Robust estimation

* Naive approach:
  - Fit model to all points
  - compute distance of each point from model
  - Discard points with largest distance
  - Fit model to remaining points

* Initial model is inaccurate and so we drift in the wrong direction.

* Approaches:
  - M-estimators
  - RANSAC

M-estimators

- Mean square Error (MSE) fitting:

$$E(\theta) = \sum d^2 (x_i ; \theta)$$

e.g. $d^2 = (l^T x_i)^2$
for line fitting

- Robust estimation:

$$E(\theta) = \sum \rho_\sigma \left(d(x_i ; \theta) \right)$$

MSE is a special case
where $\rho_\sigma(x) = x^2$

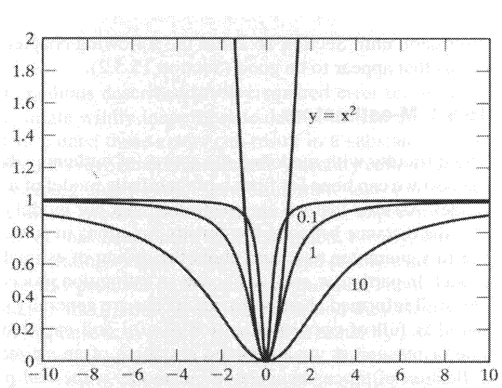
Geman-McClure estimator

$$\rho_\sigma(x) = \frac{x^2}{x^2 + \sigma^2}$$

$x >> \sigma \Rightarrow \rho_\sigma(x) = 1$

$x << \sigma \Rightarrow \rho_\sigma(x) = \frac{x^2}{\sigma^2}$

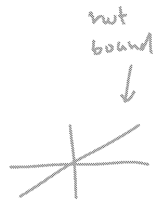
larger σ
$\Downarrow$
wider valley



Geman-McClure estimator

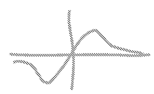
$$E(\theta) = \sum \rho_\sigma \left(d(x_i ; \theta) \right)$$

$$\nabla E(\theta) = \sum \frac{\partial}{\partial d} \rho_\sigma (d) \frac{\partial}{\partial \theta} d(\theta)$$



For $\rho_\sigma(d) = d^2 \Rightarrow \frac{\partial \rho}{\partial d} = 2d$

For $\rho_\sigma = \frac{d^2}{d+\sigma^2} \Rightarrow \frac{\partial \rho}{\partial d} = \frac{2 d \sigma^2}{(d^2+\sigma^2)^2}$



Selecting bandwidth parameter

- large $\sigma \Rightarrow$ include more points
- small $\sigma \Rightarrow$ include fewer points
- variable estimation:
  start with large σ and decrease
  as converging

$$\sigma^{(n)} = 1.5 \text{ median} \{ d(x_i ; \theta^{(n-1)}) \}$$

estimate at step n — parameters at step $n-1$

M-estimator summary

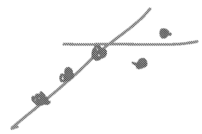
1) Draw a <u>large</u> set of points uniformly at random
2) Select initial value of σ
3) Fit model $\rightarrow \theta^{(i)}$
4) Compute $\sigma^{(i)}$ using median distance of points
5) continue while objective is decreasing

* To overcome incorrect initial guess of σ
repeat several times and select best solution
(smaller objective)

RANSAC

Random Sample Consensus (RANSAC):

- perform multiple experiments
- choose best results
- use small sets in hope that at least one set will not have outliers



Parameters:

n = # points drawn at each evaluation

d = min # points needed to estimate model

K = # trials

t = distance threshold to identify inliers

RANSAC Algorithm

* Repeat K times:
  - Draw n points uniformly at random (with replacement)
  - Fit a model to points
  - Find inliers in entire set (distance $< t$)
  - Recompute model (if at least d inliers)
  - update parameters (K, t)
* choose best solution:
  - Largest consensus set
  (- or smallest error)

Estimating RANSAC parameters

* To estimate t use median distance from model.
* To estimate K use:

  P: with probability of p at least one experiment does not have outliers (e.g. $P = 0.99$) user selected

  w: probability that a point is an inlier (initially $w = 0.5$) estimated

Estimating RANSAC parameters

probability that all k experiments failed:

$$(1-p) = (1 - w^n)^k$$

$$\log(1-p) = k \log(1-w^n)$$

$$k = \frac{\log(1-p)}{\log(1-w^n)}$$

large p → large k

small w → large k

$$w \leftarrow \frac{\#\text{ inliers}}{\#\text{ points}}$$

update w, k every iteration but set upper bound for k