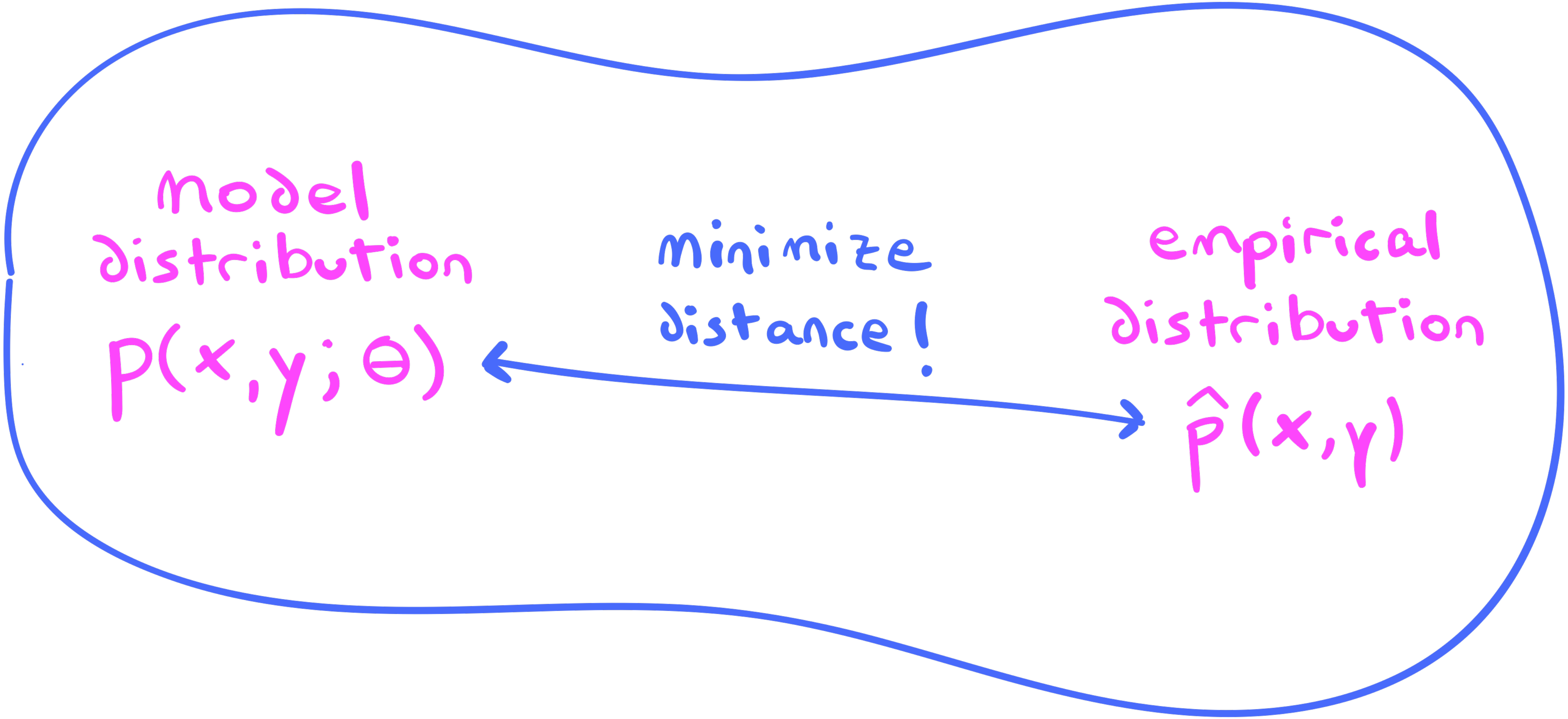
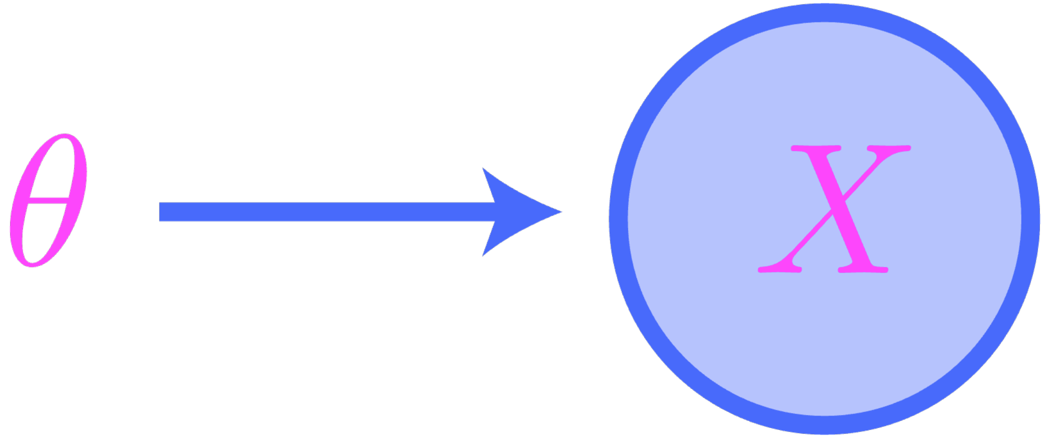
13. Learning

> **The Distance Criterion for Parameter Choice.** Given two model distributions within the same family of probabilistic models, choose the model distribution whose *distance* from the empirical distribution of the data is smaller.



model
distribution
p(x,y;θ)
minimize
distance!
empirical
distribution
p̂(x,y)

13.1. A first look at likelihood-based learning objectives


θ
X

Theorem 13.1 (Equivalent learning objectives for the univariate Bernoulli model)

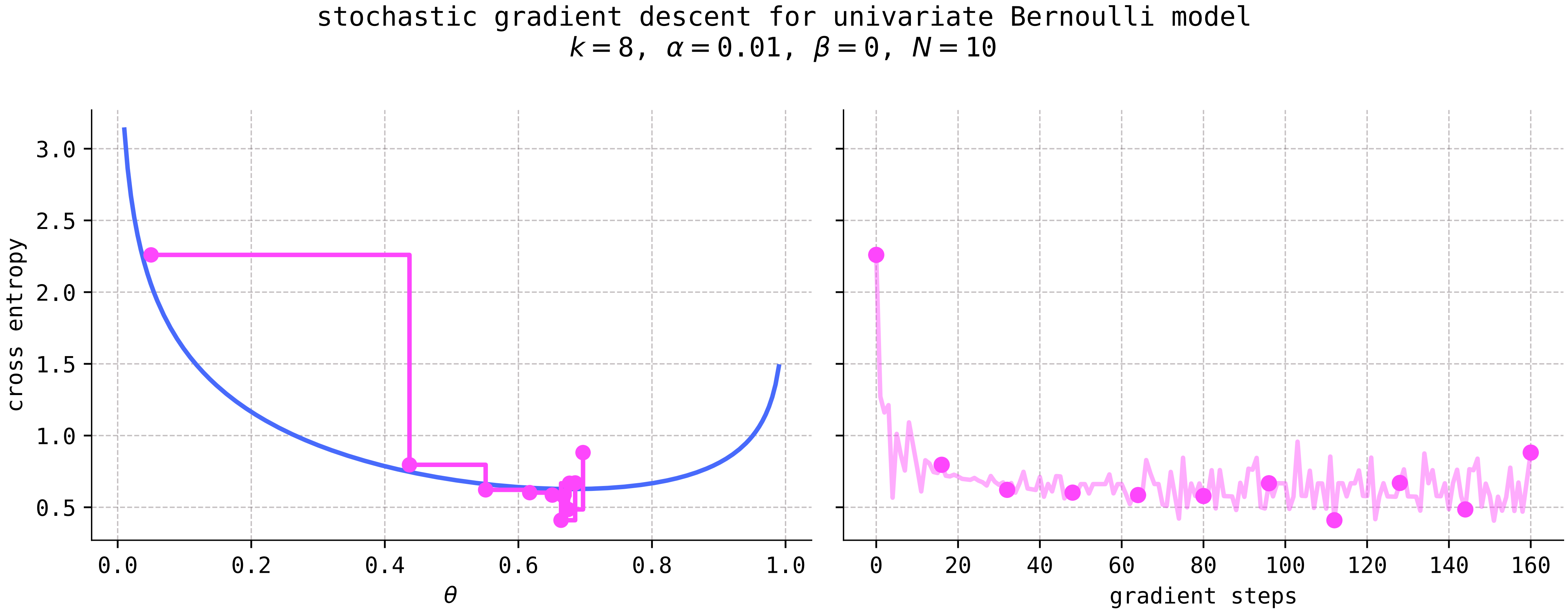
Let $x_1, x_2, \ldots, x_m \in \{0, 1\}$ be an observed dataset corresponding to a Bernoulli random variable $X \sim \mathcal{B}er(\theta)$ with unknown θ. Let P_θ be the model distribution of X and let $\hat{P}$ be the empirical distribution of the dataset. The following optimization objectives are equivalent:

1. Minimize the KL divergence $D(\hat{P} \parallel P_\theta)$ with respect to θ.
2. Minimize the cross entropy $H_{\hat{P}}(P_\theta)$ with respect to θ.
3. Minimize the data surprisal function $\mathcal{I}(\theta; x_1, \ldots, x_m)$ with respect to θ.
4. Maximize the data likelihood function $\mathcal{L}(\theta; x_1, \ldots, x_m)$ with respect to θ.

1. Minimizing the KL divergence between the empirical and model distributions has an immediate and concrete interpretation as minimizing the "distance" between these two distributions.
2. As a function of θ, the cross entropy $J(\theta) = H_{\hat{P}}(P_\theta)$ may be viewed as a stochastic objective function, since it is exactly the mean of the model surprisal function. This opens the door for applications of the stochastic gradient descent algorithm studied in Section 11.4.
3. The third optimization objective seeks the model probability distribution according to which the data is *least surprising*.
4. The fourth optimization objective seeks the model probability distribution according to which the data is *most likely*.

Theorem 13.2 (MLE for the univariate Bernoulli model)

Let $x_1, x_2, \ldots, x_m \in \{0, 1\}$ be an observed dataset corresponding to a Bernoulli random variable $X \sim \mathcal{B}er(\theta)$ with unknown θ. Then the (unique) maximum likelihood estimate $\theta^\star_{\text{MLE}}$ is the ratio $\Sigma x/m$.


stochastic gradient descent for univariate Bernoulli model
$k = 8$, $\alpha = 0.01$, $\beta = 0$, $N = 10$
cross entropy
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.2
0.4
0.6
0.8
1.0
θ
0
20
40
60
80
100
120
140
160
gradient steps