Standard Form

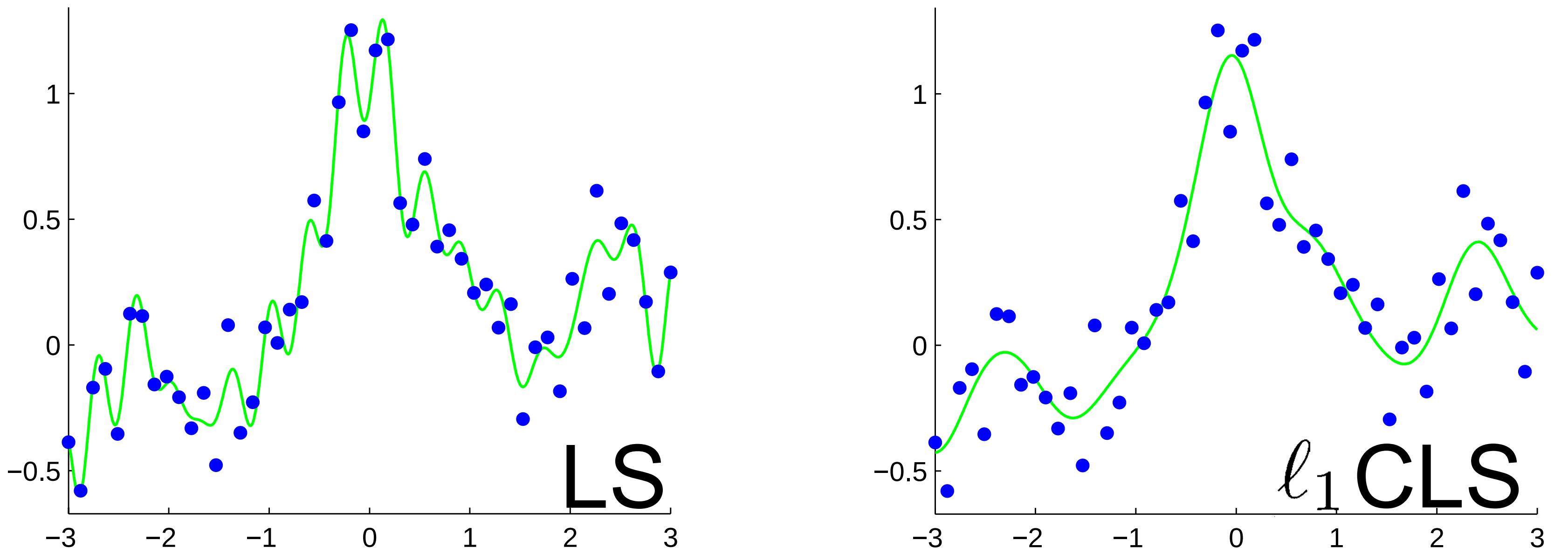
- Sparse learning can be expressed as

$$
\begin{aligned}
\boldsymbol{\beta} &= \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{u} \end{pmatrix} \\
\boldsymbol{Q} &= 2\boldsymbol{\Gamma}_{\boldsymbol{\alpha}}^{\top}\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{\Gamma}_{\boldsymbol{\alpha}} \\
\boldsymbol{q} &= -2\boldsymbol{\Gamma}_{\boldsymbol{\alpha}}^{\top}\boldsymbol{X}^{\top}\boldsymbol{y} + \lambda\boldsymbol{\Gamma}_{\boldsymbol{u}}^{\top}\mathbf{1}_p \\
\boldsymbol{V} &= \begin{pmatrix} -\boldsymbol{\Gamma}_{\boldsymbol{\alpha}} - \boldsymbol{\Gamma}_{\boldsymbol{u}} \\ \boldsymbol{\Gamma}_{\boldsymbol{\alpha}} - \boldsymbol{\Gamma}_{\boldsymbol{u}} \end{pmatrix}, \\
\boldsymbol{v} &= \mathbf{0}_{2p}, \\
\boldsymbol{G} &= \boldsymbol{O}_{2p}, \qquad \boldsymbol{\Gamma}_{\boldsymbol{\alpha}} = (\boldsymbol{I}_p, \boldsymbol{O}_p), \\
\boldsymbol{g} &= \mathbf{0}_{2p}. \qquad \boldsymbol{\Gamma}_{\boldsymbol{u}} = (\boldsymbol{O}_p, \boldsymbol{I}_p).
\end{aligned}
$$

Example of Sparse Learning

- Gaussian kernel model:

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) \qquad K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/2\right)$$



- Over-fit can be avoided by properly choosing the regularization factor.
- 27 out of 50 parameters are exactly zero.

Constrained LSs

	Sparseness	Model parameter	Parameter learning
Subspace LS	Yes	Discrete	Linear
Quadratically constrained LS	No	Continuous	Linear
ℓ_1 constrained LS	Yes	Continuous	Non-linear

Learning Methods

- Outliers
- Robust learning
- Sparse and Robust learning

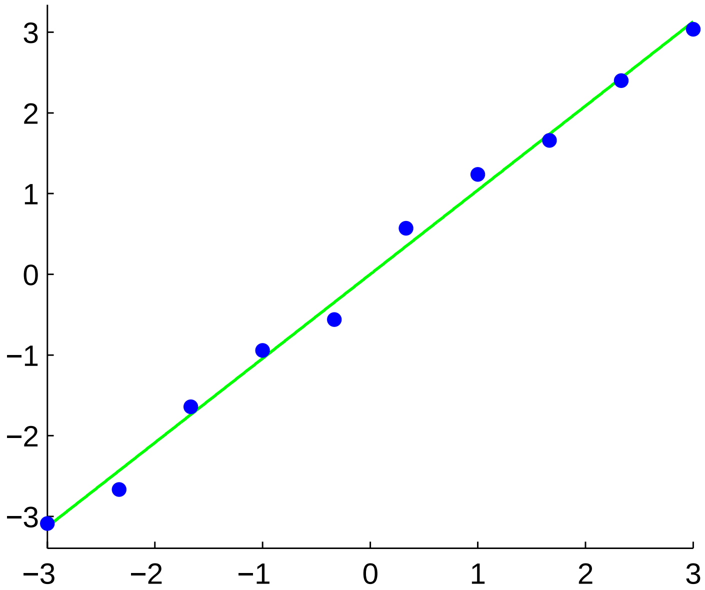
Outliers

- LS estimator is asymptotically efficient when the noise is Gaussian.
- However, in practice, the noise may not be Gaussian.
- If noise distribution has heavy tails, very large noise sometimes appears.
- Samples with such irregular values are called outliers.

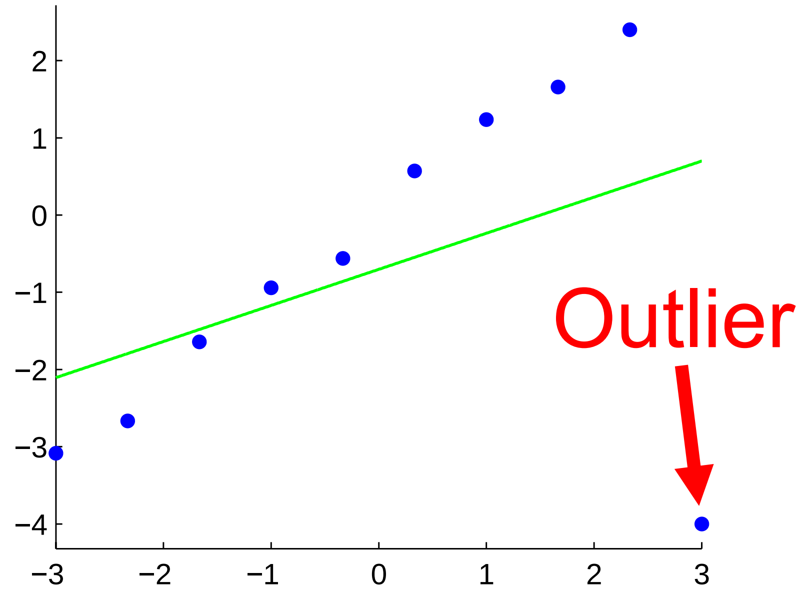
Outliers (cont.)

- LS criterion is sensitive to outliers.

$$\hat{f}(x) = \alpha_1 + \alpha_2 x$$



LS (without outlier)



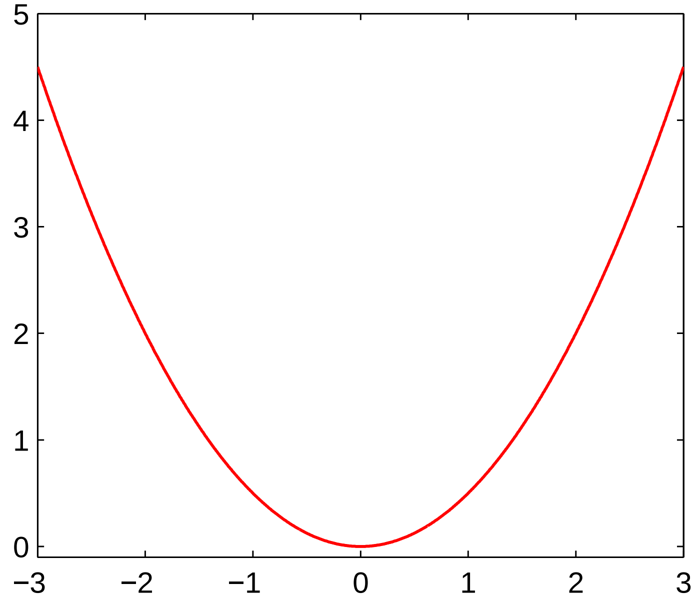
LS (with outlier)

- Single outlier can corrupt the learning result badly.

Quadratic Loss

$$J_{LS}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2$$

- In LS, goodness-of-fit is measured by the squared loss.
- Therefore, even a single outlier has quadratic power to “pull” the learned function
- The solution will be robust if the effect of outliers are deemphasized.



Huber's Loss

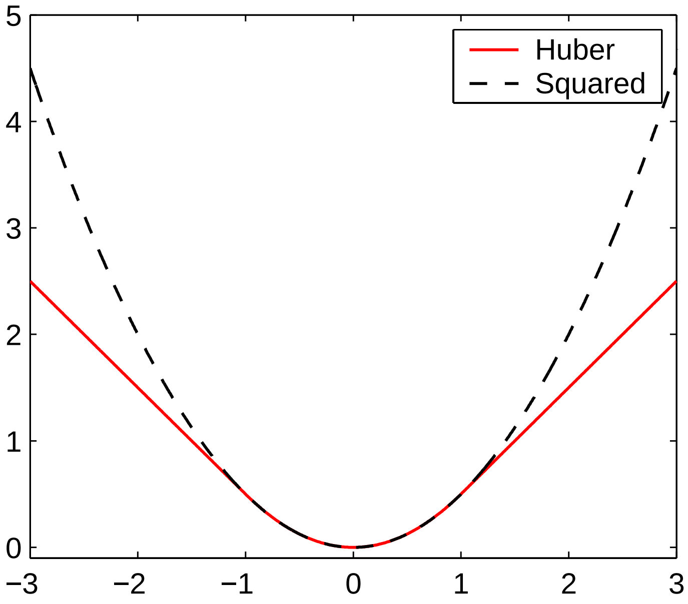
- Huber's method

$$\hat{\boldsymbol{\alpha}}_{Huber} = \underset{\boldsymbol{\alpha}}{\mathrm{argmin}}\, J_{Huber}(\boldsymbol{\alpha})$$

$$J_{Huber}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \rho\left(\hat{f}(\boldsymbol{x}_i) - y_i\right)$$

$$\rho(y) = \begin{cases} \frac{1}{2}y^2 & (|y| \leq t) \\ t|y| - \frac{1}{2}t^2 & (|y| > t) \end{cases}$$

$$t > 0$$



- Squared-loss for non-outliers with small errors.
- Linear penalty for outliers with large errors.

How to Obtain Solutions

- How to deal with Huber's loss?
- Use the following identity:

$$\rho(y) = \min_{b} \left[\frac{1}{2} b^2 + t|y - b| \right]$$

- Then

$$\hat{\boldsymbol{\alpha}}_{Huber} = \operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^p, \boldsymbol{b} \in \mathbb{R}^n} \left[\frac{1}{2} \|\boldsymbol{b}\|^2 + t \|\boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{y} - \boldsymbol{b}\|_1 \right]$$

$$\boldsymbol{X}_{i,j} = \varphi_j(\boldsymbol{x}_i)$$

How to Obtain Solutions (cont.)

- Trick to avoid absolute value:

$$\|\boldsymbol{\alpha}\|_1 = \min_{\boldsymbol{u} \in \mathbb{R}^p} \sum_{i=1}^{p} u_i$$

$$\text{subject to } -\boldsymbol{u} \leq \boldsymbol{\alpha} \leq \boldsymbol{u},$$

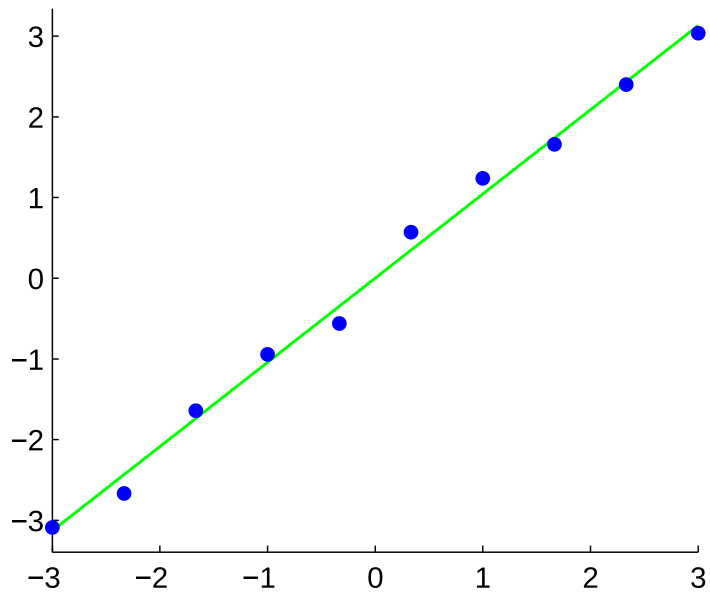
- Solution is given by

$$\hat{\boldsymbol{\alpha}}_{Huber} = \mathop{\mathrm{argmin}}_{\boldsymbol{\alpha} \in \mathbb{R}^p, \boldsymbol{b}, \boldsymbol{h} \in \mathbb{R}^n} \left[\frac{1}{2} \|\boldsymbol{b}\|^2 + t \sum_{i=1}^{n} h_i \right]$$

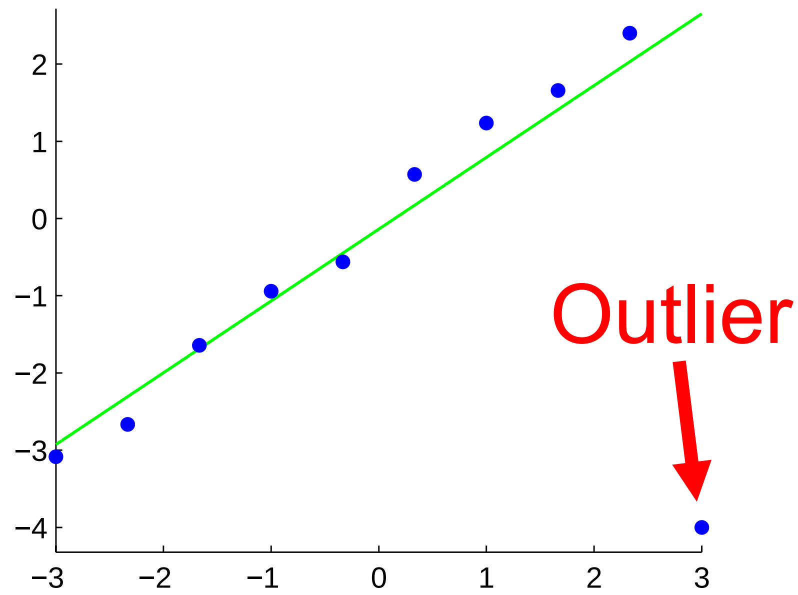
$$\text{subject to } -\boldsymbol{h} \leq \boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{y} - \boldsymbol{b} \leq \boldsymbol{h}$$

Example of Huber's Method

- Huber method can successfully suppress the influence of outliers.



Huber (without outlier)



Huber (with outlier)

Intermediate Report

- Read one/both of the following articles and write your opinions.
  - B. Schölkopf,

    Statistical learning and kernel methods.

    ftp://ftp.research.microsoft.com/pub/tr/tr-2000-23.pdf
  - D.J.C. MacKay,

    Introduction to Gaussian processes.

    ftp://wol.ra.phy.cam.ac.uk/pub/mackay/gpB.ps.gz
- Write your opinions on this course.
- Deadline: Jan. 7, 2005 (First class next year)