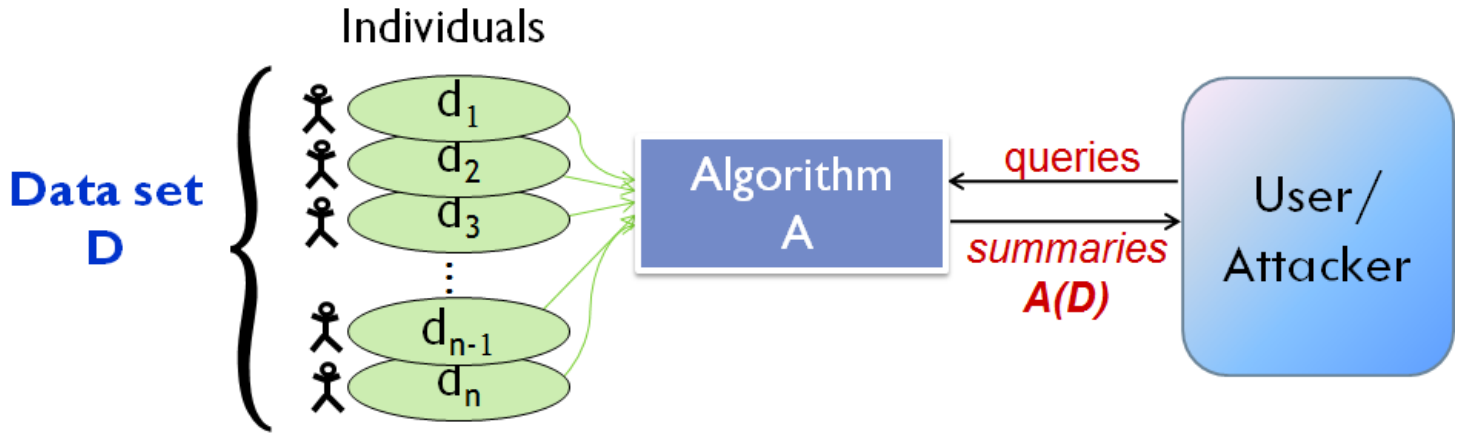
Private Model Selection via Stability Arguments and the Robustness of Lasso

Adam Smith [1] Abhradeep Thakurta[2]

[1]Pennsylvania State University

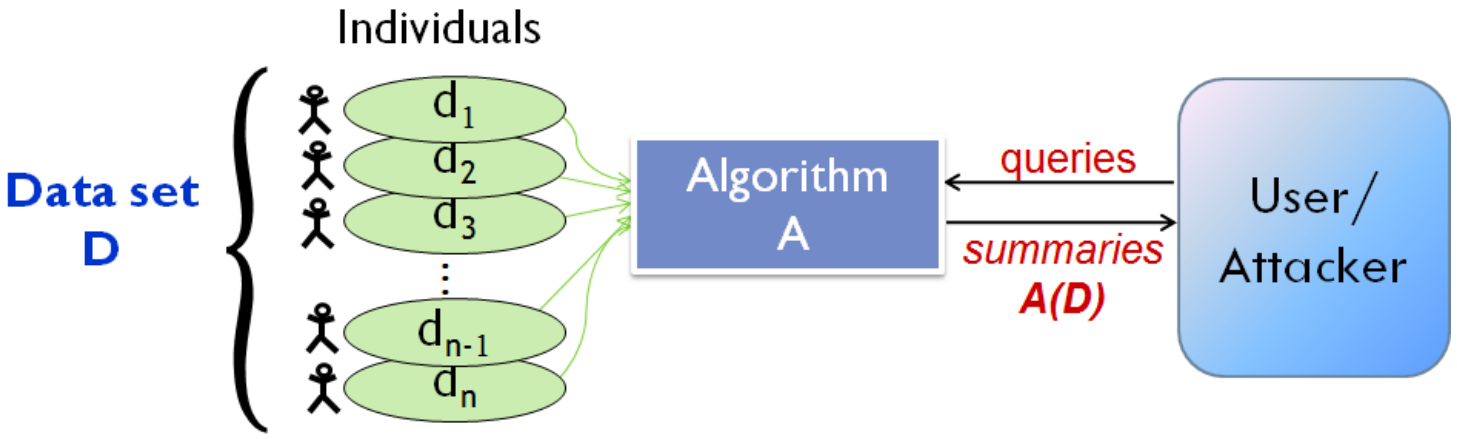
[2]Stanford University and Microsoft Research SVC

What Can We Learn Privately?



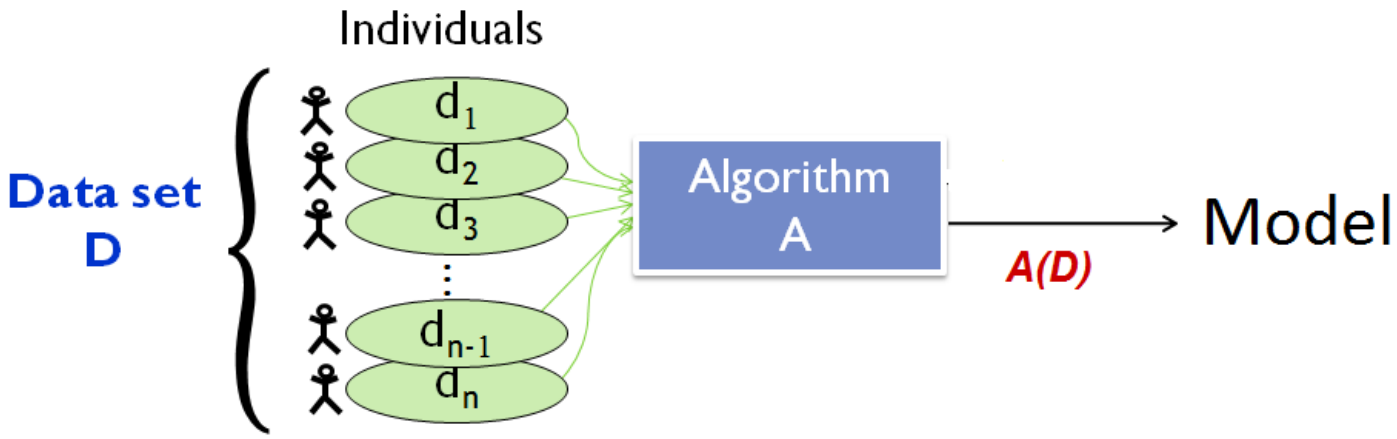
- Consider data set of sensitive individual data (e.g, medical records, purchase history)
- **Two conflicting goals:**
  - **Utility:** Release global information about $\mathcal{D}$
  - **Privacy:** Protect the privacy of individual entries

What Can We Learn Privately?



- Consider data set of sensitive individual data (e.g, medical records, purchase history)
- **Difficult problem:** Many attacks on "anonymized" data [NS08,Korolova11,CKN11,...]
- **Active research area:** Databases, Learning Theory, Programming Languages,Cryptography, Algorithms...

Model Seleciton



Models: Discrete collection of probability distribution families

$$M_1, \quad , M_2, \quad \cdots$$

Goal: Select a model that best "fits" the data set $\mathcal{D}$

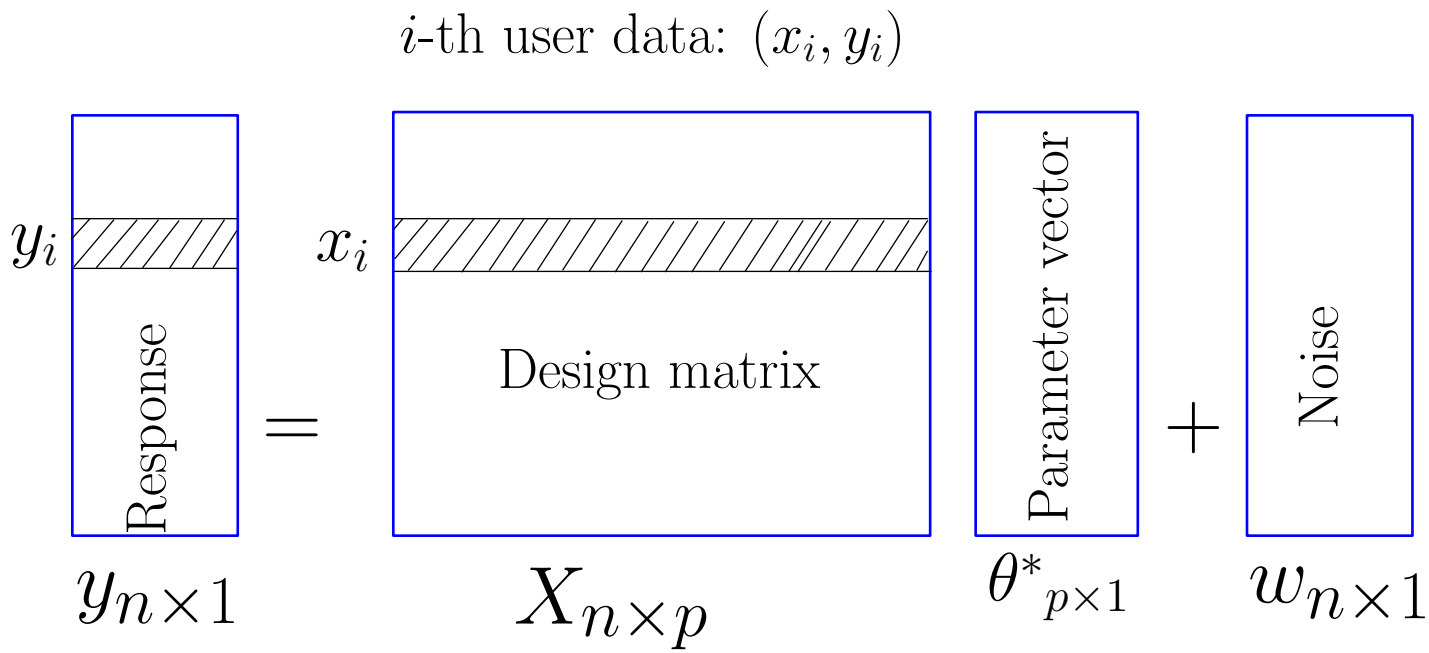
Example: Model could be a set of features

Running Example: Sparse Linear Regression

- Data point $d_i = (x_i, y_i)$, where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$

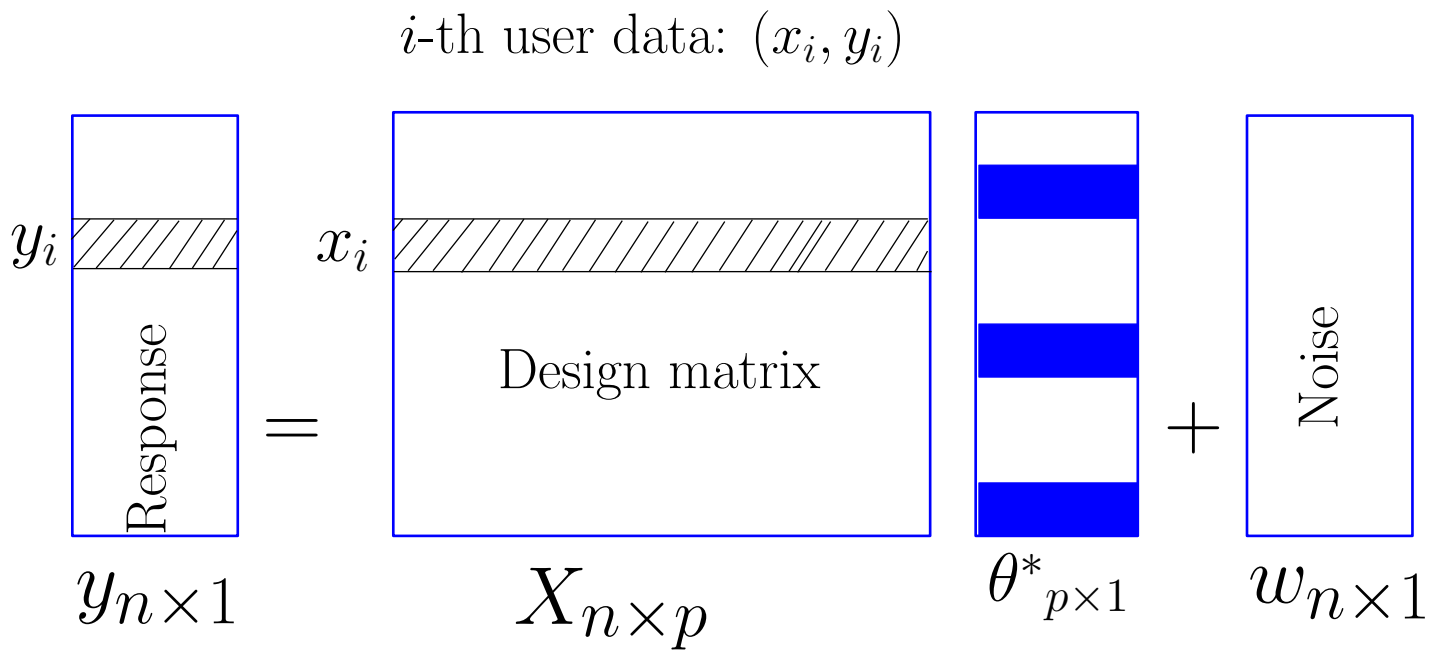
Running Example: Sparse Linear Regression

- Data point $\mathtt{d_i} = (\mathtt{x_i}, \mathtt{y_i})$, where $\mathtt{x_i} \in \mathbb{R}^{\mathtt{p}}$ and $\mathtt{y_i} \in \mathbb{R}$
- Model assumption: Data generated by noisy linear system



Running Example: Sparse Linear Regression

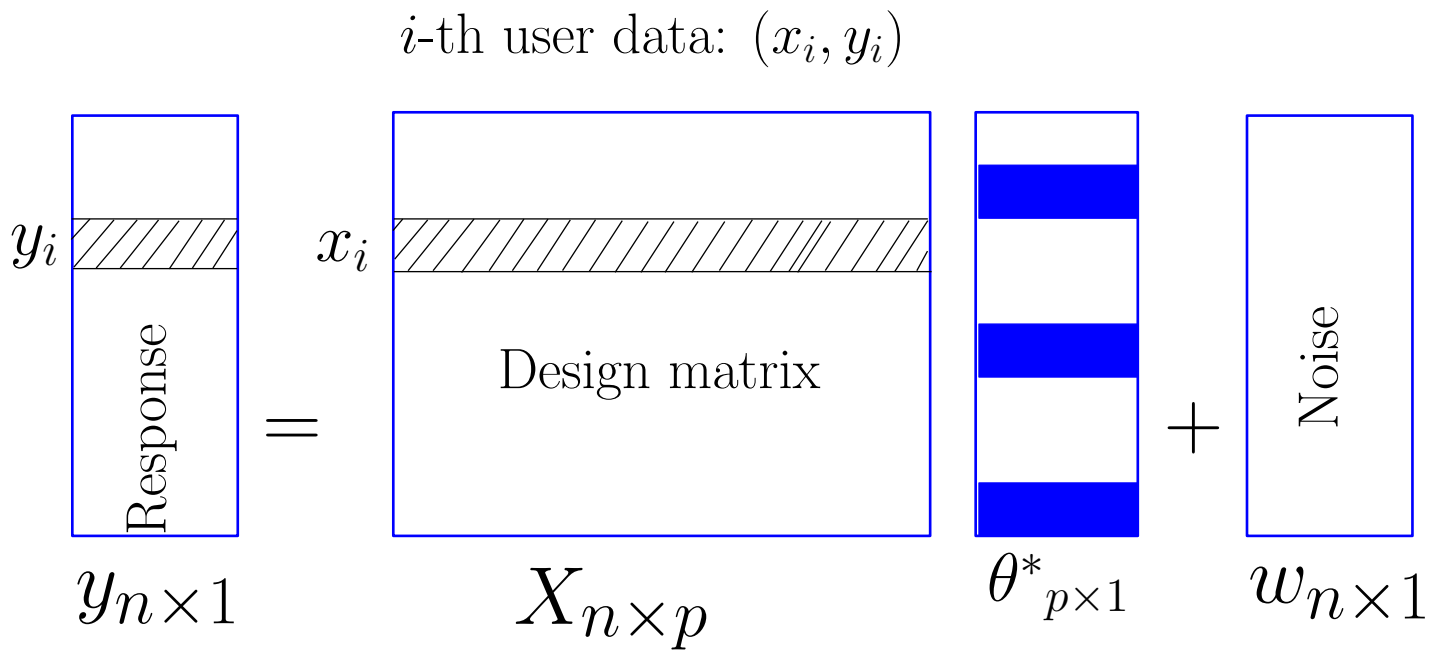
- Data point $\mathtt{d_i} = (\mathtt{x_i}, \mathtt{y_i})$, where $\mathtt{x_i} \in \mathbb{R}^{\mathtt{p}}$ and $\mathtt{y_i} \in \mathbb{R}$
- Model assumption: Data generated by noisy linear system



- **Sparsity:** θ^* has $s < n$ non-zero entries
- **High-dimensionality:** $p \gg n$

Running Example: Sparse Linear Regression

- Data point $\mathtt{d_i} = (\mathtt{x_i}, \mathtt{y_i})$, where $\mathtt{x_i} \in \mathbb{R}^{\mathtt{p}}$ and $\mathtt{y_i} \in \mathbb{R}$
- Model assumption: Data generated by noisy linear system

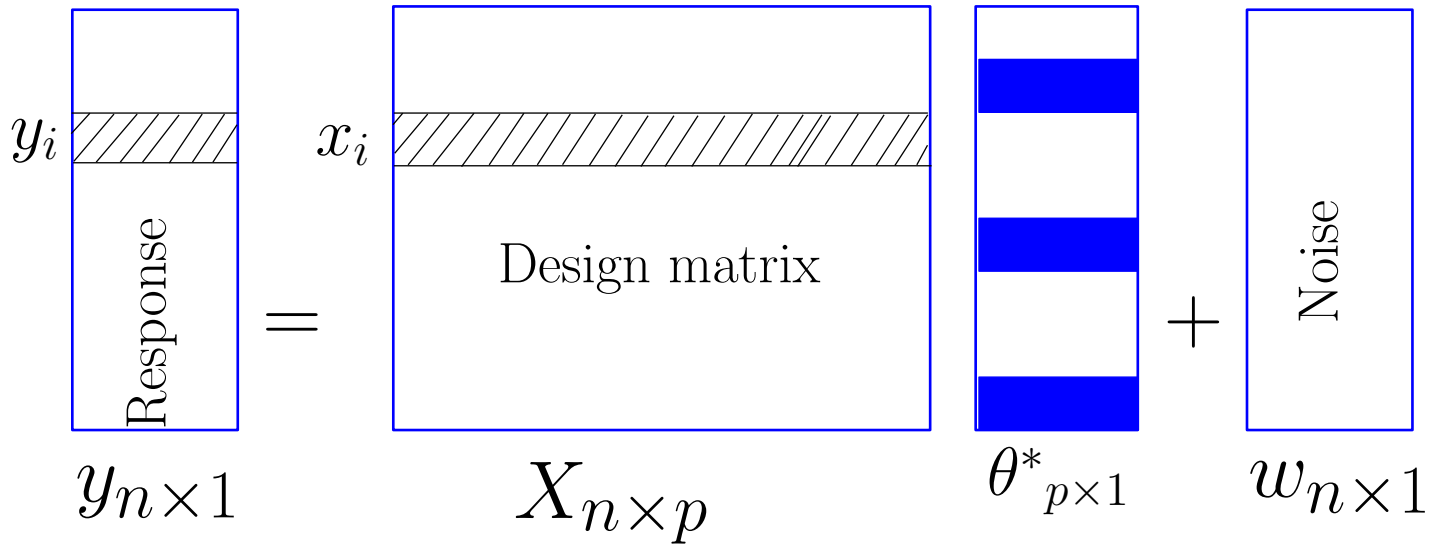


  - **Sparsity:** θ^* has $s < n$ non-zero entries
  - **High-dimensionality:** $p \gg n$

- Model selection problem: Find the support of θ^*

Running Example: Sparse Linear Regression

- Data point $\mathsf{d_i} = (\mathsf{x_i}, \mathsf{y_i})$, where $\mathsf{x_i} \in \mathbb{R}^{\mathsf{p}}$ and $\mathsf{y_i} \in \mathbb{R}$
- Model assumption: Data generated by noisy linear system



- Sparse linear regression: LASSO estimator

$$\hat{\theta} \in \arg\min_{\theta\in\mathbb{R}^p} \frac{1}{2n}\|y - X\theta\|_2^2 + \frac{\lambda}{n}\|\theta\|_1$$

This Paper:

- Differentially private model selection
  - Application: Sparse linear regression
- Stability analysis of LASSO

Why is Privacy a Concern in Model Selection?

Many model selection procedures use convex optimization

Examples of convex optimization problems:

- LASSO, mean, median and support vector machines (SVM)

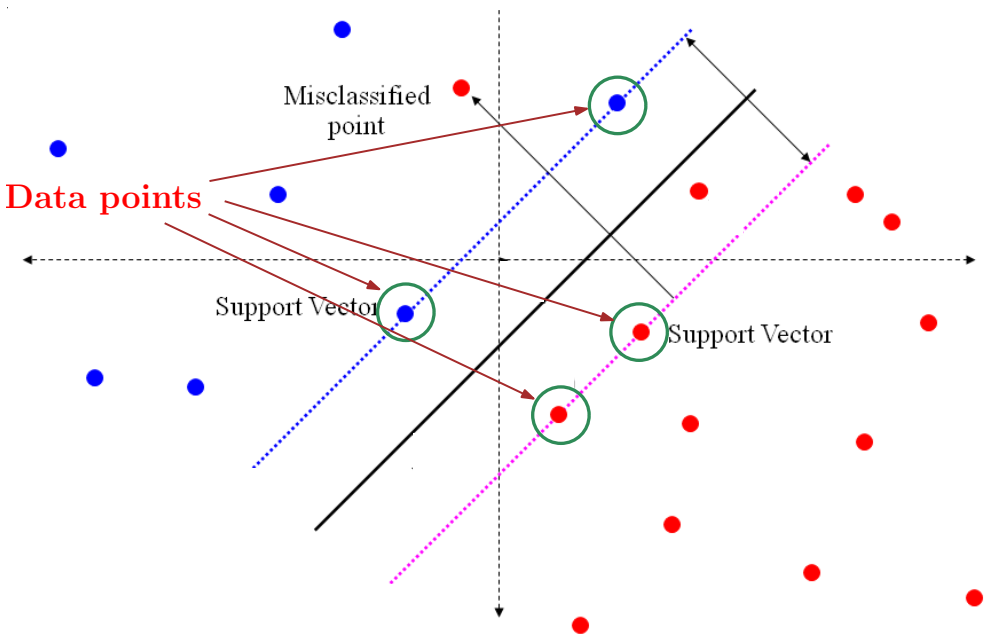
Why is Privacy a Concern in Model Selection?

Many model selection procedures use convex optimization

Examples of convex optimization problems:

- LASSO, mean, median and support vector machines (SVM)

Privacy breach: SVM example



- Reveals some data points exactly

Why is Privacy a Concern in Model Selection?

Many model selection procedures use convex optimization

Examples of convex optimization problems:

- LASSO, mean, median and support vector machines (SVM)

Question: How can we ensure privacy?

Why is Privacy a Concern in Model Selection?

Many model selection procedures use convex optimization

Examples of convex optimization problems:

- LASSO, mean, median and support vector machines (SVM)

Question: How can we ensure privacy?

Answer: Design algorithms that satisfy differential privacy

Differential Privacy [DMNS06,DKMMN06]

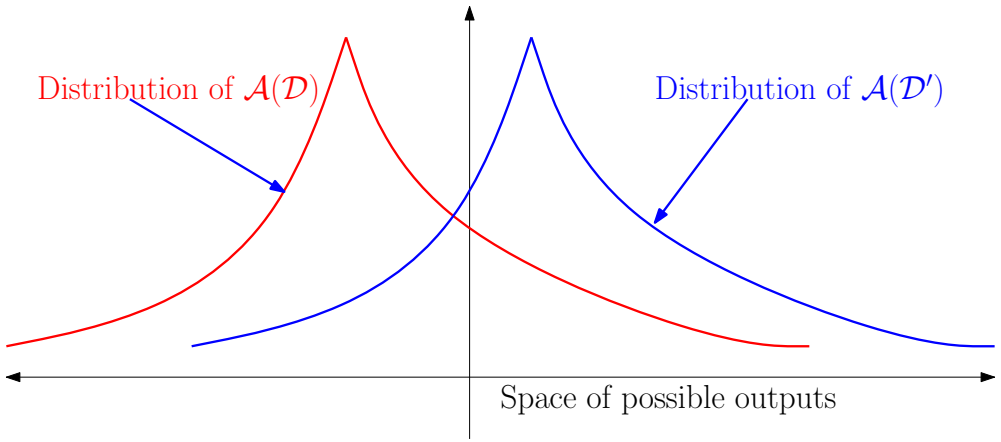
- Intuition: Users learn roughly the same things about you whether or not your data is in the data set
- It is a restriction on the algorithm $\mathcal{A}$
  - Guarantee privacy even with arbitrary side information

Differential Privacy [DMNS06,DKMMN06]

- Intuition: Users learn roughly the same things about you whether or not your data is in the data set
- It is a restriction on the algorithm $\mathcal{A}$
  - Guarantee privacy even with arbitrary side information

Requirement

Any two data sets $\mathcal{D}$ and $\mathcal{D}'$ that differ in one element induce close distributions on the space of outputs of $\mathcal{A}$

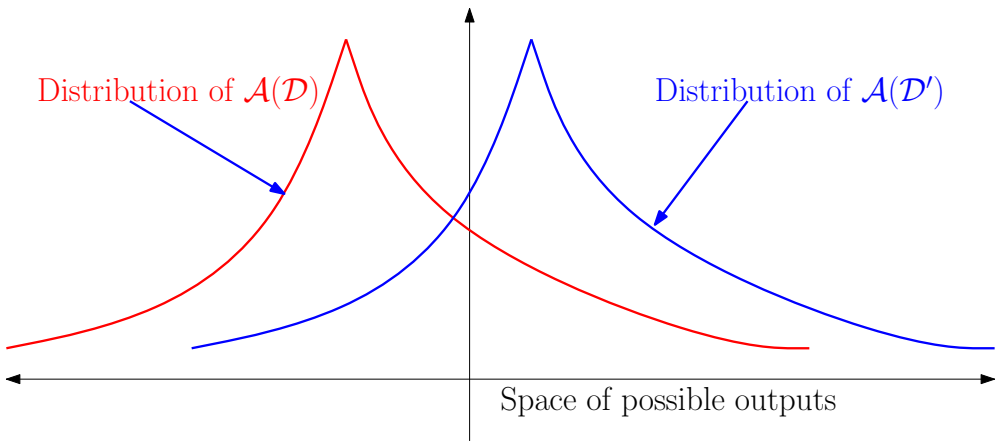


Differential Privacy [DMNS06,DKMMN06]

- Intuition: Users learn roughly the same things about you whether or not your data is in the data set
- It is a restriction on the algorithm $\mathcal{A}$
  - Guarantee privacy even with arbitrary side information

Requirement

Any two data sets $\mathcal{D}$ and $\mathcal{D}'$ that differ in one element induce close distributions on the space of outputs of $\mathcal{A}$



- Closeness is measured by parameters ϵ and δ
  - Think ϵ = small constant and $\delta = 1/\mathrm{poly}(n)$

Prior Works on Learning and Privacy

- Lots of work on differentially private learning [KLNRS08,CM08,ZLW09,CMS11,KST12,JKT12,$\cdots$]
- Almost all the results apply only to low-dimensional regime (i.e., $\mathfrak{p} < \mathfrak{n}$)
- In the context of sparse regression:
  - [ZLW09] works in the regime when $\mathfrak{p} < \mathfrak{n}$
  - [Kifer, Smith, T.'12] is the only result in $\mathfrak{p} \gg \mathfrak{n}$ regime

Our Contributions

New connections between stability and differential privacy

1. Stable nonprivate model selection
   implies private model selection
   - Two notions of stability
     - Subsampling stability [Shao96,Bach08,MB10,...]
     - Perturbation stability [BE02, SSSS10,DL09,...]
2. Algorithm based on subsampling always efficient

Our Contributions

New connections between stability and differential privacy

1. Stable nonprivate model selection
   implies private model selection
   - Two notions of stability
     - Subsampling stability [Shao96,Bach08,MB10,...]
     - Perturbation stability [BE02, SSSS10,DL09,...]
2. Algorithm based on subsampling always efficient

Stability results for the LASSO estimator

1. Consistency assumptions imply stability
2. Efficient test of perturbation stability
3. Private algorithm for sparse linear regression

Our Contributions

New connections between stability and differential privacy

1. Stable nonprivate model selection
   implies private model selection
   - Two notions of stability
     - Subsampling stability
     - Perturbation stability [This talk]
2. Algorithm based on subsampling always efficient

Stability results for the LASSO estimator

1. Consistency assumptions imply stability
2. Efficient test of perturbation stability [This talk]
3. Private algorithm for sparse linear regression

Our Results for Sparse Linear Regression

Stochastic setting [in this talk]:

- $\|\theta^*\|_\infty \leqslant 1$, $\mathrm{supp}(\theta^*) \leqslant s$, min. non zero entry of $\theta^* = \Omega(1)$
- $x_i \sim \mathcal{N}(0,1)^p$, $w \sim \mathcal{N}(0,\sigma^2)^p$

Our Results for Sparse Linear Regression

Stochastic setting [in this talk]:

- $\|\theta^*\|_\infty \leqslant 1$, $\mathrm{supp}(\theta^*) \leqslant s$, min. non zero entry of $\theta^* = \Omega(1)$
- $x_i \sim \mathcal{N}(0,1)^p$, $w \sim \mathcal{N}(0,\sigma^2)^p$

Algorithm	Sample complexity
1. Subsampling based [Kifer, Smith, T.'12]	$O\left(\frac{1}{\epsilon} s^2 \log^2 p\right)$
2. Subsampling based	$O^*\left(\frac{\log(1/\delta)}{\epsilon} s \log p\right)$
3. Perturbation stability based	$O^*\left(s \log p + \frac{\log(1/\delta)}{\epsilon}\lambda\right)$ λ : low-order terms in s and $\log p$ for small s

O^* hides poly log factors

Our Results for Sparse Linear Regression

Stochastic setting [in this talk]:

- $\|\theta^*\|_\infty \leqslant 1$, $\text{supp}(\theta^*) \leqslant s$, min. non zero entry of $\theta^* = \Omega(1)$
- $x_i \sim \mathcal{N}(0,1)^p$, $w \sim \mathcal{N}(0,\sigma^2)^p$

Algorithm	Sample complexity
1. Subsampling based [Kifer, Smith, T.'12]	$O\left(\frac{1}{\epsilon} s^2 \log^2 p\right)$
2. Subsampling based	$O^*\left(\frac{\log(1/\delta)}{\epsilon} s \log p\right)$
3. Perturbation stability based	$O^*\left(s \log p + \frac{\log(1/\delta)}{\epsilon}\lambda\right)$ λ : low-order terms in s and $\log p$ for small s

O^* hides poly log factors

- Non-private optimal sample complexity: $s \log p$

Our Contributions

New connections between stability and differential privacy

1. Stable nonprivate model selection
   implies private model selection
   - Two notions of stability
     - Subsampling stability
     - Perturbation stability
2. Algorithm based on subsampling always efficient

Stability results for the LASSO estimator

1. Consistency assumptions imply stability
2. Efficient test of perturbation stability
3. Private algorithm for sparse linear regression

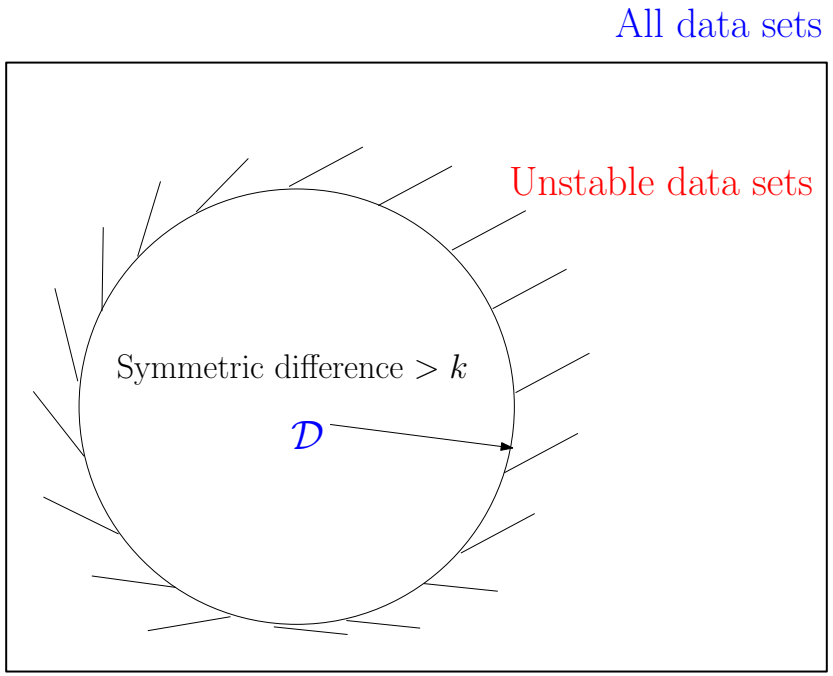
Perturbation Stability based Model Selection

Distance to Instability Framework:

- A function $\mathsf{f} : \mathsf{U}^* \to \mathcal{R}$ is k-stable at $\mathcal{D}$ if
  - For any data set $\mathcal{D}' \in \mathsf{U}^*$, with $|\mathcal{D} \Delta \mathcal{D}'| \leqslant \mathsf{k}$

$$\mathsf{f}(\mathcal{D}) = \mathsf{f}(\mathcal{D}')$$

- Objective: output $\mathsf{f}(\mathcal{D})$ while preserving privacy



Perturbation Stability based Model Selection

1. $\mathsf{dist} \leftarrow \mathsf{max}_k(\mathsf{f}(\mathcal{D})$ is k-stable)
2. $\widetilde{\mathsf{dist}} \leftarrow \mathsf{dist} + \mathsf{Lap}(\frac{1}{\epsilon})$ [Think of the noise as constant]
3. If $\widetilde{\mathsf{dist}} > \frac{\log(1/\delta)}{\epsilon}$, then **return** $\mathsf{f}(\mathcal{D})$, o.w. **return** $\perp$

Perturbation Stability based Model Selection

1. $\text{dist} \leftarrow \max_k(\mathsf{f}(\mathcal{D}) \text{ is k-stable})$
2. $\widetilde{\text{dist}} \leftarrow \text{dist} + \text{Lap}(\frac{1}{\epsilon})$ [Think of the noise as constant]
3. If $\widetilde{\text{dist}} > \frac{\log(1/\delta)}{\epsilon}$, then **return** $\mathsf{f}(\mathcal{D})$, o.w. **return** $\perp$

Privacy guarantee

Theorem (variants in [DL09, KRSY11])

The algorithm is (ϵ, δ)-differentially private

Utility guarantee

Theorem

If f *is* $\left(\frac{2\log(1/\delta)}{\epsilon}\right)$*-stable w.r.t.* $\mathcal{D}$, ***then*** *the algorithm outputs* $\mathsf{f}(\mathcal{D})$

Perturbation Stability based Model Selection

Issue:

- For a given f, distance to instability might not efficiently computable

Perturbation Stability based Model Selection

Issue:

- For a given f, distance to instability might not efficiently computable

Fix: Obtain an efficiently computable proxy distance $\hat{\mathsf{d}}$

- **Safety:** $\forall\ \mathcal{D}, \hat{\mathsf{d}}(\mathcal{D}) \leqslant$ distance to instability of f
- $\hat{\mathsf{d}}$ **is stable**: $\forall \mathcal{D}, \mathcal{D}'$ s.t. $|\mathcal{D} \Delta \mathcal{D}'| = 1$,

$$|\hat{\mathsf{d}}(\mathcal{D}) - \hat{\mathsf{d}}(\mathcal{D}')| = 1$$

Our Contributions

New connections between stability and differential privacy

1. Stable nonprivate model selection
   implies private model selection
   - Two notions of stability
     - Subsampling stability
     - Perturbation stability
2. Algorithm based on subsampling always efficient

Stability results for the LASSO estimator

1. Consistency assumptions imply stability
2. Efficient test of perturbation stability
3. Private algorithm for sparse linear regression

Perturbation Stability of LASSO

Recall: LASSO estimator

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2n}\|y - X\theta\|_2^2 + \frac{\Lambda}{n}\|\theta\|_1$$

Theorem (Consistency [...,Wainwright06])

In the stochastic setting, under suitable choice of Λ *and* $n = \omega(s \log p)$, *w.h.p.* $supp(\hat{\theta}) = supp(\theta^*)$

Perturbation Stability of LASSO

Recall: LASSO estimator

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2n}\|y - X\theta\|_2^2 + \frac{\Lambda}{n}\|\theta\|_1$$

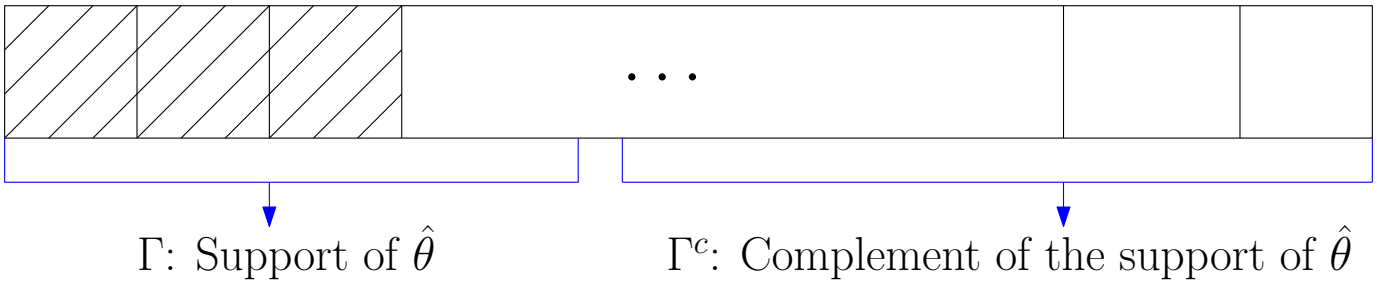
Theorem (Stability of LASSO)

Consistency assumptions on LASSO are sufficient to guarantee perturbation stability

Proof Idea:

- Analyze the KKT condition at $\hat{\theta}$
- Show that $\text{supp}(\hat{\theta})$ is stable by constructing "dual certificate" for nearby instances

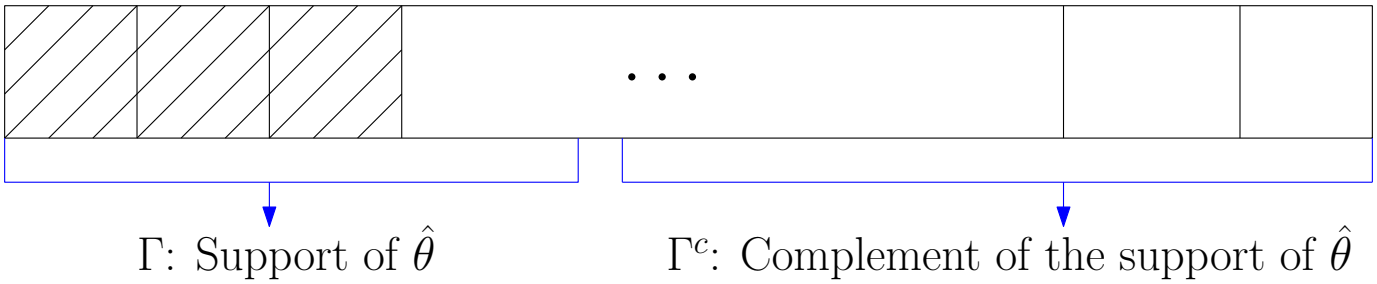
Stability Test for LASSO



Test for the following (real test is more complicated)

- Restricted Strong Convexity (RSC): Restricted to subspace induced by Γ, the least-squared loss is strongly convex
  - **Note:** We only need to test strong convexity at Γ

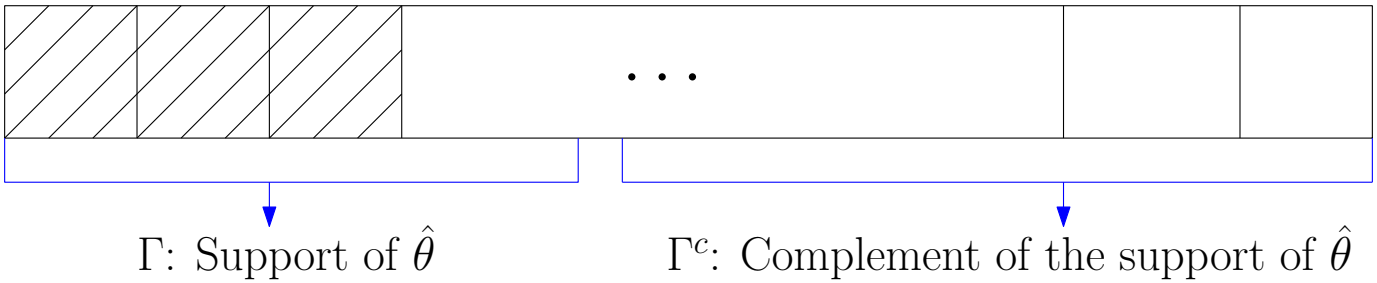
Stability Test for LASSO



Test for the following (real test is more complicated)

- Restricted Strong Convexity (RSC): Restricted to subspace induced by Γ, the least-squared loss is strongly convex
  - **Note:** We only need to test strong convexity at Γ
- "Strong stability": The coordinates of the gradient of the least-squared loss in Γ^c are $\ll \Lambda$

Stability Test for LASSO



Test for the following (real test is more complicated)

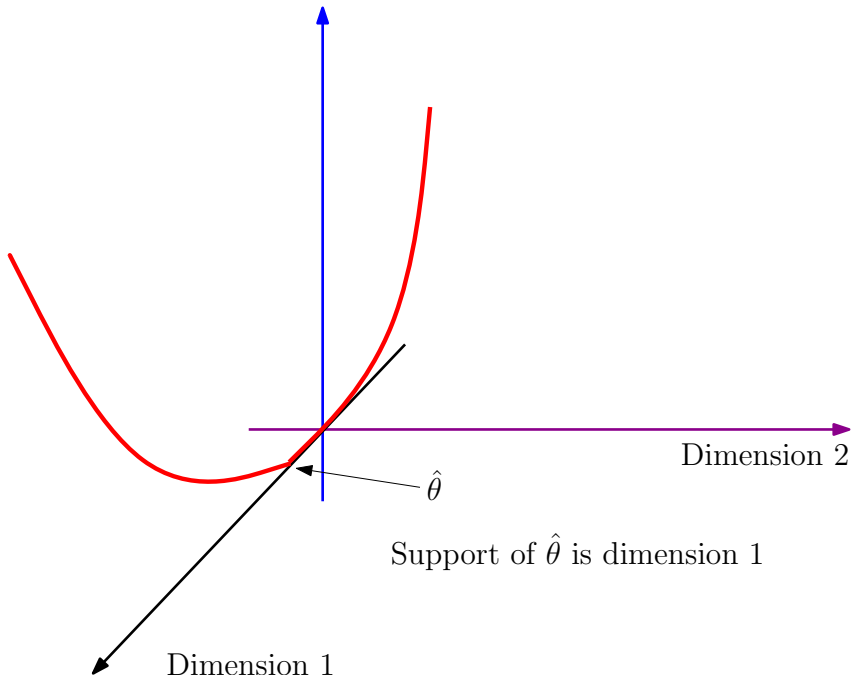
- Restricted Strong Convexity (RSC): Restricted to subspace induced by Γ, the least-squared loss is strongly convex
  - **Note:** We only need to test strong convexity at Γ
- "Strong stability": The coordinates of the gradient of the least-squared loss in Γ^c are $\ll \Lambda$

These conditions hold w.h.p. in the stochastic setting

Geometry of LASSO

Intuition:

- Strong sparsity and RSC implies stability of $\mathrm{supp}(\hat{\theta})$
  - i.e., $\hat{\theta}$ does not shift *too much* by changing one entry in $\mathcal{D}$

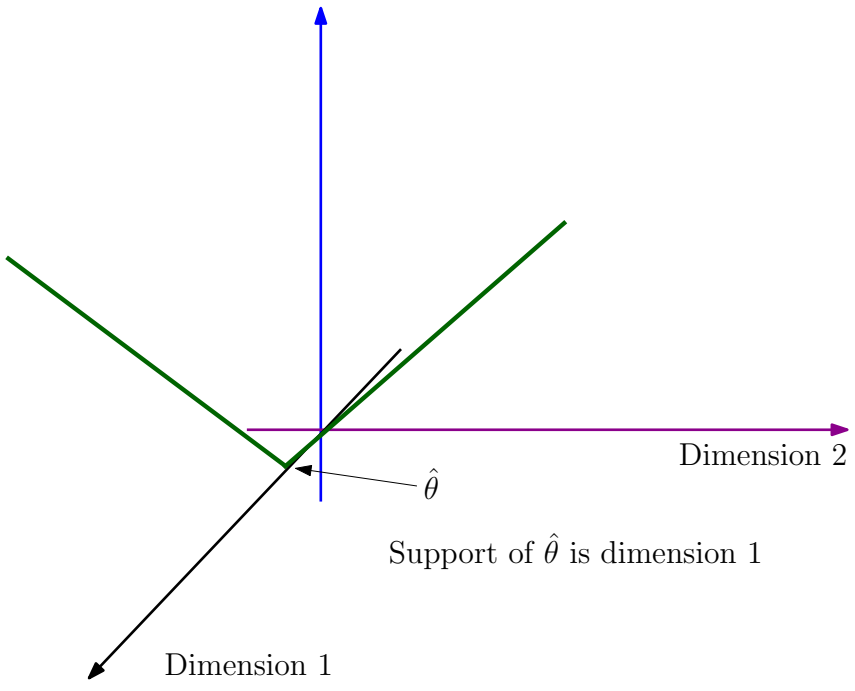


Strong convexity gives stability along *dimension 1*

Geometry of LASSO

Intuition:

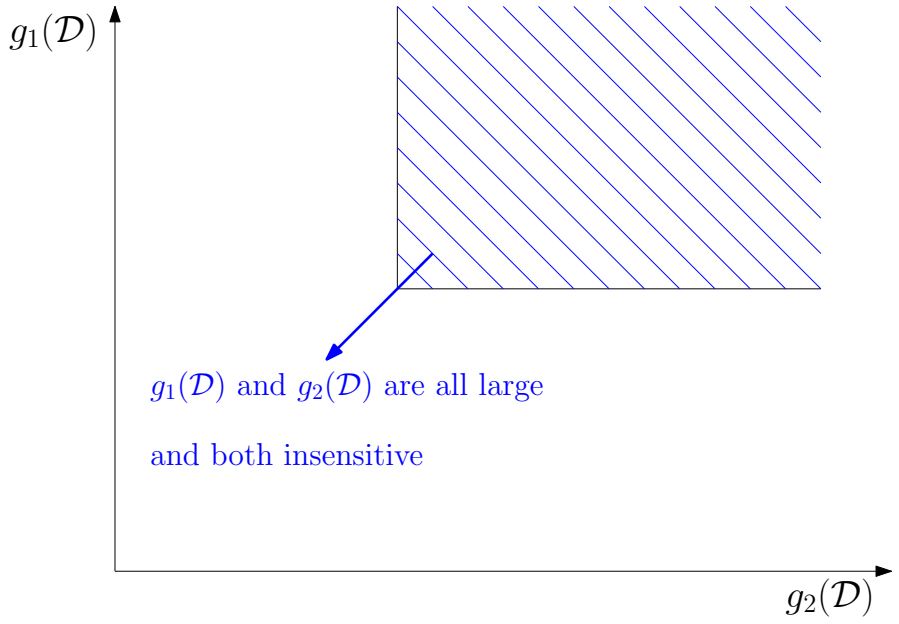
- Strong sparsity and RSC implies stability of $\mathrm{supp}(\hat{\theta})$
  - i.e., $\hat{\theta}$ does not shift *too much* by changing one entry in $\mathcal{D}$



L_1-regularization gives stability along *dimension 2*

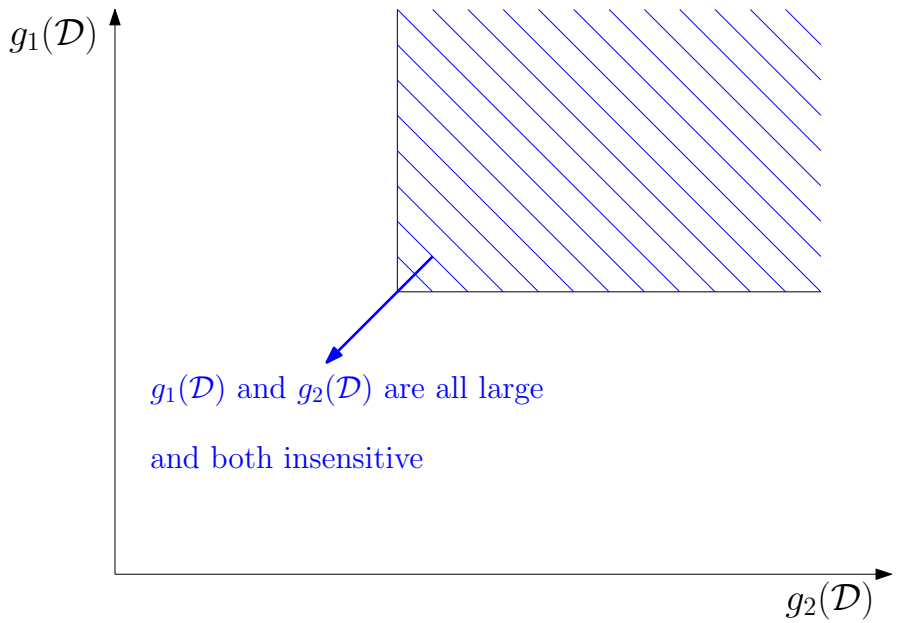
Making the Test Private (Simplified)

- Test for sparsity g_1: Negative of the absolute value of $(s+1)$-th entry of $X^T(y - X\hat{\theta})$
- Test for strong convexity g_2: Min. eigenvalue of $X_\Gamma^T X_\Gamma$



Making the Test Private (Simplified)

- Test for sparsity g_1: Negative of the absolute value of $(s+1)$-th entry of $X^T(y - X\hat{\theta})$
- Test for strong convexity g_2: Min. eigenvalue of $X_\Gamma^T X_\Gamma$



Our result:

- Can privately test if g_1 and g_2 are both large
- Serves as a proxy for the distance to instability $\hat{d}$

Putting the Pieces Together

1. Compute $\hat{d}$ = function of $g_1(\mathcal{D})$ and $g_2(\mathcal{D})$
2. $\widetilde{d} \leftarrow \hat{d} + \mathsf{Lap}(\frac{1}{\epsilon})$ [Think of the noise as constant]
3. If $\hat{d} > \frac{\log(1/\delta)}{\epsilon}$, then **return** $f(\mathcal{D})$, o.w. **return** $\perp$

Privacy guarantee

Theorem (variants in [DL09, KRSY11])

The algorithm is (ϵ, δ)-differentially private

Utility guarantee

Theorem

If $\hat{d}$ *is at least* $\frac{2\log(1/\delta)}{\epsilon}$, ***then*** *w.h.p. the algorithm outputs the support of the underlying parameter* θ^*

Revisit: Our Contributions

New connections between stability and differential privacy

1. Stable nonprivate model selection
   implies private model selection
   - Two notions of stability
     - Subsampling stability
     - Perturbation stability
2. Algorithm based on subsampling always efficient

Stability results for the LASSO estimator

1. Consistency assumptions imply stability
2. Efficient test of perturbation stability
3. Private algorithm for sparse linear regression

Future Work

Future work

- Analyze privacy/stability properties of iterative sparse estimation algorithms (e.g., least angle regression)
- Privacy for high-dimensional learning where "good" model is not unique
  - Low rank matrix approximation
  - Sparse representation with many "good" sparse vectors