Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2011

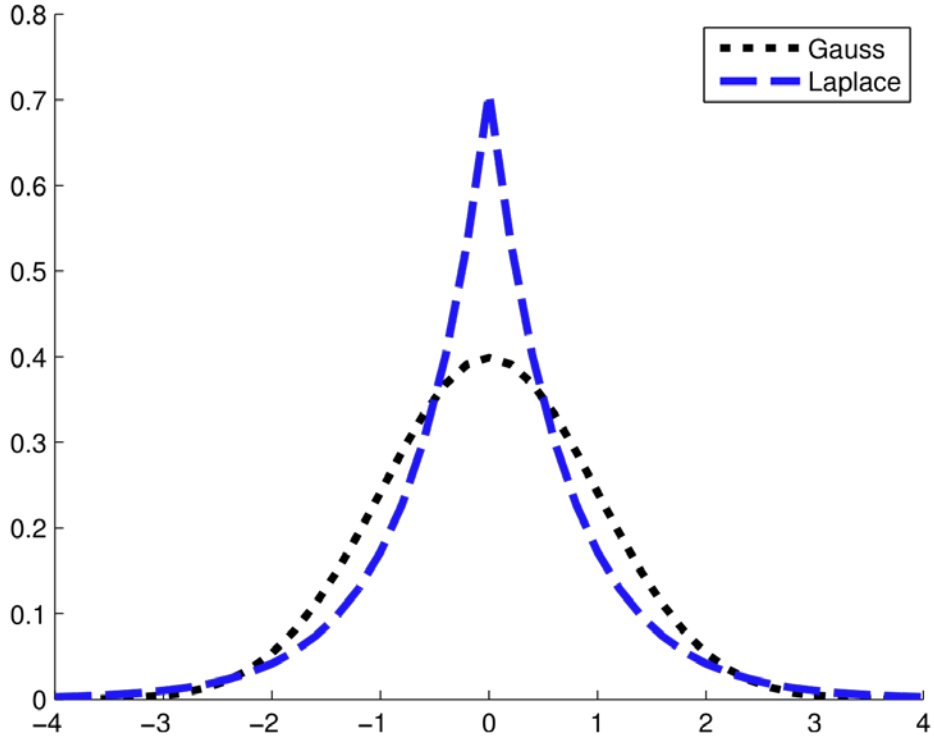
Prof. Erik Sudderth

Lecture 14: L_1 Optimization, Kernel Methods, Gaussian Process Regression

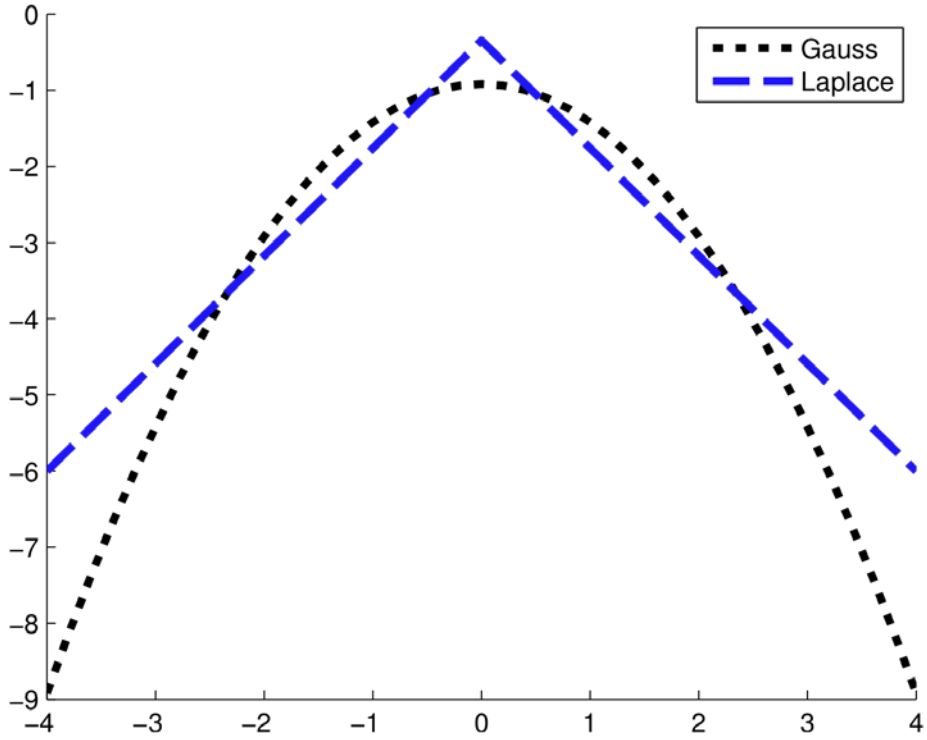
Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

Laplace Distribution

Probability Densities



Log Probability Densities



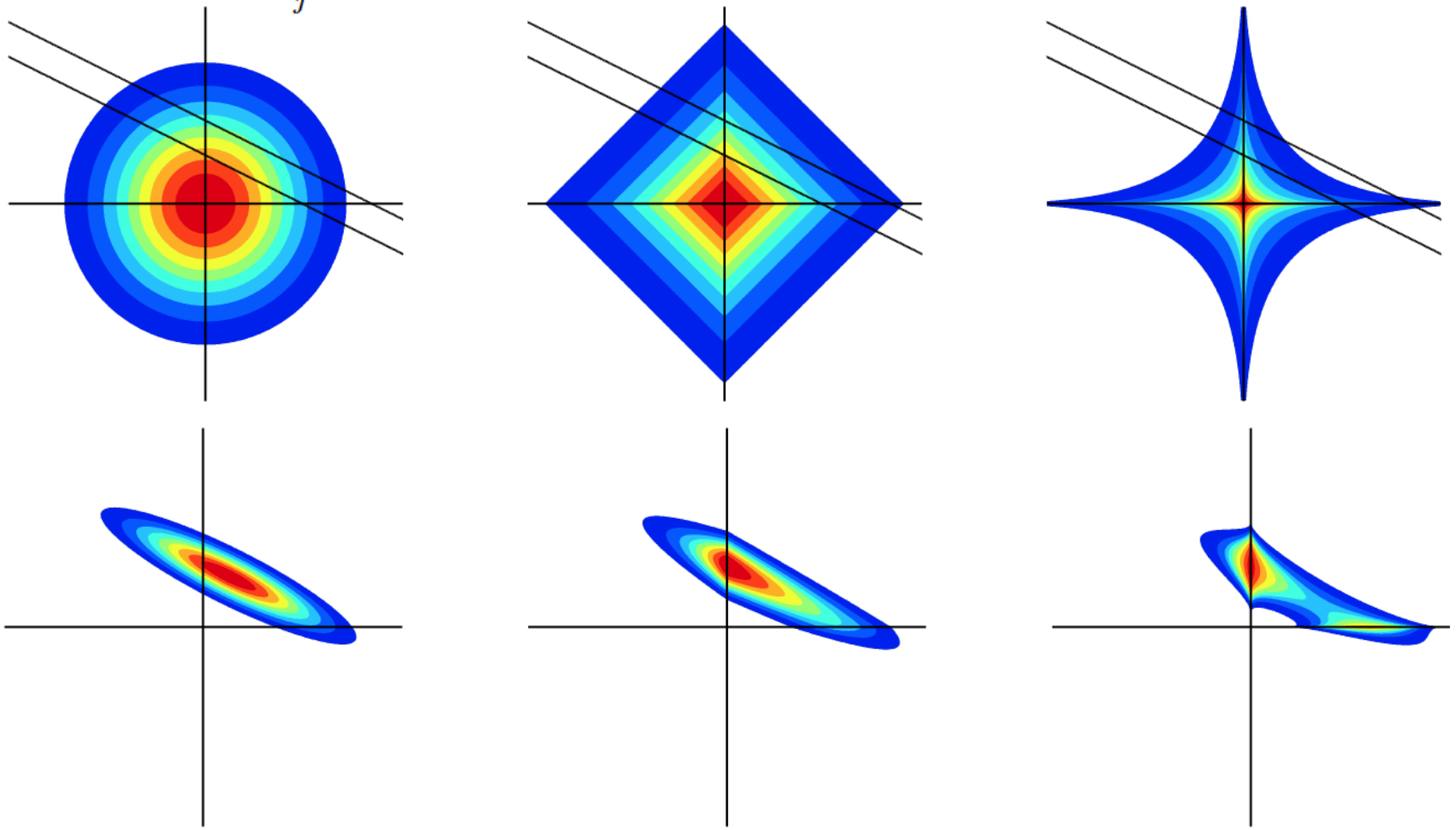
Relative to Gaussian distributions with equal variance:

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

Comparing Regression Posteriors

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

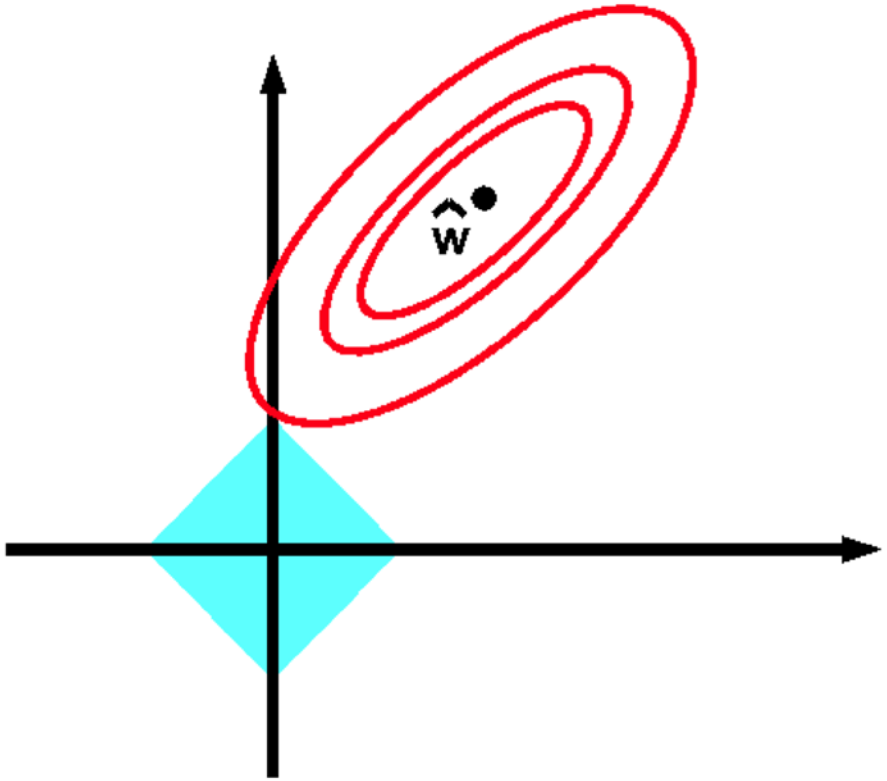
$$\mathrm{ExpPower}(w|\mu, a, b) := \frac{b}{2a\Gamma(1/b)} \exp(-\frac{|x-\mu|}{a})^b$$



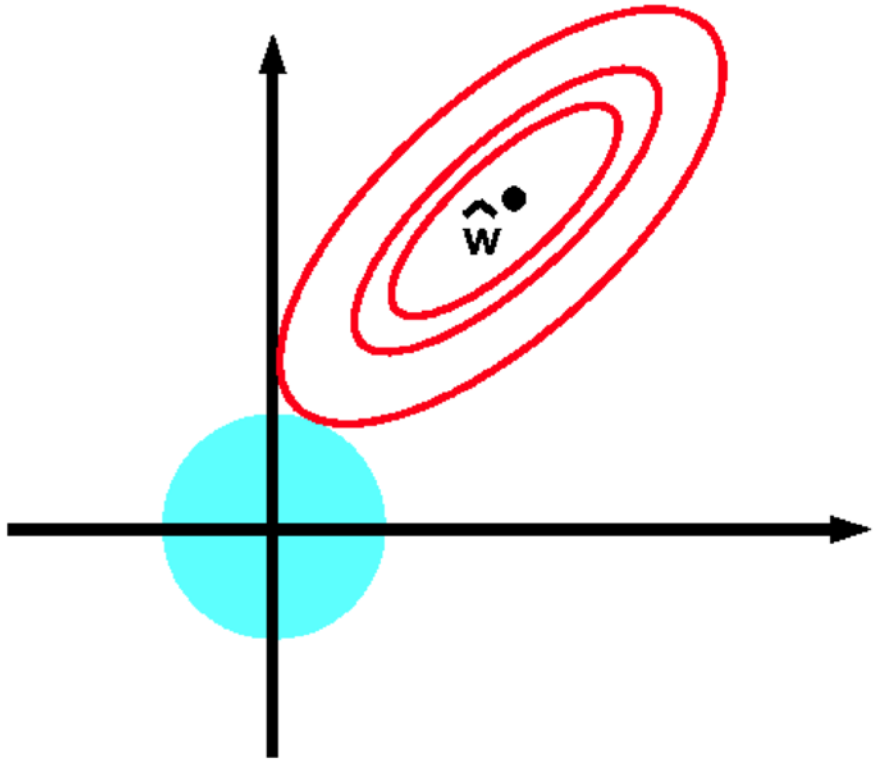
b=2 b=1 b=0.4

Constrained Optimization

Laplacian prior
L_1 regularization
Lasso regression



Gaussian prior
L_2 regularization
Ridge regression



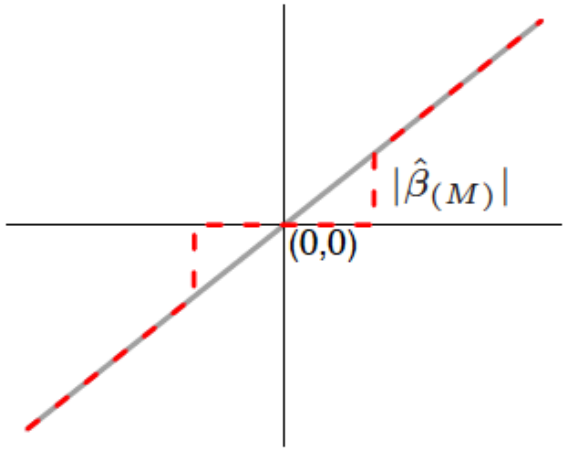
Where do level sets of the quadratic regression cost function first intersect the constraint set?

Shrinkage for Orthonormal Features

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

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

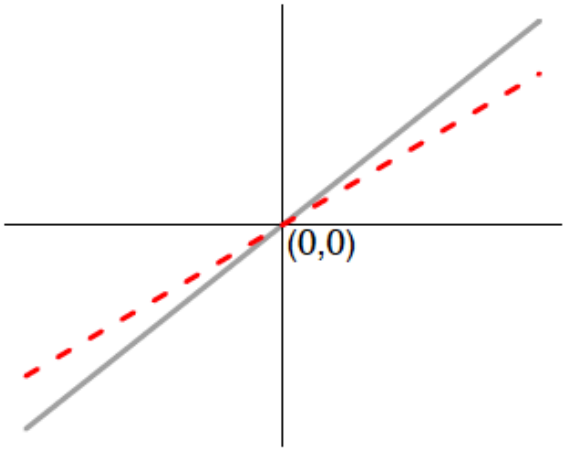
Best Subset



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

Hard thresholding:
Goal of discrete feature selection

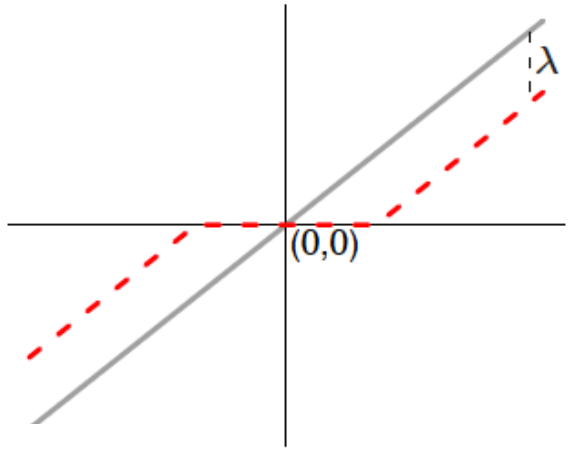
Ridge



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

Linear Shrinkage:
All coefficients remain non-zero

Lasso

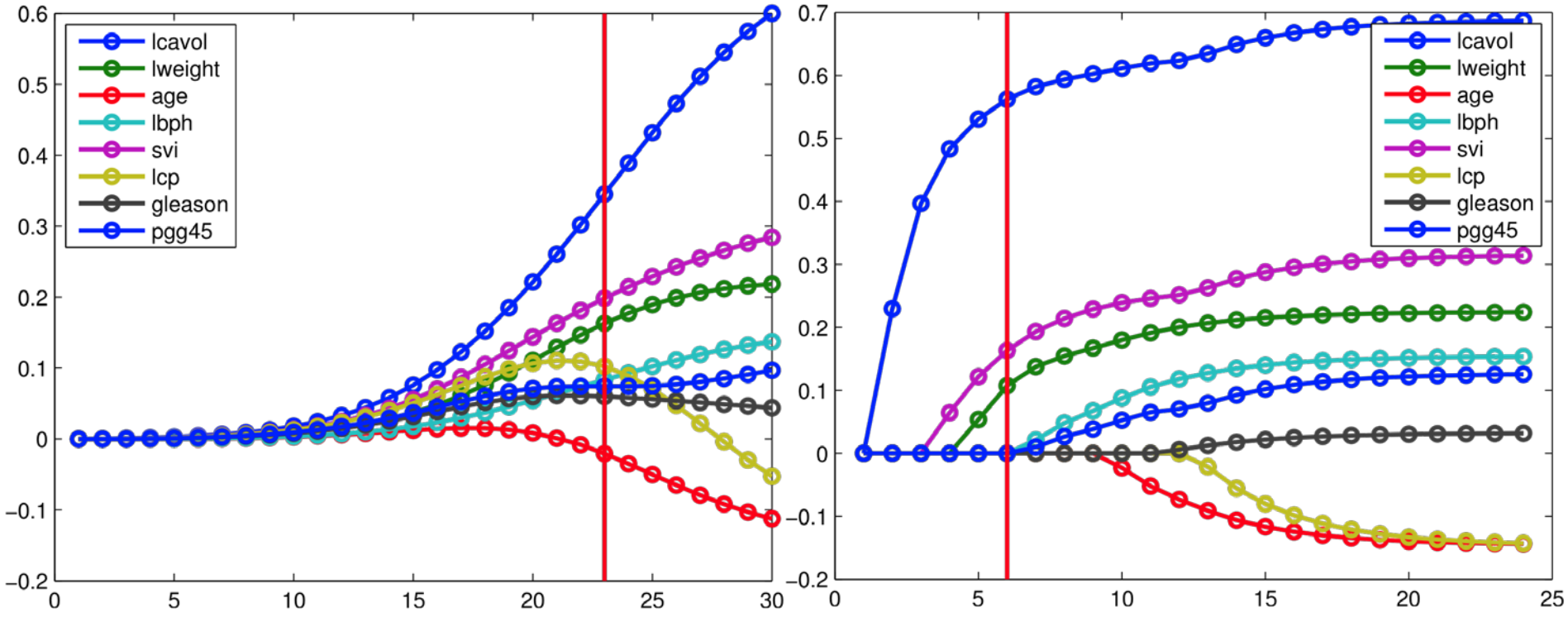


$$\hat{w}_k^{lasso} = \text{sign}(\hat{w}_k^{OLS})\left(|\hat{w}_k^{OLS}| - \frac{\lambda}{2}\right)_+$$

Soft thresholding:
"Least Absolute Selection & Shrinkage Operator"

Regularization Paths

Prostate Cancer Dataset with N=67, D=8



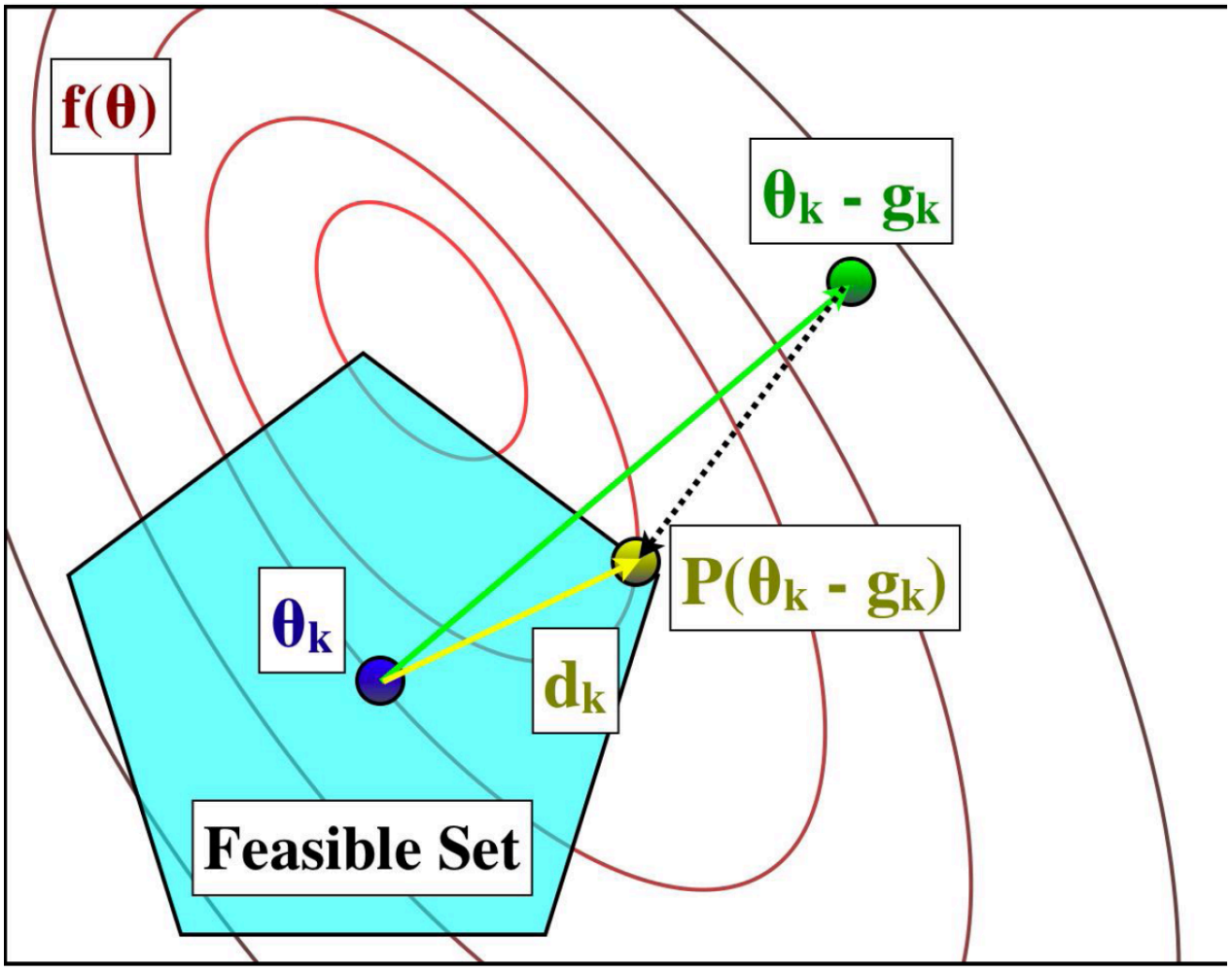
***Ridge**: Bound on L_2 norm*

***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…