# ColumbiaX: Machine Learning: Week1 : maximum likelihood estimation

### A commom approach for inferring the hidden parameters of probabilistic model

Given the observed data, we have defined the joint distribution on them,which represents a probabilistic model. But we don’t know $\theta$ yet. How do we choose the parameter? For a specific objective function, we can arbitrarily pick values of $\theta$ and evaluate the objective function on the data to see if it decreases or increases. Or we can have a principled way to do this.

> $\nabla \prod_{i=1}^{n} \mathbf{p}(\mathbf{x}_i|\theta) = 0$

which basically says that the gradient of the joint distribution, when setting to 0 , gained $\theta$ is the optimal value to maximize the value of the joint distribution.It tells us what’s the probability of this parametrized probabilistic model generating the observed data. Usually n is large and computing the derivative manually is a very complicated thing to do. So we use the logarithm trick to do this.

> We want to $\mathbf{argmax}_\theta \prod_{i=1}^{n} \mathbf{p}(\mathbf{x}_i|\theta)$
> it is equal to $\mathbf{argmax}_\theta \mathbf{ln}(\prod_{i=1}^{n} \mathbf{p}(\mathbf{x}_i|\theta))$

When the log function reaches its largest value, its input x will also reaches its largest value, in this case, x is the joint distribution, due to the log function is monotically increasing.

> furthermore, it is equal to $\mathbf{argmax}_\theta \sum_{i=1}^{n} \mathbf{ln}(\mathbf{p}(\mathbf{x}_i|\theta))$
> In this way, the original gradient(1st equation) turns out to be
> $\nabla \sum_{i=1}^{n} \mathbf{ln}(\mathbf{p}(\mathbf{x}_i|\theta)) = 0$

what the above equations done is to transform a hard-to-solve equation to a easy equation.

for a chosen model, we can solve this by
1\ analytically - an accurate solution
2\ numerically - via an iterative algorithm using different equations
3\ approximately