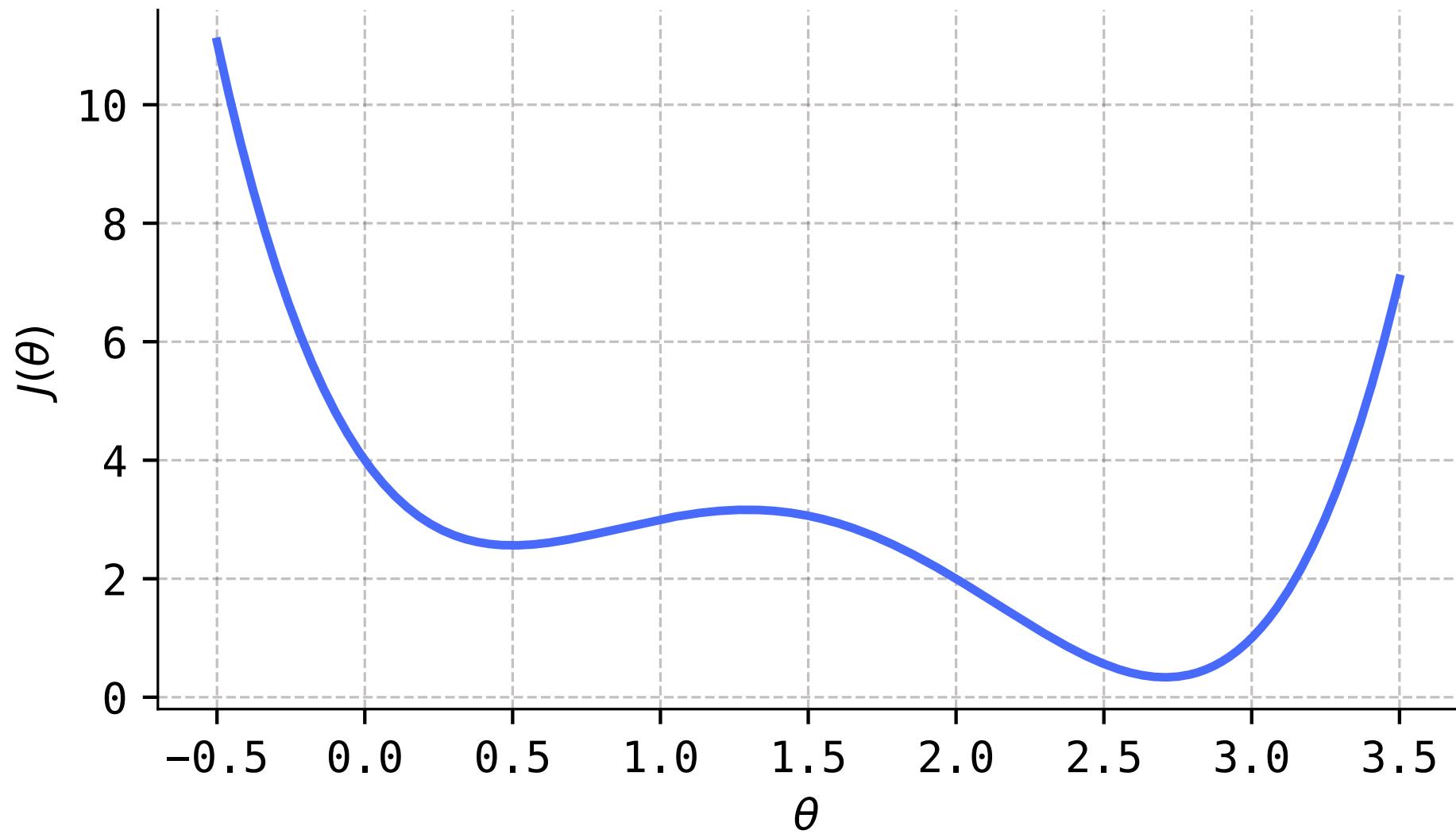


# 11. Optimization

## **11.1. Gradient descent in one variable**





## Definition 11.1

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

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

for all  $\boldsymbol{\theta}$  in a neighborhood of  $\boldsymbol{\theta}^*$ ; if this inequality holds for *all*  $\boldsymbol{\theta}$ , then  $\boldsymbol{\theta}^*$  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  $J(\boldsymbol{\theta}^*)$  of the function where  $\boldsymbol{\theta}^*$  is an extremizer are called *extrema* or *extreme values*.

### Algorithm 11.1 (Single-variable gradient descent)

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

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

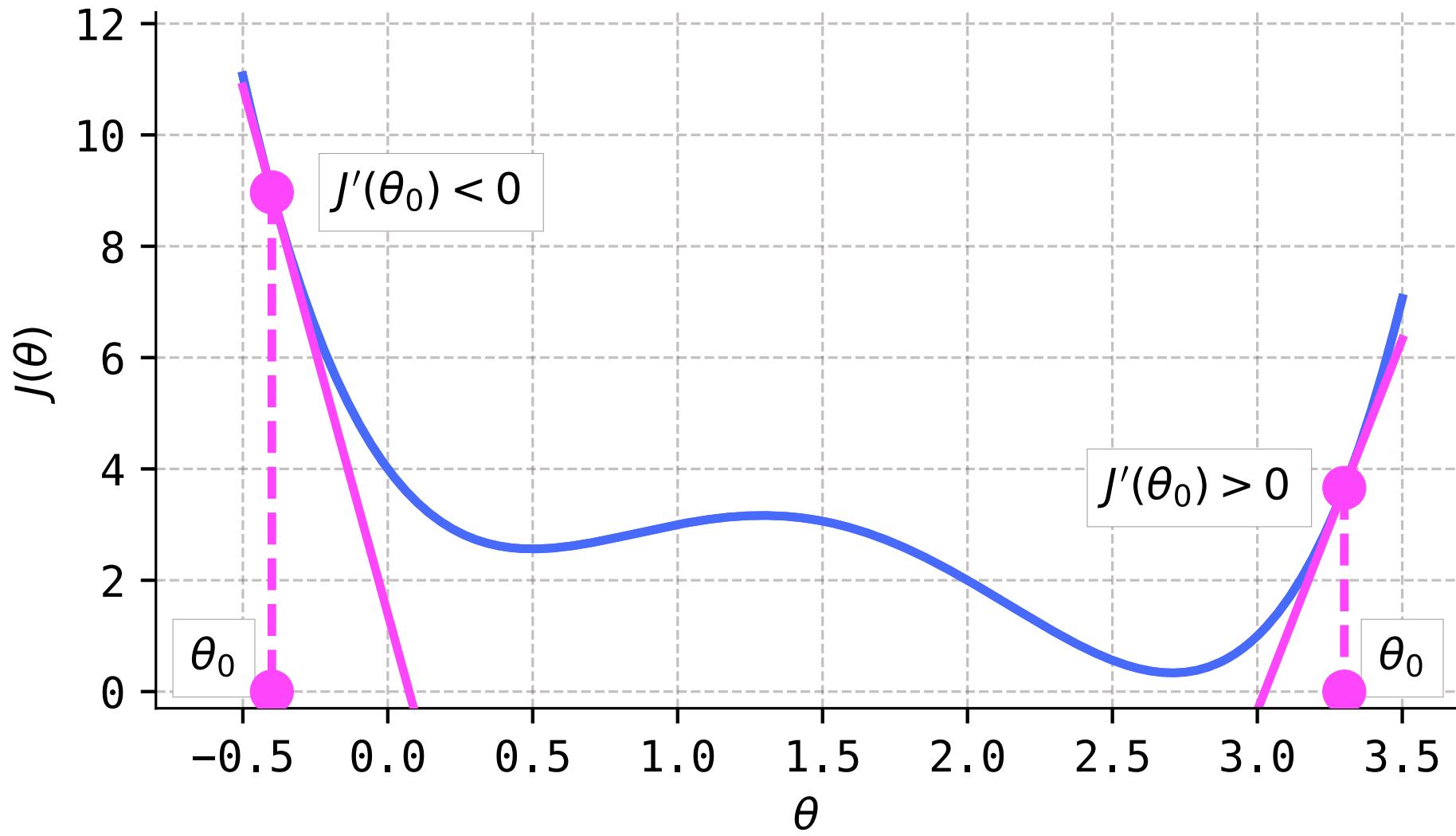
---

$$\theta := \theta_0$$

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

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

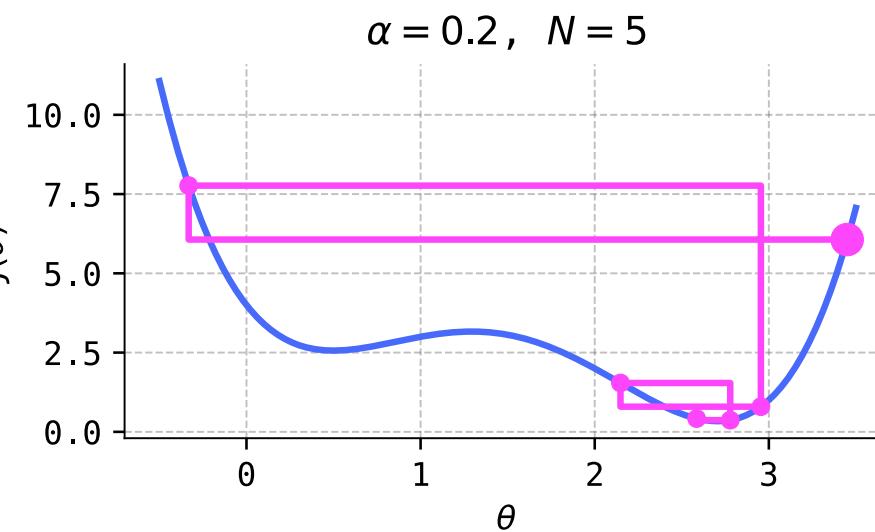
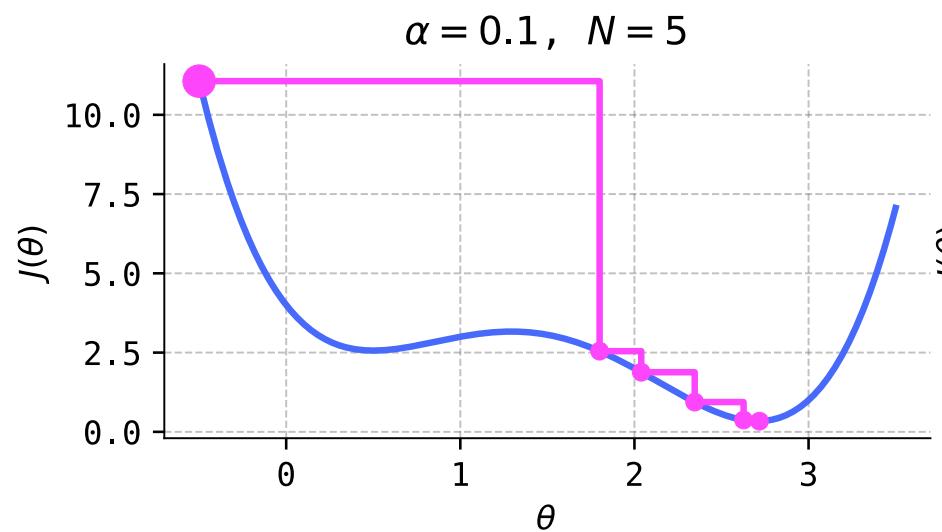
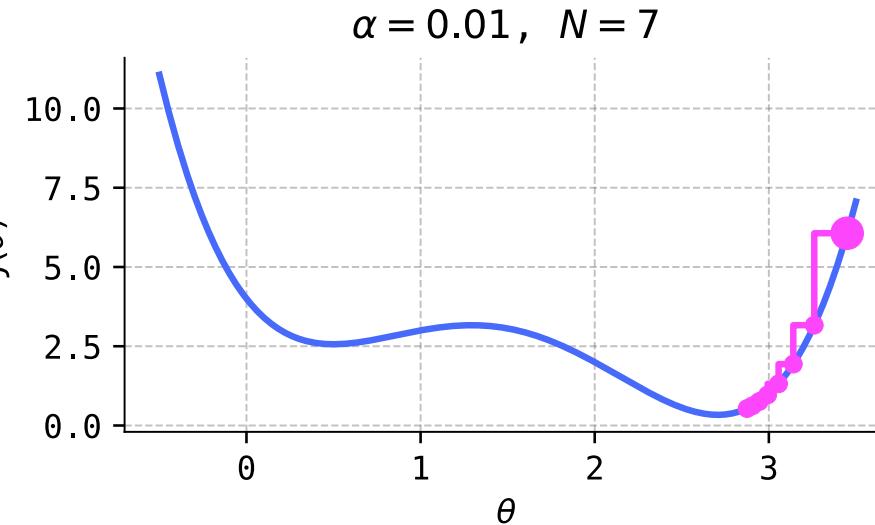
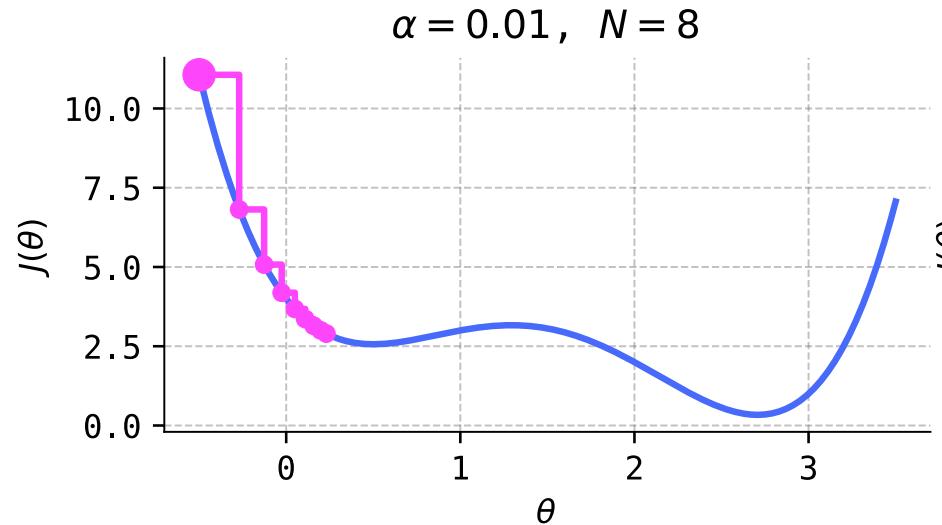
Return  $\theta$





### Observation 11.1

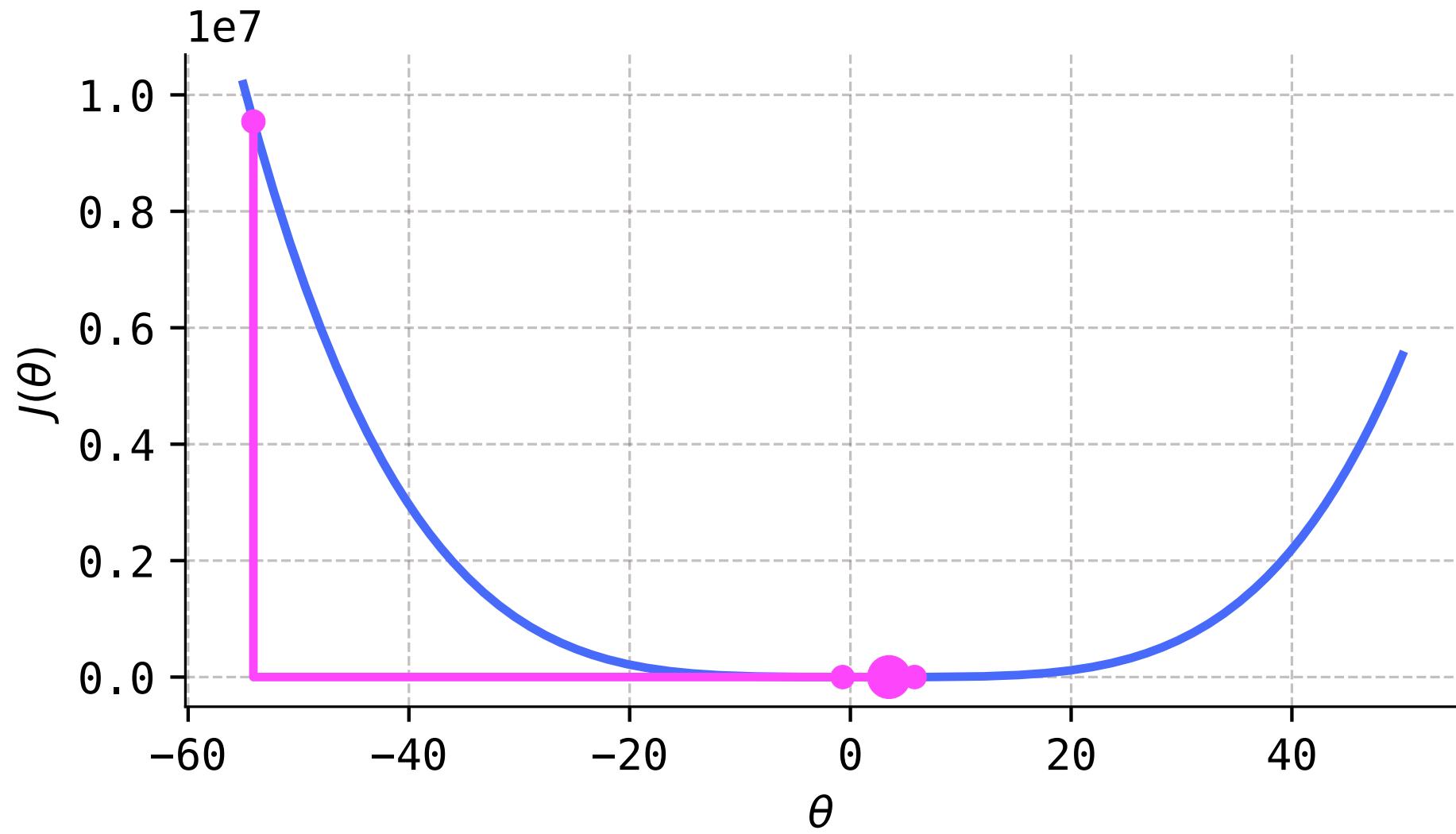
- The negative derivative  $-J'(\theta)$  always “points downhill.”
- When the gradient descent algorithm works, it locates a minimizer by following the negative derivative “downhill.”





### Problem Prompt

Do problem 1 on the worksheet.





### 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  $J : \mathbb{R} \rightarrow \mathbb{R}$ , an initial guess  $\theta_0 \in \mathbb{R}$  for a local minimizer  $\theta^*$ , 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  $\theta^*$ .

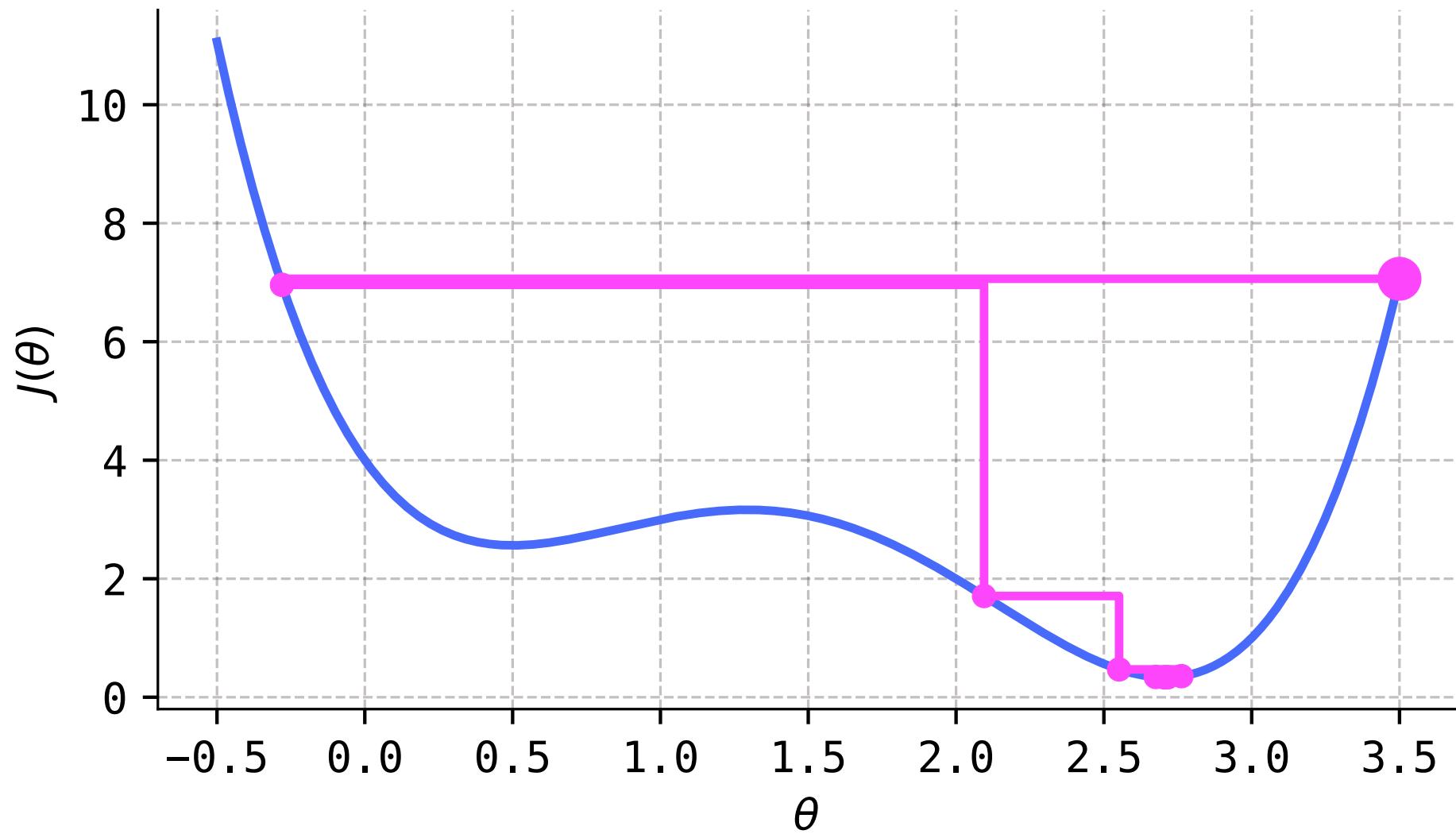
---

$$\theta := \theta_0$$

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

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

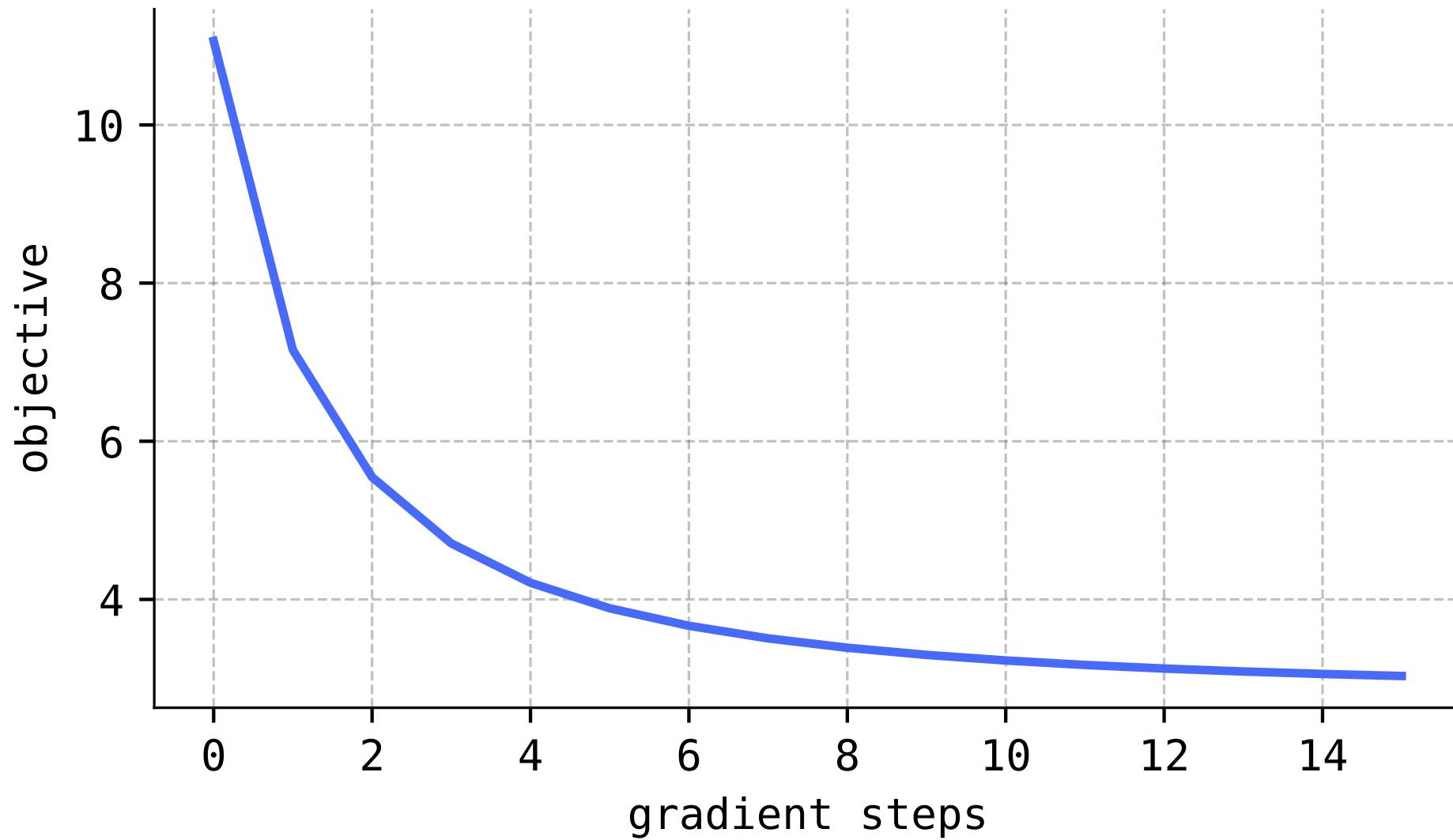
Return  $\theta$





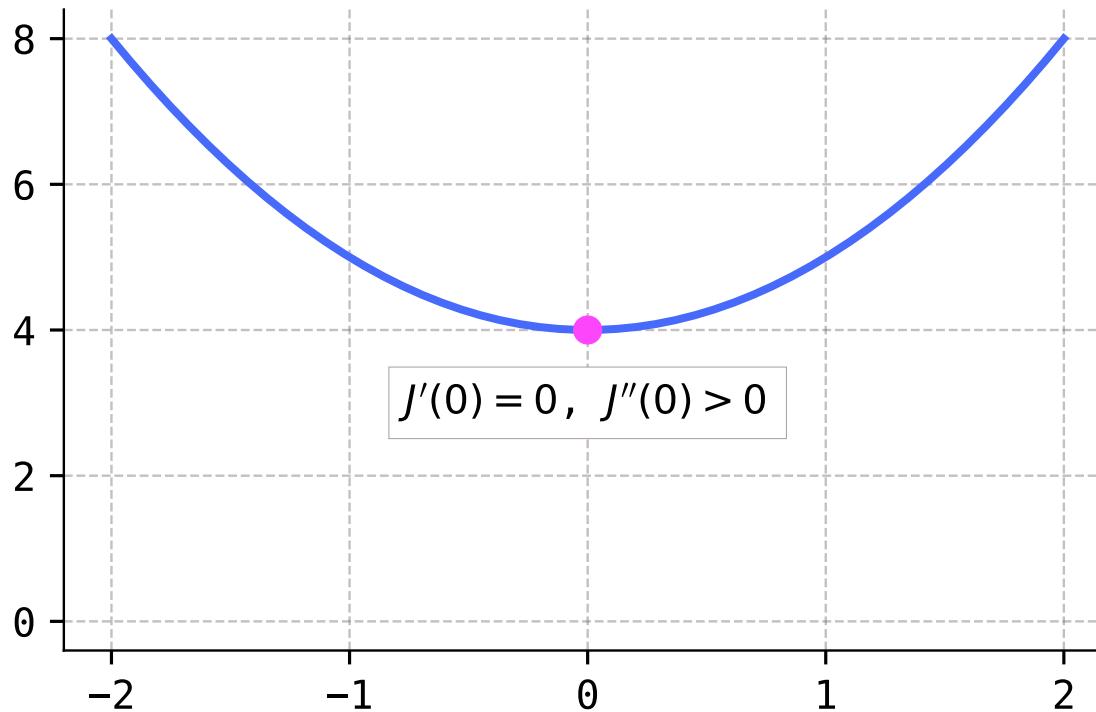
### Problem Prompt

Do problem 4 on the worksheet.

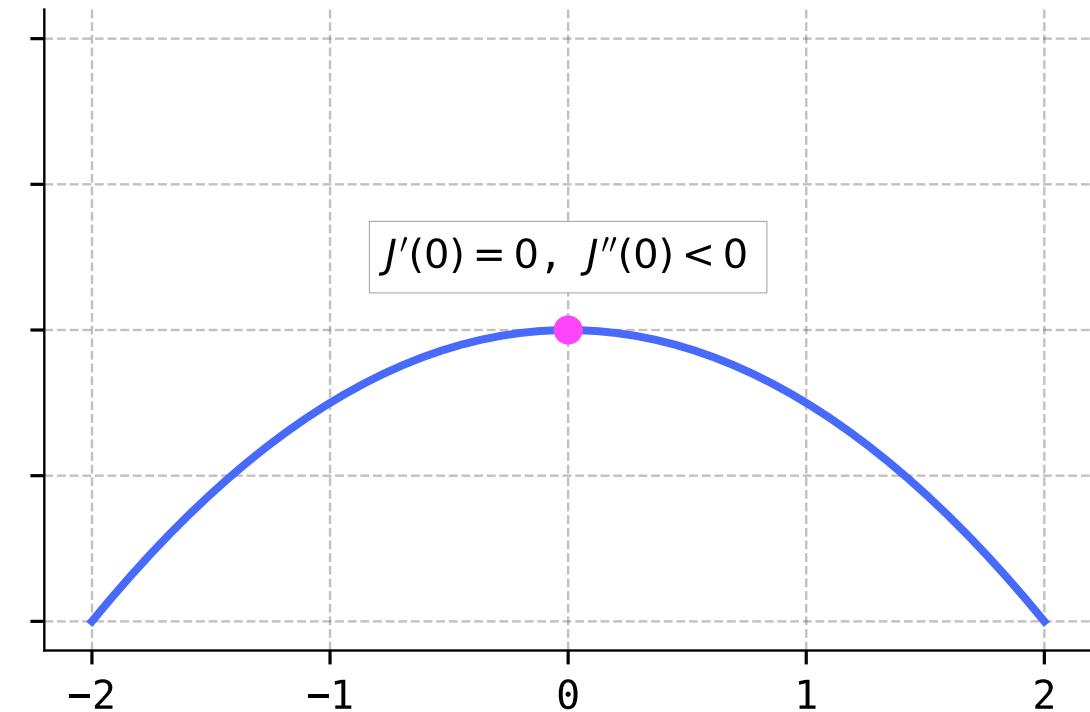


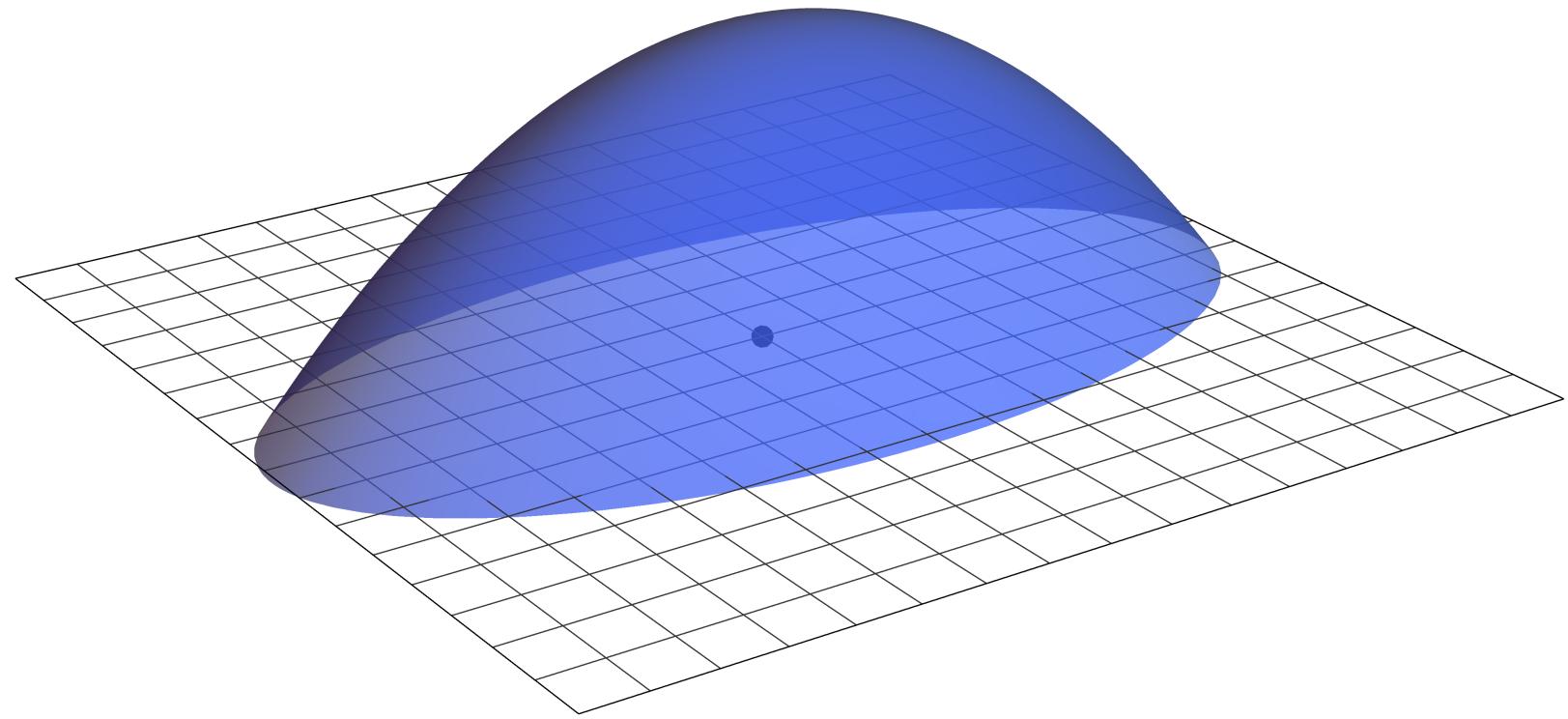
## 11.2. Differential geometry

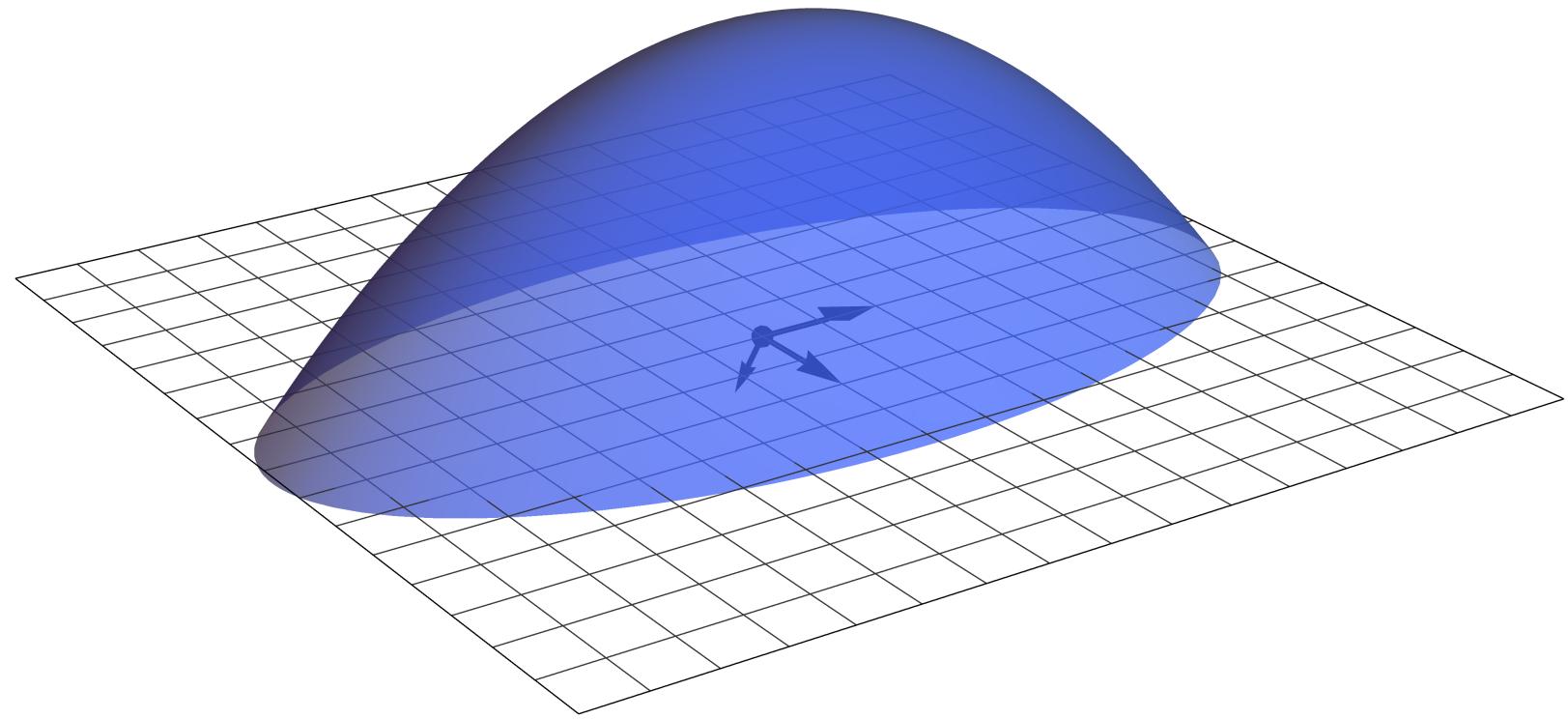
convex  $\Rightarrow$  minimizer

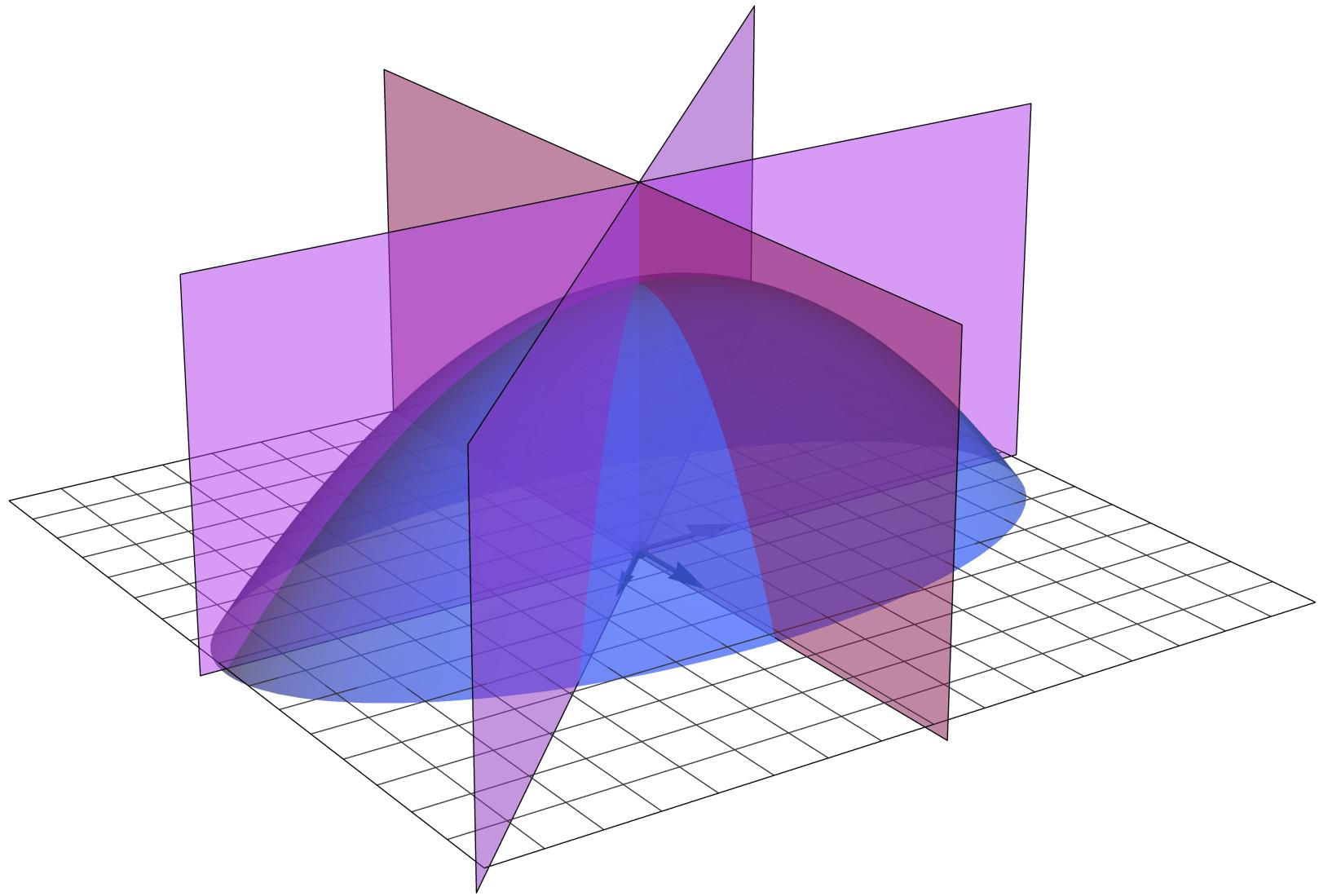


concave  $\Rightarrow$  maximizer









## Definition 11.2

Let  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  be a function of class  $C^2$ ,  $\theta \in \mathbb{R}^n$  a point, and  $\mathbf{v} \in \mathbb{R}^n$  a vector. We define the *directional first derivative of  $J$  at  $\theta$  in the direction  $\mathbf{v}$*  to be

$$J'_{\mathbf{v}}(\theta) \stackrel{\text{def}}{=} \frac{d}{dt} \Big|_{t=0} J(t\mathbf{v} + \theta),$$

while we define the *directional second derivative* to be

$$J''_{\mathbf{v}}(\theta) \stackrel{\text{def}}{=} \frac{d^2}{dt^2} \Big|_{t=0} J(t\mathbf{v} + \theta).$$

In this context, the vector  $\mathbf{v}$  is called the *directional vector*.



### Problem Prompt

Do problem 5 on the worksheet.

### 🔔 Definition 11.3

Let  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  be a function of class  $C^2$  and  $\boldsymbol{\theta} \in \mathbb{R}^n$  a point. We define the *gradient vector* to be

$$\nabla J(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \left[ \frac{\partial J}{\partial \theta_i}(\boldsymbol{\theta}) \right] = \begin{bmatrix} \frac{\partial J}{\partial \theta_1}(\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial J}{\partial \theta_n}(\boldsymbol{\theta}) \end{bmatrix} \in \mathbb{R}^n,$$

while we define the the *Hessian matrix* to be

$$\nabla^2 J(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \left[ \frac{\partial^2 J}{\partial \theta_i \partial \theta_j}(\boldsymbol{\theta}) \right] = \begin{bmatrix} \frac{\partial^2 J}{\partial \theta_1^2}(\boldsymbol{\theta}) & \cdots & \frac{\partial^2 J}{\partial \theta_1 \partial \theta_n}(\boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 J}{\partial \theta_n \partial \theta_1}(\boldsymbol{\theta}) & \cdots & \frac{\partial^2 J}{\partial \theta_n^2}(\boldsymbol{\theta}) \end{bmatrix} \in \mathbb{R}^{n \times n}.$$



### Theorem 11.1 (Slopes, curvatures, and partial derivatives)

Let  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  be a function of class  $C^2$ ,  $\theta \in \mathbb{R}^n$  a point, and  $\mathbf{v} \in \mathbb{R}^n$  a directional vector.

1. We have

$$J'_{\mathbf{v}}(\theta) = \mathbf{v}^\top \nabla J(\theta).$$

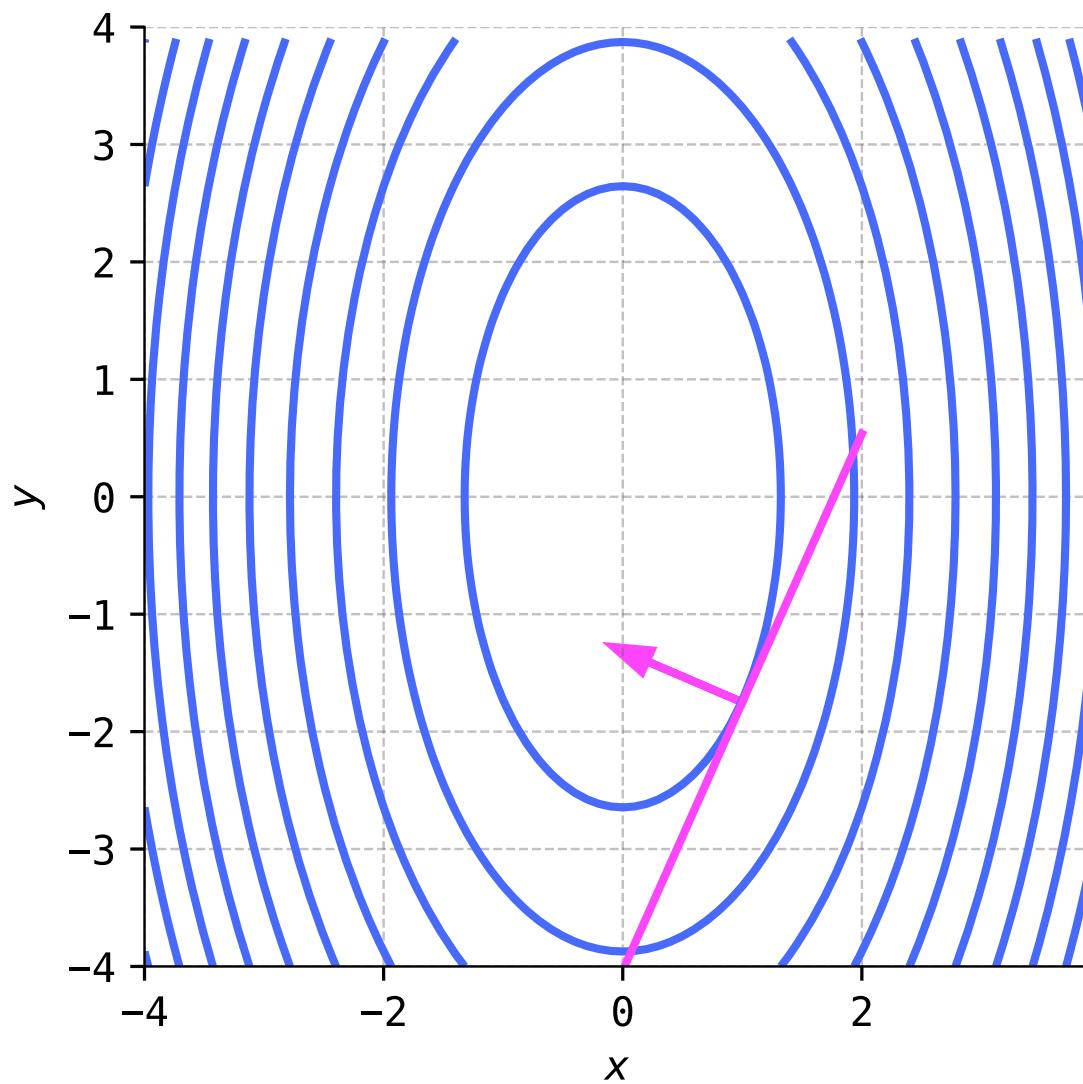
2. We have

$$J''_{\mathbf{v}}(\theta) = \mathbf{v}^\top \nabla^2 J(\theta) \mathbf{v}.$$



### Problem Prompt

Do problem 6 on the worksheet.





### Theorem 11.2 (Properties of gradient vectors)

Let  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  be a function of class  $C^2$ ,  $\theta \in \mathbb{R}^n$  a point, and suppose the gradient vector  $\nabla J(\theta)$  is nonzero.

1. The gradient vector  $\nabla J(\theta)$  points in the direction of maximum rate of change.
2. The negative gradient vector  $-\nabla J(\theta)$  points in the direction of minimum rate of change.
3. The gradient vector  $\nabla J(\theta)$  is orthogonal to level surfaces.



### Problem Prompt

Do problem 7 on the worksheet.



### Theorem 11.3 (Second Derivative Test)

Let  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  be a function of class  $C^2$  and  $\boldsymbol{\theta}^* \in \mathbb{R}^n$  a stationary point.

1. If the Hessian matrix  $\nabla^2 J(\boldsymbol{\theta}^*)$  is positive definite, then  $\boldsymbol{\theta}^*$  is a local minimizer.
2. If the Hessian matrix  $\nabla^2 J(\boldsymbol{\theta}^*)$  is negative definite, then  $\boldsymbol{\theta}^*$  is a local maximizer.



### Problem Prompt

Do problem 8 on the worksheet.



### Theorem 11.4 (Eigenvalues, eigenvectors, and local curvature)

Let  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  be a function of class  $C^2$  and  $\boldsymbol{\theta} \in \mathbb{R}^n$  a point with positive definite Hessian matrix  $\nabla^2 J(\boldsymbol{\theta})$ . Suppose we linearly order the eigenvalues of the Hessian matrix as

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n. \quad (11.8)$$

Then:

1. The directional curvature  $J''_{\mathbf{v}}(\boldsymbol{\theta})$  is maximized exactly when  $\mathbf{v}$  lies in the eigenspace of  $\lambda_n$ , in which case  $J''_{\mathbf{v}}(\boldsymbol{\theta}) = \lambda_n$ .
2. The directional curvature  $J''_{\mathbf{v}}(\boldsymbol{\theta})$  is minimized exactly when  $\mathbf{v}$  lies in the eigenspace of  $\lambda_1$ , in which case  $J''_{\mathbf{v}}(\boldsymbol{\theta}) = \lambda_1$ .



### Problem Prompt

Do problem 9 on the worksheet.

### Definition 11.4

Let  $\mathbf{A}$  be an  $n \times n$  square matrix.

1. The *spectrum* of  $\mathbf{A}$ , denoted  $\sigma(\mathbf{A})$ , is the set of eigenvalues of  $\mathbf{A}$ .
2. The *spectral radius* of  $\mathbf{A}$ , denoted  $\rho(\mathbf{A})$ , is given by

$$\rho(\mathbf{A}) \stackrel{\text{def}}{=} \max_{\lambda \in \sigma(\mathbf{A})} |\lambda|.$$

3. If  $\mathbf{A}$  is positive definite, the *condition number* of  $\mathbf{A}$ , denoted  $\kappa(\mathbf{A})$ , is the ratio

$$\kappa(\mathbf{A}) \stackrel{\text{def}}{=} \frac{\lambda_{\max}}{\lambda_{\min}}$$

of the largest eigenvalue of  $\mathbf{A}$  to the smallest.



### Problem Prompt

Do problem 10 on the worksheet.

## 11.3. Gradient descent in multiple variables



### Algorithm 11.3 (Multi-variable gradient descent with learning rate decay)

**Input:** A differentiable function  $J : \mathbb{R}^n \rightarrow \mathbb{R}$ , an initial guess  $\theta_0 \in \mathbb{R}^n$  for a local minimizer  $\theta^*$ , 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  $\theta^*$ .

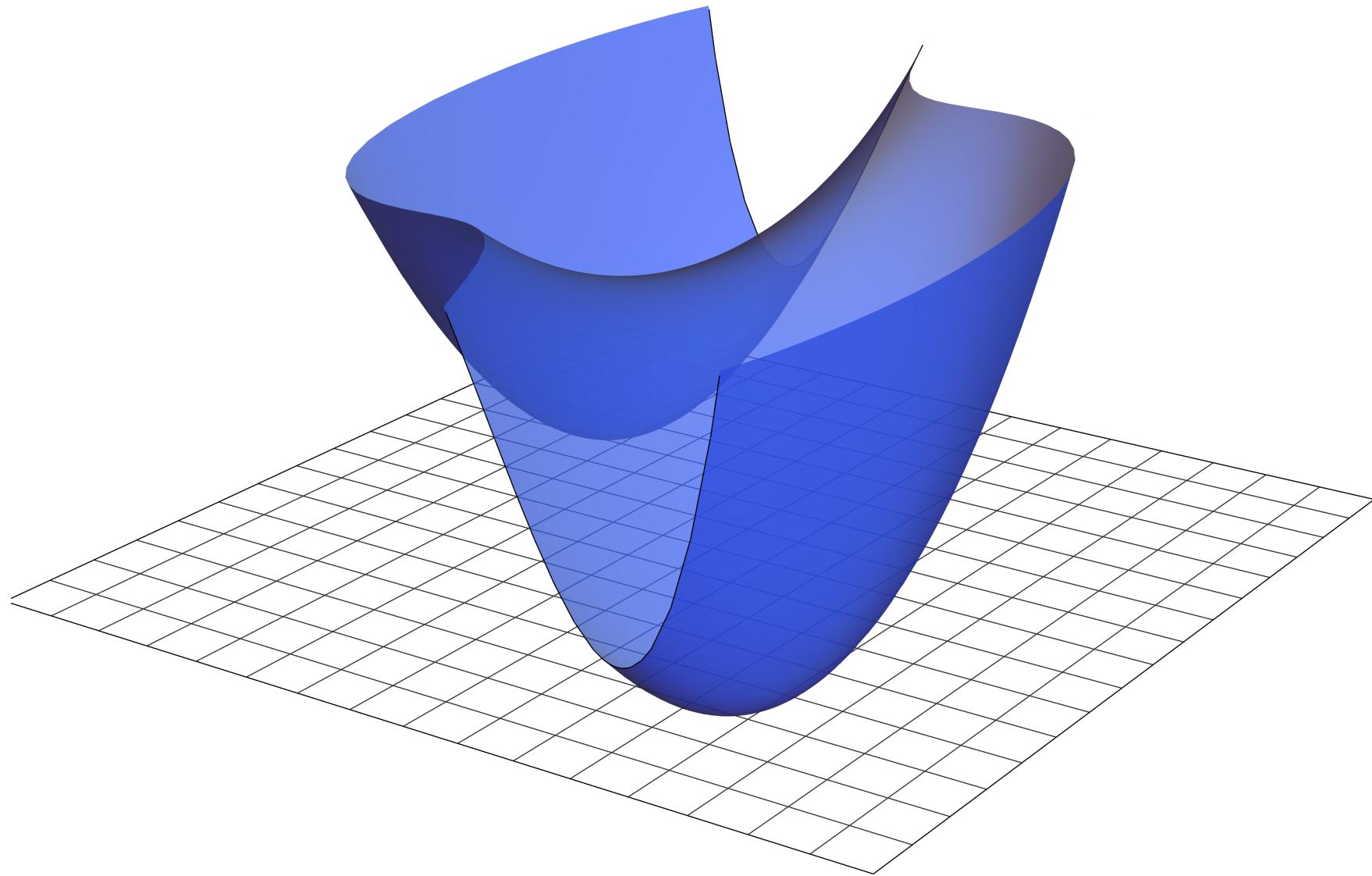
---

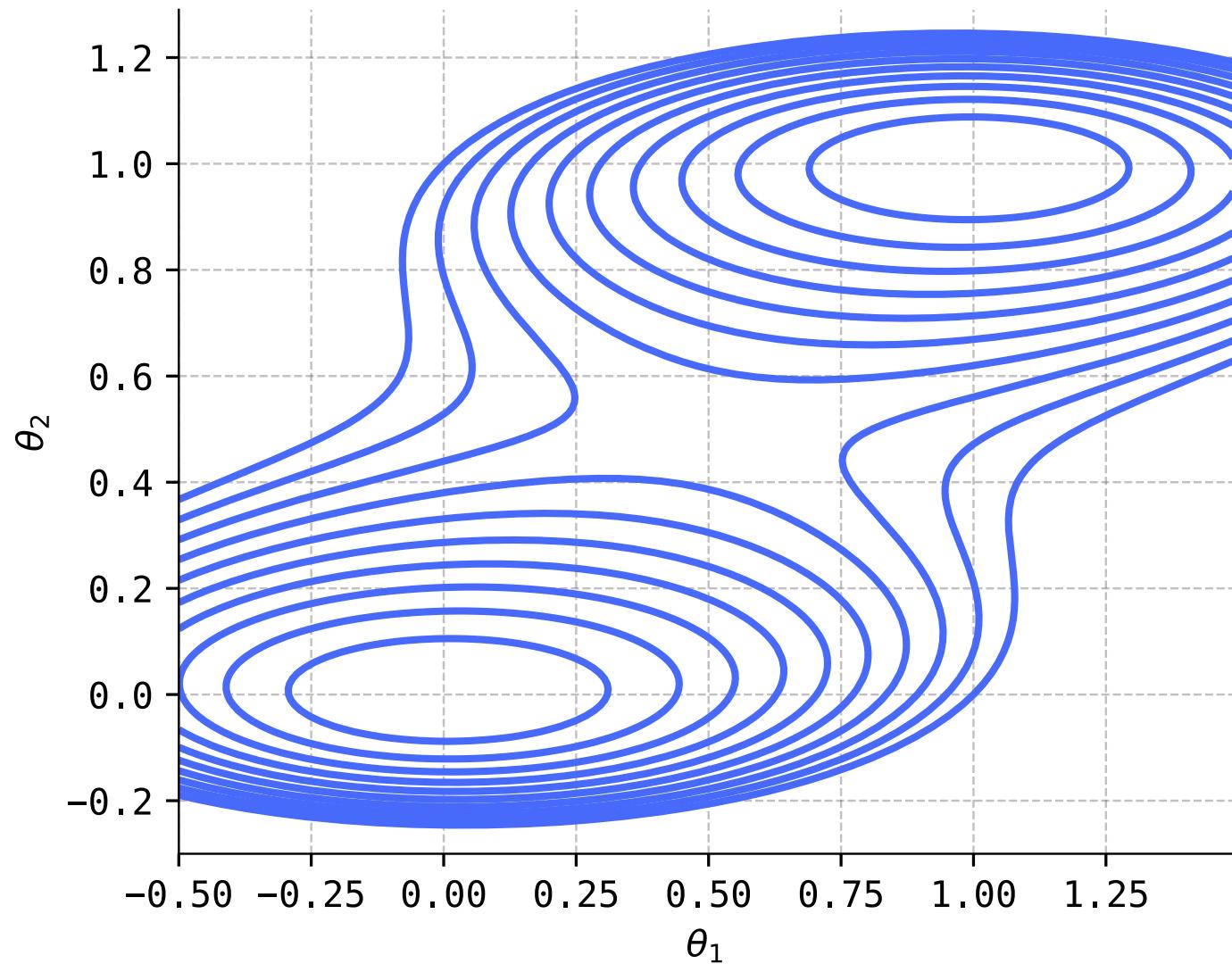
$$\theta := \theta_0$$

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

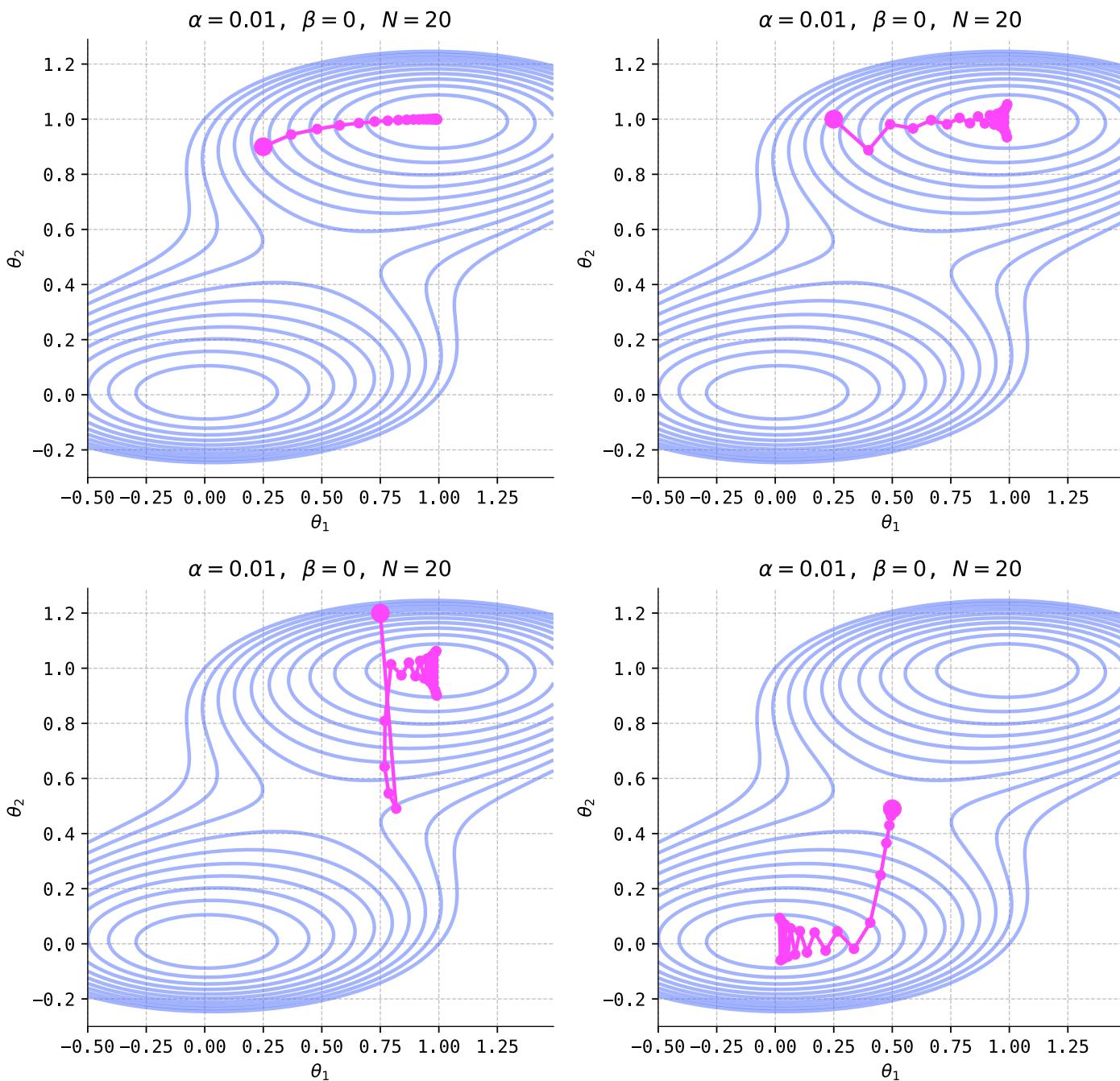
$$\theta := \theta - \alpha(1 - \beta)^{t+1} \nabla J(\theta)$$

Return  $\theta$ .

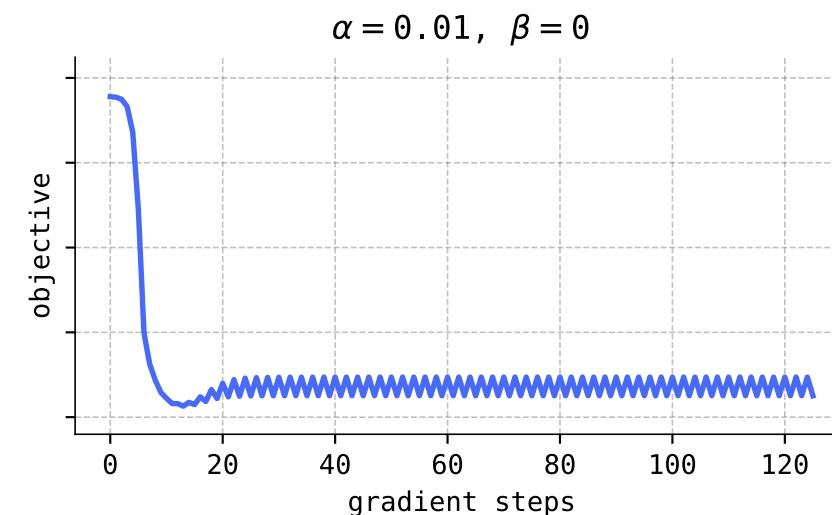
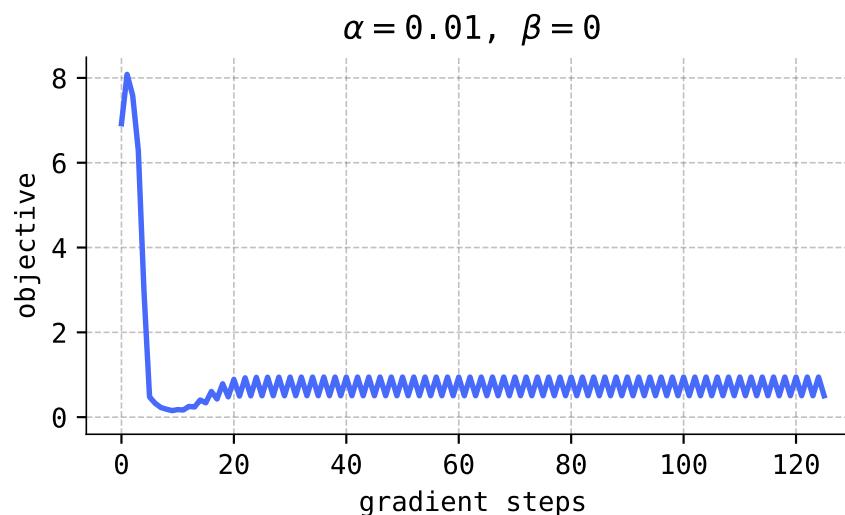
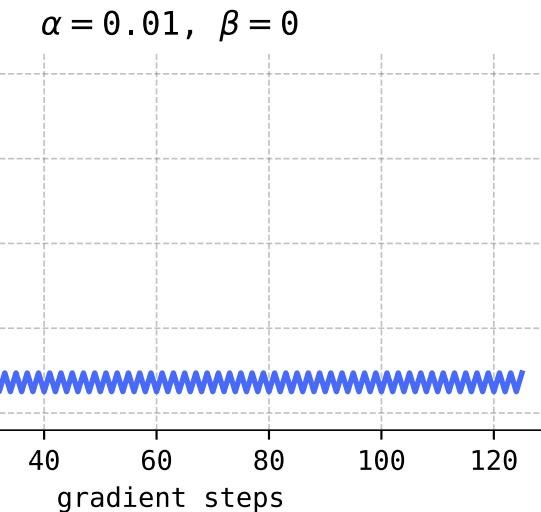
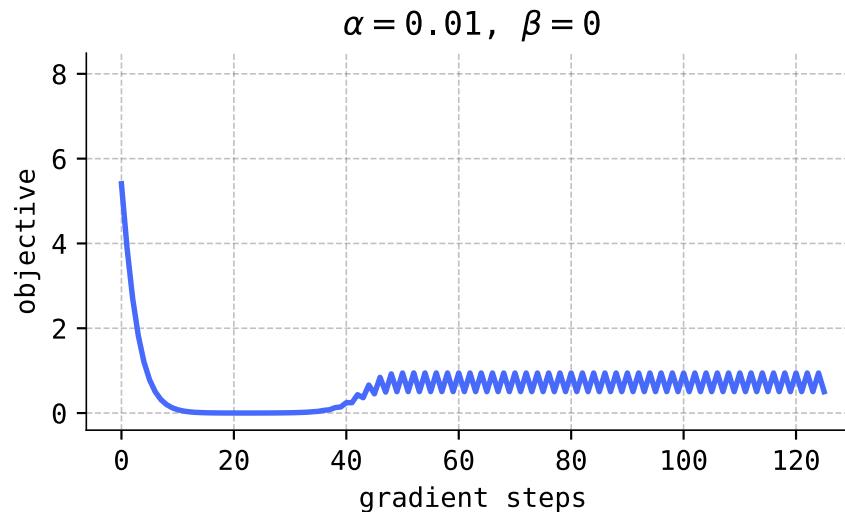




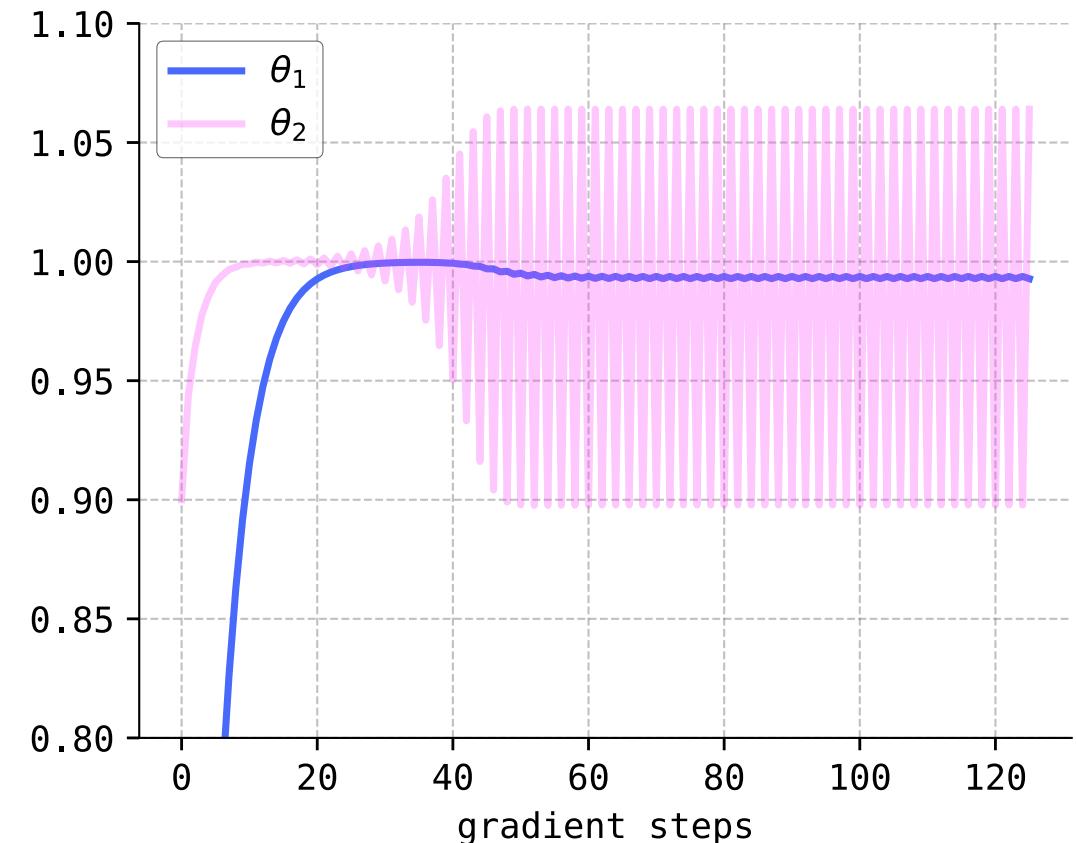
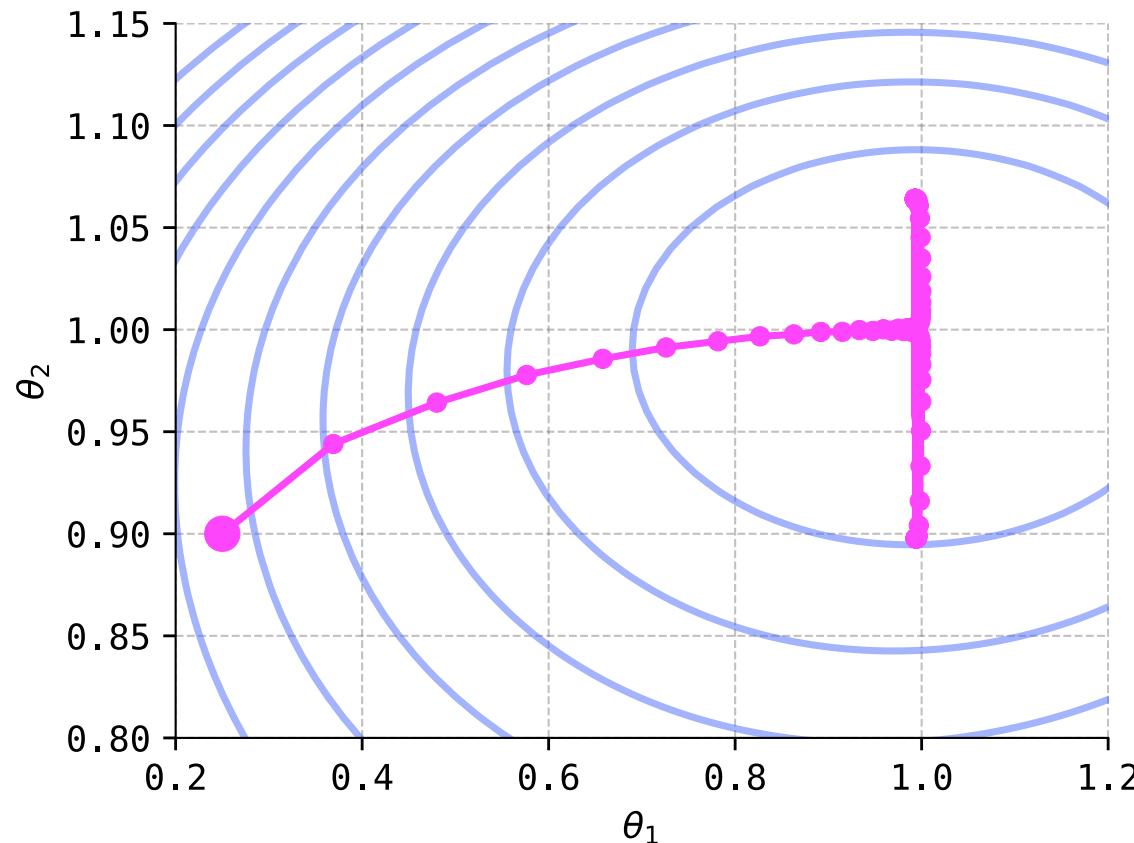
first runs of gradient descent



first runs of gradient descent



$$\alpha = 0.01, \beta = 0, N = 125$$





### Problem Prompt

Do problem 11 on the worksheet.



### Theorem 11.5 (Curvature and gradient descent)

Let  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  be a function of class  $C^2$  and  $\boldsymbol{\theta}^*$  a local minimizer with positive definite Hessian matrix  $\mathbf{H} = \nabla^2 J(\boldsymbol{\theta}^*)$  with minimum and maximum eigenvalues  $\lambda_{\min}$  and  $\lambda_{\max}$ . For initial guesses  $\boldsymbol{\theta}_0$  sufficiently near  $\boldsymbol{\theta}^*$  to allow a degree-2 Taylor polynomial approximation, the gradient descent algorithm with learning rate

$$\alpha = \frac{2}{\lambda_{\min} + \lambda_{\max}} \quad \text{and} \quad \beta = 0$$

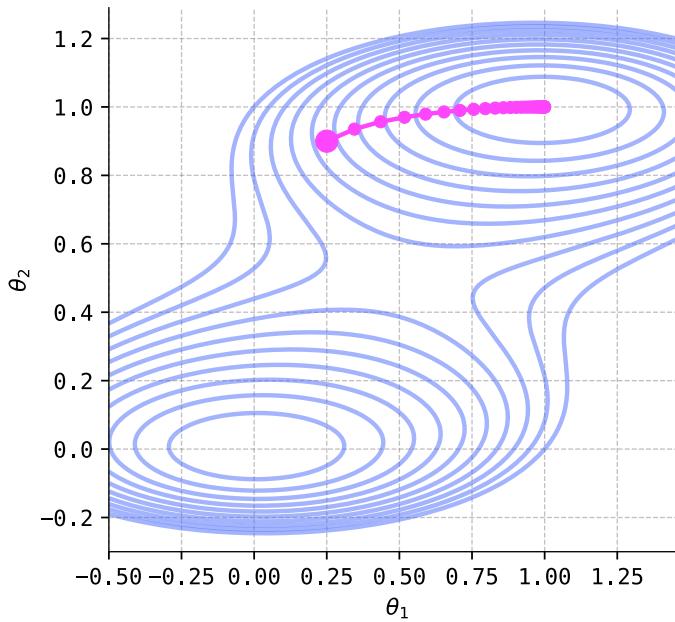
converges to  $\boldsymbol{\theta}^*$  exponentially fast, with

$$|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*| \leq \left(1 - \frac{2}{1 + \kappa(\mathbf{H})}\right)^t |\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*|$$

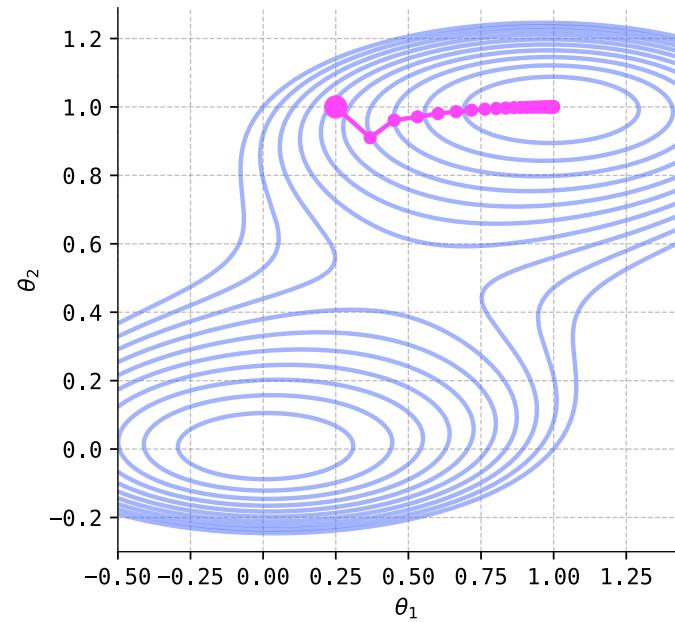
for each  $t \geq 1$ . Here,  $\kappa(\mathbf{H})$  is the condition number of  $\mathbf{H}$ .

second runs of gradient descent with smaller learning rates

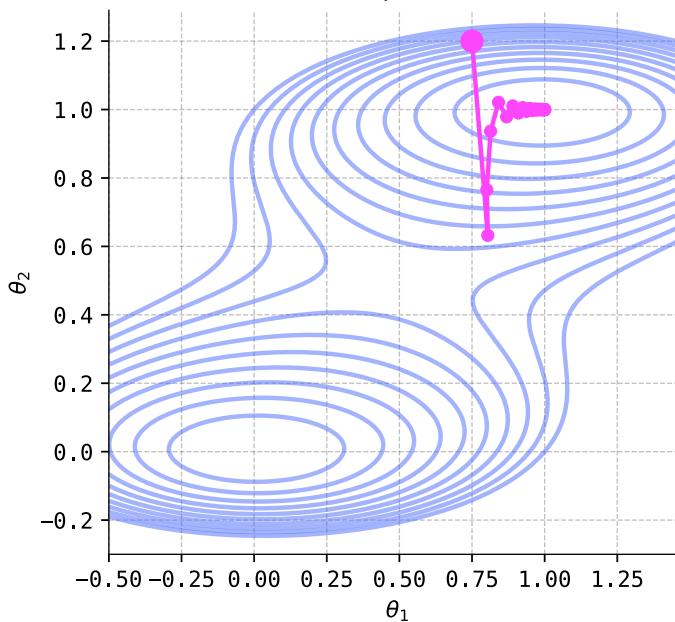
$\alpha = 0.008, \beta = 0, N = 40$



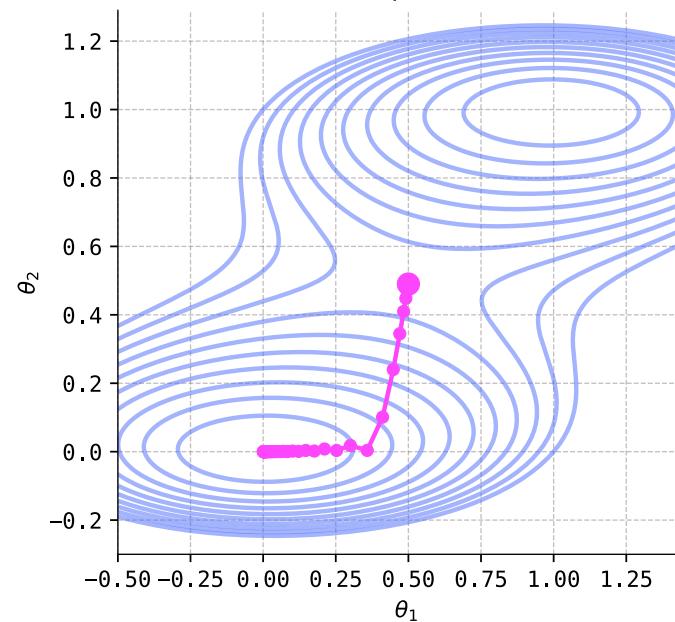
$\alpha = 0.008, \beta = 0, N = 40$



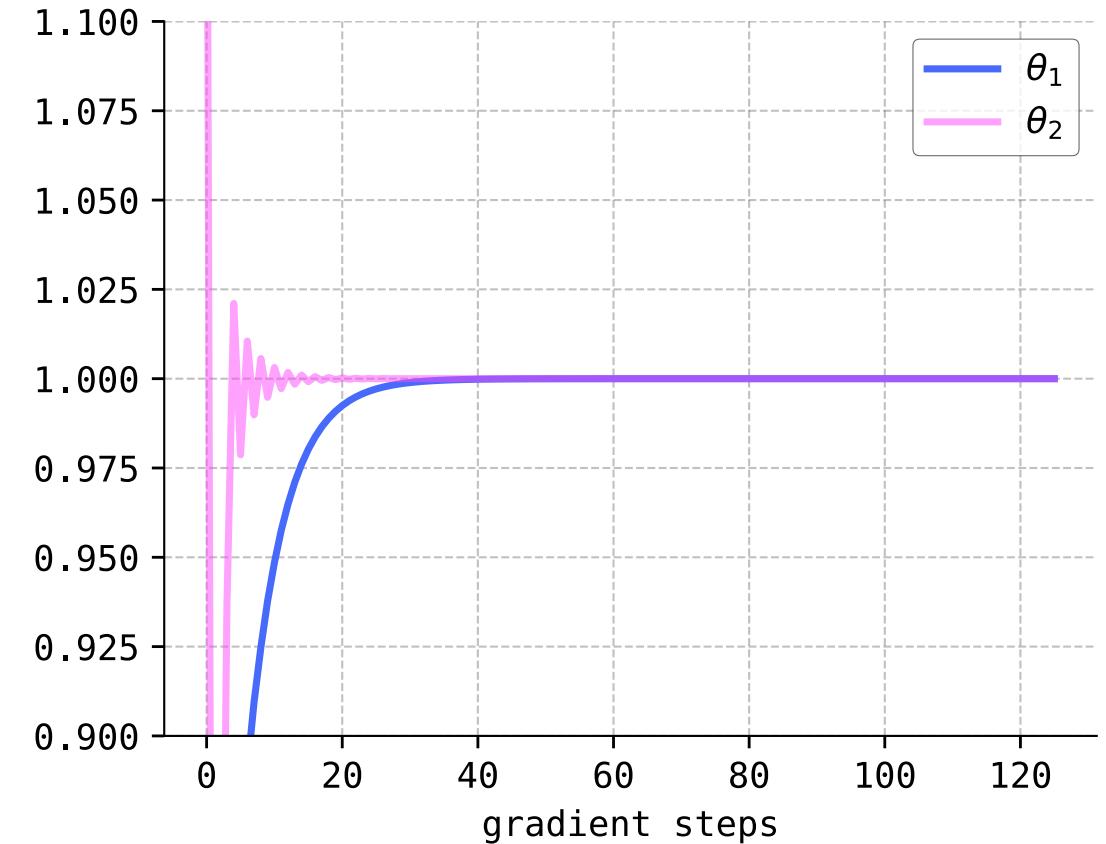
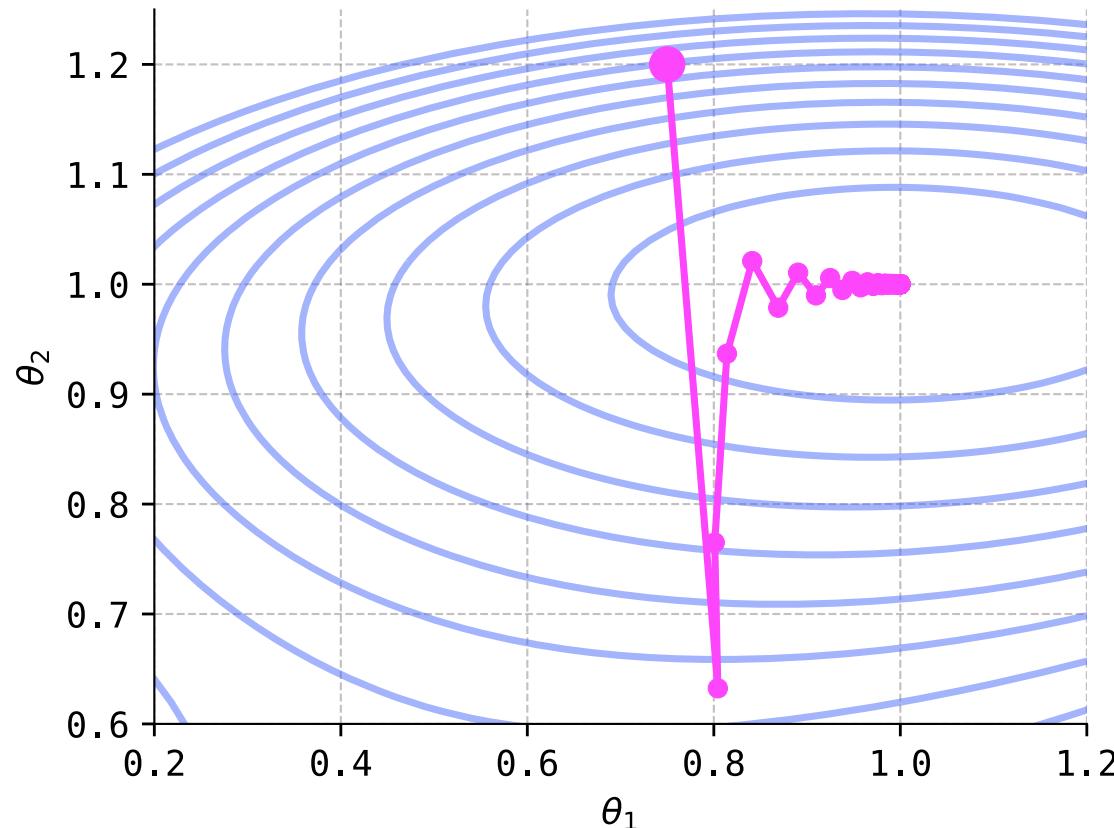
$\alpha = 0.008, \beta = 0, N = 40$



$\alpha = 0.008, \beta = 0, N = 40$

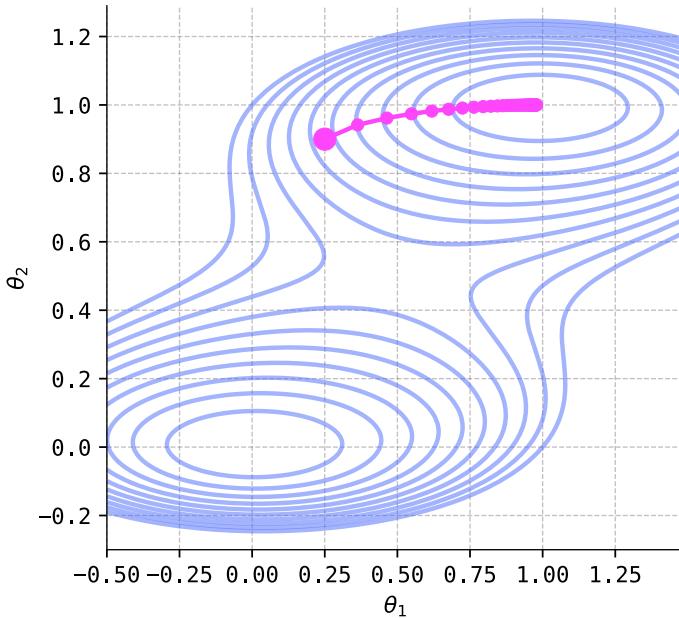


$$\alpha = 0.008, \beta = 0, N = 125$$

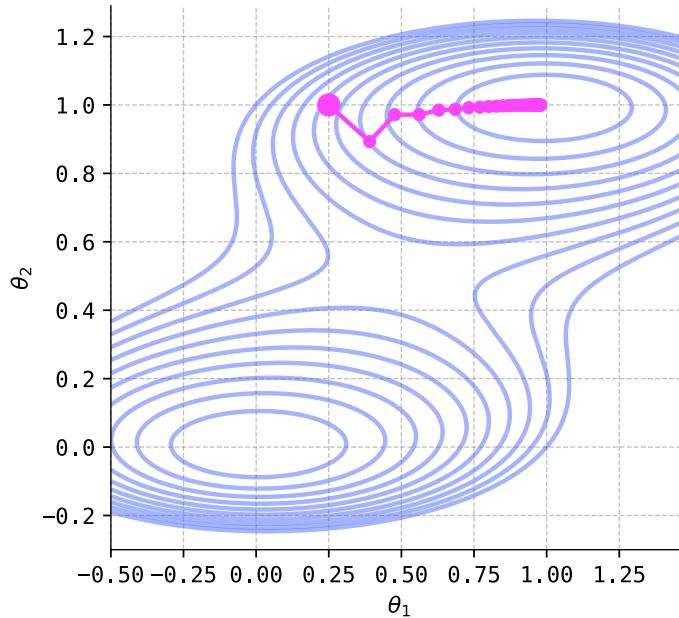


third runs of gradient descent with learning rate decay

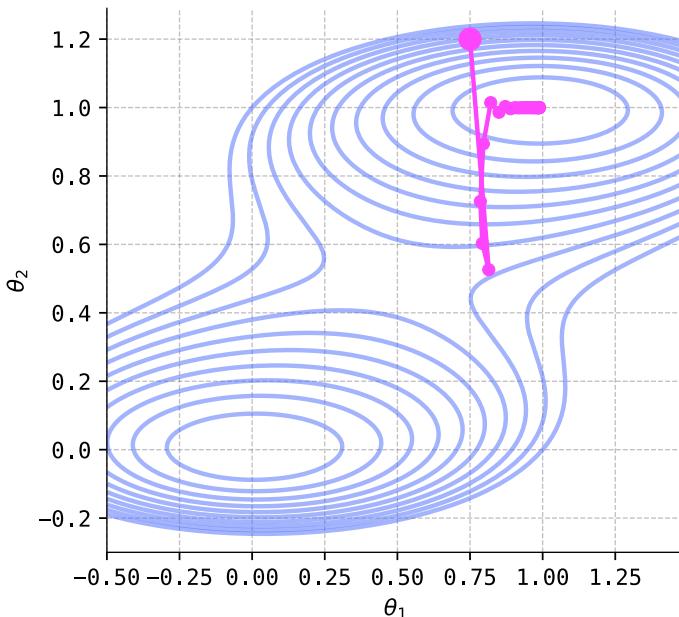
$\alpha = 0.01, \beta = 0.05, N = 40$



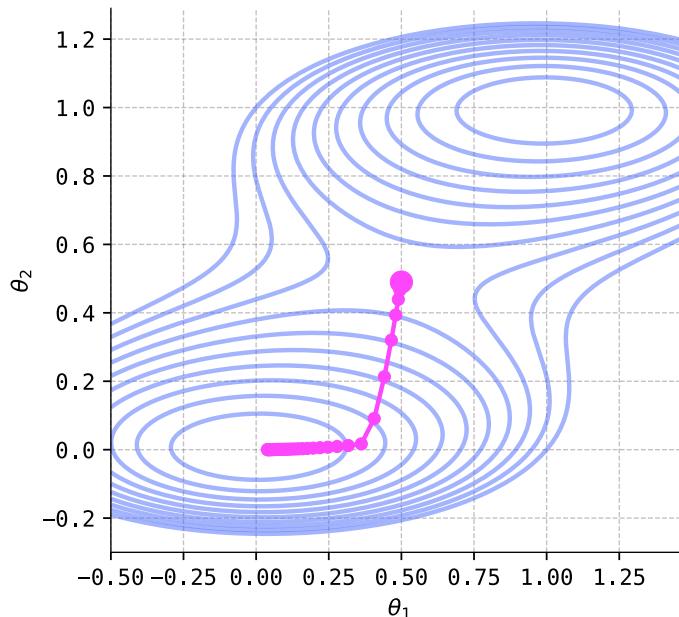
$\alpha = 0.01, \beta = 0.05, N = 40$



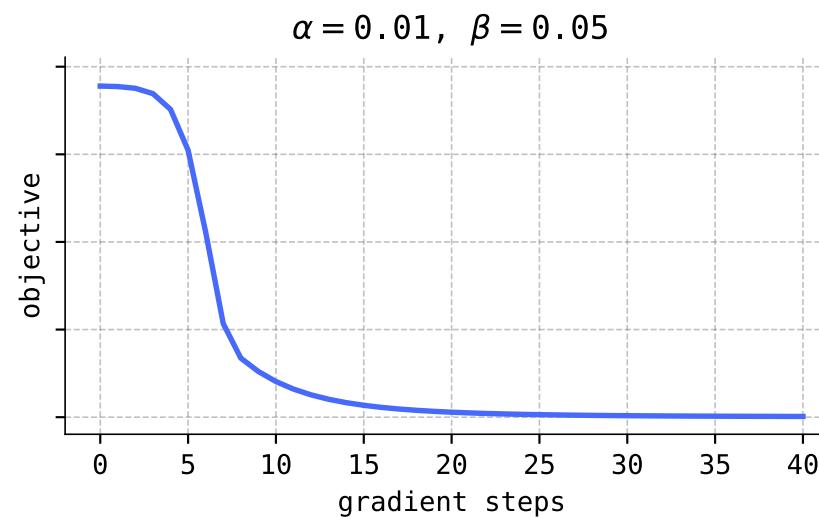
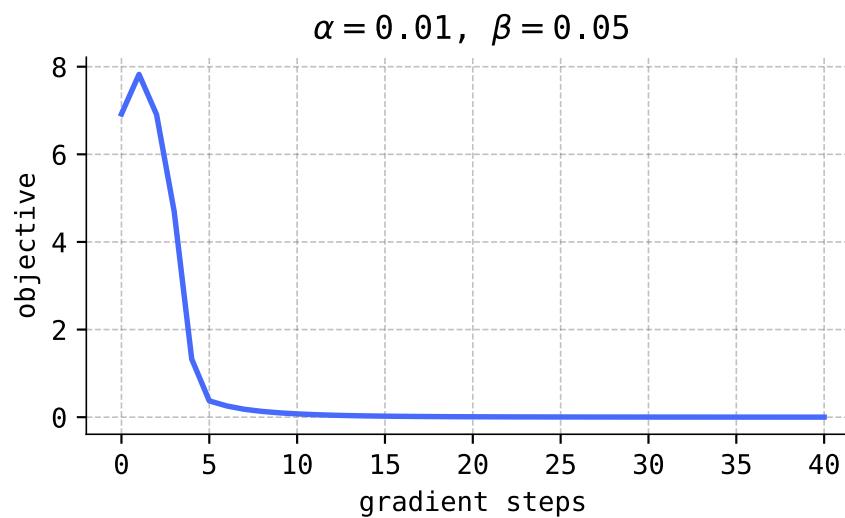
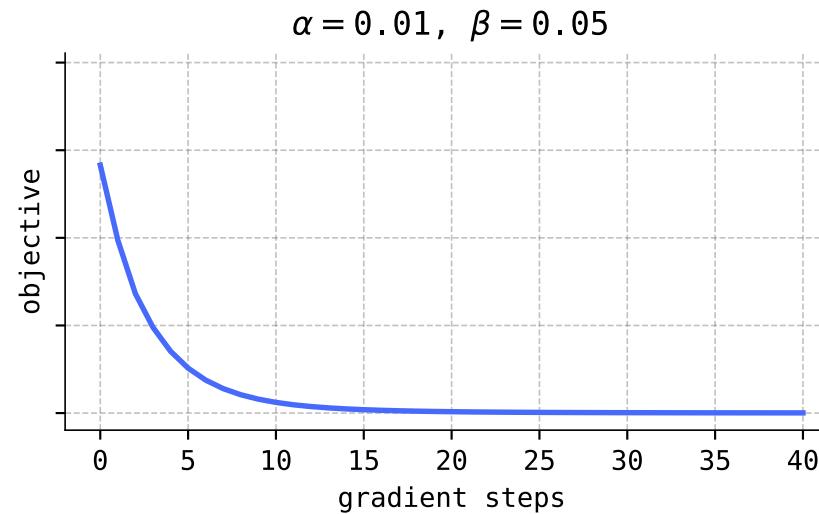
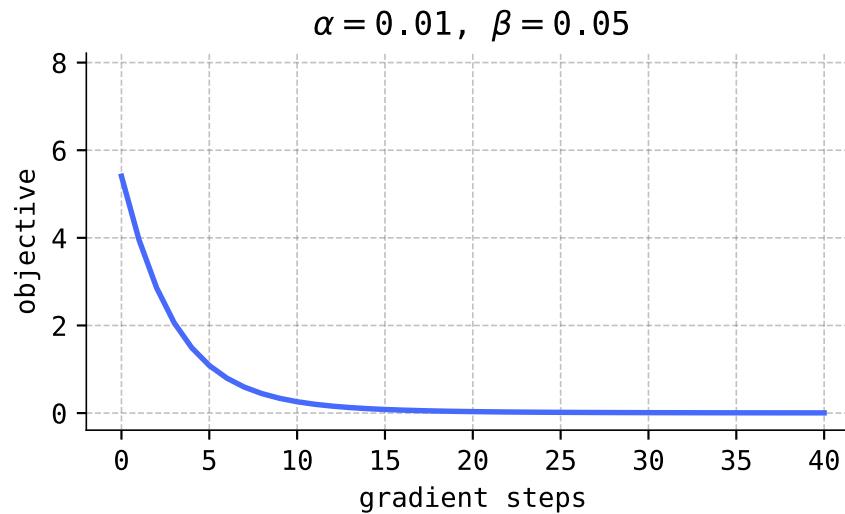
$\alpha = 0.01, \beta = 0.05, N = 40$



$\alpha = 0.01, \beta = 0.05, N = 40$



third runs of gradient descent with learning rate decay

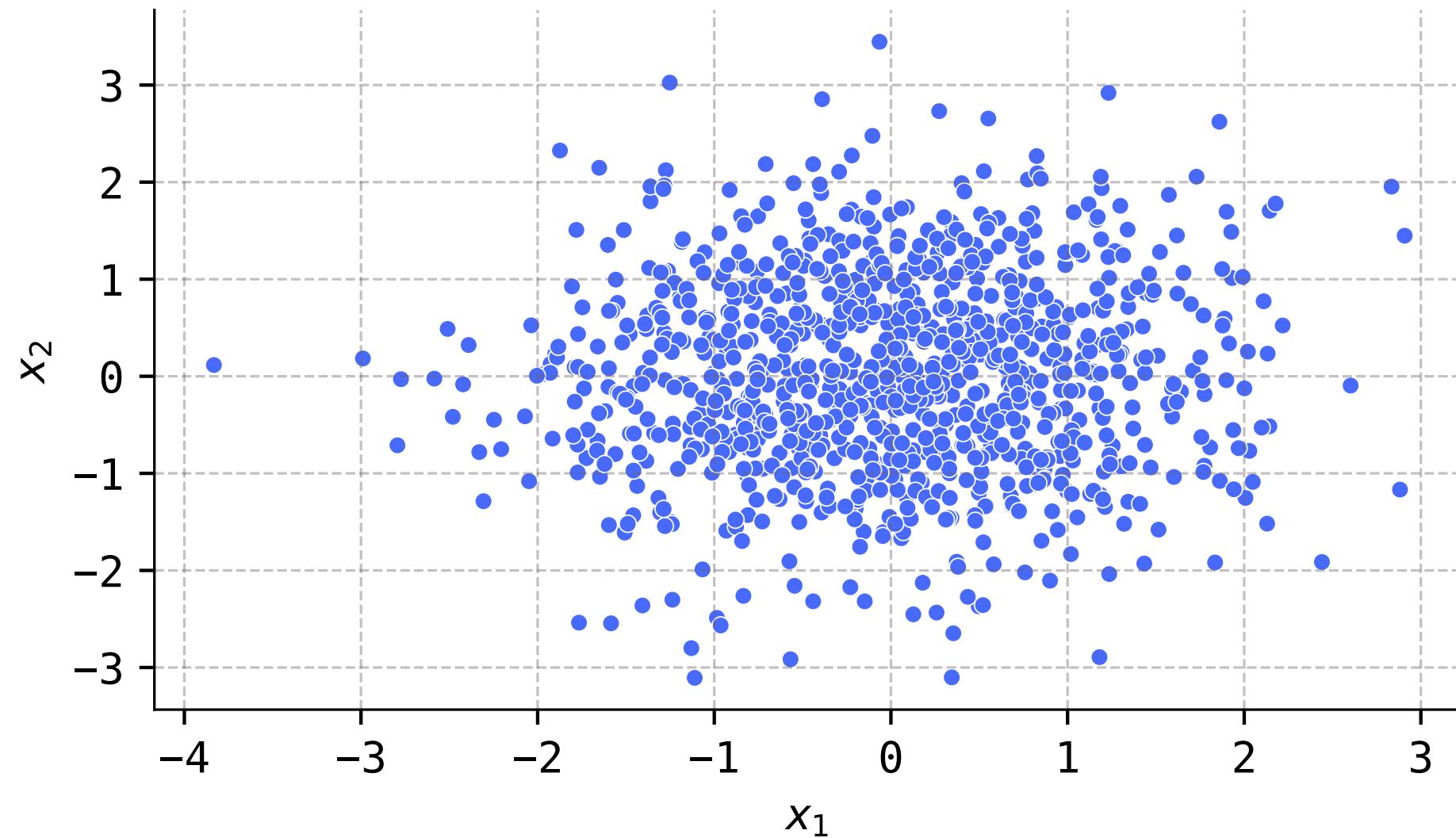


## 11.4. Stochastic gradient descent



### Definition 11.5

The *batch gradient descent algorithm* is the gradient descent algorithm applied to a stochastic objective function of the form [\(11.18\)](#).





### Problem Prompt

Do problem 12 on the worksheet.