

1 Gradient Decent Optimization

1.1 Illustration

At one iteration, for each variable, **while fixing all other variables at their temporal values**, increases or decrease by a small step-length so that the loss can be decreased. And for each variable, this step-length is determined by how much can it decrease the loss: the more (decrement) the longer (step-length), so that the efficiency can be maximized.

2 Bayesian

2.1 Prior-Posterior Iteration of Encoding

Bayesian approach is an iterative process of encoding. It is encoding since it encodes the prior assumptions based on human illustration and the likelihood based on data. And it is iterative since when adding new data, the prior of the of the new posterior is the posterior gained by the old data.

2.2 Comparing to MAP

Traditional MAP approach gains a best fit point only on parameter-space, and being irrelevant to the size of dataset. Contrarily, Bayesian approach gains a distribution, which encodes the human prior, data character, and importantly the size of data, encoded as the confidence level in the posterior. So, Bayesian approach extract more information from data.

2.3 Overcoming Over-Fitting

This is one aspect of how over-fitting is overcome in Bayesian approach. When the size of dataset is small, the confidence level decreases, the sampling in parameter-space becomes dispersing; and when the size is large, the confidence level increases, the sampling in parameters becomes concentrated. In one word, the uncertainty caused by the lack of data is explicitly reflected in Bayesian approach, as it should be. (However, this uncertainty is absent in traditional MAP approach.)

Another aspect of how over-fitting is overcome in Bayesian approach is that it can encode the higher but not the highest peaks in posterior. That is, encoding more information of data, which, even though not being the most important, can still be significant for inference. Contrarily, the traditional MAP approach can only encode the highest peak in posterior.