MLE

The MLE has three key properties:

- Consistency ($\hat{\theta}^{MLE} \xrightarrow{\mathbb{P}} \theta^*$)
- Asymptotically normal ($\sqrt{n}(\hat{\theta}^{MLE} - \theta^*) \rightsquigarrow \mathcal{N}(0, \sigma^2_{MLE})$)
- Asymptotic efficiency (for any other estimator that is asymptotically normal, its asymptotic variance is at least that of the MLE estimator)

Note:

- The MLE is not always asymptotically normal!
  - If the log likelihood is differentiable, then it is asymptotically normal
  - Otherwise, it is not guaranteed
- As an example, the log likelihood for a uniform distribution between $[0, \theta]$ is not differentiable, but the MLE is just the max of the datapoints
  - Then, $\sqrt{n}(\hat{\theta}^{MLE} - \theta^*) \rightsquigarrow 0$, and does not become normal
  - Can see this because the term in parentheses will always be negative, and normal variables are both positive and negative

MLE Alternative Origin

- For each θ, we want to compute the "distance" between the distribution P_θ and P_{θ^*}
- We can't compute this exactly since we only have samples, so we will approximate the distance, which we will denote as $\widehat{\text{dist}}(P_\theta, P_{\theta^*})$
- We need this distance metric to satisfy:
  - Approximately computable from samples
  - Minimized only at θ^*
    - Should be 0 at θ^* and > 0 everywhere else

KL Divergence

Let f_{θ^*} and f_θ be the pdfs associated with P_{θ^*} and P_θ respectively. Then, the KL divergence is defined as:

$$D_{KL}(P_{\theta^*} \| P_\theta) = \int f_{\theta^*} \log\left(\frac{f_{\theta^*}(x)}{f_\theta(x)}\right) dx$$

Notes:

- Isn't symmetric and doesn't satisfy triangle inequality, so not really a "distance"
- This function is always non-negative and is 0 only when $P_\theta = P_{\theta^*}$
  - If the parameter is identifiable, then this means $\theta = \theta^*$ is the unique minimizer of the KL divergence

We can rewrite the KL divergence as:

$$D_{KL}(P_{\theta^*} \| P_\theta) = \int f_{\theta^*}(x) \log f_{\theta^*}(x) dx - \int f_\theta(x) \log f_\theta(x) dx$$

The first term is actually fixed, so minimizing this is equivalent to just maximizing:

$$\int f_{\theta^*}(x) \log f_\theta(x) dx = \mathbb{E}_{\theta^*}\left[\log f_\theta(X)\right]$$

An expectation can be approximated with a sample average, so this just becomes maximizing:

$$\frac{1}{n} \sum_{i=1}^{n} \log f_\theta(X_i)$$

This is just our original MLE definition but with an extra $1/n$ (which can be safely discarded)!

Population Log-Likelihood

$$\ell(\theta) = \mathbb{E}_{\theta^*}\left[\log f_\theta(X)\right]$$

This is known as the **population log-likehood** and it is maxmized at θ^*