June 2025

Empirical Risk Minimization and Regularization

Julio Antonio Soto Vicente

IE University (demo class)

Outline

1. Empirical Risk Minimization

2. L_2 regularization

3. L_1 regularization

Empirical Risk Minimization

A tale of two terms

$$\min_{\mathbf{w},b} \quad \underbrace{\sum_{n=1}^{N} \ell(h_{\mathbf{w},b}(\mathbf{x}_n), y_n)}_{\text{Loss}}$$

$$\text{s.t.} \quad \underbrace{r(\mathbf{w}) \leq C}_{\text{Regularization}}$$

The regulzarization generates a **constraint** that limits learned parameter values in some way, in order to favour simpler solutions

A tale of two terms

$$\min_{\mathbf{w},b} \quad \underbrace{\sum_{n=1}^{N} \ell(h_{\mathbf{w},b}(\mathbf{x}_n), y_n)}_{\text{Loss}}$$

$$\text{s.t.} \quad \underbrace{r(\mathbf{w}) \leq C}_{\text{Regularization}}$$

The regulzarization generates a **constraint** that limits learned parameter values in some way, in order to favour simpler solutions

We can also write the same in Lagrangian form:

$$\min_{\mathbf{w},b} \quad \underbrace{\sum_{n=1}^{N} \ell(h_{\mathbf{w},b}(\mathbf{x}_n), y_n)}_{\text{Loss}} + \underbrace{\lambda r(\mathbf{w})}_{\text{Regularization}}$$

A tale of two terms

$$\min_{\mathbf{w},b} \quad \underbrace{\sum_{n=1}^{N} \ell(h_{\mathbf{w},b}(\mathbf{x}_n), y_n)}_{\text{Loss}}$$

$$\text{s.t.} \quad \underbrace{r(\mathbf{w}) \leq C}_{\text{Regularization}}$$

The regulzarization generates a **constraint** that limits learned parameter values in some way, in order to favour simpler solutions

We can also write the same in Lagrangian form:

$$\min_{\mathbf{w},b} \quad \underbrace{\sum_{n=1}^{N} \ell(h_{\mathbf{w},b}(\mathbf{x}_n), y_n)}_{\text{Loss}} + \underbrace{\lambda r(\mathbf{w})}_{\text{Regularization}} \qquad C \downarrow \quad \lambda \uparrow$$

Empirical Risk Minimization

$$\min_{\mathbf{w},b} \underbrace{\sum_{n=1}^{N} \ell(h_{\mathbf{w},b}(\mathbf{x}_n), y_n)}_{\text{Loss}} + \underbrace{\lambda r(\mathbf{w})}_{\text{Regularization}}$$

Lots of ML theory is built on top of this framework: linear models, Support Vector Machines, neural networks...

$$\min_{\mathbf{w},b} \quad \underbrace{\sum_{n=1}^{N} \ell(h_{\mathbf{w},b}(\mathbf{x}_n), y_n)}_{\text{Loss}} + \underbrace{\lambda r(\mathbf{w})}_{\text{Regularization}}$$

Lots of ML theory is built on top of this framework: linear models, Support Vector Machines, neural networks...

The regularization term can take many different forms. The most popular ones are:

- L_2 regulatization
- L_1 regularization

L_2 regularization

L_2 regularization

The regularizer forces the model to **minimize the squared L_2 norm of the weight vector** → $r(\mathbf{w}) = \|\mathbf{w}\|_2^2$:

$$\min_{\mathbf{w},b} \quad \sum_{n=1}^{N} \ell(h_{\mathbf{w},b}(\mathbf{x}_n), y_n) + \lambda \|\mathbf{w}\|_2^2$$

L_2 regularization

The regularizer forces the model to **minimize the squared L_2 norm of the weight vector** → $r(\mathbf{w}) = \|\mathbf{w}\|_2^2$:

$$\min_{\mathbf{w},b} \quad \sum_{n=1}^{N} \ell(h_{\mathbf{w},b}(\mathbf{x}_n), y_n) + \lambda \|\mathbf{w}\|_2^2$$

Where

$$\|\mathbf{w}\|_2^2 = \left(\|\mathbf{w}\|_2\right)^2 = \left(\sqrt{w_1^2 + w_2^2 + \cdots + w_d^2}\right)^2 = \sum_{i=1}^{d} w_i^2$$

L_2 regularization

The regularizer forces the model to **minimize the squared L_2 norm of the weight vector** → $r(\mathbf{w}) = \|\mathbf{w}\|_2^2$:

$$\min_{\mathbf{w},b} \quad \sum_{n=1}^{N} \ell(h_{\mathbf{w},b}(\mathbf{x}_n), y_n) + \lambda \|\mathbf{w}\|_2^2$$

Where

$$\|\mathbf{w}\|_2^2 = \left(\|\mathbf{w}\|_2\right)^2 = \left(\sqrt{w_1^2 + w_2^2 + \cdots + w_d^2}\right)^2 = \sum_{i=1}^{d} w_i^2$$

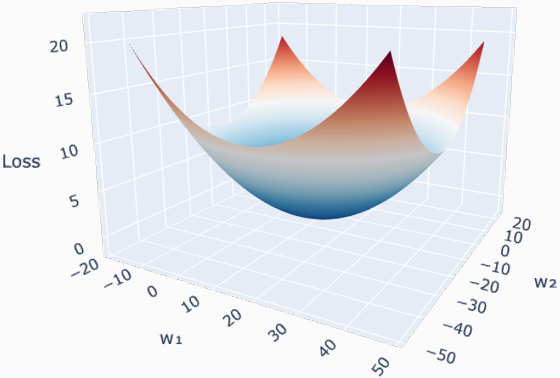
λ is the *regularization strength* → controls the tradeoff between loss and regularization

L_2 regularization

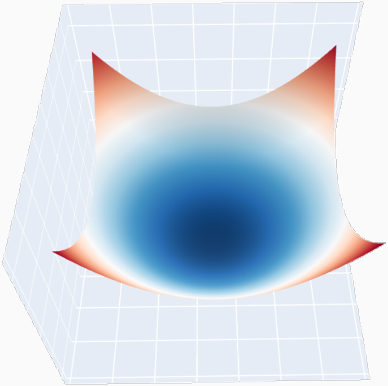
L_2 regularization has the effect of **shrinking** the estimated parameters $\mathbf{w}$ to smaller values

Smaller values for $\mathbf{w}$ generate more *conservative* predictions, potentially preventing overfitting

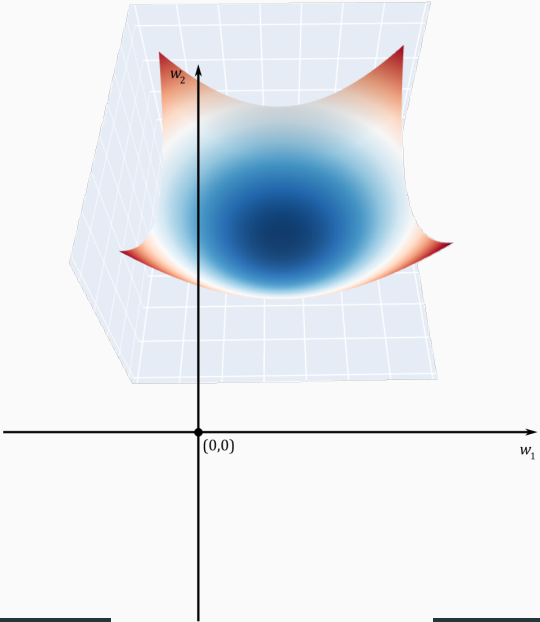
L_2 regularization



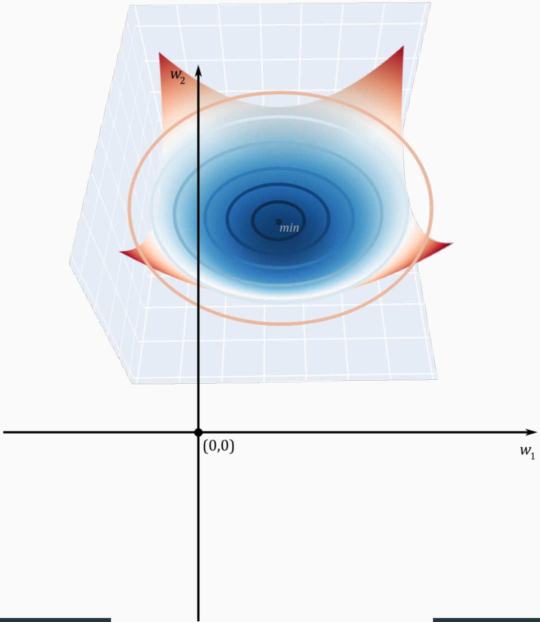
L_2 regularization



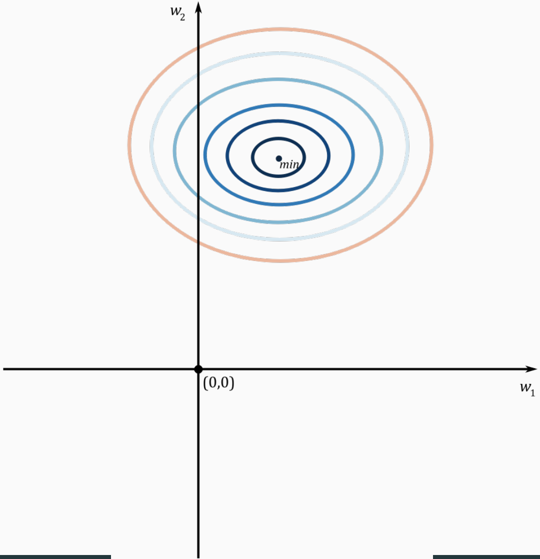
L_2 regularization



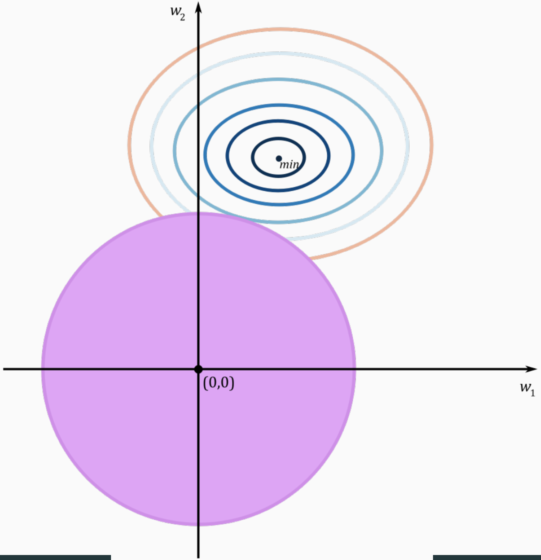
L_2 regularization



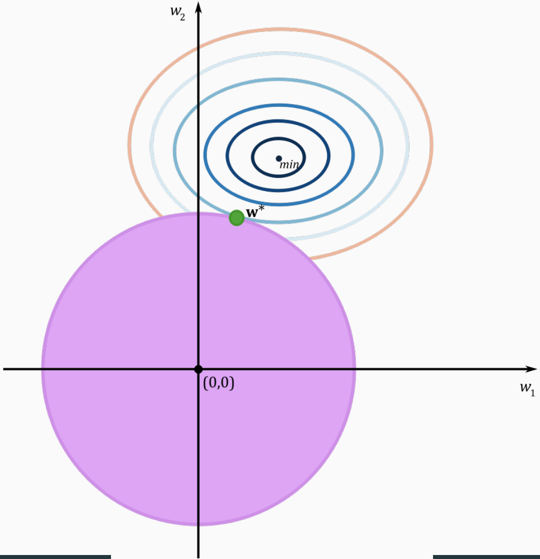
L_2 regularization



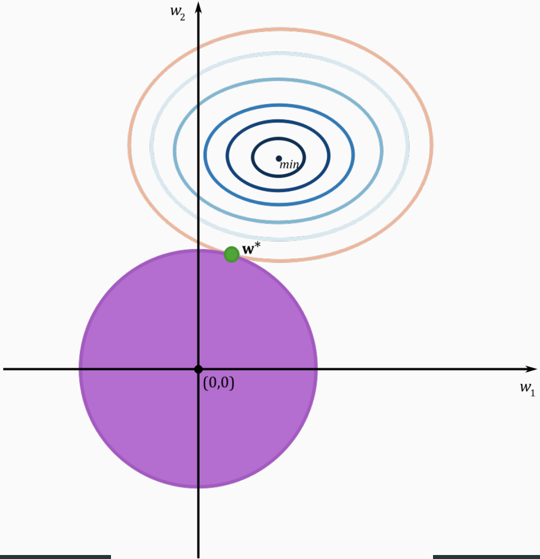
L_2 regularization



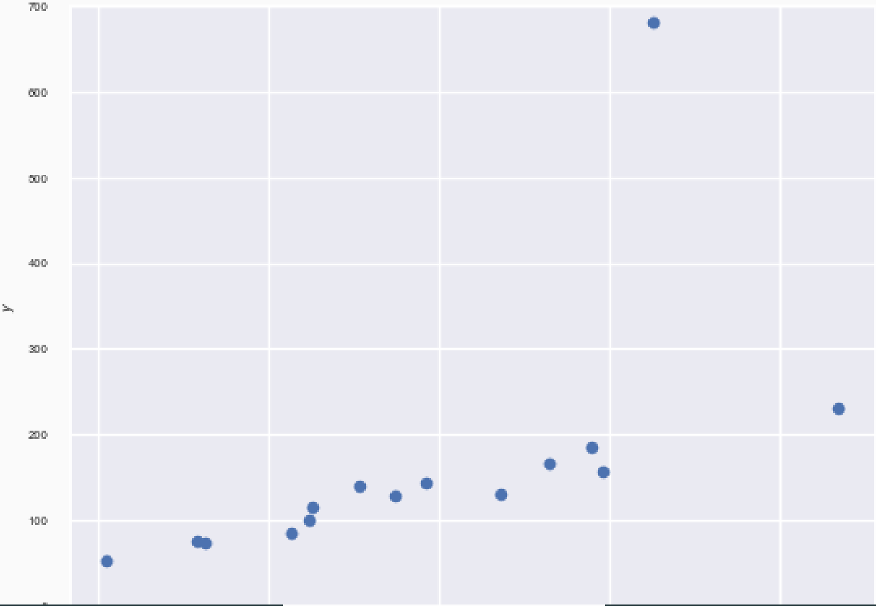
L_2 regularization



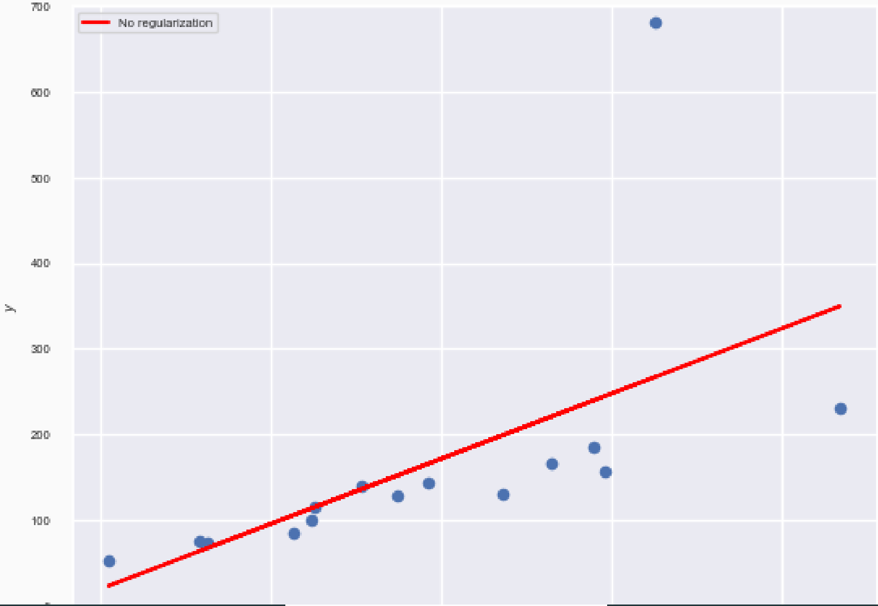
What happens if we increase λ?



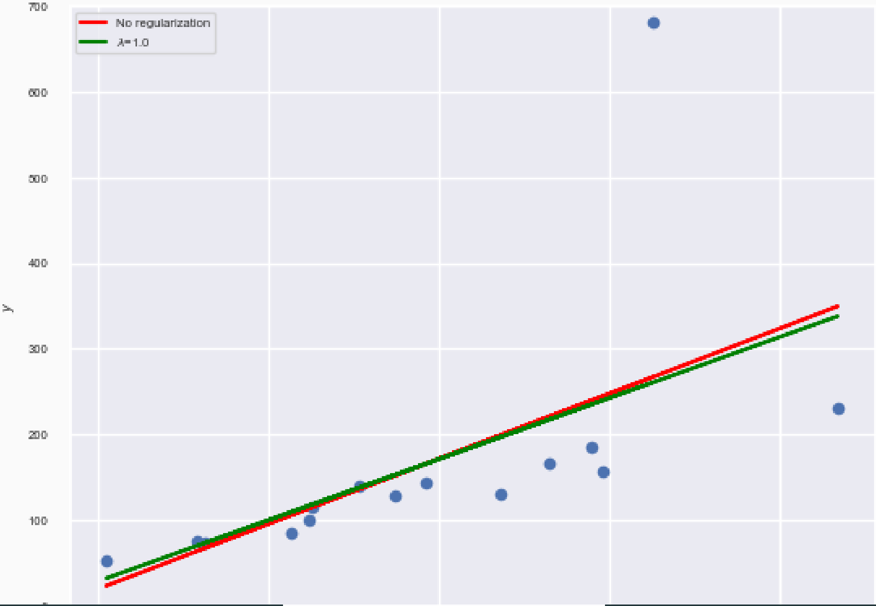
Sample dataset



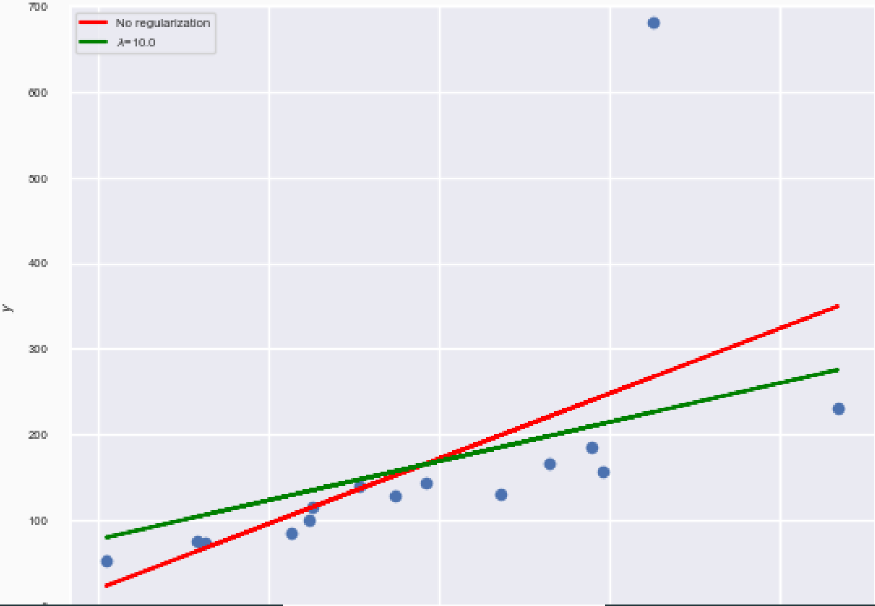
Linear regression



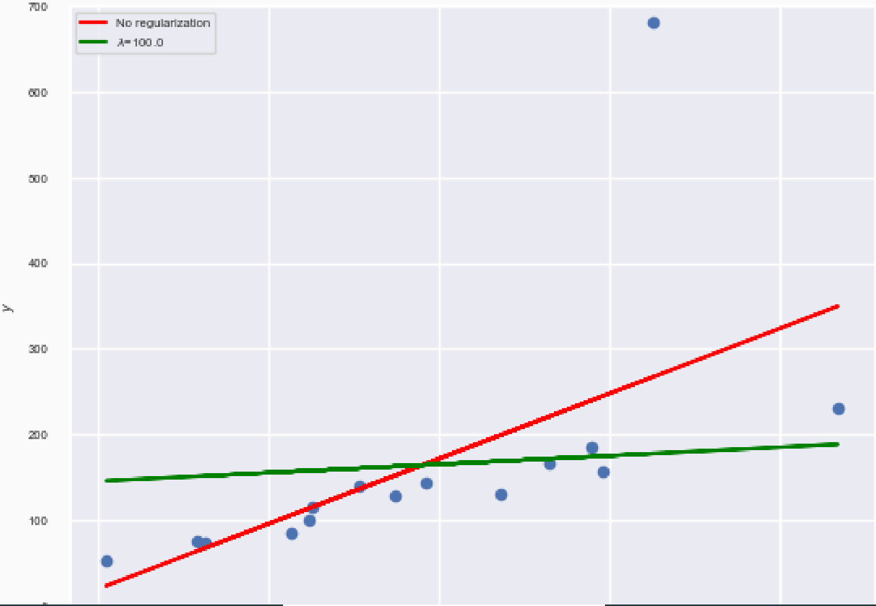
L_2-regularized linear regression



L_2-regularized linear regression



L_2-regularized linear regression



L_1 regularization

L_1 regularization

The regularizer forces the model to **minimize the L_1 norm of the weight vector** → $r(\mathbf{w}) = \|\mathbf{w}\|_1$:

$$\min_{\mathbf{w},b} \quad \sum_{n=1}^{N} \ell(h_{\mathbf{w},b}(\mathbf{x}_n), y_n) + \lambda \|\mathbf{w}\|_1$$

L_1 regularization

The regularizer forces the model to **minimize the L_1 norm of the weight vector** → $r(\mathbf{w}) = \|\mathbf{w}\|_1$:

$$\min_{\mathbf{w},b} \quad \sum_{n=1}^{N} \ell(h_{\mathbf{w},b}(\mathbf{x}_n), y_n) + \lambda \|\mathbf{w}\|_1$$

Where

$$\|\mathbf{w}\|_1 = \sum_{i=1}^{d} |w_i|$$

L_1 regularization has the effect of turning some of the parameters in $\mathbf{w}$ to exactly 0

This induces **model sparsity**, where some of the features/dimensions are completely disregarded—a w_i of 0 makes the feature it multiplies to have no effect on the final prediction

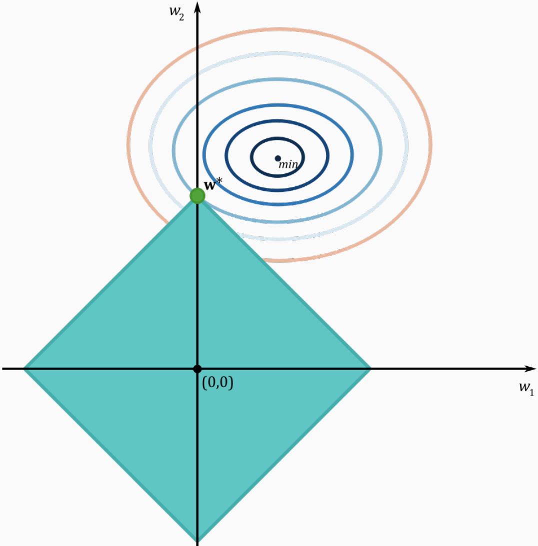
L_1 regularization

L_1 regularization has the effect of turning some of the parameters in $\mathbf{w}$ to exactly 0

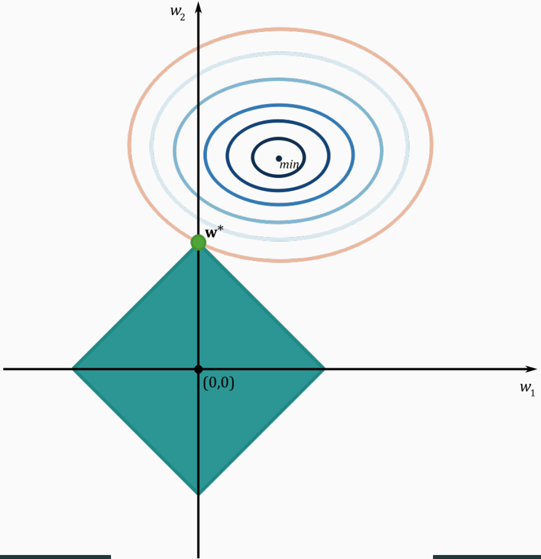
This induces **model sparsity**, where some of the features/dimensions are completely disregarded—a w_i of 0 makes the feature it multiplies to have no effect on the final prediction

Therefore, L_1 regularization performs **feature selection** implicitly

L_1 regularization



What happens if we increase λ?



L_2 and L_1 regularization

For the specific case of linear regression:

- Linear regression with L_2 regularization → *Ridge*
- Linear regression with L_1 regularization → *LASSO*
- Linear regression with both L_1 and L_2 regularization → *Elastic Net*

L_2 and L_1 regularization

For the specific case of linear regression:

- Linear regression with L_2 regularization → *Ridge*
- Linear regression with L_1 regularization → *LASSO*
- Linear regression with both L_1 and L_2 regularization → *Elastic Net*

$$\text{Elastic Net} \;=\; \min_{\mathbf{w},b} \; \frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}^\top\mathbf{x}_n + b - y_n)^2 + \lambda(\underbrace{\alpha\,\|\mathbf{w}\|_1}_{\substack{L_1\\ \text{regularization}}} + \underbrace{\left(\frac{1-\alpha}{2}\right)\|\mathbf{w}\|_2^2}_{\substack{L_2\\ \text{regularization}}})$$

$$\alpha \in [0,1)$$

Questions?