Bayesian approach to statistics

Uncertainty interpretation

A frequentist would say: A coin flip has 1/2 heads probs, 1/2 tails probs.

A bayesian practicioner would say: We can parameterize the angle of the coin flip, velocity, thumb acceleration, etc.
The outcome of the experiment can be predicted!

Frequentist

Bayesian

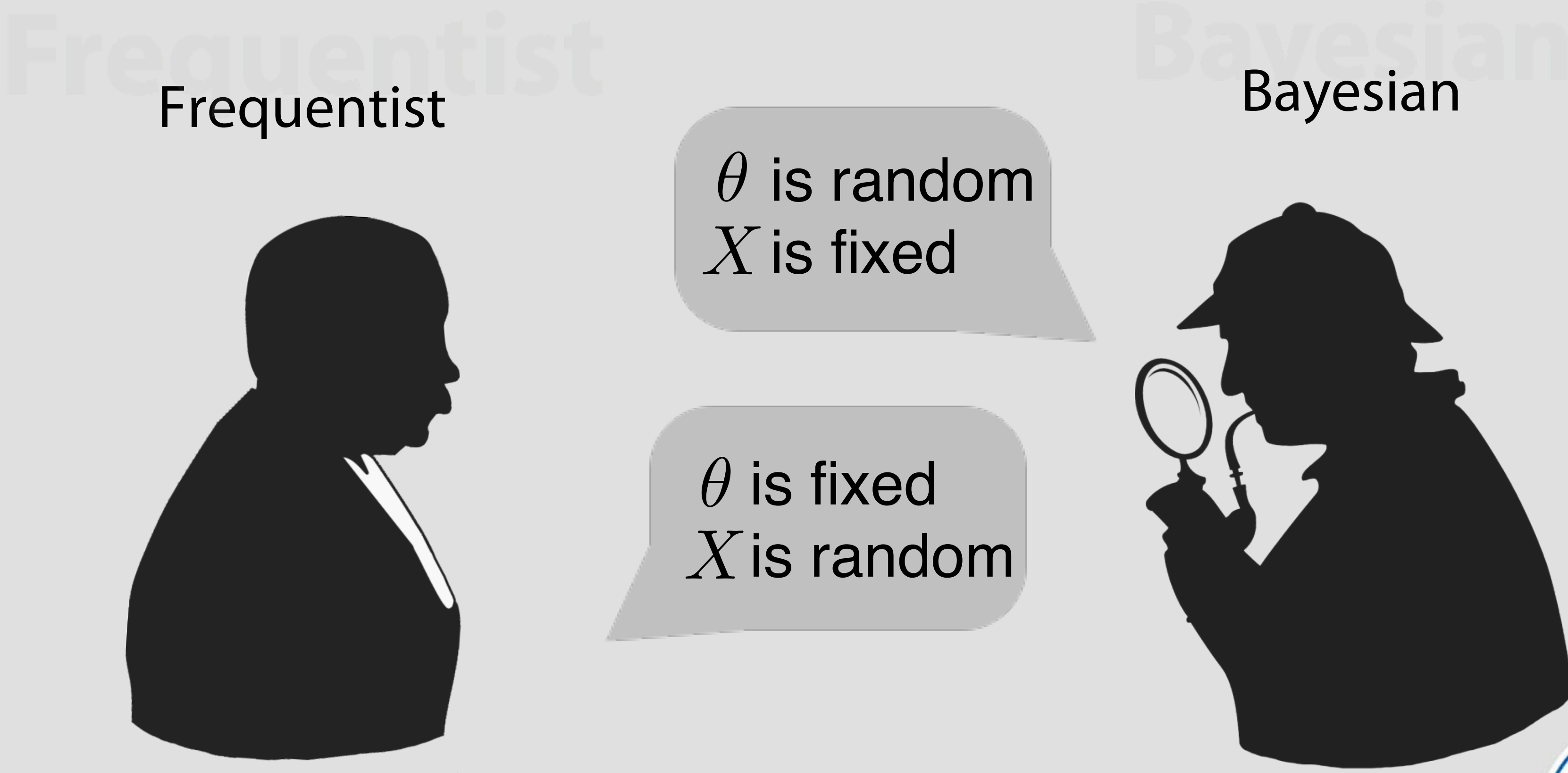
Subjective

Objective

Data and parameters

Theta is parameters. X is the data.

The bayesian approach sorts of make sense.
When we 're training a machine learning model for example,
our data (X) is fixed, yet the parameters (theta) are random.



Data and parameters

Frequentist

Bayesian

number of data doesn't matter.

For any $|X|$

$|X| \gg |\theta|$

Datapoints must be significantly larger
than number of parameters.

Training

Find theta such that we get the highest probability of data given parameters theta.

Frequentist

Maximum Likelihood:

$$\widehat{\theta} = arg\max_{\theta} P(X|\theta)$$

Bayesian

Training

Frequentist

We compute the posterior.

Bayesian

Frequentist

Bayesian

Bayes theorem:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Classification

For the bayesian method,

Training:

$$P(\theta|X_{\mathrm{tr}}, y_{\mathrm{tr}}) = \frac{P(y_{\mathrm{tr}}|X_{\mathrm{tr}}, \theta)P(\theta)}{P(y_{\mathrm{tr}}|X_{\mathrm{tr}})}$$

Prediction:

$$P(y_{\mathrm{ts}}|X_{\mathrm{ts}}, X_{\mathrm{tr}}, y_{\mathrm{tr}}) = \int P(y_{\mathrm{ts}}|X_{\mathrm{ts}}, \theta)P(\theta|X_{\mathrm{tr}}, y_{\mathrm{tr}})d\theta$$

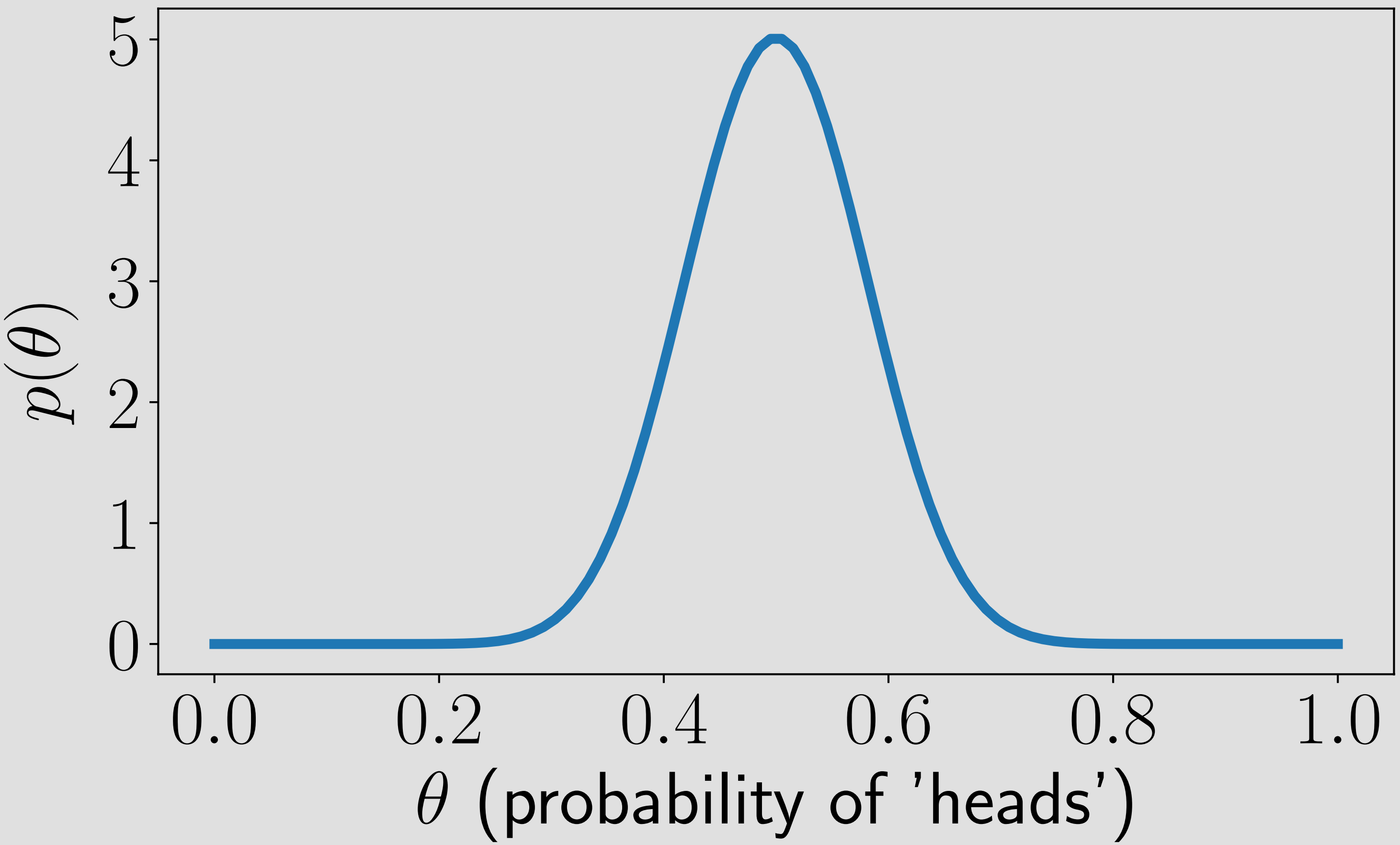
Regularization

Treat prior probability as a regularizer.

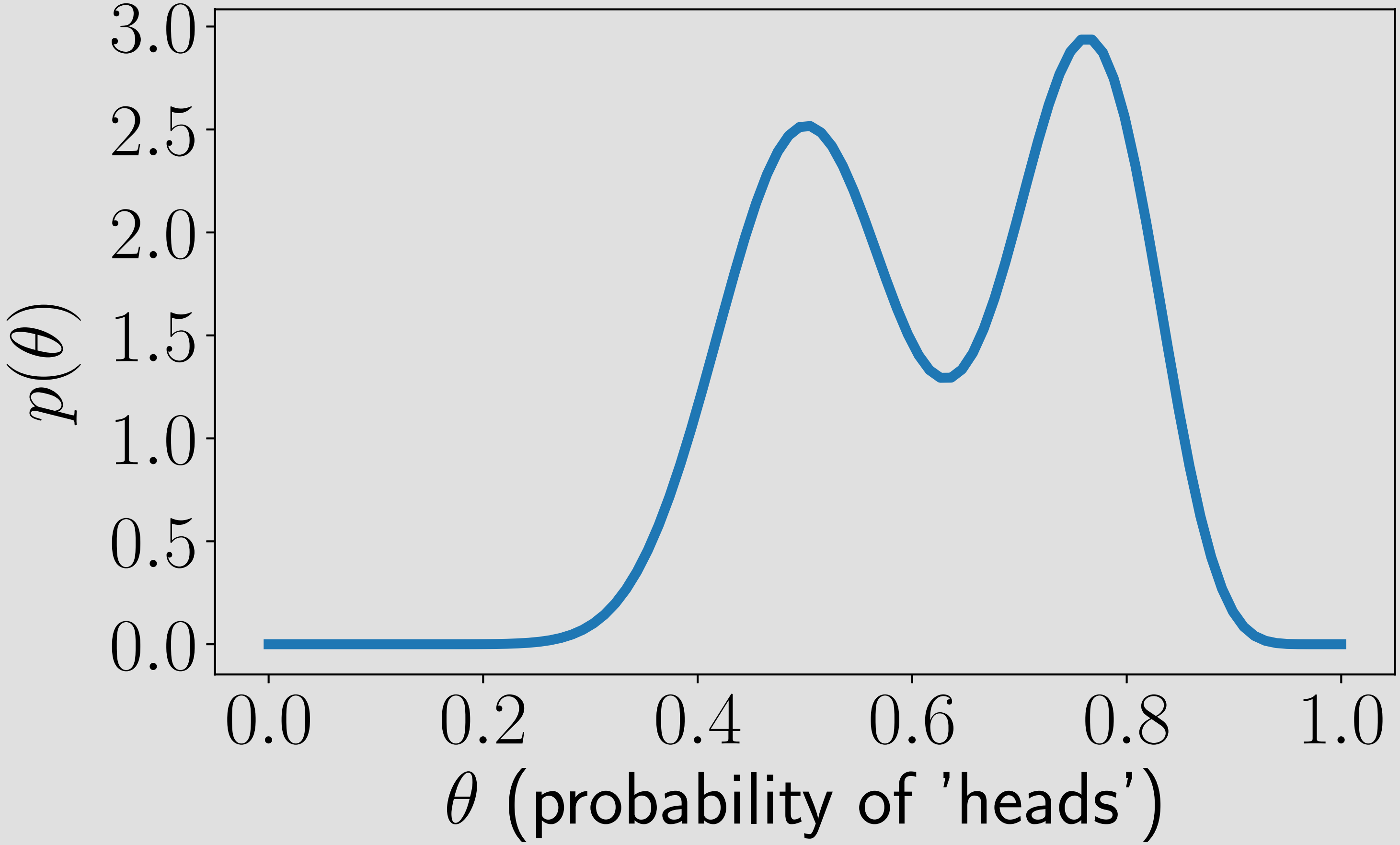
Regularizer

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Regularization



Regularization



On-line learning

New prior

Likelihood

Prior

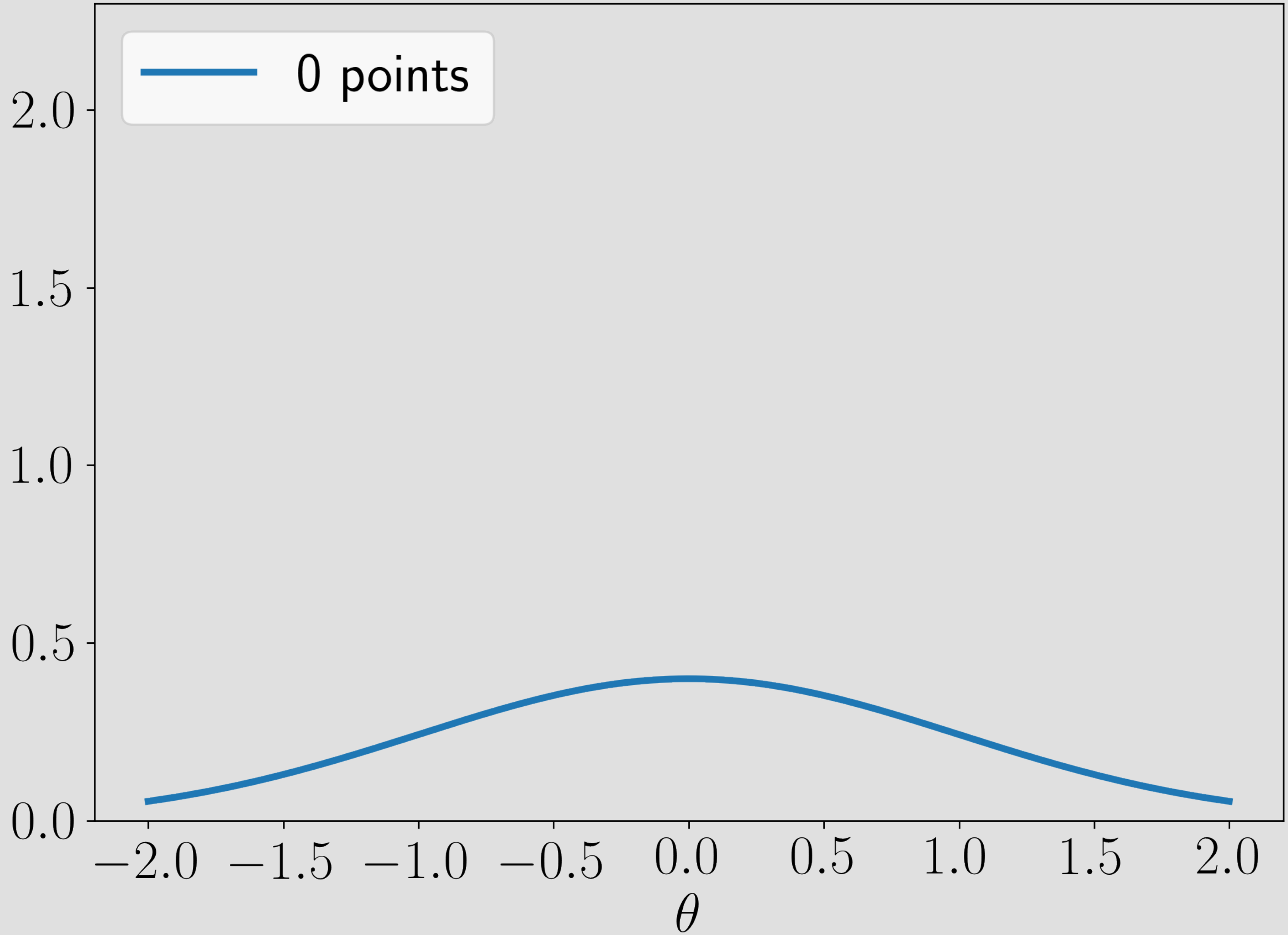
$$P_k(\theta) = P(\theta|x_k) = \frac{P(x_k|\theta)P_{k-1}(\theta)}{P(x_k)}$$

Posterior

Set prior to posterior.

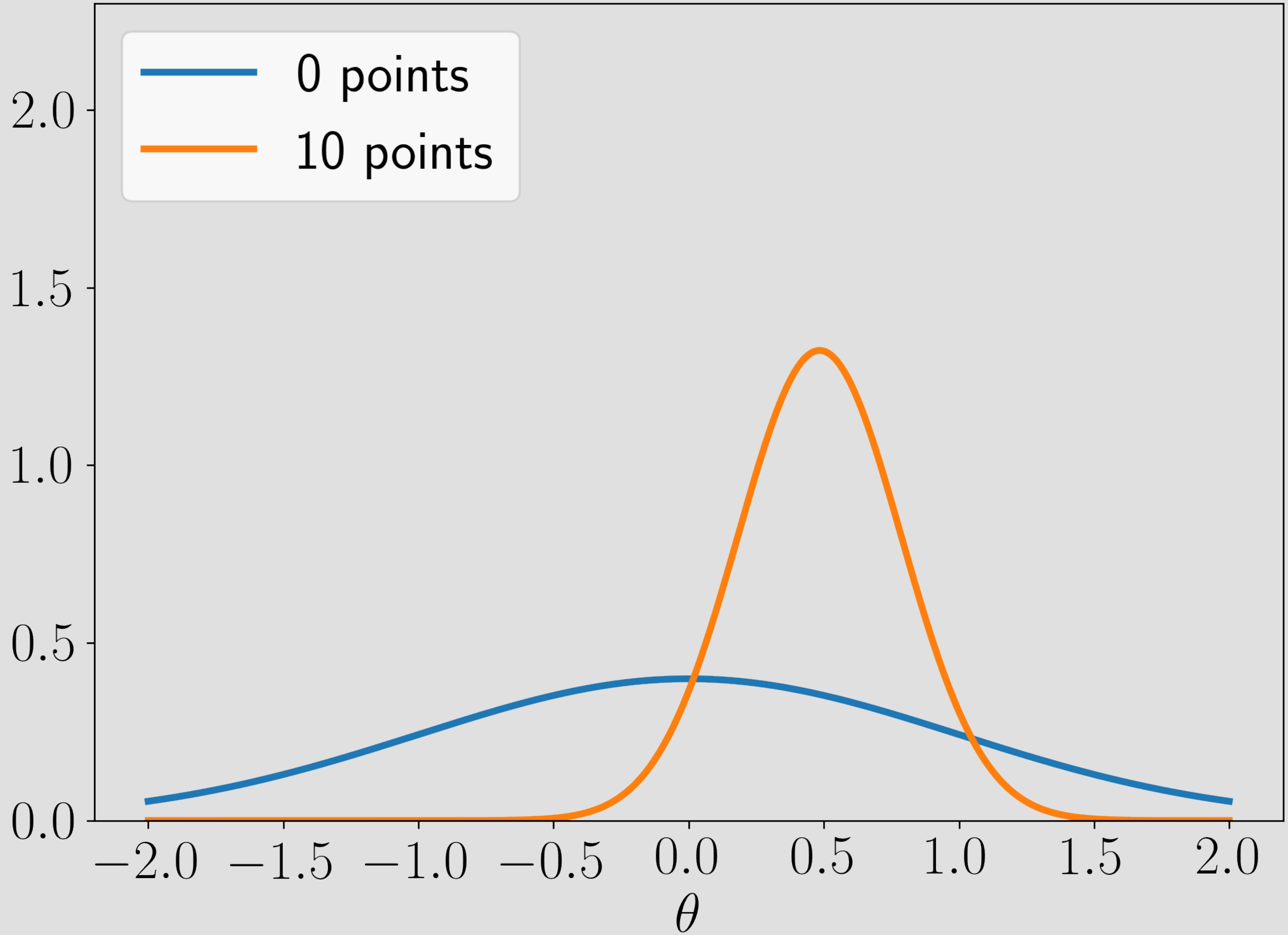
On-line learning

Suppose we want to estimate the parameter theta, and our prior is as follows:



On-line learning

When we update our prior = posterior, we get a new shape. The variance of this distribution is less, and mean has changed.



On-line learning