high-dimensional estimation via sum-of-squares proofs

David Steurer (ETH Zurich)

based on ICM proceedings article with **Prasad Raghavendra (UC Berkeley)** and **Tselil Schramm (Harvard / MIT)**

SIAM AG, Polynomial Optimization and its Applications, Bern, 2019

estimation: class of problems studied in *signal processing*, *statistics*, and *machine learning*

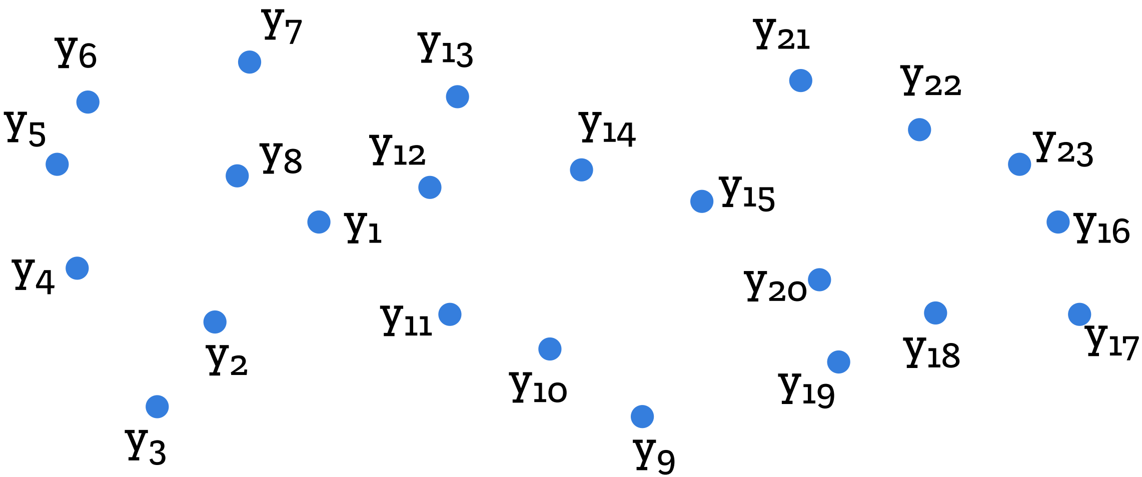
estimation problem (aka inference / inverse problem)

given: output Y of known randomized process for some unknown input X^*

goal: recover (approximately) this input X^*

$$\text{parameter / signal } X^* \xrightarrow{\text{process}} \text{measurement / observation } Y$$

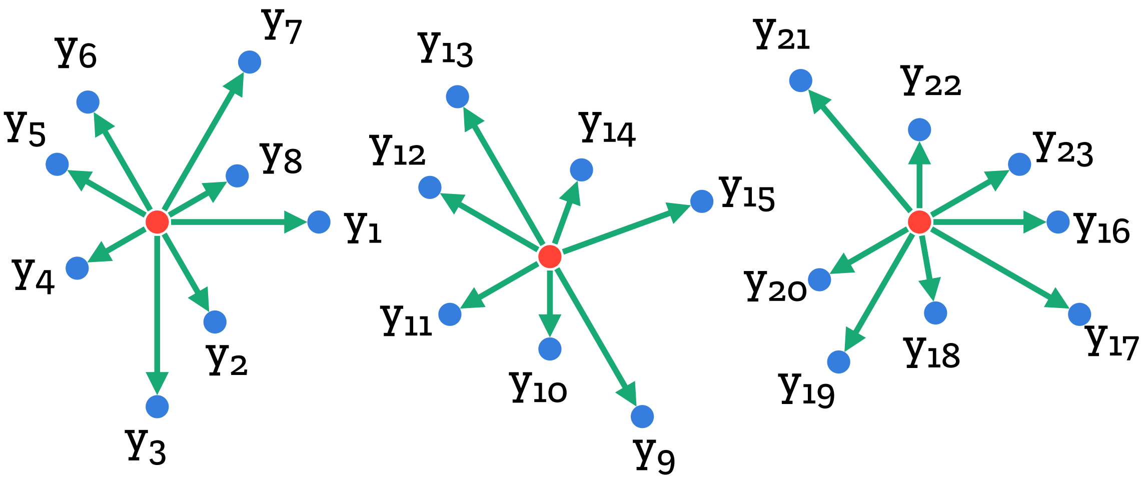
example: *clustering* as estimation problem



k-clustering problem

given: vectors $y_1, \ldots, y_n \in \mathbb{R}^d$

example: *clustering* as estimation problem

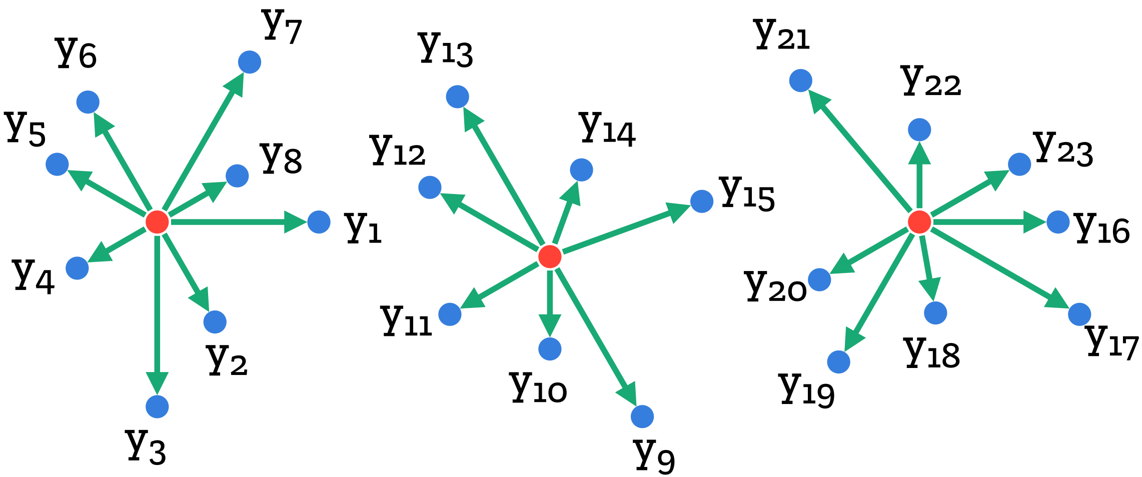


k-clustering problem

given: vectors $y_1, \ldots, y_n \in \mathbb{R}^d$ with $y_i = x_i^* + w_i$, where

- $\{x_1^*, \ldots, x_n^*\}$ consists of $\leqslant k$ different vectors *("centers")*
- $w_1, \ldots, w_n$ are iid standard Gaussian vectors *("noise")* and

example: *clustering* as estimation problem



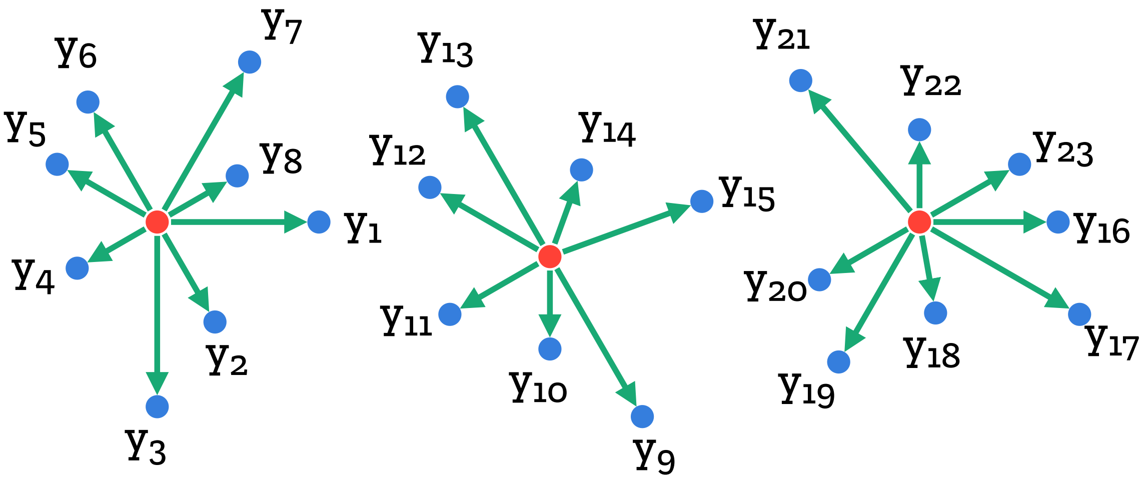
k-clustering problem

given: vectors $y_1, \dots, y_n \in \mathbb{R}^d$ with $y_i = x_i^* + w_i$, where

- $\{x_1^*, \dots, x_n^*\}$ consists of $\leqslant k$ different vectors *("centers")*
- $w_1, \dots, w_n$ are iid standard Gaussian vectors *("noise")* and

goal: recover $x_1^*, \dots, x_n^*$ *(center for each data point)*

example: *clustering* as estimation problem



k-clustering problem (aka Gaussian mixture model)

given: vectors $y_1, \ldots, y_n \in \mathbb{R}^d$ with $y_i = x_i^* + w_i$, where

- $\{x_1^*, \ldots, x_n^*\}$ consists of $\leqslant k$ different vectors *("centers")*
- $w_1, \ldots, w_n$ are iid standard Gaussian vectors *("noise")* and

goal: recover $x_1^*, \ldots, x_n^*$ *(center for each data point)*

one of the most extensively studied statistical models

k-clustering problem (aka Gaussian mixture model)

given: vectors $y_1, \ldots, y_n \in \mathbb{R}^d$ with $y_i = x_i^* + w_i$, where

- $\{x_1^*, \ldots, x_n^*\}$ consists of $\leqslant k$ different vectors *("centers")*
- $w_1, \ldots, w_n$ are iid standard Gaussian vectors *("noise")* and

goal: recover $x_1^*, \ldots, x_n^*$ *(center for each data point)*

concise matrix form:

$$X^* \in \mathbb{R}^{d\times n} \xrightarrow{\text{additive Gaussian noise } W} Y = X^* + W \in \mathbb{R}^{d\times n}$$

where X^* is restricted to have $\leqslant k$ different columns

k-clustering problem (aka Gaussian mixture model)

given: vectors $y_1, \ldots, y_n \in \mathbb{R}^d$ with $y_i = x_i^* + w_i$, where

- $\{x_1^*, \ldots, x_n^*\}$ consists of $\leqslant k$ different vectors *("centers")*
- $w_1, \ldots, w_n$ are iid standard Gaussian vectors *("noise")* and

goal: recover $x_1^*, \ldots, x_n^*$ *(center for each data point)*

concise matrix form:

$$X^* \in \mathbb{R}^{d \times n} \xrightarrow{\text{additive Gaussian noise } W} Y = X^* + W \in \mathbb{R}^{d \times n}$$

where X^* is restricted to have $\leqslant k$ different columns

similar examples (different constraints on X^*)

linear regression: X^* belongs to known linear subspace

principal component analysis (PCA): X^* is rank-1 matrix

tensor PCA: X^* is rank-1 tensor (order 3 or higher)

k-clustering problem (aka Gaussian mixture model)

given: matrix $Y \in \mathbb{R}^{d \times n}$ with $Y = X^* + W$, where

- W has iid standard Gaussian entries
- X^* has $\leqslant k$ different columns

goal: recover X^*

k-clustering problem (aka Gaussian mixture model)

given: matrix $Y \in \mathbb{R}^{d \times n}$ with $Y = X^* + W$, where

- W has iid standard Gaussian entries
- X^* has $\leqslant k$ different columns

goal: recover X^*

desired: estimator $\hat{X}(Y)$ with small *error* $\frac{1}{n}\|\hat{X}(Y) - X^*\|_{\mathrm{F}}^2$

(with high probability over W; even for worst X^*)

k-clustering problem (aka Gaussian mixture model)

given: matrix $Y \in \mathbb{R}^{d \times n}$ with $Y = X^* + W$, where

- W has iid standard Gaussian entries
- X^* has $\leqslant k$ different columns

goal: recover X^*

desired: estimator $\hat{X}(Y)$ with small *error* $\frac{1}{n}\|\hat{X}(Y) - X^*\|_{\mathrm{F}}^2$
(with high probability over W; even for worst X^*)

best known theoretical guarantees: (until recently)

statistical: $\Theta(\log k)$ (exponential time) [Regev–Vijayaraghavan'17]

computational: $O(k^{1/2})$ (polynomial time) [Vempala–Wang]

k-clustering problem (aka Gaussian mixture model)

given: matrix $Y \in \mathbb{R}^{d\times n}$ with $Y = X^* + W$, where

- W has iid standard Gaussian entries
- X^* has $\leqslant k$ different columns

goal: recover X^*

desired: estimator $\hat{X}(Y)$ with small *error* $\frac{1}{n}\|\hat{X}(Y) - X^*\|_{\mathrm{F}}^2$

best known theoretical guarantees: (until recently)

statistical: $\Theta(\log k)$ (exponential time) [Regev–Vijayaraghavan'17]

computational: $O(k^{1/2})$ (polynomial time) [Vempala–Wang]

inherent gap between statistical and computational error?

k-clustering problem (aka Gaussian mixture model)

given: matrix $Y \in \mathbb{R}^{d\times n}$ with $Y = X^* + W$, where

- W has iid standard Gaussian entries
- X^* has $\leqslant k$ different columns

goal: recover X^*

desired: estimator $\hat{X}(Y)$ with small *error* $\frac{1}{n}\|\hat{X}(Y) - X^*\|_{\mathrm{F}}^2$

best known theoretical guarantees: (until recently)

statistical: $\Theta(\log k)$ (exponential time) [Regev–Vijayaraghavan'17]

computational: $O(k^{1/2})$ (polynomial time) [Vempala–Wang]

inherent gap between statistical and computational error?

what about gap for other estimation problems, e.g., tensor PCA?

overview

overview

meta-algorithm: ***sum-of-squares (sos)***

- applies *in canonical way* to wide range of est. problems

overview

meta-algorithm: *sum-of-squares (sos)*

- applies *in canonical way* to wide range of est. problems

***clustering* via sos:**

[Hopkins-Li, Kothari-Steinhardt-S., Diakonikolas-Kane-Stewart]

- achieve improved computational error to nearly match optimal statistical error
- technique: ***proof-to-algorithm paradigm***

overview

meta-algorithm: *sum-of-squares (sos)*

- applies *in canonical way* to wide range of est. problems

***clustering* via sos:**

[Hopkins-Li, Kothari-Steinhardt-S., Diakonikolas-Kane-Stewart]

- achieve improved computational error to nearly match optimal statistical error
- technique: ***proof-to-algorithm paradigm***

***strong limitations* of sos:**

[Barak-Hopkins-Kelner-Kothari-Moitra-Potechin, Hopkins-Kothari-Potechin-R.-Schramm-S.]

- concrete evidence for inherent gap between statistical and computational error for some est. problems, e.g., *tensor PCA*
- technique: ***pseudo-calibration***

setup for sum-of-squares meta-algorithm

setup for sum-of-squares meta-algorithm

let $\Omega \subseteq \mathbb{R}^N$ be set of possible signals (e.g., set of matrices with $\leqslant k$ different columns or set of rank-1 tensors of order 3)

setup for sum-of-squares meta-algorithm

let $\Omega \subseteq \mathbb{R}^N$ be set of possible signals (e.g., set of matrices with $\leqslant k$ different columns or set of rank-1 tensors of order 3)

given: $Y = X^* + W$ for $X^* \in \Omega$ and std. Gaussian W

goal: estimate X^*

setup for sum-of-squares meta-algorithm

let $\Omega \subseteq \mathbb{R}^N$ be set of possible signals (e.g., set of matrices with $\leqslant k$ different columns or set of rank-1 tensors of order 3)

given: $Y = X^* + W$ for $X^* \in \Omega$ and std. Gaussian W

goal: estimate X^*

starting point: *low-deg. polynomial constraints* $\{\vec{p}(X, Y) \geqslant 0\}$;

let $\Omega \subseteq \mathbb{R}^N$ be set of possible signals

given: $Y = X^* + W$ for $X^* \in \Omega$ and std. Gaussian W

goal: estimate X^*

starting point: *low-deg. polynomial constraints* $\{\vec{p}(X, Y) \geqslant 0\}$;
for all $X^* \in \Omega$ and most measurements $Y = X^* + W$:

let $\Omega \subseteq \mathbb{R}^N$ be set of possible signals

given: $Y = X^* + W$ for $X^* \in \Omega$ and std. Gaussian W

goal: estimate X^*

starting point: *low-deg. polynomial constraints* $\{\vec{p}(X, Y) \geqslant 0\}$;
for all $X^* \in \Omega$ and most measurements $Y = X^* + W$:

1. signal-measurement pair satisfies $\vec{p}(X^*, Y) \geqslant 0$

let $\Omega \subseteq \mathbb{R}^N$ be set of possible signals

given: $Y = X^* + W$ for $X^* \in \Omega$ and std. Gaussian W

goal: estimate X^*

starting point: *low-deg. polynomial constraints* $\{\vec{p}(X, Y) \geqslant 0\}$;
for all $X^* \in \Omega$ and most measurements $Y = X^* + W$:

1. signal-measurement pair satisfies $\vec{p}(X^*, Y) \geqslant 0$
2. constraints are *"ε-identifying"*:

$$\forall X.\ \vec{p}(X, Y) \geqslant 0 \Rightarrow \|X - X^*\|^2 \leqslant \varepsilon$$

let $\Omega \subseteq \mathbb{R}^N$ be set of possible signals

given: $Y = X^* + W$ for $X^* \in \Omega$ and std. Gaussian W

goal: estimate X^*

starting point: *low-deg. polynomial constraints* $\{\vec{p}(X, Y) \geqslant 0\}$;
for all $X^* \in \Omega$ and most measurements $Y = X^* + W$:

1. signal-measurement pair satisfies $\vec{p}(X^*, Y) \geqslant 0$
2. constraints are *"ε-identifying"*:

$$\forall X.\ \ \vec{p}(X, Y) \geqslant 0 \Rightarrow \|X - X^*\|^2 \leqslant \varepsilon$$

solutions to this polynomial system are
estimates with error $\leqslant \varepsilon$ ☺

let $\Omega \subseteq \mathbb{R}^N$ be set of possible signals

given: $Y = X^* + W$ for $X^* \in \Omega$ and std. Gaussian W

goal: estimate X^*

starting point: *low-deg. polynomial constraints* $\{\vec{p}(X,Y) \geqslant 0\}$;
for all $X^* \in \Omega$ and most measurements $Y = X^* + W$:

1. signal-measurement pair satisfies $\vec{p}(X^*,Y) \geqslant 0$
2. constraints are *"ε-identifying"*:

$$\forall X.\ \vec{p}(X,Y) \geqslant 0 \Rightarrow \|X - X^*\|^2 \leqslant \varepsilon$$

solutions to this polynomial system are estimates with error $\leqslant \varepsilon$ ☺

*estimates a-priori **not efficiently computable** because polynomial systems are NP-hard to solve in worst-case ☹*

let $\Omega \subseteq \mathbb{R}^N$ be set of possible signals

given: $Y = X^* + W$ for $X^* \in \Omega$ and std. Gaussian W

goal: estimate X^*

starting point: *low-deg. polynomial constraints* $\{\vec{p}(X, Y) \geqslant 0\}$;
for all $X^* \in \Omega$ and most measurements $Y = X^* + W$:

1. signal-measurement pair satisfies $\vec{p}(X^*, Y) \geqslant 0$
2. constraints are *"ε-identifying"*:

$$\forall X.\ \vec{p}(X, Y) \geqslant 0 \Rightarrow \|X - X^*\|^2 \leqslant \varepsilon$$

many classical estimation algorithms can be viewed as
specialized solvers *to exploit structure of polynomial system;*
often notoriously hard to analyze

let $\Omega \subseteq \mathbb{R}^N$ be set of possible signals

given: $Y = X^* + W$ for $X^* \in \Omega$ and std. Gaussian W

goal: estimate X^*

starting point: *low-deg. polynomial constraints* $\{\vec{p}(X, Y) \geqslant 0\}$;

for all $X^* \in \Omega$ and most measurements $Y = X^* + W$:

1. signal-measurement pair satisfies $\vec{p}(X^*, Y) \geqslant 0$
2. constraints are *"ε-identifying"*:

$$\forall X.\ \vec{p}(X, Y) \geqslant 0 \Rightarrow \|X - X^*\|^2 \leqslant \varepsilon \qquad (*)$$

many classical estimation algorithms can be viewed as ***specialized solvers*** *to exploit structure of polynomial system;* ***often notoriously hard to analyze***

emerging approach: *often* ***proof of*** $(*)$ *can be turned systematically into efficient estimation algorithm*

let $\Omega \subseteq \mathbb{R}^N$ be set of possible signals

given: $Y = X^* + W$ for $X^* \in \Omega$ and std. Gaussian W

goal: estimate X^*

starting point: *low-deg. polynomial constraints* $\{\vec{p}(X, Y) \geqslant 0\}$;
for all $X^* \in \Omega$ and most measurements $Y = X^* + W$:

1. signal-measurement pair satisfies $\vec{p}(X^*, Y) \geqslant 0$
2. constraints are *"ε-identifying"*:

$$\forall X.\ \vec{p}(X, Y) \geqslant 0 \Rightarrow \|X - X^*\|^2 \leqslant \varepsilon \qquad (*)$$

sum-of-squares meta-algorithm: *computationally-efficient estimator with error* $\leqslant \varepsilon$ *whenever* $\exists$ *"low-degree proof" for* $(*)$

sum-of-squares proofs

let $S = \{X \in \mathbb{R}^N \mid p_1(X) \geqslant 0, ..., p_m(X) \geqslant 0\}$

sum-of-squares proofs

let $S = \{X \in \mathbb{R}^N \mid p_1(X) \geqslant 0, \ldots, p_m(X) \geqslant 0\}$

definition: *degree-ℓ sum-of-squares proof* of the statement

$$\forall X \in S.\ q(X) \geqslant 0$$

sum-of-squares proofs

let $S = \{X \in \mathbb{R}^N \mid p_1(X) \geqslant 0, \ldots, p_m(X) \geqslant 0\}$

definition: *degree-ℓ sum-of-squares proof* of the statement

$$\forall X \in S.\ q(X) \geqslant 0$$

expresses $q(X)$ as *sum of nonnegative polynomials over* S with degree $\leqslant \ell$ of the form

sum-of-squares proofs

let $S = \{X \in \mathbb{R}^N \mid p_1(X) \geqslant 0, ..., p_m(X) \geqslant 0\}$

definition: *degree-ℓ sum-of-squares proof* of the statement

$$\forall X \in S.\ q(X) \geqslant 0$$

expresses $q(X)$ as *sum of nonnegative polynomials over* S with degree $\leqslant \ell$ of the form

$$r(X)^2 \cdot p_1(X)^{b_1} \cdots p_m(X)^{b_m},$$

where $r(X)$ is polynomial and $b_1, \ldots, b_m \in \{0, 1\}$

let $S = \{X \in \mathbb{R}^N \mid p_1(X) \geqslant 0, ..., p_m(X) \geqslant 0\}$

definition: *degree-ℓ sum-of-squares proof* of the statement

$$\forall X \in S.\ q(X) \geqslant 0$$

expresses $q(X)$ as *sum of nonnegative polynomials over* S with degree $\leqslant \ell$ of the form

$$r(X)^2 \cdot p_1(X)^{b_1} \cdots p_m(X)^{b_m},$$

where $r(X)$ is polynomial and $b_1, \ldots, b_m \in \{0, 1\}$

assuming compactness of S, *every true statement has sum-of-squares proof* (of potentially large degree)

let $S = \{X \in \mathbb{R}^N \mid p_1(X) \geqslant 0, \ldots, p_m(X) \geqslant 0\}$

definition: *degree-ℓ sum-of-squares proof* of the statement

$$\forall X \in S.\ q(X) \geqslant 0$$

expresses $q(X)$ as *sum of nonnegative polynomials over S* with degree $\leqslant \ell$ of the form

$$r(X)^2 \cdot p_1(X)^{b_1} \cdots p_m(X)^{b_m},$$

where $r(X)$ is polynomial and $b_1, \ldots, b_m \in \{0, 1\}$

assuming compactness of S, *every true statement has sum-of-squares proof* (of potentially large degree)

related to:

- *Hilbert's 17th problem* (expressing positive polynomials as sum of squares of rational functions) [Artin]
- *Positivstellensatz* in real algebraic geometry [Krivine, Stengle]

let $S = \{X \in \mathbb{R}^N \mid p_1(X) \geqslant 0, \ldots, p_m(X) \geqslant 0\}$

definition: *degree-ℓ sum-of-squares proof* of the statement

$$\forall X \in S.\ q(X) \geqslant 0$$

expresses $q(X)$ as *sum of nonnegative polynomials over* S with degree $\leqslant \ell$ of the form

$$r(X)^2 \cdot p_1(X)^{b_1} \cdots p_m(X)^{b_m},$$

where $r(X)$ is polynomial and $b_1, \ldots, b_m \in \{0, 1\}$

assuming compactness of S, *every true statement has sum-of-squares proof* (of potentially large degree)

key property: many *"useful" statements require only low degree* (independent of dimension N)

let $S = \{X \in \mathbb{R}^N \mid p_1(X) \geqslant 0, \ldots, p_m(X) \geqslant 0\}$

definition: *degree-ℓ sum-of-squares proof* of the statement

$$\forall X \in S.\, q(X) \geqslant 0$$

expresses $q(X)$ as *sum of nonnegative polynomials over* S with degree $\leqslant \ell$ of the form

$$r(X)^2 \cdot p_1(X)^{b_1} \cdots p_m(X)^{b_m},$$

where $r(X)$ is polynomial and $b_1, \ldots, b_m \in \{0, 1\}$

key property: [Grötschel–Lovász–Schrijver and Shor, Parrilo, Lasserre]
given $p_1(X), \ldots, p_m(X), q(X)$, if degree-$\ell$ proof exists, can find it efficiently by *convex optimization* in time $N^{O(\ell)}$

let $S = \{X \in \mathbb{R}^N \mid p_1(X) \geqslant 0, \ldots, p_m(X) \geqslant 0\}$

definition: *degree-ℓ sum-of-squares proof* of the statement

$$\forall X \in S.\ q(X) \geqslant 0$$

expresses $q(X)$ as *sum of nonnegative polynomials over S* with degree $\leqslant \ell$ of the form

$$r(X)^2 \cdot p_1(X)^{b_1} \cdots p_m(X)^{b_m},$$

where $r(X)$ is polynomial and $b_1, \ldots, b_m \in \{0, 1\}$

key property: [Grötschel–Lovász–Schrijver and Shor, Parrilo, Lasserre]
given $p_1(X), \ldots, p_m(X), q(X)$, if degree-$\ell$ proof exists, can find it efficiently by *convex optimization* in time $N^{O(\ell)}$

underlies computational efficiency of
sum-of-squares meta-algorithm

back to clustering

back to clustering

polynomial constraints: $\vec{p}(X, Y) \geqslant 0$ holds if and only if

back to clustering

polynomial constraints: $\vec{p}(X, Y) \geqslant 0$ holds if and only if

1. $X \in \mathbb{R}^{d \times n}$ has at most k different columns

back to clustering

polynomial constraints: $\vec{p}(X, Y) \geqslant 0$ holds if and only if

1. $X \in \mathbb{R}^{d \times n}$ has at most k different columns
2. $W = Y - X \in \mathbb{R}^{d \times n}$ *"looks Gaussian"*

back to clustering

polynomial constraints: $\vec{p}(X, Y) \geqslant 0$ holds if and only if

1. $X \in \mathbb{R}^{d \times n}$ has at most k different columns
2. $W = Y - X \in \mathbb{R}^{d \times n}$ *"looks Gaussian"*, that is, $\forall t \leqslant \ell$,

$$\frac{1}{n} \sum_{i=1}^{n} w_i^{\otimes t} \approx \mathbb{E}_{w \sim N(0, \mathrm{Id})} w^{\otimes t}$$

back to clustering

polynomial constraints: $\vec{p}(X, Y) \geqslant 0$ holds if and only if

1. $X \in \mathbb{R}^{d \times n}$ has at most k different columns
2. $W = Y - X \in \mathbb{R}^{d \times n}$ *"looks Gaussian"*, that is, $\forall t \leqslant \ell$,

$$\frac{1}{n} \sum_{i=1}^{n} w_i^{\otimes t} \approx \mathbb{E}_{w \sim N(0,\mathrm{Id})} w^{\otimes t}$$

theorem: if $n \geqslant d^{\ell}$, constraints are $\ell \cdot k^{1/\ell}$*-identifying* and the proof requires only *degree* ℓ

polynomial constraints: $\vec{p}(X, Y) \geqslant 0$ holds if and only if

1. $X \in \mathbb{R}^{d\times n}$ has at most k different columns
2. $W = Y - X \in \mathbb{R}^{d\times n}$ *"looks Gaussian"*, that is, $\forall t \leqslant \ell$,

$$\frac{1}{n}\sum_{i=1}^{n} w_i^{\otimes t} \approx \mathbb{E}_{w\sim N(0,\mathrm{Id})}\, w^{\otimes t}$$

theorem: if $n \geqslant d^{\ell}$, constraints are $\ell \cdot k^{1/\ell}$*-identifying* and the proof requires only *degree* ℓ

sum-of-squares meta-algorithm: *existence of low-deg. proof* directly implies *efficient* d^{ℓ^2} *-time estimator with error* $\ell \cdot k^{1/\ell}$

polynomial constraints: $\vec{p}(X,Y) \geqslant 0$ holds if and only if

1. $X \in \mathbb{R}^{d\times n}$ has at most k different columns
2. $W = Y - X \in \mathbb{R}^{d\times n}$ *"looks Gaussian"*, that is, $\forall t \leqslant \ell$,

$$\frac{1}{n}\sum_{i=1}^{n} w_i^{\otimes t} \approx \mathbb{E}_{w\sim N(0,\mathrm{Id})}\, w^{\otimes t}$$

theorem: if $n \geqslant d^{\ell}$, constraints are $\ell \cdot k^{1/\ell}$*-identifying* and the proof requires only *degree* ℓ

sum-of-squares meta-algorithm: *existence of low-deg. proof* directly implies *efficient* d^{ℓ^2} *-time estimator with error* $\ell \cdot k^{1/\ell}$

for $\ell = \log k$, match *optimal statistical error* (at cost of quasi-polynomial time and sample size)

polynomial constraints: $\vec{p}(X, Y) \geqslant 0$ holds if and only if

1. $X \in \mathbb{R}^{d \times n}$ has at most k different columns
2. $W = Y - X \in \mathbb{R}^{d \times n}$ *"looks Gaussian"*, that is, $\forall t \leqslant \ell$,

$$\frac{1}{n} \sum_{i=1}^{n} w_i^{\otimes t} \approx \mathbb{E}_{w \sim N(0, \mathrm{Id})} w^{\otimes t}$$

theorem: if $n \geqslant d^{\ell}$, constraints are $\ell \cdot k^{1/\ell}$*-identifying* and the proof requires only *degree* ℓ

strategy: find *"regular proof"* that constraints are identifying; then, turn this proof *step-by-step* into low-degree one

polynomial constraints: $\vec{p}(X, Y) \geqslant 0$ holds if and only if

1. $X \in \mathbb{R}^{d \times n}$ has at most k different columns
2. $W = Y - X \in \mathbb{R}^{d \times n}$ *"looks Gaussian"*, that is, $\forall t \leqslant \ell$,

$$\frac{1}{n} \sum_{i=1}^{n} w_i^{\otimes t} \approx \mathbb{E}_{w \sim N(0, \mathrm{Id})} w^{\otimes t}$$

theorem: if $n \geqslant d^\ell$, constraints are $\ell \cdot k^{1/\ell}$*-identifying* and the proof requires only *degree* ℓ

key ingredient: *stability of mean* of Gaussian $\mathbf{w} \sim N(0, \mathrm{Id}_d)$ under restrictions

polynomial constraints: $\vec{p}(X, Y) \geqslant 0$ holds if and only if

1. $X \in \mathbb{R}^{d\times n}$ has at most k different columns
2. $W = Y - X \in \mathbb{R}^{d\times n}$ *"looks Gaussian"*, that is, $\forall t \leqslant \ell$,

$$\frac{1}{n}\sum_{i=1}^{n} w_i^{\otimes t} \approx \mathbb{E}_{w\sim N(0,\mathrm{Id})}\, w^{\otimes t}$$

theorem: if $n \geqslant d^\ell$, constraints are $\ell \cdot k^{1/\ell}$*-identifying* and the proof requires only *degree* ℓ

key ingredient: *stability of mean* of Gaussian $\mathbf{w} \sim N(0, \mathrm{Id}_d)$ under restrictions

$$\forall A \subseteq \mathbb{R}^d. \; \left\|\mathbb{E}[\mathbf{w} \mid A]\right\|^2 \leqslant \log(1/\mathbb{P}(A))$$

polynomial constraints: $\vec{p}(X, Y) \geqslant 0$ holds if and only if

1. $X \in \mathbb{R}^{d \times n}$ has at most k different columns
2. $W = Y - X \in \mathbb{R}^{d \times n}$ *"looks Gaussian"*, that is, $\forall t \leqslant \ell$,

$$\frac{1}{n} \sum_{i=1}^{n} w_i^{\otimes t} \approx \mathbb{E}_{w \sim N(0,\mathrm{Id})} w^{\otimes t}$$

theorem: if $n \geqslant d^{\ell}$, constraints are $\ell \cdot k^{1/\ell}$*-identifying* and the proof requires only *degree* ℓ

key ingredient: *stability of mean* of Gaussian $\mathbf{w} \sim N(0, \mathrm{Id}_d)$ under restrictions

$$\forall A \subseteq \mathbb{R}^d. \;\; \left\|\mathbb{E}[\mathbf{w} \mid A]\right\|^2 \leqslant \log(1/\mathbb{P}(A))$$

get bound $\ell \cdot \mathbb{P}(A)^{1/\ell}$ if just *first* ℓ *moments* of $\mathbf{w}$ are Gaussian

polynomial constraints: $\vec{p}(X, Y) \geqslant 0$ holds if and only if

1. $X \in \mathbb{R}^{d \times n}$ has at most k different columns
2. $W = Y - X \in \mathbb{R}^{d \times n}$ *"looks Gaussian"*, that is, $\forall t \leqslant \ell$,

$$\frac{1}{n} \sum_{i=1}^{n} w_i^{\otimes t} \approx \mathbb{E}_{w \sim N(0,\mathrm{Id})} w^{\otimes t}$$

theorem: if $n \geqslant d^{\ell}$, constraints are $\ell \cdot k^{1/\ell}$-*identifying* and the proof requires only *degree* ℓ

key ingredient: *stability of mean* of Gaussian $\mathbf{w} \sim N(0, \mathrm{Id}_d)$ under restrictions

$$\forall A \subseteq \mathbb{R}^d. \ \left\|\mathbb{E}[\mathbf{w} \mid A]\right\|^2 \leqslant \log(1/\mathbb{P}(A))$$

get bound $\ell \cdot \mathbb{P}(A)^{1/\ell}$ if just *first* ℓ *moments* of $\mathbf{w}$ are Gaussian

this proof has degree $\leqslant 2\ell$ (in indicator function of A)

proof-to-algorithm paradigm

estimation problem

proof-to-algorithm paradigm

estimation problem

⇩ 1

low-degree polynomial constraints

proof-to-algorithm paradigm

estimation problem

⇩ **1**

low-degree polynomial constraints

⇩ **2**

property: constraints are ε-identifying

proof-to-algorithm paradigm

estimation problem

⇓ 1

low-degree polynomial constraints

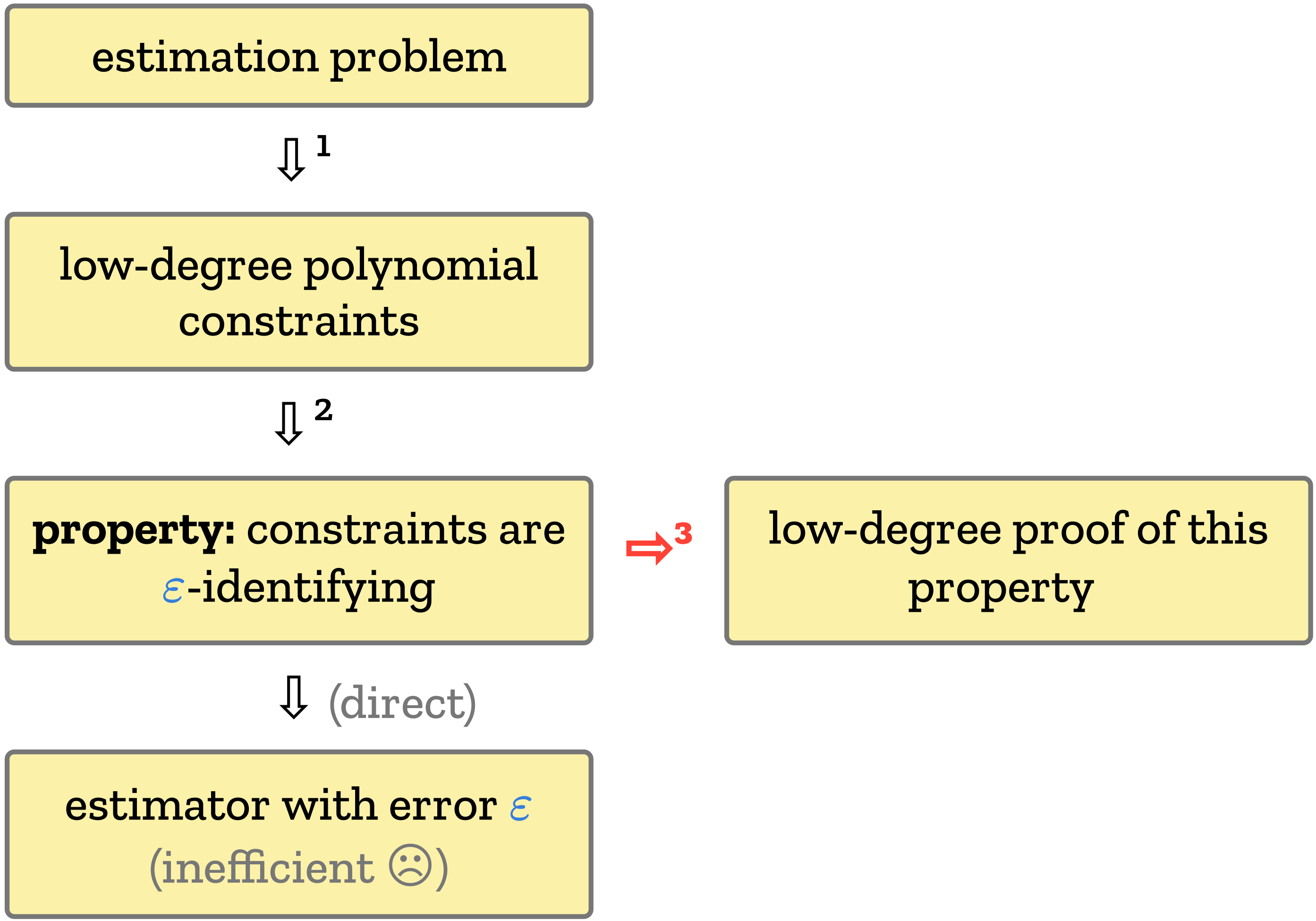
⇓ 2

property: constraints are ε-identifying

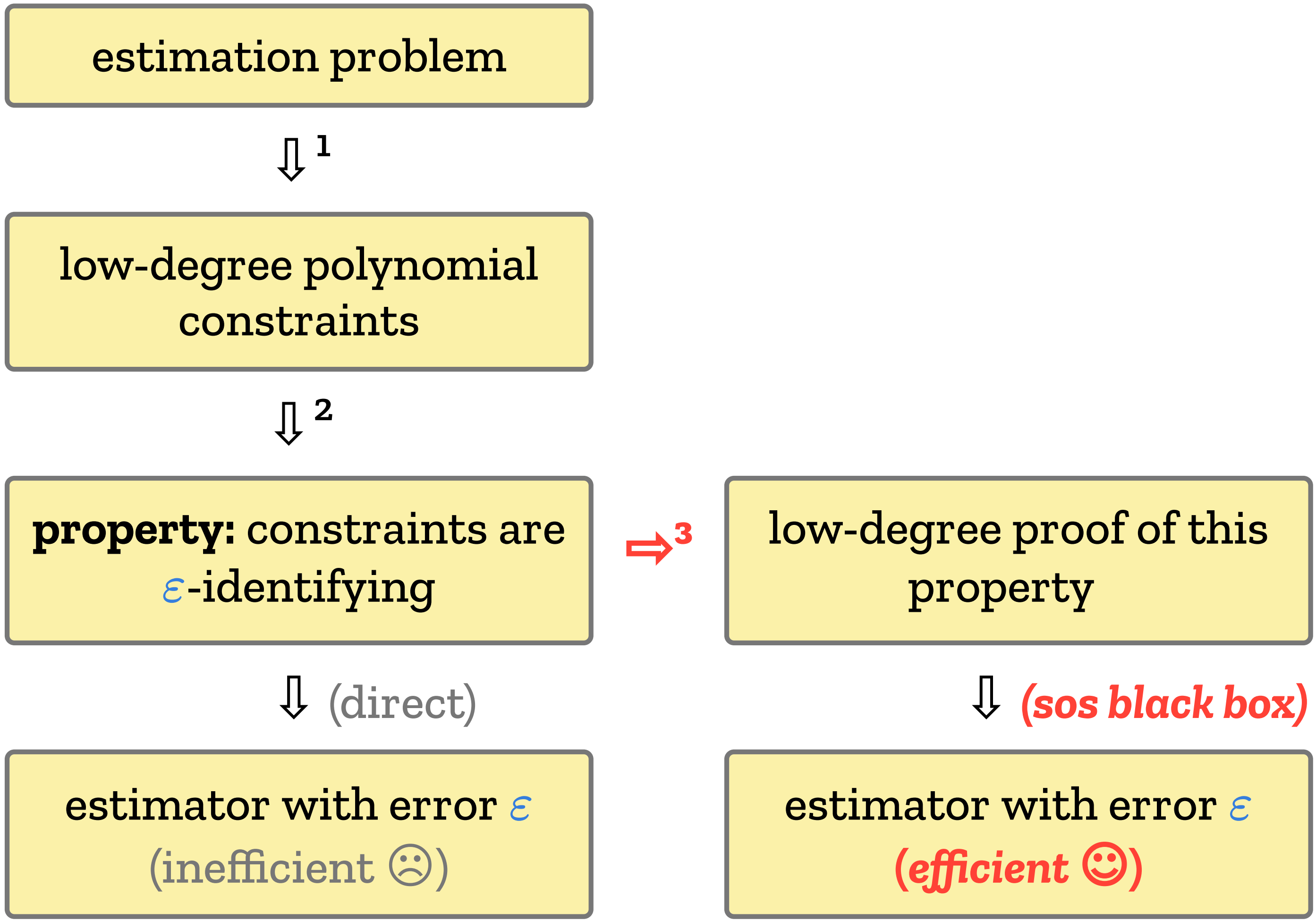
⇓ (direct)

estimator with error ε
(inefficient ☹)

proof-to-algorithm paradigm



proof-to-algorithm paradigm



proof-to-algorithm paradigm

strategy for step 3: turn "regular proof" from step 2 into low-degree proof

estimation problem

⇩ 1

low-degree polynomial constraints

⇩ 2

property: constraints are ε-identifying

⇨ 3

low-degree proof of this property

⇩ (direct)

estimator with error ε (inefficient ☹)

⇩ ***(sos black box)***

estimator with error ε ***(efficient ☺)***

proof-to-algorithm paradigm

estimation problem

⇩ 1

low-degree polynomial constraints

⇩ 2

property: constraints are ε-identifying

⇨ 3 low-degree proof of this property

⇩ (direct)

estimator with error ε (inefficient ☹)

⇩ ***(sos black box)***

estimator with error ε ***(efficient ☺)***

strategy for step 3: turn "regular proof" from step 2 into low-degree proof

growing general toolkit to carry out this modular task

(see www.sumofsquares.org)

proof-to-algorithm paradigm

estimation problem

⇩ 1

low-degree polynomial constraints

⇩ 2

property: constraints are ε-identifying

⇩ (direct)

estimator with error ε (inefficient ☹)

⇨ 3

low-degree proof of this property

⇩ *(sos black box)*

estimator with error ε *(efficient ☺)*

strategy for step 3: turn "regular proof" from step 2 into low-degree proof

growing general toolkit to carry out this modular task (see www.sumofsquares.org)

Thank you!