1 MLE and MAP for Regression (Part I)

So far, we've explored two approaches of the regression framework, Ordinary Least Squares and Ridge Regression:

$$\hat{\mathbf{w}}_{\text{OLS}} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$
$$\hat{\mathbf{w}}_{\text{RIDGE}} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

One question that arises is why we specifically use the ℓ^2 norm to measure the error of our predictions, and to penalize the model parameters. We will justify this design choice by exploring the statistical interpretations of regression — namely, we will employ Gaussians, MLE and MAP to validate what we've done so far through a different lens.

1.1 Probabilistic Model

In the context of **supervised learning**, we assume that there exists a **true underlying model** mapping inputs to outputs:

$$f : \mathbf{x} \to f(\mathbf{x})$$

The true model is unknown to us, and our goal is to find a **hypothesis model** that best represents the true model. The only information that we have about the true model is via a dataset

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$$

where $\mathbf{x}_i \in \mathbb{R}^d$ is the input and $y_i \in \mathbb{R}$ is the observation, a noisy version of the true output $f(\mathbf{x}_i)$:

$$Y_i = f(\mathbf{x}_i) + Z_i$$

We assume that $\mathbf{x}_i$ is a fixed value (which implies that $f(\mathbf{x}_i)$ is fixed as well), while Z_i is a random variable (which implies that Y_i is a random variable as well). We always assume that Z_i has zero mean, because otherwise there would be systematic bias in our observations. The Z_i's could be Gaussian, uniform, Laplacian, etc... In most contexts, we us assume that they are **independent identically distributed (i.i.d)** Gaussians: $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. We can therefore say that Y_i is a random variable whose probability distribution is given by

$$Y_i \overset{\text{iid}}{\sim} \mathcal{N}(f(\mathbf{x}_i), \sigma^2)$$

Now that we have defined the model and data, we wish to find a hypothesis model $h_{\boldsymbol{\theta}}$ (parameterized by $\boldsymbol{\theta}$) that best captures the relationships in the data, while possibly taking into account prior beliefs that we have about the true model. We can represent this as a probability problem, where the goal is to find the optimal model that maximizes our probability.

1.2 Maximum Likelihood Estimation

In **Maximum Likelihood Estimation** (MLE), the goal is to find the hypothesis model that maximizes the probability of the data. If we parameterize the set of hypothesis models with $\boldsymbol{\theta}$, we can express the problem as

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = p(\text{data} = \mathcal{D} \mid \text{ true model} = h_{\boldsymbol{\theta}})$$

The quantity $\mathcal{L}(\boldsymbol{\theta})$ that we are maximizing is also known as the **likelihood**, hence the term MLE. Substituting our representation of $\mathcal{D}$ we have

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathbf{X}, \mathbf{y}) = p(y_1, \ldots, y_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\theta})$$

Note that we implicitly condition on the $\mathbf{x}_i$'s, because we treat them as *fixed* values of the data. The only randomness in our data comes from the y_i's (since they are noisy versions of the true values $f(\mathbf{x}_i)$). We can further simplify the problem by working with the **log likelihood** $\ell(\boldsymbol{\theta}; \mathbf{X}, \mathbf{y}) = \log \mathcal{L}(\boldsymbol{\theta}; \mathbf{X}, \mathbf{y})$

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathbf{X}, \mathbf{y}) = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{X}, \mathbf{y})$$

With logs we are still working with the same problem, because logarithms are monotonic functions. In other words we have that:

$$P(A) < P(B) \iff \log P(A) < \log P(B)$$

Let's decompose the log likelihood:

$$\ell(\boldsymbol{\theta}; \mathbf{X}, \mathbf{y}) = \log p(y_1, \ldots, y_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\theta}) = \log \prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log[p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})]$$

We decoupled the probabilities from each datapoints because their corresponding noise components are independent. Note that the logs allow us to work with sums rather products, simplifying the problem — one reason why the log likelihood is such a powerful tool. Each individual term $p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})$ comes from a Gaussian

$$Y_i \mid \boldsymbol{\theta} \sim \mathcal{N}(h_{\boldsymbol{\theta}}(\mathbf{x}_i), \sigma^2)$$

Continuing with logs:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{\text{MLE}} &= \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{X}, \mathbf{y}) & (1)\\
&= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log[p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})] & (2)\\
&= \arg\max_{\boldsymbol{\theta}} -\left(\sum_{i=1}^{n} \frac{(y_i - h_{\boldsymbol{\theta}}(\mathbf{x}_i))^2}{2\sigma^2}\right) - n \log \sqrt{2\pi}\sigma & (3)\\
&= \arg\min_{\boldsymbol{\theta}} \left(\sum_{i=1}^{n} \frac{(y_i - h_{\boldsymbol{\theta}}(\mathbf{x}_i))^2}{2\sigma^2}\right) + n \log \sqrt{2\pi}\sigma & (4)
\end{aligned}$$

$$= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - h_{\boldsymbol{\theta}}(\mathbf{x}_i))^2 \tag{5}$$

Note that in step (4) we turned the problem from a maximization problem to a minimization problem by negating the objective. In step (5) we eliminated the second term and the denominator in the first term, because they do not depend on the variables we are trying to optimize over.

Now let's look at the case of regression — our hypothesis has the form $h_{\boldsymbol{\theta}}(\mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \mathbb{R}^d$, where d is the number of dimensions of our featurized datapoints. For this specific setting, the problem becomes:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\theta})^2$$

This is just the Ordinary Least Squares (OLS) problem! We just proved that OLS and MLE for regression lead to the same answer! We conclude that MLE is a probabilistic justification for why using squared error (which is the basis of OLS) is a good metric for evaluating a regression model.

1.3 Maximum a Posteriori

In **Maximum a Posteriori** (MAP) Estimation, the goal is to find the model, for which the data maximizes the probability of the model:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \; p(\text{true model} = h_{\boldsymbol{\theta}} \mid \text{data} = \mathcal{D})$$

The probability distribution that we are maximizing is known as the **posterior**. Maximizing this term directly is often infeasible, so we we use **Bayes' Rule** to re-express the objective.

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{\text{MAP}} &= \arg\max_{\boldsymbol{\theta}} \; p(\text{true model} = h_{\boldsymbol{\theta}} \mid \text{data} = \mathcal{D}) \\
&= \arg\max_{\boldsymbol{\theta}} \; \frac{p(\text{data} = \mathcal{D} \mid \text{true model} = h_{\boldsymbol{\theta}}) \cdot p(\text{true model} = h_{\boldsymbol{\theta}})}{p(\text{data} = \mathcal{D})} \\
&= \arg\max_{\boldsymbol{\theta}} \; p(\text{data} = \mathcal{D} \mid \text{true model} = h_{\boldsymbol{\theta}}) \cdot p(\text{true model} = h_{\boldsymbol{\theta}}) \\
&= \arg\max_{\boldsymbol{\theta}} \; \log p(\text{data} = \mathcal{D} \mid \text{true model} = h_{\boldsymbol{\theta}}) + \log p(\text{true model} = h_{\boldsymbol{\theta}}) \\
&= \arg\min_{\boldsymbol{\theta}} \; -\log p(\text{data} = \mathcal{D} \mid \text{true model} = h_{\boldsymbol{\theta}}) - \log p(\text{true model} = h_{\boldsymbol{\theta}})
\end{aligned}$$

We treat $p(\text{data} = \mathcal{D})$ as a constant value because it does not depend on the variables we are optimizing over. Notice that MAP is just like MLE, except we add a term $p(\text{true model} = h_{\boldsymbol{\theta}})$ to our objective. This term is the **prior** over our true model. Adding the prior has the effect of favoring certain models over others *a priori*, regardless of the dataset. Note the MLE is a special case of MAP, when the prior does not treat any model more favorably over other models. Concretely, we have that

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} -\left(\sum_{i=1}^{n} \log[p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})]\right) - \log[p(\boldsymbol{\theta})]$$

Again, just as in MLE, notice that we implicitly condition on the $\mathbf{x}_i$'s because we treat them as constants. Also, let us assume as before that the noise terms are i.i.d. Gaussians: $N_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. For the prior term $P(\mathbf{\Theta})$, we assume that the components θ_j are i.i.d. Gaussians:

$$\theta_j \overset{\text{iid}}{\sim} \mathcal{N}(\theta_{j_0}, \sigma_h^2)$$

Using this specific information, we now have:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\text{MAP}} &= \arg\min_{\boldsymbol{\theta}} \left(\frac{\sum_{i=1}^n (y_i - h_{\boldsymbol{\theta}}(\mathbf{x}_i))^2}{2\sigma^2} \right) + \left(\frac{\sum_{j=1}^d (\theta_j - \theta_{j_0})^2}{2\sigma_h^2} \right) \\ &= \arg\min_{\boldsymbol{\theta}} \left(\sum_{i=1}^n (y_i - h_{\boldsymbol{\theta}}(\mathbf{x}_i))^2 \right) + \frac{\sigma^2}{\sigma_h^2} \left(\sum_{j=1}^d (\theta_j - \theta_{j_0})^2 \right)\end{aligned}$$

Let's look again at the case for linear regression to illustrate the effect of the prior term when $\theta_{j_0} = 0$. In this context, we refer to the linear hypothesis function $h_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x}$.

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\theta})^2 + \frac{\sigma^2}{\sigma_h^2} \sum_{j=1}^d \theta_j^2$$

This is just the Ridge Regression problem! We just proved that Ridge Regression and MAP for regression lead to the same answer! We can simply set $\lambda = \frac{\sigma^2}{\sigma_h^2}$. We conclude that MAP is a probabilistic justification for adding the penalized ridge term in Ridge Regression.

1.4 MLE vs. MAP

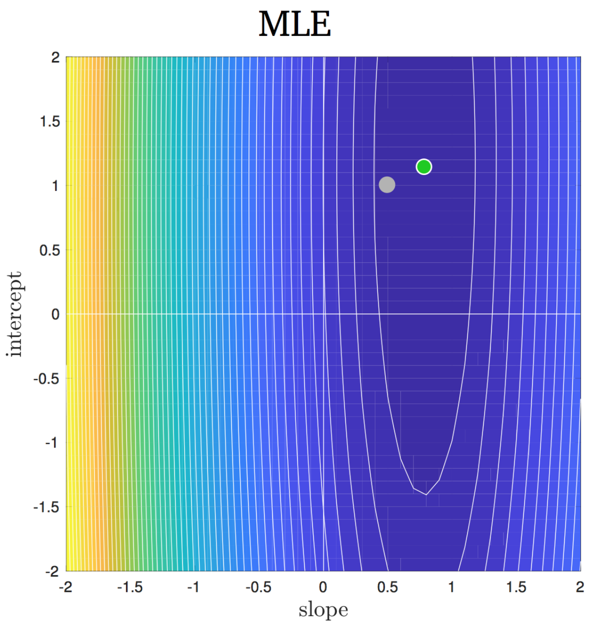
Based on our analysis of Ordinary Least Squares Regression and Ridge Regression, we should expect to see MAP perform better than MLE. But is that always the case? Let us visit a simple 2D problem where

$$f(x) = \text{slope} \cdot x + \text{intercept}$$

Suppose we already know the true underlying model parameters:

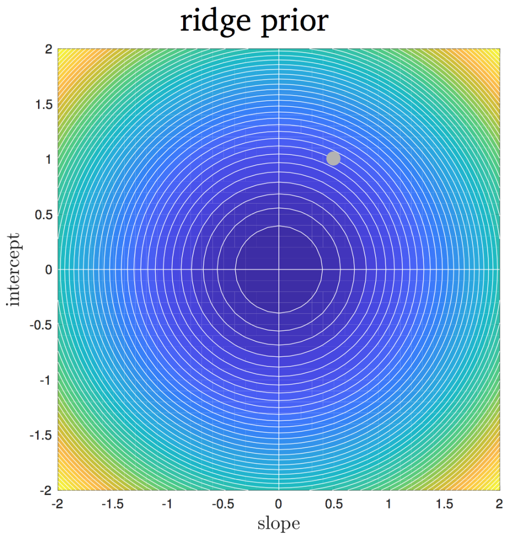
$$(\text{slope}^*, \text{intercept}^*) = (0.5, 1.0)$$

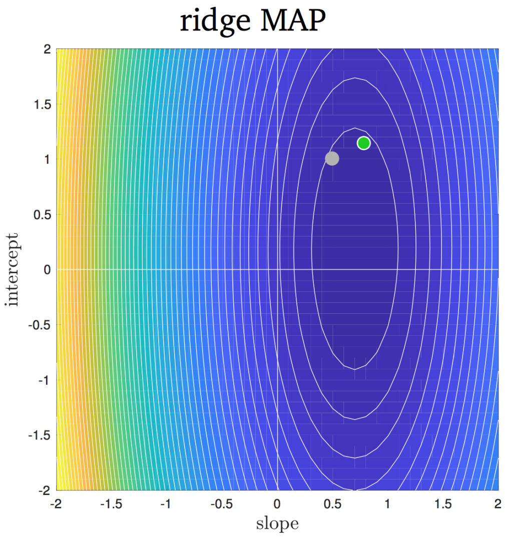
we would like to know what parameters MLE and MAP will select, after providing them with some dataset $\mathcal{D}$. Let's start with MLE:



The diagram above shows the the contours of the likelihood distribution in model space. The gray dot represents the true underlying model. MLE chooses the point that maximizes the likelihood, which is indicated by the green dot. As we can see, MLE chooses a reasonable hypothesis, but this hypothesis lies in a region on high variance, which indicates a high level of uncertainty in the predicted model. A slightly different dataset could significantly alter the predicted model.

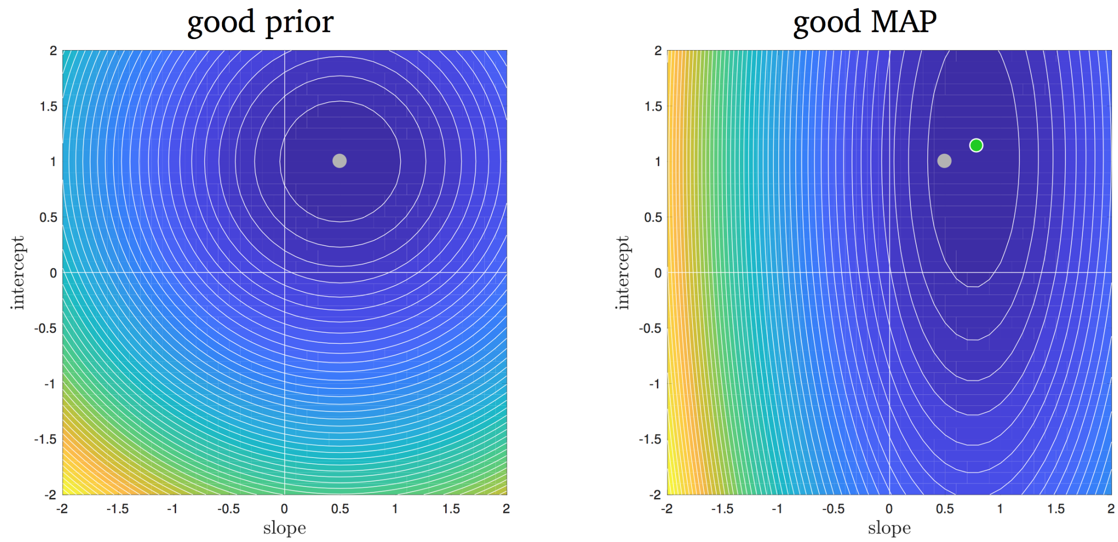
Now, let's take a look at the hypothesis model from MAP. One question that arises is where the prior should be centered and what its variance should be. This depends on our belief of what the true underlying model is. If we have reason to believe that the model weights should all be small, then the prior should be centered at zero with a small variance. Let's look at MAP for a prior that is centered at zero:



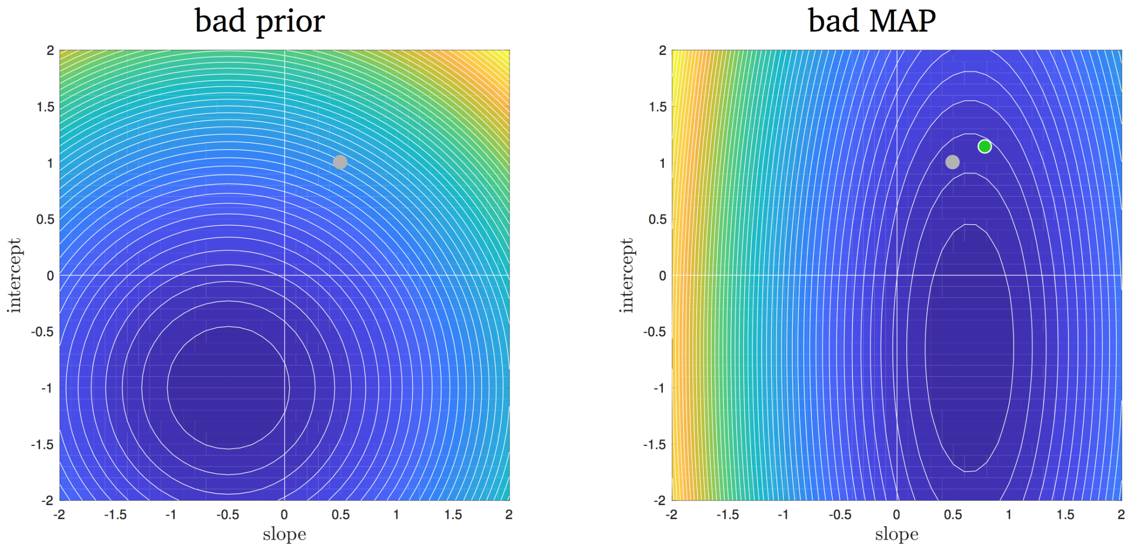


For reference, we have marked the MLE estimation from before as the green point and the true model as the gray point. The prior distribution is indicated by the diagram on the left, and the posterior distribution is indicated by the diagram on the right. MAP chooses the point that maximizes the posterior probability, which is approximately $(0.70, 0.25)$. Using a prior centered at zero leads us to skew our prediction of the model weights toward the origin, leading to a less accurate hypothesis than MLE. However, the posterior has significantly less variance, meaning that the point that MAP chooses is less likely to overfit to the noise in the dataset.

Let's say in our case that we have reason to believe that both model weights should be centered around the 0.5 to 1 range.



Our prediction is now close to that of MLE, with the added benefit that there is significantly less variance. However, if we believe the model weights should be centered around the -0.5 to -1 range, we would make a much poorer prediction than MLE.



As always, in order to compare our beliefs to see which prior works best in practice, we should use cross validation!