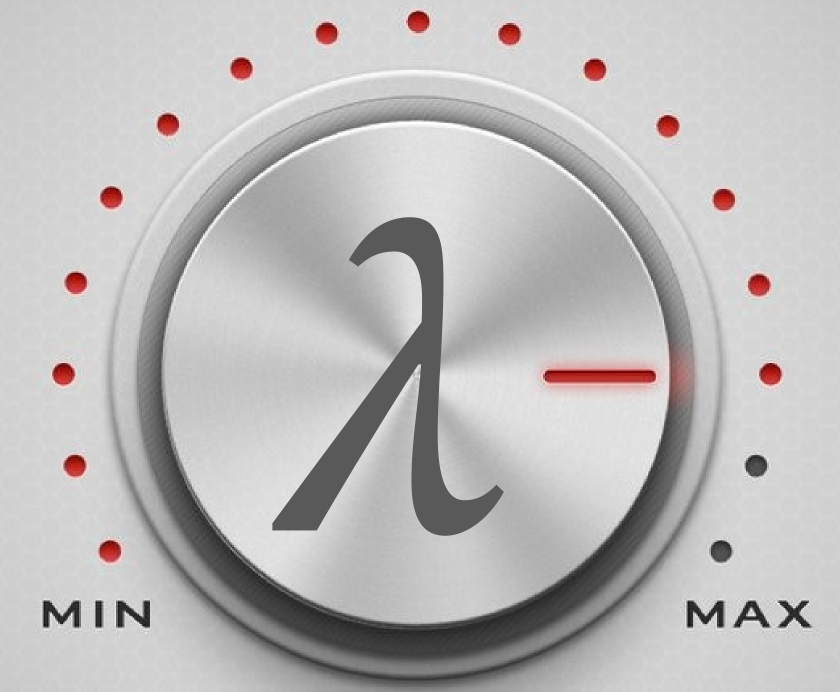
YSC2239 Lecture 19

Regularization

Controlling the
Model Complexity



Basic Idea

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathbf{Loss}\left(y_i, f_\theta(x_i)\right)$$

Such that:

f_θ does not "overfit"

Can we make this more formal?

Basic Idea

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathbf{Loss}\left(y_i, f_\theta(x_i)\right)$$

Such that:

$$\text{Complexity}(f_\theta) \leq \beta$$

Regularization Hyperparameter

How do we define this?

Idealized Notion of Complexity

$$\text{Complexity}(f_\theta) \leq \beta$$

- Focus on complexity of **linear models**:
  - Number and kinds of features
- Ideal definition:

$$\textbf{Complexity}(f_\theta) = \sum_{j=1}^{d} \mathbb{I}\left[\theta_j \neq 0\right]$$

Number of non-zero parameters

- Why?

Ideal "Regularization"

Find the best value of θ which uses fewer than β features.

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathbf{Loss}\left(y_i, f_\theta(x_i)\right)$$

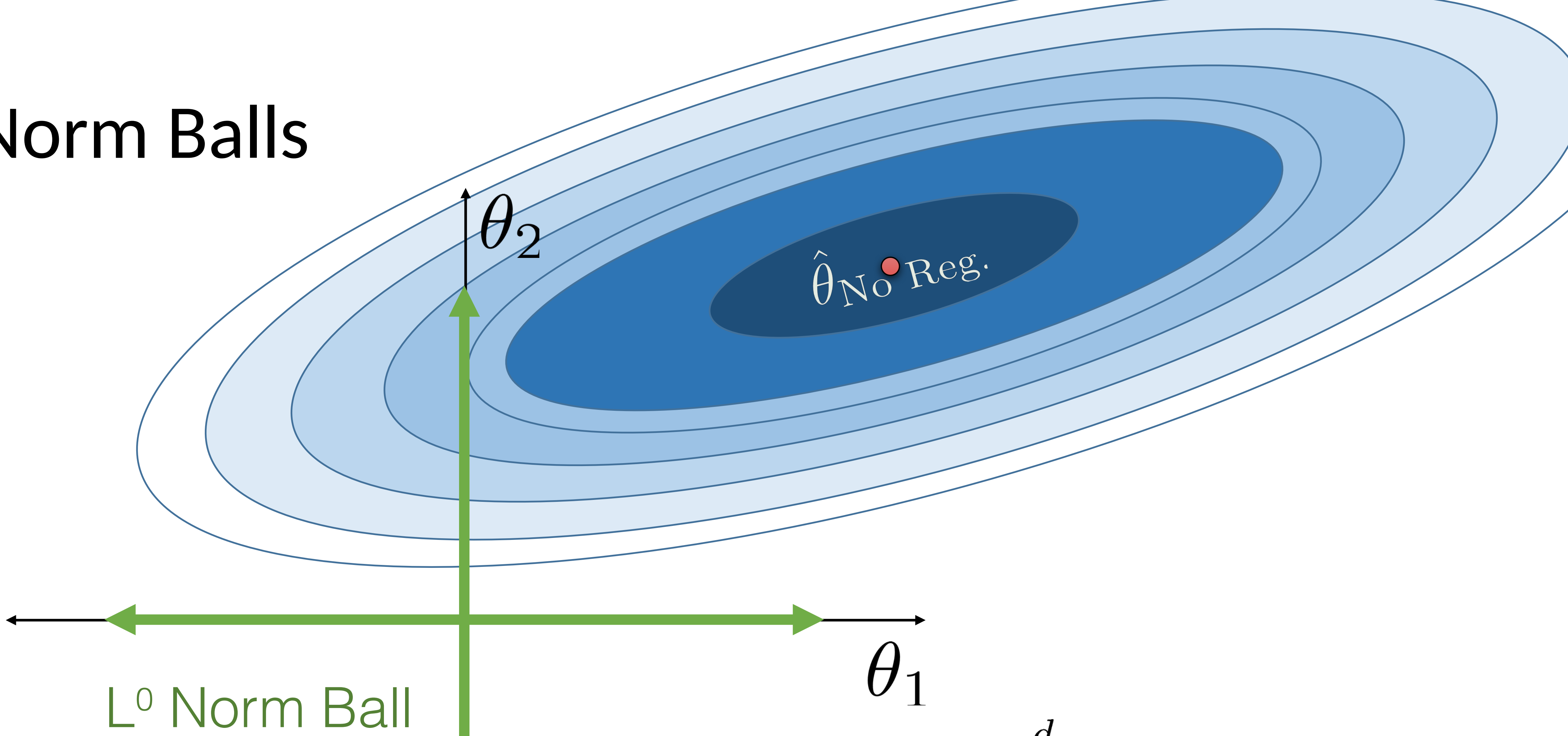
Such that:

Need an approximation!

$$\mathbf{Complexity}(f_\theta) = \sum_{j=1}^{d} \mathbb{I}\left[\theta_j \neq 0\right] \leq \beta$$

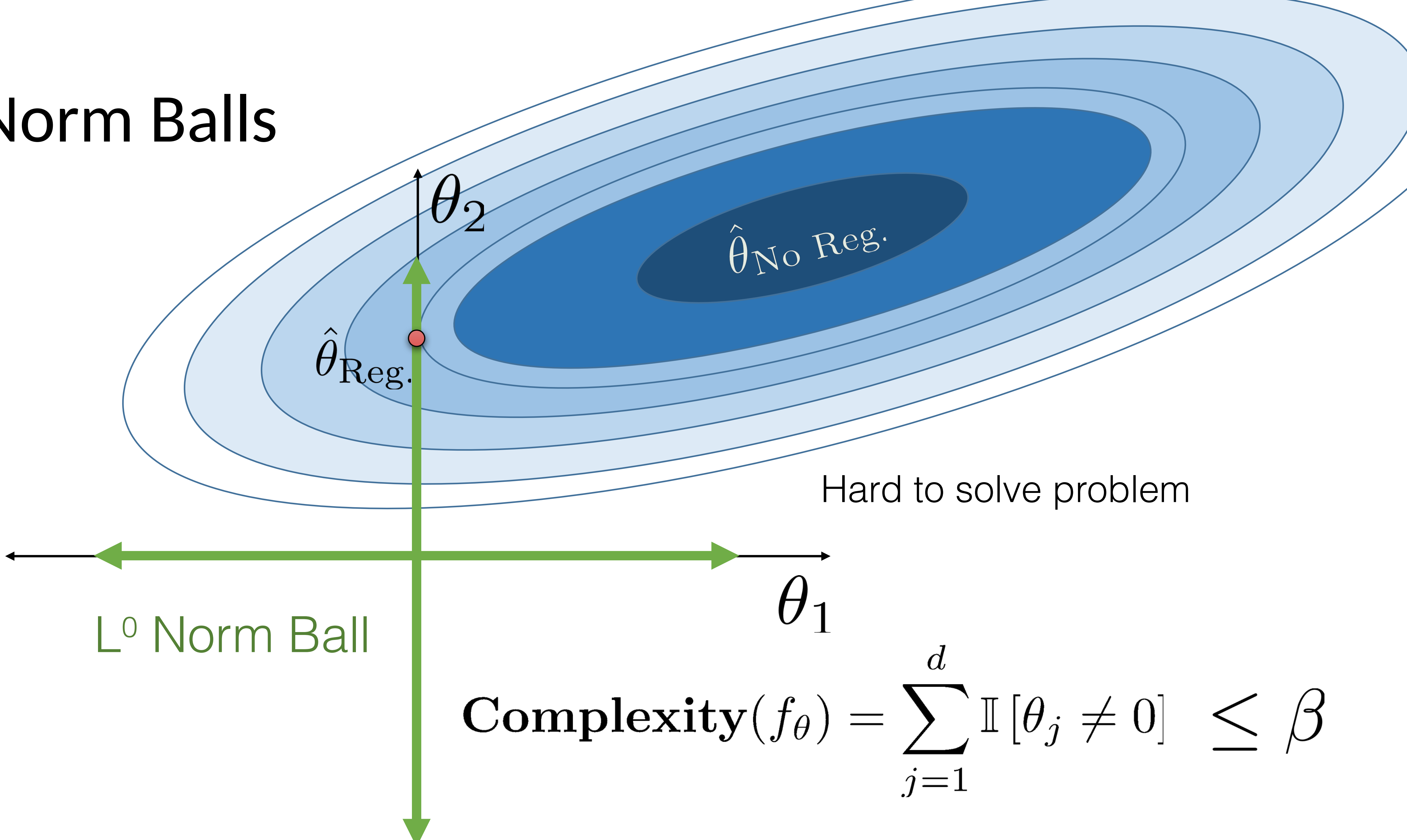
Combinatorial search problem – NP-hard to solve in general.

Norm Balls



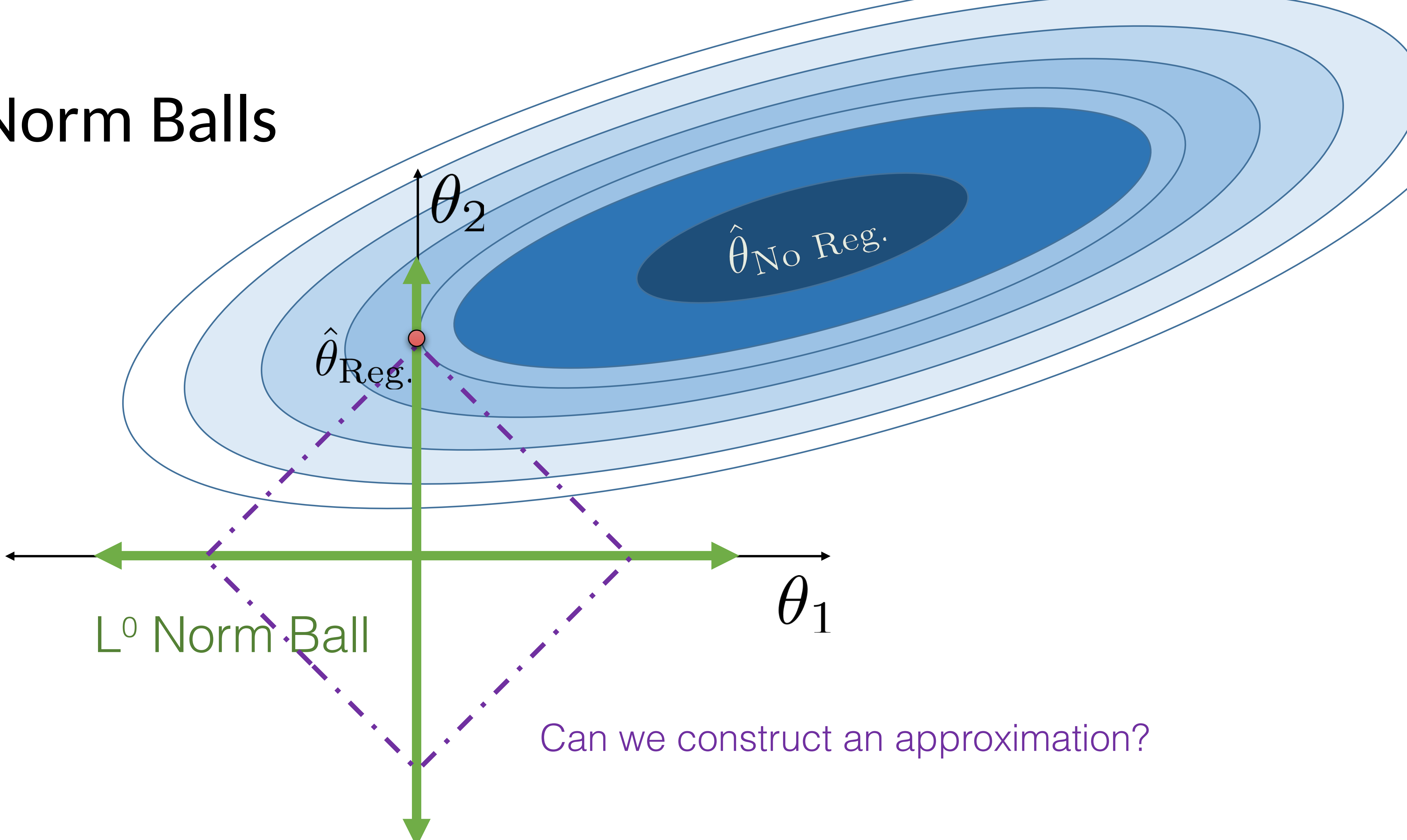
$$\textbf{Complexity}(f_\theta) = \sum_{j=1}^{d} \mathbb{I}\left[\theta_j \neq 0\right]$$

Norm Balls



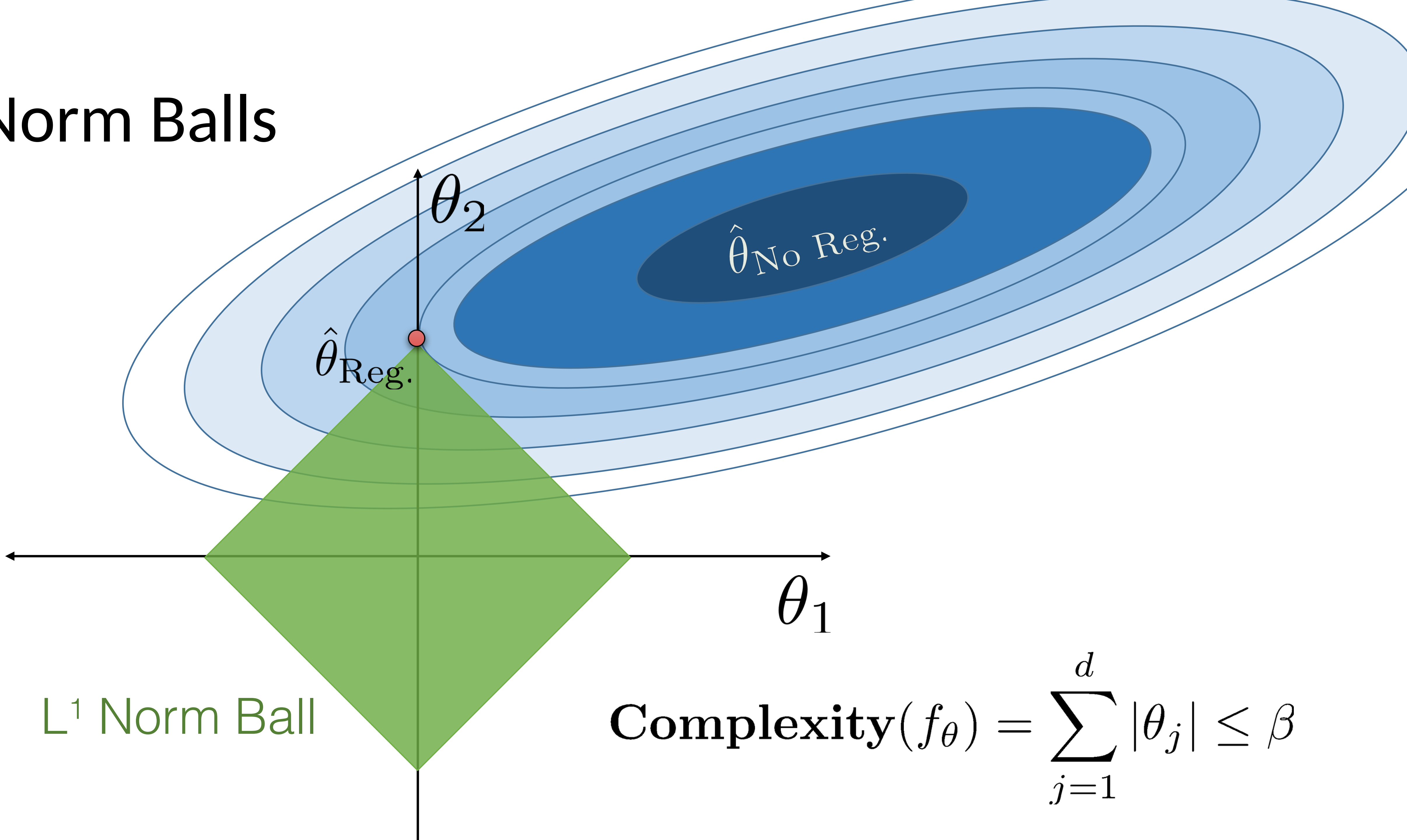
$$\textbf{Complexity}(f_\theta) = \sum_{j=1}^{d} \mathbb{I}\left[\theta_j \neq 0\right] \leq \beta$$

Norm Balls



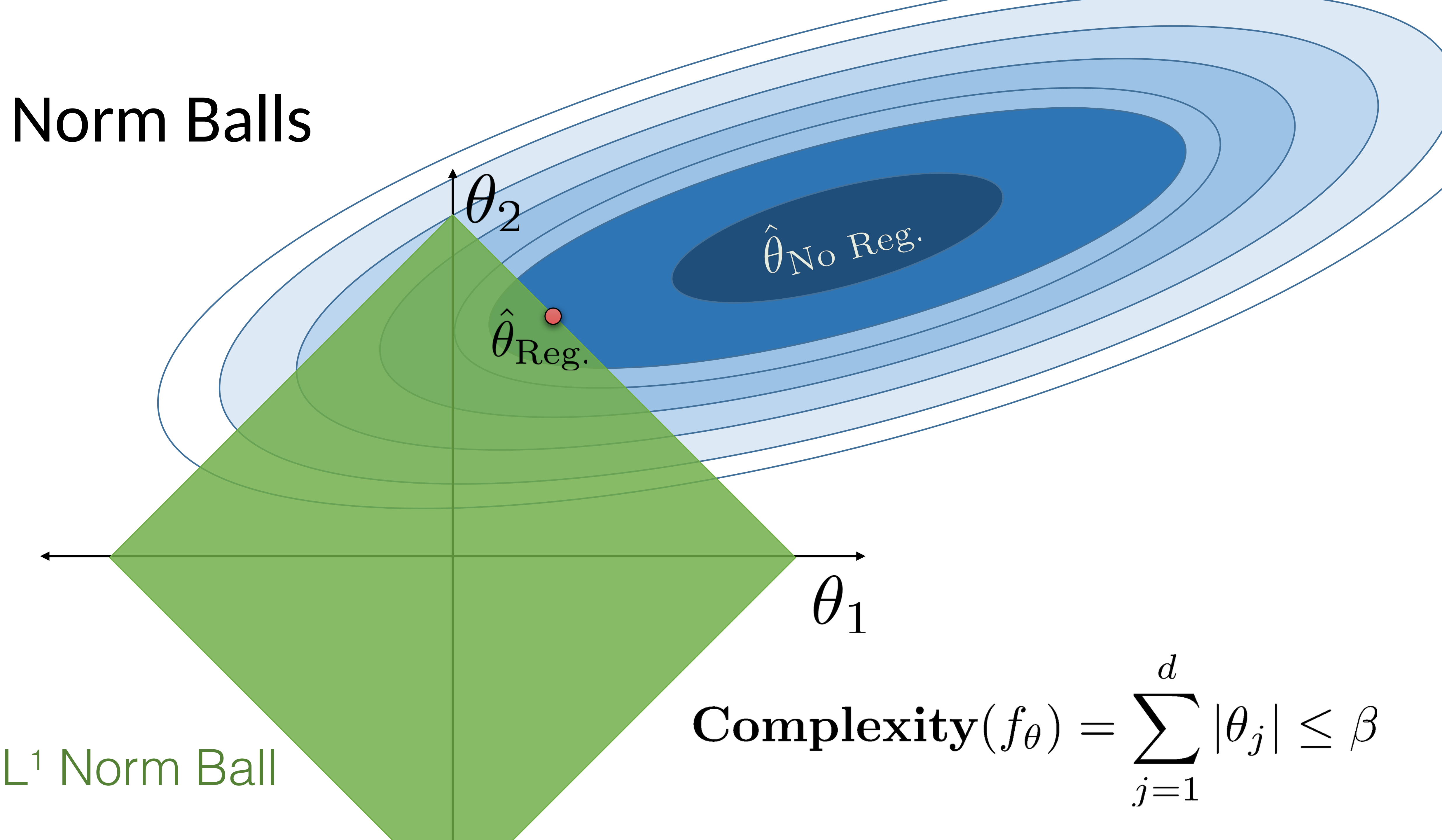
Can we construct an approximation?

Norm Balls



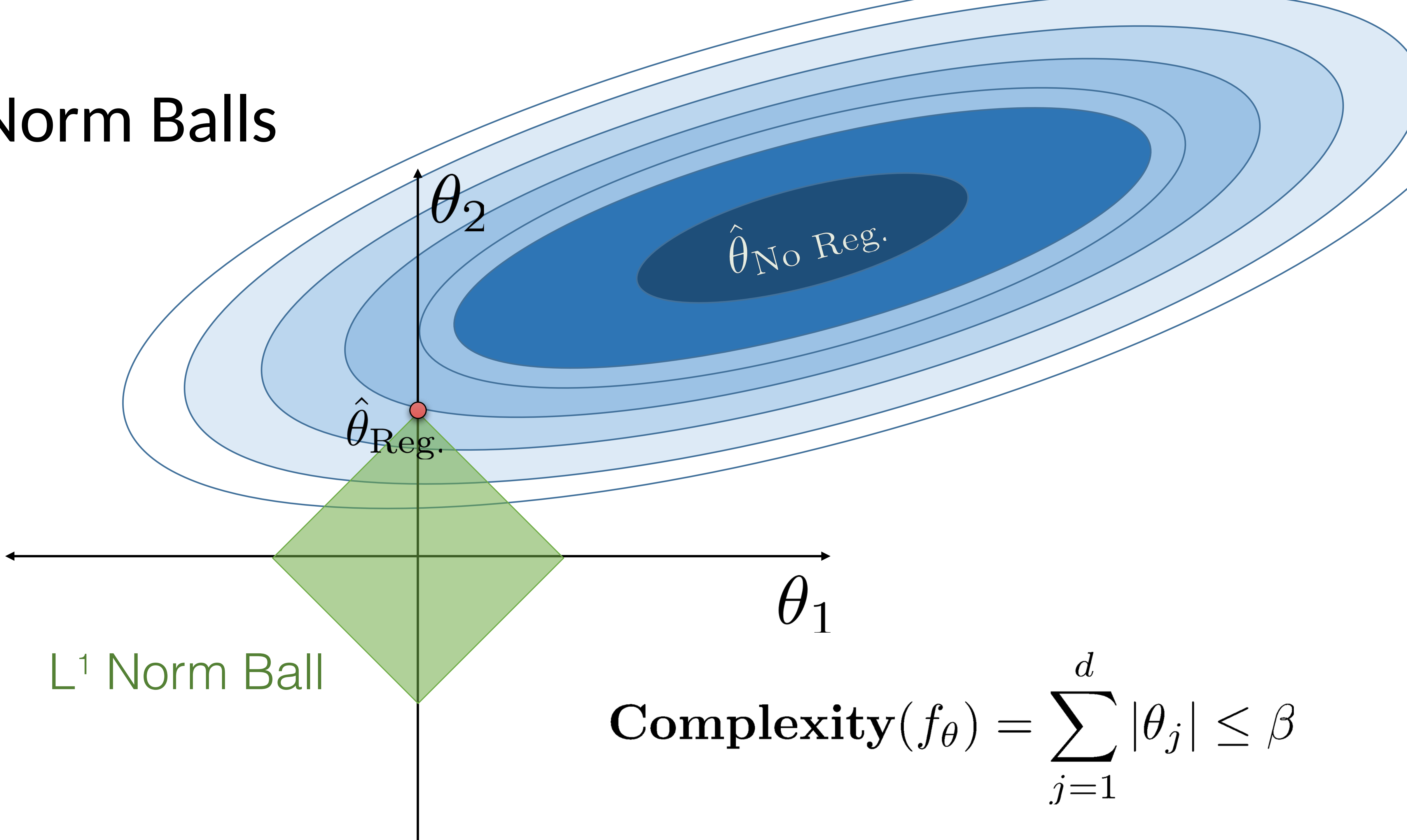
$$\textbf{Complexity}(f_\theta) = \sum_{j=1}^{d} |\theta_j| \leq \beta$$

Norm Balls



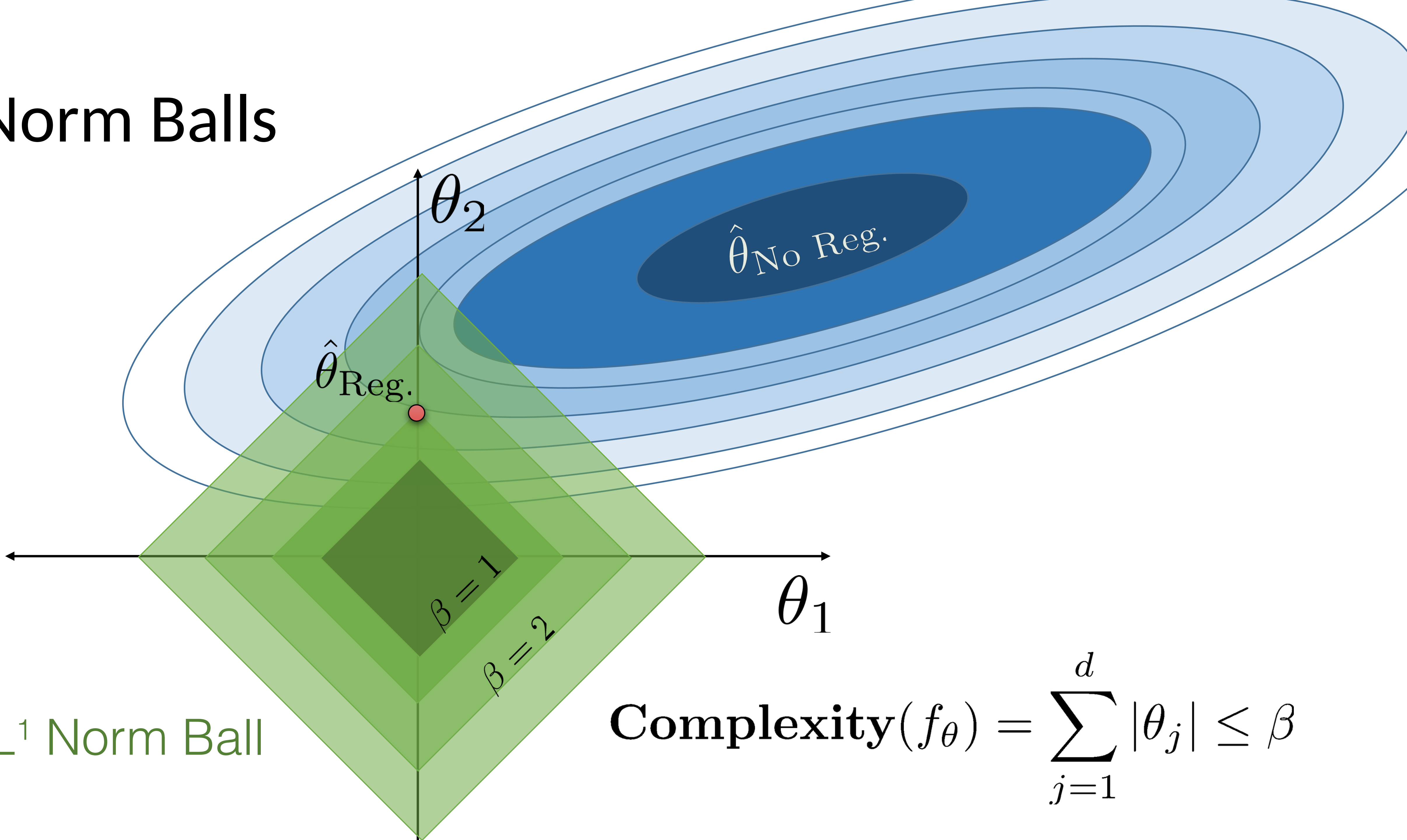
$$\textbf{Complexity}(f_\theta) = \sum_{j=1}^{d} |\theta_j| \leq \beta$$

Norm Balls



$$\textbf{Complexity}(f_\theta) = \sum_{j=1}^{d} |\theta_j| \leq \beta$$

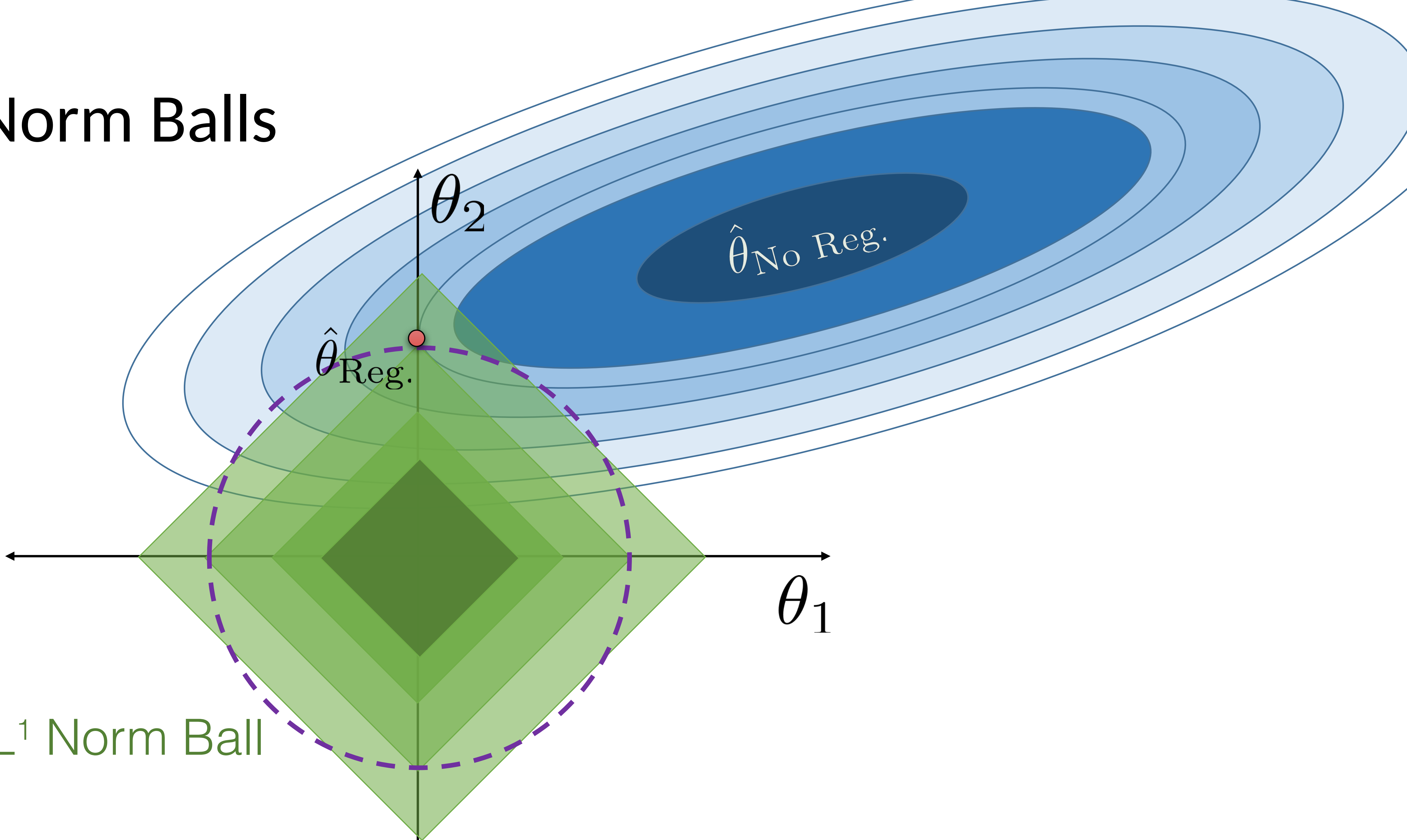
Norm Balls



L¹ Norm Ball

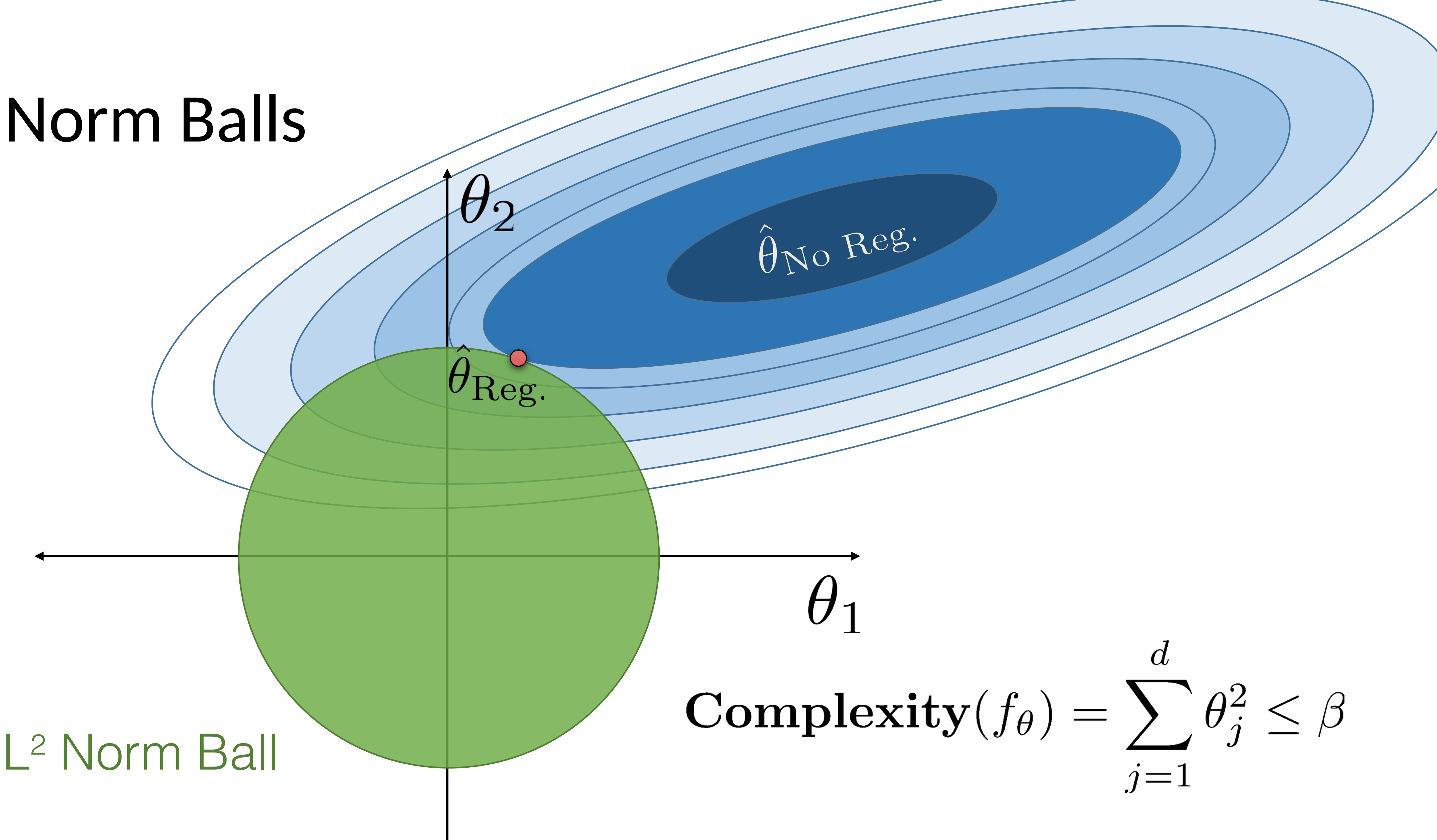
$$\textbf{Complexity}(f_\theta) = \sum_{j=1}^{d} |\theta_j| \leq \beta$$

Norm Balls



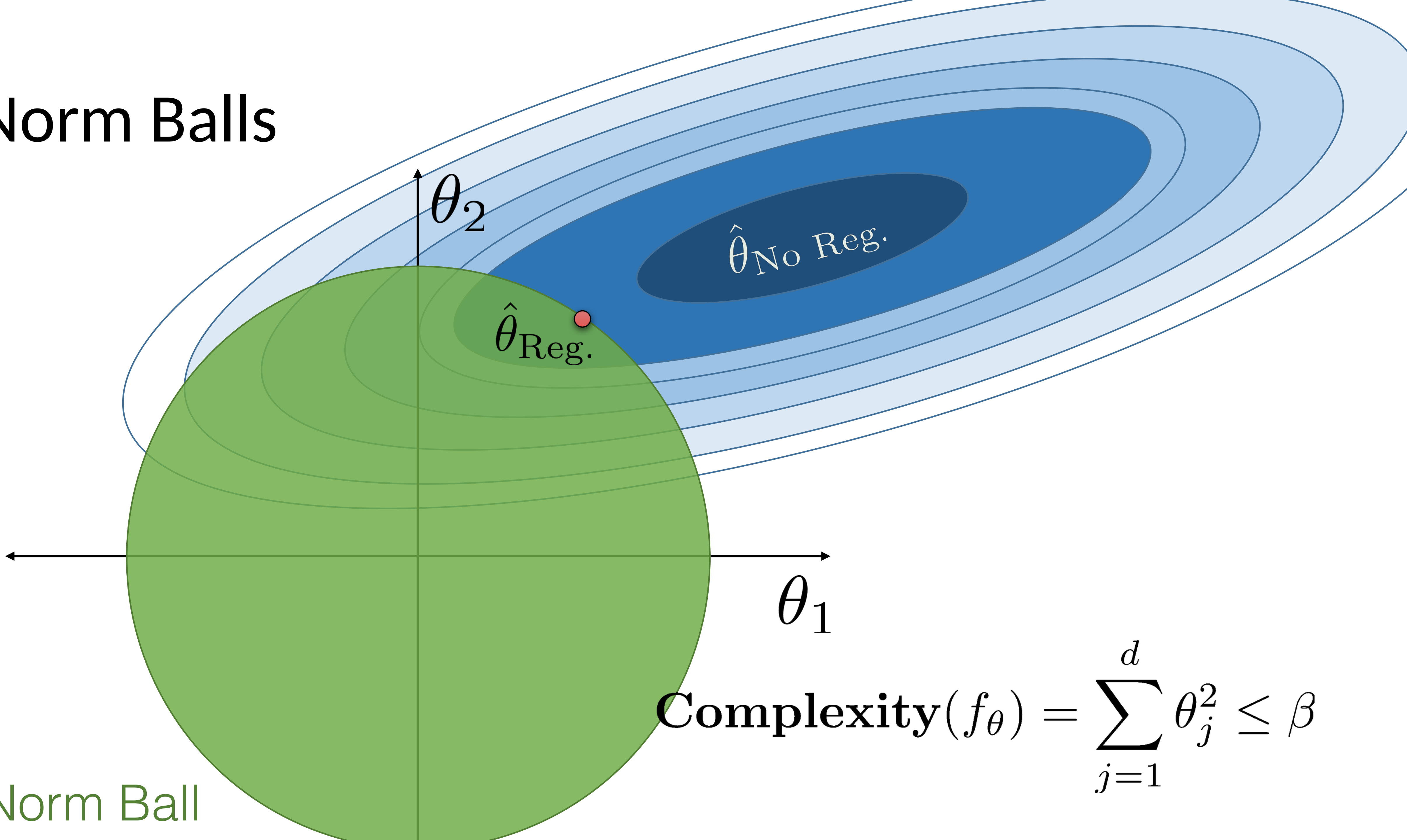
θ_2
$\hat{\theta}_{\text{No Reg.}}$
$\hat{\theta}_{\text{Reg.}}$
θ_1
L¹ Norm Ball

Norm Balls



$$\textbf{Complexity}(f_\theta) = \sum_{j=1}^{d} \theta_j^2 \leq \beta$$

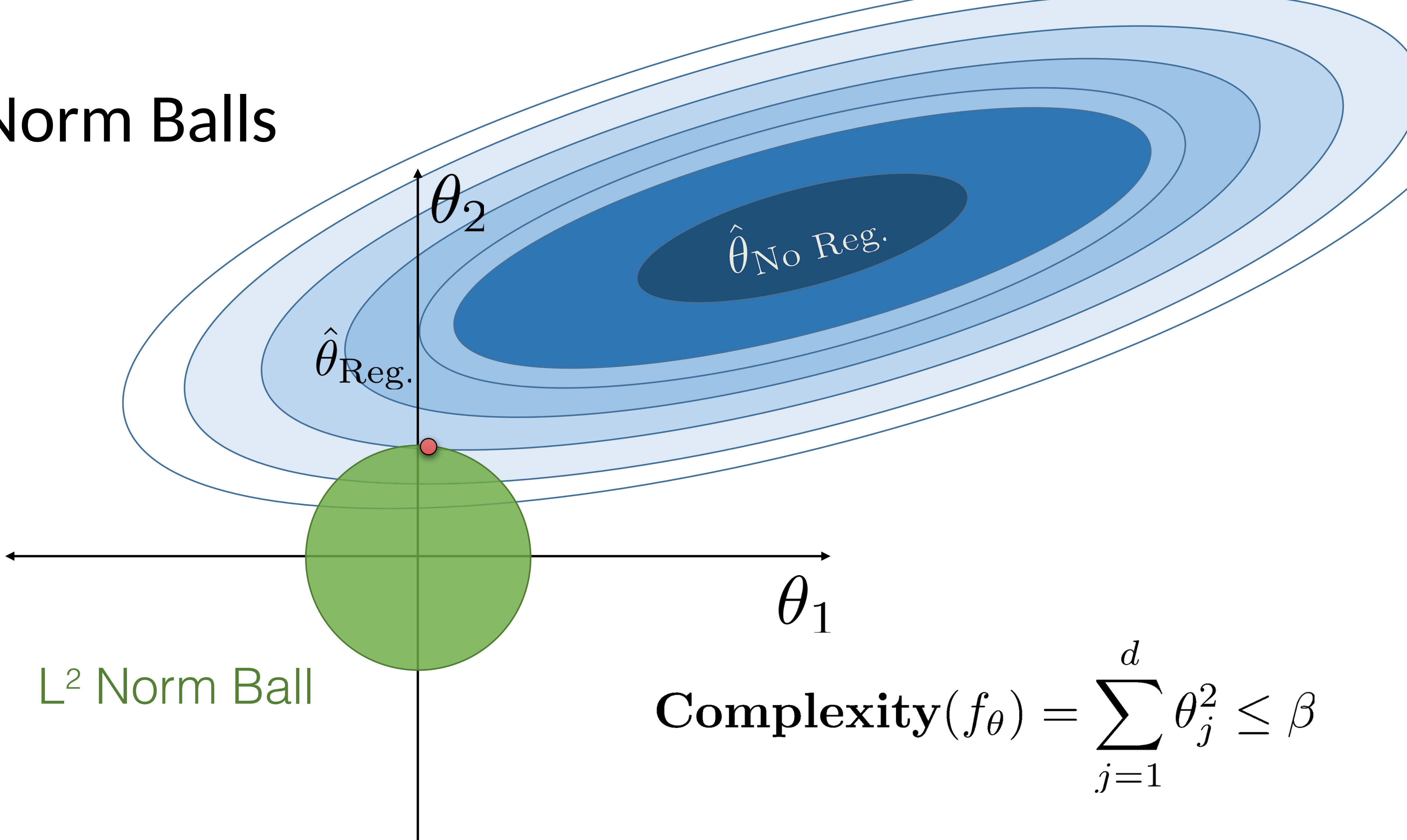
Norm Balls



$$\textbf{Complexity}(f_\theta) = \sum_{j=1}^{d} \theta_j^2 \leq \beta$$

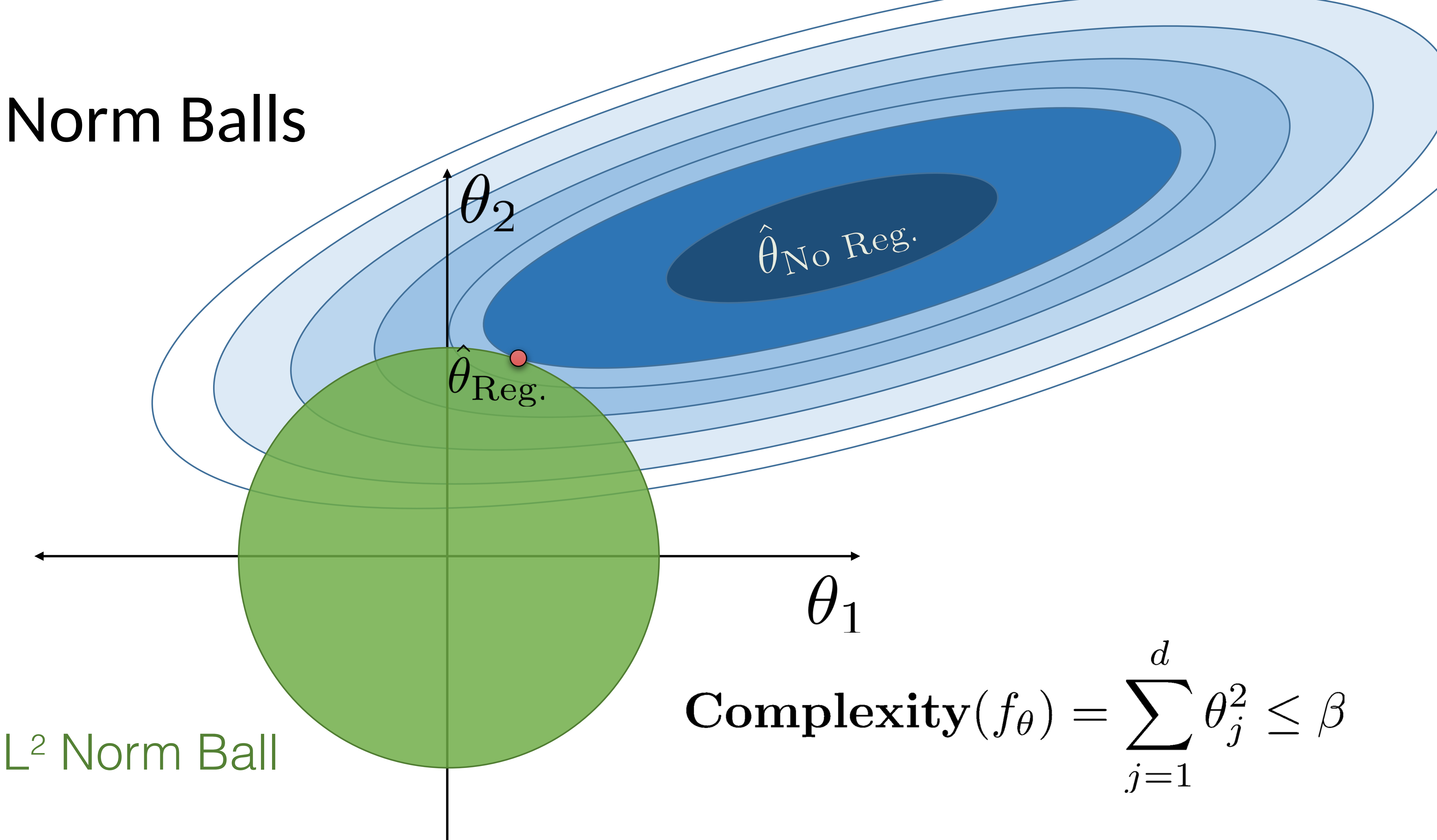
L² Norm Ball

Norm Balls



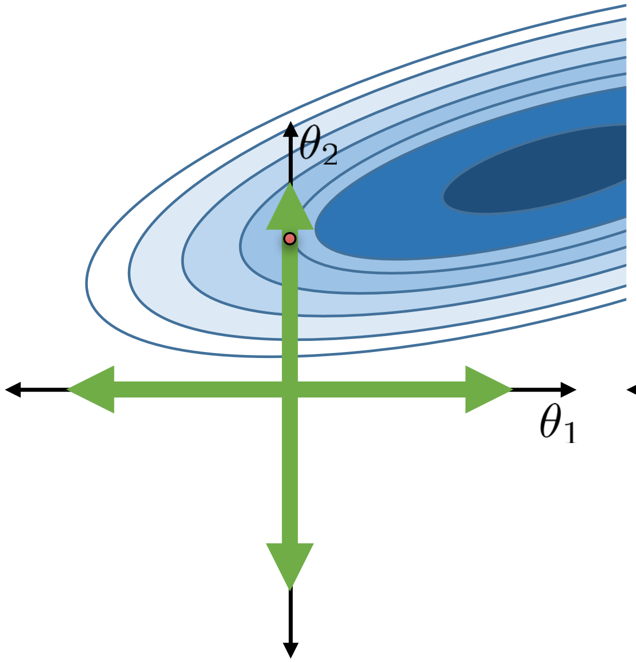
$$\textbf{Complexity}(f_\theta) = \sum_{j=1}^{d} \theta_j^2 \leq \beta$$

Norm Balls



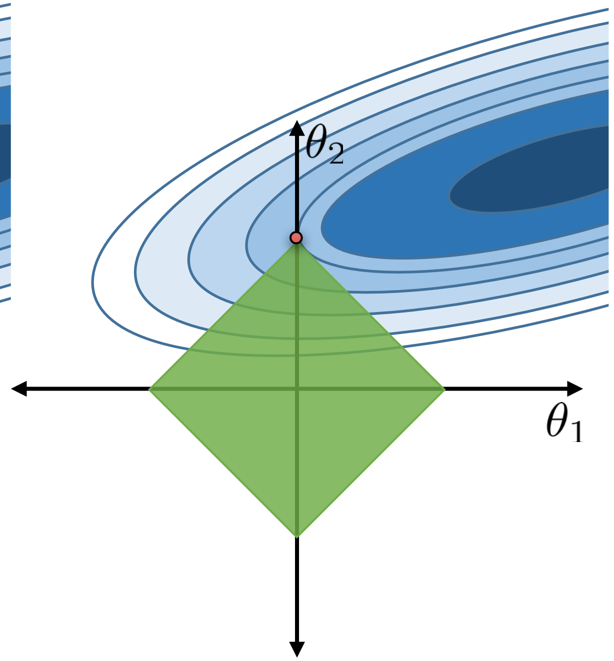
$$\textbf{Complexity}(f_\theta) = \sum_{j=1}^{d} \theta_j^2 \leq \beta$$

L^0 Norm Ball



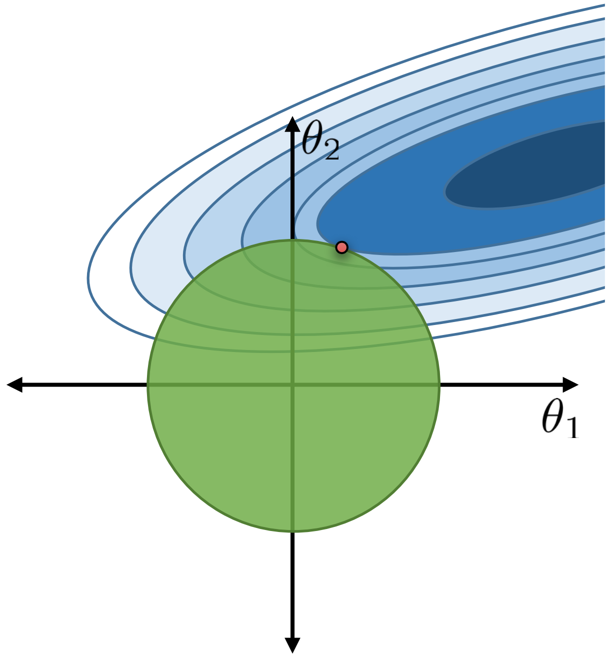
Ideal for **Feature Selection** but combinatorically difficult to optimize

L^1 Norm Ball



Encourages sparse solutions

L^2 Norm Ball



Spreads weight over features, but does not encourage sparsity

Ridge and LASSO Regression

Ridge Regression

“Ridge Regression” is a term for the following specific combination of model, loss, and regularization:

- Model: $\hat{\mathbb{Y}} = \mathbb{X}\hat{\theta}$
- Loss: Squared loss
- Regularization: L2 regularization

The **objective function** we minimize for Ridge Regression is average squared loss, plus an added penalty:

$$\hat{\theta}_{\text{ridge}} = \arg\min_{\theta} \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2 + \lambda \sum_{j=1}^{d} \theta_i^2$$

LASSO Regression

"LASSO Regression" is a term for the following specific combination of model, loss, and regularization:

- Model: $\hat{\mathbb{Y}} = \mathbb{X}\hat{\theta}$
- Loss: Squared loss
- Regularization: L1 regularization

The **objective function** we minimize for LASSO Regression is average squared loss, plus an added penalty:

$$\hat{\theta}_{\text{LASSO}} = \arg\min_{\theta} \frac{1}{n} ||\mathbb{Y} - \mathbb{X}\theta||_2^2 + \lambda \sum_{j=1}^{d} |\theta_i|$$

Summary of Regression Methods

Name	Model	Loss	Reg.	Objective
OLS	$\hat{\mathbb{Y}} = \mathbb{X}\hat{\theta}$	Squared loss	None	$\frac{1}{n}\lVert\mathbb{Y} - \mathbb{X}\theta\rVert_2^2$
Ridge Regression	$\hat{\mathbb{Y}} = \mathbb{X}\hat{\theta}$	Squared loss	L2	$\frac{1}{n}\lVert\mathbb{Y} - \mathbb{X}\theta\rVert_2^2 + \lambda\sum_{j=1}^{d}\theta_i^2$
LASSO	$\hat{\mathbb{Y}} = \mathbb{X}\hat{\theta}$	Squared loss	L1	$\frac{1}{n}\lVert\mathbb{Y} - \mathbb{X}\theta\rVert_2^2 + \lambda\sum_{j=1}^{d}\lvert\theta_i\rvert$

Hyperparameters vs. Parameters

Parameters are facts about the world that we want to *estimate*

- Commonly denoted by p, θ, θ_i

Statistics are the *estimators* of the parameters, based on our data

- Commonly denoted by $\hat{p}, \hat{\theta}, \hat{\theta}_i$

Hyperparameters are design *choices* we make in our modeling process that affect our model, but do not directly come from the data

- examples: regularization hyperparameter, degree of polynomial
- Commonly denoted by λ, α, C

Demo