1 Work with Norms

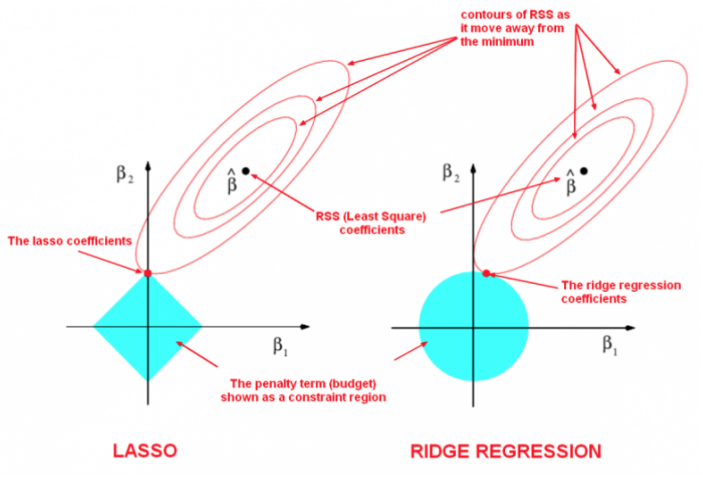


Figure 1: A depiction of ℓ_1 and ℓ_2 norms in penalized regression.

In lecture, we discussed various types of norms, and their use in regularizing our models. Recall for a vector $\mathbf{w} \in \mathbb{R}^K$:

$$\ell_0 \text{ norm: } \|\mathbf{w}\|_0 = \text{the number of non-zero elements in the vector}$$
$$\ell_1 \text{ norm: } \|\mathbf{w}\|_1 = \sum_{i=1}^{K} |w_i|$$
$$\ell_2 \text{ norm: } \|\mathbf{w}\|_2 = \sqrt{\sum_{i=1}^{K} (w_i)^2}$$

(a) Suppose we have a vector $\mathbf{w} \in \mathbb{R}^8$. We know that $\|\mathbf{w}\|_0 = 5$, $\|\mathbf{w}\|_1 = 22$, $\|\mathbf{w}\|_2 = 10$

(b) Now suppose we have a vector $\mathbf{w} \in \mathbb{R}^8$. We know that $\|\mathbf{w}\|_0 = 2$, $\|\mathbf{w}\|_1 = 17$, $\|\mathbf{w}\|_2 = 13$. Give a vector $\mathbf{w}$ satisfying these conditions.

2 MLE and MAP

To estimate probabilities from data, we can use maximum likelihood estimation (MLE) or maximum a posteriori (MAP).

In MLE, you choose the parameter θ that maximizes the likelihood of observed data, $\text{argmax}_\theta\, p(\mathcal{D}|\theta)$. In MAP, you choose the parameter θ that maximizes the likelihood of posterior probability, $\text{argmax}_\theta\, p(\theta|\mathcal{D})$.

Using the coin flipping example in class, we'll solve for θ with both approaches, where θ is the probability of flipping heads. Assume that the coin was flipped 10 times, giving 8 heads and 2 tails. Each flip is independent and identically distributed according to the Bernoulli distribution.

MLE

a. Formulate the likelihood function $\mathcal{L}(\theta)$ (represented by $p(\mathcal{D}|\theta)$)

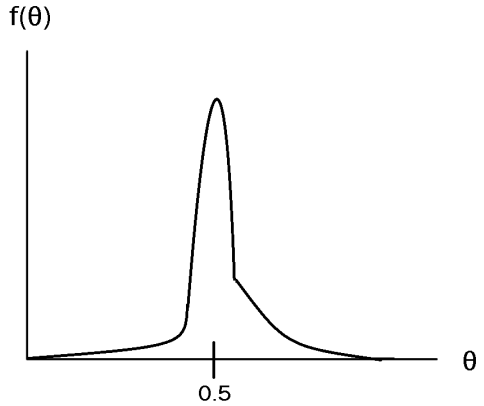
b. Take the log to obtain the log-likelihood function $\ell(\theta)$

c. Solve for θ by taking the derivative of log-likelihood function w.r.t θ and setting it to 0.

MAP

Now let's consider MAP instead. Recall that we're trying to maximize the posterior probability, which is proportional to likelihood * prior. In mathematical form, $p(\theta|\mathcal{D}) \propto \prod p(\mathcal{D}^{(n)}|\theta)p(\theta)$.

The prior probability distribution given in class is as follows:



where f(θ=0.25) = 0.1, f(θ=0.5)=0.6, and f(θ=0.75) = 0.2.

2.1 MAP estimate with 2 samples

Let us consider only the first 2 samples, which we learn are both heads. Formally solving for θ is similar to the process for MLE, except the prior probability is also introduced. For simplicity of this exercise, just find and compare the posterior probabilities corresponding to $\theta = 0.25$, 0.5 and 0.75. Which θ gives you the highest probability?

2.2 MAP estimate with 10 samples

Now consider all 10 samples, and find the corresponding posterior probabilities for the three θ values again. Which θ gives you the highest probability?

2.3 Effect of number of samples on MAP estimate

How do the θ values from parts a and b compare? In which case does prior probability play a bigger role, and why?