Regression in the Presence of Additive Oblivious Corruptions

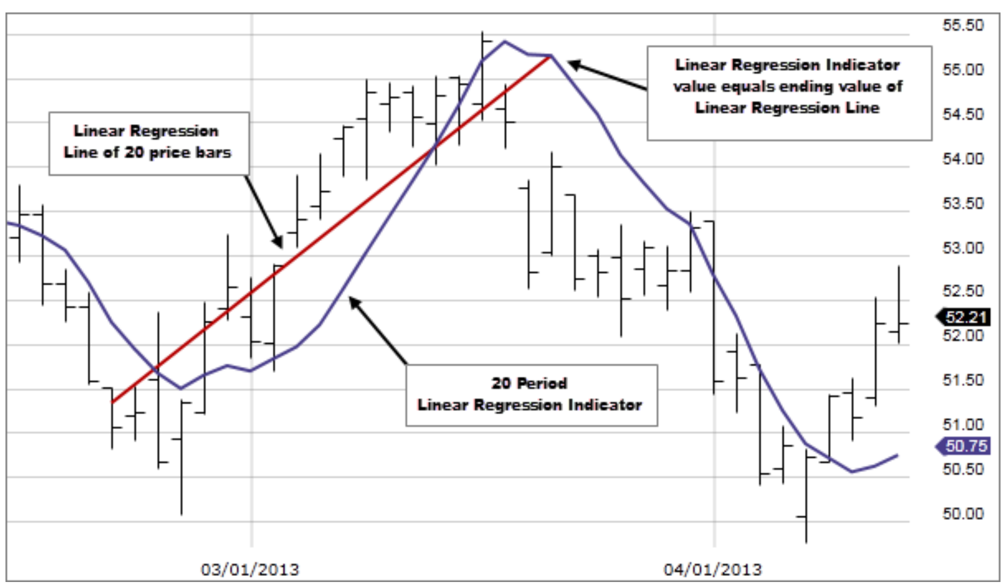
Sushrut Karmalkar
UW-Madison

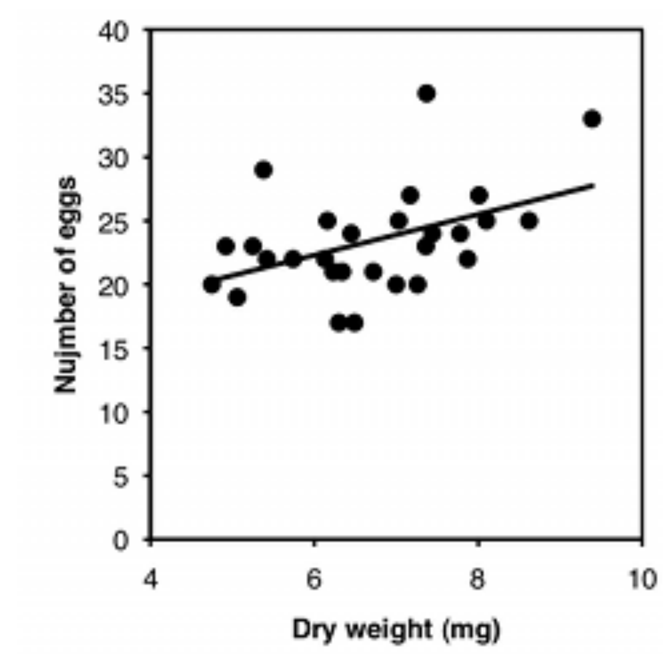
Linear Regression

Given: n samples $\{(x_1, y_1), \ldots, (x_n, y_n)\} \in \mathbb{R}^d \times \mathbb{R}$ s.t.

$$y_i = w^* \cdot x_i + \epsilon_i \text{ where } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Goal: Recover w^*.



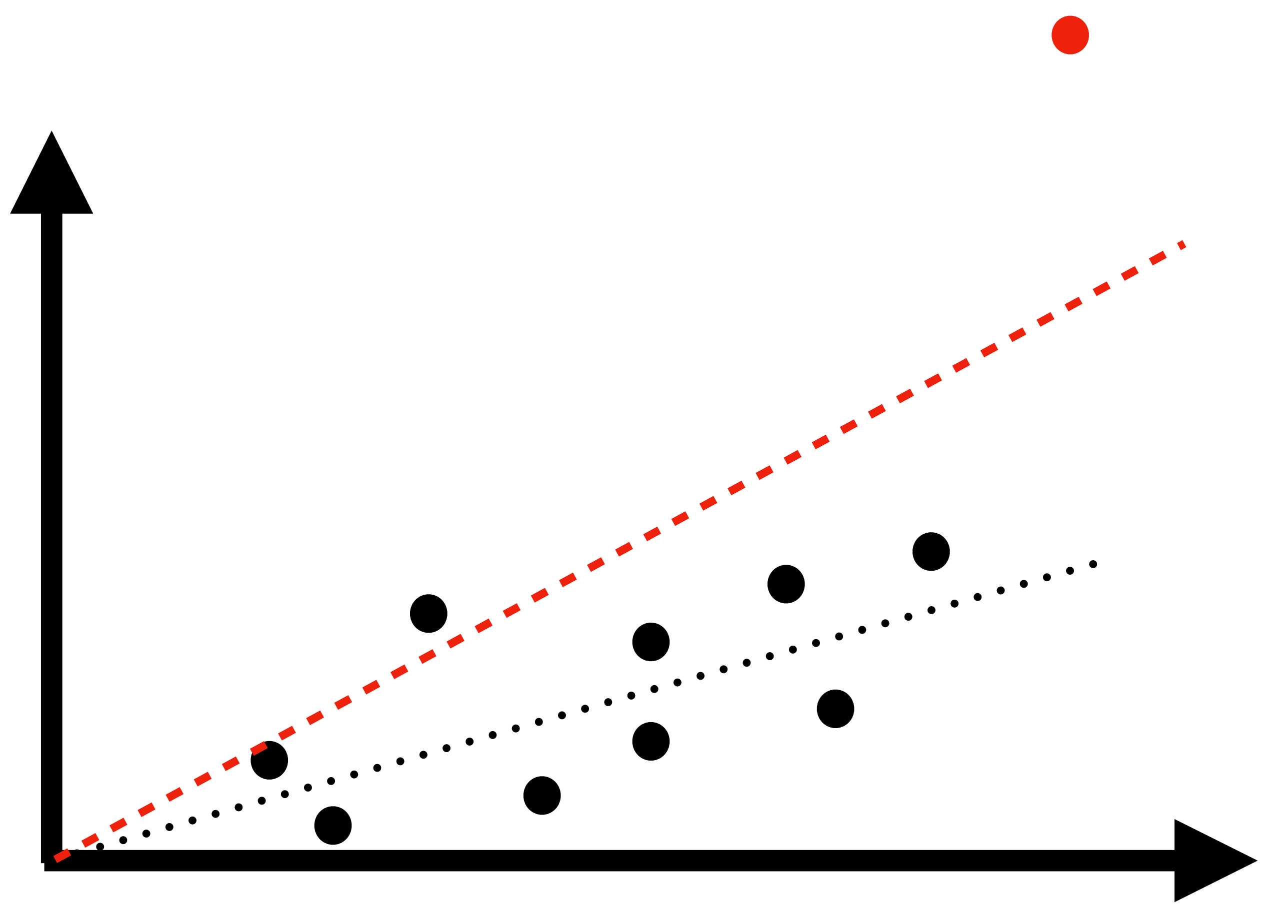


Classic approach: Least Squares Estimator

$$\text{Return the minimizer of } \frac{1}{n}\sum_{i=1}^{n} (y_i - w \cdot x_i)^2$$

Linear Regression

Issue with least squares: Sensitive to even a single outlier!



Can we design efficient and robust estimators?

How do we model corruption?

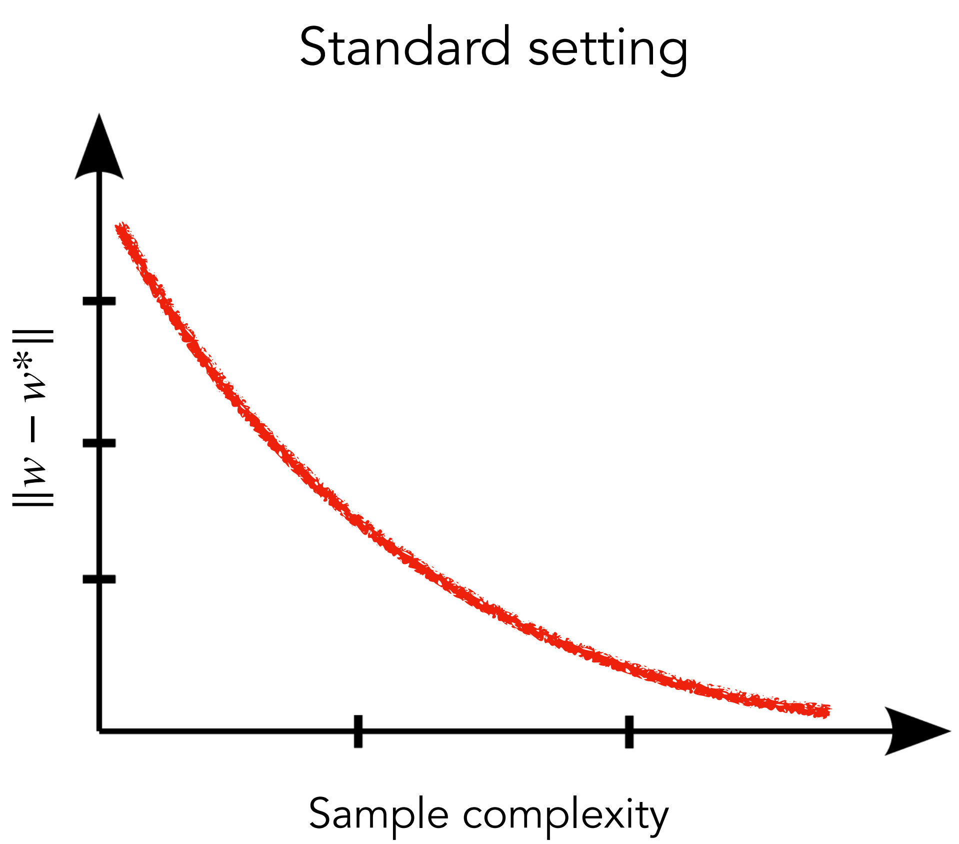
Huber Contamination Model:

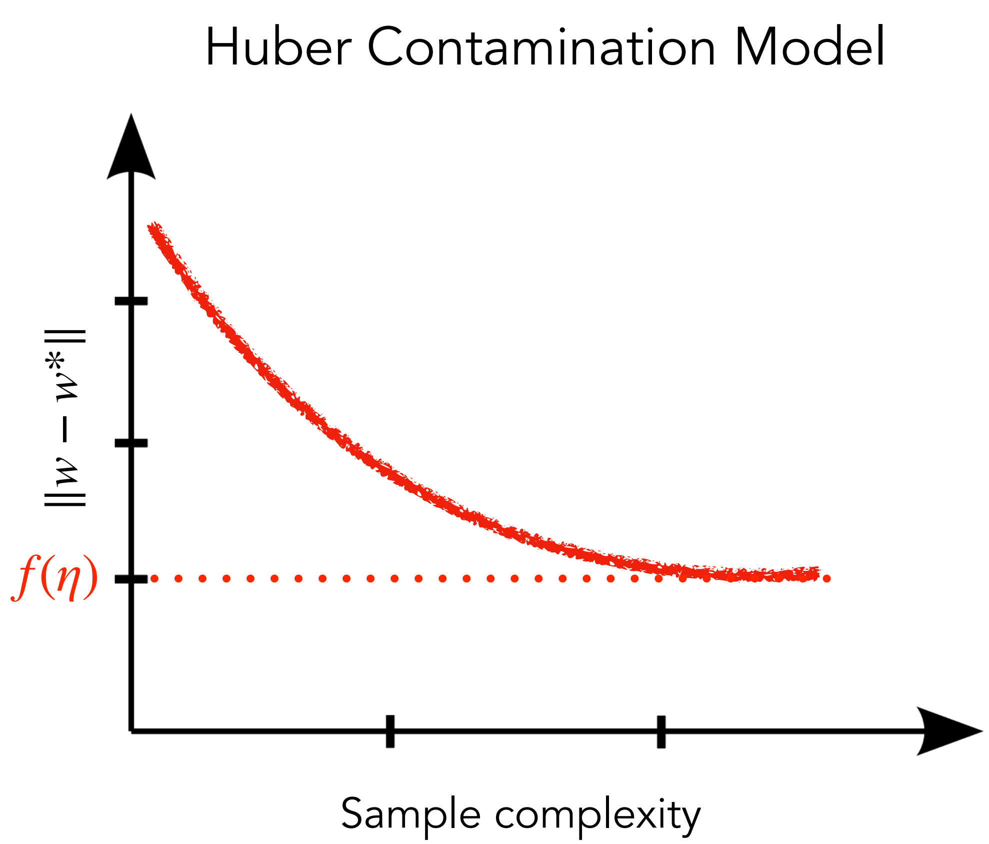
A set of n samples is η**-corrupted** if they are drawn from $(1-\eta)\mathcal{I} + \eta\mathcal{O}$ where,

- $\mathcal{I}$ is the ``inlier distribution'' from some known class of distributions
- $\mathcal{O}$ is an arbitrary and unknown outlier distribution.

Information Theoretic Optimal Error: $\|w - w^*\| \leq O(\sigma\eta)$

Consistency





Algorithm achieving $\approx$ 0 error

Algorithm achieving error $f(\eta) > 0$

Consistency: More data → Improved Accuracy

Is there a setting that allows for the following simultaneously?

- Arbitrary (label) outliers
- Consistency
- Efficient recovery

Oblivious Noise

Given: Independent samples $\{(x_1, y_1), \ldots, (x_n, y_n)\} \in \mathbb{R}^d \times \mathbb{R}$.

$$y_i = w^* \cdot x_i + \epsilon_i + \xi_i$$

where $\xi_i \sim D_\xi$, $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$ drawn i.i.d. and $\Pr[\xi_i = 0] \geq \beta$

Goal: Recover $\hat{w}$ s.t. $\mathbb{E}_x[(\hat{w} \cdot x - w^* \cdot x)^2]$ is small

Measurement Noise

Oblivious Noise

Oblivious Noise

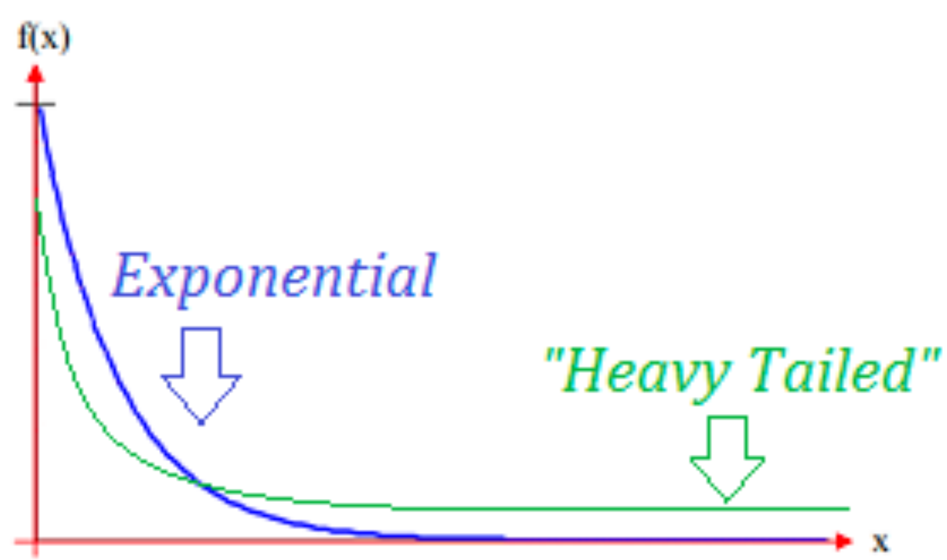
Given: Independent samples $\{(x_1, y_1), \ldots, (x_n, y_n)\} \in \mathbb{R}^d \times \mathbb{R}$.

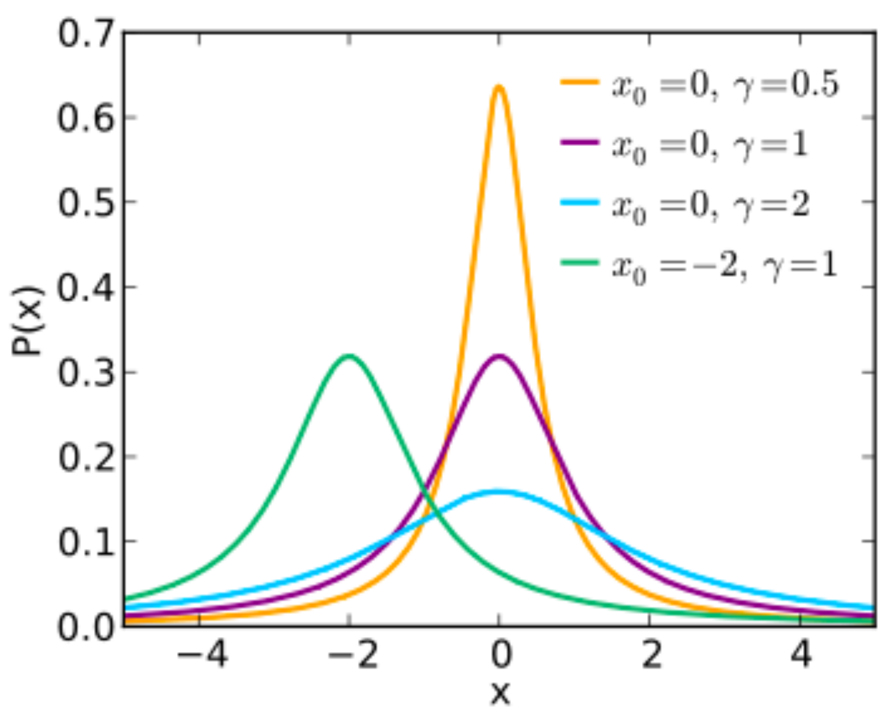
$$y_i = w^* \cdot x_i + \epsilon_i + \xi_i$$

where $\xi_i \sim D_\xi$, $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$ drawn i.i.d. and $\Pr[\xi_i = 0] \geq \beta$

Goal: Recover $\hat{w}$ s.t. $\mathbb{E}_x[(\hat{w} \cdot x - w^* \cdot x)^2]$ is small

Captures a wide range of heavy-tailed and asymmetric noises!





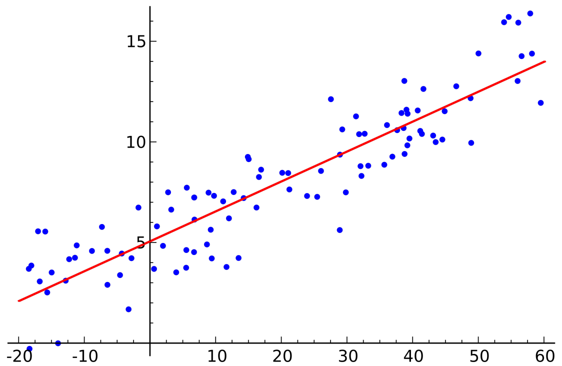
Parameters of Interest

- Inlier probability (β)
- Sample complexity (n) and runtime
- Final error
- Assumptions on noise (ξ) and features.

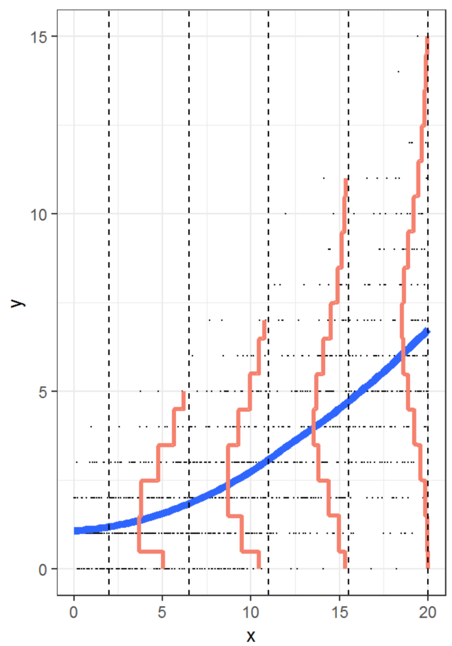
Problems Studied

Supervised Learning

$y = f(w^*, x) + \xi$



Linear Regression



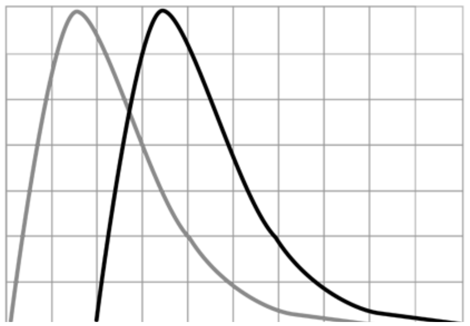
GLM Regression

Unsupervised Learning

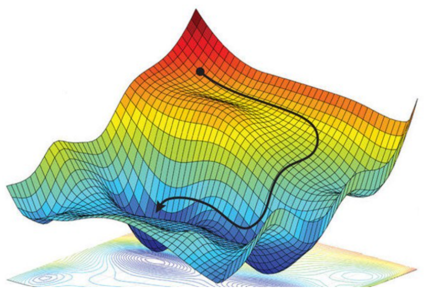
$Y = W^* + \xi$

$\xi \in \mathbb{R}^d$

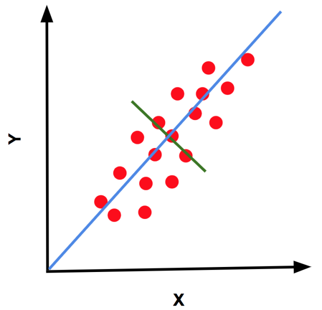
$\Pr[\xi = \vec{0}] \geq \beta$



Location estimation



Stochastic Convex Optimization

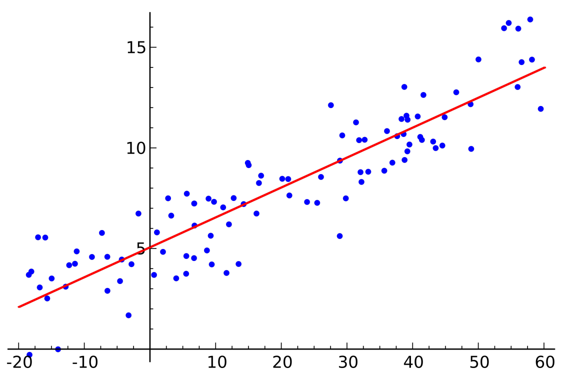


Principal Component Analysis

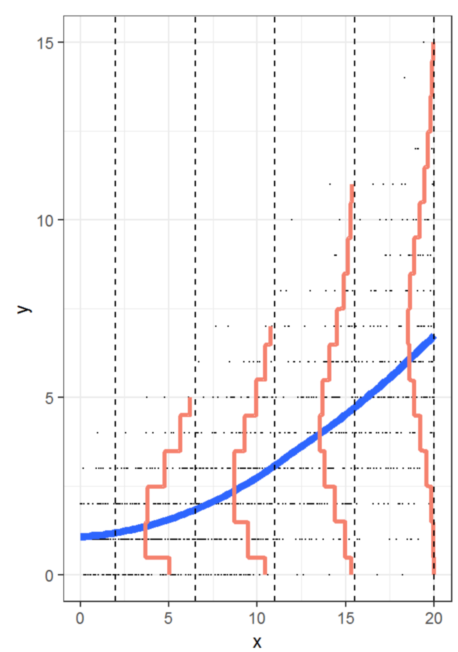
Today

Supervised Learning

$y = f(w^*, x) + \xi$



Linear Regression



GLM Regression

No proofs :(

Discuss **simple algorithms** and some of the core ideas involved

Outline

- Linear Regression with Oblivious Noise
    - Hard-thresholding Based Algorithm
    - Simple(r) algorithms for Gaussian Features
- Learning GLMs with Oblivious Noise

Biased Survey: Linear Regression

- **[Bhatia-Jain-Kamalaruban-Kar'17]:** $\beta \geq 0.99$, $n = \tilde{O}(d)$ and X satisfies some strong-convexity and smoothness conditions.
- **[Suggala-Bhatia-Ravikumar-Jain'19]:** $\beta > 1/\log\log(n)$, $n = \tilde{O}(d)$ same assumptions.
- **[Tsakonas-Jaldén-Sidiropoulos-Ottersten'14]:** $\beta > 1/\sqrt{n}$, but $n = \tilde{O}(d^2)$ and $x \sim \mathcal{N}(0, I_d)$. By minimizing Huber loss.
- **[Pesme-Flammarion'20]:** $x \sim \mathcal{N}(0, I_d)$. First algorithm in the streaming setting (SGD on ℓ_1-loss).
- **[d'Orsi-Novikov-Steurer'21]:** For symmetric oblivious noise and more general feature distributions.
- **[Norman-Weinberger-Levy'22]:** First analysis for $\Sigma \succcurlyeq 0$.

Summary

Paper	Features	Inlier Rate	Error Rate	Estimator
[BJKK'17]	$\mathcal{N}(0,\Sigma); \Sigma \succ 0$ *	> 0.99	$\tilde{O}(d/n\beta^2)$	HT
[SBRJ'19]	$\mathcal{N}(0,\Sigma); \Sigma \succ 0$ *	$> 1/\log\log(n)$	$\tilde{O}(d/n\beta^2)$	HT
[TJSO'14]	$\mathcal{N}(0,I_d)$	$> 1/\sqrt{n}$	$O_{d,\beta}(1/n)$	Huber Loss
[PF'20]	$\mathcal{N}(0,\Sigma); \Sigma \succ 0$	$> 1/\sqrt{n}$	$O(d/n\beta^2)$	L1 Loss
[d'ONS'21]	Non-centered; Mild anticoncentration	$> 1/\sqrt{n}$	$O(d/n\beta^2)$	Huber Loss
[NWL'22]	Subgaussian, $\Sigma \succcurlyeq 0$	$> 1/\sqrt{n}$	$O(d/\beta\sqrt{n})$	Huber Loss

Also results for sparse signals and showing optimality.

* Also for more general classes

Today

Paper	Features	Inlier Rate	Error Rate	Estimator
[BJKK'17]	$\mathcal{N}(0,\Sigma); \Sigma \succ 0$	> 0.99	$\tilde{O}(d/n\beta^2)$	HT
[d'ONS'21]	Non-centered; Mild anticoncentration	$> 1/\sqrt{n}$	$O(d/n\beta^2)$	

Further assume features are Gaussian

Hard-thresholding Based Algorithm

BJKK Theorem

Features: $x \sim \mathcal{N}(0,\Sigma)$

Noise: $\Pr[\xi = 0] \geq \beta \geq 0.99$

Theorem [BJKK'17]: For any $\epsilon, \delta > 0$ and $\beta > 1 - 10^{-5}$, there is a polynomial time algorithm that draws n samples, runs in time poly$(d, n, \log \|\xi\|, \log(1/\epsilon))$ and recovers $\hat{w}$ satisfying

$$\|\hat{w} - w^*\| \leq \epsilon + \tilde{O}_{d,\delta}\left(\frac{\sigma}{\sqrt{\lambda_{min}(\Sigma)}} \cdot \sqrt{\frac{d}{n}}\right)$$

Runtime depends on $\log(\|\xi\|_2)$

Improved in their follow-up work.

[Bhatia-Jain-Kamalaruban-Kar'17]

BJKK Algorithm

Approach: Recover the noise as well as signal.

Problem: $$\min_{w\in\mathbb{R}^d,\ \|\xi\|_0\le(1-\beta)n} \|X^\top w-(y-\xi)\|_2^2 \equiv (1)$$

For a fixed ξ the minimizing w is $w=(XX^\top)^{-1}X(y-\xi)$.

Let $P_X := X^\top(XX^\top)^{-1}X$ and $f(\xi) := \|(I-P_X)(y-\xi)\|_2^2$

$$(1) \equiv \min_{\|\xi\|_0\le(1-\beta)n} f(\xi) \equiv \min_{\|\xi\|_0\le(1-\beta)n} \|(I-P_X)(y-\xi)\|_2^2$$

Algorithm: Gradient-descent on $f(\xi)$ with hard thresholding

BJKK Algorithm

For $v \in \mathbb{R}^n$, $\mathrm{HT}_k(v)$ zeros out the smallest $n-k$ entries of v

$\xi_0 = 0,\ P_X = X^\top (XX^\top)^{-1} X,\ k \geq 2(1-\beta)n$

While $\|\xi^t - \xi^{t-1}\| \geq \tau$

$\quad \xi^{t+1} \leftarrow \mathrm{HT}_k\left(\xi^t - \nabla f(\xi^t)\right)$

Return $w^t \leftarrow (XX^\top)^{-1} X (y - \xi^t)$

Simple(r) Algorithms for Gaussian Features

Assumptions

Assumption: $x \sim \mathcal{N}(0,1)$ and the oblivious noise is symmetric

We can transform the data to satisfy this

- Let $z_i \sim \{+1, -1\}$ uniformly at random
- $y_i \to y_i' = z_i y_i = w^* \cdot (z_i x_i) + (z_i \xi_i) + (z_i \epsilon_i)$
- $x_i \to x_i' = z_i x_i$

Gaussian Features: 1-dimension

Assumption: $x \sim \mathcal{N}(0,1)$ and the oblivious noise is symmetric

Theorem [d'ONS'21]: Given $\tau > 0$, there is an algorithm taking,

- $n \geq \tau/\beta^2$ samples,
- Runs in $O(n)$ time,

And with probability $1 - 2\exp(-\Omega(\tau))$ recovers $\hat{w}$ satisfying,

$$|\hat{w} - w^*|^2 \leq \frac{\tau}{n \cdot \beta^2}.$$

[d'Orsi-Novikov-Steurer'21]

Gaussian Features: 1-dimension

$$(y_i/x_i) = w^* + (\epsilon_i + \xi_i)/x_i.$$

Symmetric

Estimator: $\hat{w} = \text{median}\left(\{y_i/x_i : |x_i| \geq 1/2\}_{i=1}^{n}\right)$

- Anticoncentration: $\Pr_{x_i \sim \mathcal{N}(0,1)}[\,|x_i| \geq 1/2] \geq \Omega(1).$
- $(y_i/x_i) - w^*$ is symmetric and concentrated around 0.

$$\Pr[\,|(\epsilon_i + \xi_i)/x_i| \leq \tau] \geq \Pr[\,|\epsilon_i + \xi_i| \leq \tau/2] \geq \beta\tau/20.$$

What about higher dimensions?

Gaussian Features: d-dimensions

If oblivious noise is symmetric, can extend one-dimensional case

Assumption: $x \sim \mathcal{N}(0, I_d)$ and the oblivious noise is symmetric

Theorem [d'ONS'21]: Given $\Delta > 10 + \|w^*\|$, there is a polytime algorithm that draws $n \geq \tilde{\Omega}_{\Delta,d}(d/\beta^2)$ samples and with probability $1 - d^{-10}$ recovers $\hat{w}$ satisfying

$$\|\hat{w} - w^*\| \leq \tilde{O}\left(\frac{d}{n\beta^2}\right).$$

[d'Orsi-Novikov-Steurer'21]

Ideas

Apply one-d estimator coordinate-wise. For coordinate k,

$$\frac{y_i}{x_k} = w_k^* + \frac{1}{x_k}\left(\epsilon_i + \sum_{j \neq k} w_j^* \cdot x_j + \xi_i\right).$$

Recovers w_k^* to an additive error of $O\left(\frac{(1+\|w^*\|^2)\ \log(d)}{n\beta^2}\right)$.

How do we deal with dependence on $\|w^*\|$?

Bootstrap!

- Let $w^{(i)}$ be the i-th estimate and $\{(x_j', y_j')\}$ be fresh samples.
- Construct $\{(x_j', y_j' - w^{(i)} \cdot x_j')\}$ with signal $w^* - w^{(i)}$ and norm $\ll \|w^*\|/2$.
- Repeat to get improved estimate.

Learning Generalized Linear Models with Oblivious Noise

Regression with Oblivious Noise

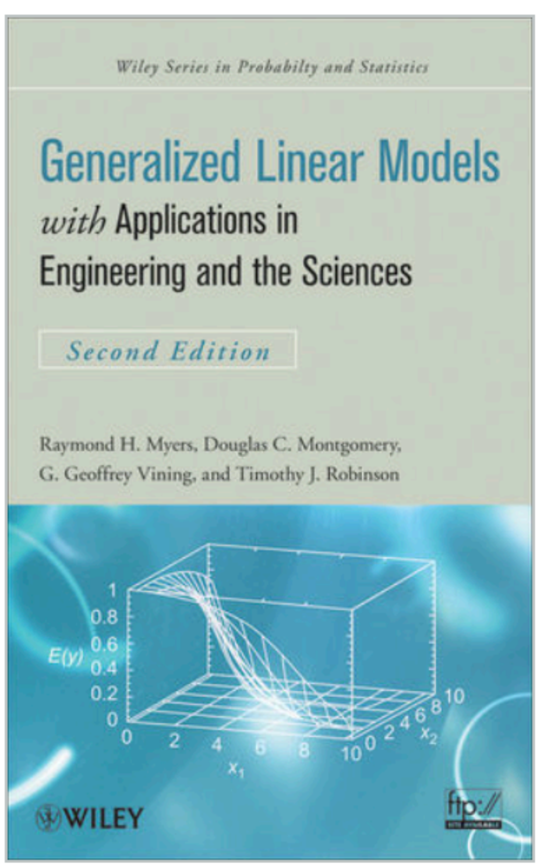
Given: independent samples $\{(x_1, y_1), \ldots, (x_n, y_n)\} \in \mathbb{R}^d \times \mathbb{R}$.

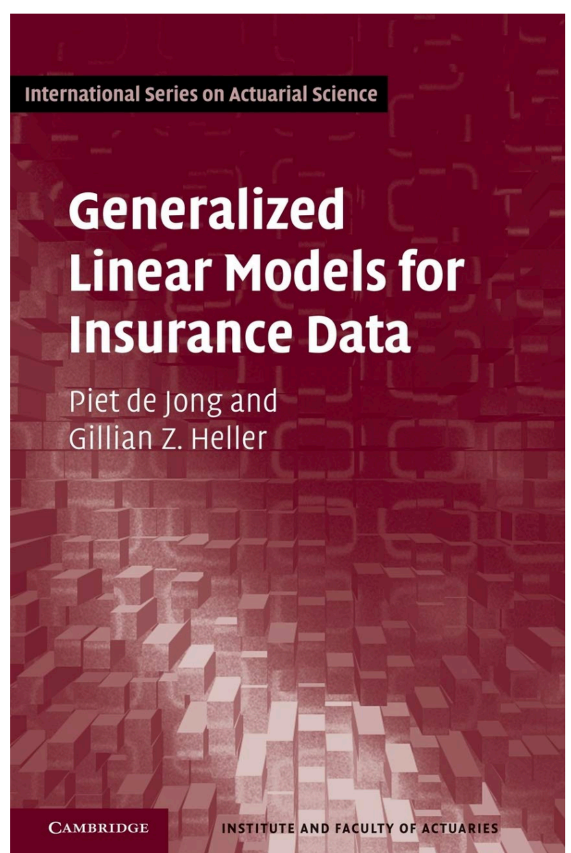
$$y_i = g(w^* \cdot x_i) + \epsilon_i + \xi_i$$

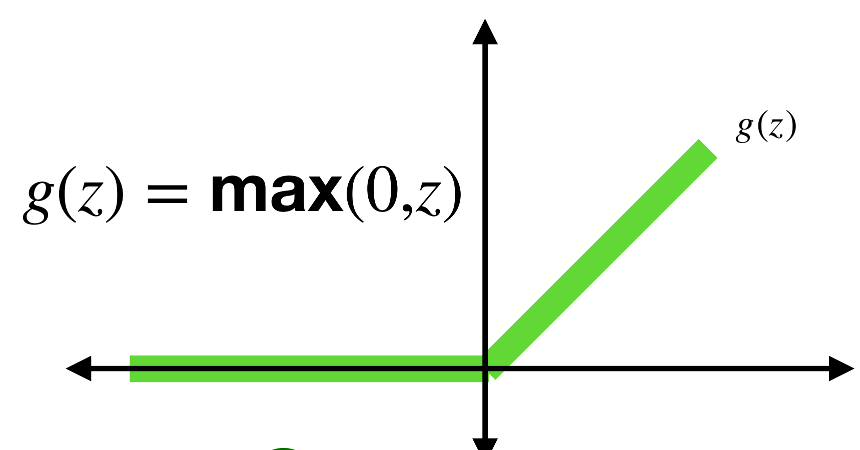
where $\xi_i \sim \mathscr{D}$, $\epsilon_i \sim \mathscr{N}(0, \sigma^2)$ drawn i.i.d. and $\Pr[\xi_i = 0] \geq \beta$

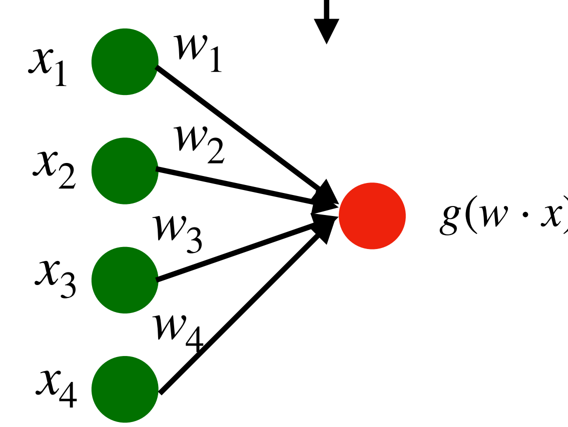
Goal: Recover $\hat{w}$ s.t. $\mathbb{E}_x[(g(\hat{w} \cdot x) - g(w^* \cdot x))^2]$ is small

We assume g (link) is monotonically increasing and Lipschitz









Generality of our setting

Our Goal: First algorithm for GLM regression with oblivious noise s.t. $n \to \infty$ implies error $\to 0$

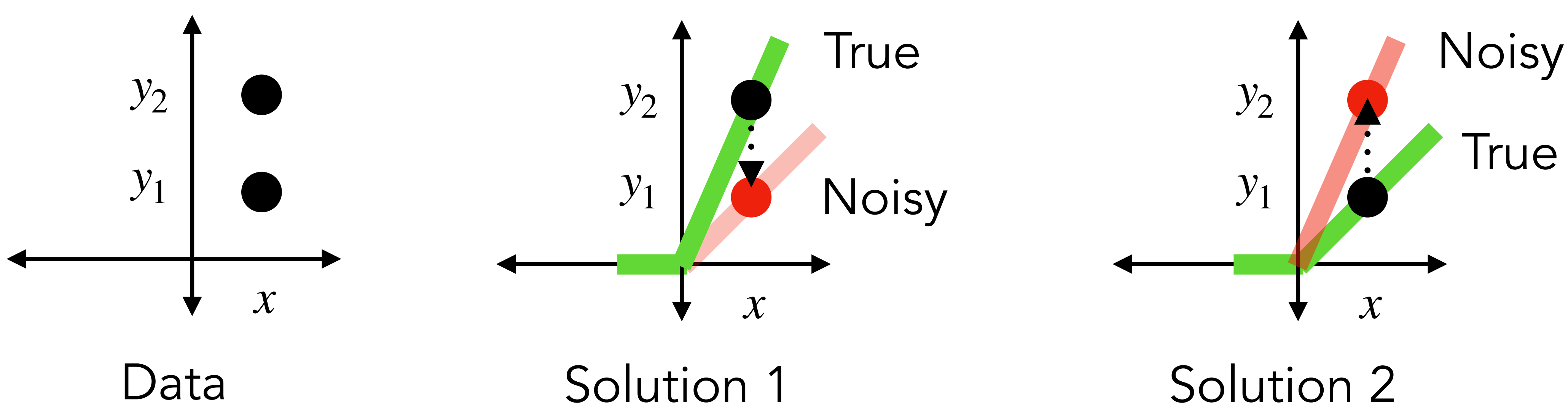
Setting: $\|x\|, \|w^*\| \leq \text{poly}(d)$. No further assumptions on ξ.

Can't symmetrize the noise while preserving the problem

$$-\sigma(w^* \cdot x_j) \neq \sigma(w^* \cdot -x_j)$$

Setting sometimes not uniquely identifiable.

$g(z) = \textbf{max}(0,z) = \textbf{ReLU}(z)$



Generality of our setting

Our Goal: First algorithm for GLM regression with oblivious noise s.t. $n \to \infty$ implies error $\to 0$

Setting: $\|x\|, \|w^*\| \leq$ poly(d). No further assumptions on ξ.

Can't symmetrize the noise while preserving the problem

$$-\sigma(w^* \cdot x_j) \neq \sigma(w^* \cdot -x_j)$$

Setting sometimes not uniquely identifiable.

↓

In this case, we output a list!

Our Result

Theorem [DKPT'23]: There exists an algorithm which,

- Draws polynomially many samples.
- Runs in polynomial time.
- **If uniquely identifiable:** Recovers an estimate for $g(w^* \cdot x)$
  Else: returns a list containing an estimate for $g(w^* \cdot x)$.

Today:

- What to do when median$(\xi) = 0$.
- How we prune candidates.

[Diakonikolas-K-Park-Tzamos'23]

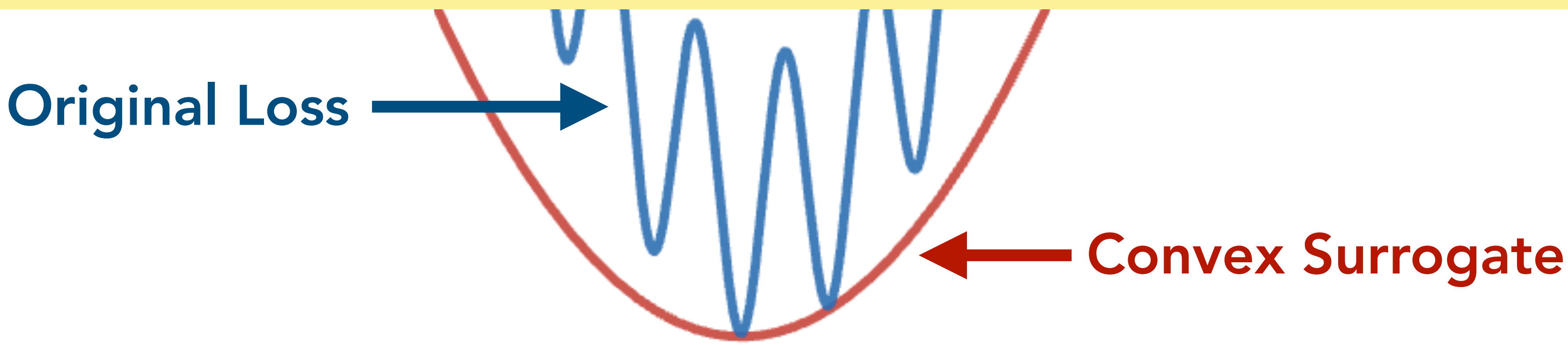
Median 0 Oblivious Noise

Without $g(\cdot)$: minimize ℓ_1-loss$(w) = \frac{1}{n}\sum_i |w \cdot x_i - y_i|$

What happens when g comes into the picture?

g makes standard losses non-convex (e.g. $\frac{1}{n}\sum_i |g(w \cdot x_i) - y_i|$)

Landscape Design: Find a convex surrogate for nonconvex loss.



$$\frac{1}{n}\sum_i (g(w \cdot x_i) - y_i)^2 \to \frac{1}{n}\sum_i \left(\int_0^{w \cdot x_i} g(t) - y_i \, dt\right)$$

Squared loss → Matching loss*

*Dating back to Auer, Herbster, Warmuth'95

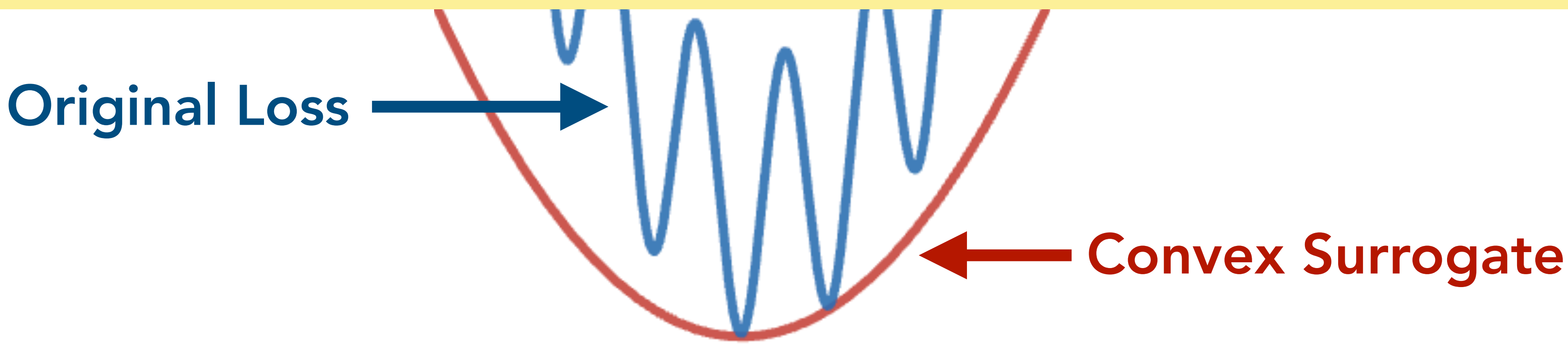
Median 0 Oblivious Noise

Without $g(\cdot)$: minimize ℓ_1-loss$(w) = \frac{1}{n}\sum_i |w \cdot x_i - y_i|$

What happens when g comes into the picture?

g makes standard losses non-convex (e.g. $\frac{1}{n}\sum_i |g(w \cdot x_i) - y_i|$)

Landscape Design: Find a convex surrogate for nonconvex loss.



Solution: Find w minimizing $\frac{1}{n}\sum_i \int_0^{w \cdot x_i} \text{sign}(g(t) - y_i)\, dt$

median(ξ)≠0: Family of similar losses + pruning procedure

One for each possible quantile

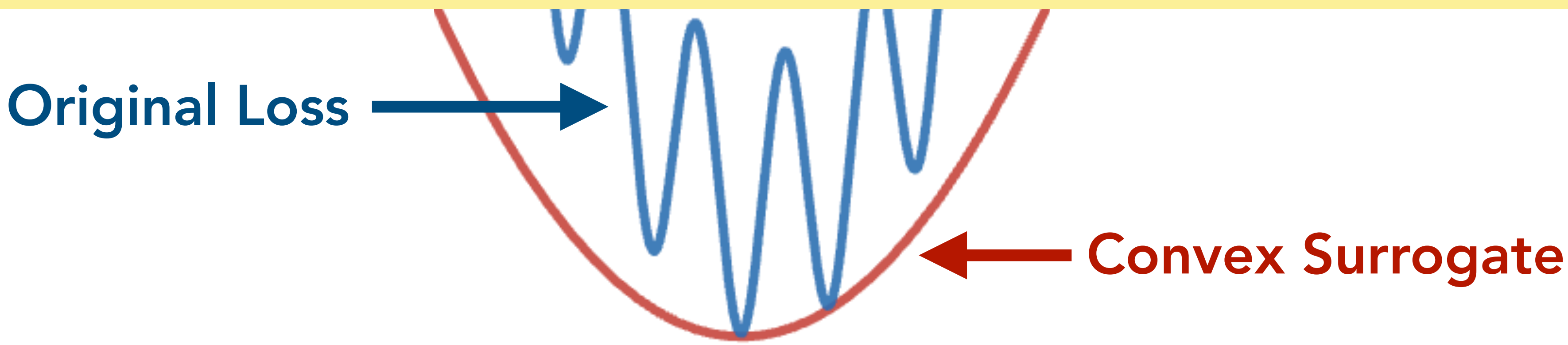
Median 0 Oblivious Noise

Without $g(\cdot)$: minimize ℓ_1-loss$(w) = \frac{1}{n}\sum_i |w \cdot x_i - y_i|$

What happens when g comes into the picture?

g makes standard losses non-convex (e.g. $\frac{1}{n}\sum_i |g(w \cdot x_i) - y_i|$)

Landscape Design: Find a convex surrogate for nonconvex loss.



Solution: Find w minimizing $\frac{1}{n}\sum_i \int_0^{w \cdot x_i} \text{sign}(g(t) - y_i)\, dt$

median(ξ)≠0: Family of similar losses + **pruning procedure**

How do we prune?

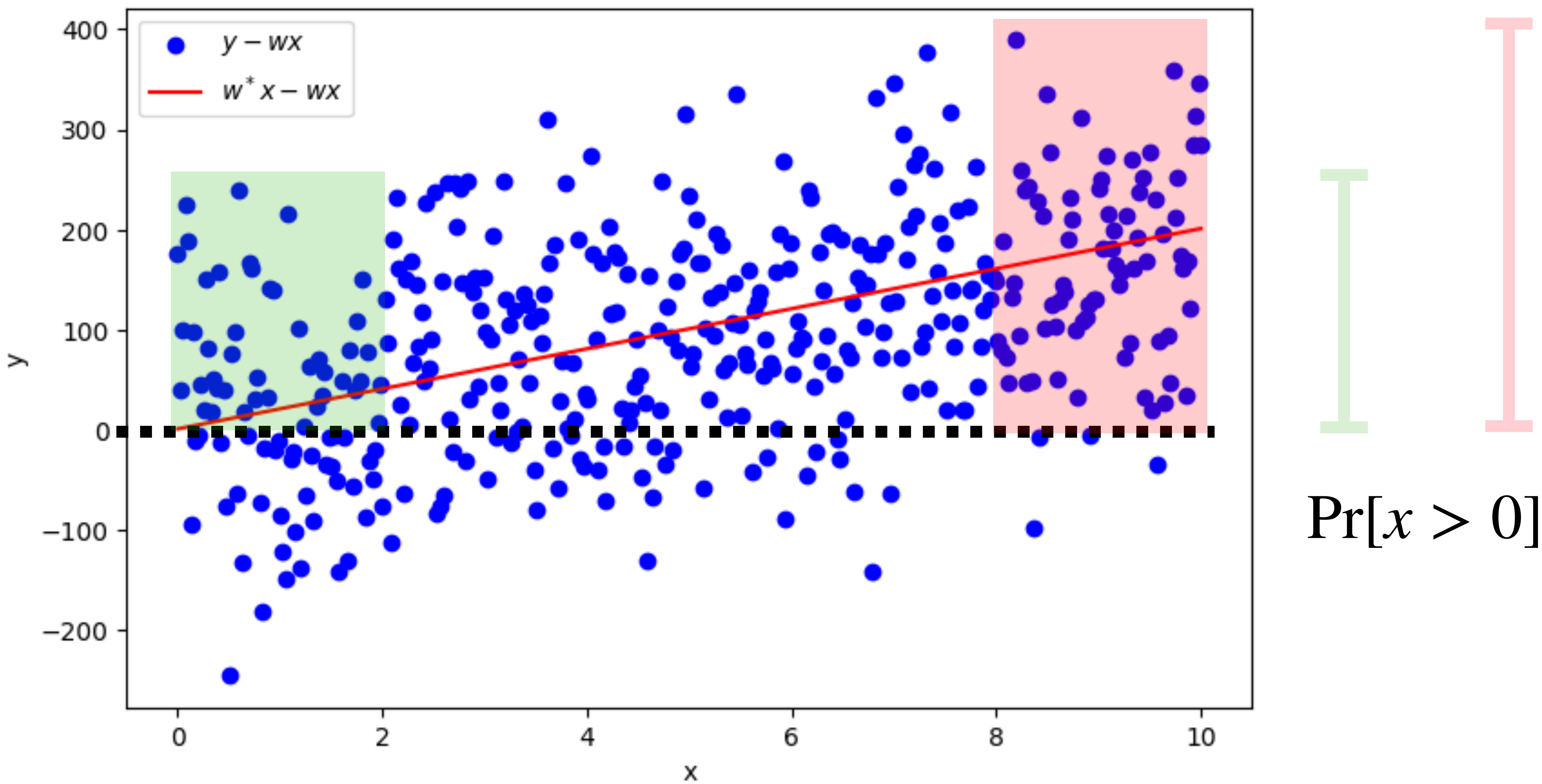
One-dimensional Pruning

Stylized one-dimensional setting:

$g(t) = t$, $\sigma = 1$ and $\text{pdf}_{D_x}(x) \geq c$ for $x \in (8,10) \cup (0,2)$

Given w, how do we check that w is a solution?

Based on quantiles of $y_i - w \cdot x_i = (w^* - w) \cdot x_i + (\xi_i + \epsilon_i)$.



$\Pr[x > 0]$

High-dimensional Pruning

In higher dimesnions not as clear which regions to condition on

Stylized setting: Assume x is anticoncentrated

Given: $L = \{w_1, \ldots, w_q\}$ such that $w^* \in L$

Recover: w^* from L.

Tournament-style algorithm:

- For each $w, w' \in L$:
  Partition $\mathbb{R}^d$ depending on value of $v(x) := (w \cdot x) - (w' \cdot x)$.
- Prune if you can identify 2 regions s.t. the quantiles are sufficiently different.
- Since $w^* \in L$, if w is to be eliminated, such regions will be identified.

Summary

- Oblivious noise: Captures a broad range of additive independent noise models.
- Today: A biased subsampling of the literature and a result on GLMs with oblivious noise.
- Open questions:
  - What are the optimal rates for learning GLMs with oblivious noise?
  - Open questions in the context of location estimation, stochastic convex optimization, etc.