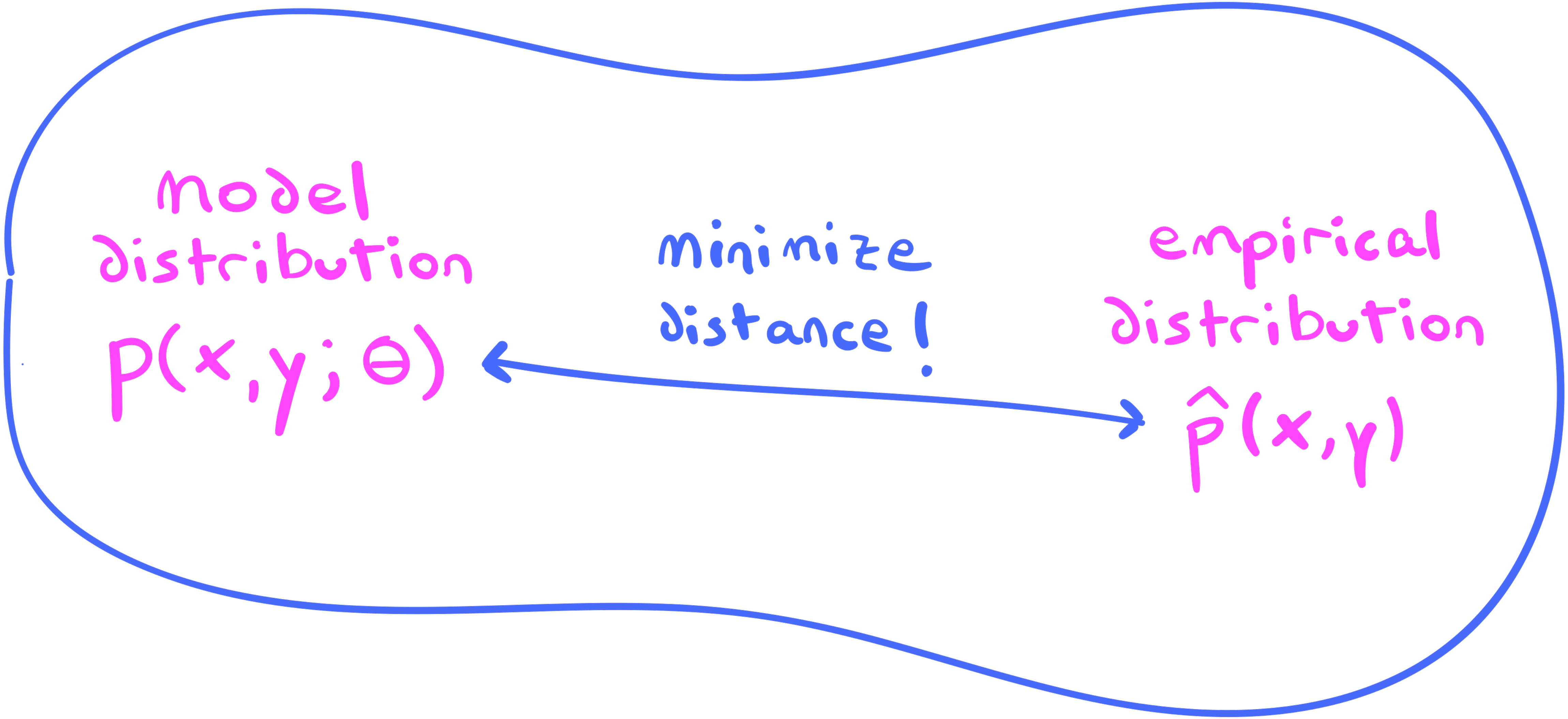
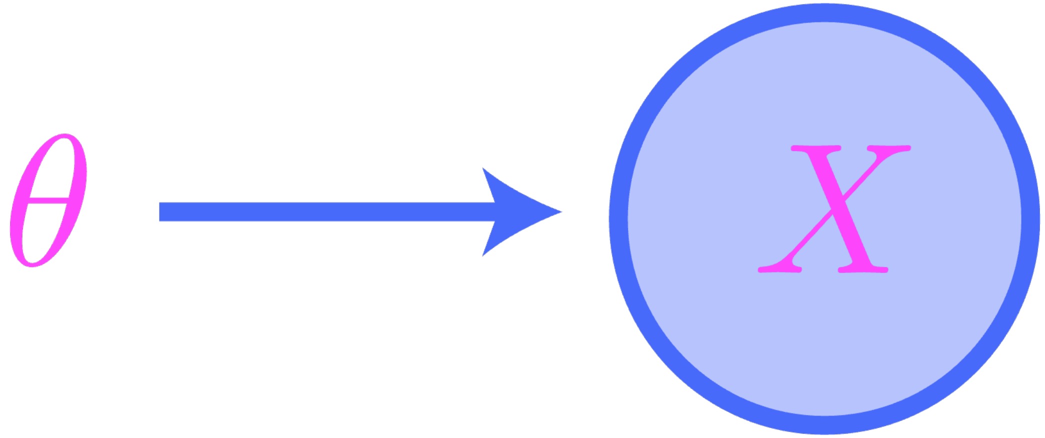
13. Learning

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
$p(x,y;\theta)$
minimize distance!
empirical distribution
$\hat{p}(x,y)$

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:

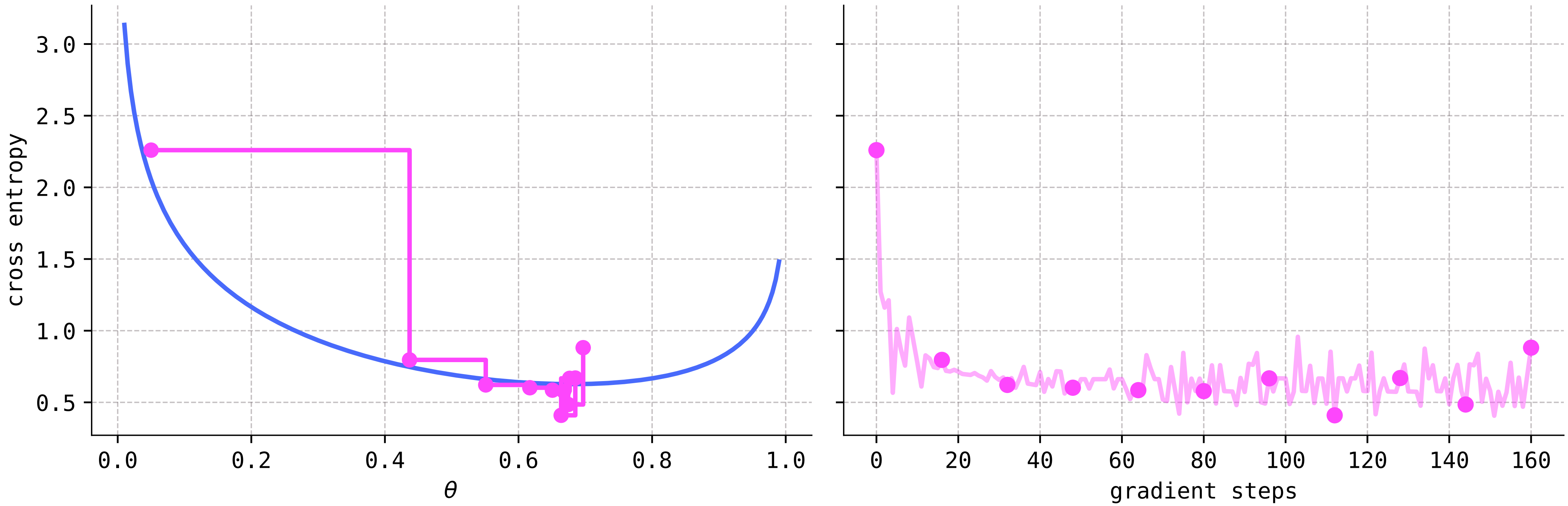
1. Minimize the KL divergence $D(\hat{P} \parallel P_\theta)$ with respect to θ.
2. Minimize the cross entropy $H_{\hat{P}}(P_\theta)$ with respect to θ.
3. Minimize the data surprisal function $\mathcal{I}(\theta; x_1, \ldots, x_m)$ with respect to θ.
4. Maximize the data likelihood function $\mathcal{L}(\theta;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

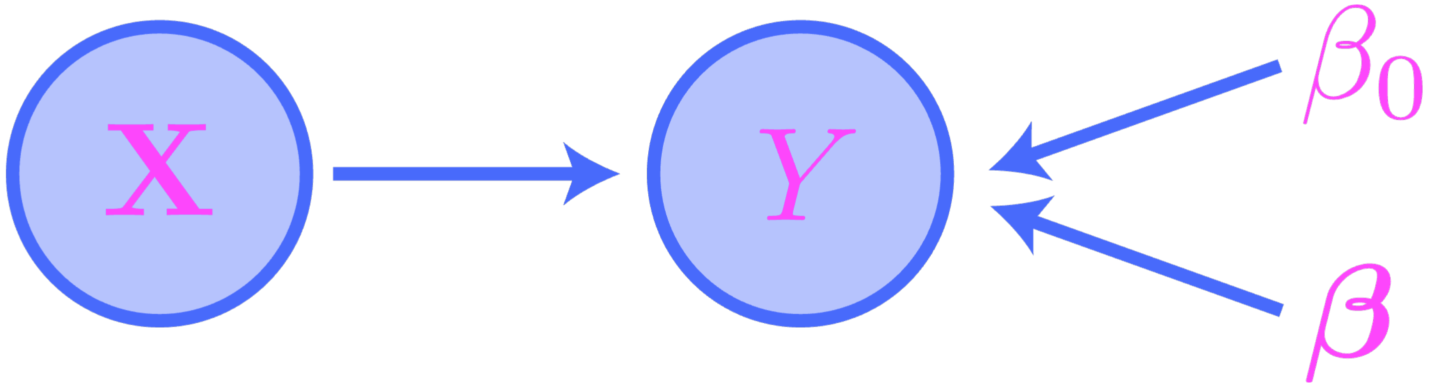
stochastic gradient descent for univariate Bernoulli model
$k = 8$, $\alpha = 0.01$, $\beta = 0$, $N = 10$



Problem Prompt

Do problem 1 on the worksheet.

13.3. MLE for linear regression



$\mathbf{X}$
Y
β_0
$\boldsymbol{\beta}$

Theorem 13.8 (MLEs for linear regression models with known variance)

Consider a linear regression model with *fixed* variance σ^2, and let

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m) \in \mathbb{R}^n \times \mathbb{R}$$

be an observed dataset. Supposing

$$\mathbf{x}_i^\intercal = (x_{0i}, x_{i1}, \ldots, x_{in}) = (1, x_{i1}, \ldots, x_{in})$$

for each $i = 1, \ldots, m$, let

$$\boldsymbol{\mathcal{X}} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\intercal & \rightarrow \\ \vdots & \vdots & \vdots \\ \leftarrow & \mathbf{x}_m^\intercal & \rightarrow \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$$

Provided that the $(n+1) \times (n+1)$ square matrix $\boldsymbol{\mathcal{X}}^\intercal \boldsymbol{\mathcal{X}}$ is invertible, maximum likelihood estimates for the parameters β_0 and $\boldsymbol{\beta}$ are given by

$$\boldsymbol{\theta}_{\text{MLE}}^\star = (\boldsymbol{\mathcal{X}}^\intercal \boldsymbol{\mathcal{X}})^{-1} \boldsymbol{\mathcal{X}}^\intercal \mathbf{y}.$$

Corollary 13.1 (MLEs for simple linear regression models with known variance)

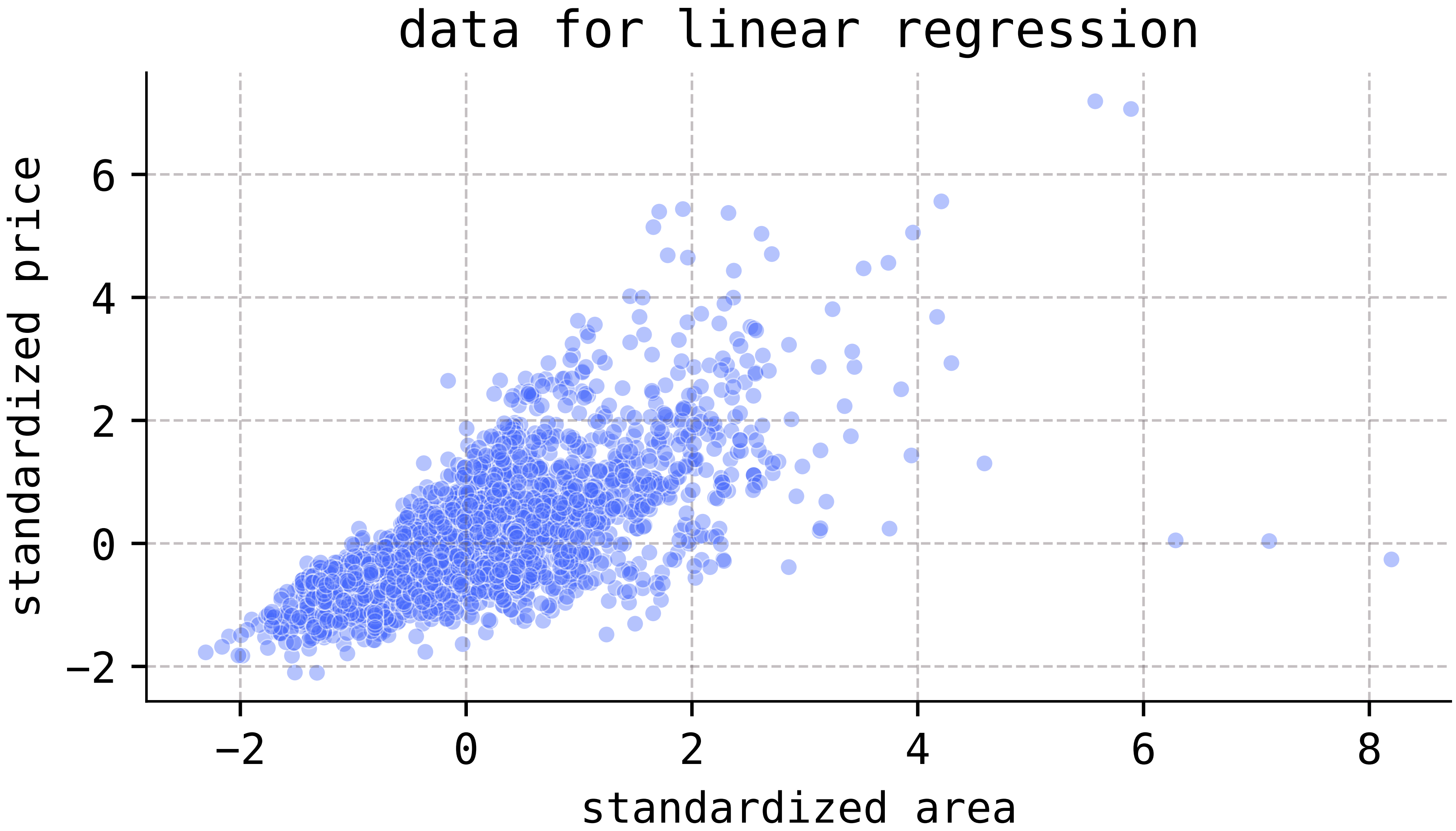
Let the notation be as in Theorem 13.8, but assume that $\mathbf{X}$ is 1-dimensional, equal to a random variable X. Then MLEs for the parameters β_0 and β_1 are given by

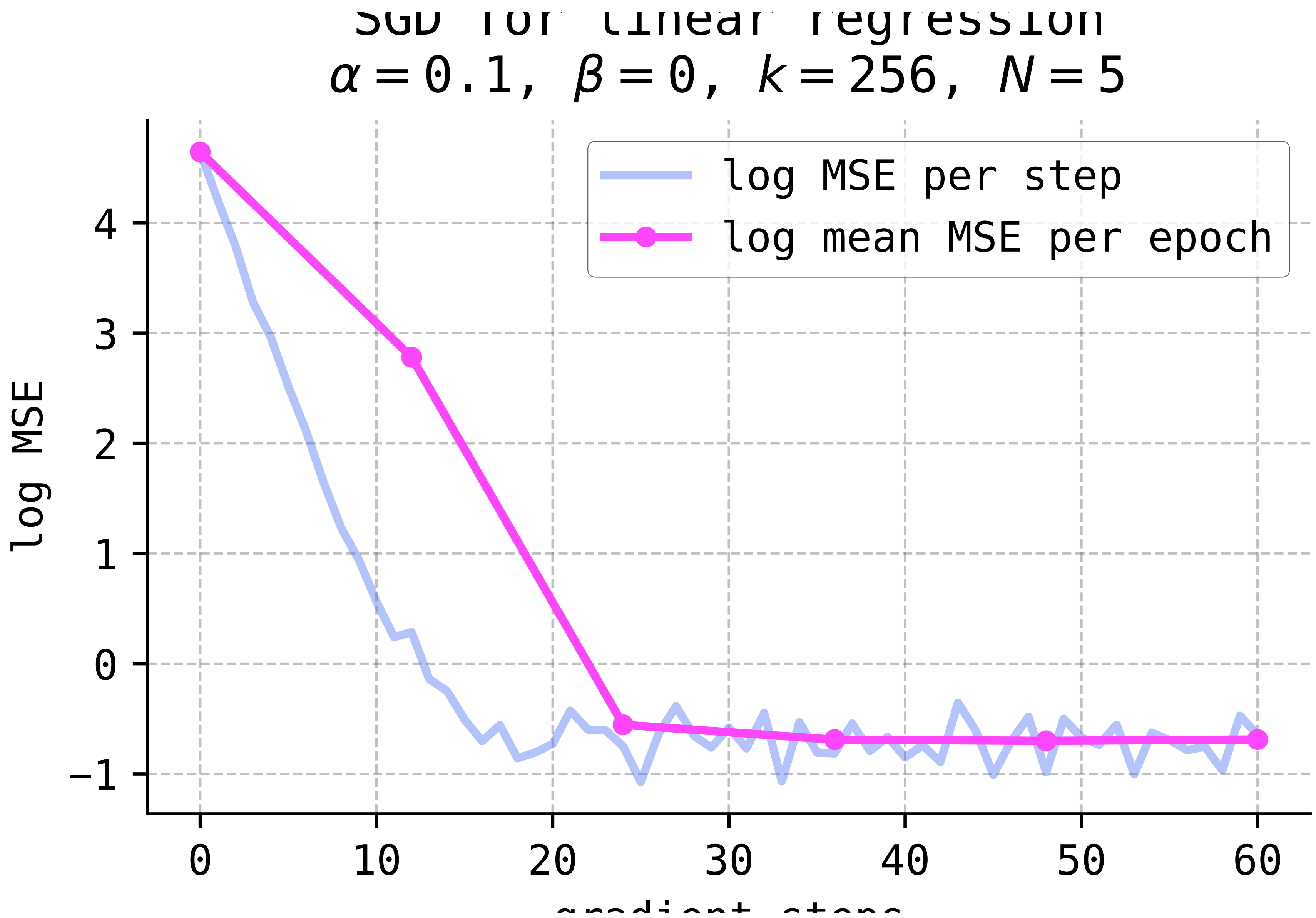
$$
\begin{aligned}
(\beta_1)^\star_{\text{MLE}} &= \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^m (x_i - \bar{x})^2}, \\
(\beta_0)^\star_{\text{MLE}} &= \bar{y} - (\beta_1)^\star_{\text{MLE}}\bar{x},
\end{aligned}
$$

where $\bar{x} = \frac{1}{m}\sum_{i=1}^m x_i$ and $\bar{y} = \frac{1}{m}\sum_{i=1}^m y_i$ are the empirical means.

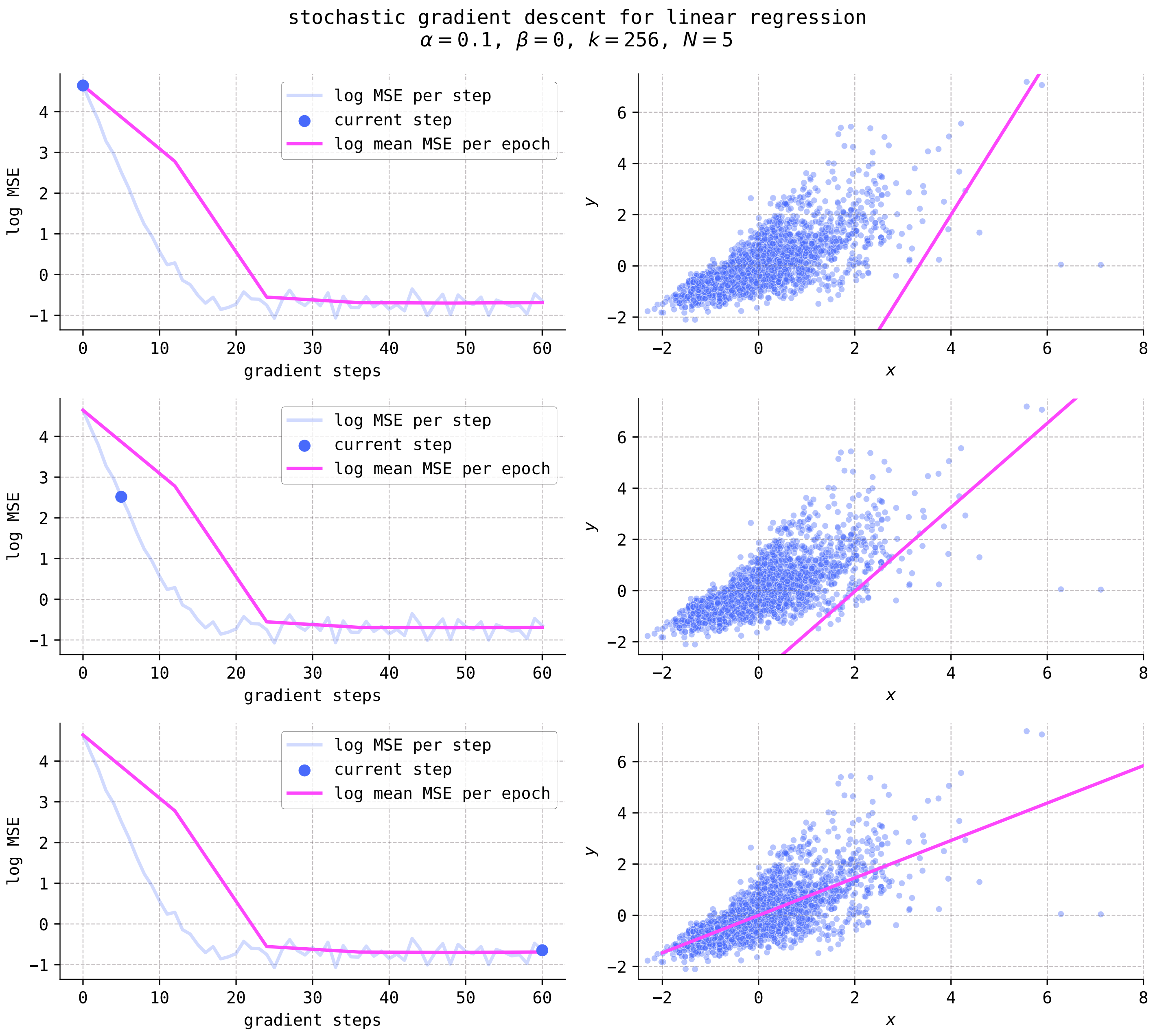
Problem Prompt

Do problem 2 on the worksheet.


data for linear regression
standardized price
6
4
2
0
−2
−2
0
2
4
6
8
standardized area



SGD for linear regression
$\alpha = 0.1,\ \beta = 0,\ k = 256,\ N = 5$
log MSE per step
log mean MSE per epoch
log MSE
gradient steps
4
3
2
1
0
−1
0
10
20
30
40
50
60


stochastic gradient descent for linear regression
α = 0.1, β = 0, k = 256, N = 5
log MSE per step
current step
log mean MSE per epoch
log MSE
gradient steps
x
y