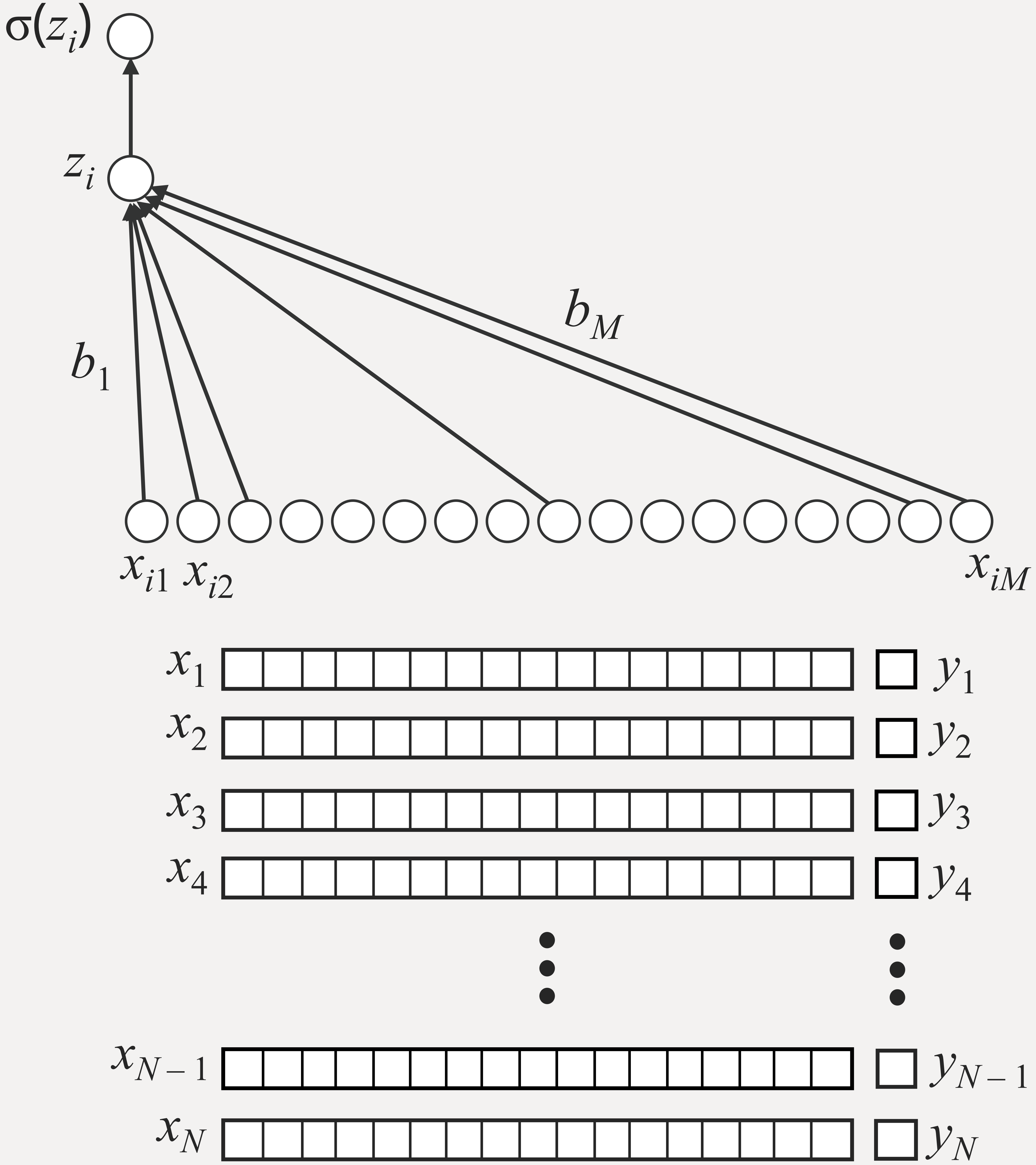
How Do We Handle Big Data?

Gradient Descent

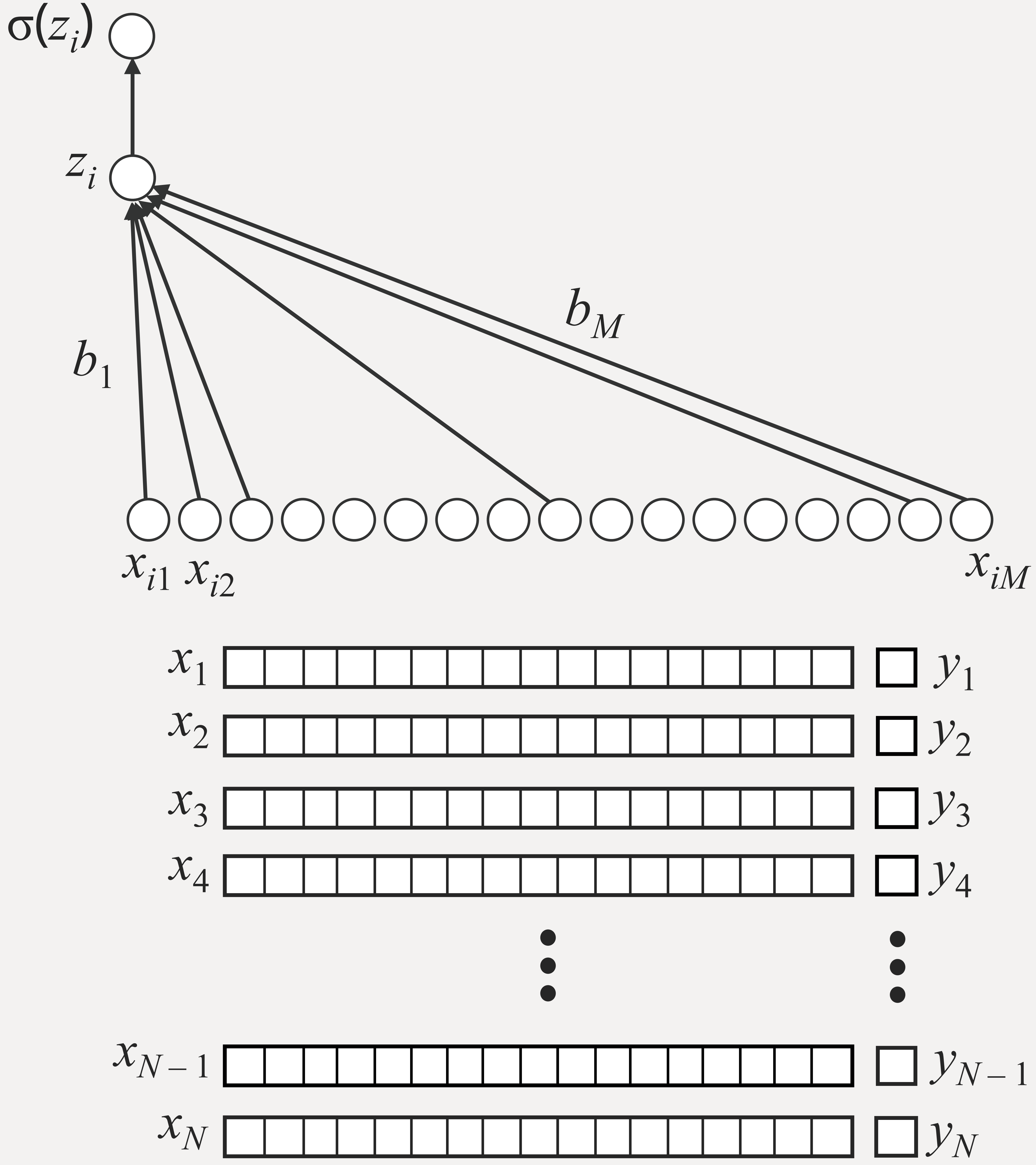
How can we handle big data from an optimization perspective?

$$b^* = \arg\min_b \frac{1}{N}\sum_i^N \ell(y_i, \sigma(z_i))$$



Calculating the gradient requires looking at every single data point

$$\nabla \frac{1}{N}\sum_{i}^{N} \ell(y_i, \sigma(z_i)) = \frac{1}{N}\sum_{i}^{N} \nabla \ell(y_i, \sigma(z_i))$$



MNIST Data Set

- ~60,000 images

Data Source: MNIST Dataset by LeCun et al. (1999) / CC-by-SA 3.0

Looking at every data point in the parameters is not scalable

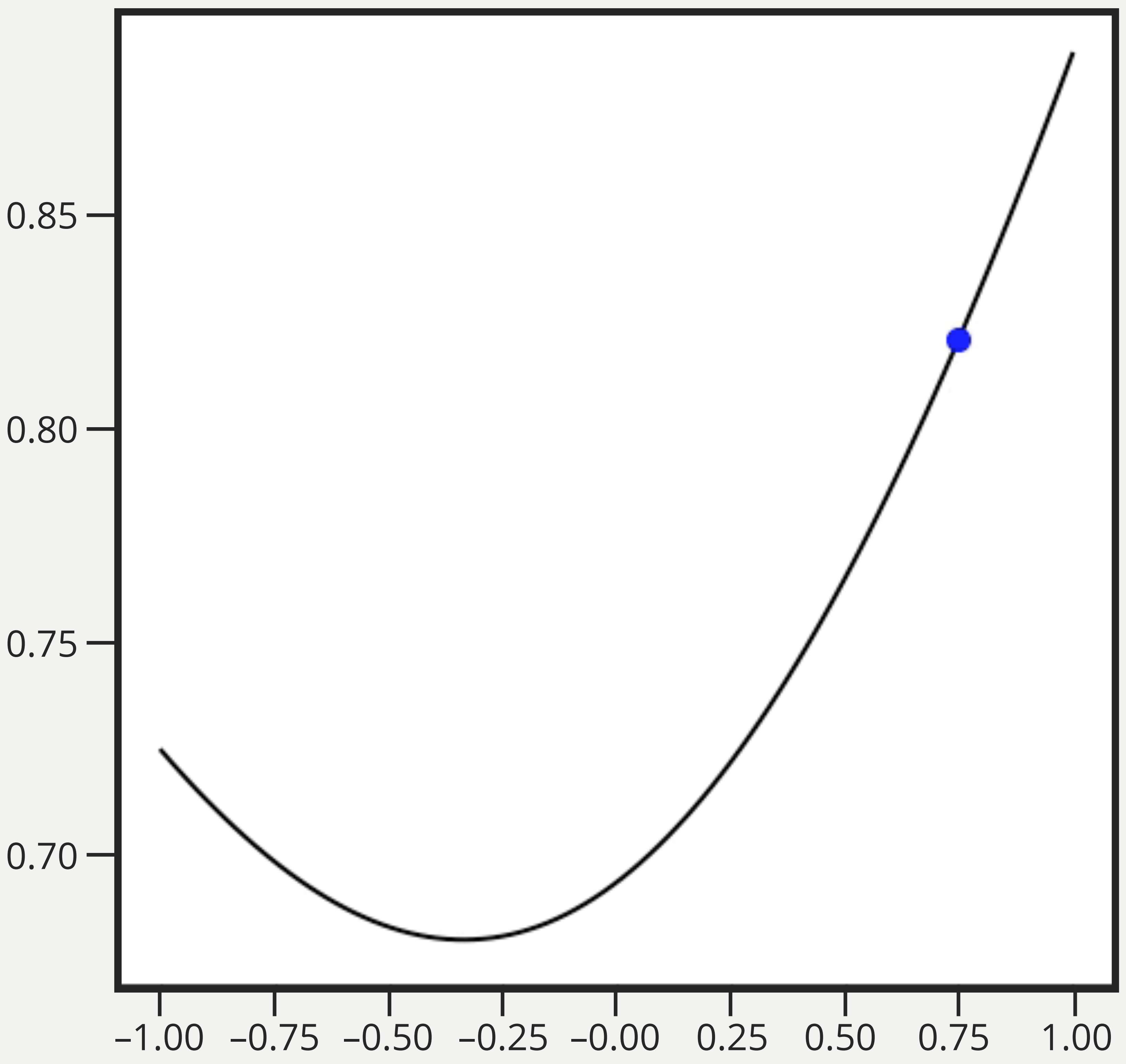
$j\,\square$

$$\nabla \ell(y_j, \sigma(z_j)) \simeq \frac{1}{N}\sum_{i=1}^{N} \nabla \ell(y_i, \sigma(z_i))$$

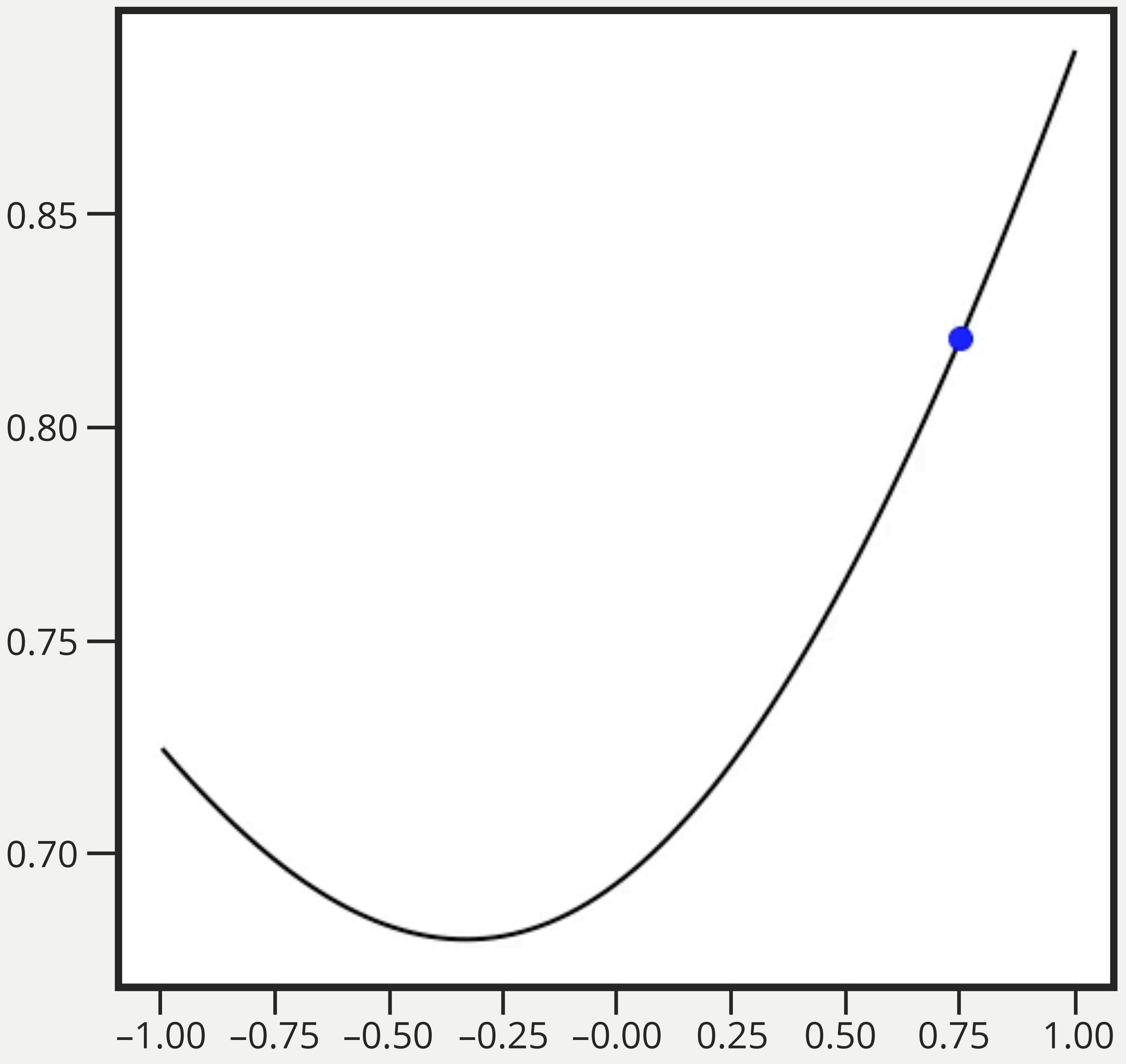
Does this work?

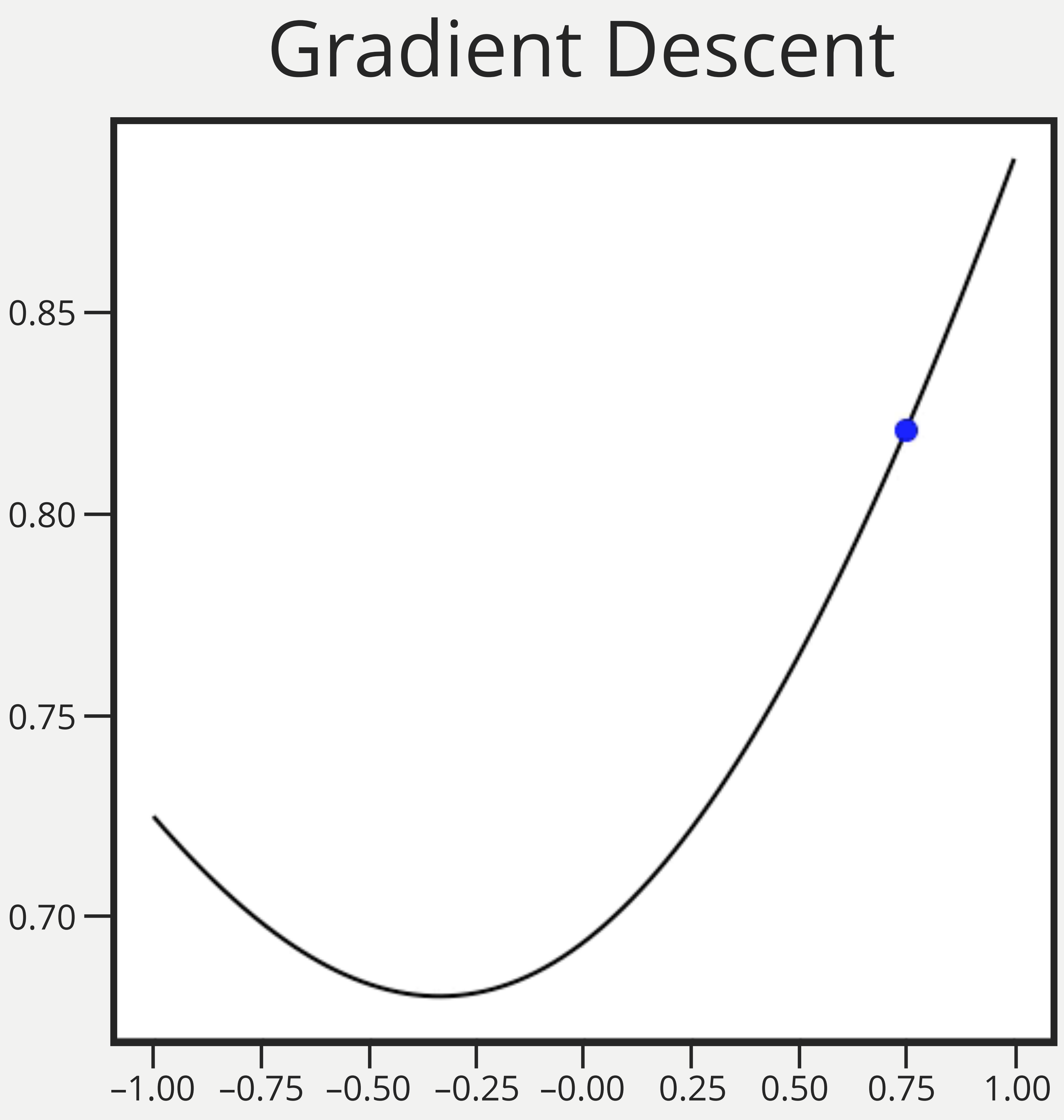
What would this look like?

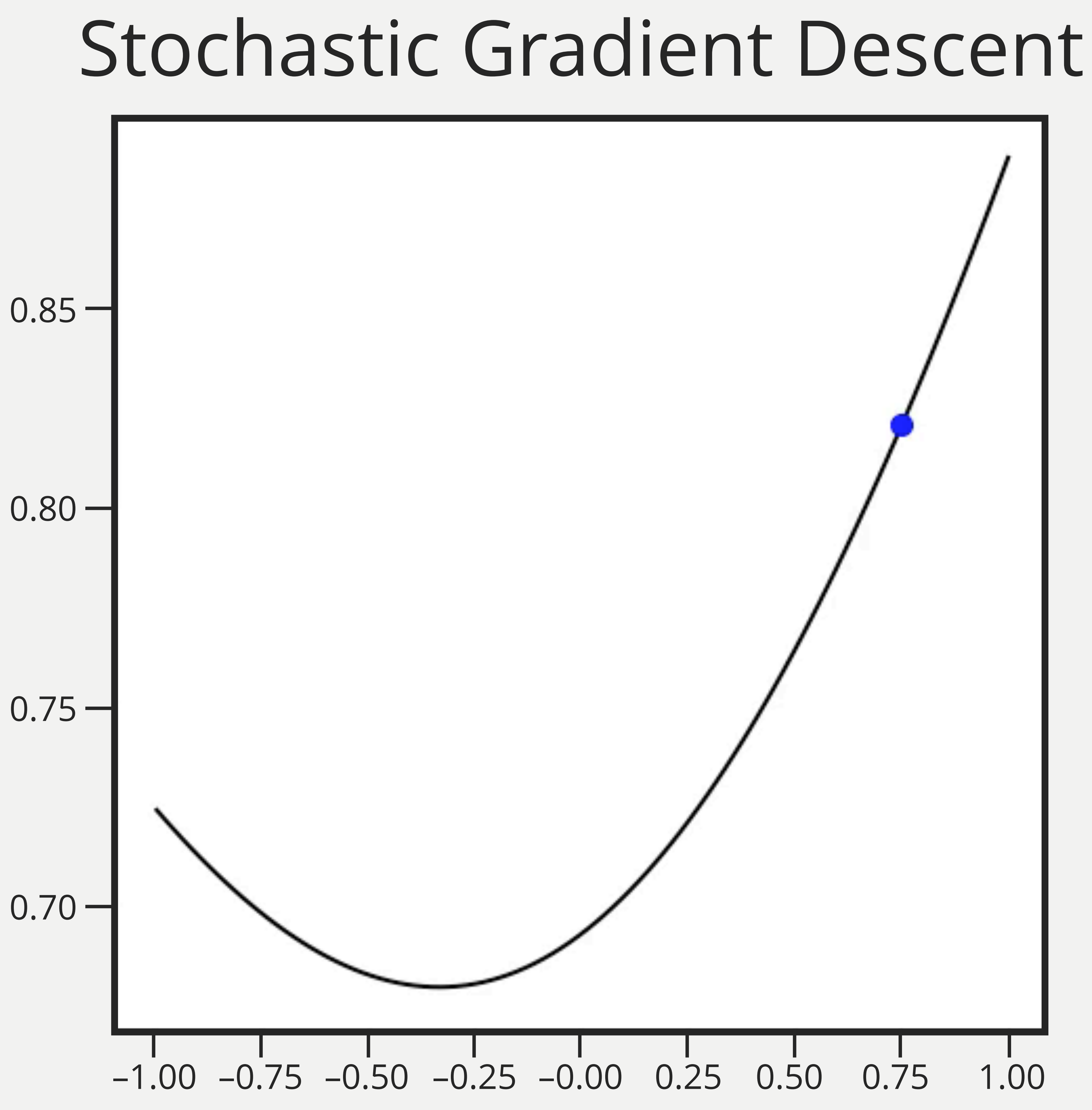
Gradient Descent



Stochastic Gradient Descent




Gradient Descent
0.85
0.80
0.75
0.70
−1.00
−0.75
−0.50
−0.25
−0.00
0.25
0.50
0.75
1.00


Stochastic Gradient Descent
0.85
0.80
0.75
0.70
−1.00
−0.75
−0.50
−0.25
−0.00
0.25
0.50
0.75
1.00

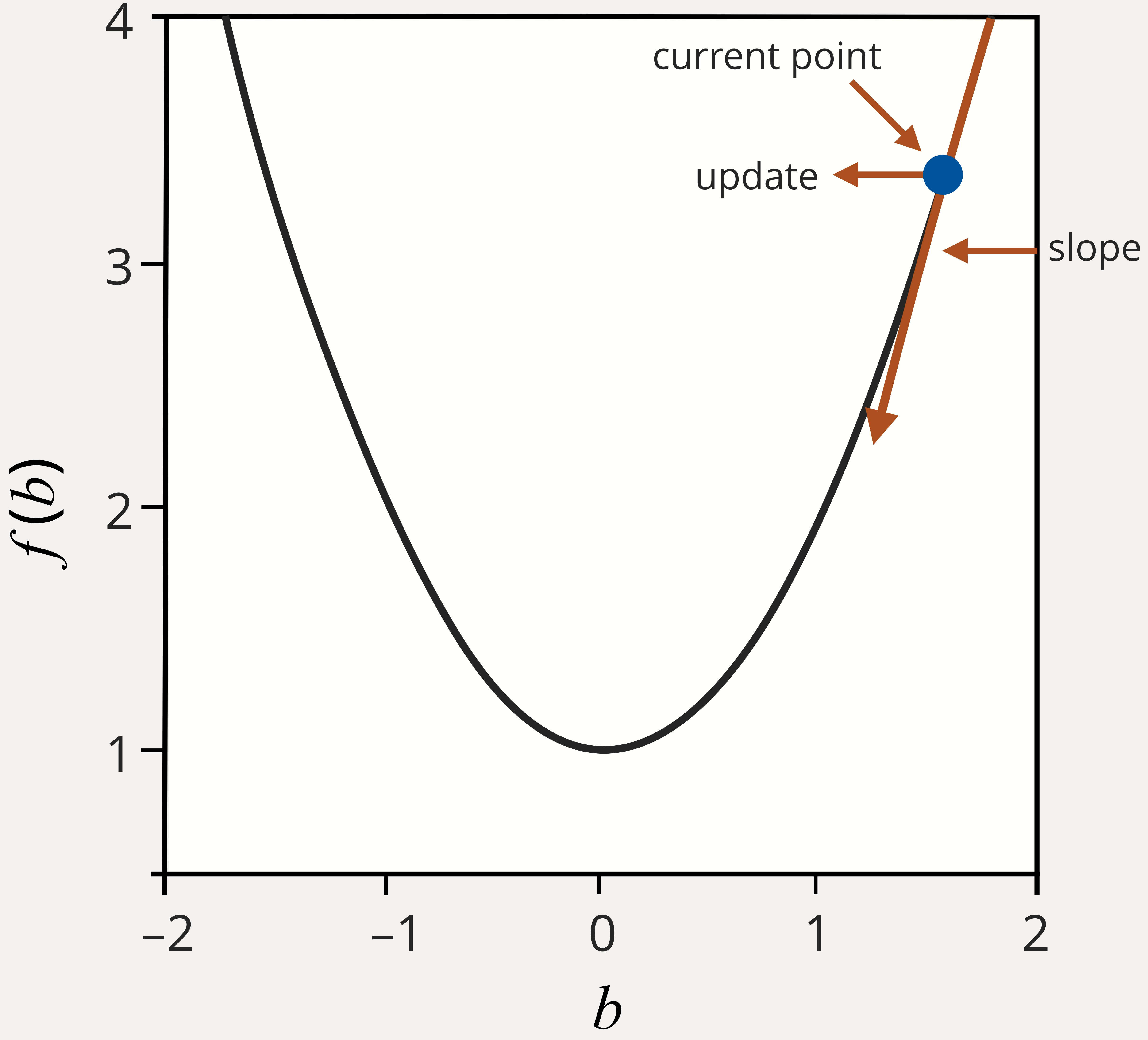
Why does this work?

Data is often redundant

MNIST Data Set

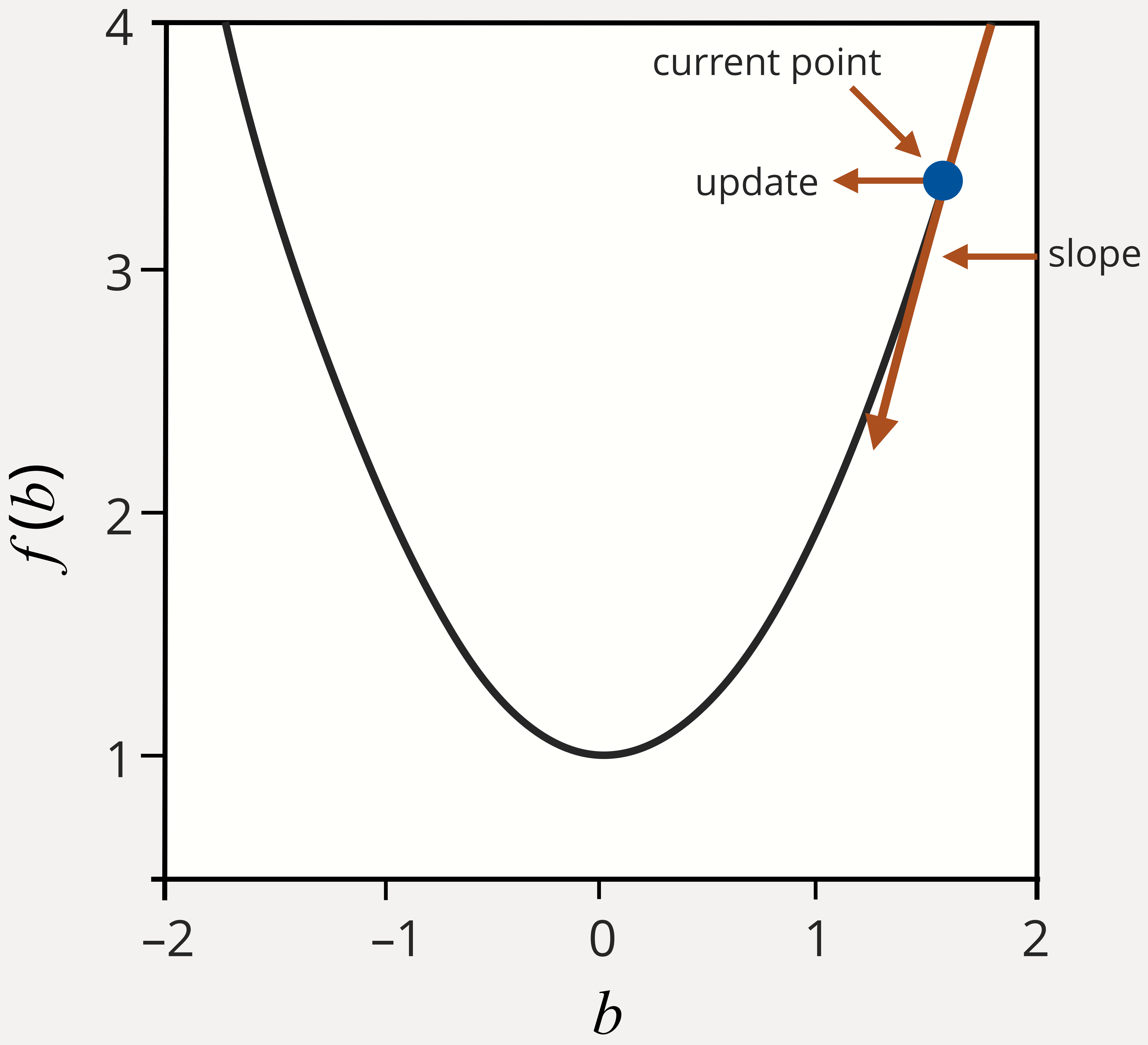
- ~60,000 images
- Only have 10 **types** of images

Data Source: MNIST Dataset by LeCun et al. (1999) / CC-by-SA 3.0



Gradient Descent

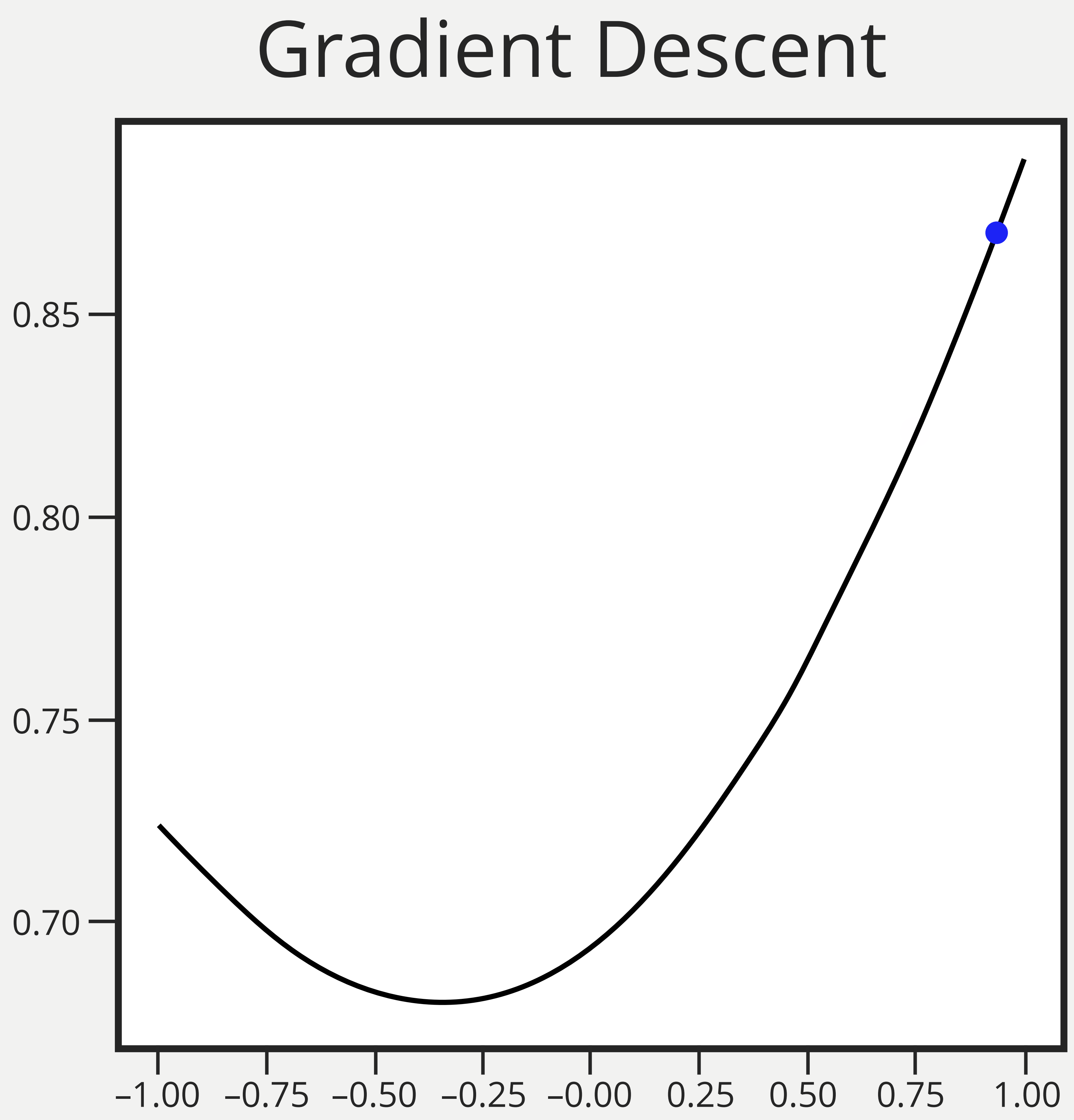
1. Start with initial value b^0
2. Calculate gradient $\nabla f(b^k)$ over **all** data
3. Iteratively update: $b^{k+1} \leftarrow b^k - \alpha^k \nabla f(b^k)$
4. Repeat 2-3 until solution is good enough

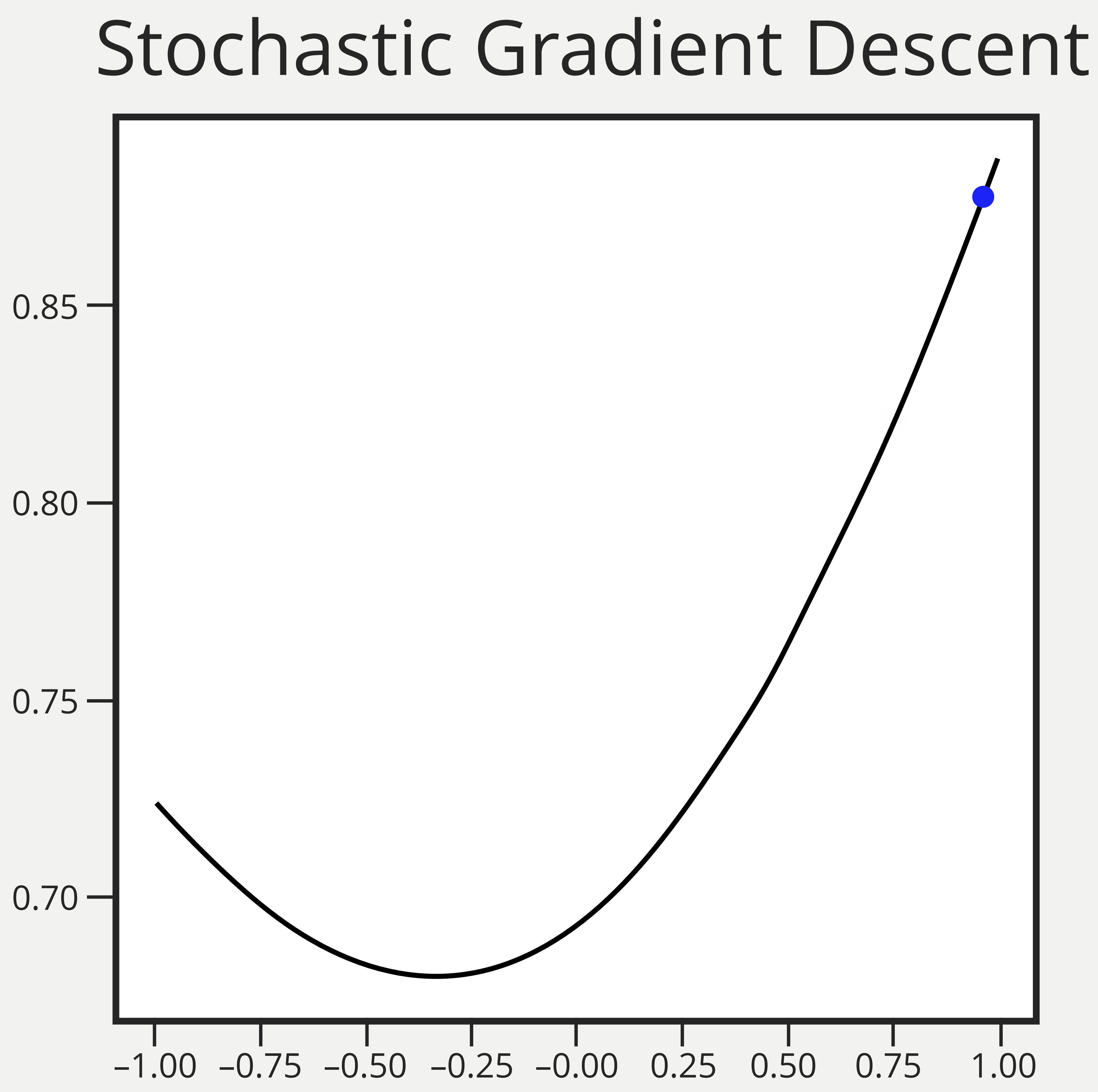


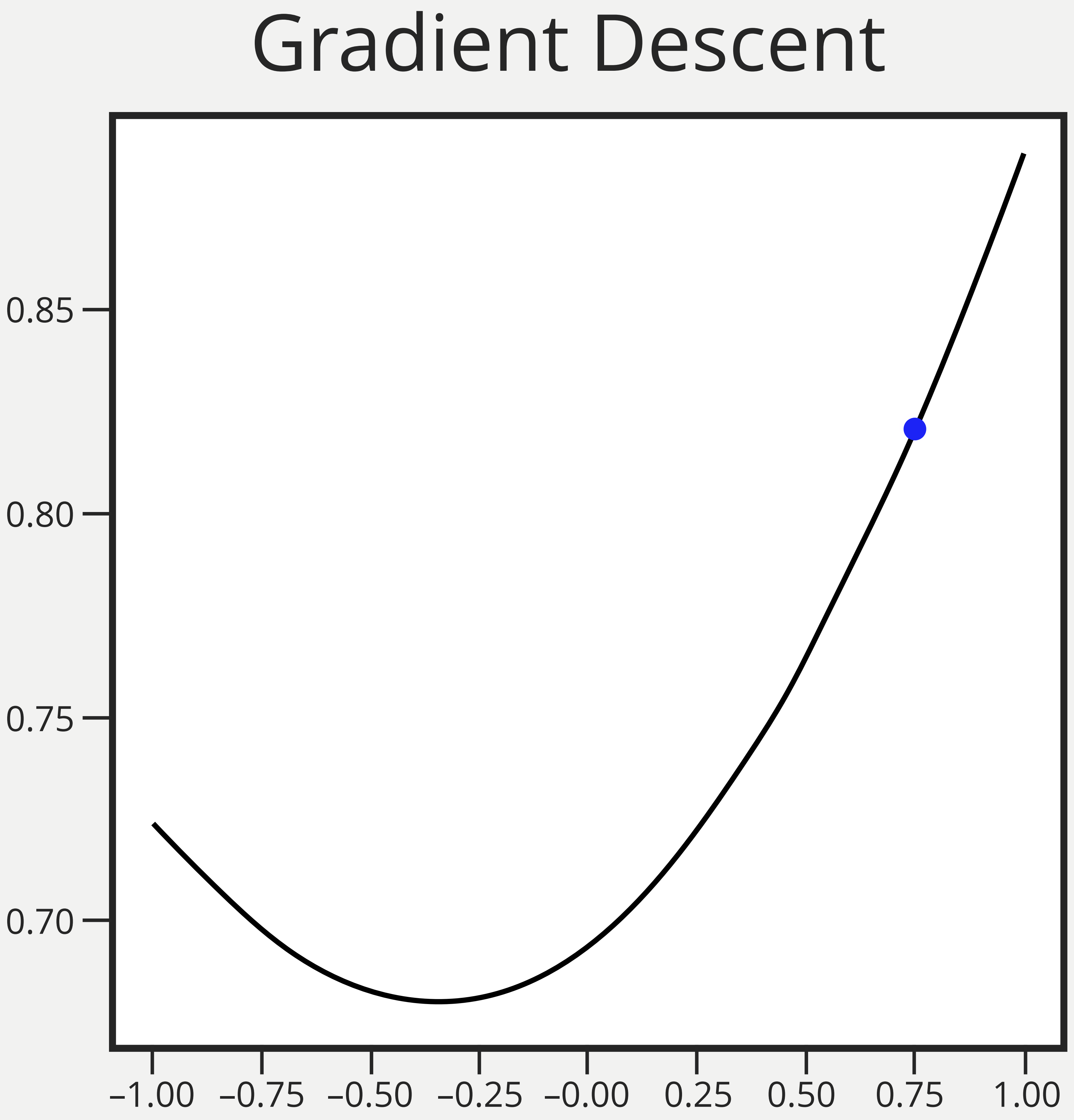
Stochastic Gradient Descent

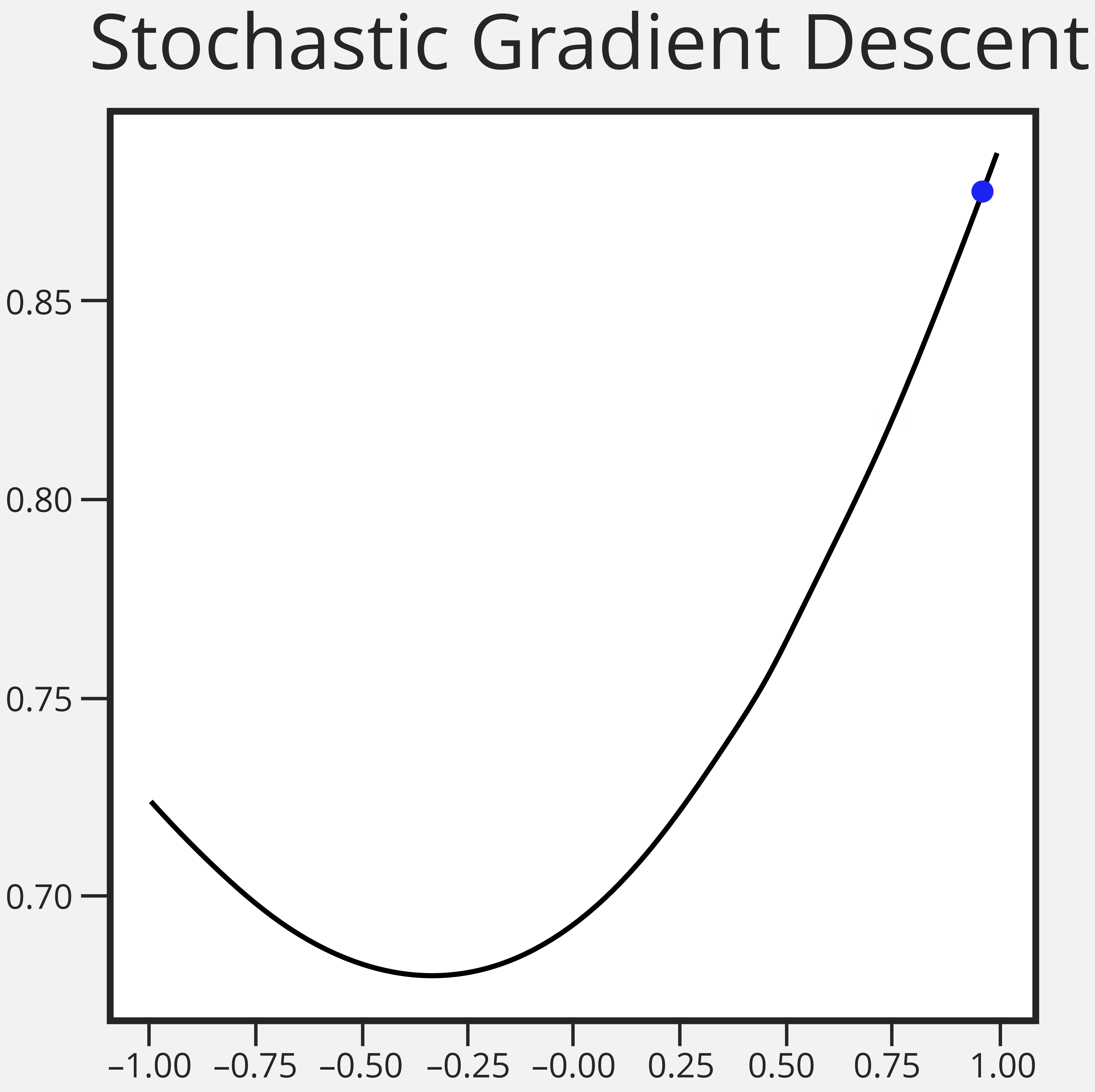
1. Start with initial value b^0
2. Choose a **random** data entry j
3. Estimate gradient $\widehat{\nabla f}(b^k)$ by data point j
4. Iteratively update: $b^{k+1} \leftarrow b^k - \alpha^k \widehat{\nabla f}(b^k)$
5. Repeat 2-4 until solution is good enough

$$\mathrm{E}_{j \sim Unif(1,\ldots,\mathrm{N})}\left[\nabla f_j(b)\right] = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(b)$$

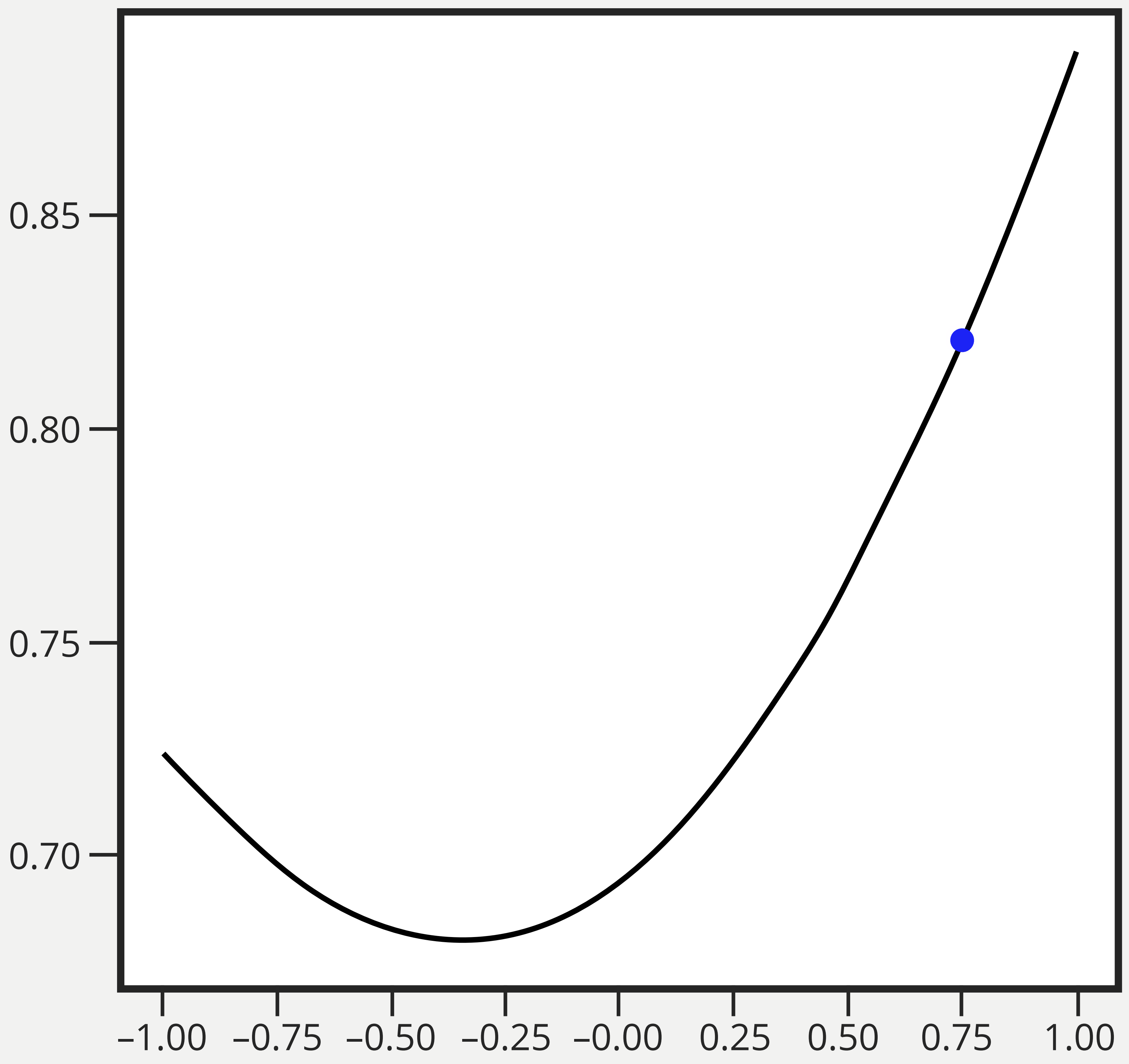

Gradient Descent
0.85
0.80
0.75
0.70
−1.00 −0.75 −0.50 −0.25 −0.00 0.25 0.50 0.75 1.00


Stochastic Gradient Descent
0.85
0.80
0.75
0.70
−1.00 −0.75 −0.50 −0.25 −0.00 0.25 0.50 0.75 1.00

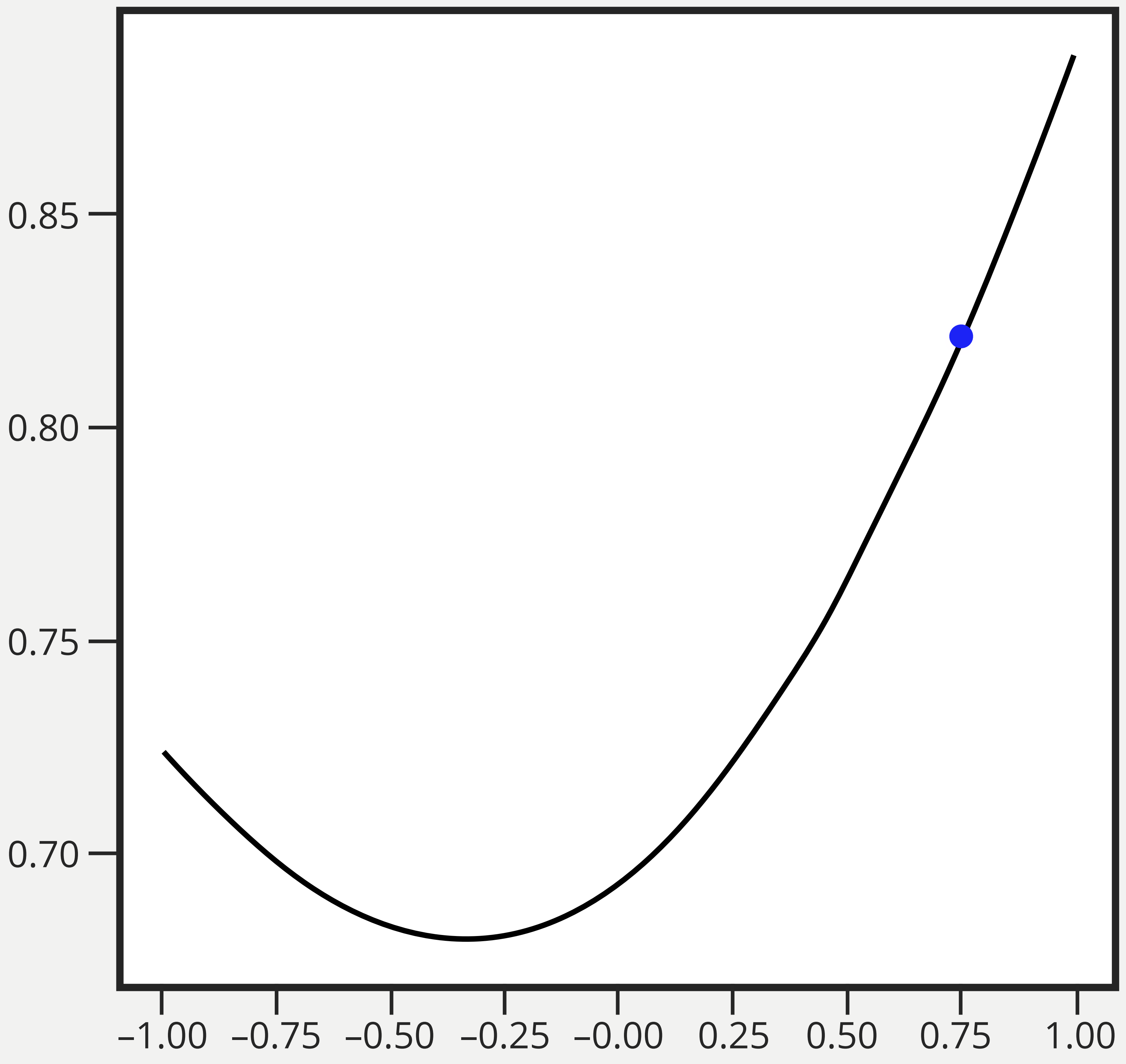

Gradient Descent
0.85
0.80
0.75
0.70
−1.00
−0.75
−0.50
−0.25
−0.00
0.25
0.50
0.75
1.00


Stochastic Gradient Descent
0.85
0.80
0.75
0.70
−1.00
−0.75
−0.50
−0.25
−0.00
0.25
0.50
0.75
1.00

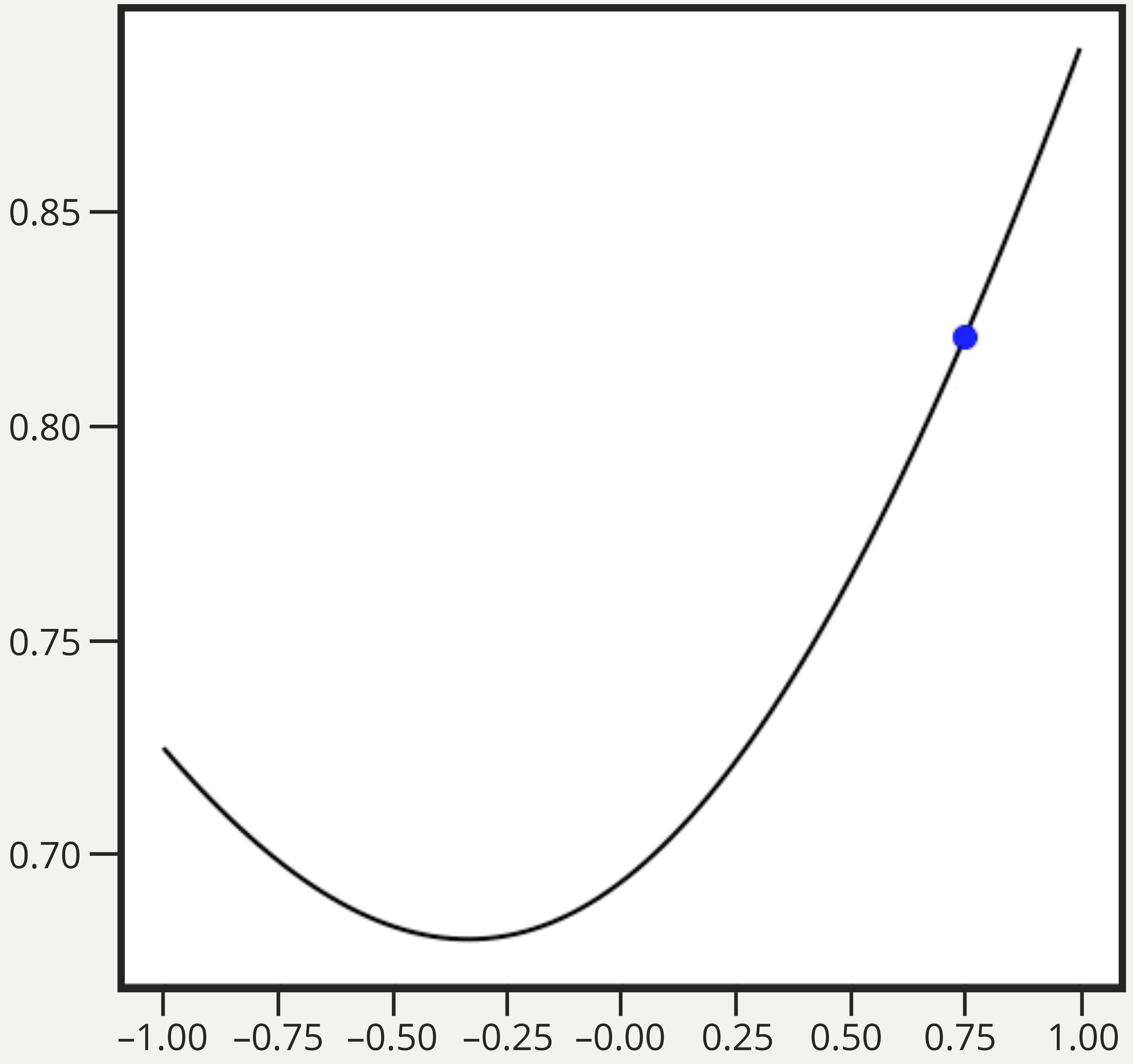
Gradient Descent



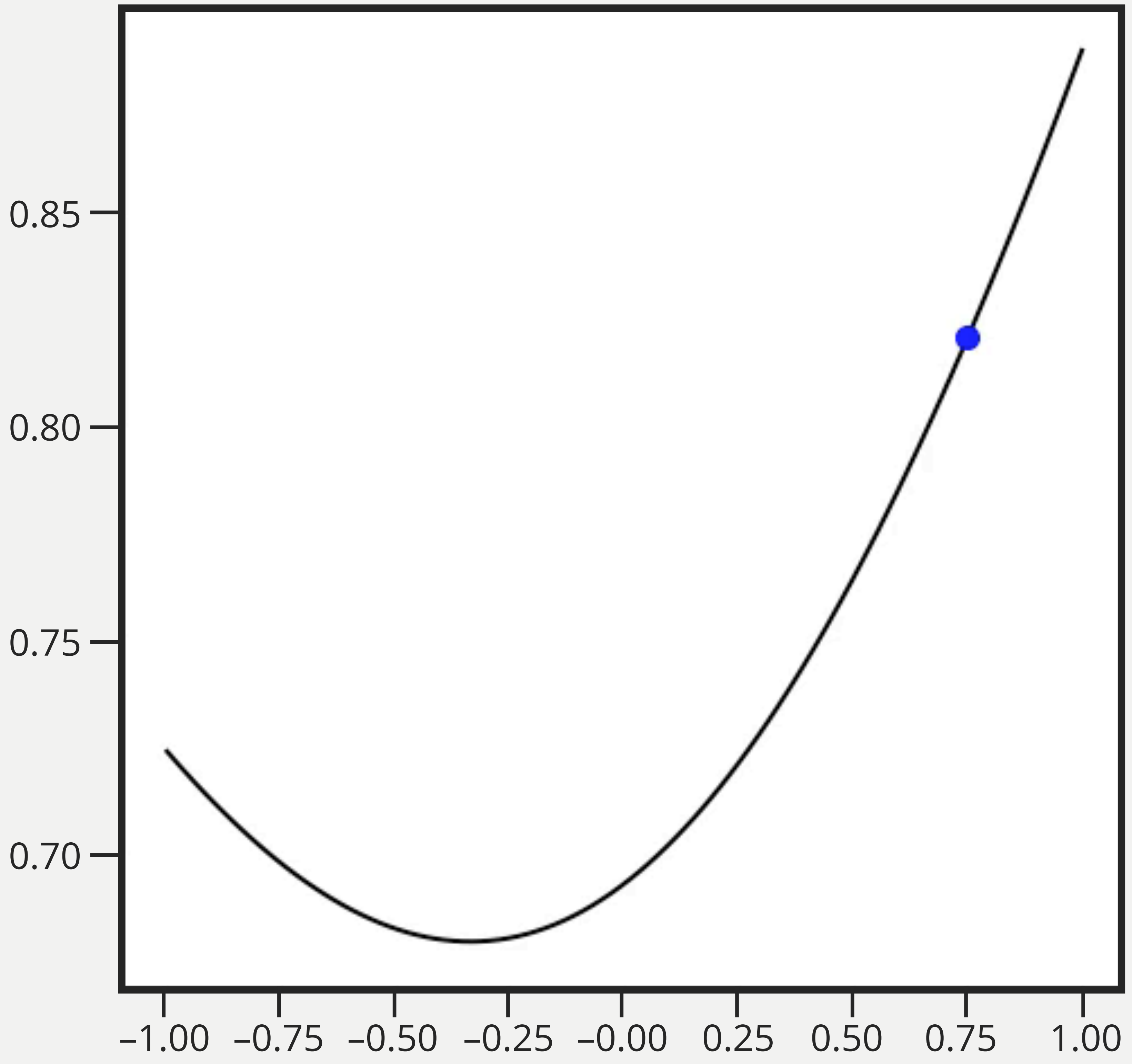
Stochastic Gradient Descent



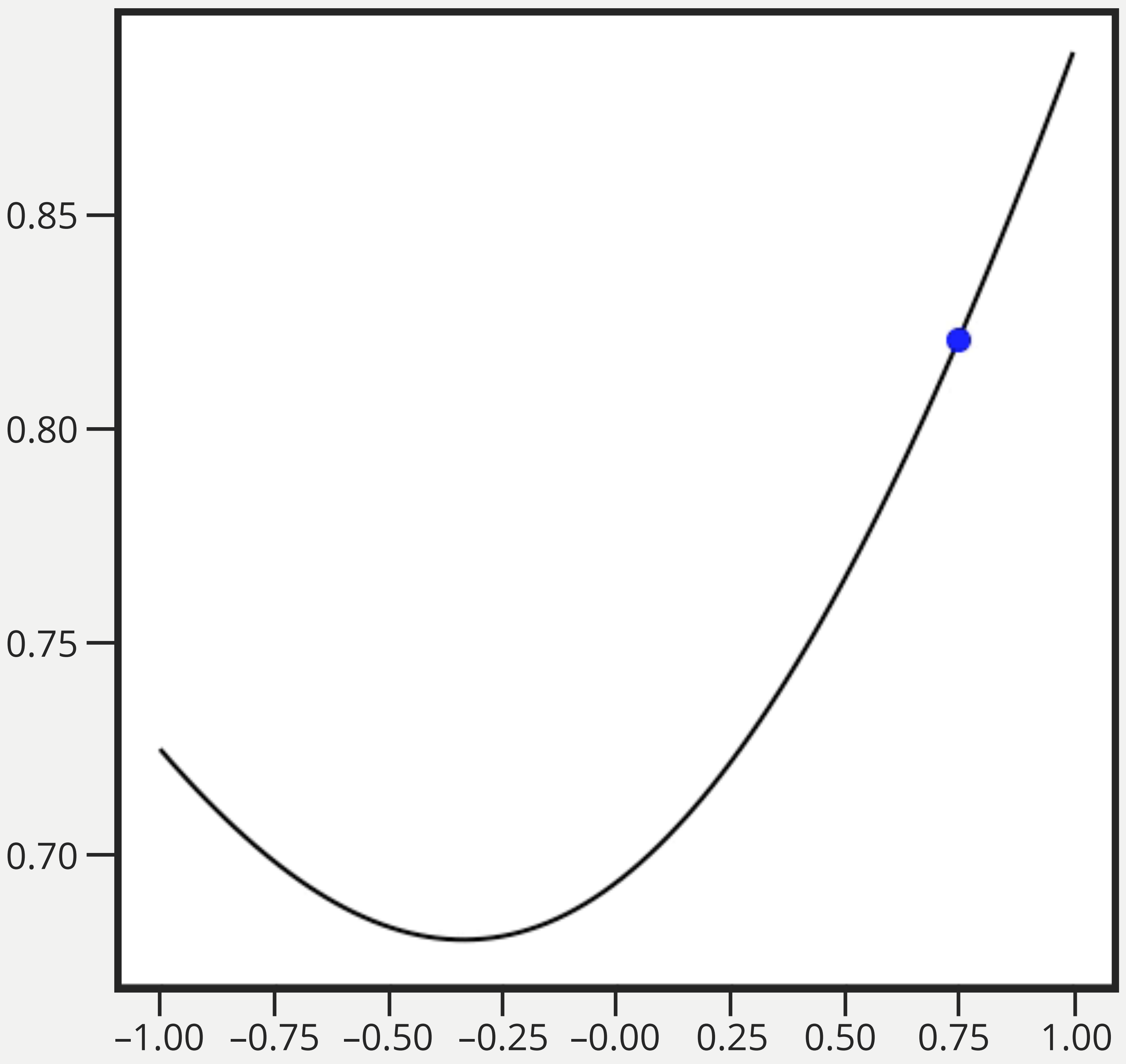
Gradient Descent



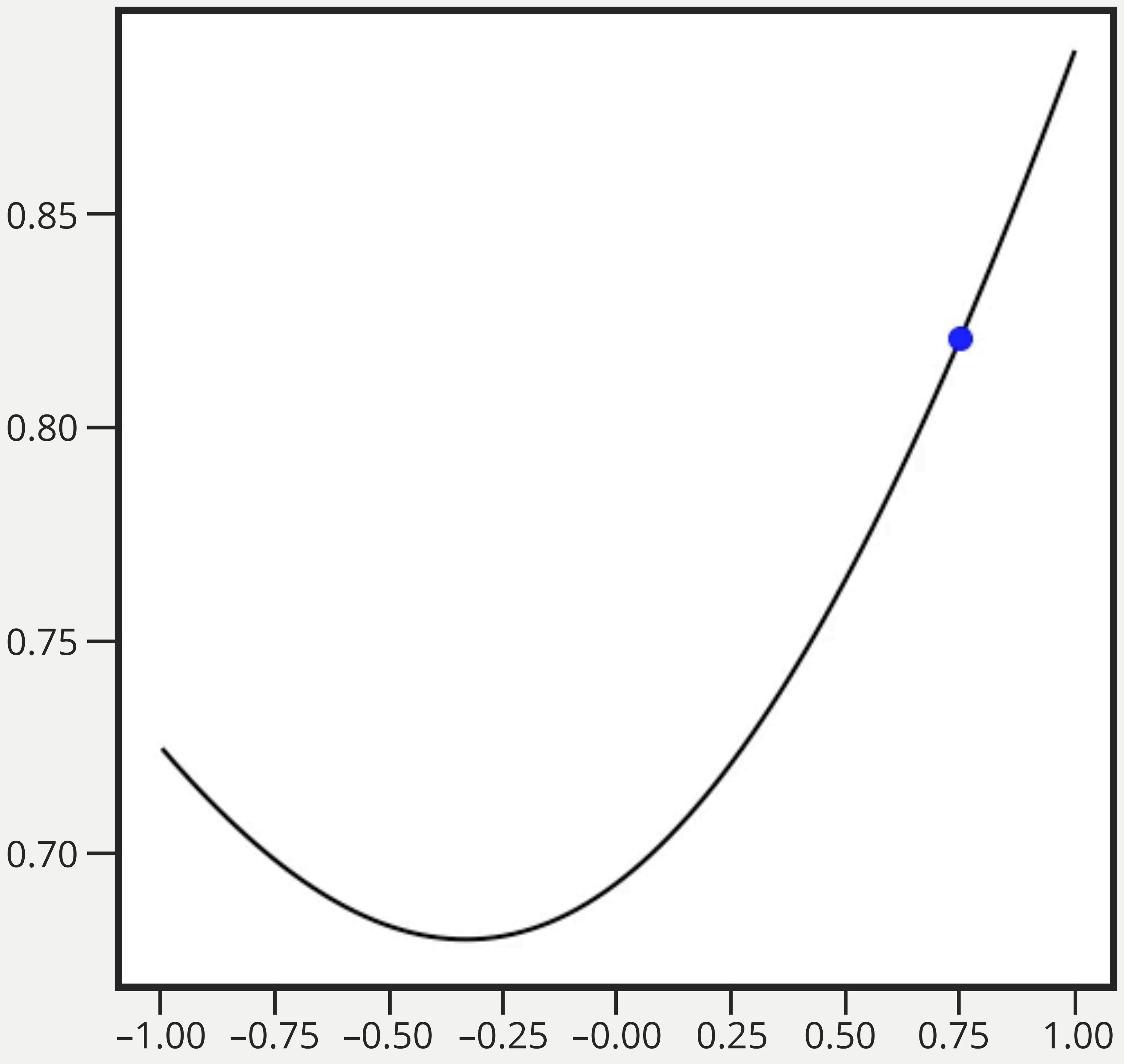
Stochastic Gradient Descent



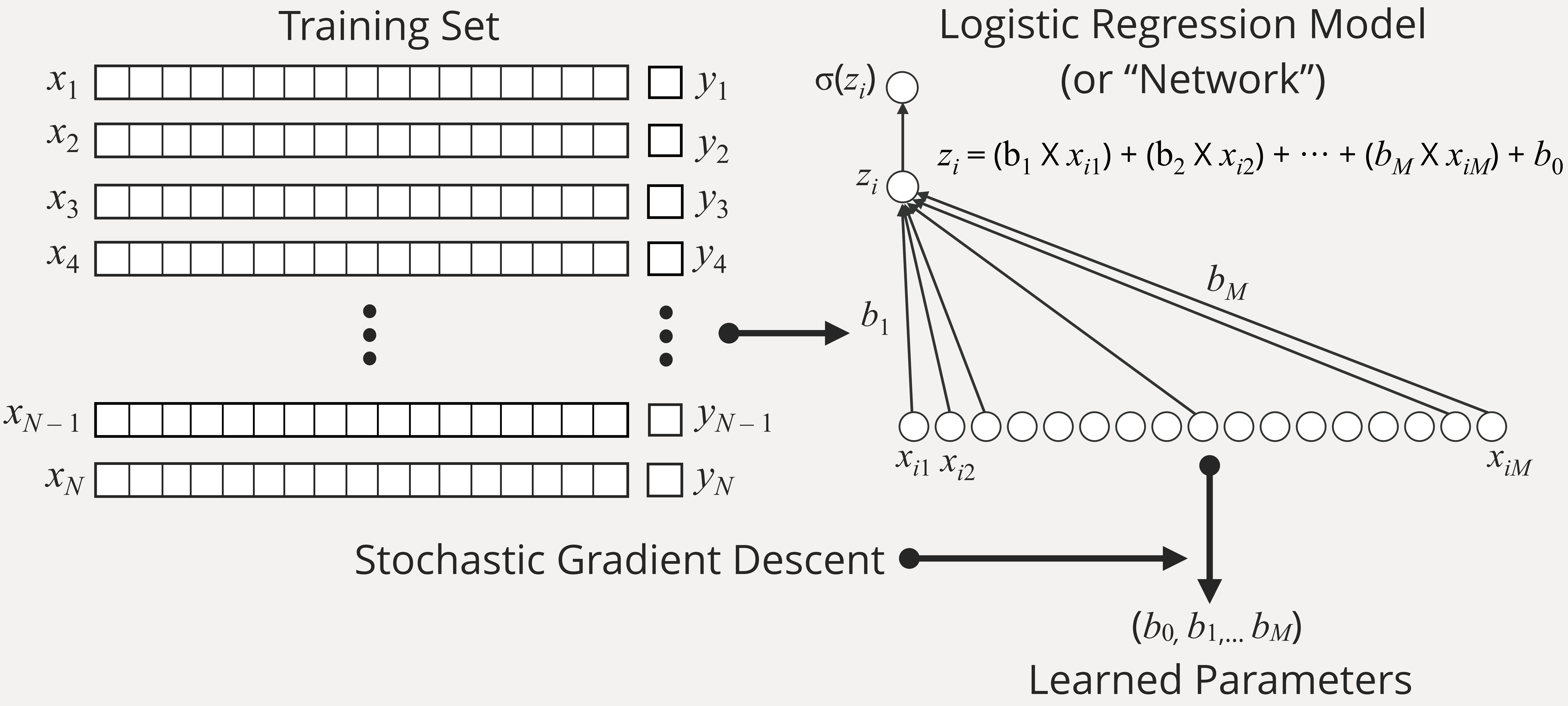
Gradient Descent



Stochastic Gradient Descent



Learned Model Parameters



Stochastic Gradient Descent can update many more times than Gradient Descent

Will get near the solution quickly, using Stochastic Gradient Descent

Stochastic Gradient Descent allows for scaling to big data

Credits

MNIST Dataset of Handwritten Digits (Images)

Yann LeCun (Courant Institute, NYU) and Corinna Cortes (Google Labs, New York)
http://yann.lecun.com/exdb/mnist/