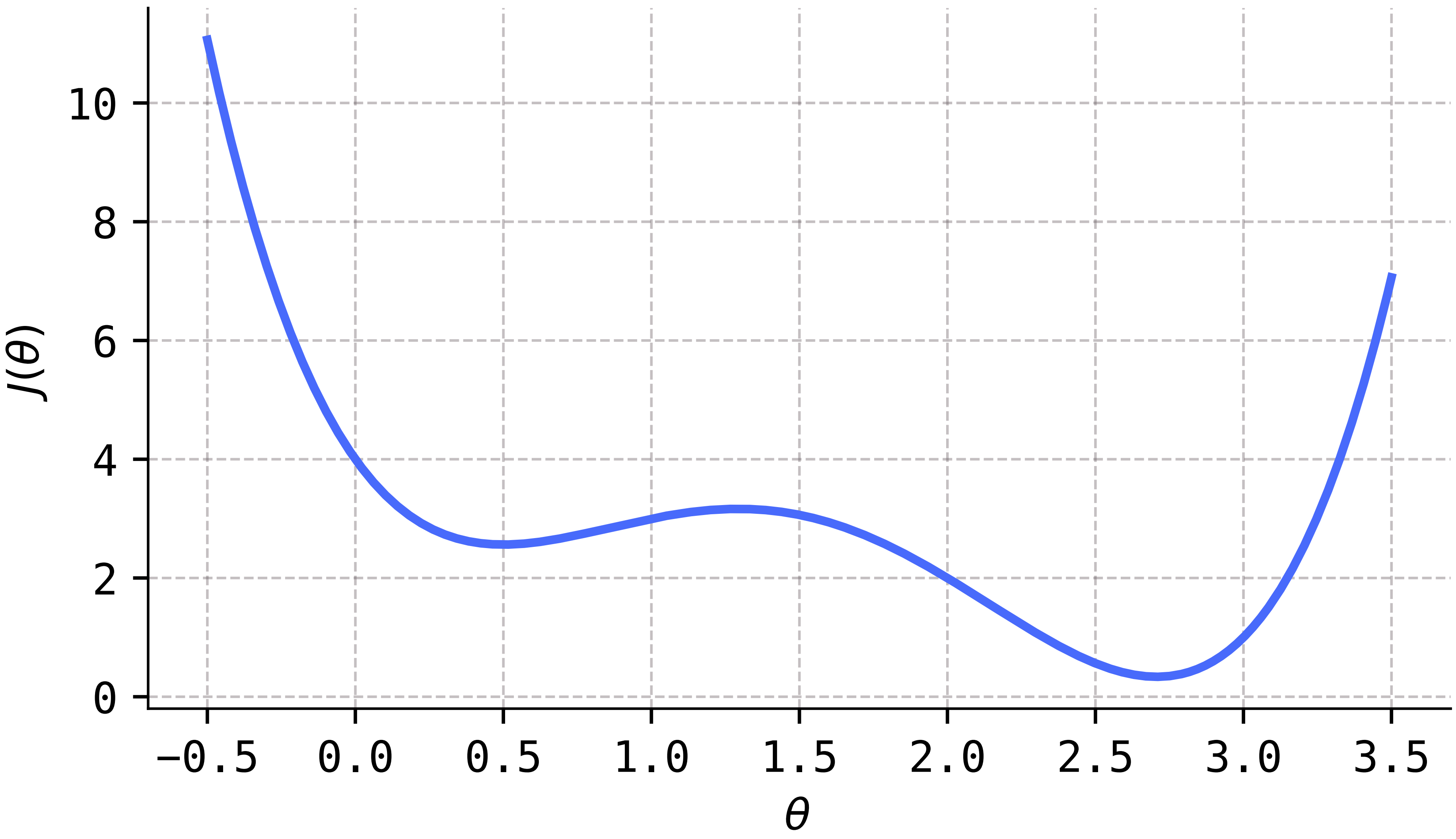
11. Optimization



$J(\theta)$
10
8
6
4
2
0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
θ

Definition 11.1

Let $J : \mathbb{R}^n \to \mathbb{R}$ be a function. A vector $\boldsymbol{\theta}^\star$ is a *local minimizer* of $J(\boldsymbol{\theta})$ provided that

$$J(\boldsymbol{\theta}^\star) \leq J(\boldsymbol{\theta})$$

for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}^\star$; if this inequality holds for *all* $\boldsymbol{\theta}$, then $\boldsymbol{\theta}^\star$ is called a *global minimizer* of $J(\boldsymbol{\theta})$. If we flip the inequality the other direction, then we obtain the definitions of *local* and *global maximizers*. Collectively, local and global minimizers and maximizers of $J(\boldsymbol{\theta})$ are called *extremizers*, and the values $J(\boldsymbol{\theta}^\star)$ of the function where $\boldsymbol{\theta}^\star$ is an extremizer are called *extrema* or *extreme values*.

Algorithm 11.1 (Single-variable gradient descent)

Input: A differentiable objective function $J : \mathbb{R} \to \mathbb{R}$, an initial guess $\theta_0 \in \mathbb{R}$ for a local minimizer $\theta^\star$, a learning rate $\alpha > 0$, and the number N of gradient steps.

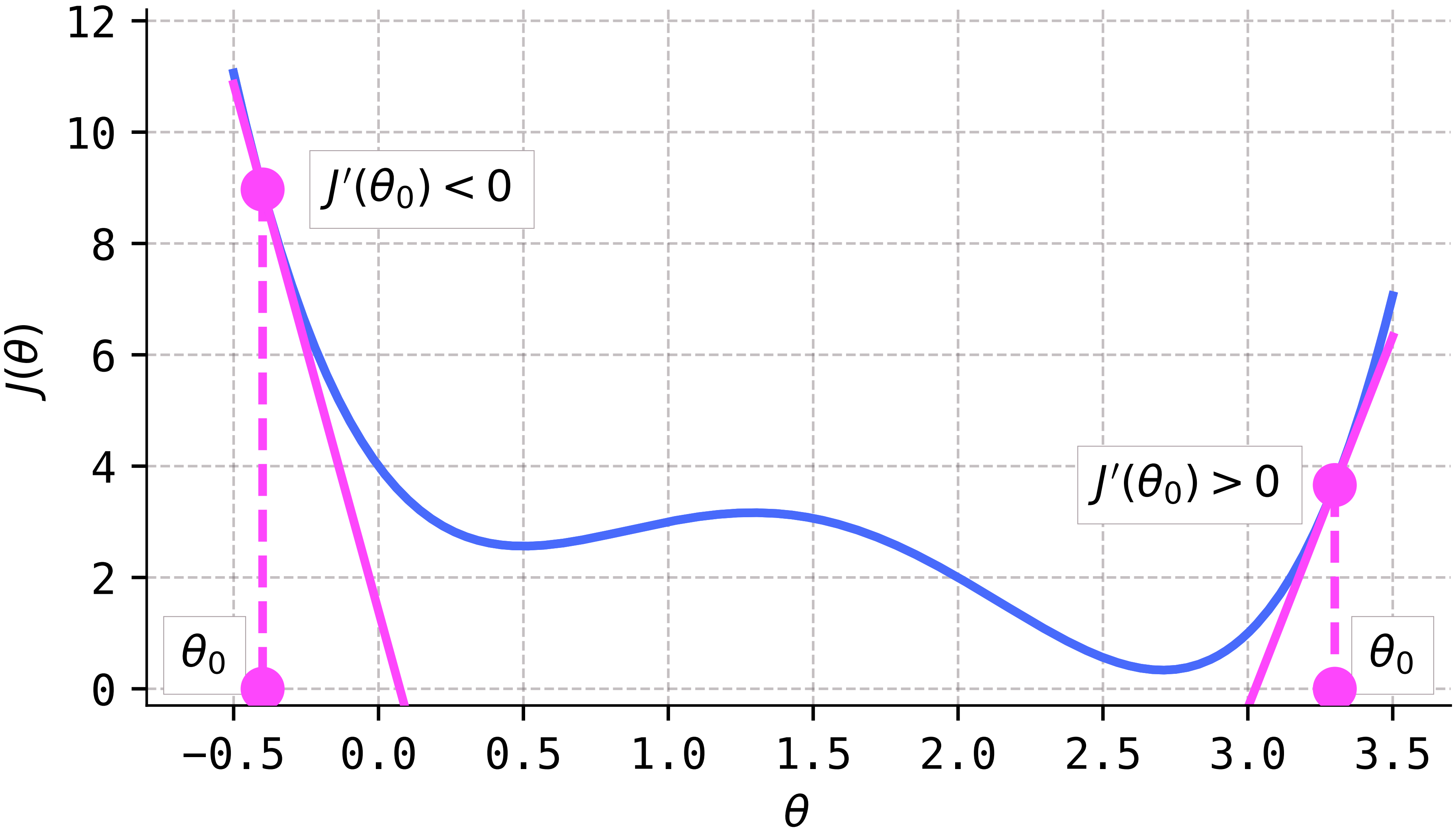
Output: An approximation to a local minimizer $\theta^\star$.

$\theta := \theta_0$

For t from 0 to $N-1$, do:

    $\theta := \theta - \alpha J'(\theta)$

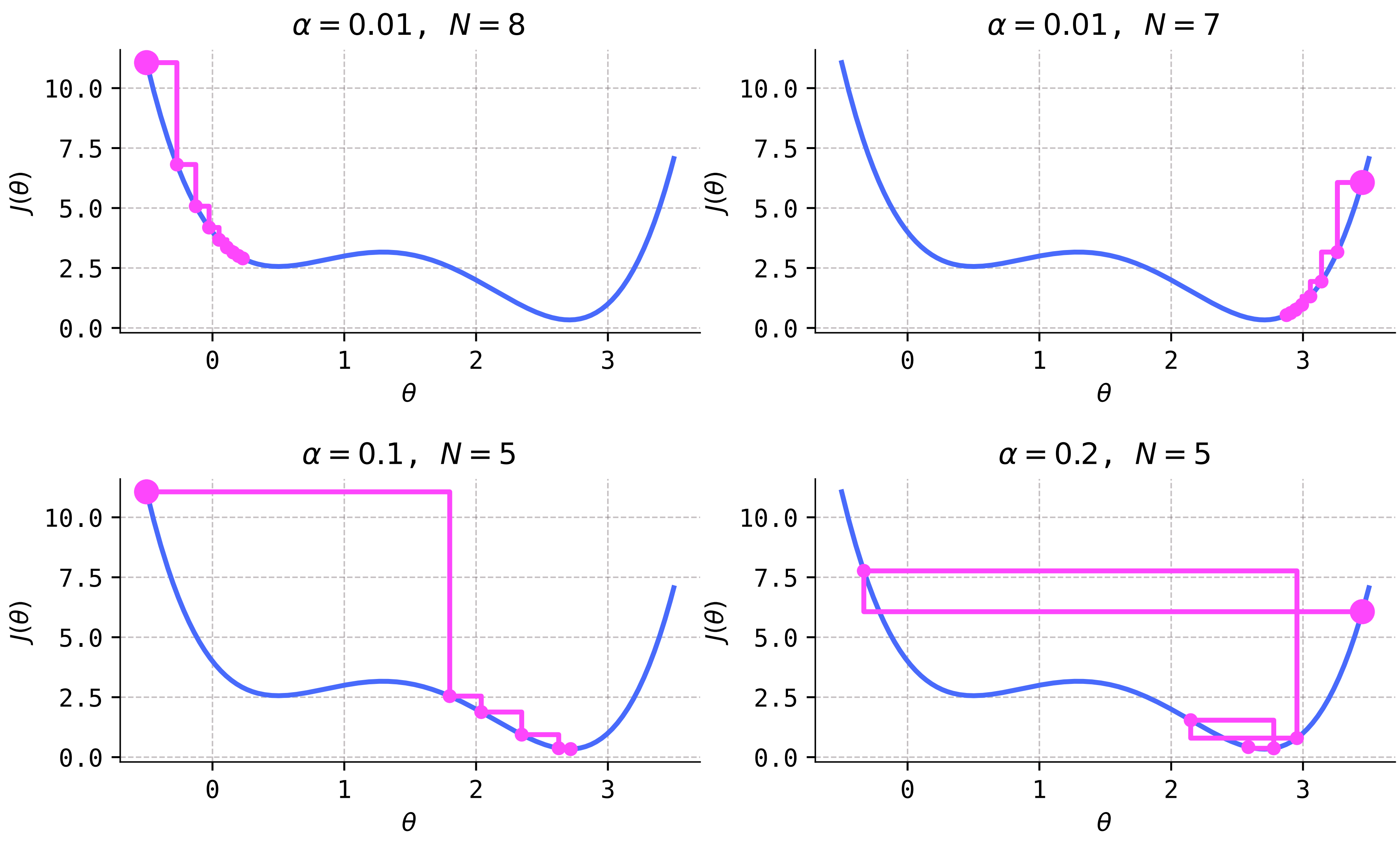
Return θ



$J'(\theta_0) < 0$
$J'(\theta_0) > 0$
θ_0
θ_0
$J(\theta)$
θ
12
10
8
6
4
2
0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

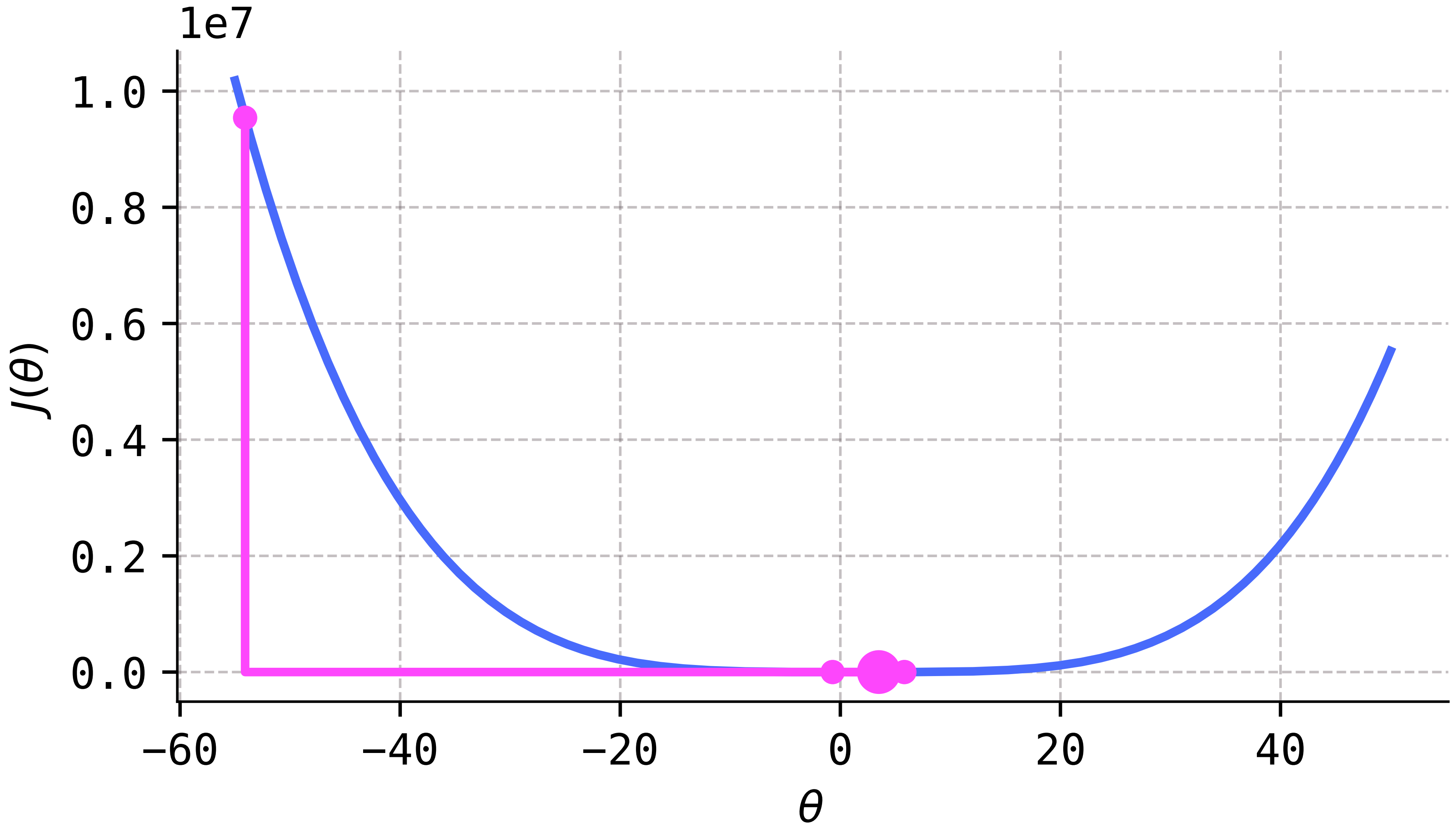
Observation 11.1

- The negative derivative $-J'(\theta)$ always "points downhill."
- When the gradient descent algorithm works, it locates a minimizer by following the negative derivative "downhill."


$\alpha = 0.01, \ N = 8$
$\alpha = 0.01, \ N = 7$
$\alpha = 0.1, \ N = 5$
$\alpha = 0.2, \ N = 5$
$J(\theta)$
θ
0.0
2.5
5.0
7.5
10.0
0
1
2
3

Problem Prompt

Do problem 1 on the worksheet.


1e7
1.0
0.8
0.6
0.4
0.2
0.0
J(θ)
−60
−40
−20
0
20
40
θ

Problem Prompt

Do problems 2 and 3 on the worksheet.

Algorithm 11.2 (Single-variable gradient descent with learning rate decay)

Input: A differentiable objective function $J : \mathbb{R} \to \mathbb{R}$, an initial guess $\theta_0 \in \mathbb{R}$ for a local minimizer $\theta^\star$, a learning rate $\alpha > 0$, a decay rate $\beta \in [0, 1)$, and the number N of gradient steps.

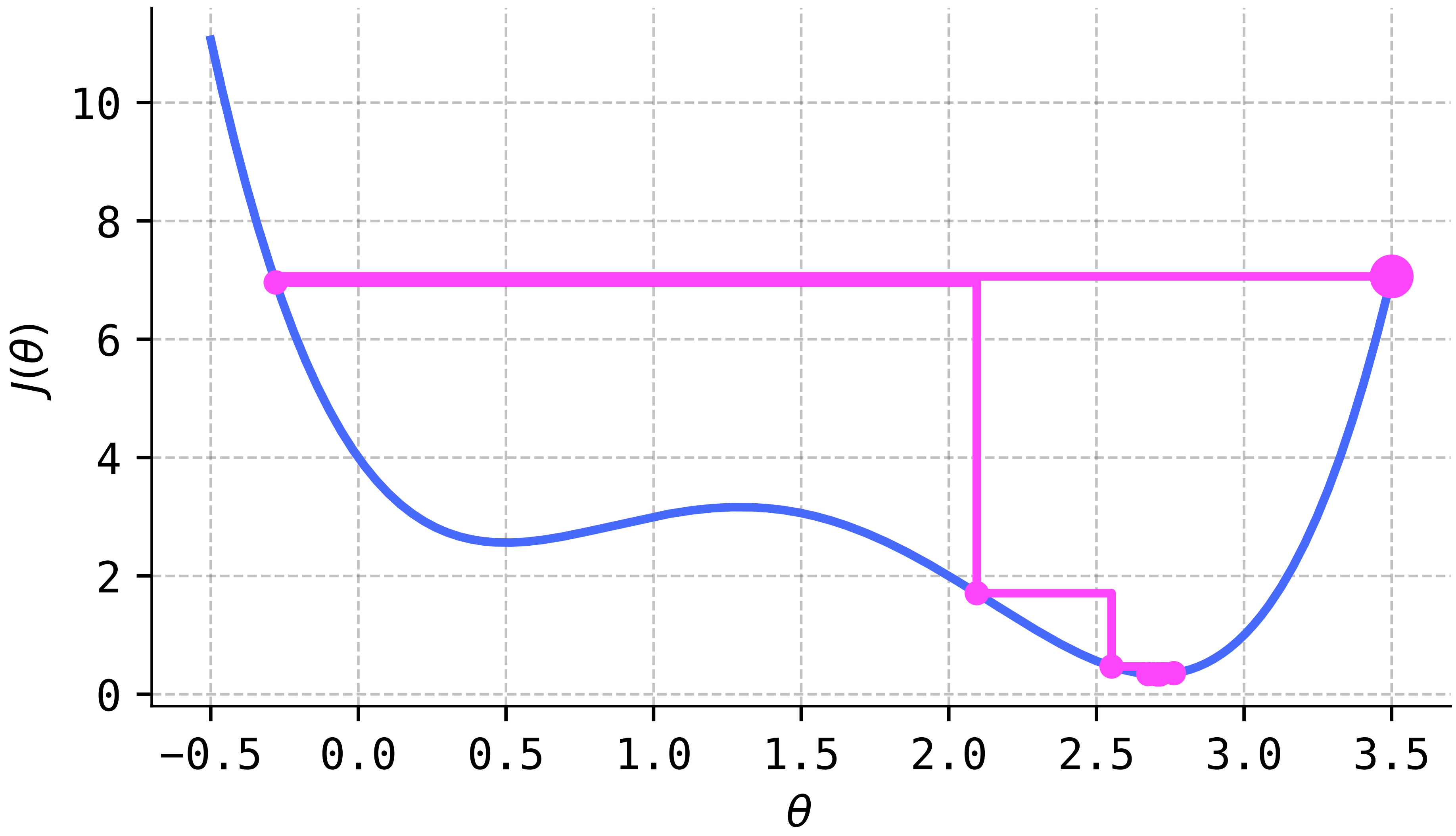
Output: An approximation to a local minimizer $\theta^\star$.

$\theta := \theta_0$

For t from 0 to $N - 1$, do:

$\qquad \theta := \theta - \alpha(1 - \beta)^{t+1} J'(\theta)$

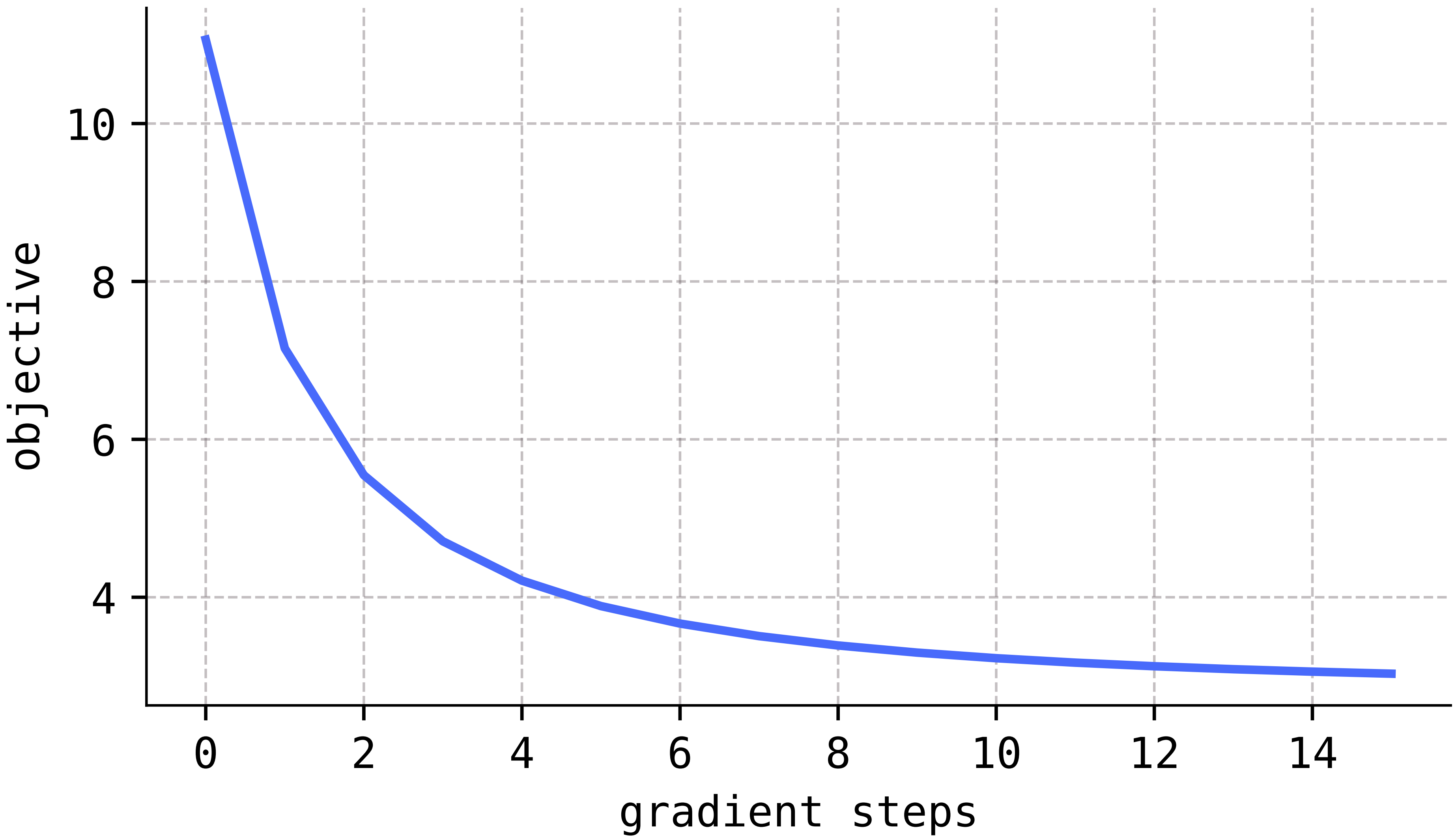
Return θ



10
8
6
4
2
0
$J(\theta)$
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
θ

Problem Prompt

Do problem 4 on the worksheet.


objective
10
8
6
4
0
2
4
6
8
10
12
14
gradient steps