Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2012
Prof. Erik Sudderth

Lecture 14:
Sparsity & L_1 Regularization: The Lasso

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

Feature Selection for Regression

$$f(w) = ||y - \Phi w||_2^2 + \lambda ||w||_0$$

$$||w||_p = \left(\sum_{j=1}^{D} |w_j|^p \right)^{1/p} \qquad ||w||_0 = \textit{number of nonzero entries}$$

- L_0 penalty comes from Bernoulli prior on feature usage indicators, and large variance prior on non-zero weights:

$$p(\gamma) = \prod_{j=1}^{D} \mathrm{Ber}(\gamma_j | \pi_0) = \pi_0^{||\gamma||_0} (1 - \pi_0)^{D - ||\gamma||_0} \qquad ||\gamma||_0 = \textstyle\sum_{j=1}^{D} \gamma_j$$

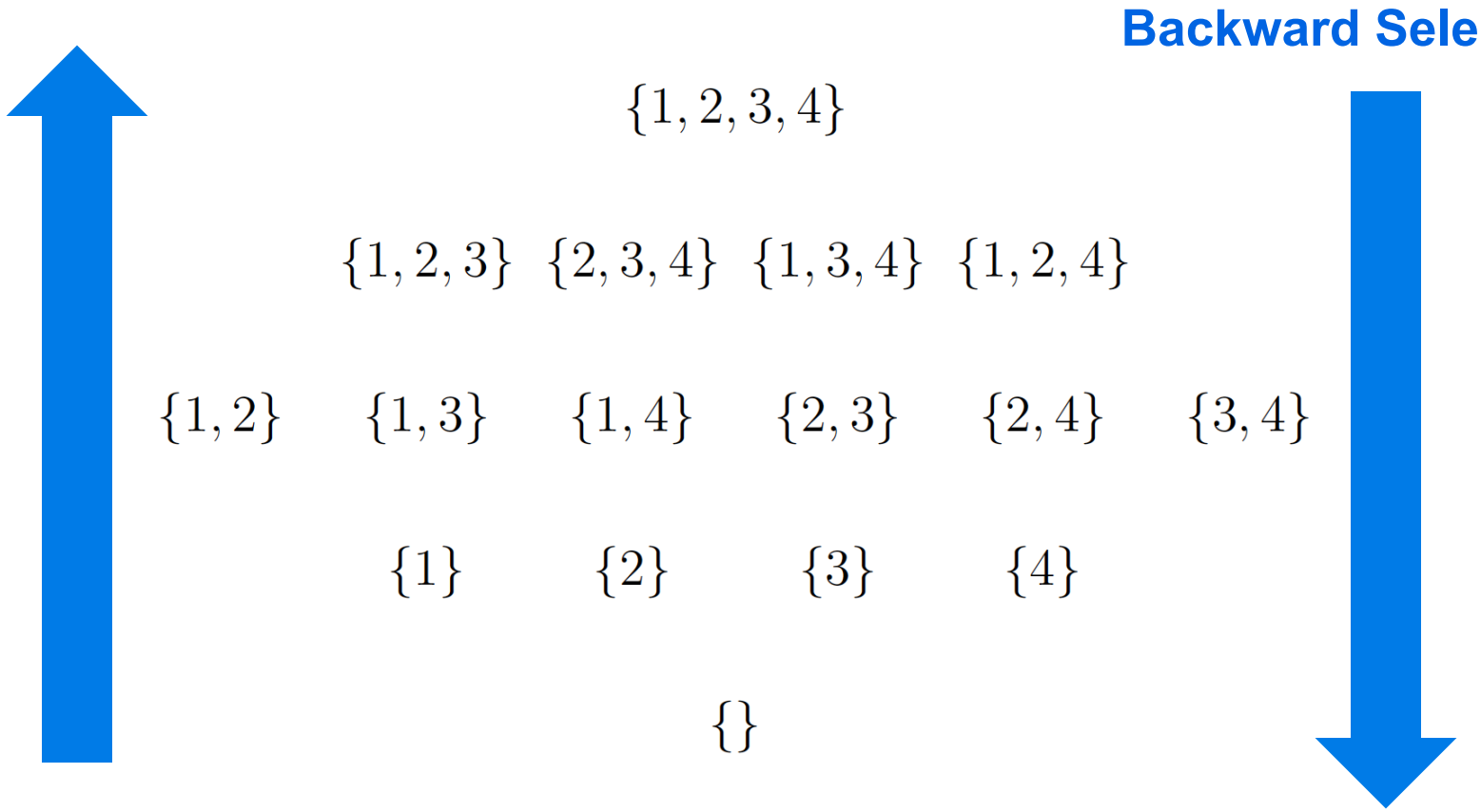
$$w_j \sim \mathcal{N}(0, \sigma_w^2) \qquad \sigma_w^2 \to \infty$$

- L_2 penalty comes from Gaussian regression likelihood:

$$p(y_i \mid x_i, w, \gamma, \sigma^2) = \mathcal{N}(y_i \mid \sum_{j=1}^{D} \gamma_j w_j \phi_j(x_i), \sigma^2)$$

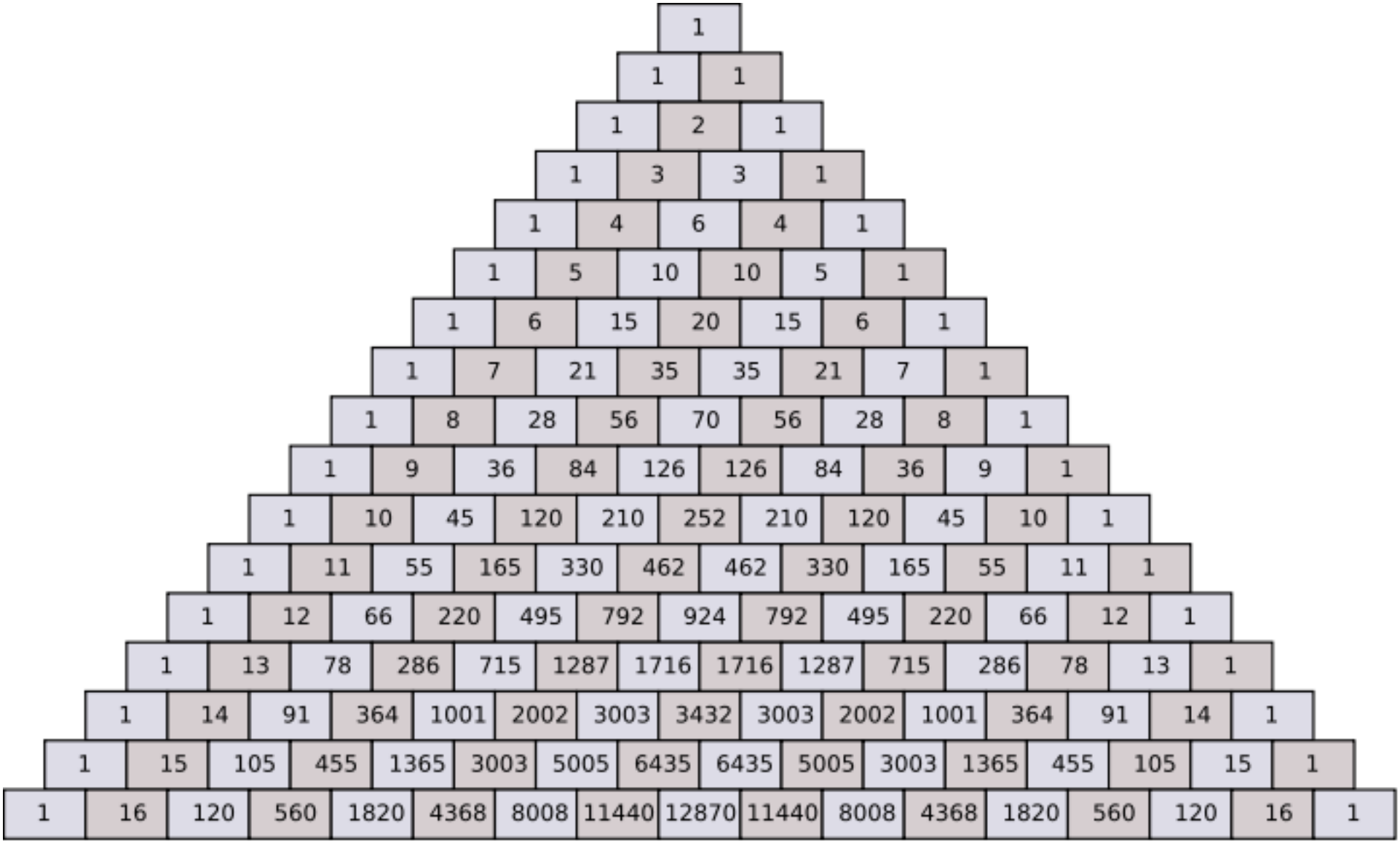
- ***Problem:*** Optimization is hard combinatorial problem!

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???*

Too Many Models

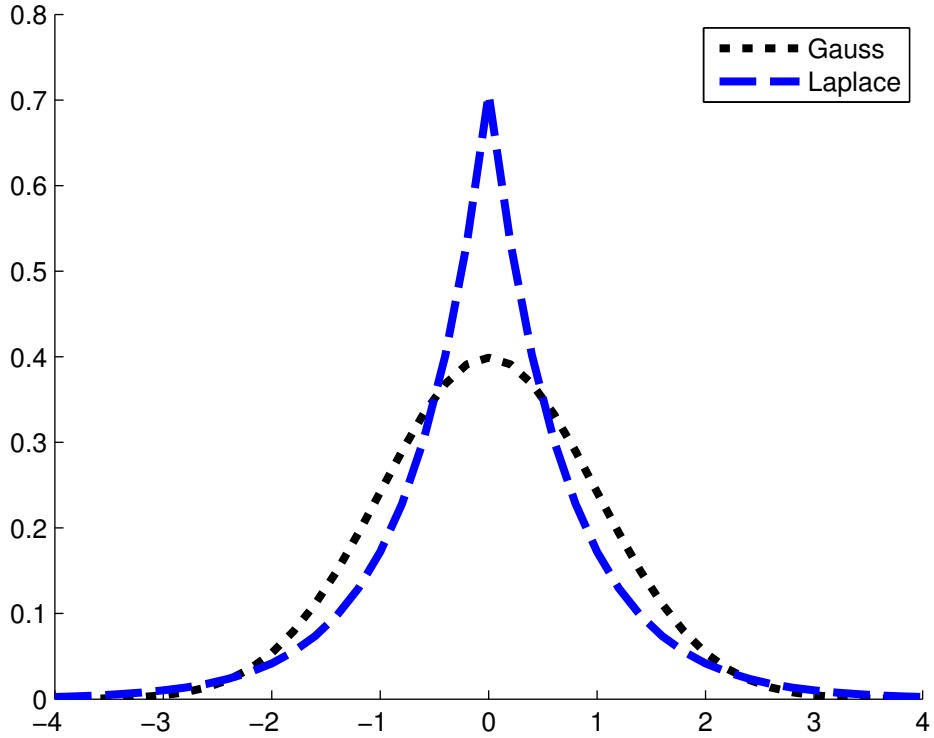


Pascal's Triangle *(http://www.mathwarehouse.com/)*

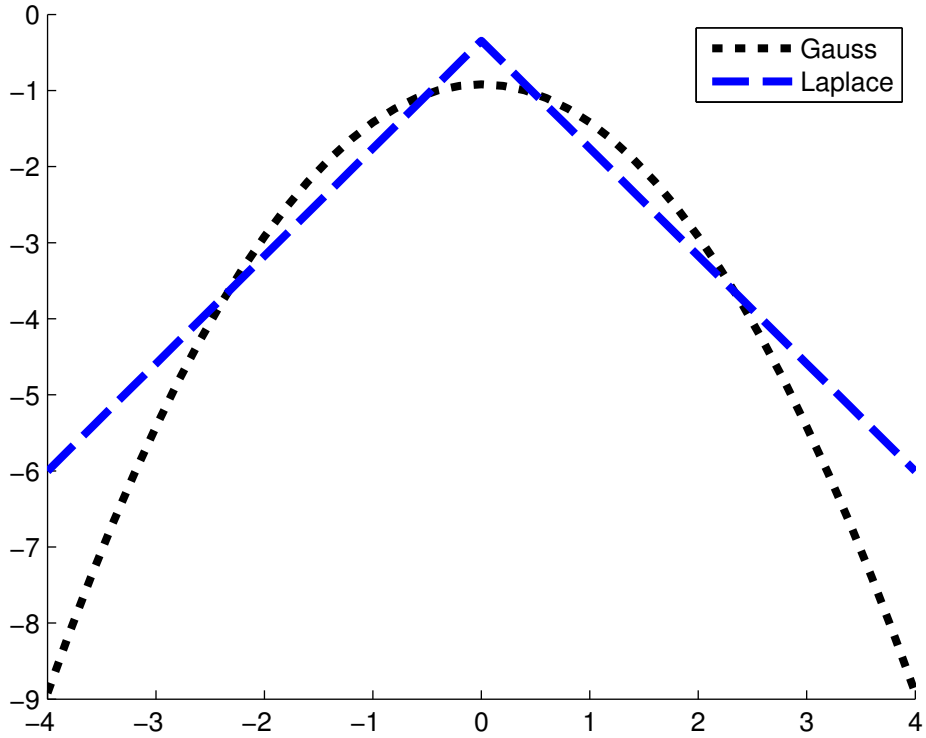
Laplace Distribution

$$\mathrm{Lap}(x \mid \mu, \lambda) = \frac{\lambda}{2} \exp(-\lambda |x - \mu|)$$

Probability Densities



Log Probability Densities

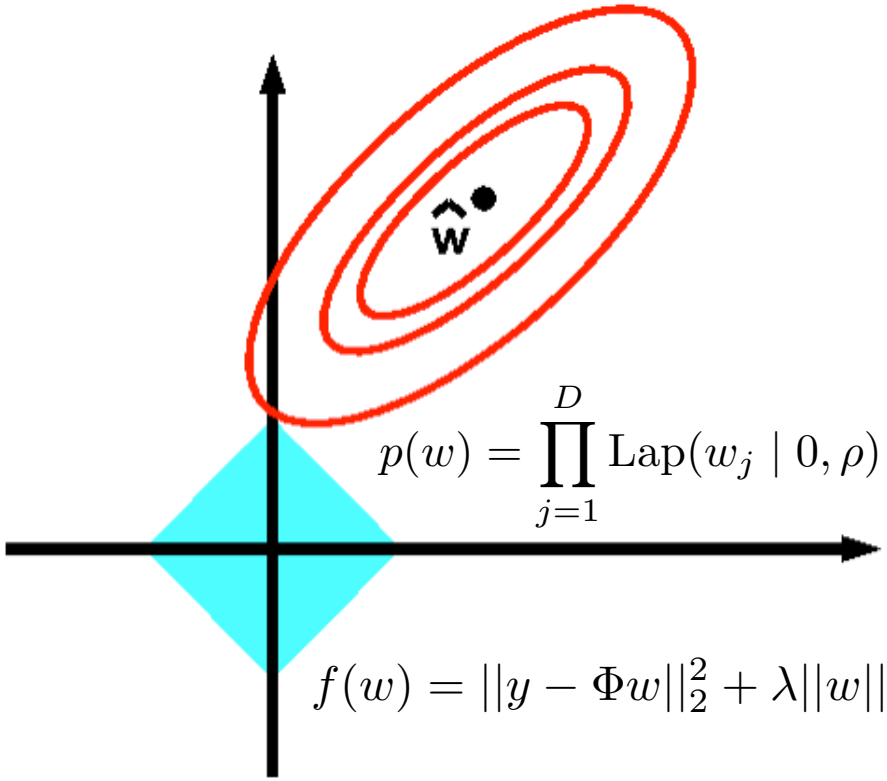


When used as a prior on vectors of model parameters:

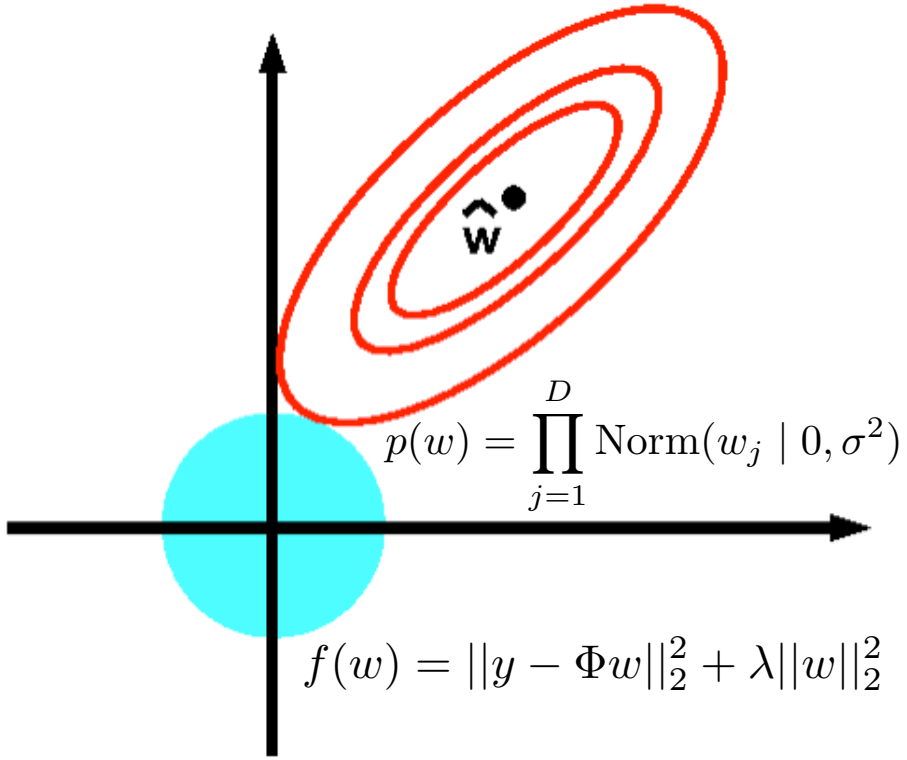
- Compared to Gaussian, stronger bias that many near zero
- When find MAP estimate, some weights are *exactly* zero
- Learning harder than for Gaussian, but still *convex*

Constrained Optimization

Laplacian prior
L_1 regularization
Lasso regression



Gaussian prior
L_2 regularization
Ridge regression

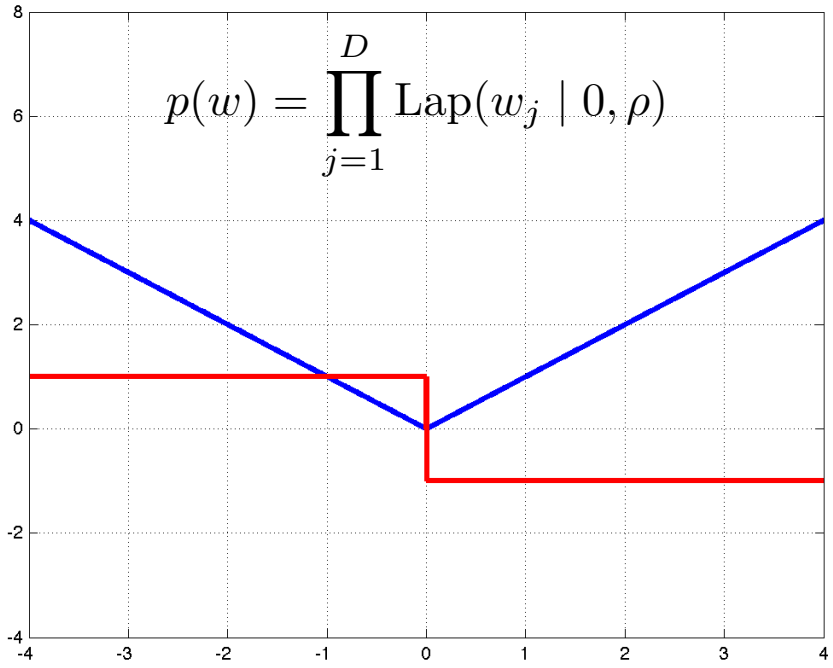


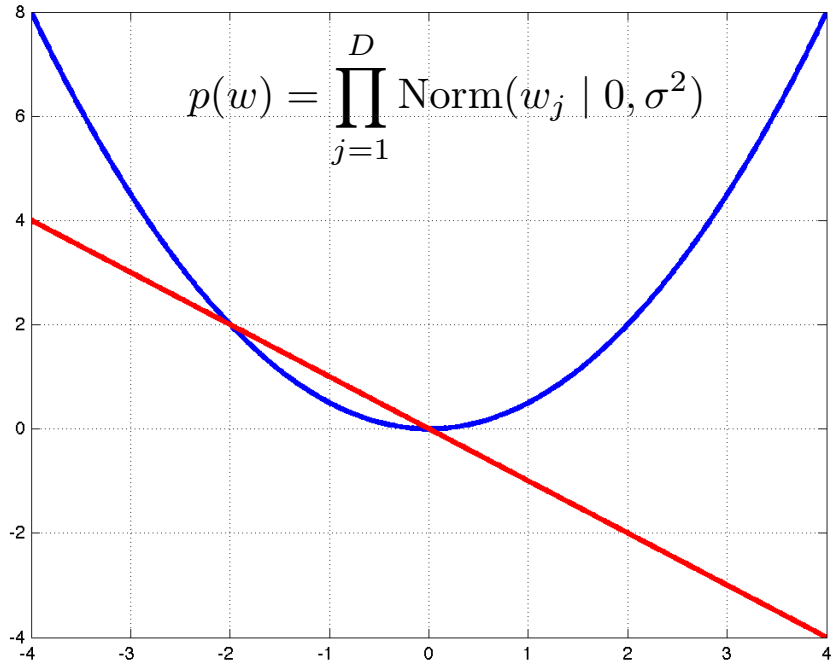
Where do level sets of the quadratic regression cost function first intersect the constraint set?

Gradient-Based Optimization

Laplacian prior
L_1 regularization
Lasso regression

Gaussian prior
L_2 regularization
Ridge regression





Objective Function: $f(w) = -\log p(w)$

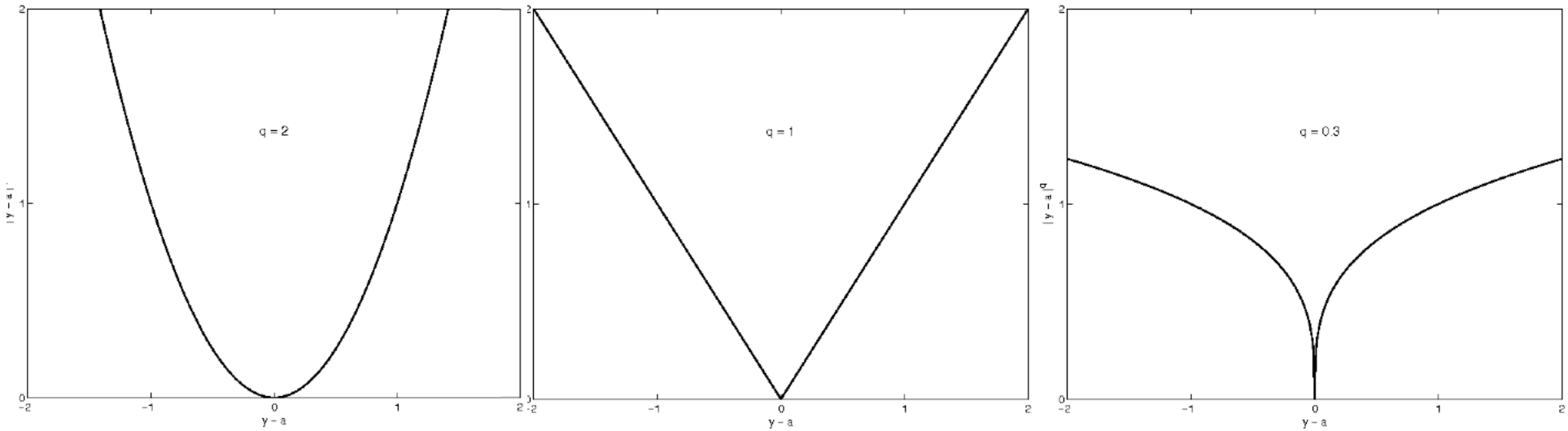
Negative Gradient: $-f'(w)$

(Informal intuition: Gradient of L_1 objective not defined at zero)

Generalized Norms: Bridge Regression

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

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

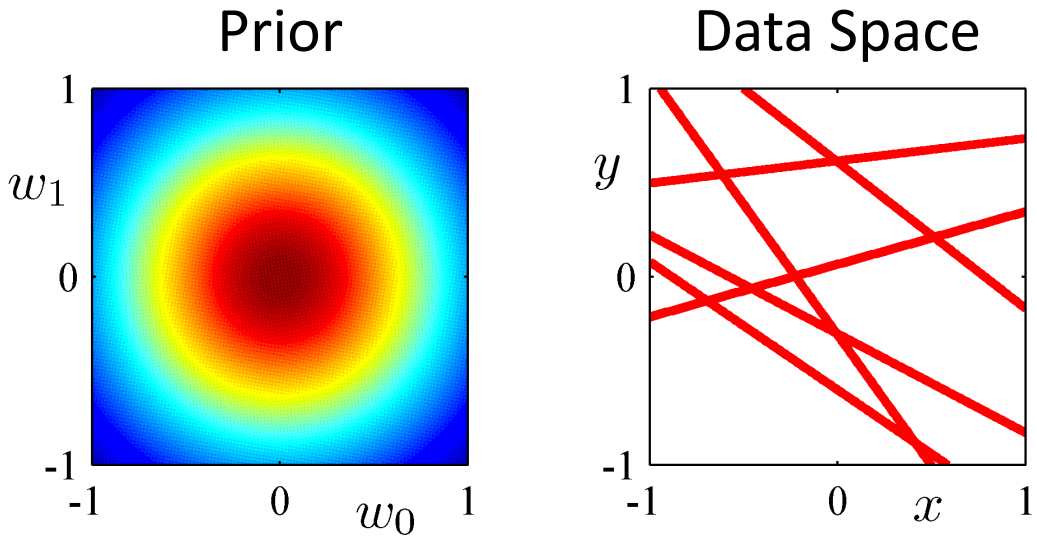


b=2 b=1 b=0.3

- 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

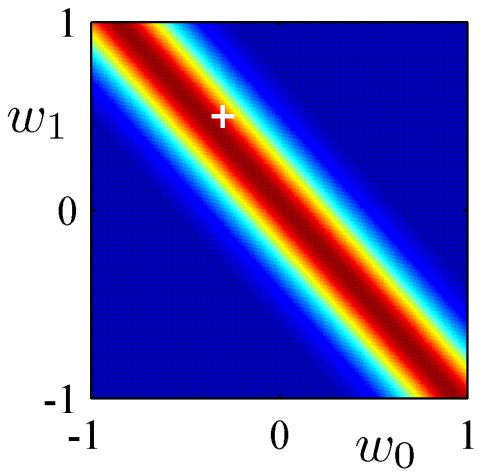
0 data points observed



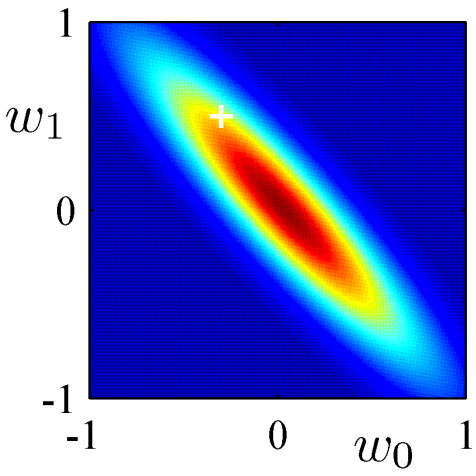
Bayesian Linear Regression

1 data point observed

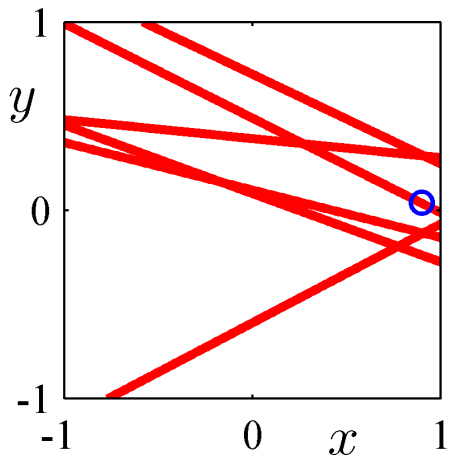
Likelihood



Posterior



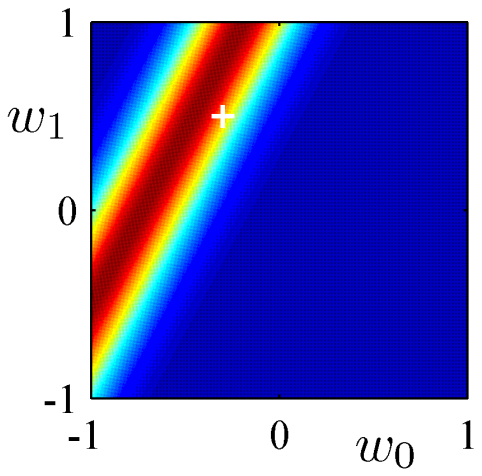
Data Space



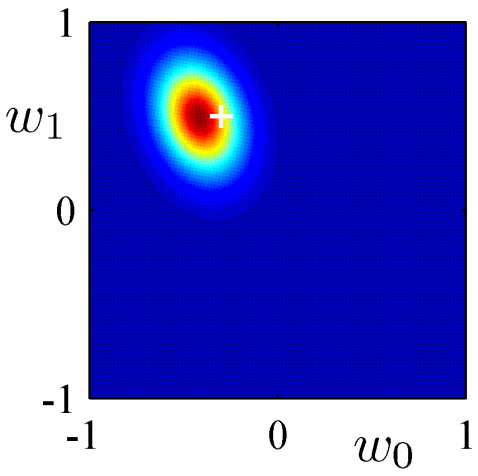
Bayesian Linear Regression

2 data points observed

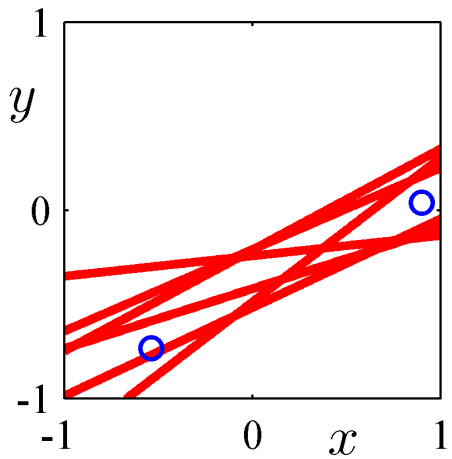
Likelihood



Posterior



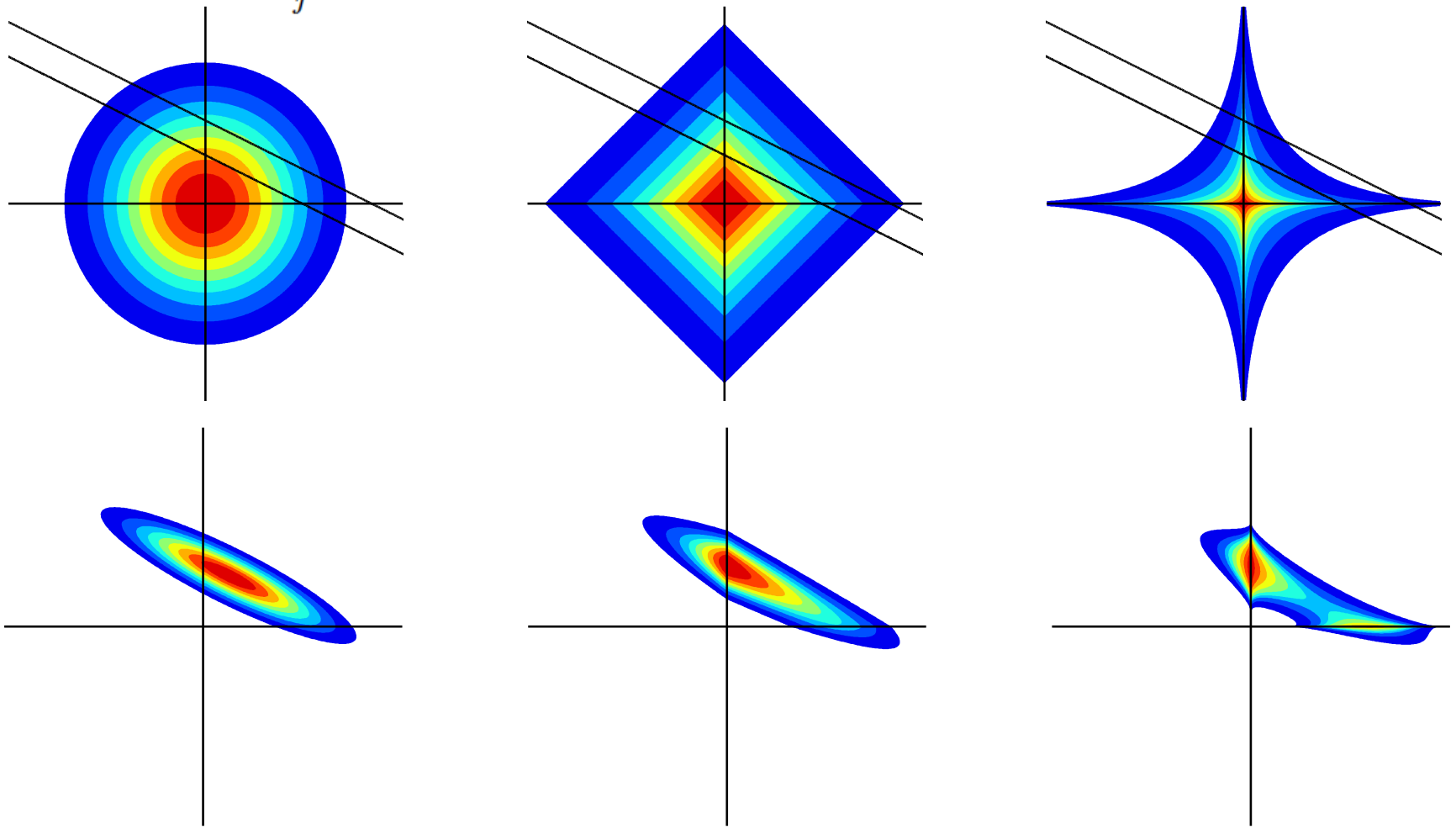
Data Space



Comparing Regression Posteriors

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

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



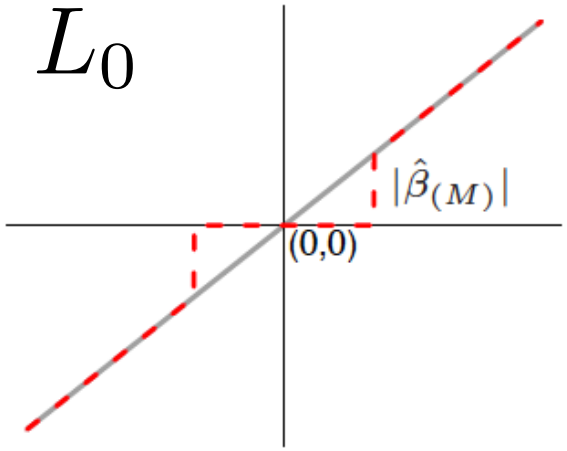
b=2 b=1 b=0.4

Shrinkage for Orthonormal Features

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

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

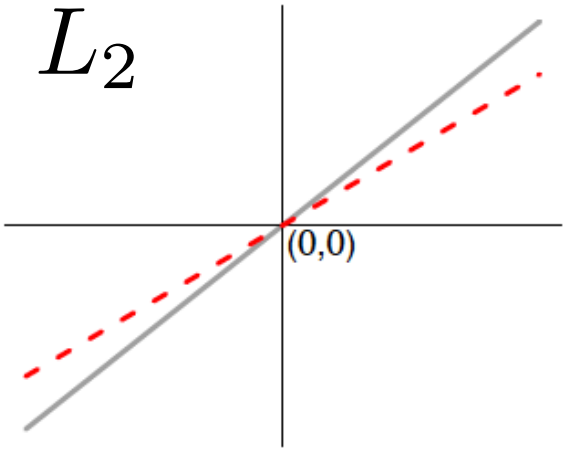
Best Subset



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

Hard thresholding:
Goal of discrete feature selection

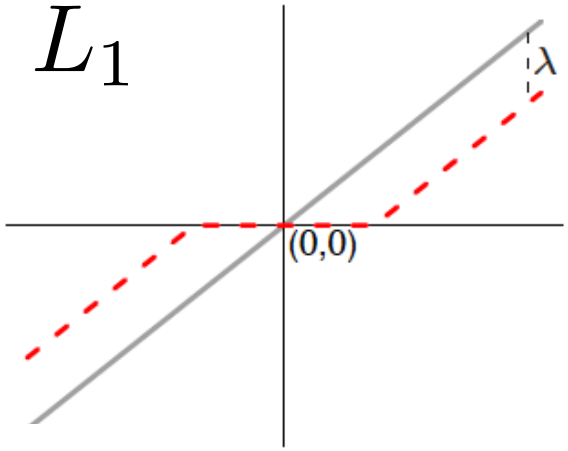
Ridge



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

Linear Shrinkage:
All coefficients remain non-zero

Lasso

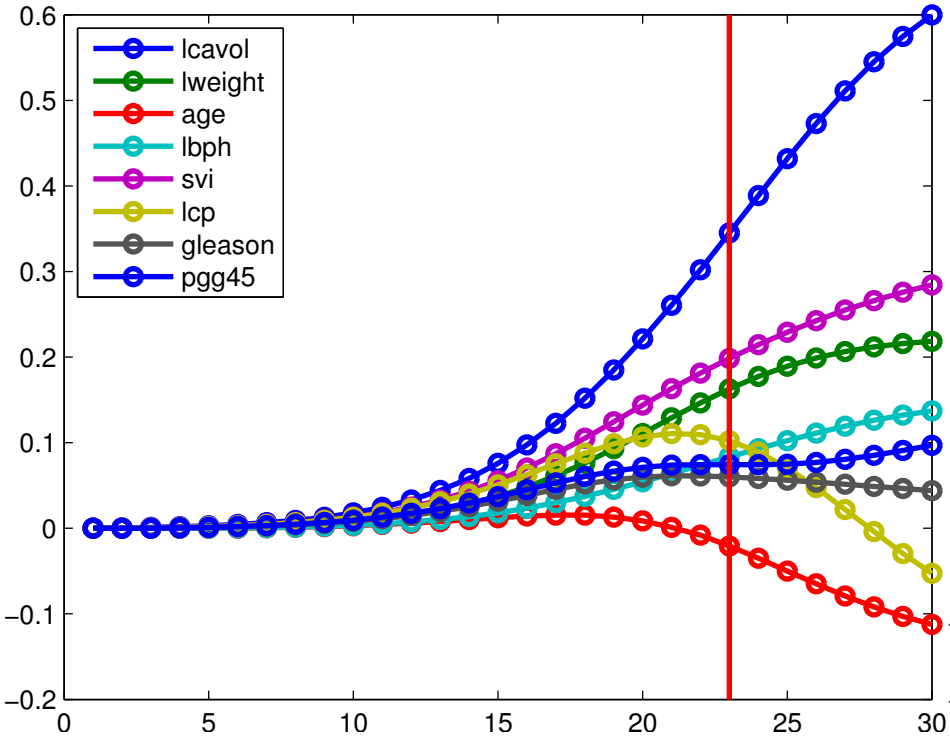


$$\hat{w}_k^{lasso} = \text{sign}(\hat{w}_k^{OLS})\left(|\hat{w}_k^{OLS}| - \frac{\lambda}{2}\right)_+$$

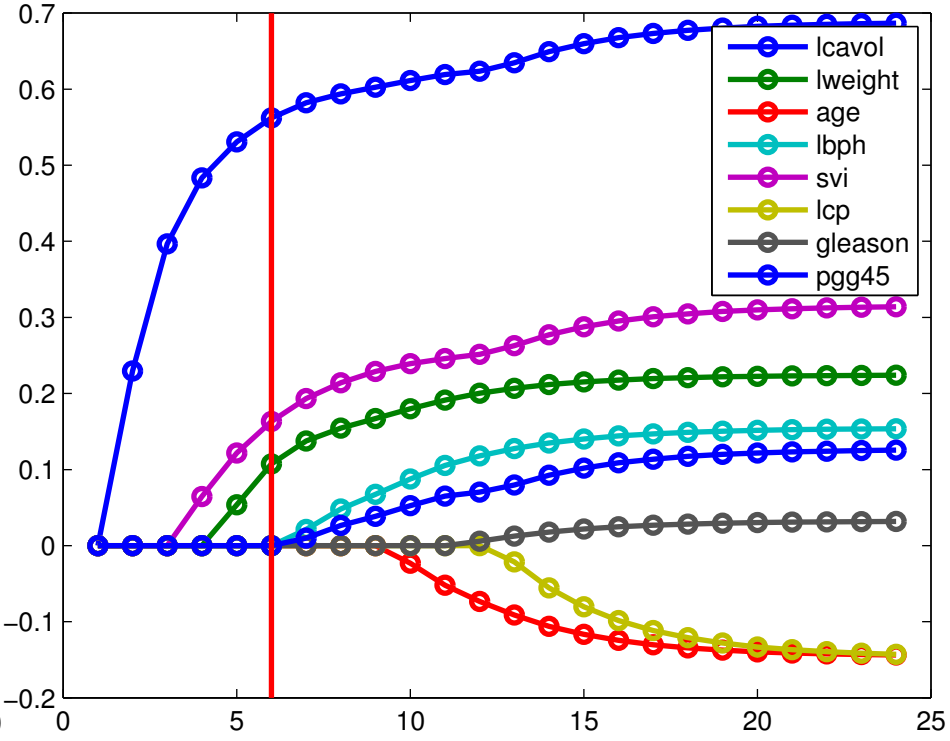
Soft thresholding:
"Least Absolute Selection & Shrinkage Operator"

Regularization Paths

Prostate Cancer Dataset with N=67, D=8



***Ridge**: Bound on L_2 norm*

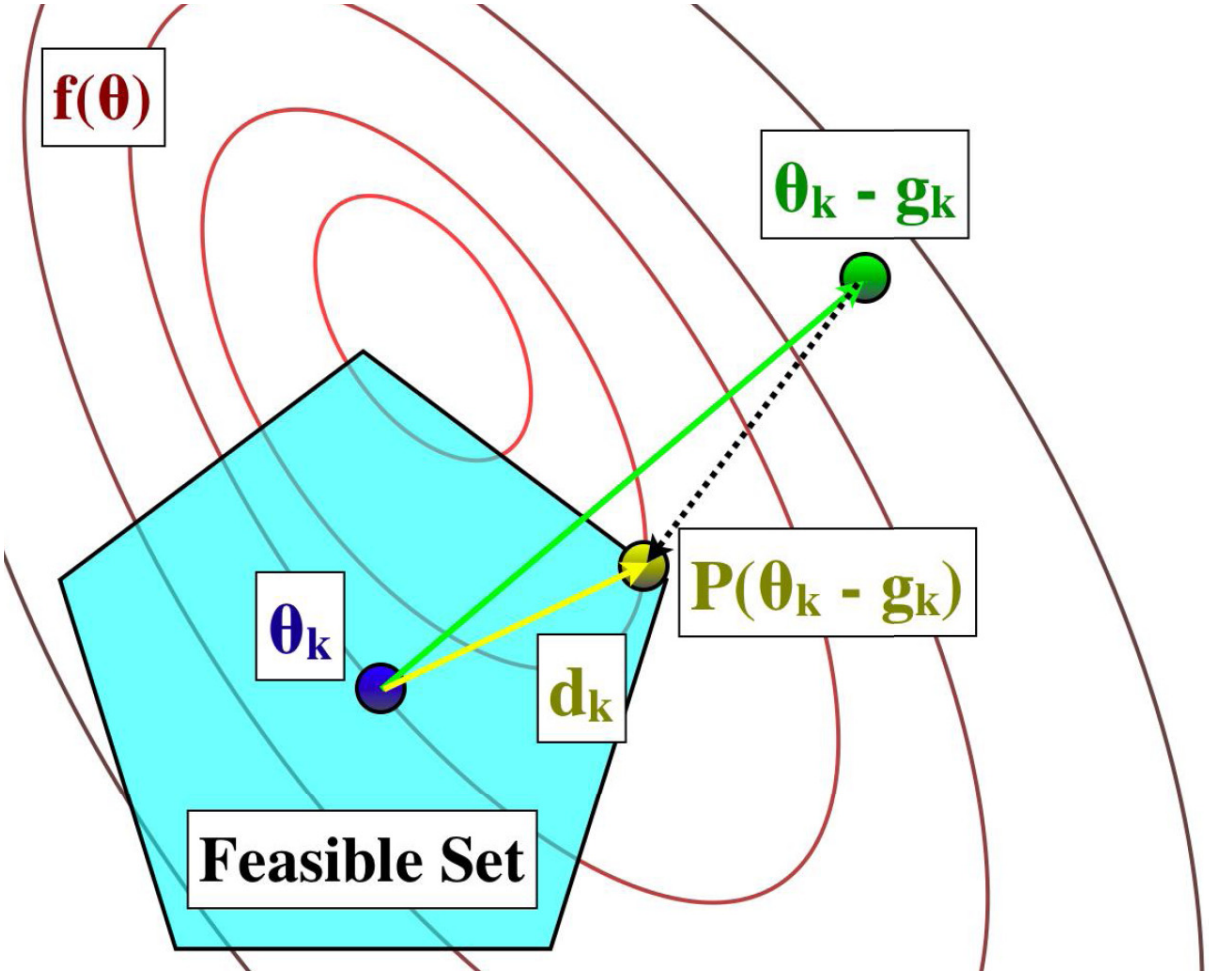


***Lasso**: Bound on L_1 norm*

Vertical lines are models chosen by cross-validation

Optimization: Projected Gradient

$$\min_{\mathbf{w}} NLL(\mathbf{w}) + \lambda||\mathbf{w}||_1 \Leftrightarrow \min_{\mathbf{w}^+,\mathbf{w}^-} NLL(\mathbf{w}^+ - \mathbf{w}^-) + \lambda \sum_j (w_j^+ + w_j^-) \text{ s.t. } w_j^+ \geq 0, w_j^- \geq 0$$



Generic method based on gradient & projection operators:

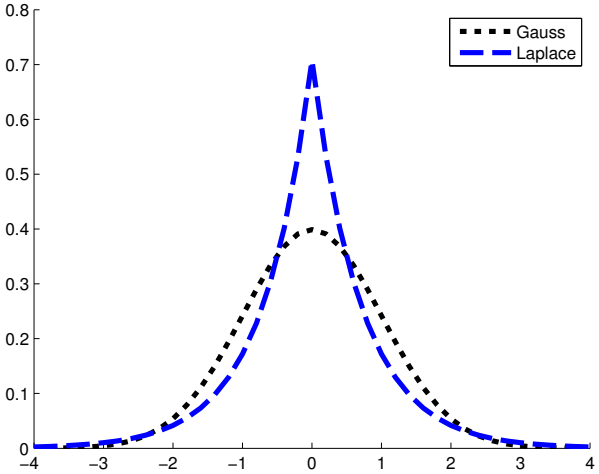
$$\begin{aligned} \mathbf{w}_{k+1} &= \mathbf{w}_k + \eta_k \mathbf{d}_k \\ \mathbf{d}_k &= \text{proj}(\mathbf{w}_k - \eta_k \mathbf{g}_k) - \mathbf{w}_k \end{aligned}$$

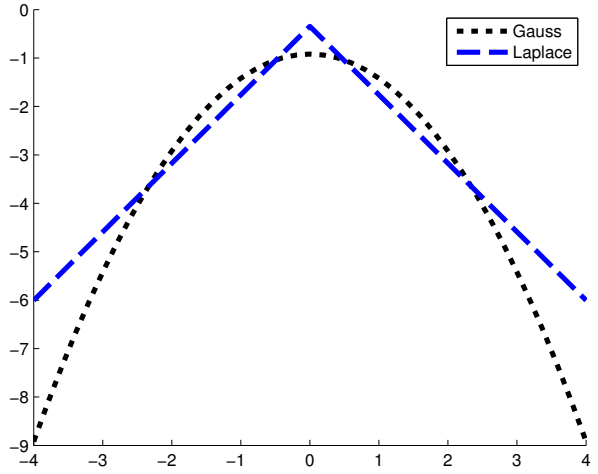
Projection onto non-negativity constraint is trivial:

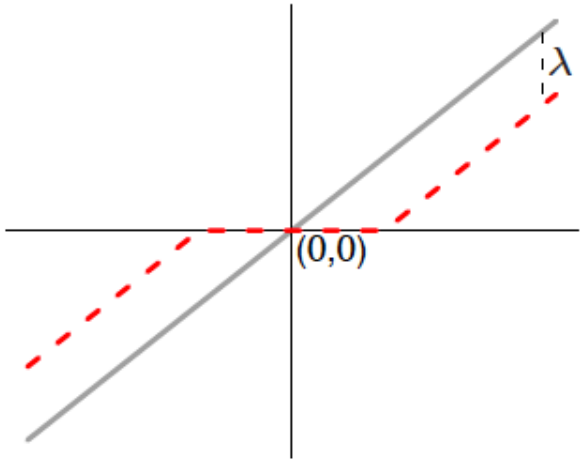
$$w_i := \max(w_i, 0)$$

Good properties, extensions choose even better descent directions…

Potential Issues with the Lasso







- *MAP estimate is atypical:* Samples from the prior have all weights non-zero with probability one
  - Full Bayesian learning requires richer prior models (often perform better at greater computational cost)
- In addition to setting some coefficients exactly to zero, non-zero coefficients are significantly biased towards zero
  - *Two-stage estimators* estimate weight vector support, then re-estimate weights with a less strong prior
- Theory guarantees *support recovery* in some conditions, but can be *unstable* when features are strongly correlated