Lecture 5: Regularization Methods

ENEE 691 – Machine Learning and Photonics @ UMBC

Ergun Simsek, Ph.D. and Masoud Soroush, Ph.D.

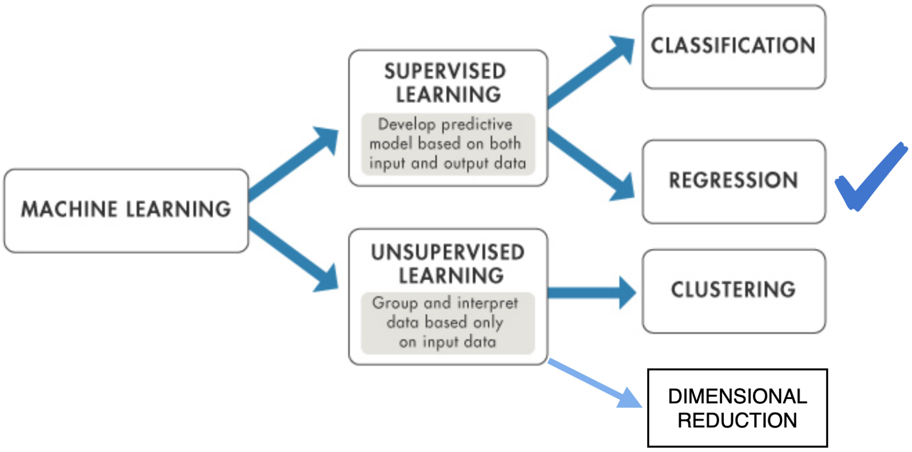
March 1, 2023

1 Regularization Methods

In today's lecture[1], we introduce regularization methods to reduce overfitting. From previous lecture, we recall that one common source of error in statistical analysis is due to variance. Regularization methods are introduced to reduce the effect of errors that are caused by variance. We introduce regularization in the context of regression analysis. However, as we will see later on, regularization methods are employed to reduce the variance in the context of classification algorithms as well.

1.1 Background

Just as the previous lecture, our discussion is in the realm of *supervised learning* with a *continuous target variable*. The ultimate goal of the model is to predict the continuous target variable.



As observed last time, there are different sources to account for errors in statistical analyses. Errors that are due to *high bias* often lead to an *underfit* of the training data. In contrast, errors caused by *high variance* often lead to an *overfit* of the training data. There is a third type of error in statistical

[1] This product is a property of UMBC, and no distribution is allowed. These notes are solely for your own use, as a registered student in this class.

analyses, called *irreducible* error. An irreducible error is caused by the noise in the data itself. The only way to reduce this type of error is to clean up the data. Today, we focus on errors due to high variance.

1.2 What Is Regularization?

To set the stage, let us review some basic facts we derived last time. Our starting point is a dataset in which the features (d of them) are denoted by $\vec{x} \in \mathbb{R}^d$, and y denotes the target variable. The dataset consists of n observations, and both $\vec{x}$ and y are continuous variables.

Observation	$\vec{x} \in \mathbb{R}^d$	y
1	$\vec{x}^{(1)}$	y_1
2	$\vec{x}^{(2)}$	y_2
3	$\vec{x}^{(3)}$	y_3
$\vdots$	$\vdots$	$\vdots$
n	$\vec{x}^{(n)}$	y_n

Associated with the above dataset, we constructed an $n \times (d+1)$ matrix, $\mathbb{X}$, that captures the observations in the above dataset:

$$\mathbb{X} = (\mathbb{1}, x_1, x_2, \cdots, x_d) = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \cdots & x_d^{(n)} \end{pmatrix} . \tag{1}$$

The linear regression model assumes a linear relation between the target and the features. We represented the linear relation in the following compact matrix form

$$\hat{\mathbb{Y}} = \mathbb{X} \cdot \omega^{\mathsf{T}} , \tag{2}$$

where matrix $\omega = (\omega_0, \omega_1, \cdots, \omega_d)$. As usual, the hatted quantities are the predictions of the model. We then defined the cost function $J(\omega)$

$$J(\omega) = \frac{1}{2}(\mathbb{Y} - \hat{\mathbb{Y}})^{\mathsf{T}} \cdot (\mathbb{Y} - \hat{\mathbb{Y}}) , \tag{3}$$

and showed that *minimizing the cost function* $J(\omega)$ in equation (3) determines the coefficients ω of the linear model as follows:

$$\omega^{\mathsf{T}} = \left(\mathbb{X}^{\mathsf{T}}\mathbb{X}\right)^{-1} \mathbb{X}^{\mathsf{T}}\mathbb{Y} . \tag{4}$$

The above linear model has $d+1$ **degrees of freedom**. If one uses polynomial regression, then one would have even more degrees of freedom. For instance, if one uses quadratic regression, one would find $\dfrac{d(d+1)}{2}$ more degrees of freedom (in addition to $d+1$ linear degrees of freedom). *The greater degrees of freedom a model possesses, the easier it will overfit the data.* Equivalently, the fewer degrees of freedom a model has, the harder it will be for it to overfit the data. Therefore, in order to avoid overfitting, one has to reduce the degrees of freedom of the model. **Regularization is a method to reduce degrees of freedom of a model to avoid overfitting the data**.

It is important to note that in regularizing a model, we *do not eliminate any of the features by hand*. We rather *constrain the model*. The model will then learn through the process of training to identify irrelevant (or less relevant) features.

1.3 Different Types of Regularization

There are several types of regularization methods. These methods have similarities and differences with one another. In here, we consider three common types of regularizations, namely the *Ridge*, *Lasso*, and *Elastic Net* regressions. In these regularization methods, a constraint on the weights ω (*i.e.* parameters of the model) is imposed. The cost function is then minimized in the presence of the introduced constraint.

Before we continue with the details of the constraints used in regularization methods, we need to distinguish between *parameters* and *hyperparameters* of machine learning models.

Parameter: In machine learning, a parameter is a factor that is introduced by the model, and its optimal value is determined through the training process once the training is complete. In other words, *parameters of a model are learned* by through the training process. For instance, coefficients of linear regression ω are parameters of the linear model, and their optimal values are determined by minimizing the cost function through the training process.

Hyperparameter: A hyperparameter is a factor whose value is set by a user before the training process starts. Unlike parameters, hyperparameters are not learned by the model, and their values are are fixed before training starts. Without specifying fixed values for hyperparameters, the training process cannot start. An important example of a hyperparameter is the *train-test split ratio*.

The constraint that is introduced to the model should possess the following important properties:

- The constraint is **independent of the training data**. This implies that the feature matrix $\mathbb{X}$ and the target matrix $\mathbb{Y}$ do not show up in the constraint. This is important as we desire a general constraint that does not depend on the details and specifics of the model.

- The constraint involves a **positive definite term** that only depends on the magnitude of the weights ω.

- The constraint puts an **upper bound on the magnitude of the parameters** of the model (in this case the coefficients of linear model ω).

- To incorporate and control the positive definite constraint in the minimization of the cost function, a hyperparameter λ is introduced to the model. The hyperparameter λ is positive semi-definite (*i.e.* $\lambda \geq 0$). This hyperparameter λ is called the **regularization factor** or the **penalty factor**.

- The case $\lambda = 0$ corresponds to the unconstrained linear model we studied last time. The greater λ is, the stronger the penalty term will be. A stronger penalty term forces the coefficients ω to be smaller in magnitude.

We now introduce three regularization methods and check that the above conditions are fulfilled by each regularization scheme.

1.3.1 Ridge Regularization

The first regularization scheme is the Ridge regularization (sometimes called Tikhonov regularization) that takes the advantage of the L_2-norm. Let us denote the weight vector $\tilde{\omega}$ to be $\tilde{\omega} = (\omega_1, \cdots, \omega_d)$. Notice that this weight vector $\tilde{\omega}$ does not include the intercept of regression ω_0 (*i.e.* the bias term). Then, in Ridge regression, we minimize the cost function $J(\omega)$ in the presence of a constraint. The constraint is that the sum of squares of components of $\tilde{\omega}$ (*i.e.* the L_2-norm of $\tilde{\omega}$) must be smaller than a positive quantity (s^2). Note that in Ridge regression, we minimize the same function (*i.e.* cost function) as in the case of ordinary regression. The difference is that in the Ridge regression, we have to satisfy a constraint (*i.e.* L_2-norm of the weight vector $\tilde{\omega}$ must be smaller than a chosen upper bound). In terms of equations, the Ridge regression is summarized as follows:

$$\boxed{\begin{aligned} &\underset{\omega}{\text{argmin}}\Big\{ J(\omega) = \frac{1}{2}(\mathbb{Y} - \hat{\mathbb{Y}})^{\mathsf{T}} \cdot (\mathbb{Y} - \hat{\mathbb{Y}}) \Big\} \\ &\text{subject to: } (\|\tilde{\omega}\|_2)^2 = \sum_{i=1}^{d} \omega_i^2 \leq s^2 \,. \end{aligned}} \tag{5}$$

Important Note: Note that only weights associated with features (*i.e.* $\omega_1, \omega_2, \cdots, \omega_d$) are regularized. In other words, the *bias term ω_0 is never regularized*. The sum in (5) starts from 1, not 0!

To incorporate the constraint in the minimization process of the cost function, we can take the advantage of a known calculus trick (known as Lagrange Multiplier Method) that allows us to get rid of the constraint at the cost of adding a term to the function that is being optimized. Instead of minimizing the cost function $J(\omega)$ in (5), we minimize $\tilde{J}(\omega)$ in the following equation

$$\tilde{J}(\omega) = \frac{1}{2}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 + \frac{\lambda}{2}\sum_{i=1}^{d}\omega_i^2 = \frac{1}{2}(\mathbb{Y} - \hat{\mathbb{Y}})^{\mathsf{T}} \cdot (\mathbb{Y} - \hat{\mathbb{Y}}) + \frac{\lambda}{2}\,\tilde{\omega}\,\tilde{\omega}^{\mathsf{T}} \,. \tag{6}$$

However, the benefit is that for minimizing $\tilde{J}(\omega)$ in (6), we do not have any constraints, and hence, we follow the usual optimization process. The coefficient λ in (6) is called the penalty factor. If we set $\lambda = 0$, $\tilde{J}(\omega)$ reduces to the cost function $J(\omega)$, and we will have the ordinary regression. The greater the penalty factor λ is, the greater the effect of regularization will be. The extreme case $\lambda \to \infty$ corresponds to the most severe penalty in which case all weights $\omega_i \to 0$ for all $i = 1, 2, \cdots, d$ (because that is the only way to keep $\tilde{J}(\omega)$ finite). In summary, the coefficient λ *controls the severity of the imposed penalty* on the cost function.

Question: How does regularization treat weakly correlated features?

Answer: Weakly correlated features to the target are typically variables whose regression coefficients are very small. However, regression algorithm uses these small coefficients to minimize the cost function to its least residual standard error. Ordinary regression has no way to distinguish weak features from noise. But once you consider a penalty term for the cost function, the regression algorithm sets the coefficients of weakly correlated features to zero. That is a benefit that is obtained by regularizing the regression.

Side Question: We saw that it was possible to solve the regression problem exactly. Is it still possible to obtain an exact solution for the Ridge regression when a penalty term is present?

Answer: The answer is yes! We can derive an exact formula analogous to the ordinary regression. We do not include the derivation in here and simply state the result. To find the exact result, we need to do two things. First, we need to place the origin of the coordinate system at the mean of the features (*i.e.* $\frac{1}{n}\sum_{i=1}^{n}\vec{x}^{(i)}$). After translating the origin of the coordinate system to the mean of the features, we define matrix $\tilde{\mathbb{X}}$ as follows:

$$\tilde{\mathbb{X}} = (x_1 - \bar{x}_1, \cdots, x_d - \bar{x}_d) = \begin{pmatrix} x_1^{(1)} - \bar{x}_1 & x_2^{(1)} - \bar{x}_2 & \cdots & x_d^{(1)} - \bar{x}_d \\ x_1^{(2)} - \bar{x}_1 & x_2^{(2)} - \bar{x}_2 & \cdots & x_d^{(2)} - \bar{x}_d \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(n)} - \bar{x}_1 & x_2^{(n)} - \bar{x}_2 & \cdots & x_d^{(n)} - \bar{x}_d \end{pmatrix} . \tag{7}$$

The exact result for the weights $\tilde{\omega}$ and the bias term ω_0 is then given by

$$\boxed{\omega_0 = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y^{(i)}\,, \quad \tilde{\omega}^{\mathsf{T}} = \left(\tilde{\mathbb{X}}^{\mathsf{T}}\tilde{\mathbb{X}} + \lambda \mathbb{1}_{d\times d}\right)^{-1}\tilde{\mathbb{X}}^{\mathsf{T}}\mathbb{Y}\,,} \tag{8}$$

where $\mathbb{1}_{d\times d}$ is the $d \times d$ identity matrix. Notice that equation (8) states that bias term ω_0 is nothing but the *mean of the target variable* y. Moreover, note the presence of the penalty factor λ in the exact result for $\tilde{\omega}$. The exact formula (8) shows that the greater λ is, the smaller $\tilde{\omega}$ will be. In other words, equation (8) shows that $\tilde{\omega} \to \vec{0}$ as $\lambda \to \infty$.

Question: Does scikit-learn use the exact formula (8) to calculate the weights of the Ridge regression?

Answer: No! For the same reason as the ordinary regression, scikit-learn library does not use the exact formula. It rather uses a gradient descent approach which is much faster for big datasets!

1.3.2 Lasso Regularization

The second regularization method that we will discuss here is LASSO (Least Absolute Shrinkage and Selection Operator) regression. In many ways, Lasso regression is similar to the Ridge regression. However, the key difference between the two approaches is that Lasso considers an L_1-norm constraint (rather than L_2-norm constraint in the case of Ridge). In terms of equations, the Lasso regression is defined by

$$\boxed{\begin{aligned} &\underset{\omega}{\operatorname{argmin}}\left\{J(\omega) = \frac{1}{2}(\mathbb{Y} - \hat{\mathbb{Y}})^{\mathsf{T}} \cdot (\mathbb{Y} - \hat{\mathbb{Y}})\right\} \\ &\text{subject to: } \|\tilde{\omega}\|_1 = \sum_{i=1}^{d} |\omega_i| \leq s\,. \end{aligned}} \tag{9}$$

Note that the L_1-norm concerns the absolute value of the weights ω_i. As in the Ridge regression, the bias term ω_0 is never regularized. As is clear from equations (5) and (9), the same cost function is shared by Ridge and Lasso regressions. The difference between the two concerns the imposed constraints. As in the Ridge case, we can employ the Lagrange multiplier method to facilitate the minimization of the cost function by adding an appropriate penalty term.

$$\tilde{J}(\omega) = \frac{1}{2}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 + \frac{\lambda}{2}\sum_{i=1}^{d}|\omega_i| \, . \tag{10}$$

Again, setting $\lambda = 0$ reduces the problem to the ordinary regression. As in the Ridge regression, the greater the penalty factor λ is, the smaller the weights $\tilde{\omega}$ will be.

1.3.3 Elastic Net Regularization

The third regularization method is the Elastic Net regularization which is a mixture of both the Ridge and the Lasso regressions. In Elastic Net regularization, two separate constraints are imposed on the minimization of the cost function $J(\omega)$. One constraint considers the L_1-norm and the other considers the L_2-norm.

$$\boxed{\begin{array}{l} \underset{\omega}{\text{argmin}}\Big\{ J(\omega) = \frac{1}{2}(\mathbb{Y}-\hat{\mathbb{Y}})^{\mathsf{T}} \cdot (\mathbb{Y}-\hat{\mathbb{Y}}) \Big\} \\ \text{subject to: } \begin{cases} \|\tilde{\omega}\|_1 = \sum_{i=1}^{d} |\omega_i| \leq s_1 \\ (\|\tilde{\omega}\|_2)^2 = \sum_{i=1}^{d} \omega_i^2 \leq s_2 \end{cases} \end{array}} \tag{11}$$

Employing the same Lagrange multiplier technique, we get rid of the constraints at the cost of adding two penalty terms to the cost function (one for each constraint)

$$\tilde{J}(\omega) = \frac{1}{2}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 + \frac{\lambda_1}{2}\sum_{i=1}^{d}\omega_i^2 + \frac{\lambda_2}{2}\sum_{i=1}^{d}|\omega_i| \, . \tag{12}$$

The Elastic Net regression possesses two independent penalty factors, λ_1 and λ_2. Setting $\lambda_1 = 0$ reduces the Elastic Net regression to the Lasso regression, and setting $\lambda_2 = 0$, reduces the Elastic Net regression to the Ridge case. Moreover, setting $\lambda_1 = 0$ and $\lambda_2 = 0$ reduces the problem to the ordinary regression. As in previous regularization methods, the greater λ_1 and λ_2 are, the smaller the weights $\tilde{\omega}$ will be.

1.4 Regularization Methods Comparison

Now that we have introduced three different regularization methods, it would be appropriate to compare them with one another and identify their differences.

First, let us compare the Ridge and Lasso regressions. As mentioned above, introducing a penalty term (Ridge or Lasso) puts an upper bound on the magnitude of the weights $\tilde{\omega}$, and the greater the penalty factor λ is, the smaller the weights $\tilde{\omega}$ will be. Despite this similarity between the Ridge and Lasso regressions, the two regularization methods have an important difference:

Difference between Ridge and Lasso Regularizations: Although the Ridge regression shrinks all the weights toward smaller values (*i.e.* close to zero) as the penalty factor λ increases, it will not set any of the weights ω_i exactly to zero (unless you set $\lambda \to \infty$). This may not be an issue for the accuracy of the model, but it can pose a challenge in interpreting the model in which the number of features is quite large. In contrary, the Lasso regularization has the effect of forcing some of the weights ω_i to be exactly zero when the penalty factor λ is sufficiently large. In other words, *Lasso regularization can be used as a feature selection criterion*, and hence, it offers a simpler model to interpret!

Question: How does the Lasso regularization set some of the weights *exactly* to zero? And why can't the Ridge regularization do the same thing?

To answer this question, and in order to keep the analysis simple, we consider a special case. Suppose we have a model that possesses only two features x_1 and x_2, and a target variable y. In this case, $\tilde{\omega} = (\omega_1, \omega_2)$, and the cost function $J(\omega)$ is given by

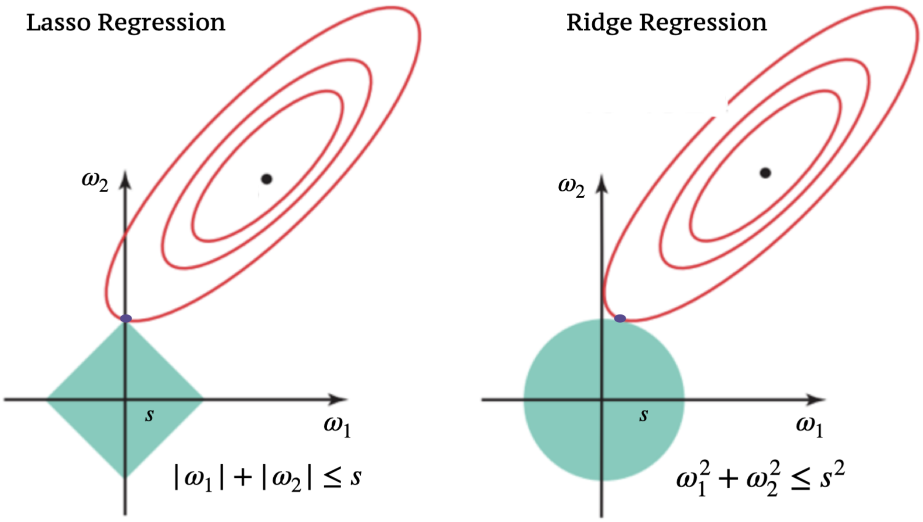
$$\begin{aligned} J(\omega_0, \omega_1, \omega_2) &= \frac{1}{2}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 = \frac{1}{2}\sum_{i=1}^{n}(y^{(i)} - \omega_1 x_1^{(i)} - \omega_2 x_2^{(i)} - \omega_0)^2 \\ &= A\,\omega_1^2 + B\,\omega_2^2 + C\,\omega_1\omega_2 + D\,\omega_1 + E\,\omega_2 + F\,, \end{aligned} \tag{13}$$

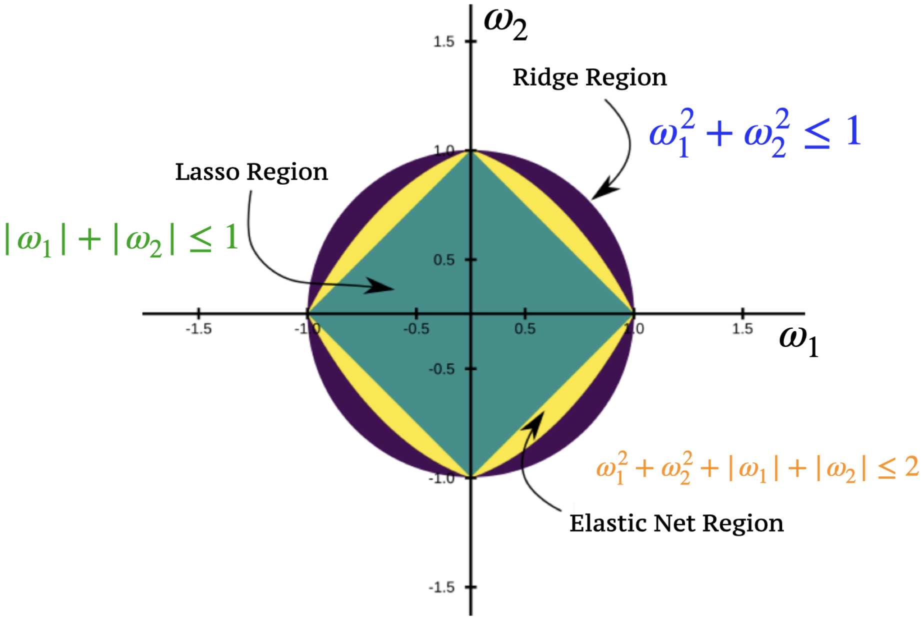
where the constant coefficients A, B, C, D, E, and F can be easily expressed in terms of the values of the tabular dataset and ω_0 (*e.g.* $A = \frac{1}{2}\sum_{i=1}^{n}(x_1^{(i)})^2$). Now, consider a two-dimensional space ($\mathbb{R}^2$) with coordinate system (ω_1, ω_2). Setting the cost function (13) to a constant, we obtain the equation of an ellipse in (ω_1, ω_2)-space. Changing this constant value, we find a family of **(cocentric) ellipses**, one ellipse associated with each constant (In calculus, the curves obtained in this manner are called the *contour plots*). Now, in order to apply the Ridge and Lasso regularizations, we need to apply an appropriate constraint for each case in the (ω_1, ω_2)-space. For the Ridge case, we have to apply $\omega_1^2 + \omega_2^2 \leq s^2$, and for the Lasso case, $|\omega_1| + |\omega_2| \leq s$. These two regions correspond to a disc and the interior of a rhombus, respectively.

To minimize the cost function in (13), we need to intersect the contour plots (*i.e.* the red ellipses) with the regions corresponding to the constraints. It is evident from the above picture that the ellipses intersect the disc (in the Ridge case) at a point for which ω_1 is very small, but not exactly zero. On the other hand, the ellipses in the Lasso case intersect one of the vertices of the rhombus for which ω_1 is exactly zero!

Moral of the Story: In applying a regularization method, if the corresponding constraint region possesses **sharp corners on the coordinate axes** (like the rhombus in Lasso case), the regularization method is more likely to set some of the weights to *exactly* zero! In other words, the sharper the corners of the constraint region are, the greater likelihood for setting features to exactly zero value will be. This conclusion generalizes to more number of features (*i.e.* $d > 2$) straightforwardly.

Now, you may wonder what can be said about Elastic Net regularization. In order to see what Elastic Net does, we plot the its constraint region in the context of the above setup (*i.e.* with only two features). The below picture depicts the constraint regions of all three regularizations together with their equations.


Lasso Regression
ω_2
s
ω_1
$|\omega_1| + |\omega_2| \leq s$
Ridge Regression
ω_2
s
ω_1
$\omega_1^2 + \omega_2^2 \leq s^2$


ω_2
1.5
Ridge Region
$\omega_1^2 + \omega_2^2 \leq 1$
1.0
Lasso Region
$|\omega_1| + |\omega_2| \leq 1$
0.5
-1.5
-1.0
-0.5
0.5
1.0
1.5
ω_1
-0.5
$\omega_1^2 + \omega_2^2 + |\omega_1| + |\omega_2| \leq 2$
-1.0
Elastic Net Region
-1.5

As is clear from the above picture, the constraint region associated with the Elastic Net is a region in between of the Ridge and the Lasso. Elastic Net region does possess four sharp corners, but the corners are not as sharp as the corners of the Lasso region. Therefore, we expect the Elastic Net regularization to operate similar to the Lasso when the penalty factor of the L_1-term is sufficiently large.

Note: Although we studied regularization techniques in the context of regression analysis, they are applied to classification problems as well. The philosophy for adding regularization terms in classification problems is completely analogous to what we saw in this lecture. In future lectures, we will use regularization techniques in the context of classification problems as well.