Computational and Statistical Tradeoffs via Convex Relaxation

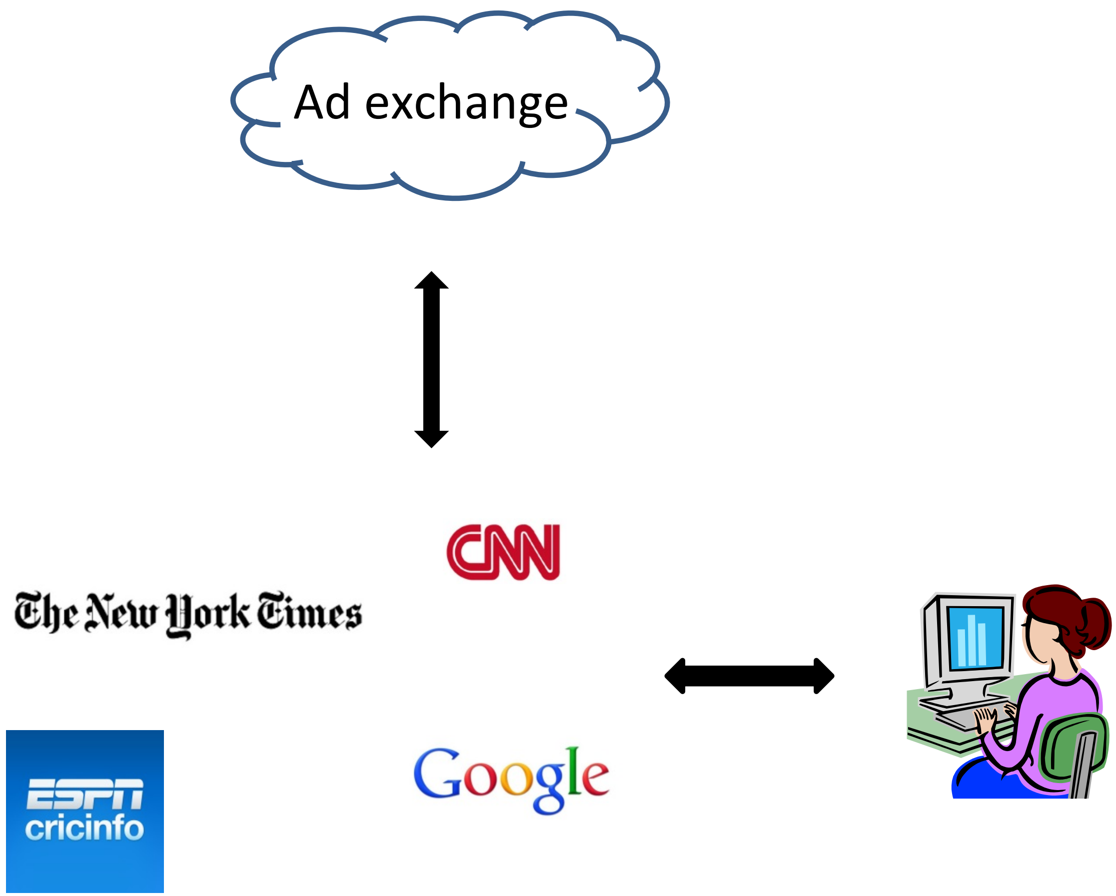
Venkat Chandrasekaran

Caltech

Joint work with **Michael Jordan**

Time-constrained Inference

- Require decision after a fixed (usually small) amount of time



Classical Analysis in Statistics

- Previously, key bottleneck: amount of data
- Consider ``best'' estimator without much regard for computational considerations
  - Minimax analysis
- More recently, time is key bottleneck
  - Data is plentiful in several domains
- Need to incorporate time constraints

A Thought Experiment

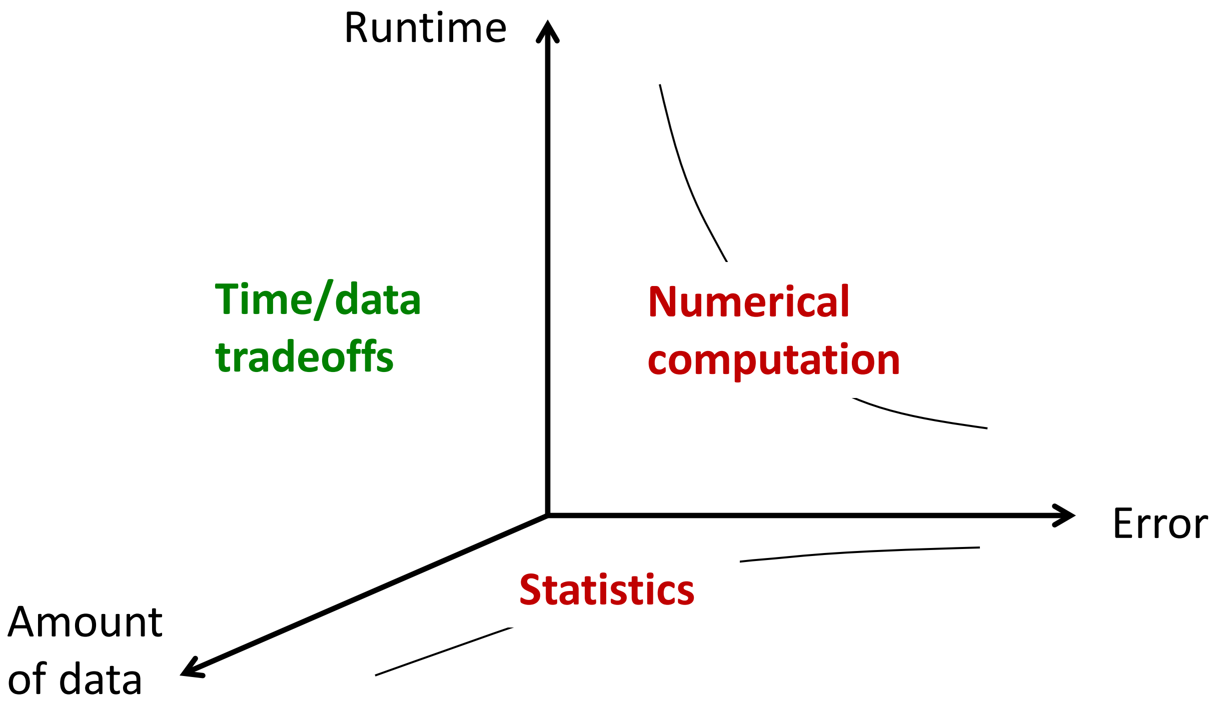
- Consider a typical inference scenario
  - 1 hour for inference task with n = 5000, risk = 0.03
  - 20 days for same task with n = 500000, risk = 0.0003

- Happy with risk = 0.03, but given n = 500000
  - Don't care about small improvements in risk
  - Statistical models are only approximations of reality

A Thought Experiment

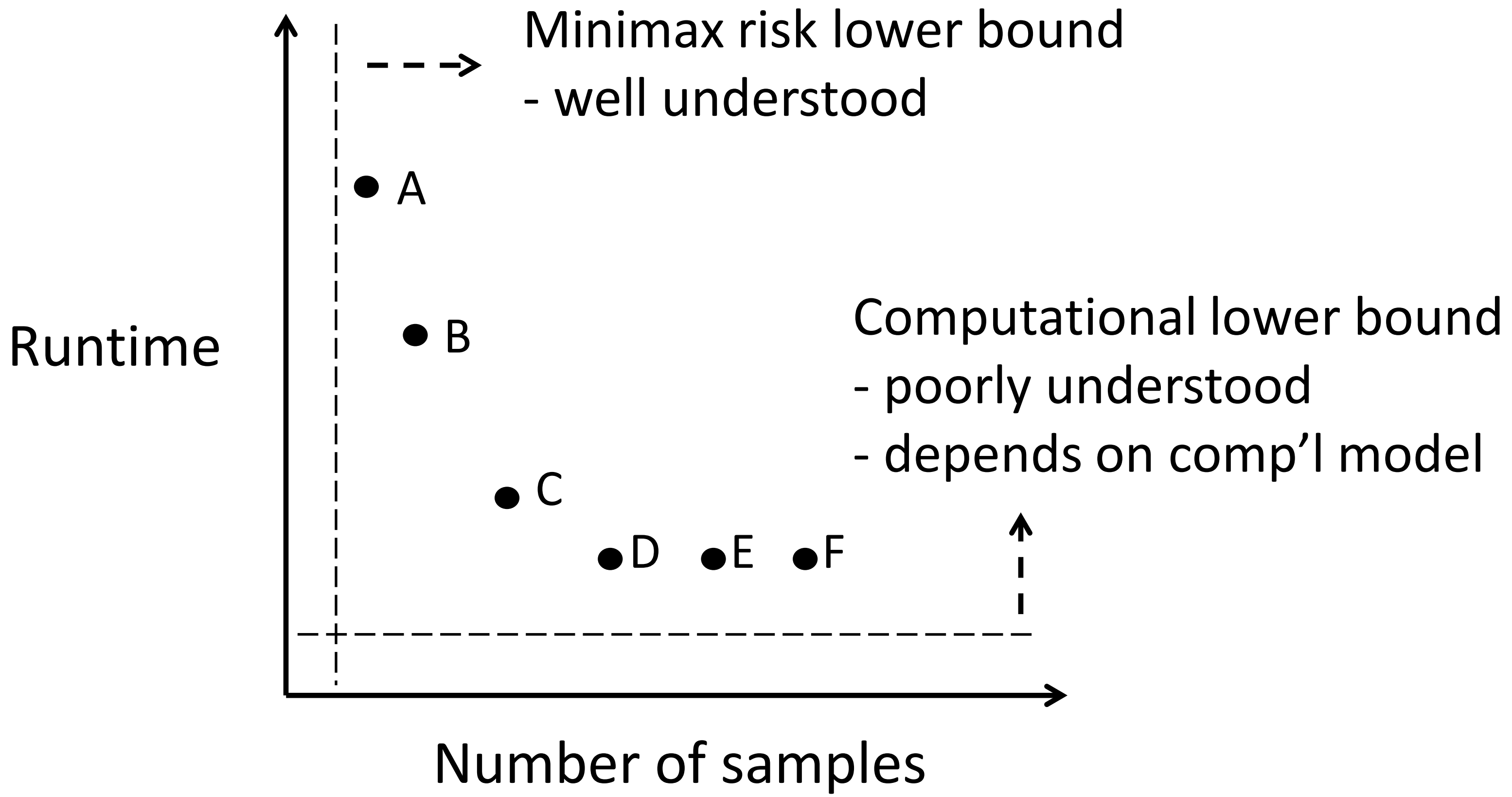
- Consider a typical inference scenario
  - 1 hour for inference task with n = 5000, risk = 0.03
  - 20 days for same task with n = 500000, risk = 0.0003
- Happy with risk = 0.03, but given n = 500000
  - Don’t care about small improvements in risk
  - Statistical models are only approximations of reality
- More data useful for less computation?

Computer Science vs. Statistics



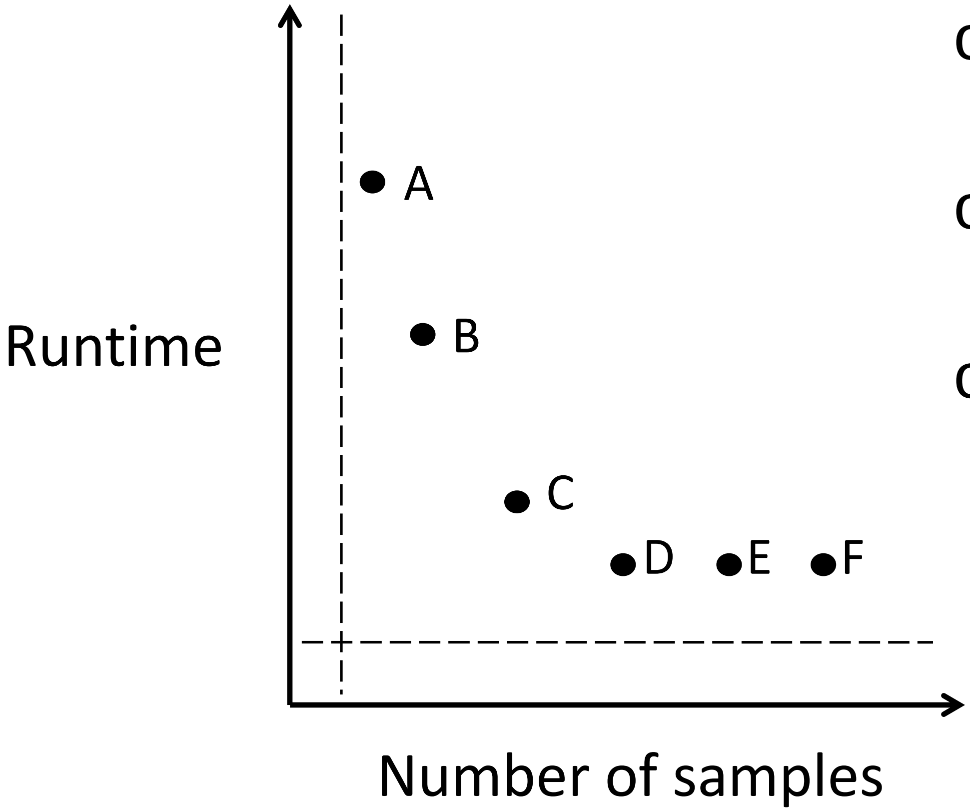
Time-Data Tradeoffs

- Consider an inference problem with ***fixed*** risk
- Inference procedures viewed as points in plot



Time-Data Tradeoffs

- Consider an inference problem with ***fixed*** risk



- Need “**weaker**” algorithms for larger datasets
- At some stage, throw away data
- Tradeoff runtime ***upper bounds***
  - More data means smaller runtime upper bound

An Estimation Problem

- Signal $\mathbf{x}^* \in \mathcal{S} \subset \mathbb{R}^p$ from known (bounded) set
- Noise $\mathbf{z} \sim \mathcal{N}(0, I_{p \times p})$
- Observation model

$$\mathbf{y} = \mathbf{x}^* + \sigma \mathbf{z}$$

- Observe n i.i.d. samples $\{\mathbf{y}_i\}_{i=1}^n$

Convex Programming Estimator

- Sample mean $\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i$ is sufficient statistic

- Natural M-estimator

$$\hat{\mathbf{x}}_n(\mathcal{S}) = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \ \tfrac{1}{2}\,\|\bar{\mathbf{y}} - \mathbf{x}\|_{\ell_2}^2 \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{S}$$

- Convex programming M-estimator

$$\hat{\mathbf{x}}_n(C) = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \ \tfrac{1}{2}\,\|\bar{\mathbf{y}} - \mathbf{x}\|_{\ell_2}^2 \quad \text{s.t.} \quad \mathbf{x} \in C$$

  – C is a **convex** set such that $\mathcal{S} \subset C$

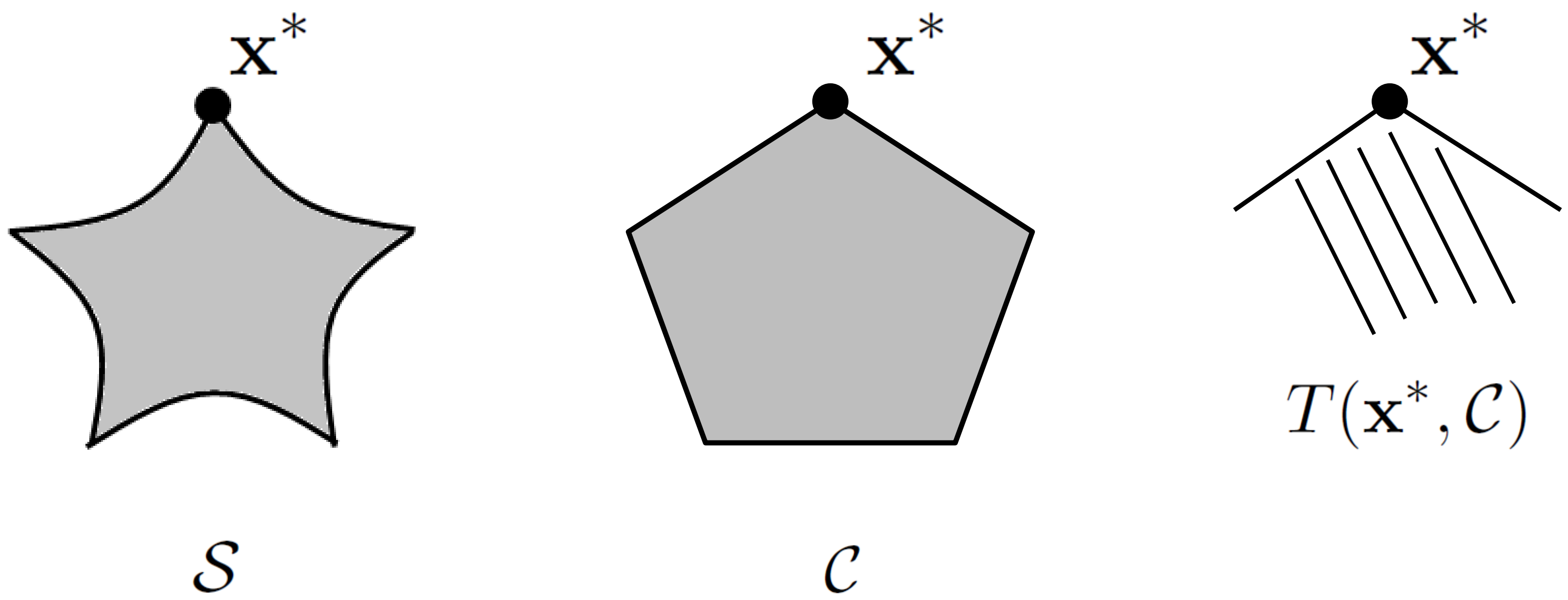
Convex Programming Estimator

- Long history of shrinkage estimation in statistics
  - James, Stein (1961)
  - Donoho, Johnstone (early 1990s)
  - Shrinkage onto convex sets for tractability
- Many surprises in high dimensions, i.e., large p
- More recently
  - L1 norm, trace norm, max norm, …

Statistical Performance of Estimator

- <u>Defn 1</u>: The ***cone of feasible directions*** into a convex set C is defined as

$$T(\mathbf{x}^*, C) = \text{cone}\{w - \mathbf{x}^* | w \in C\}$$



Statistical Performance of Estimator

- Defn 1: The ***cone of feasible directions*** into a convex set C is defined as

$$T(\mathbf{x}^*, C) = \text{cone}\{w - \mathbf{x}^* | w \in C\}$$

- Defn 2: The ***Gaussian (squared) complexity*** of a cone T is defined as

$$g(T) = \mathbb{E}\left[\sup_{\delta \in T, ||\delta||_{\ell_2} \leq 1} \langle \mathbf{z}, \delta \rangle^2\right]$$

Statistical Performance of Estimator

- Prop: The risk of the estimator $\hat{\mathbf{x}}_n(C)$ is

$$\mathbb{E}\left[\|\hat{\mathbf{x}}_n(C) - \mathbf{x}^*\|_{\ell_2}^2\right] \leq \frac{\sigma^2}{n}\, g\Big(T(\mathbf{x}^*, C)\Big)$$

- Proof: Apply optimality conditions

- Intuition: Only consider error in feasible cone

Weakening via Convex Relaxation

- Prop: The risk of the estimator $\hat{\mathbf{x}}_n(C)$ is

$$\mathbb{E}\left[\|\hat{\mathbf{x}}_n(C) - \mathbf{x}^*\|_{\ell_2}^2\right] \leq \frac{\sigma^2}{n}\, g\Big(T(\mathbf{x}^*, C)\Big)$$

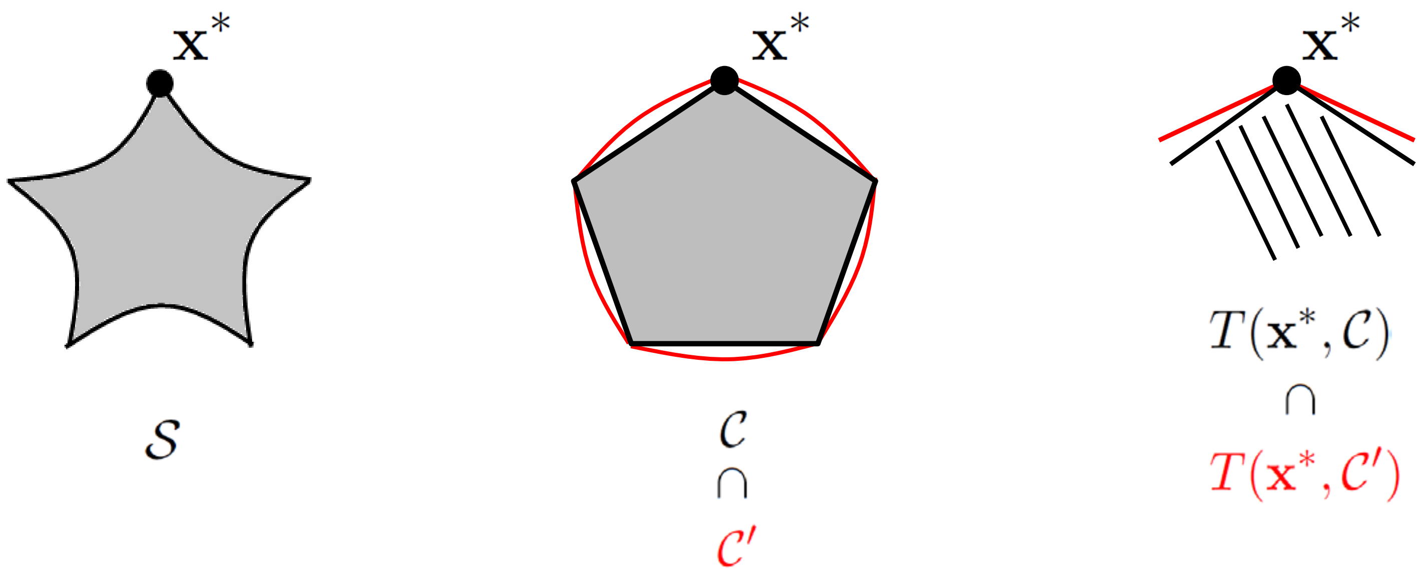
- Corr: To obtain risk of at most 1,

$$n \geq \sigma^2\, g\Big(T(\mathbf{x}^*, C)\Big)$$

Weakening via Convex Relaxation

o <u>Corr</u>: To obtain risk of at most 1,

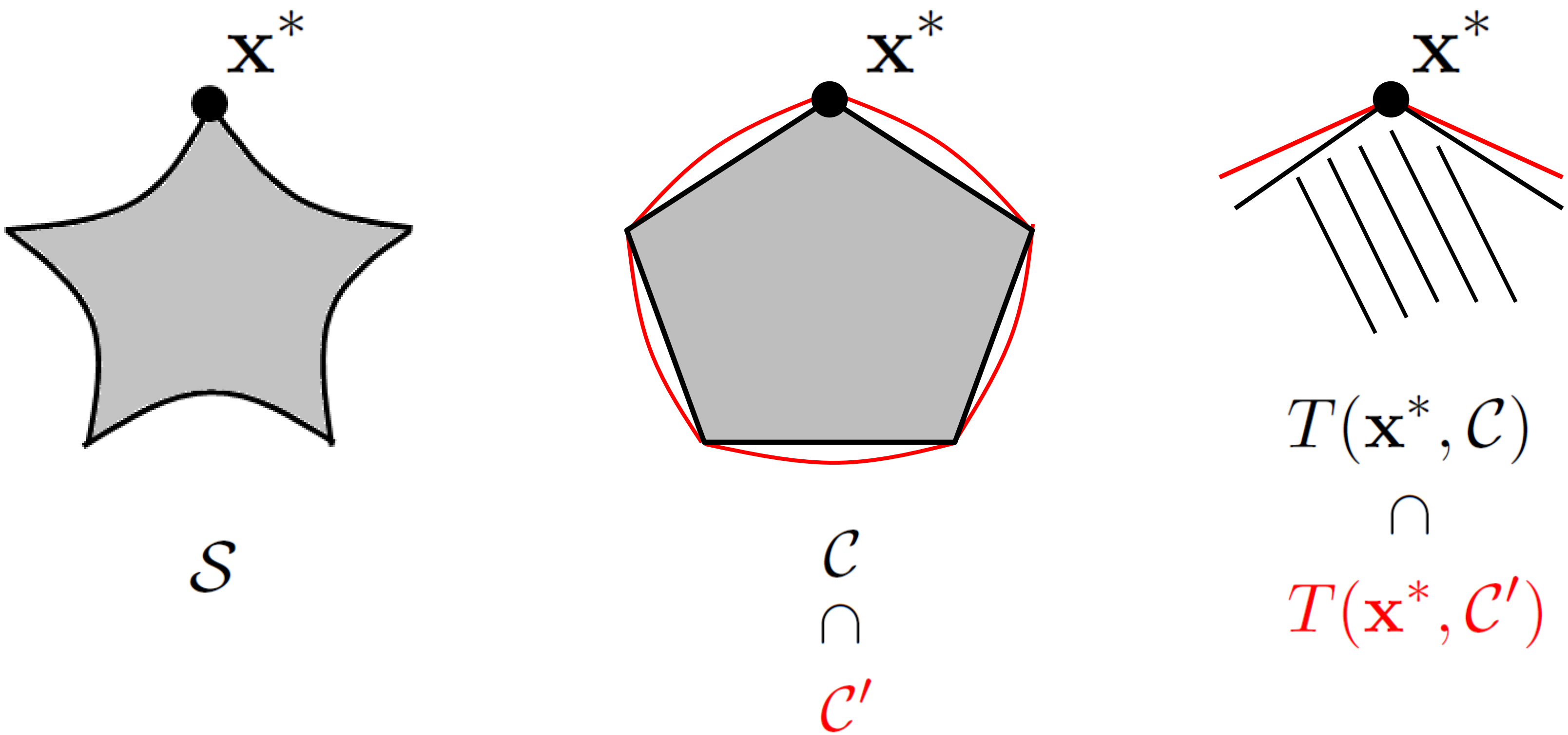
$$n \geq \sigma^2 \; \underbrace{g\Big(T(\mathbf{x}^*, C)\Big)}_{\text{Monotonic in } C}$$



Weakening via Convex Relaxation

If we have access to larger n, can use larger C

→ Obtain "weaker" estimation algorithm



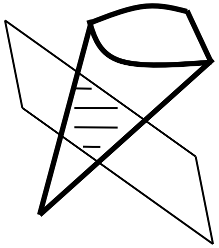
Hierarchy of Convex Relaxations

- If $\mathcal{S}$ "algebraic", then one can obtain family of outer convex approximations

$$\mathrm{conv}(\mathcal{S}) \subseteq \cdots \subset C_3 \subset C_2 \subset C_1$$

  - Polyhedral, semidefinite, hyperbolic relaxations (Sherali-Adams, Boyd, Parrilo, Lasserre, Renegar)

- Sets $\{C_i\}$ ordered by *computational complexity*
  - Central role played by **lift-and-project**



Contrast to Previous Work

- Binary classifier learning
  - Decatur et al. [1998], Servedio [2000], Shalev-Shwarz et al. [2008, 2012], Perkins & Hallett [2010]
  - Lots of extra data required for simpler algorithms
  - Our examples: modest extra data for simpler algorithms

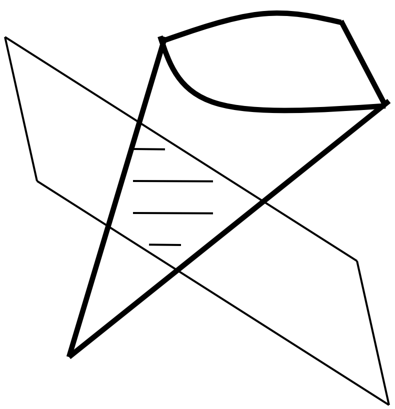
- Sparse PCA, clustering, network inference
  - Amini & Wainwright [2009], Kolar et al. [2011]
- Model selection
  - Agarwal et al. [2011]

Contrast to Previous Work

- **Our work**: Emphasis on ***algorithm weakening***

- Convex relaxation is a principled, general way to do this

$$\mathrm{conv}(\mathcal{S}) \subseteq \cdots \subset C_3 \subset C_2 \subset C_1$$



Example 1

- $\mathcal{S}$ consists of cut matrices

$$\mathcal{S} = \{\mathbf{aa}' \mid \mathbf{a} \text{ consists of } \pm 1's\}$$

- E.g., collaborative filtering, clustering

C	Runtime	n
conv($\mathcal{S}$) (cut polytope)	super-poly(p)	$c_1\sqrt{p}$
elliptope	$p^{2.25}$	$c_2\sqrt{p}$
nuclear norm ball	$p^{1.5}$	$c_3\sqrt{p}$

$(c_1 < c_2 < c_3)$

Example 2

- Signal set $\mathcal{S}$ consists of all perfect matchings in complete graph
- E.g., network inference

C	Runtime	n
$\text{conv}(\mathcal{S})$	p^5	$c_1\sqrt{p}\log(p)$
hypersimplex	$p^{1.5}\log(p)$	$c_2\sqrt{p}\log(p)$

$(c_1 < c_2)$

Example 3

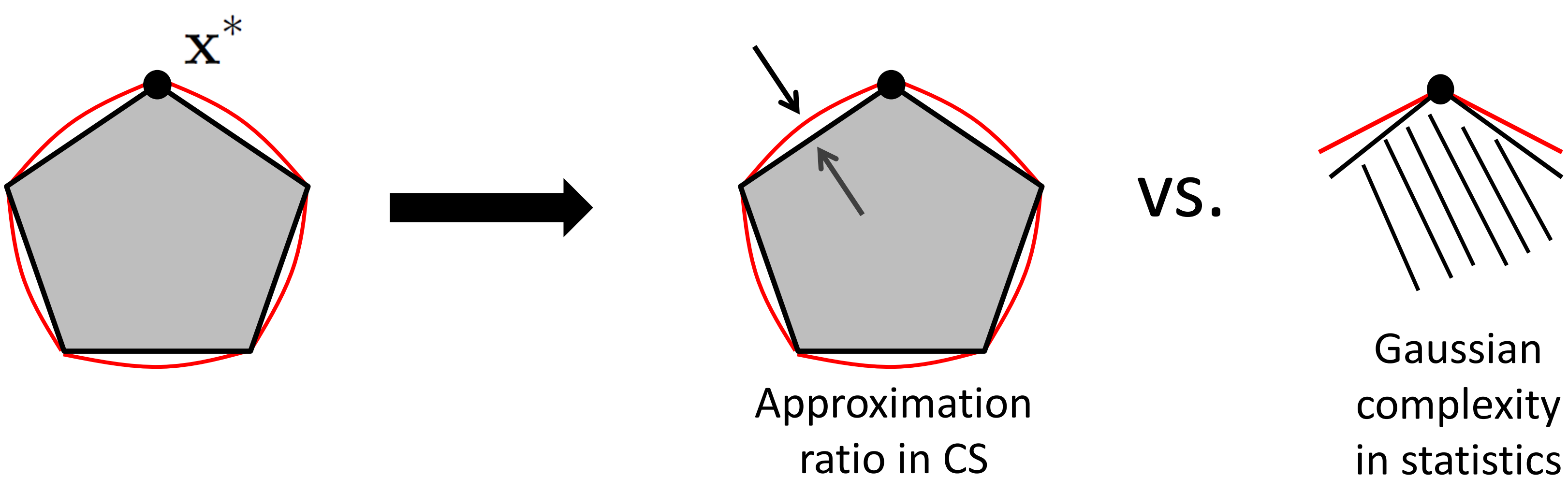
- $\mathcal{S}$ consists of all adjacency matrices of graphs with only a clique on square-root of the nodes
- E.g., sparse PCA, gene expression patterns
- Kolar et al. (2010)

C	Runtime	n
$\text{conv}(\mathcal{S})$	super-poly(p)	$\sim p^{0.25}\log(p)$
nuclear norm ball	$p^{1.5}$	$\sim \sqrt{p}$

Some Questions

- In several examples, not too many extra samples required for really simple algorithms

- Quality of approximation of convex sets
  - ***Approximation ratio*** is focus in theoretical CS
  - ***Gaussian complexities*** in statistical inference



Summary

- Challenges with massive datasets
- Considered simple denoising problem
- Time-data tradeoffs via convex relaxation

- Future work:
  - Gaussian complexities of LP/SDP hierarchies
  - Other methods to weaken algorithms
  - Other inference problems

users.cms.caltech.edu/~venkatc