
On the Least Median Square Problem

Jeff Erickson

University of Illinois

Sariel Har-Peled

University of Illinois

David Mount

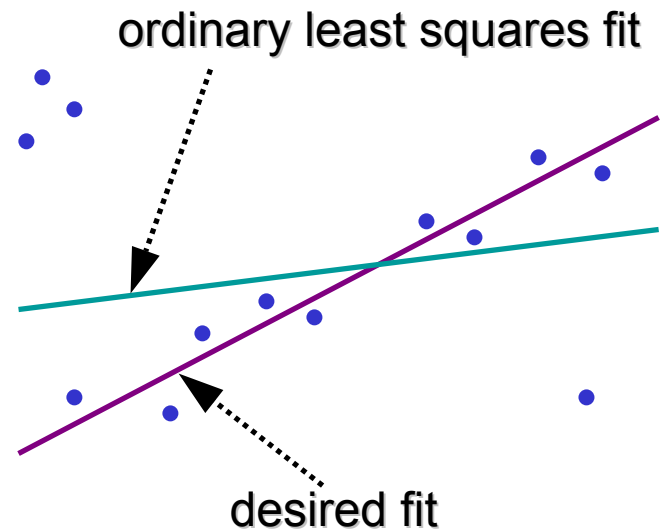
University of Maryland

Robust Linear Regression

Linear Regression: Given a set of n points $P = \{p_1, p_2, \dots, p_n\}$ in \mathbb{R}^d , fit a $(d-1)$ -dimensional hyperplane to these points.

Robust Regression: Some fraction (up to 50%) of points may be arbitrarily far from the hyperplane. Ideally, an estimator should not be biased by these **outliers**.

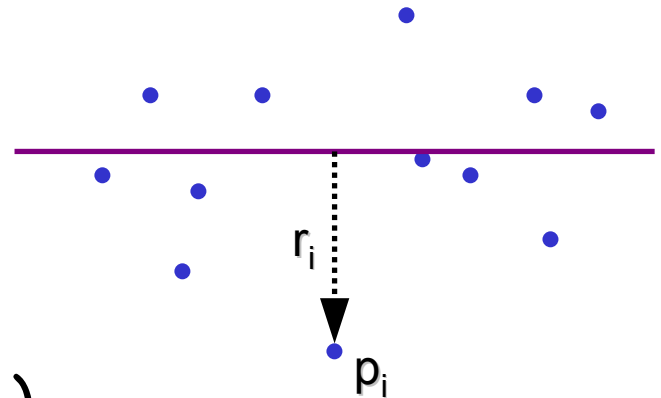
Breakdown Point: The fraction of outliers (up to 50%) that can bias a given estimator.



LMS/LQS Regression

Residual: Given a parameter vector $\theta = (\theta_1, \dots, \theta_d)$, define the i -th residual to be the **vertical distance** from the hyperplane to p_i :

$$r_i = x_{i,d} - (x_{i,1}\theta_1 + \dots + x_{i,d-1}\theta_{d-1} + \theta_d)$$



LMS Estimator (Least Median of Squares): The hyperplane that minimizes the median squared residual [Rou84]. A 50% breakdown estimator.

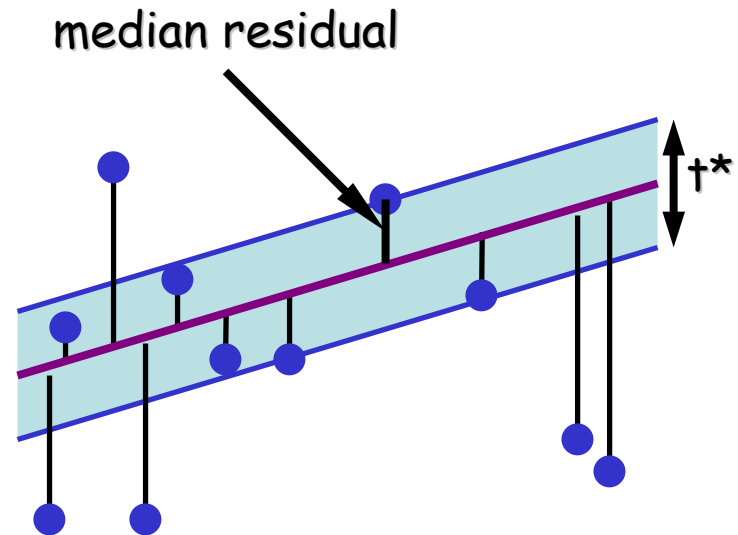
LQS (Least Quantile of Squares): Given integer k , the hyperplane that minimizes the k -th smallest squared residual.

LMS/LQS: Geometric Formulation

Slab: The region bounded by two parallel hyperplanes. LMS is equivalent to computing the **slab** of minimum width that encloses at least 50% of the points.

The **vertical height** t^* of the slab is twice the median absolute residual. The **central hyperplane** of the slab is the LMS estimator.

Vertical height or Perpendicular Width? Our results apply to both cases. We will only present the vertical case.



Prior Results

Exact:

Plane: $O(n^2)$ time and $O(n)$ space by plane sweep.
[Edelsbrunner, Souvaine-90]

d-Space: $O(n^{d+1} \log n)$ by enumeration of elemental sets.
[Rousseeuw, Leroy-87] [Stromberg-93]

Few outliers: $O(n(n-k)^{d+1})$ by LP with few violations [Matousek-95] [Chan-02] and related methods. [H-P, Wang-02]

Approximation: (to the optimum slab height t^*)

Practical Heuristics: Random sampling and branch-and-bound search. [MNPSW-97]

Factor-2 approximation: $O(n^{d-1} \log n)$. [Olson-97]

Our Results

What is the computational complexity of LMS and LQS?

	LMS	LQS (k outliers)	Lower Bound	Prior
Exact	$n^d \log n$		$\Omega(n^d)^*$	$n^{d+1} \log n$ [LR87] $n(n-k)^{d+1}$ [Mat95]
ϵ-Approx	$(n^{d-1}/\epsilon) \log n$	$(n^d/k\epsilon) \log^2 n$	$\Omega(n^{d-1})^*$	

Affine Degeneracy: Given n points, are any $d+1$ coplanar?

***Assumptions:** Affine degeneracy in dimension d requires $\Omega(n^d)$ time, d is a constant, and $\min(k, n-k)$ is $\Omega(n)$.

Overview

Remainder of the presentation:

- Geometric preliminaries
- Exact algorithm for LMS
- ε -Approximation for LMS
- Hardness of exact LMS
- Concluding Remarks

See paper for:

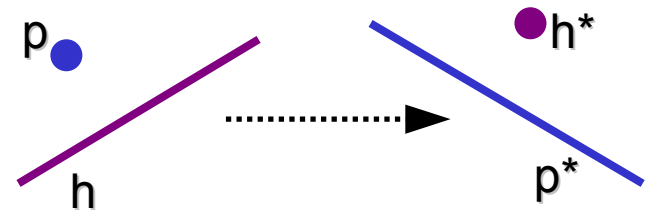
- Generalizations to LQS
- Results for perpendicular slab width
- Hardness results on approximating LMS
- Hardness results for LQS

Geometric Preliminaries

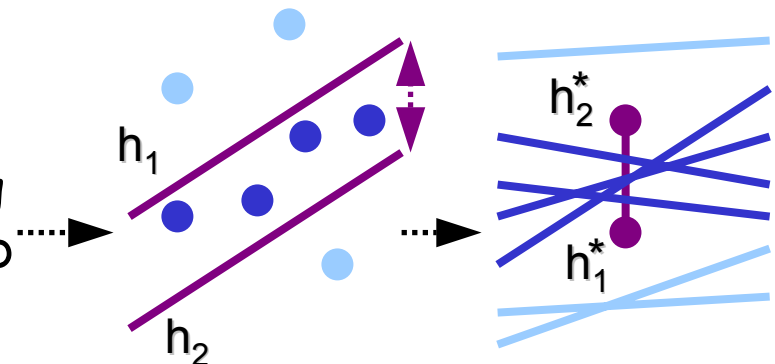
Duality Transformation: Maps point $p=(a_1, \dots, a_d)$ in \mathbf{R}^d to a $(d-1)$ -dim hyperplane:

$$p^*: x_d = a_1 x_1 + \dots + a_{d-1} x_{d-1} - a_d$$

and vice versa.



Slab: The dual of a slab containing k points is a **vertical segment** stabbing k hyperplanes. The height of the slab equals the length of the segment.



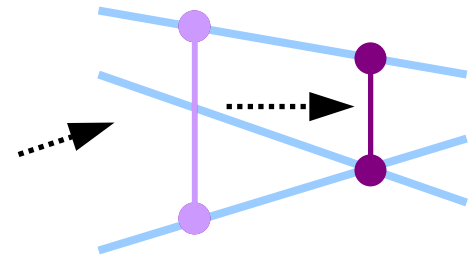
Exact Algorithm for LMS/LQS

Theorem: Given a set H of n hyperplanes in \mathbb{R}^d and an integer k , the shortest vertical segment that stabs k hyperplanes can be computed in $O(n^d \log n)$ time, with high probability.

Approach: Randomized parametric search. Let t^* be the (unknown) length of the shortest such segment.

Decision Problem: Given any length t , determine whether $t < t^*$.

Discrete candidate values: For a segment to be minimal, its endpoints must together be incident to at least $d+1$ hyperplanes. $O(n^{d+1})$ candidates result by considering all subsets of $d+1$ hyperplanes.

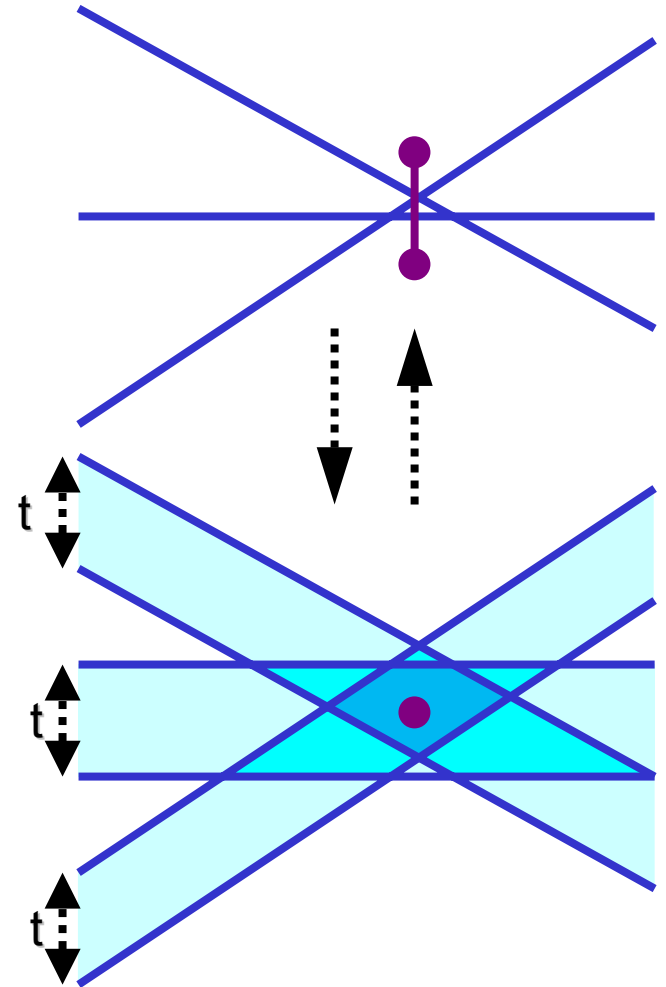


The Decision Procedure

Decision Procedure: Given any length t , in $O(n^d)$ time we can determine whether $t < t^*$.

Proof:

- Replace each hyperplane h of H with a **slab**, bounded by h and a vertical translation of h by t .
- Construct the **arrangement** of these slabs in $O(n^d)$ time.
- Determine whether there is any cell of this arrangement whose **slab depth** is k or more. This is true iff $t \geq t^*$.



Exact Algorithm: Sample and Sweep

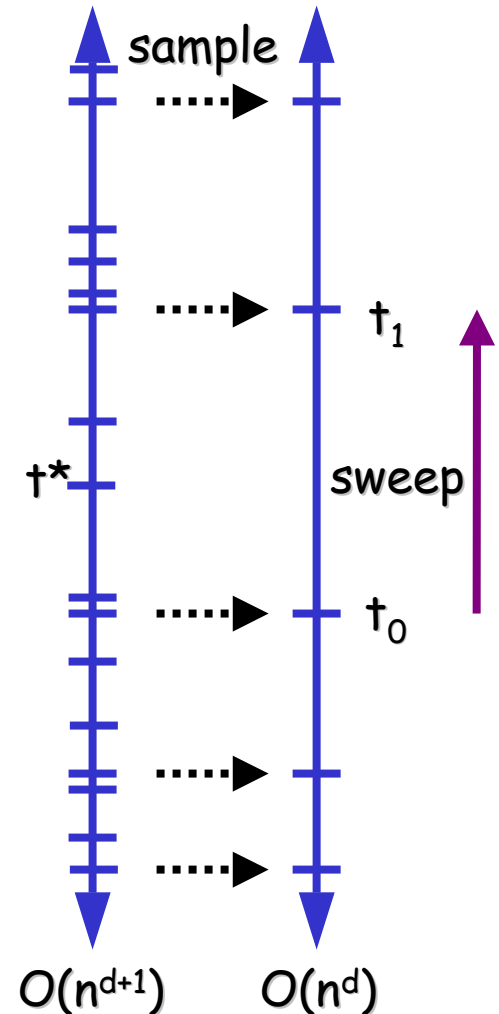
Sample:

- Take a **random sample** of $O(n^d)$ subsets of $(d+1)$ hyperplanes.
- Compute the associated t values.
- Using the **decision procedure** and **binary search**, find consecutive sample values such that t^* lies in the interval $[t_0, t_1]$.
- With high probability, the expected number of candidate values in the interval $[t_0, t_1]$ is:

$$O((n^{d+1} / n^d) \log n) = O(n \log n)$$

Sweep: Consider the **parametric arrangement** of slabs of height t , as t varies over $[t_0, t_1]$. Sweep this arrangement as a function of t .

Total Time: $O(n^d \log n)$.

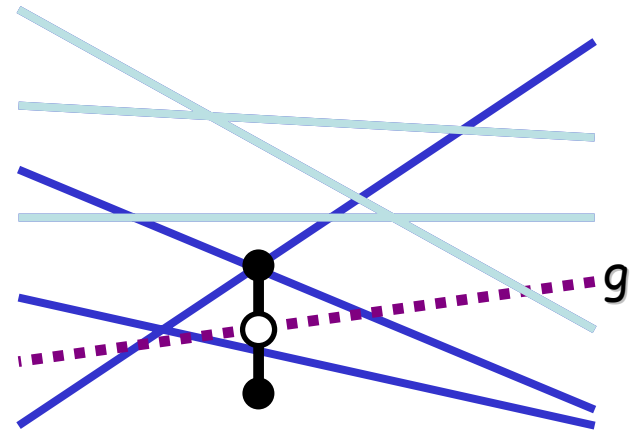


Approximation Algorithm for LMS

Theorem: Given a set of n hyperplanes in \mathbb{R}^d , an integer k and $\varepsilon > 0$, we can compute a vertical segment that stabs $n/2$ hyperplanes whose length is at most $(1+\varepsilon)$ times optimum in $O((n^{d-1}/\varepsilon)\log^2 n)$ time, with high probability.

Approach: Reduce to the following **conditional problem**.

Conditional problem: Given a set H of n hyperplanes in \mathbb{R}^d and a hyperplane g (not necessarily in H), compute the shortest vertical segment that stabs $n/2$ hyperplanes and whose midpoint lies on g .



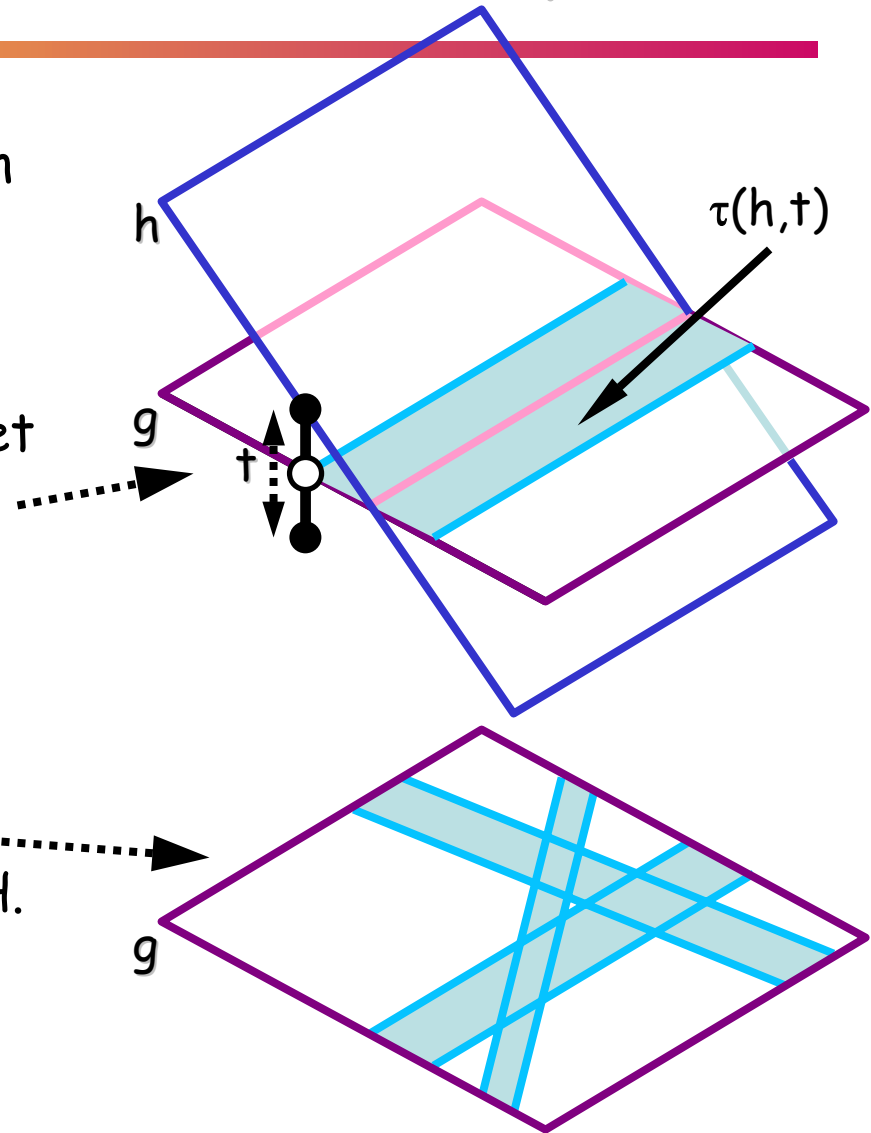
Solving the Conditional Problem

Lemma: The conditional problem can be solved in $O(n^{d-1} \log n)$ time.

Parametric Search: Let t^* be the optimum segment length for the conditional problem. For $h \in H$, let $\tau(h, t)$ be the set of points of g such that a segment of length t centered here stabs h . This is a slab on g .

Decision Problem ($t \geq t^*$): Construct the $(d-1)$ -dimensional arrangement of $\tau(h, t)$ for all $h \in H$. If the slab depth of any point exceeds $n/2$, then $t \geq t^*$.

Sample and sweep: As before.



Approximation Algorithm (cont)

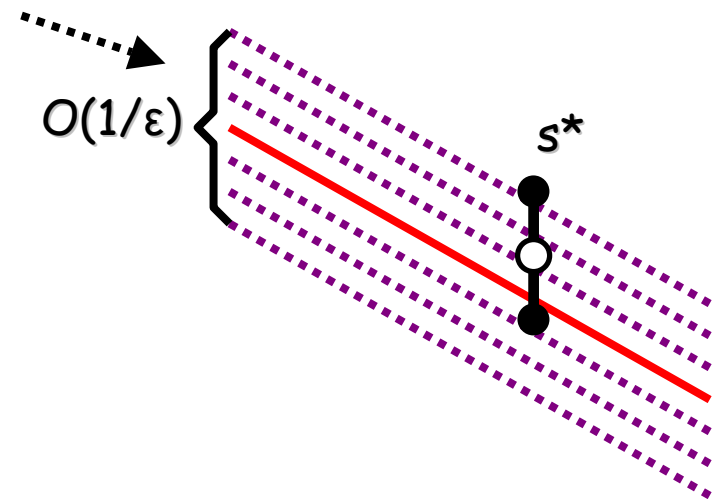
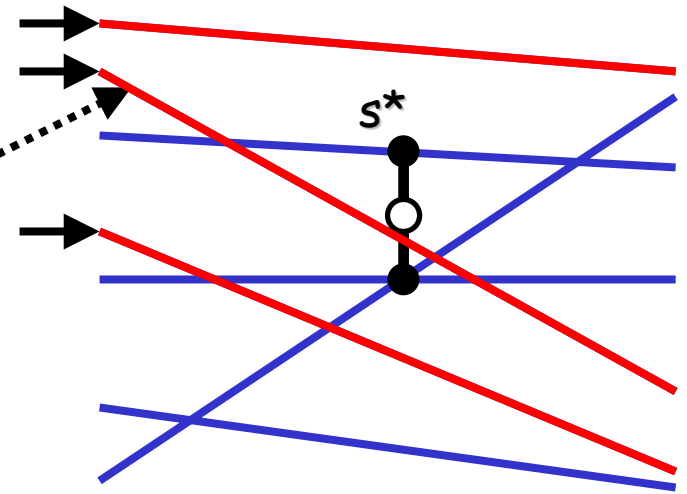
Let s^* be the shortest vertical segment that stabs $n/2$ hyperplanes of H .

Sample a set R of $O(\log n)$ hyperplanes of H . s^* stabs at least one of these with high probability.

Solve the **conditional problem** for each $g \in R$. The overall minimum length t is at most twice optimal, $t \leq 2t^*$.

Let $\delta = \epsilon t/4$. For each $g \in R$, construct $O(1/\epsilon)$ **vertical translates** of g in increments of δ . With high probability, at least one passes within $\epsilon t^*/2$ of the midpoint of s^* .

Solve the **conditional problem** on each such translate. One of the solutions will be the required ϵ -approximation.



Hardness of Exact LMS

Affine Degeneracy (AD):

- Given n points, are any $d+1$ coplanar?
- Conjectured to require $\Omega(n^d)$ time.

Approach:

- We show that AD is **reducible** to LMS in $O(n)$ time, implying that LMS is at least as hard.

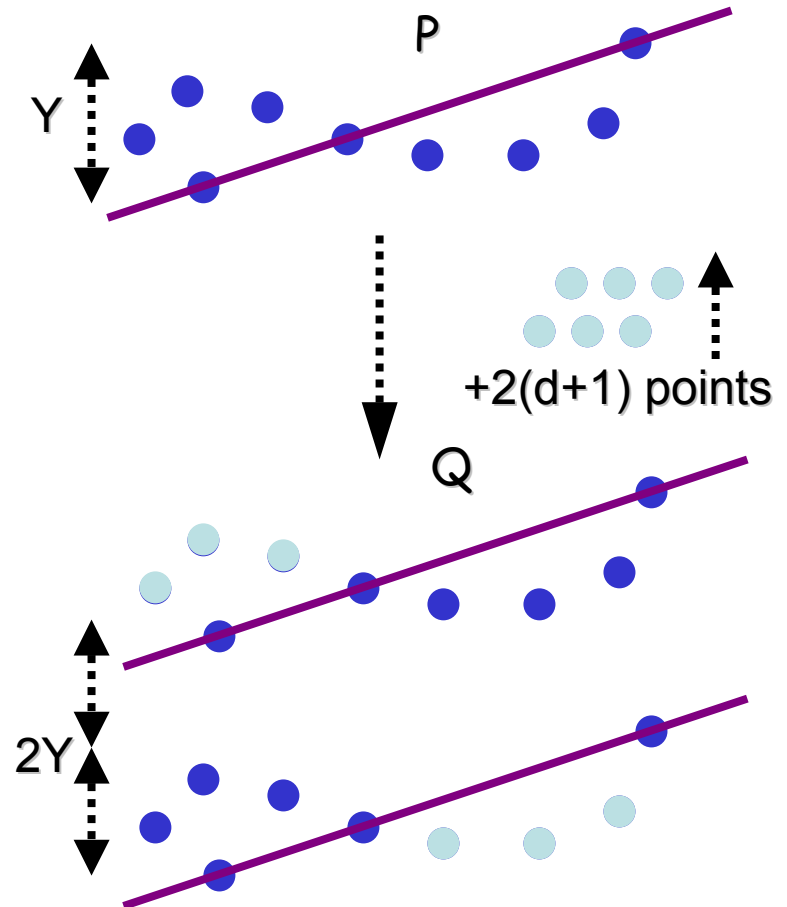
Hardness of Exact LMS

Reduction: Given a point set P of size $m = n/2 - (d+1)$ for AD. Let Y be the height of the set.

Q consists of:

- One copy of P .
- One copy of P translated vertically by $2 \cdot Y$.
- $n - 2m = 2(d+1)$ additional points placed way above.

Correctness: $d+1$ points of P are coplanar **iff** there is a slab containing $m + (d+1) = n/2$ points of Q .



Concluding Remarks

Presented exact and approximation algorithms for LMS and LQS:

- Can solve LMS/LQS in $O(n^d \log n)$ time with high probability .
- An ε -approximation to LMS/LQS in $O((n^d/k\varepsilon) \text{polylog } n)$ time. For fixed ε and $k = \Omega(n)$, this is $O(n^{d-1} \text{polylog } n)$.
- Shown that these running times are within a polylog factor of optimal, assuming the hardness of affine degeneracy.

Open Problems:

- Can space bounds be reduced from $O(n^d)$?
- How practical? Can this be combined with branch-and-bound?
- Applicable to related estimators, such as least trimmed squares (LTS)?

Thank you
