MLE & Regularization

Maximum Likelihood Estimation (MLE) -

- MLE is a probabilistic approach to find the optimal parameters of a model.
- Earlier we used Least Squares to solve Linear Regression. MLE is an alternate way of doing the same.
- It provides a way to find the “best-fitting” parameters that maximize the likelihood of observing the given data.

How to use MLE in Linear Regression?

$$\hat{y}_i = x_i w$$

We can define the error for the i^{th} data point as -

$$e = y_i - \hat{y}_i$$
$$e = y_i - x_i w$$

Assumption - The error is normally distributed with mean 0 and constant variance (σ^2)

So the error is -

$$N(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
$$e = P(x_i, y_i; w, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x_i w)^2}{2\sigma^2}\right)$$

Now we can define our likelihood function as -

$$L(x, y; w, \sigma) = P(x_1 x_2 x_3 ... x_n | w, \sigma)$$

How can we simplify this equation?

We assume that the data is **independently and identically distributed (i.i.d assumptions).**

$$
\begin{aligned}
L(x, y; w, \sigma) &= \prod_{i=1}^{N} P(x_i | w, \sigma) \\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x_i w)^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\sum_{i=1}^{N} \frac{(y_i - x_i w)^2}{2\sigma^2}\right)
\end{aligned}
$$

Our goal is to maximize this likelihood function, or find the optimal parameters w and σ.

To make the math easier, we will instead maximize the **log likelihood -**

$$
log(L(x, y; w, \sigma)) = -\frac{N}{2}(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (y_i - x_i w)^2
$$

We can do this because logarithm is a monotonically increasing function.

Now instead of maximizing the above, we can minimize this equation -

$$
-log(L(x, y; w, \sigma)) = \frac{N}{2}(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum (y_i - x_i w)^2
$$

To get the optimal parameters w and σ, differentiate the negative log likelihood w.r.t. w and equate it to 0. Do the same for σ.

Or use gradient descent.

- Does the result match to that of Least Square's method?
- How are MLE and Least Squares equivalent?

Regularization -

How to address overfitting?

1. Cross-Validation
2. Reduce number of features

   — Manually select which features to keep.

— Feature reduction algorithms (later in course).

3. Regularization

— Reduce the magnitude/values of model parameters.

— Forces the model to be simpler (or reduces complexity).

Why are large parameters bad?

Table 1.1 Table of the coefficients $\mathbf{w}^\star$ for polynomials of various order. Observe how the typical magnitude of the coefficients increases dramatically as the order of the polynomial increases.

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

- Numerical issues like overflowing.
- How does a small change in x affect $1075461.69 * x^5$ as opposed to $4.5 * x^5$?
- Very large weights usually indicates that the model tried to fit the data points exactly.
- What happens to variance? And bias?

Regularization -

$$L(w) = \sum_{i=1}^{N} (y_i - x_i w)^2 + \lambda R(w)$$

Here $R(w)$ is the *penalty term* and λ is the *regularization parameter*.

- What will be the affect of adding weights to the loss function?
- Hyperparameter vs parameter?
- How does the value of λ affect the model parameters?

- How will this affect the variance and bias?

L1 (Lasso) Regression - v

$$L_1(w) = \sum_{i=1}^{N}(y_i - x_i w)^2 + \lambda \sum_{i=1}^{N} |w|$$

- Penalty is the sum of the absolute values of the regression coefficients.
- It penalizes the model for having large coefficients.
- Performs **feature selection** by reducing some weights to zero. So if you have a lot of features, and you suspect that not all of them are important, try applying Lasso.
- λ penalty is same for all the weights, so it is necessary to **standardize** the data first. This means that Lasso (and Ridge) regression is NOT scale-invariant.

L2 (Ridge Regression) -

$$L_2(w) = \sum_{i=1}^{N}(y_i - x_i w)^2 + \lambda \sum_{i=1}^{N} w^2$$

- Penalty is the sum of squared values of the coefficients.
- Tends to keep all features in the model but with smaller weights.
- If you have only few features and you think all of them are important, try applying Ridge.

▼ Closed form solution to Ridge Regression -

$$\begin{aligned} L_2(w) &= \frac{1}{2}(y - Xw)^T(y - Xw) + \frac{\lambda}{2} w^T w \\ L_2'(w) &= X^T X w - X^T y + \lambda w \\ w_{opt} &= (X^T X + \lambda . I)^{-1} X^T y \end{aligned}$$

Geometric visualization -

We can rewrite the equation for Ridge regression like this -

$$L_2(w) = (y_i - x_i w)^2 \quad s.t. \quad |w|^2 \leq c$$

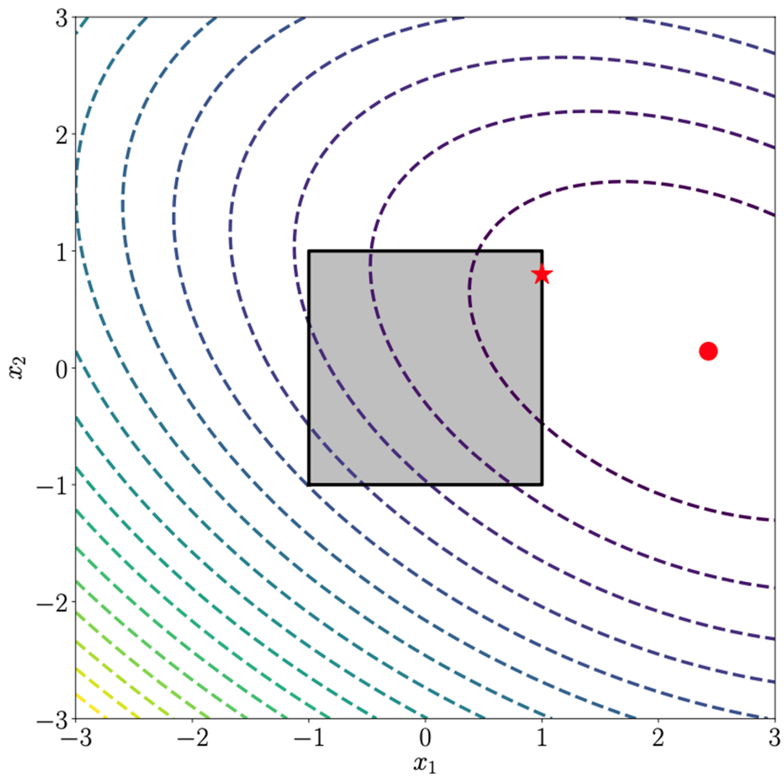
If we have only 2 parameters then the constraint would be -

$$|w_0|^2 + |w_1|^2 \leq c$$

Similarly for lasso regression -

$$|w_1| + |w_2| \leq c$$

What is constrained optimization?



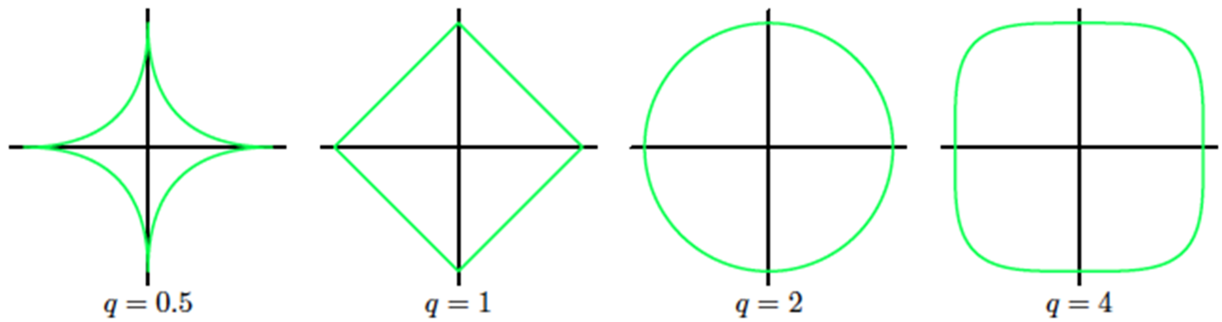
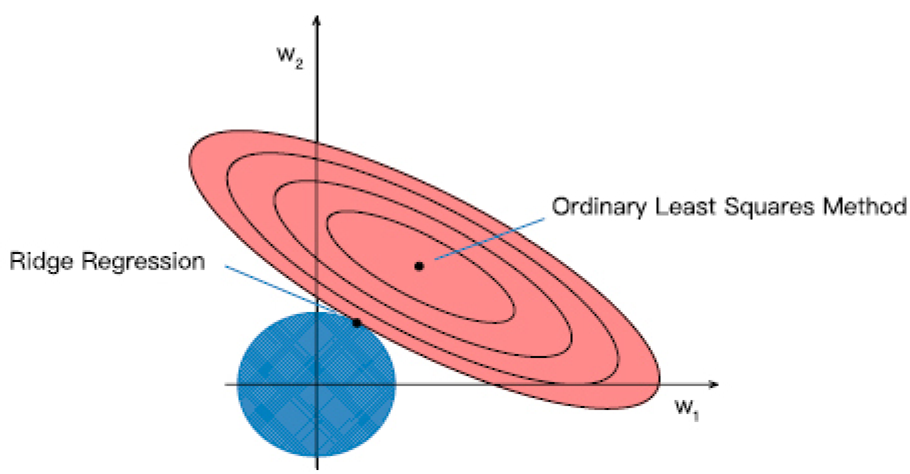
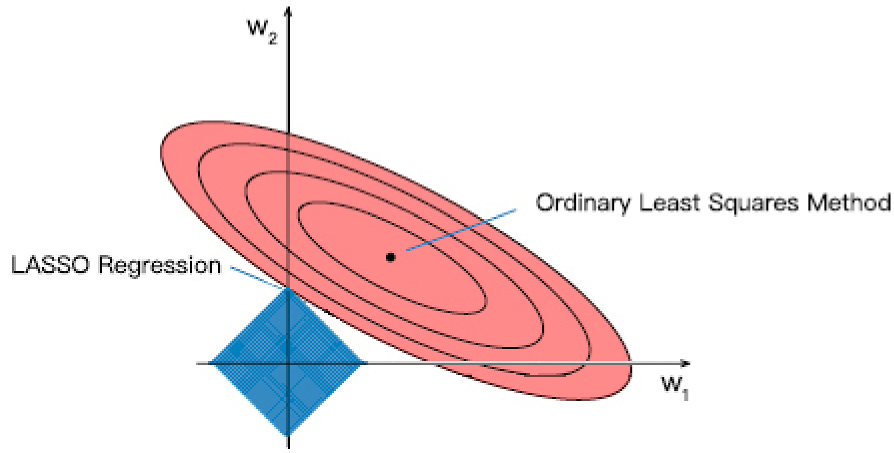


Figure 3.3 Contours of the regularization term in (3.29) for various values of the parameter q.

The optimal solution lies where the OLS contour is tangent to the constraint curve.




w2
Ordinary Least Squares Method
LASSO Regression
w1