Check: $1 \geq \theta \geq 0$ (no constraints necessary)

- MLE gives the explanation of the data you observed.
- If n is large and your model/distribution is correct (that is $\mathcal{H}$ includes the true model), then MLE finds the **true** parameters.
- But the MLE can overfit the data if n is small. It works well when n is large.
- If you do not have the correct model (and n is small) then MLE can be terribly wrong!

For example, suppose you observe H,H,H,H,H. What is $\hat{\theta}_{MLE}$?

Simple scenario: coin toss with prior knowledge

Assume you have a hunch that θ is close to $\theta' = 0.5$. But your sample size is small, so you don't trust your estimate.

Simple fix: Add m imaginery throws that would result in θ' (e.g. $\theta = 0.5$). Add m Heads and m Tails to your data.

$$\hat{\theta} = \frac{n_H + m}{n_H + n_T + 2m}$$

For large n, this is an insignificant change. For small n, it incorporates your "prior belief" about what θ should be.

Can we derive this formally?

The Bayesian Way

Model θ as a **random variable**, drawn from a distribution $P(\theta)$. Note that θ is **not** a random variable associated with an event in a sample space. In frequentist statistics, this is forbidden. In Bayesian statistics, this is allowed.
Now, we can look at $P(\theta \mid D) = \frac{P(D|\theta)P(\theta)}{P(D)}$ (recall Bayes Rule!), where

- $P(D \mid \theta)$ is the **likelihood** of the data given the parameter(s) θ,
- $P(\theta)$ is the **prior** distribution over the parameter(s) θ, and
- $P(\theta \mid D)$ is the **posterior** distribution over the parameter(s) θ.

Now, we can use the Beta distribution to model $P(\theta)$:

$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the normalization constant. Note that here we only need a distribution over a binary random variable. The multivariate generalization of the Beta distribution is the Dirichlet distribution.

Why using the Beta distribution?

- it models probabilitis (θ lives on $[0,1]$ and $\sum_i \theta_i = 1$)
- it is of the same distributional family as the binomial distribution (**conjugate prior**) $\longrightarrow$ the math will turn out nicely:

$$P(\theta \mid D) \propto P(D \mid \theta)P(\theta) \propto \theta^{n_H+\alpha-1}(1-\theta)^{n_T+\beta-1}$$

Note taht in general θ are the parameters of our model. For the coin flipping scenario $\theta = P(H)$.
So far, we have a distribution over θ. How can we get an estimate for θ?