EM ALGORITHM

What is it?

An iterative algorithm that finds **local** maximum likelihood estimates.

CURRENT USES

- Image Processing
- Data Mining
- Bayesian Statistics
- Machine Learning
- Statistical Genetics

Why do we care?

EM algorithm is most helpful when we have a variable dependent on our parameters that we can't directly observe or measure. This is called a **Latent** variable

HOW DOES THIS WORK?

LATENT VARIABLE EXAMPLE

Mo sells ice cream in the park everyday. He keeps a log of the ice creams he sells

Mo's icecream log: $X_1, X_2 \dots X_n$
where X_i = number of icecreams sold that day

Mo does this illegally, and some days he is kicked out of the park by the police, which can be represented by the random variable Z.

Mo wants to estimate how much money he would make per day without getting kicked out, but he doesn't know which days he got kicked out. We can't estimate θ just using X because our latent variable Z impacts our data in an unknown way. Unless Mo uses the EM algorithm

FIND OUT how the EM algorithm CAN HELP →

The EM algorithm uses **M**aximum **L**ikelihood **E**stimation

We learned MLE that takes in complete data to estimate a parameter θ by taking the maximum of the likelihood function.

complete data → $L(\theta; y)$ → $\arg\max L(\theta; y)$ → θ estimate

When we use the EM algorithm we DON'T have the complete data. We use something called the Q function instead. The Q function uses an estimate of θ and the incomplete data to estimate the complete data, then using this estimate of the complete data we can maximize the Q function to get a new estimate for θ

old θ → Q function → estimate of complete data → Q function $\arg\max$ → new θ_1

observed data → Q function

we can do this iteratively until our θ converges

θ_1 → Q function → estimate of complete data 2 → Q function $\arg\max$ → new θ_2

observed data → Q function → even better estimate