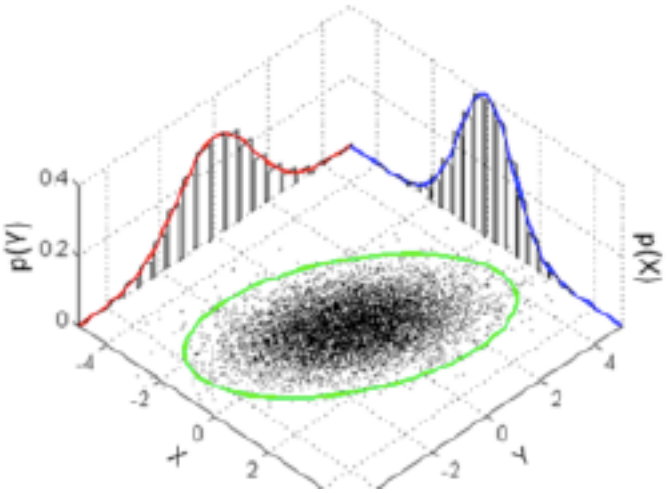
I have the data but not the distribution.
I choose how to answer the questions.

Adversary

I have the distribution.
I choose questions T.

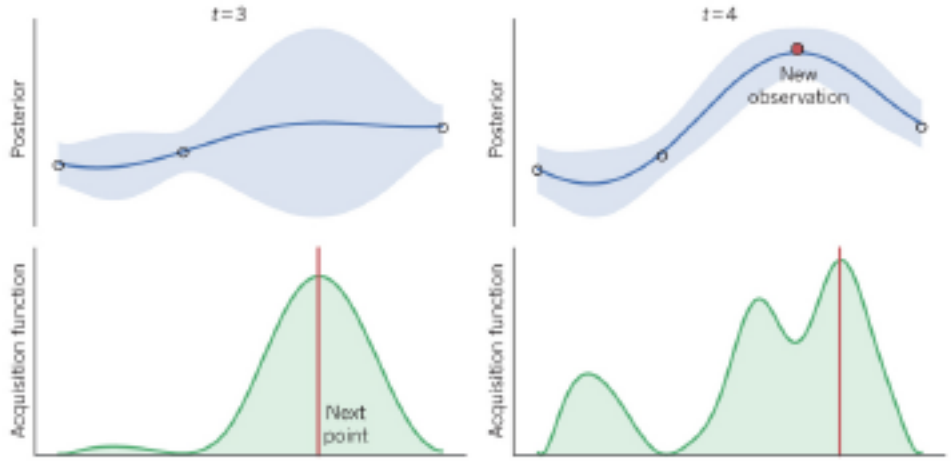
Russo and Zou. "Controlling Bias in Adaptive Data Analysis Using Information Theory." AISTATS-2016.

◇ Example: Unit projection of multivariate Gaussian.



- $X \sim \mathcal{N}(\mu, \sigma^2 I_d)$
- $\mathcal{T} = \left\{t \in \mathbb{R}^d : \|t\|_2 \leq 1\right\}$ is the class of all unit vectors
- $\phi_t(X) = \langle t, X \rangle$.
- For any $t \in \mathcal{T}$, $\mathrm{Var}(\phi_t(X)) \leq \frac{\sigma^2}{n}$.

◇ Example: Bayesian optimization for hyperparameter tuning.



- X validation set.
- $\mathcal{T} = [0,1]^d$, d-dimensional hyperparameter.
- $\phi_t(X)$: Validation error of the fitted model with hyperparameter t.
- $\phi_{\mathcal{T}}$ is assumed to be a Gaussian process.

Main results: a minimax lower bound

Theorem 1 (Unknown distribution) *Assume* $|\mathcal{T}| > k - 1 + 2^{k-1}$ *and* $\mathcal{D}$ *can induce distribution* $\phi_{\mathcal{T}} \sim \mathcal{N}(\mu, \Sigma)$ *for any* μ, Σ *satisfying* $\Sigma_{i,i} \leq \sigma^2$. *Then*

$$\inf_{A_{1:k}} \sup_{\mathcal{D}(\phi_{\mathcal{T}})} \sup_{T_{1:k}} \left(\max_i \mathbb{E}[(A_i - \mu_{T_i})^2] \right) = \Omega(\sqrt{k}\sigma^2)$$

- Plug-in estimators: $k\sigma^2$
- Independent noise adding: $\sqrt{k}\sigma^2$

Our result says: "Independent noise adding is rate-optimal."

Per-instance lower bound

Theorem 2 (Fixed distribution) *For any fixed pair of $(\mathcal{D}, \mathcal{T})$ that obeys the same joint Gaussian assumption, and in addition are sufficiently rich. Then*

$$\inf_{\text{Natural } A_{1:k}} \sup_{T_{1:k}} \left(\max_i \mathbb{E}[(A_i - \mu_{T_i})^2] \right) = \Omega(\sqrt{k}\sigma^2).$$

- The same lower bound holds for each data distribution separately.
- If we restrict the class of player strategies somewhat.
- The hardness is dense within the class!

A note on the lower bound construction

- Approximately least-favorable Prior:
  - Uniform or Gaussian prior on the mean.
  - Some structured correlations
- Adversary strategy:
  - Explore first, then exploit.
  - Sign inference attack.
- Player strategy:
  - Posterior mean is optimal for square loss.
  - Optimal noise for obfuscating sign inference attack..

Summary Part III: still a long way to go

- Lower bound does not handle iid data
- In Gaussian projection:
  - Upper bound is meaningful up to k= d^2.
  - Lower bound is tight only up to k = d.
- Beyond joint Gaussian models.

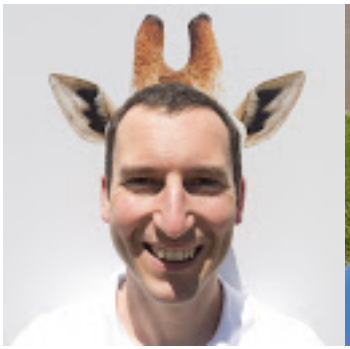
Conclusion

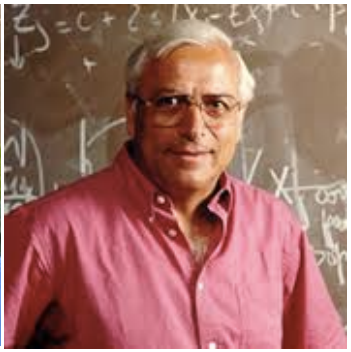
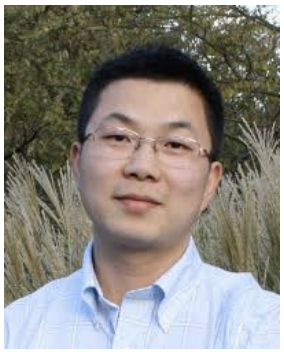
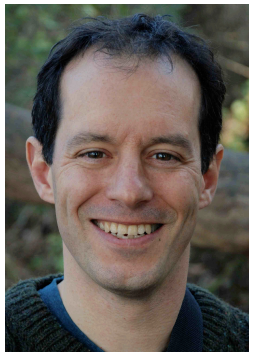
- Data/Privacy challenge:
  - Differentially private machine learning
  - Slowly transforming DP into a practical technology
- Modeling challenge:
  - locally adaptive function classes
  - And their minimax estimation error
- Sequential learning:
  - Some progress with a reasonably strong lowerbound
  - But a lot more problems to solve than solved

Future work

- Further pursue pDP. Build practical DP tools into:
  - my open source project: pyDiffPriv
- Investigating nonparametric postprocessing for DP-releases, and private nonparametric regression
- Sequential/Online Trend filtering
- Sequential estimation beyond joint Gaussian observations

Acknowledgments






Supplementary slides

Details on the direct analysis of OPS

Theorem 15 (The adaptivity of OPS in Linear/Ridge Regression)**.** *Consider the algorithm that samples from*

$$p(\theta|X,\mathbf{y}) \propto e^{-\frac{\gamma}{2}\left(\|\mathbf{y}-X\theta\|^2+\lambda\|\theta\|^2\right)}.$$

Let $\hat{\theta}$ and $\hat{\theta}'$ be the ridge regression estimate with data set $X \times \mathbf{y}$ and $[X,x] \times [\mathbf{y},y]$ and defined the out of sample leverage score $\mu := x^T(X^TX+\lambda I)^{-1}x = x^TH^{-1}x$ and in-sample leverage score $\mu' := x^T[(X')^TX'+\lambda I]^{-1}x = x^T(H')^{-1}x$. Then for every $\delta > 0$, privacy target (x,y), the algorithm is (ϵ,δ)-pDP with

$$\epsilon(Z,z) \leq \frac{1}{2}\left|-\log(1+\mu)+\frac{\gamma\mu}{(1+\mu)}(y-x^T\hat{\theta})^2\right| + \frac{\mu}{2}\log(2/\delta) + \sqrt{\gamma\mu\log(2/\delta)}|y-x^T\hat{\theta}| \tag{5}$$

$$= \frac{1}{2}\left|-\log(1-\mu')-\frac{\gamma\mu'}{1-\mu'}(y-x^T\hat{\theta}')^2\right| + \frac{\mu'}{2}\log(2/\delta) + \sqrt{\gamma\mu'\log(2/\delta)}|y-x^T\hat{\theta}'|. \tag{6}$$

Information-theoretic lower bound

- **Due to [Bassily et. al., 14]**
  - Lipschitz $\sqrt{d}/\epsilon$
  - Lipschitz and Strongly convex $d/(\epsilon^2 \lambda_{\min})$

- We match both

$$F(\tilde{\theta}) - F(\theta^*) \leq \frac{d \log(d/\delta) \log(2/\delta)}{[\lambda + \lambda_{\min}(X^T X)]\epsilon^2} + \lambda \|\theta^*\|_2^2$$

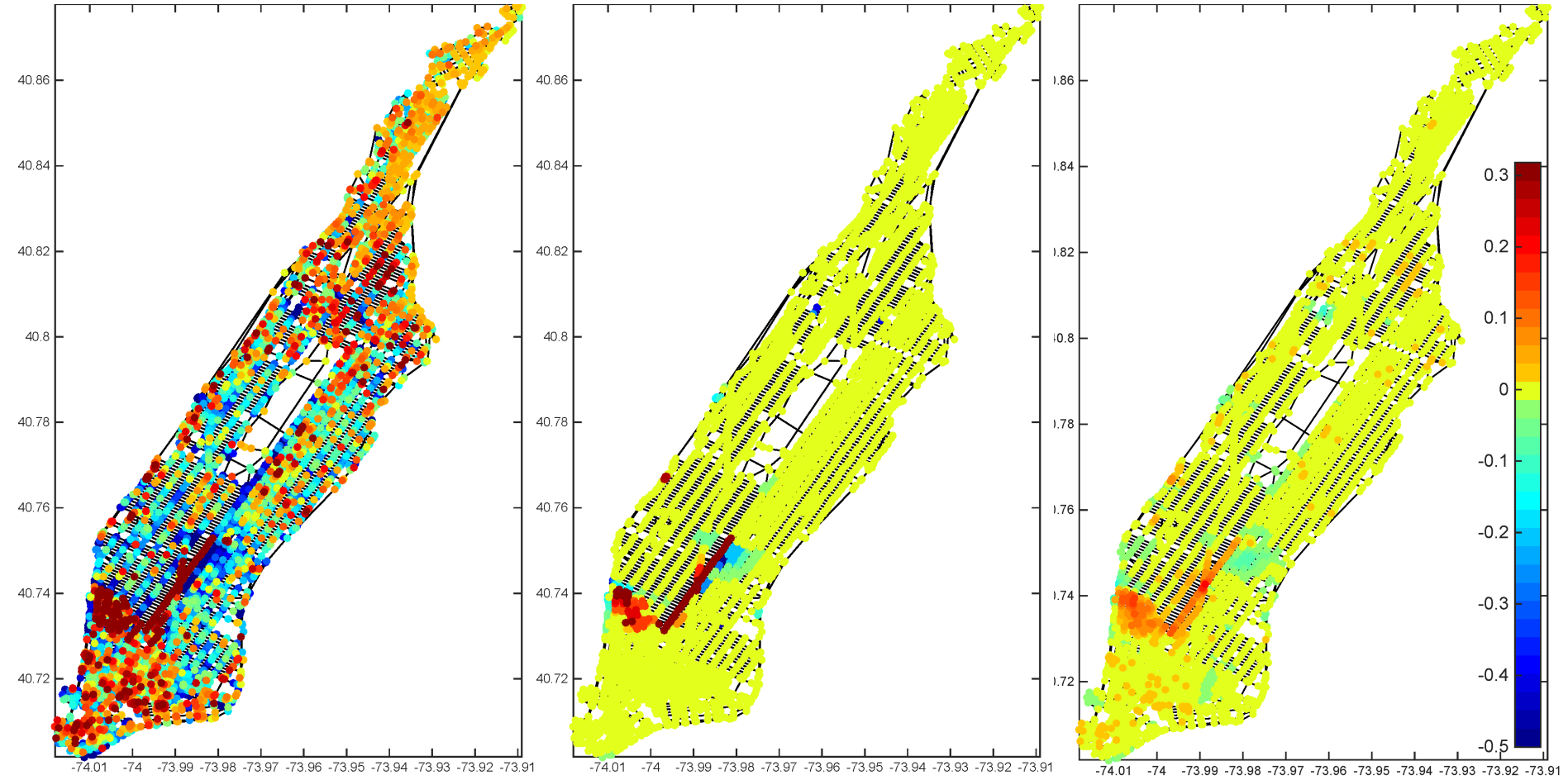
In fact, we only need local Lipschitz.

Variants of differential privacy

	Data set	private target	probability metric	parametrized by
Pure-DP[12]	$\sup_Z$	$\sup_z$	$D_\infty(P\|\|Q)$	$\mathcal{A}$ only
Approx-DP[10]	$\sup_Z$	$\sup_z$	$D_\infty^\delta(P\|\|Q)$	$\mathcal{A}$ only
(z/m)-CDP[14, 5]	$\sup_Z$	$\sup_z$	$D_{\text{subG}}(P\|\|Q)$	$\mathcal{A}$ only
Rényi-DP[25]	$\sup_Z$	$\sup_z$	$D_\alpha(P\|\|Q)$	$\mathcal{A}$ only
Personal-DP[16, 21]	$\sup_Z$	fixed z	$D_\infty^\delta(P\|\|Q)$	$\mathcal{A}$ and z
TV-privacy[2]	$\sup_Z$	$\sup_z$	$\|P-Q\|_{TV}$	$\mathcal{A}$ only
KL-privacy[2]	$\sup_Z$	$\sup_z$	$D_{KL}(P\|\|Q)$	$\mathcal{A}$ only
On-Avg KL-privacy[33]	$\mathbb{E}_{Z\sim\mathcal{D}^n}$	$\mathbb{E}_{z\sim\mathcal{D}}$	$D_{KL}(P\|\|Q)$	$\mathcal{A}$ and $\mathcal{D}$
Per-instance DP	fixed Z	fixed z	$D_\infty^\delta(P\|\|Q)$	$\mathcal{A}$, Z and z

Table 1: Comparing variances of differential privacy.

Example: New York City Taxi data



Data science for social good?

Privacy law

Food safety

Understand how brain works

Better treatment

Better education

Medical diagnosis

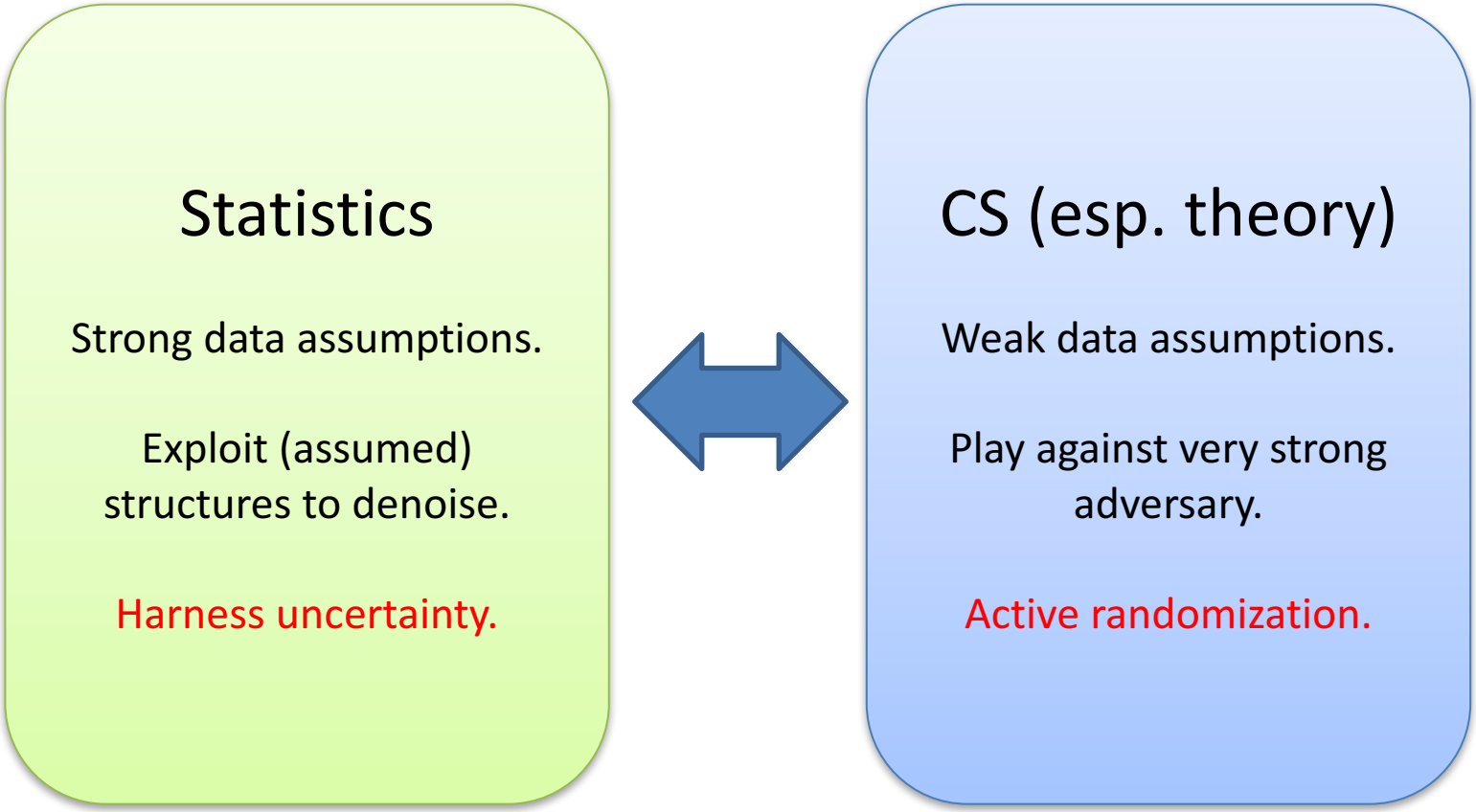
Clean energy

Epidemic forecast/prevention

Fraud detection

Increasing privacy awareness

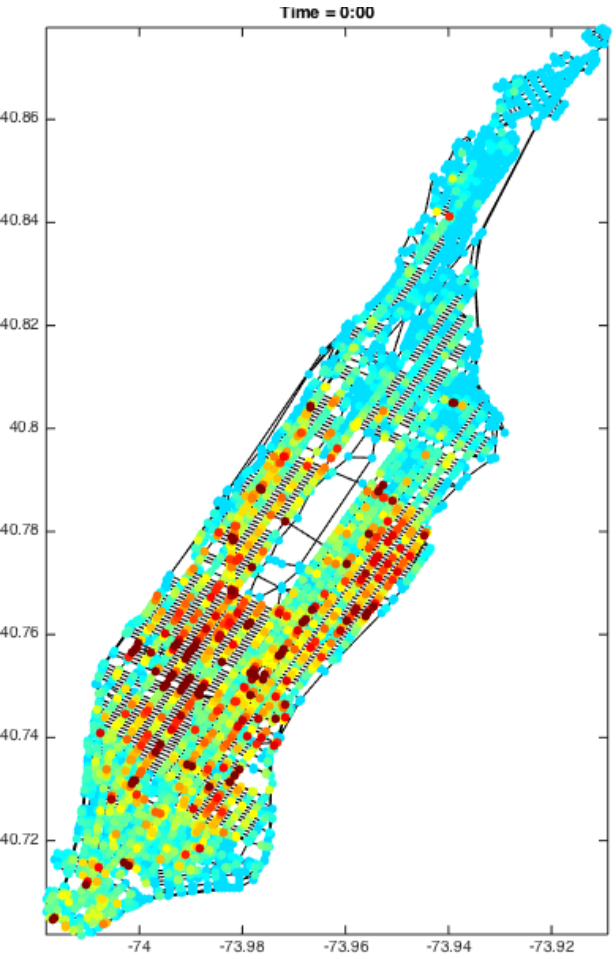
Interplay of Statistics and CS



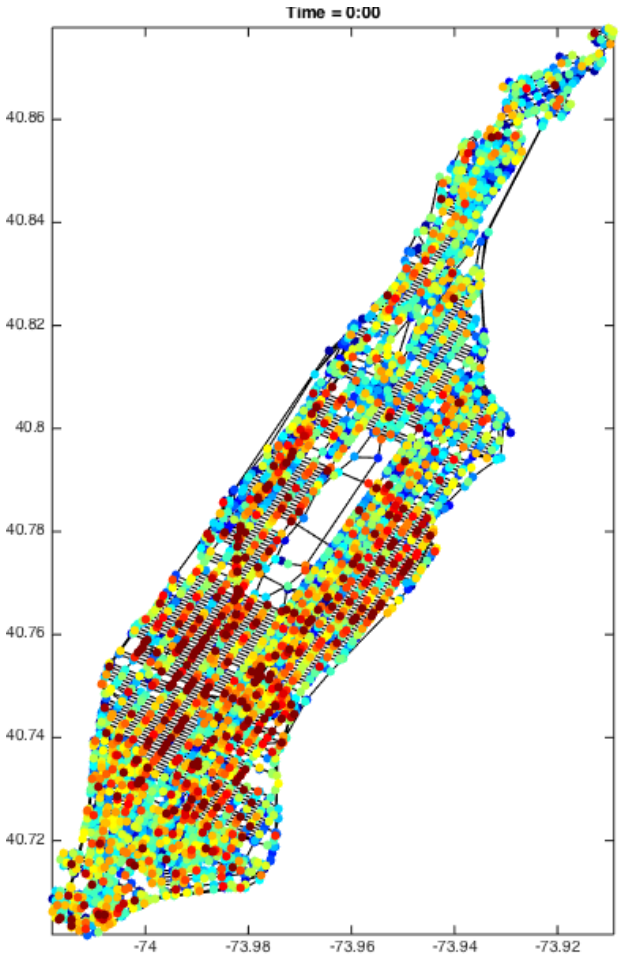
Not satisfactory because assumptions often not true.

Also not satisfactory because Guarantees are too conservative.

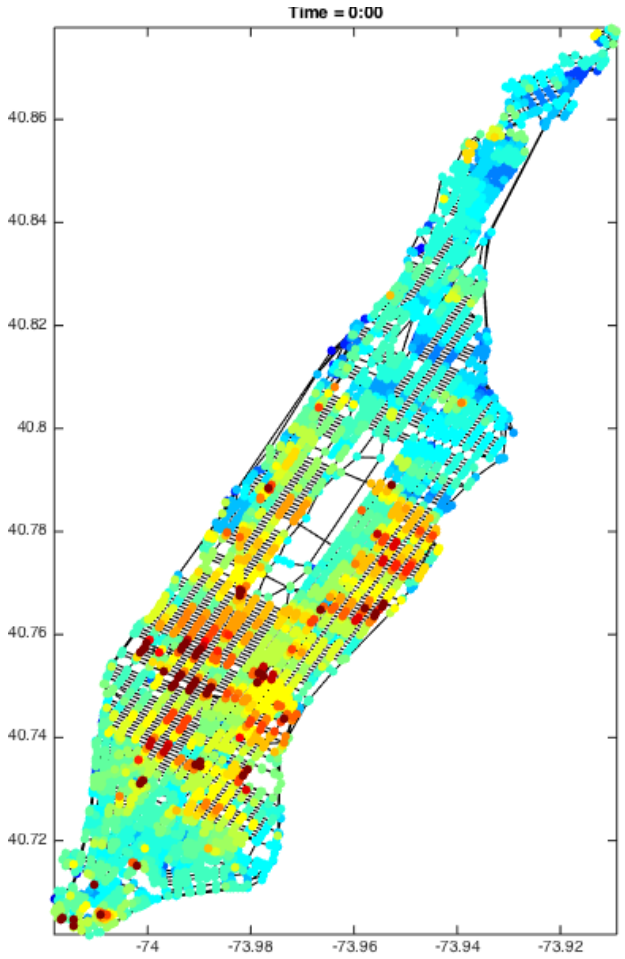
Example of Graph Trend Filtering (WSST-15) on DP release of spatial statistics



Original data

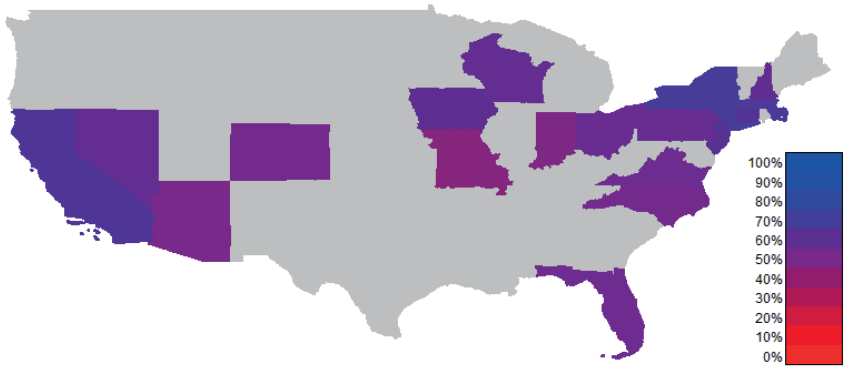


DP release

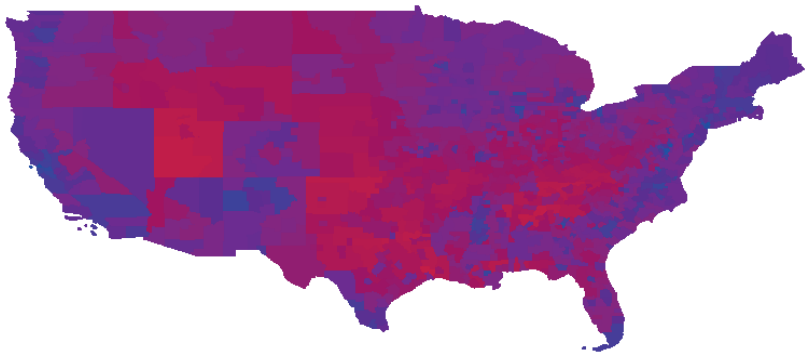


Denoised DP release

Example of Ecological inference (used only county-Level aggregates)



(a) Exit poll results for women



(c) Ecological regression results for women

- Input: Vote proportions + US Census microdata.
- Output: Subpopulation prediction/inferences.

Flaxman, Sutherland, W. and Teh. "Understanding the 2016 US Presidential Election using ecological inference and distribution regression with census microdata." manuscript (2016).

Flaxman, W. and Smola. "Who supported Obama in 2012?: Ecological inference through distribution regression." KDD'2015. **Best Student Paper.**