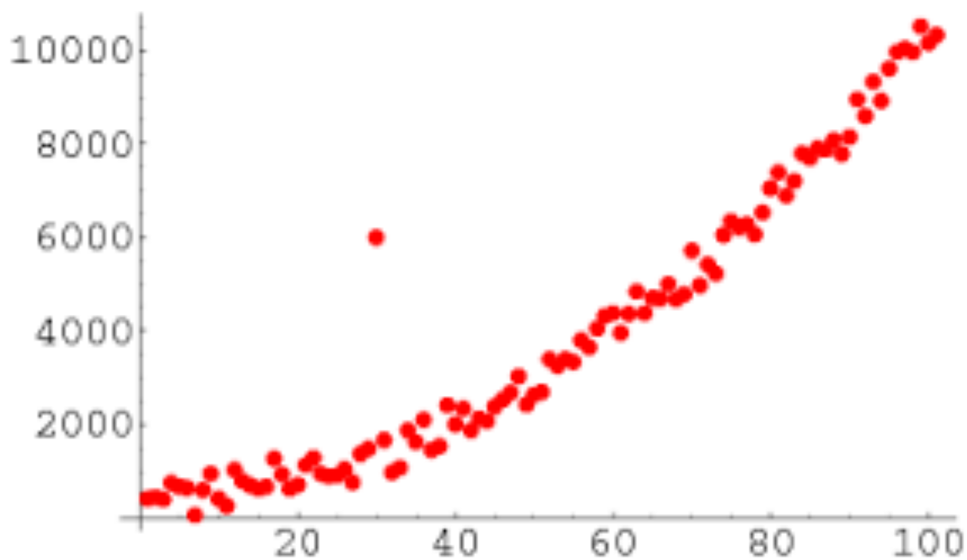
CSI 436/536
Introduction to Machine Learning

Robust regression, RWLLSE, RANSAC

Professor Siwei Lyu
Computer Science
University at Albany, State University of New York

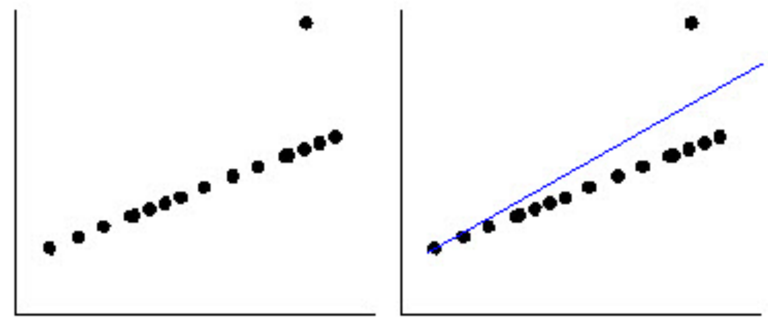
Handling outliers

- Outliers are common in real-life data
  - No formal definition of outliers exists, but they are usually defined as a small fraction of data that deviate from the common noise model
- Sources of outliers are usually input mistakes

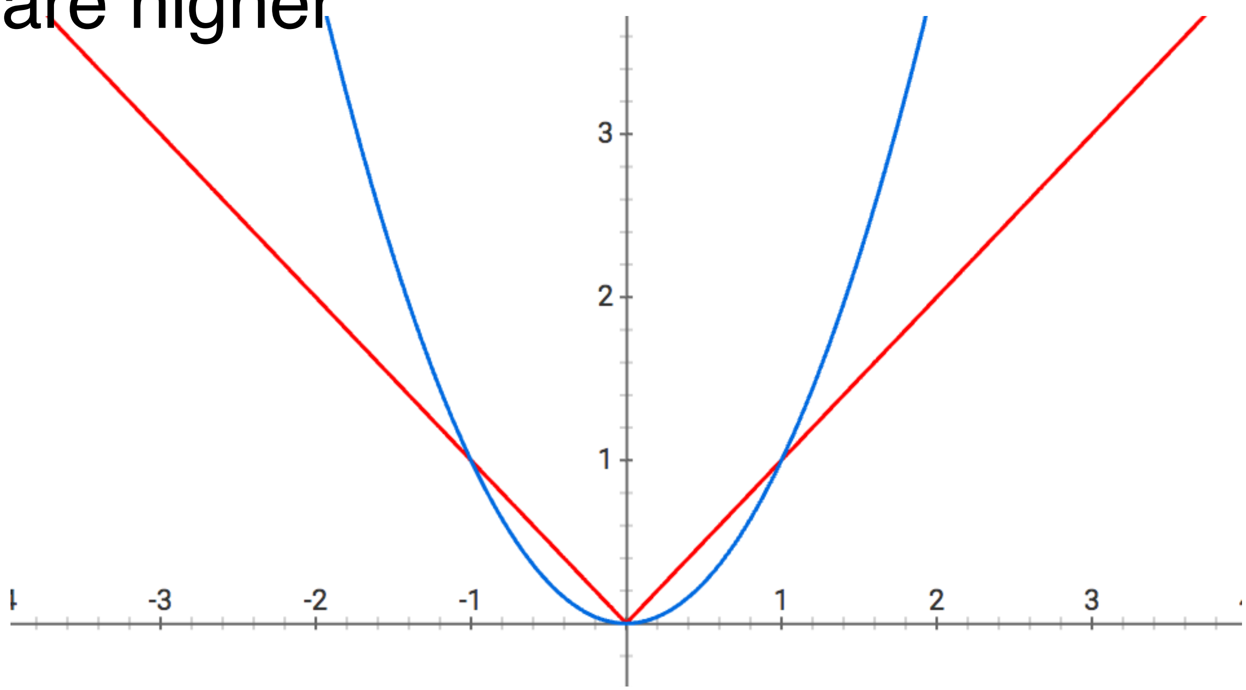


Issues

- Squared L2 loss is sensitive to outliers in training data
  - Using L1 loss is more robust to outliers in training data



- Squared L2 loss gives much larger penalty to large values with the same deviation, so the weight of outliers are higher



Minimizing L1 loss for regression

- Known as minimizing absolute difference (MAD) optimization
  - LLSE loss: $\min_w \|y - X^T w\|_2^2$
  - LMAD loss: $\min_w \|y - X^T w\|_1$
- Problem: how to minimize the LMAD objective function, given that L1 norm is not differentiable
- Non-differential optimization problem is a very important topic in machine learning
- L1 norm is a very commonly used non-differential loss function (we will see it again later in LASSO)

LMAD optimization approaches

- LMAD loss: $\min_w \|y - X^T w\|_1$
  - Direct reformulation
    $\min_{w,c} 1^T c$, s.t. $c \geq 0, -c \leq y - X^T w \leq c$
    this is a linear programming (LP) problem
  - Sub-gradient method — extend the gradient to non-differentiable functions
  - Splitting method: write the objective function
    $\min_{w,z} \|z\|_1 + \frac{\lambda}{2}\|y - X^T w - z\|_2^2$
    this will lead to soft-thresholding operation and we will see it again in LASSO
  - Smoothing method: approximate L1 norm with a differentiable function

Smoothing

- A simple math fact: $$|x| = \min_{z>0} \frac{1}{2}\left(\frac{x^2}{z} + z\right)$$
- Extend this to L1 norm

$$\|x\|_1 = \min_{z>0} \frac{1}{2} \sum_{i=1}^{n} \left(\frac{x_i^2}{z_i} + z_i\right)$$

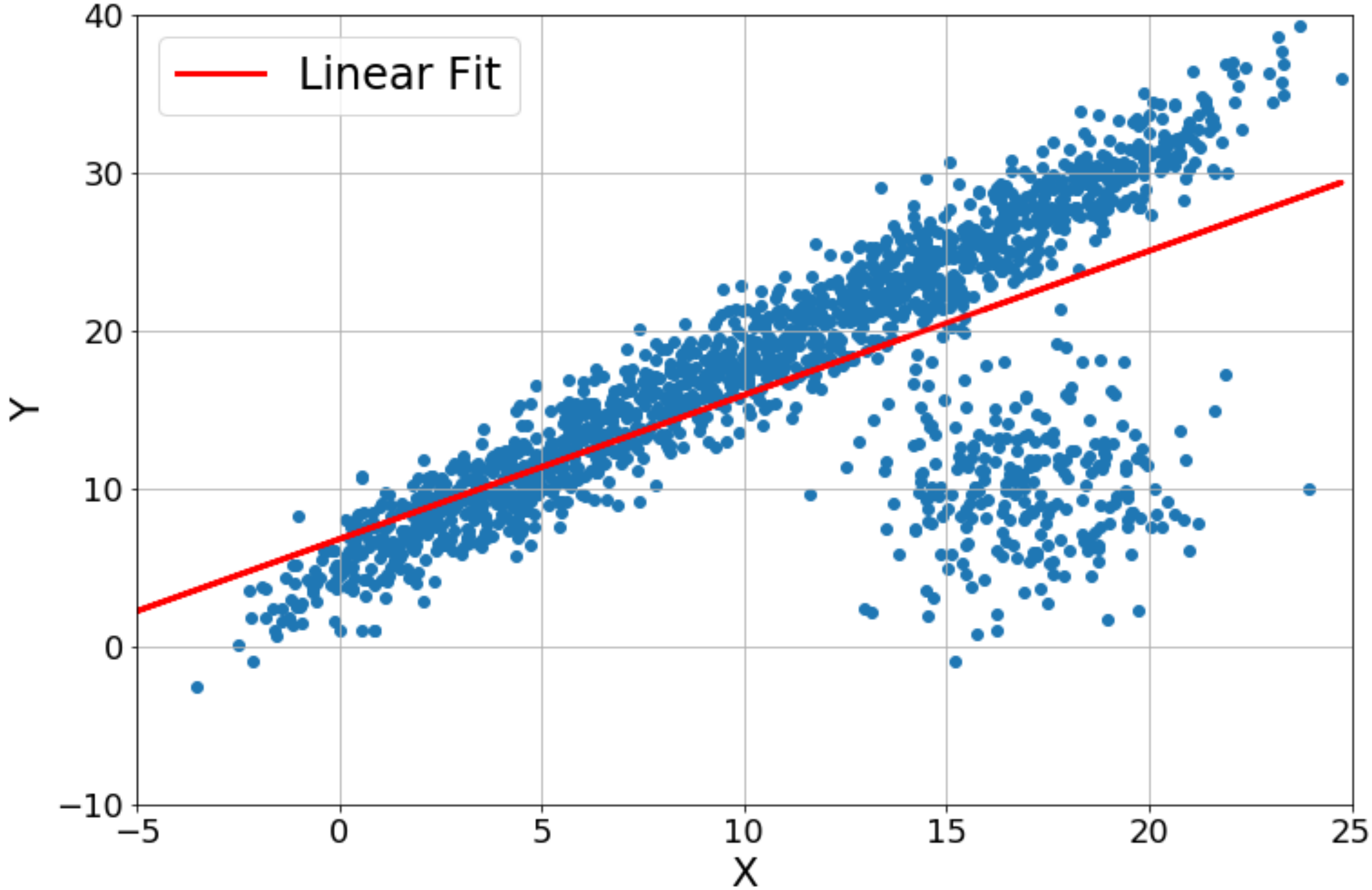
- So we have

$$\min_w \|y - X^T w\|_1 = \frac{1}{2} \min_{w,z} (y - X^T w)^T \text{diag}(z)^{-1} (y - X^T w) + 1^T z$$

  - We iterative solving w and z (known as coordinate descent method)
  - Solving w: a weighted LLSE
  - Solving z: $z_i = |x_i|$
- This algorithm is known as reweighed LLSE

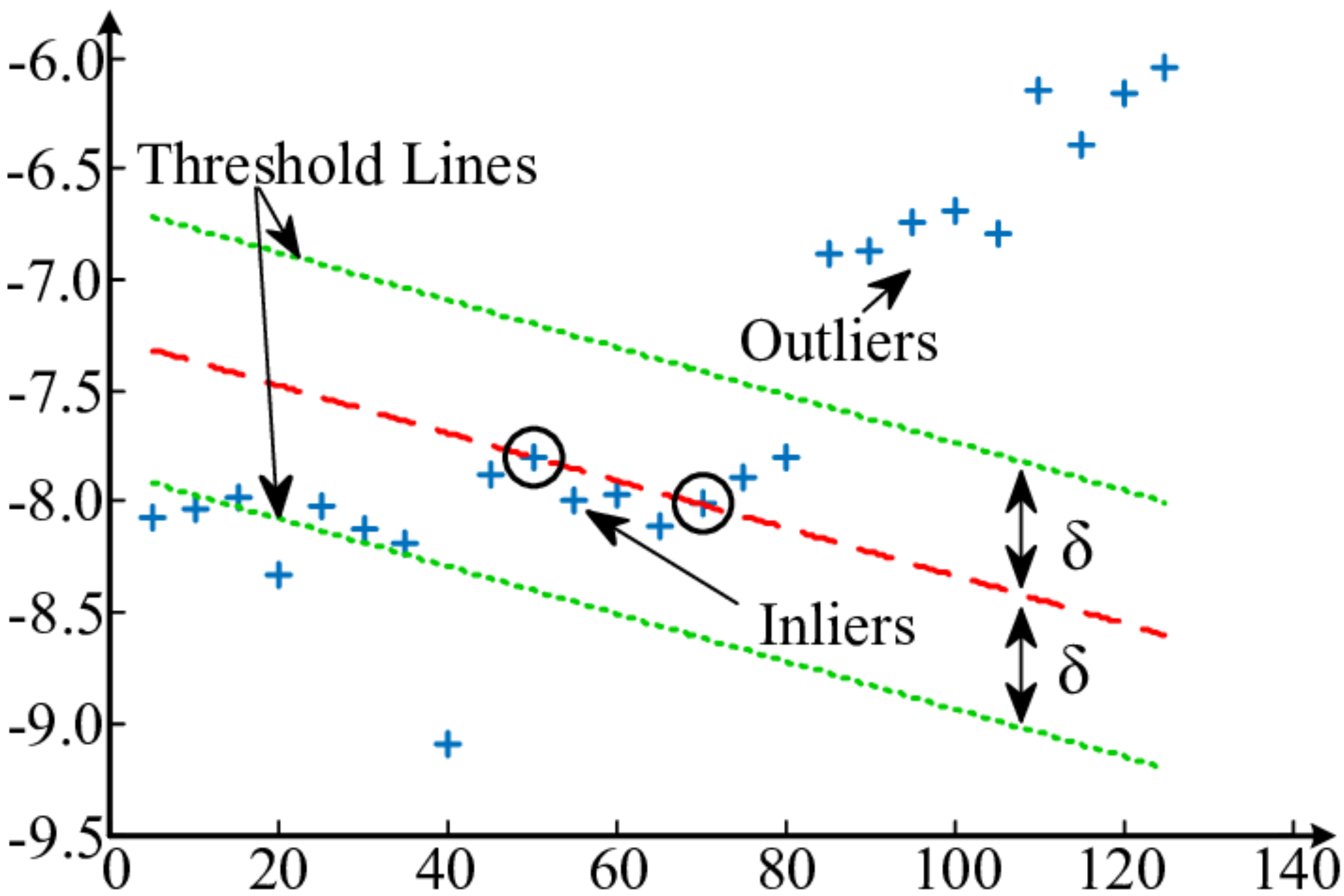
RANSAC

- RANSAC stands for RANdom SAmple Consensus



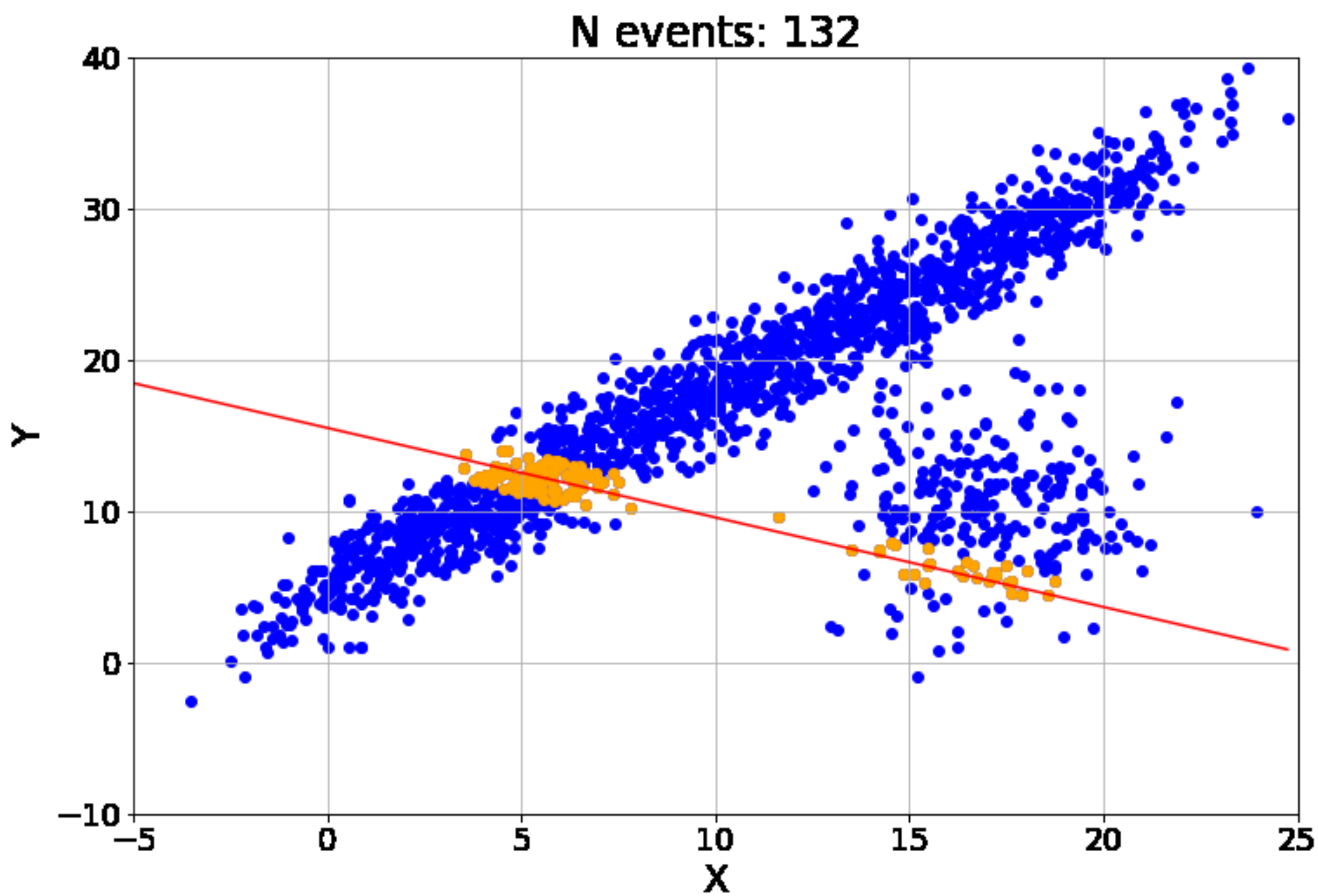
RANSAC

- Outliers vs. inliers



RANSAC

- Random sample and fit lines
- Consensus voting for the model with highest ratio of inliers



RANSAC fitting results