Introduction to Machine Learning

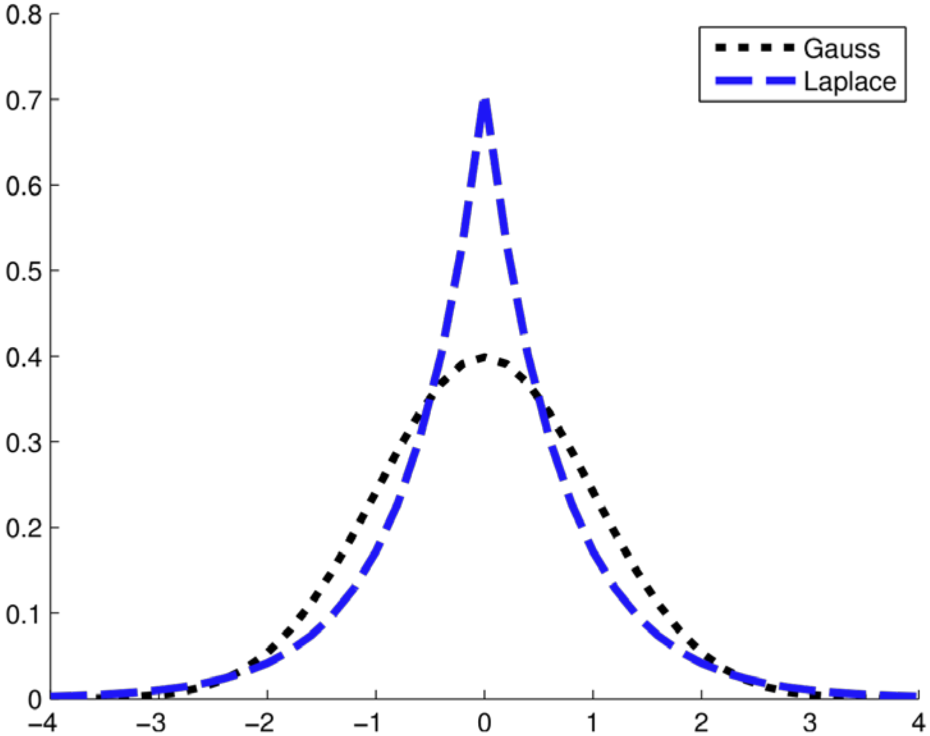
Brown University CSCI 1950-F, Spring 2011
Prof. Erik Sudderth

Lecture 13: Robust Regression, Feature Selection & Search, L_1 Regularization

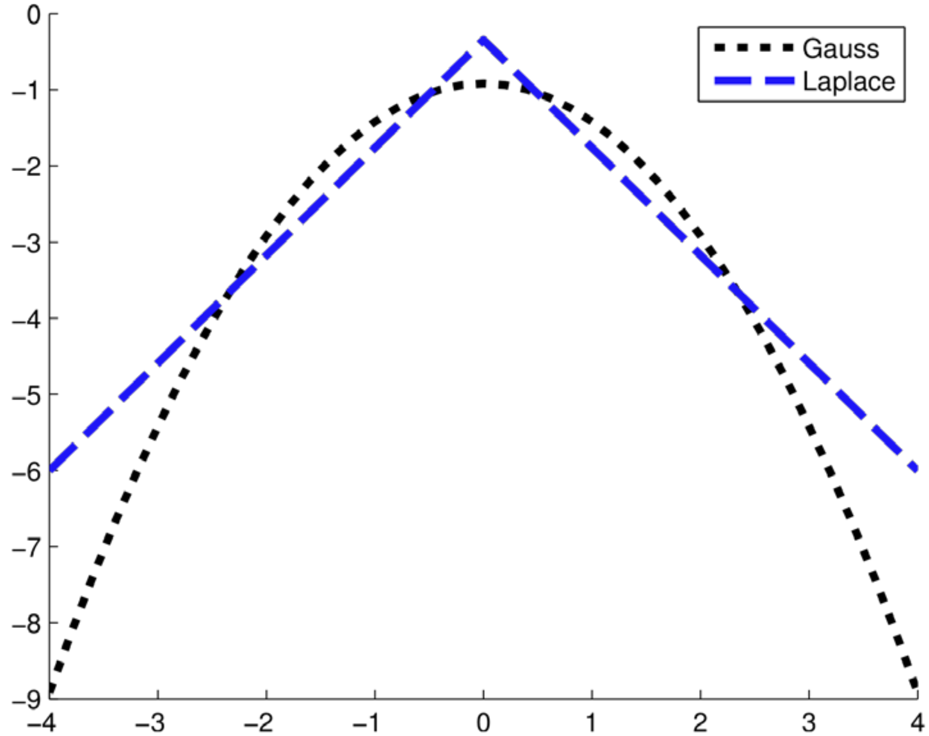
Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

Laplace Distribution

Probability Densities



Log Probability Densities

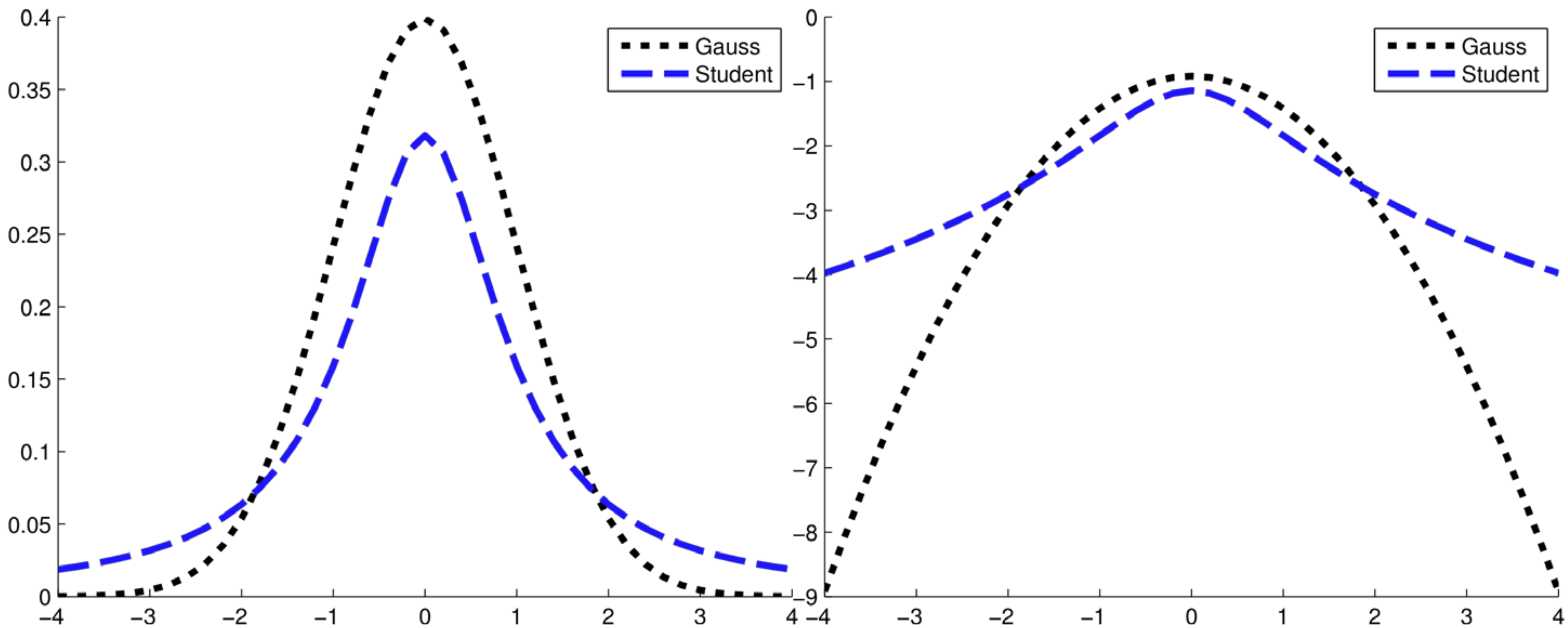


Relative to Gaussian distributions with equal variance:

- Many samples are near zero
- Occasional large-magnitude samples are far more likely
- Negative log probability density is *convex but not smooth*

Student T Distribution

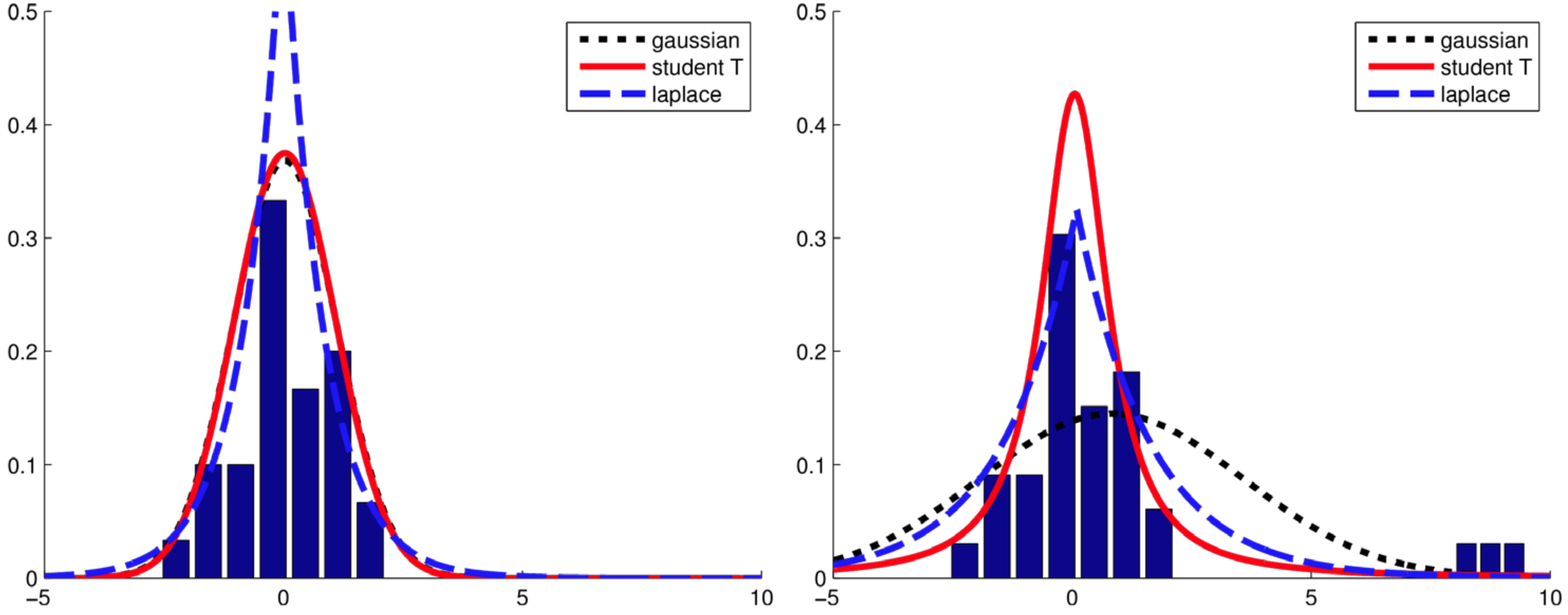
Probability Densities | Log Probability Densities



Relative to Gaussian distributions with equal variance:

- Approaches Gaussian as DOF parameter approaches infinity
- For small DOF, large-magnitude samples are far more likely
- Negative log probability density is *smooth but not convex*

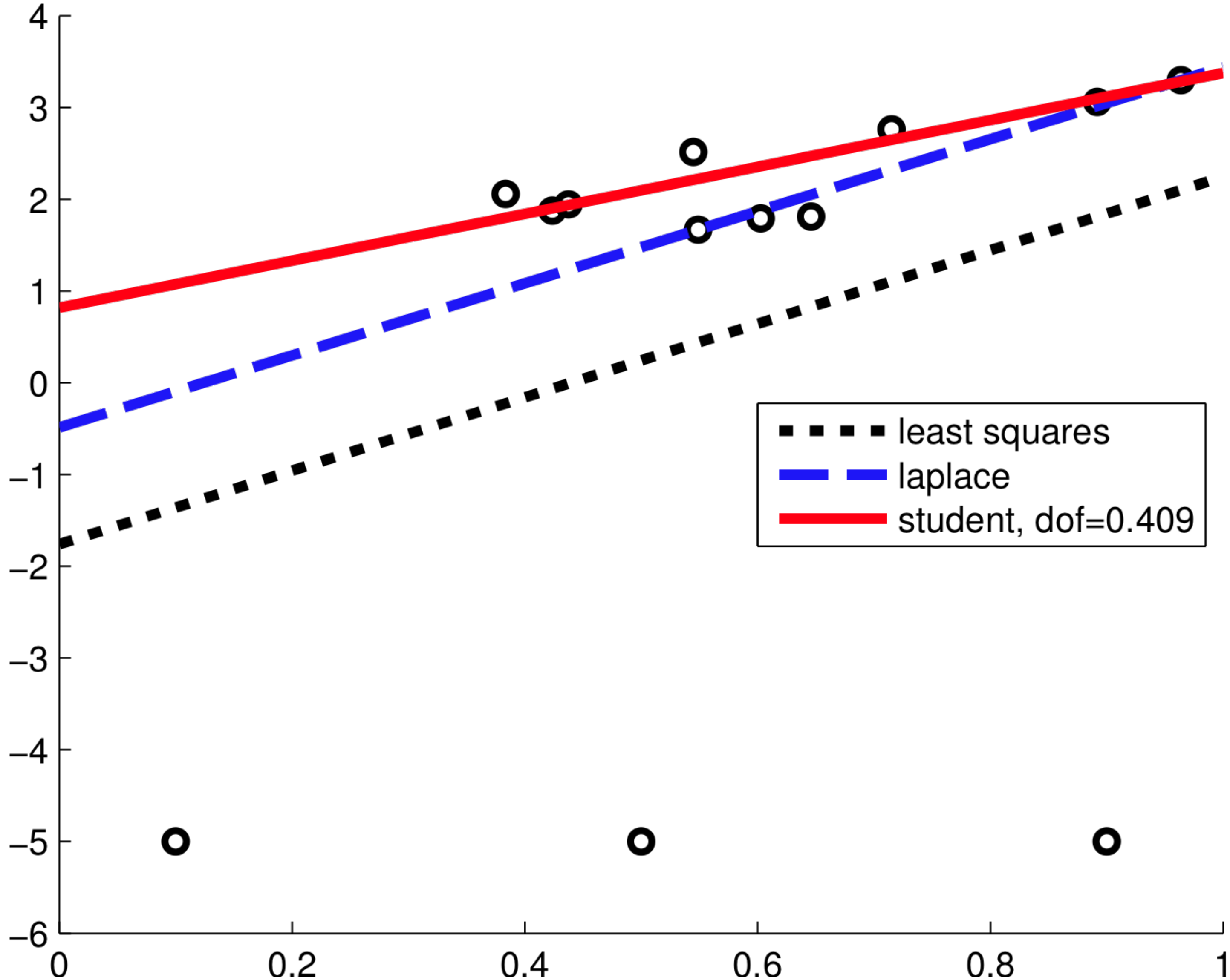
Outliers & ML Estimation



Maximum likelihood estimates of mean parameters:

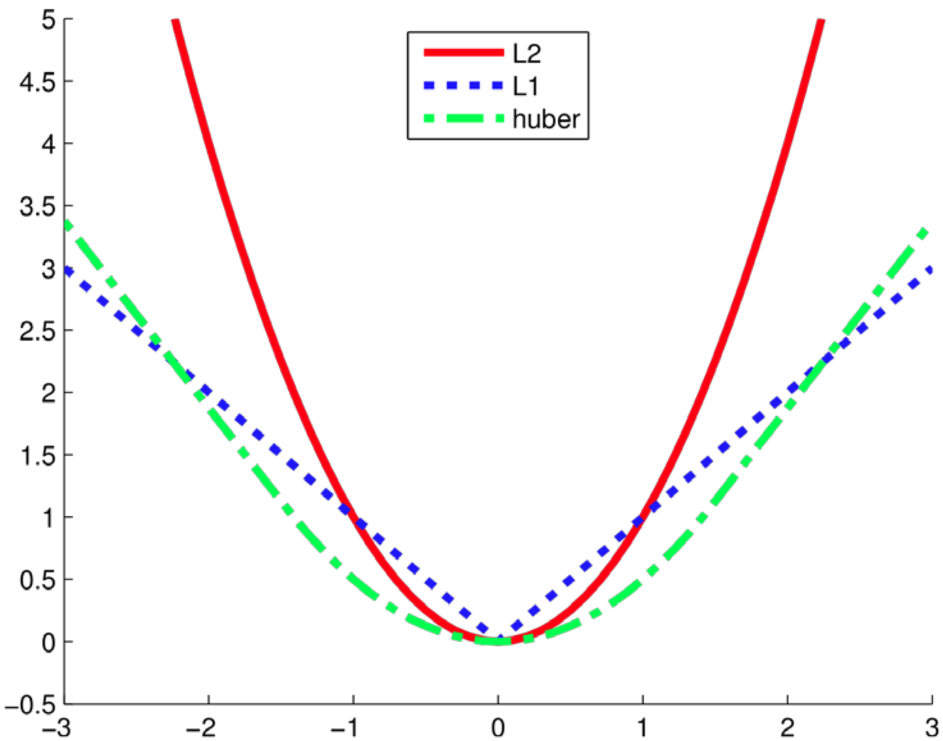
- Gaussian: Sample mean of data
- Laplacian: Sample median of data
- Student T: No closed form, optimize via gradient methods

Outliers & Linear Regression

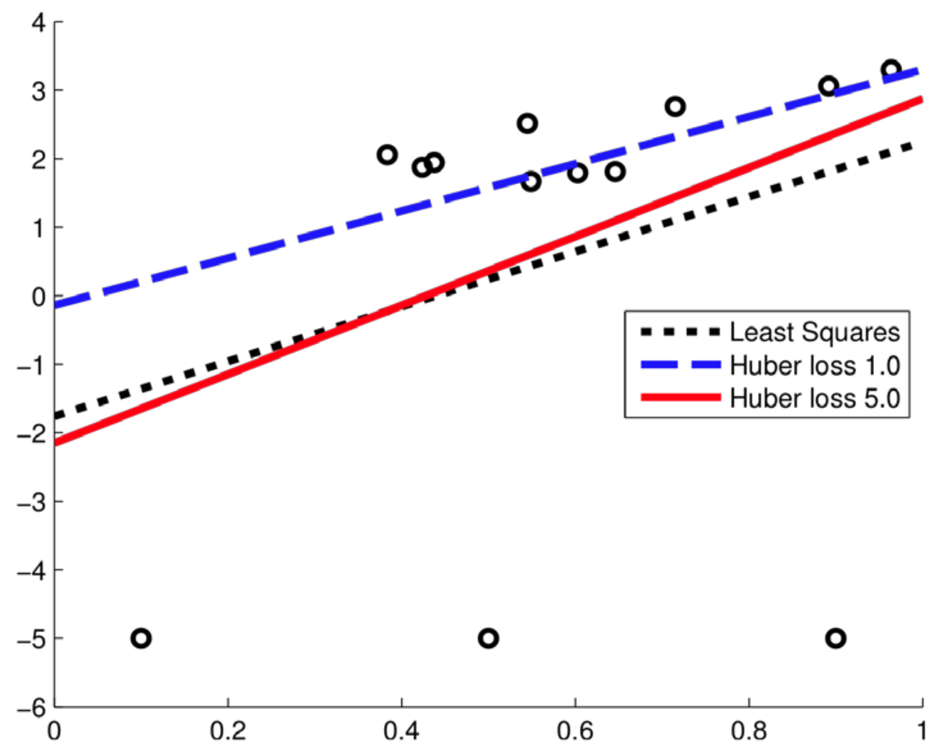


Huber Loss Function

Negative Log Probabilities



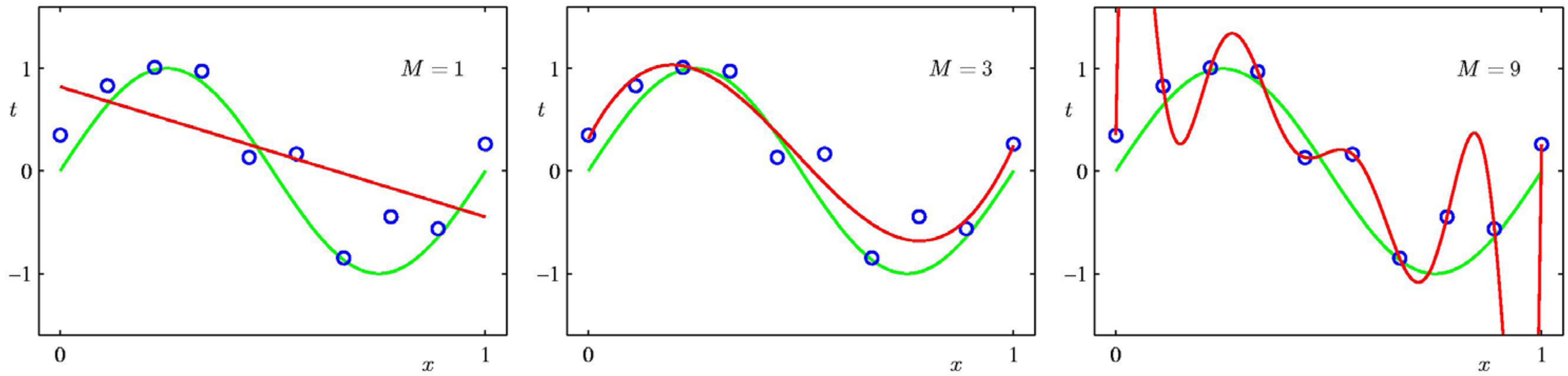
Robust Linear Regression



Relative to Gaussian distributions with equal variance:

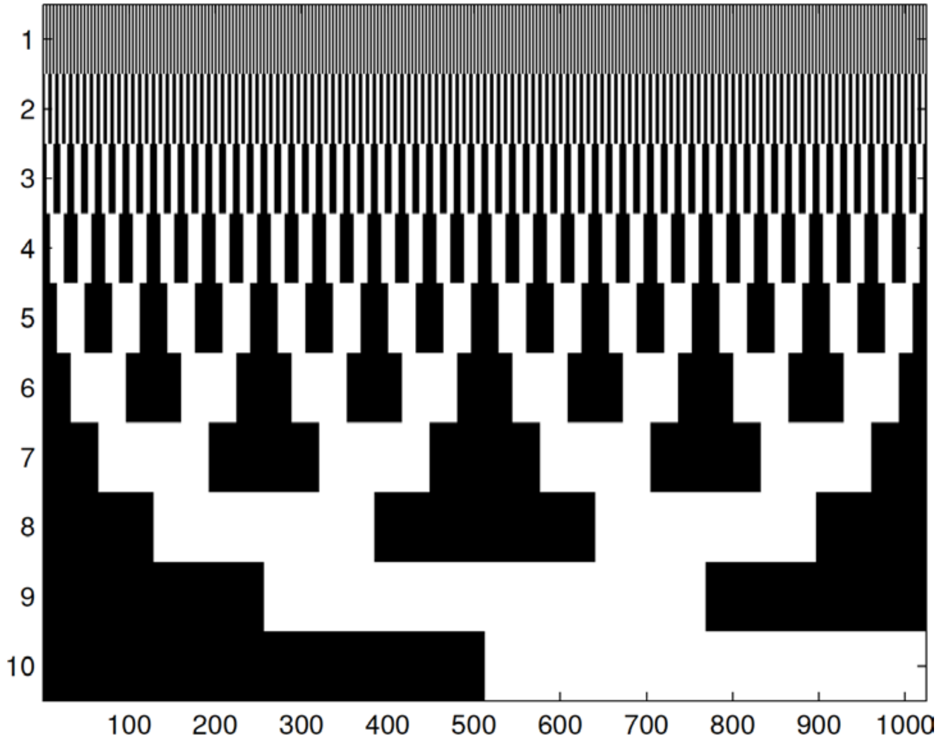
- Behaves like Gaussian near origin ("non-outliers")
- Behaves like Laplacian far from origin (robustness)
- Negative log probability density is *smooth and convex*

Regularization in Regression

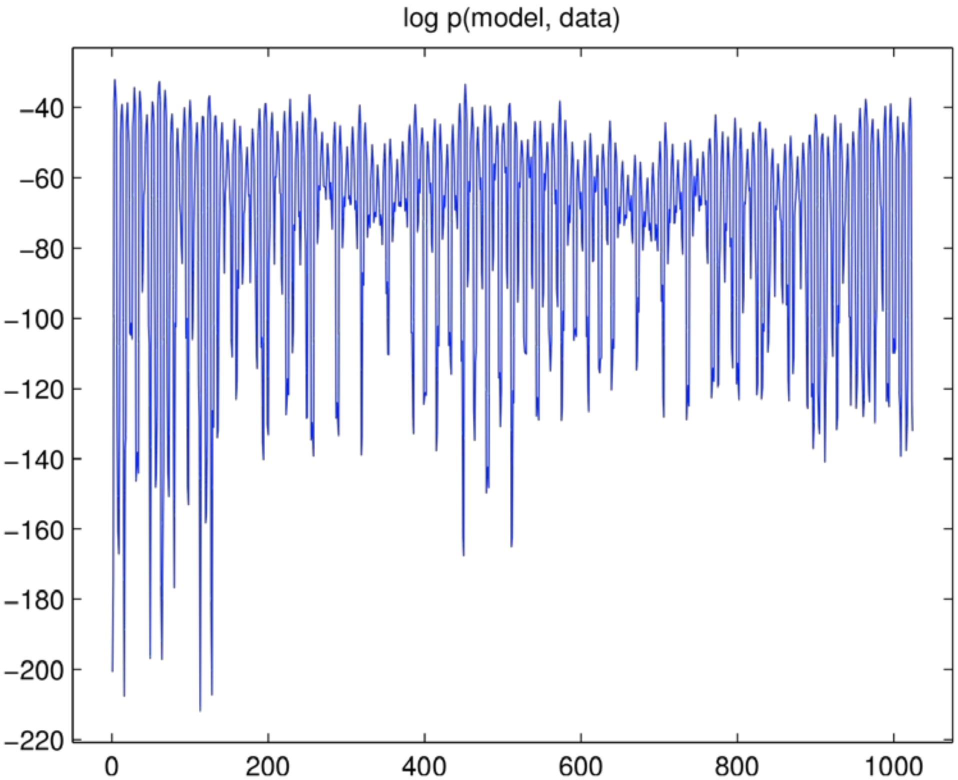


- Basic model selection: Coefficients are ordered, and only the first M are non-zero
  - Classical example: polynomial regression
  - What if my features aren't easy to interpret?
- Gaussian prior (L_2 regularization): Coefficients are small
  - Computation & storage: Expensive for many features
  - Interpretability: Doesn't identify important features
- Many applications: Only some of my features are relevant, but I don't know how many or which ones

Feature Selection: Regression



All combinations of 10 features
(2^{10} in Gray code order)



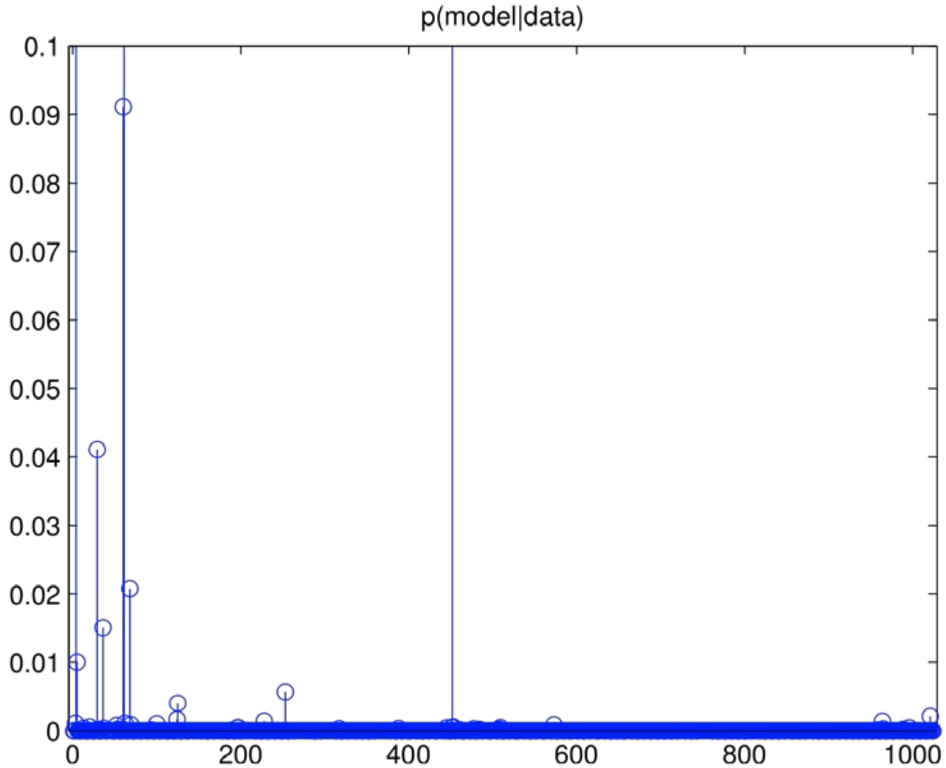
Log posterior probabilities

$$p(\boldsymbol{\gamma}|\omega) = \omega^{||\boldsymbol{\gamma}||_0}(1-\omega)^{D-||\boldsymbol{\gamma}||_0}$$

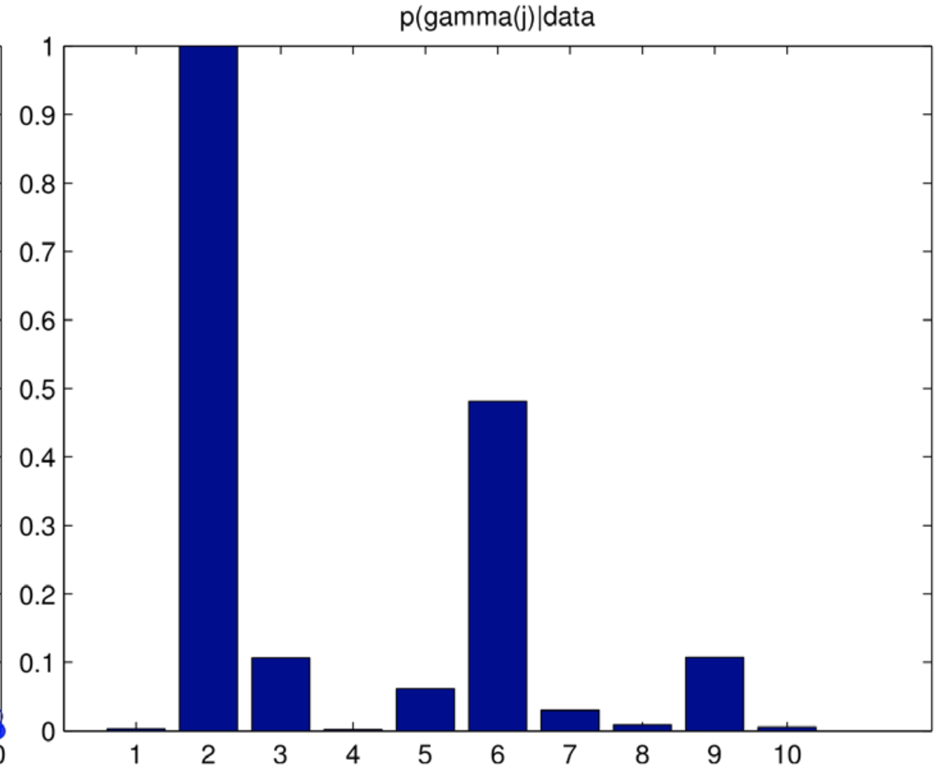
Dataset: N=10 samples based on linear regression weights

$$\mathbf{w} = (0.00, -1.67, 0.13, 0.00, 0.00, 1.19, 0.00, -0.04, 0.33, 0.00)$$

Feature Selection: Regression



Posterior Probabilities



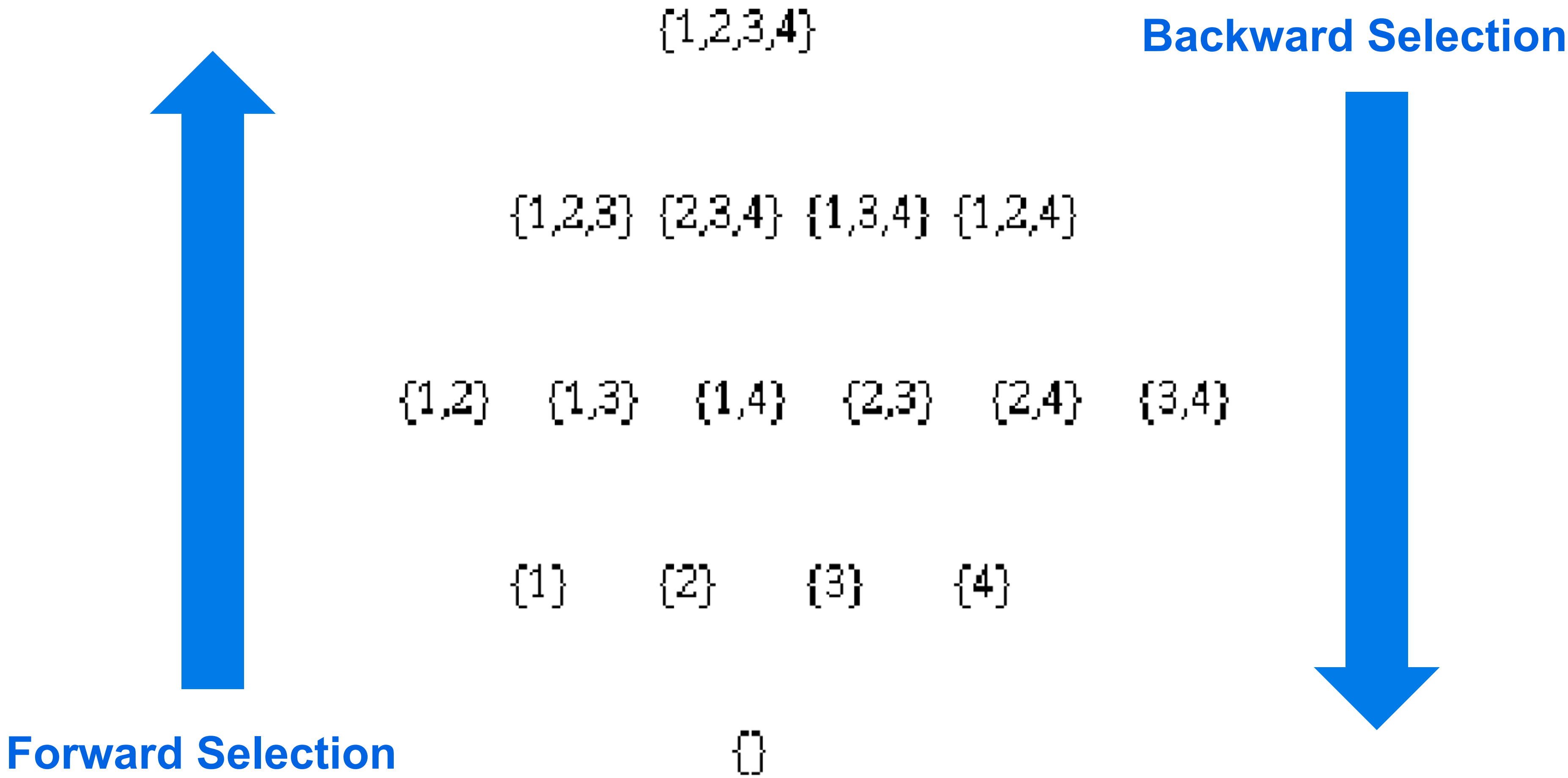
Marginal Inclusion Probabilities

Most likely models: *{2}, {2,6}, {2,6,9}, …*

Dataset: N=10 samples based on linear regression weights

$$\mathbf{w} = (0.00, -1.67, 0.13, 0.00, 0.00, 1.19, 0.00, -0.04, 0.33, 0.00)$$

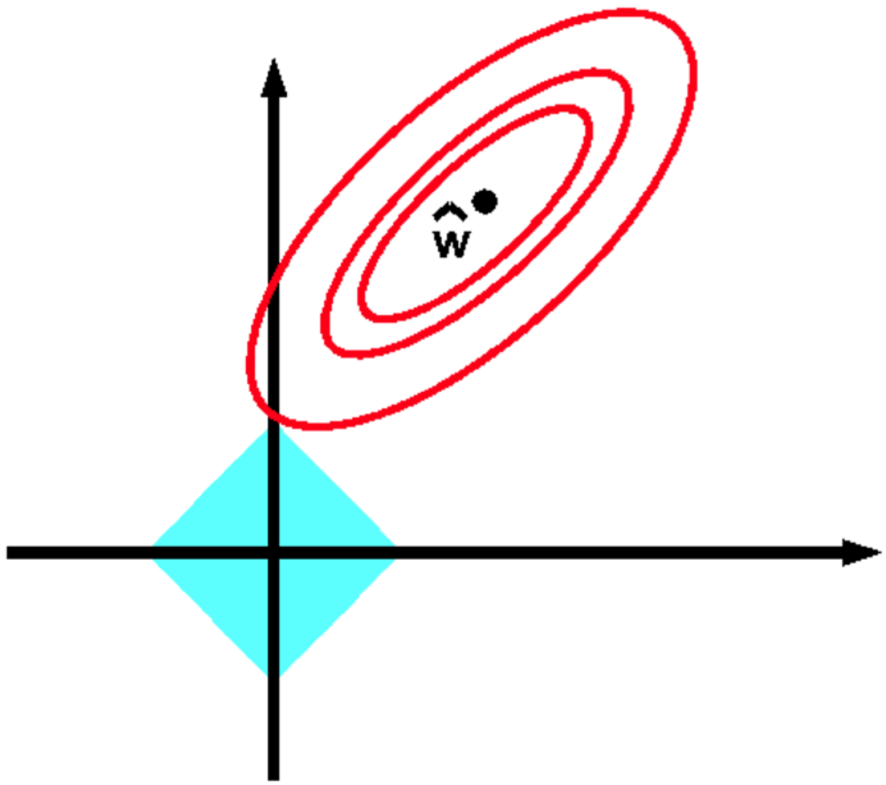
Greedy Deterministic Search



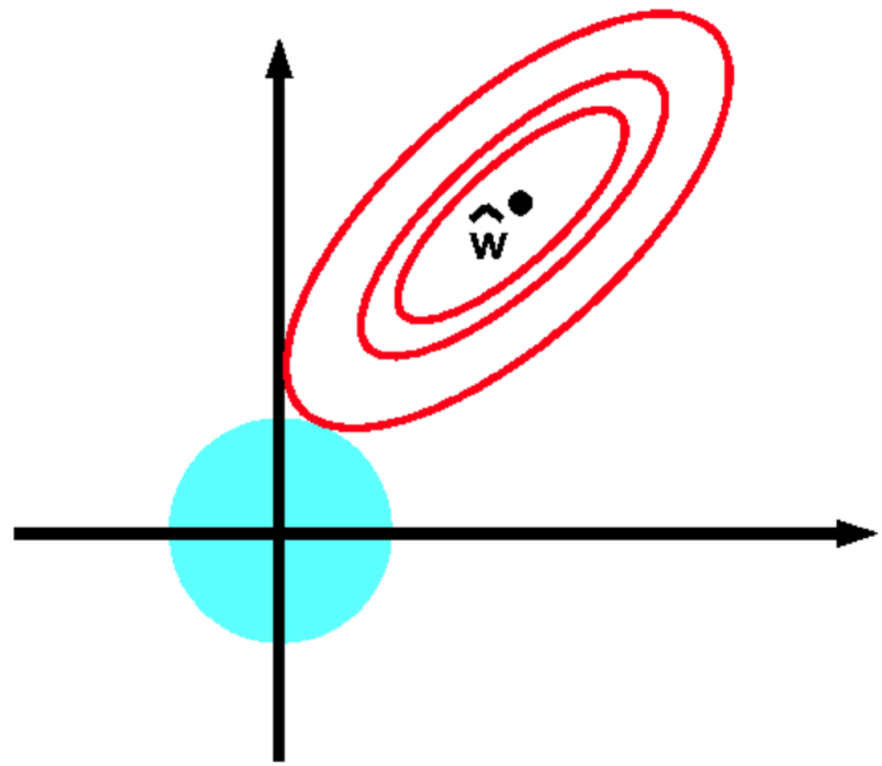
- Consider all possible ways of adding *(forward selection)* or removing *(backward selection)* one feature
- Add or remove the best feature, or stop if the current model is best
- Wrapper method: Can be applied to any objective. *Guarantees???*

Constrained Optimization

Laplacian prior
L_1 regularization
Lasso regression



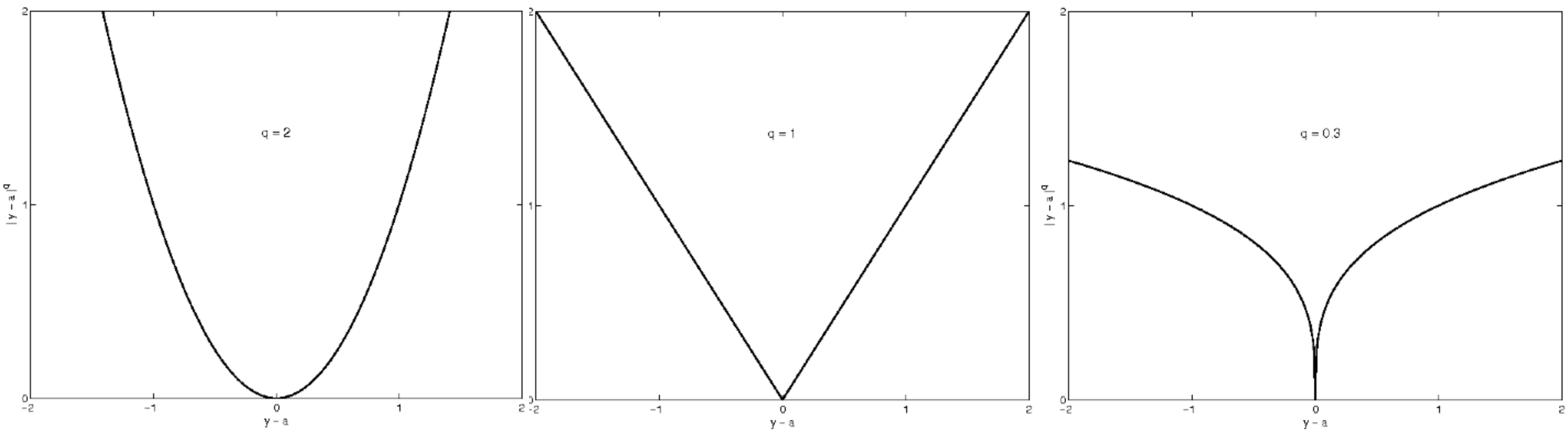
Gaussian prior
L_2 regularization
Ridge regression



Where do level sets of the quadratic regression cost function first intersect the constraint set?

Generalized Norms: Bridge Regression

$$\mathrm{NLL}(\mathbf{w}) + \lambda \sum_j |w_j|^b \qquad \mathrm{ExpPower}(w|\mu, a, b) := \frac{b}{2a\Gamma(1/b)} \exp(-\frac{|x-\mu|}{a})^b$$

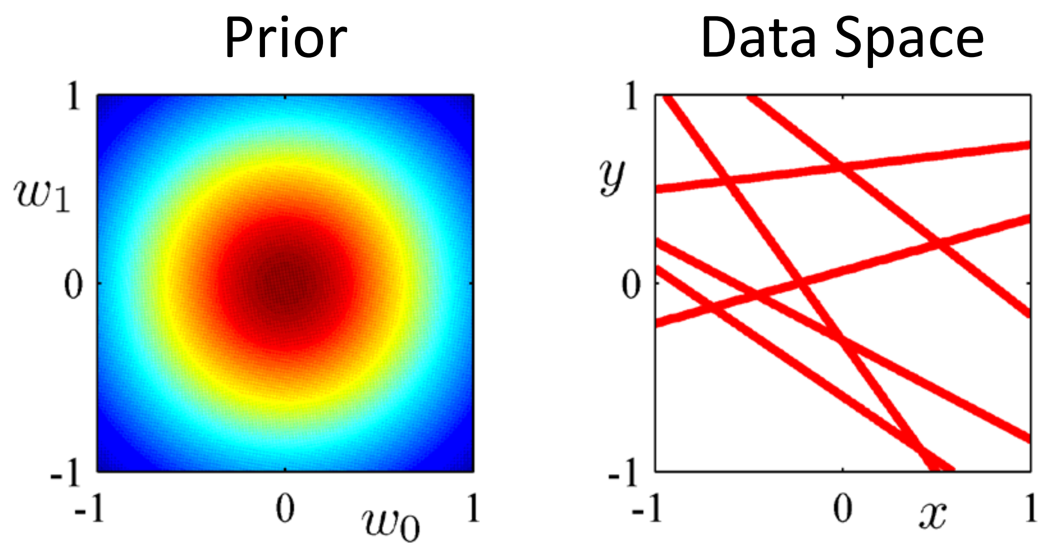


b=2 b=1 b=0.3

- Convex objective function (true norm): b ≥ 1
- Encourages sparse solutions (cusp at zero): b ≤ 1
- Lasso/Laplacian (convex & sparsifying): b = 1
- Ridge/Gaussian (classical, closed form solutions): b = 2
- Sparsity via discrete counts (greedy search): b ➔ 0

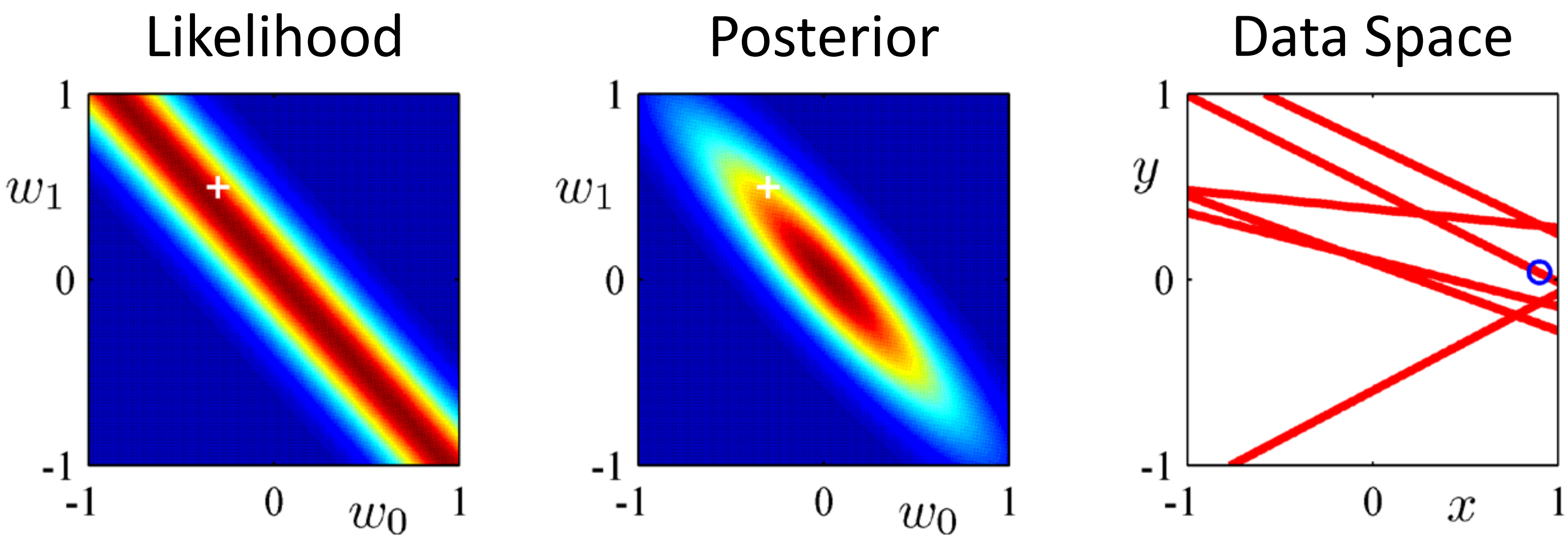
Bayesian Linear Regression

0 data points observed



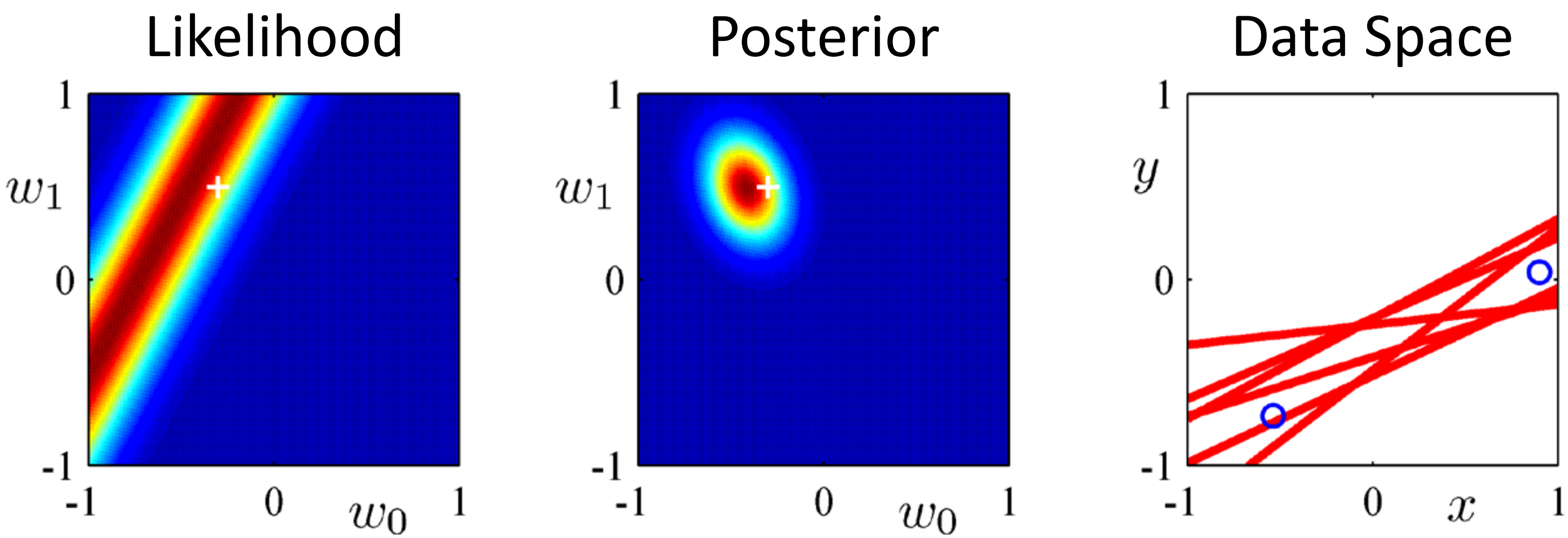
Bayesian Linear Regression

1 data point observed



Bayesian Linear Regression

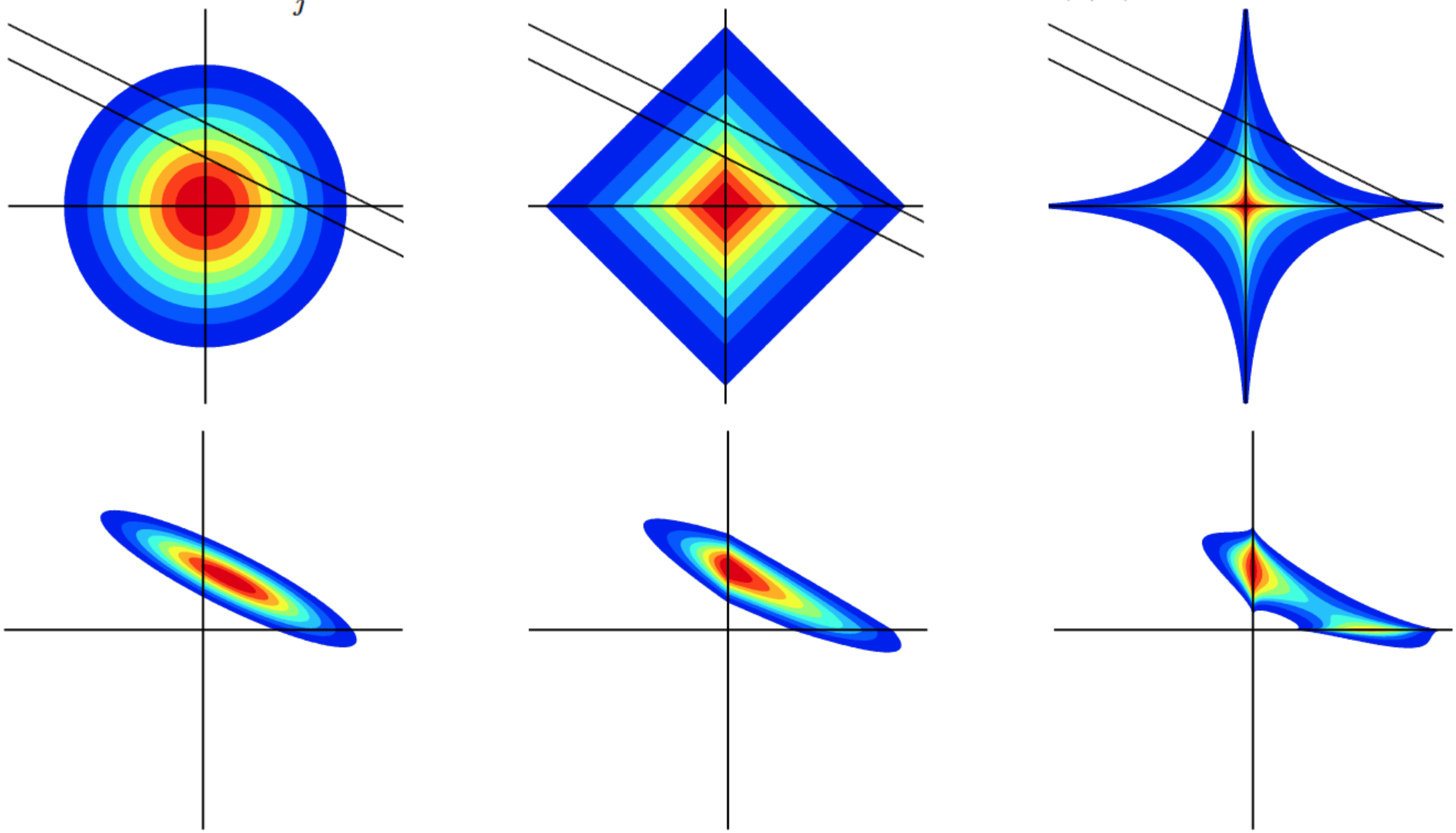
2 data points observed



Comparing Regression Posteriors

$$\mathrm{NLL}(\mathbf{w}) + \lambda \sum_j |w_j|^b$$

$$\mathrm{ExpPower}(w|\mu, a, b) := \frac{b}{2a\Gamma(1/b)} \exp(-\frac{|x-\mu|}{a})^b$$



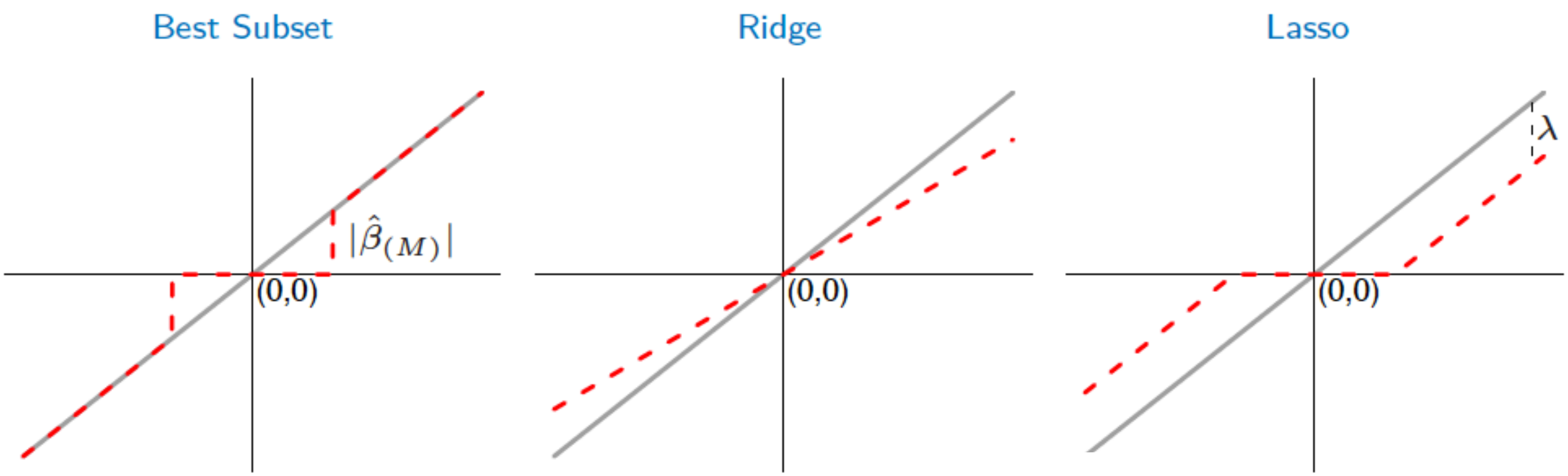
b=2 b=1 b=0.4

Shrinkage for Orthonormal Features

$$\begin{aligned} RSS(\mathbf{w}) &= ||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 = \mathbf{y}^T\mathbf{y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{x} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{y} \\ &= \text{const} + \sum_k w_k^2 - 2\sum_k\sum_i w_k x_{ik} y_i \end{aligned}$$

$$\mathbf{X}^T\mathbf{X} = \mathbf{I}$$

$$\hat{w}_k^{OLS} = \mathbf{x}_{:k}^T\mathbf{y}$$



$$\hat{w}_k^{SS} = \begin{cases} \hat{w}_k^{OLS} & \text{if rank}(|w_k|) \leq K \\ 0 & \text{otherwise} \end{cases}$$

Hard thresholding:
Goal of discrete feature selection

$$\hat{w}_k^{ridge} = \frac{\hat{w}_k^{OLS}}{1+\lambda}$$

Linear Shrinkage:
All coefficients remain non-zero

$$\hat{w}_k^{lasso} = \text{sign}(\hat{w}_k^{OLS})\left(|\hat{w}_k^{OLS}| - \frac{\lambda}{2}\right)_+$$

Soft thresholding:
“Least Absolute Selection & Shrinkage Operator”