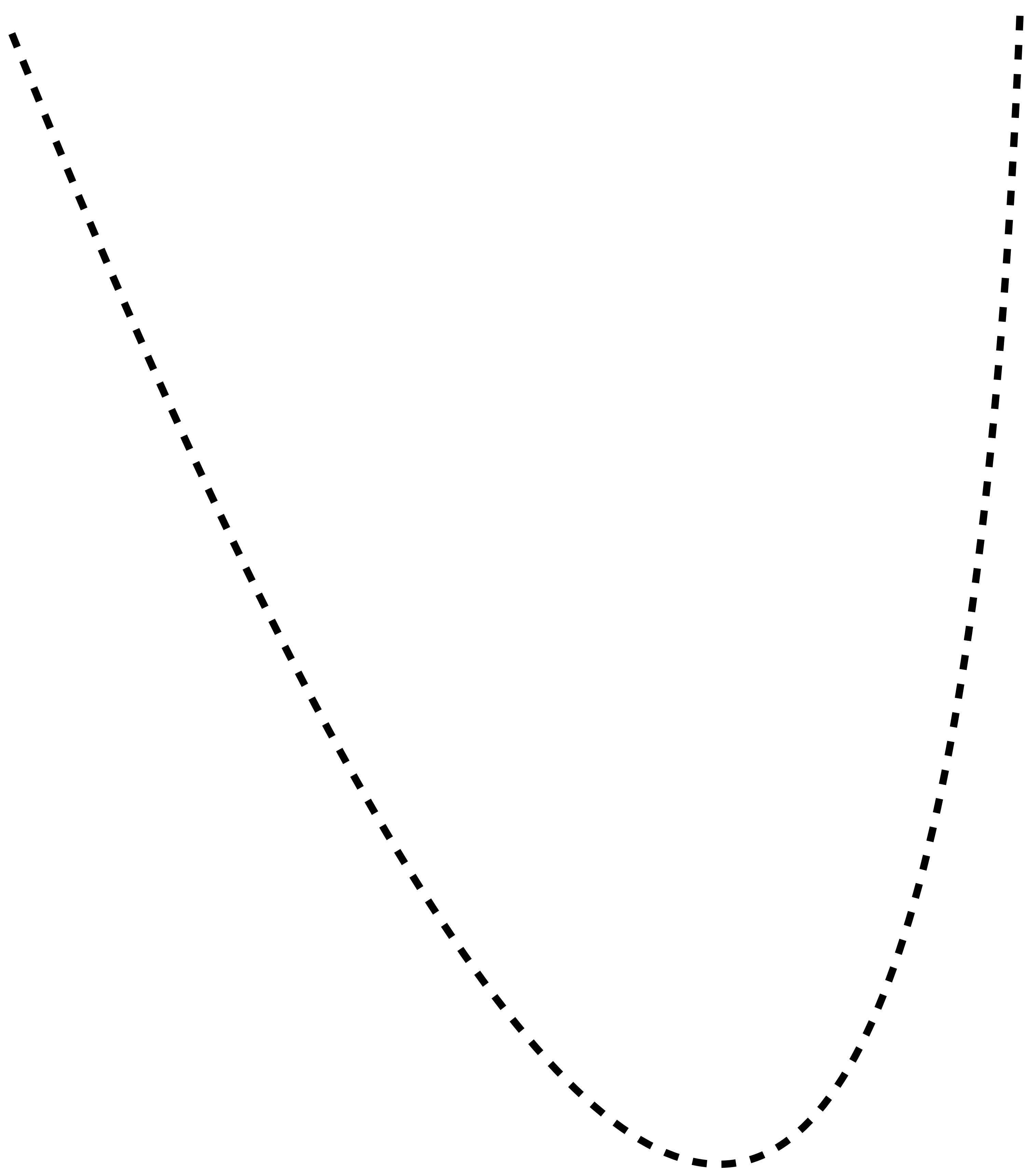
Adaptive Bounding Subgradient Method for Convex Functions

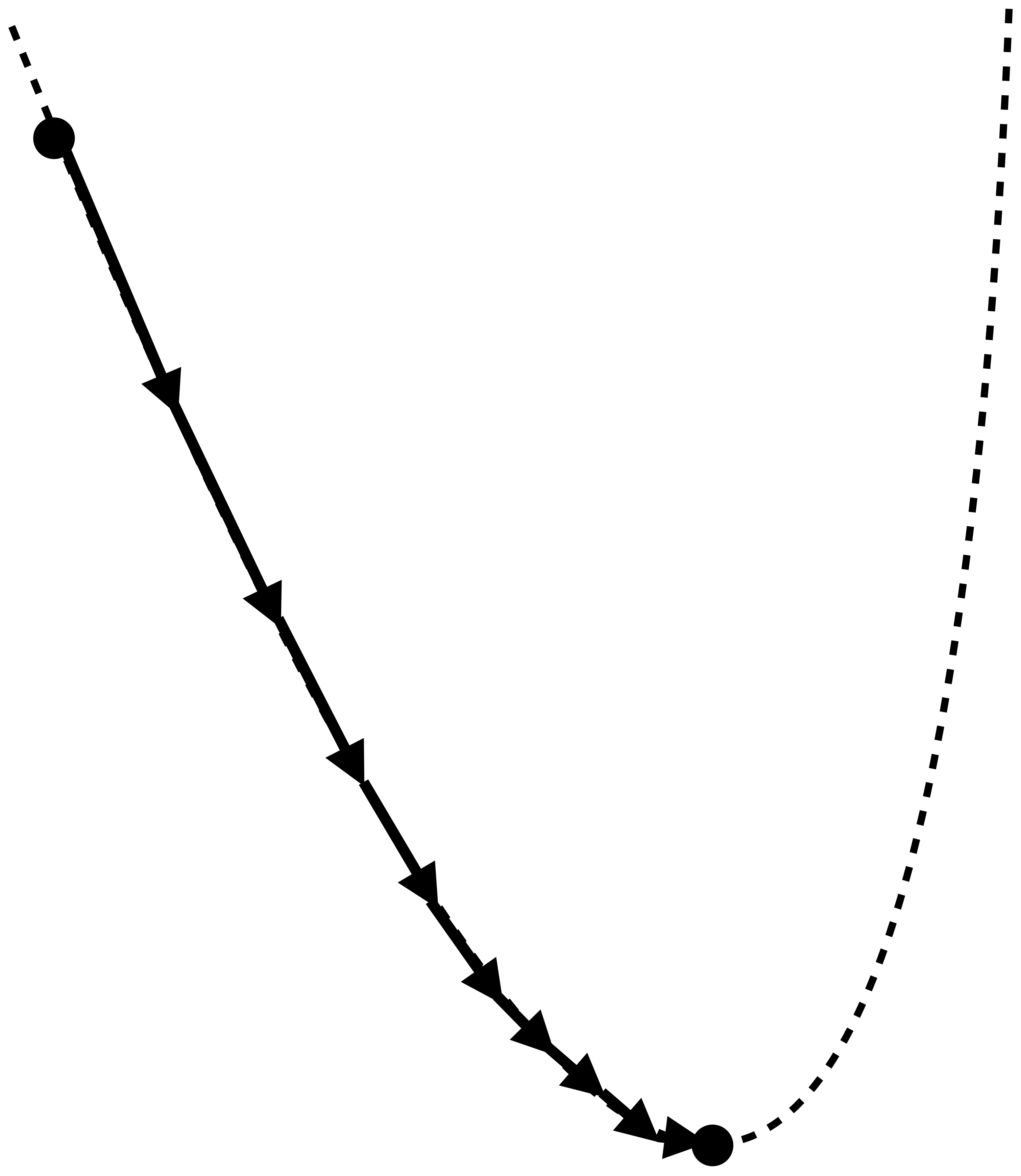
Nicholas Dwork, Charles Tsao, Cody Coleman, John M. Pauly
Radiology, University of California in San Francisco
Electrical Engineering, Stanford University

$$x^{(k+1)} = x^{(k)} - \alpha\, g\left(x^{(k)}\right)$$

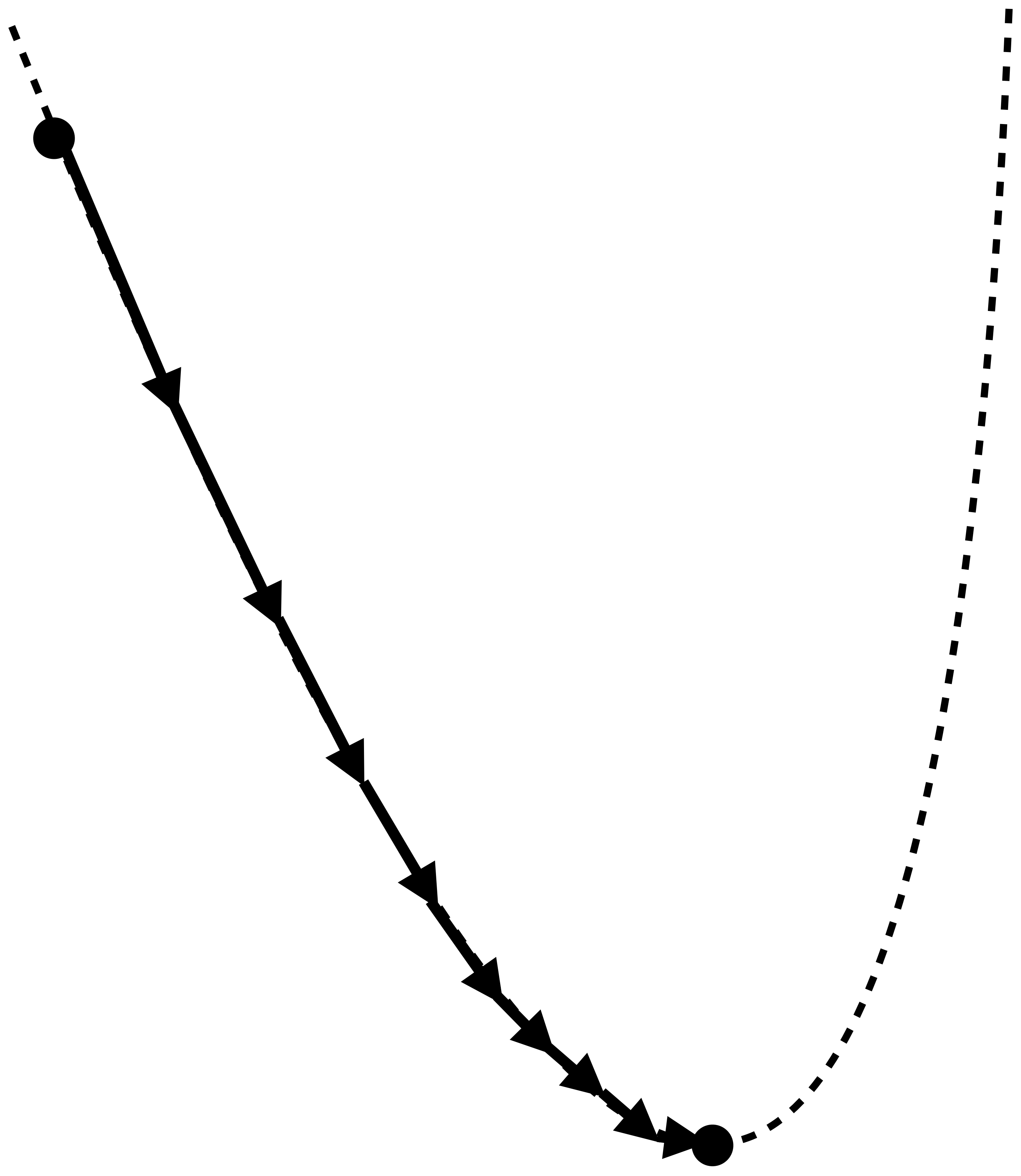
$$x^{(k+1)} = x^{(k)} - \alpha\, g\left(x^{(k)}\right)$$

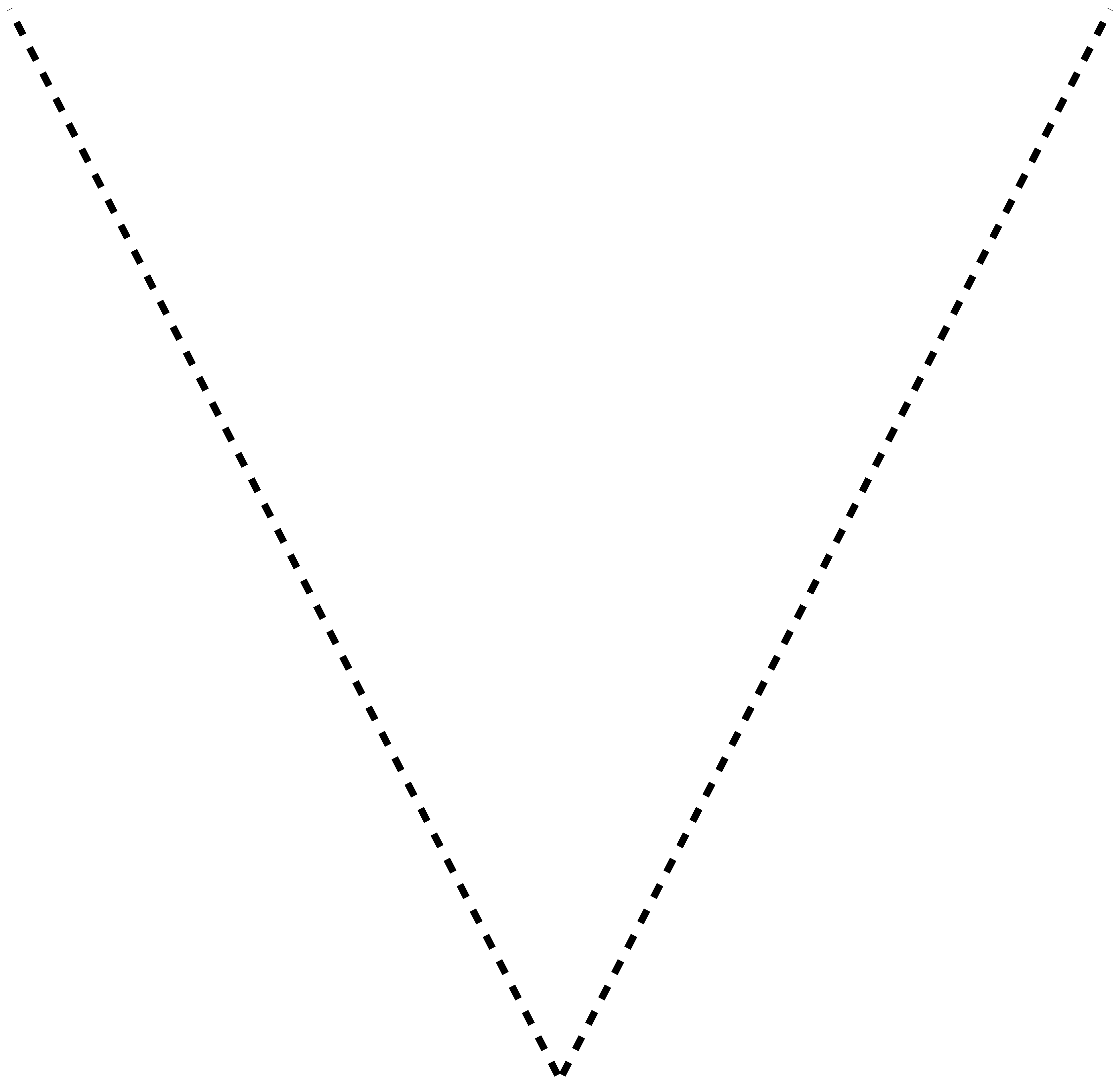


$$x^{(k+1)} = x^{(k)} - \alpha\, g\left(x^{(k)}\right)$$

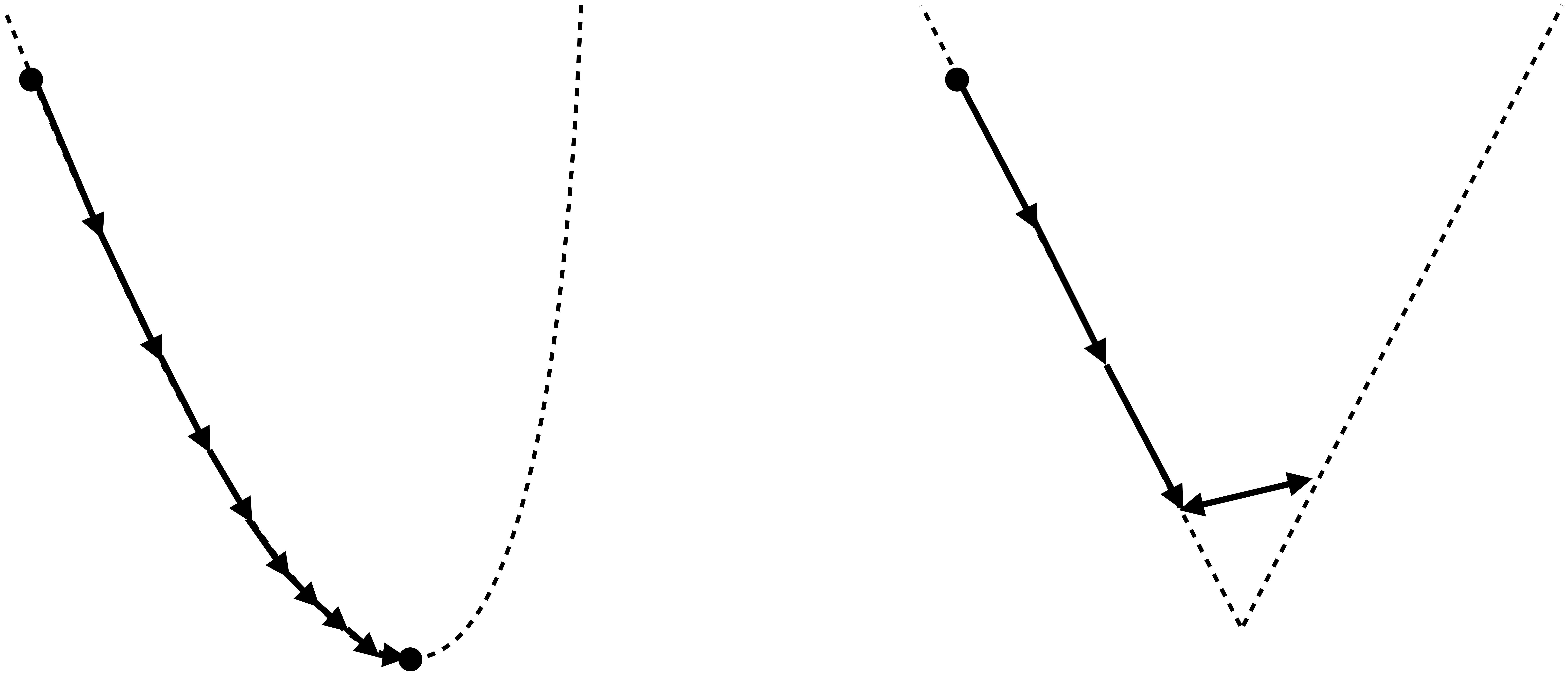


$$x^{(k+1)} = x^{(k)} - \alpha\, g\left(x^{(k)}\right)$$

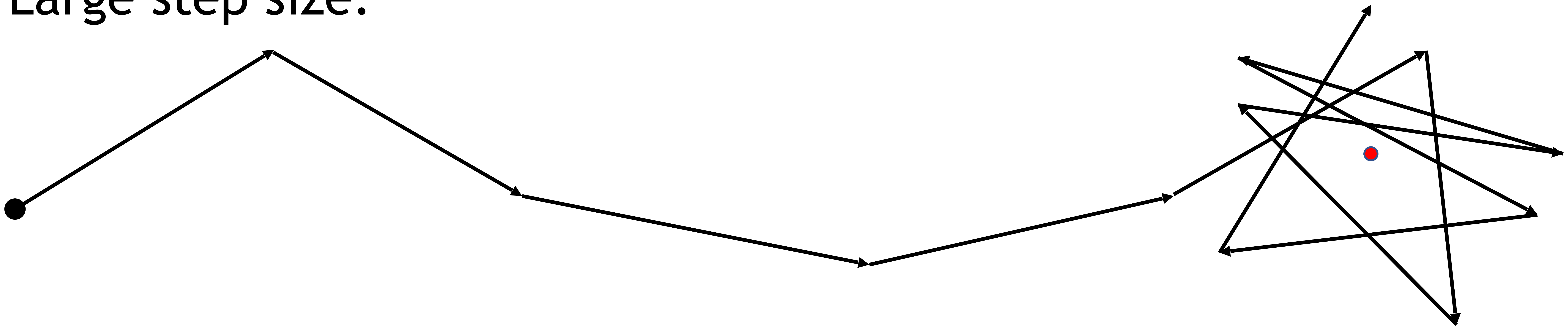




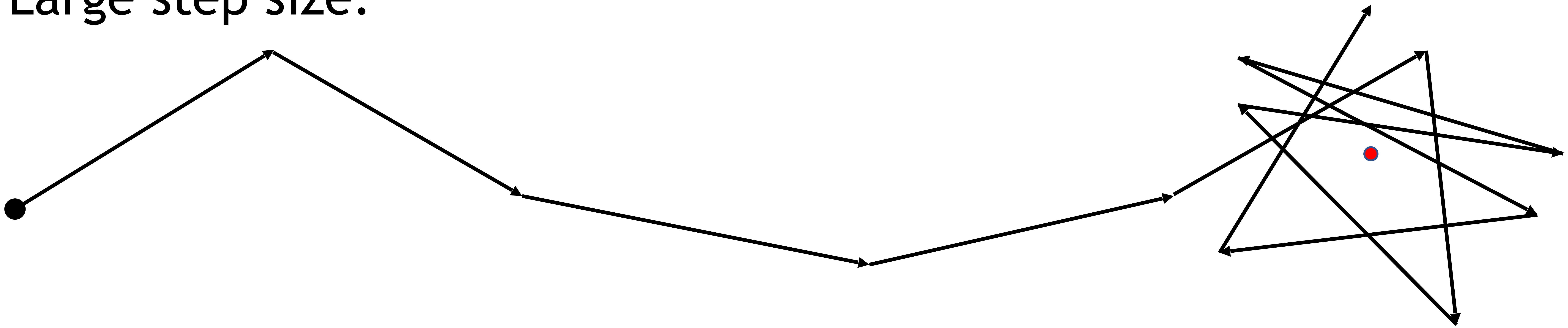
$$x^{(k+1)} = x^{(k)} - \alpha\, g\left(x^{(k)}\right)$$



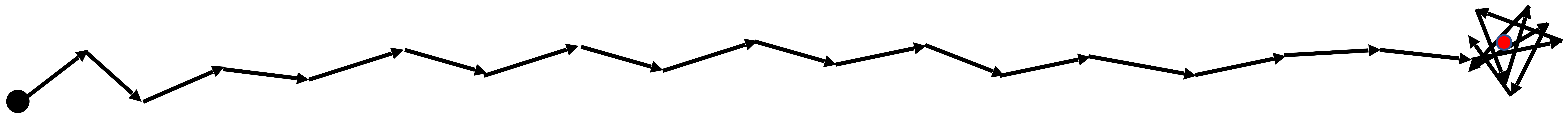
Large step size:



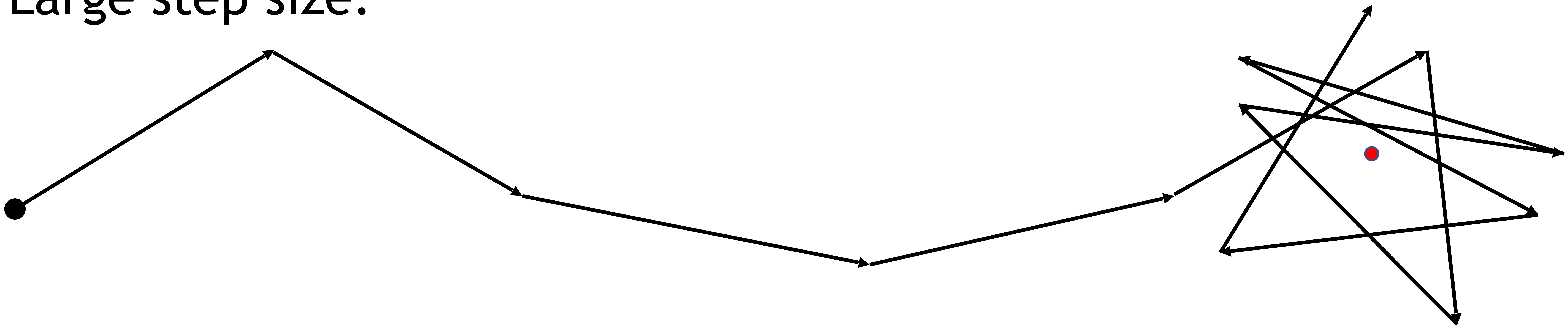
Large step size:



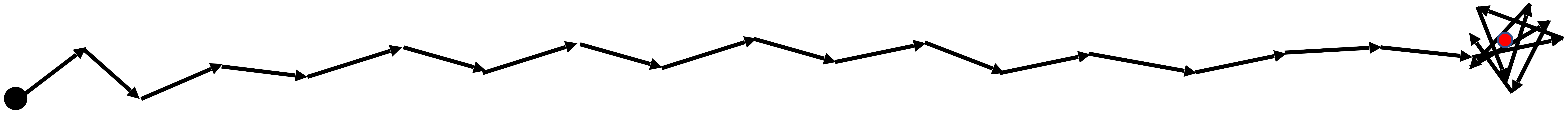
Small step size:



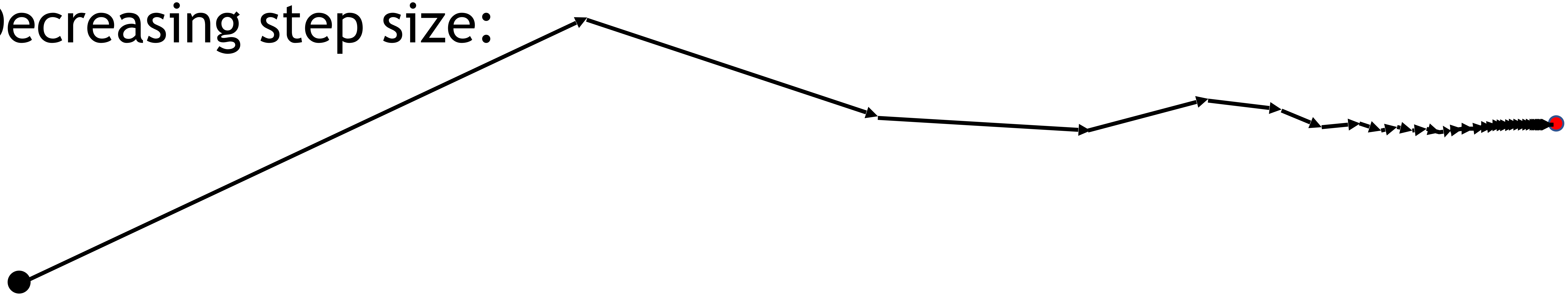
Large step size:



Small step size:



Decreasing step size:



Theorem: if α is small enough, the next point in the trajectory is closer to the optimal set.

Theorem: if α is small enough, the next point in the trajectory is closer to the optimal set.

Proof:

Theorem: if α is small enough, the next point in the trajectory is closer to the optimal set.

Proof: $x^{(k+1)} = x^{(k)} - \alpha\, g_k$

Theorem: if α is small enough, the next point in the trajectory is closer to the optimal set.

Proof: $x^{(k+1)} = x^{(k)} - \alpha\, g_k$

$$\|x^{(k+1)} - x^\star\|_2^2 = \|x^{(k)} - \alpha g_k - x^\star\|_2^2$$

Theorem: if α is small enough, the next point in the trajectory is closer to the optimal set.

Proof: $x^{(k+1)} = x^{(k)} - \alpha\, g_k$

$$\begin{aligned} \|x^{(k+1)} - x^\star\|_2^2 &= \|x^{(k)} - \alpha g_k - x^\star\|_2^2 \\ &= \|x^{(k)} - x^\star\|_2^2 - 2\alpha\langle g_k, x^{(k)} - x^\star\rangle + \alpha^2\|g_k\|_2^2 \end{aligned}$$

Theorem: if α is small enough, the next point in the trajectory is closer to the optimal set.

Proof: $x^{(k+1)} = x^{(k)} - \alpha\, g_k$

$$\begin{aligned}\|x^{(k+1)} - x^\star\|_2^2 &= \|x^{(k)} - \alpha g_k - x^\star\|_2^2 \\ &= \|x^{(k)} - x^\star\|_2^2 - 2\alpha\langle g_k, x^{(k)} - x^\star\rangle + \alpha^2\|g_k\|_2^2\end{aligned}$$

By definition of g_k, $f\left(x^\star\right) \geq f\left(x^{(k)}\right) + \langle g_k, x^\star - x^{(k)}\rangle$

Theorem: if α is small enough, the next point in the trajectory is closer to the optimal set.

Proof: $x^{(k+1)} = x^{(k)} - \alpha\, g_k$

$$\begin{aligned}\|x^{(k+1)} - x^\star\|_2^2 &= \|x^{(k)} - \alpha g_k - x^\star\|_2^2 \\ &= \|x^{(k)} - x^\star\|_2^2 - 2\alpha\langle g_k, x^{(k)} - x^\star\rangle + \alpha^2\|g_k\|_2^2\end{aligned}$$

By definition of g_k, $f\left(x^\star\right) \geq f\left(x^{(k)}\right) + \langle g_k, x^\star - x^{(k)}\rangle$

$$\|x^{(k+1)} - x^\star\|_2^2 \leq \|x^{(k)} - x^\star\|_2^2 - 2\alpha\left(f\left(x^{(k)}\right) - f(x^\star)\right) + \alpha^2\|g_k\|_2^2$$

Theorem: if α is small enough, the next point in the trajectory is closer to the optimal set.

Proof: $x^{(k+1)} = x^{(k)} - \alpha\, g_k$

$$\|x^{(k+1)} - x^\star\|_2^2 = \|x^{(k)} - \alpha g_k - x^\star\|_2^2$$
$$= \|x^{(k)} - x^\star\|_2^2 - 2\alpha\langle g_k, x^{(k)} - x^\star\rangle + \alpha^2\|g_k\|_2^2$$

By definition of g_k, $f\left(x^\star\right) \geq f\left(x^{(k)}\right) + \langle g_k, x^\star - x^{(k)}\rangle$

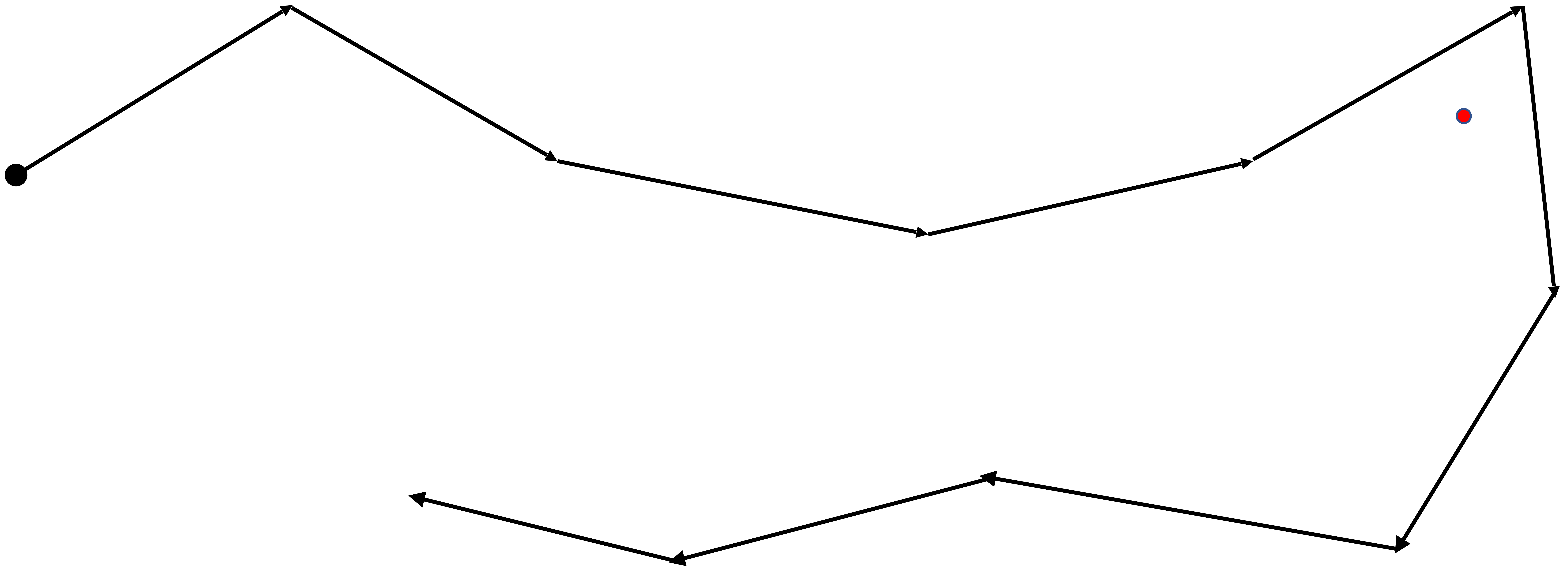
$$\|x^{(k+1)} - x^\star\|_2^2 \leq \|x^{(k)} - x^\star\|_2^2 - 2\alpha\left(f\left(x^{(k)}\right) - f(x^\star)\right) + \alpha^2\|g_k\|_2^2$$

$\therefore\ \|x^{(k+1)} - x^\star\|_2^2 < \|x^{(k)} - x^\star\|_2^2$ when $\alpha < 2\left(f\left(x^{(k)}\right) - f^\star\right)/\|g_k\|_2^2$

Theorem: if α is small enough, the next point in the trajectory is closer to the optimal set.

Theorem: if α is small enough, the next point in the trajectory is closer to the optimal set.

Does the trajectory stay close, or can it move far away?

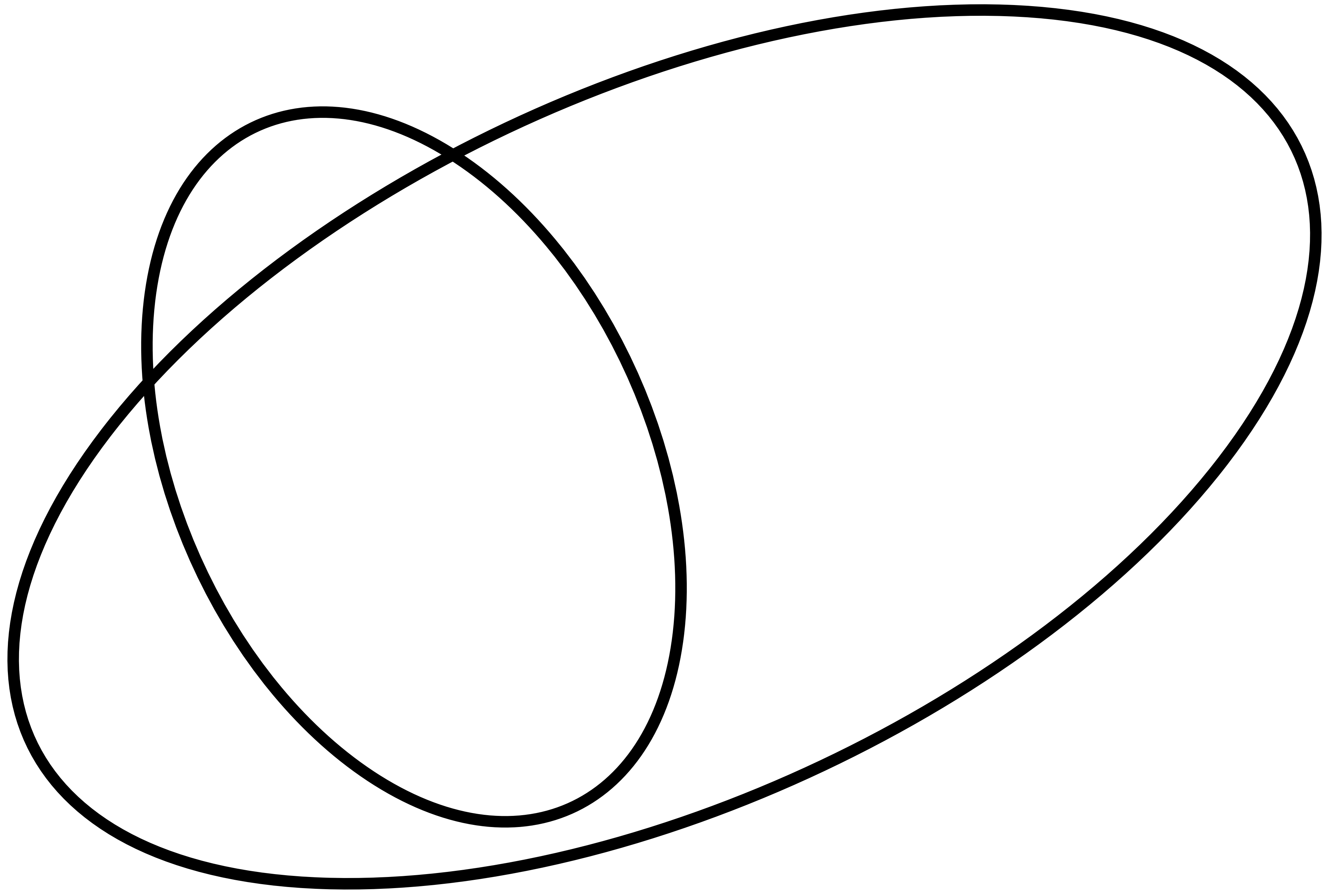


Hausdorff Distance

$$H(X, Y) = \max\left(\sup_{x \in X^\star} \inf_{y \in S_k} \|x - y\|_2, \sup_{y \in S_k} \inf_{x \in X^\star} \|x - y\|_2\right)$$

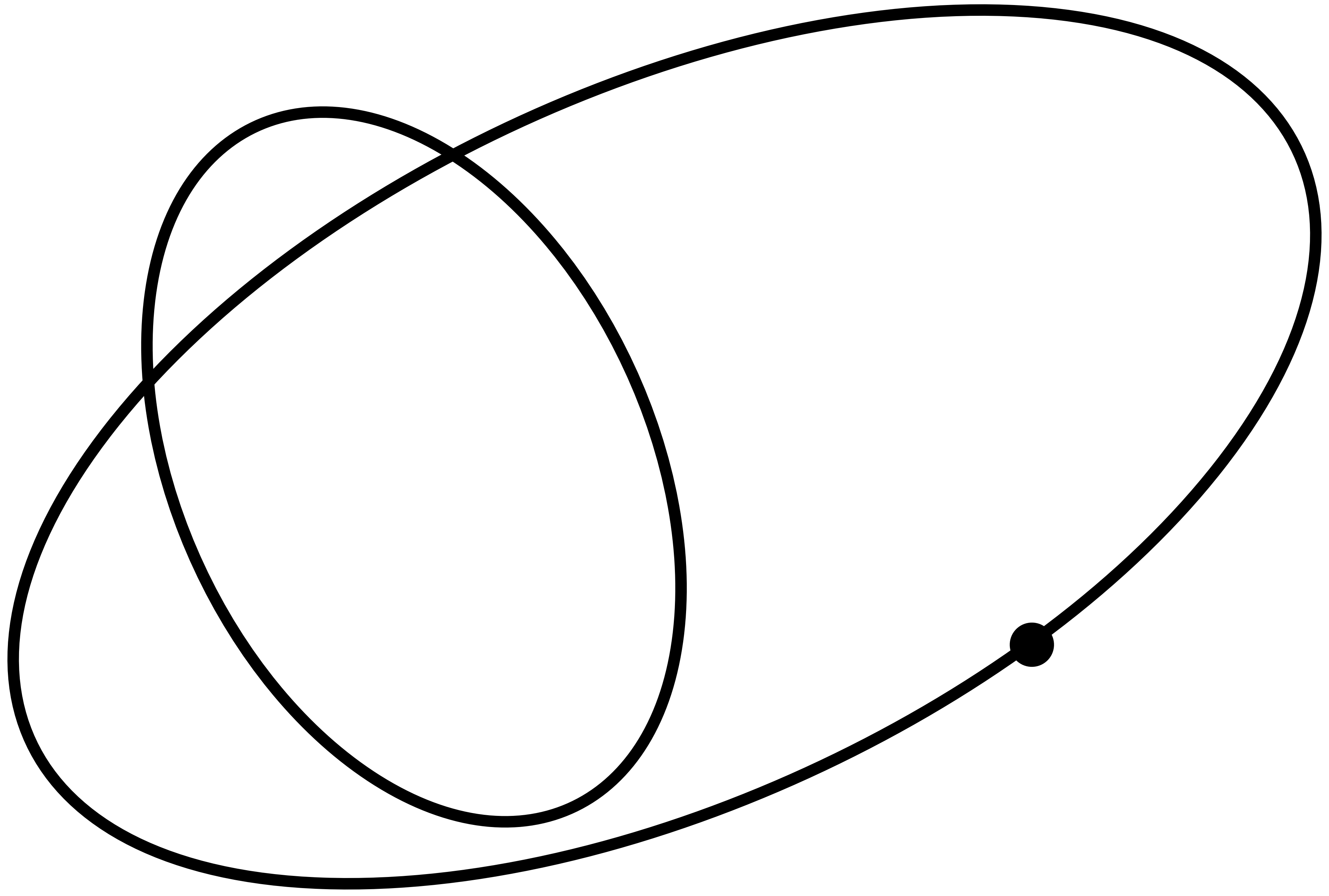
Hausdorff Distance

$$H(X,Y) = \max\left(\sup_{x\in X^\star}\inf_{y\in S_k}\|x-y\|_2, \sup_{y\in S_k}\inf_{x\in X^\star}\|x-y\|_2\right)$$



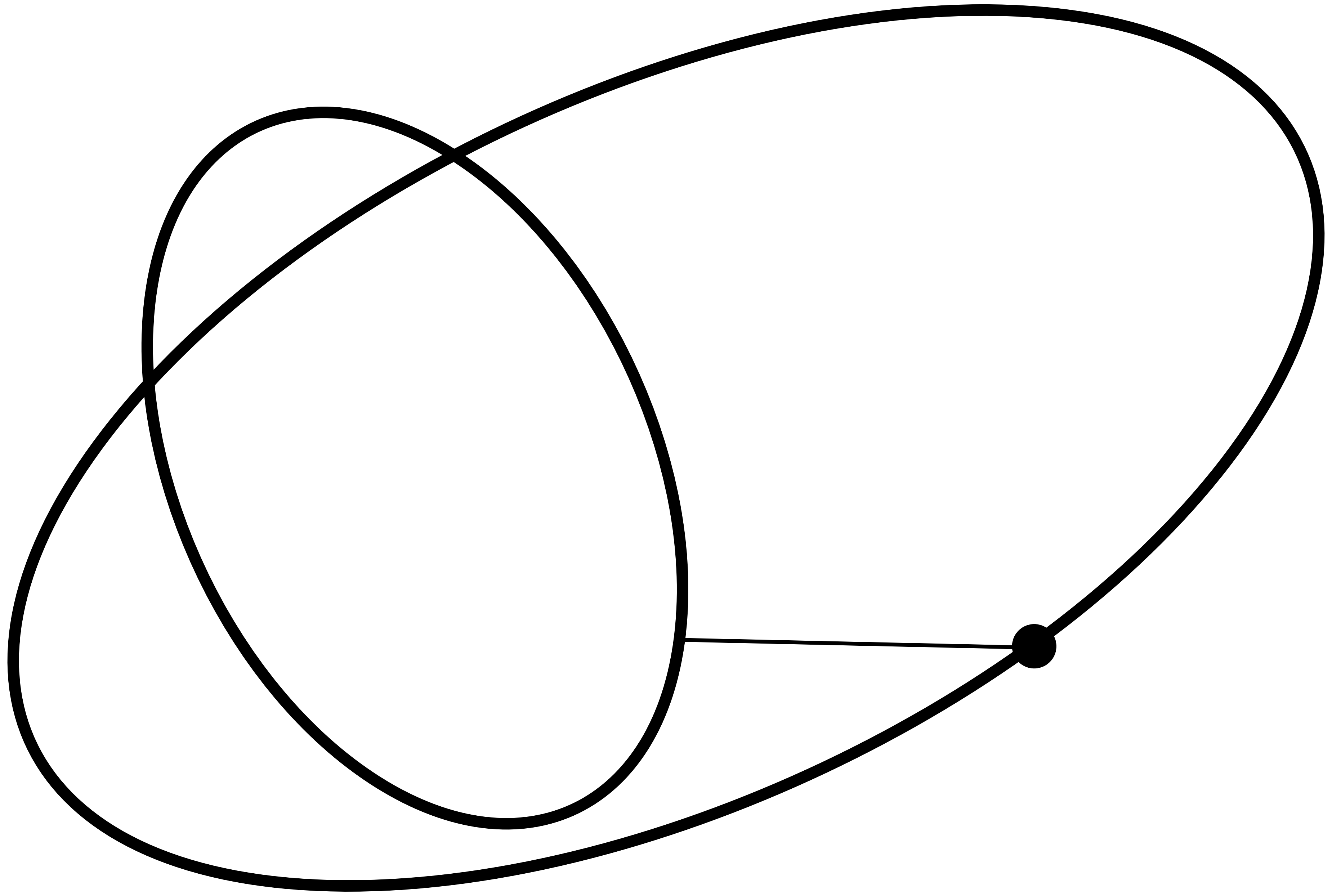
Hausdorff Distance

$$H(X,Y) = \max\left(\sup_{x\in X^\star}\inf_{y\in S_k}\|x-y\|_2, \sup_{y\in S_k}\inf_{x\in X^\star}\|x-y\|_2\right)$$



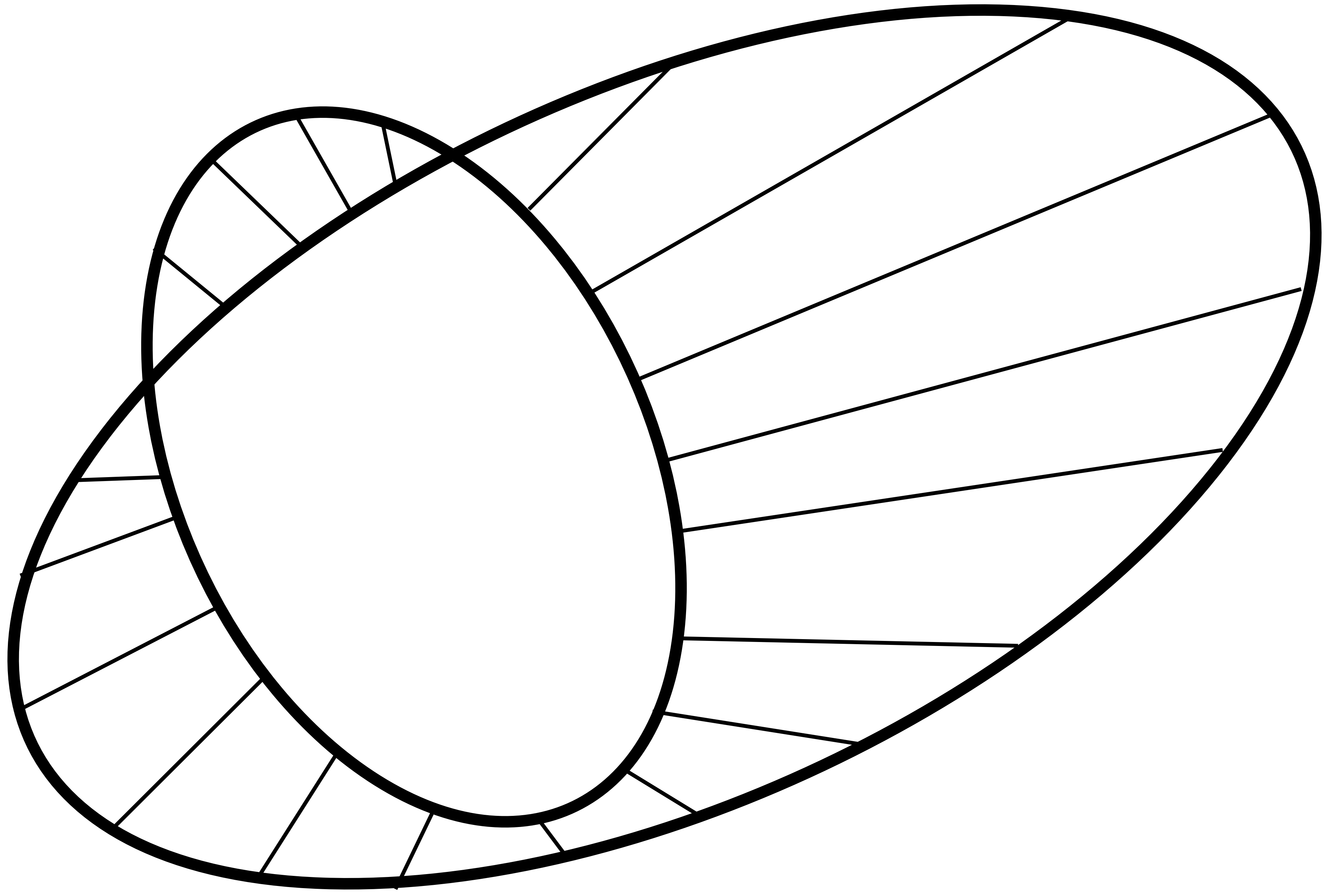
Hausdorff Distance

$$H(X,Y) = \max\left(\sup_{x \in X^\star} \inf_{y \in S_k} \|x - y\|_2, \sup_{y \in S_k} \inf_{x \in X^\star} \|x - y\|_2\right)$$



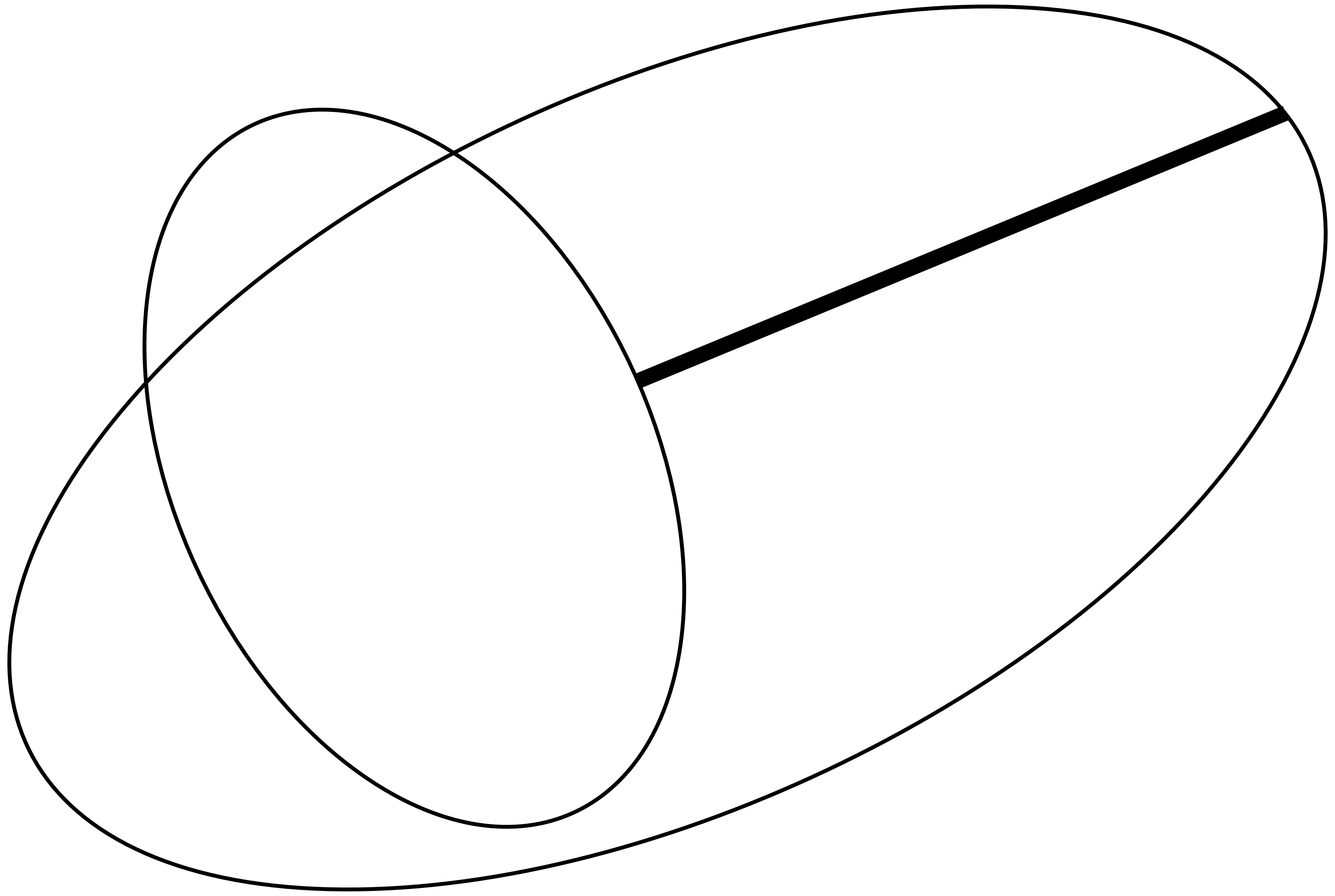
Hausdorff Distance

$$H(X, Y) = \max\left(\sup_{x \in X^\star} \inf_{y \in S_k} \|x - y\|_2, \sup_{y \in S_k} \inf_{x \in X^\star} \|x - y\|_2\right)$$



Hausdorff Distance

$$H(X, Y) = \max \left(\sup_{x \in X^\star} \inf_{y \in S_k} \|x - y\|_2, \sup_{y \in S_k} \inf_{x \in X^\star} \|x - y\|_2 \right)$$



Sublevel set: $S_k = \{x \in \mathbb{R}^N \mid f(x) \leq f(x^{(k)})\}$

Sublevel set: $$S_k = \{x \in \mathbb{R}^N \mid f(x) \leq f(x^{(k)})\}$$

$$H_k = H\left(S_k, X^\star\right)$$

Sublevel set: $S_k = \{x \in \mathbb{R}^N \mid f(x) \leq f(x^{(k)})\}$

$$H_k = H\left(S_k, X^\star\right)$$

Theorem: the trajectory always stays within a bounded distance of H_k.

Assume: $\|g\|_2 < G \quad \forall \quad g \in \partial f$

$$\alpha < \left(f(x^{(k)}) - f^\star\right)/G^2$$

Assume: $\|g\|_2 < G \quad \forall \quad g \in \partial f$

$$\alpha < \left(f(x^{(k)}) - f^\star\right)/G^2$$

Theorem: $\operatorname{dist}\left(x^{(m)}, X^\star\right) < H_k + 2\left(f\left(x^{(k)}\right) - f^\star\right)/G$

Assume: $\|g\|_2 < G \quad \forall \quad g \in \partial f$

$$\alpha < \left(f(x^{(k)}) - f^\star\right)/G^2$$

Theorem: $\operatorname{dist}\left(x^{(m)}, X^\star\right) < H_k + 2\left(f\left(x^{(k)}\right) - f^\star\right)/G$

Proof (by induction):

Assume: $\|g\|_2 < G \quad \forall \quad g \in \partial f$

$$\alpha < \left(f(x^{(k)}) - f^\star \right) / G^2$$

Theorem: $\mathrm{dist}\left(x^{(m)}, X^\star \right) < H_k + 2 \left(f \left(x^{(k)} \right) - f^\star \right) / G$

Proof (by induction):

Base step: since $m = k$, $x^{(k)} \in S_k$. Therefore the theorem is true by the previous theorem.

Inductive step:

Case 1: $x^{(k)} \in S_k$

Inductive step:

Case 1: $x^{(k)} \in S_k$

Since $\alpha < \left(f(x^{(k)}) - f^\star \right) / G^2,$

$$\|x^{(m+1)} - x^{(m)}\|_2 = \alpha \|g^{(m)}\|_2$$

Inductive step:

Case 1: $x^{(k)} \in S_k$

Since $\alpha < \left(f(x^{(k)}) - f^\star\right)/G^2$,

$$\|x^{(m+1)} - x^{(m)}\|_2 = \alpha\|g^{(m)}\|_2 \leq aG$$

Inductive step:

Case 1: $x^{(k)} \in S_k$

Since $\alpha < \left(f(x^{(k)}) - f^\star\right)/G^2$,

$$\|x^{(m+1)} - x^{(m)}\|_2 = \alpha\|g^{(m)}\|_2 \leq aG < 2\left(f(x^{(k)}) - f^\star\right)/G$$

Inductive step:

Case 1: $x^{(k)} \in S_k$

Since $\alpha < \left(f(x^{(k)}) - f^\star \right) / G^2,$

$$\|x^{(m+1)} - x^{(m)}\|_2 = \alpha \|g^{(m)}\|_2 \leq aG < 2 \left(f(x^{(k)}) - f^\star \right) / G$$

By the triangle inequality,

$$\mathrm{dist}\left(x^{(m+1)}, X^\star \right) \leq \mathrm{dist}\left(x^{(m+1)}, x^{(m)} \right) + \mathrm{dist}\left(x^{(m)}, X^\star \right)$$

Inductive step:

Case 1: $x^{(k)} \in S_k$

Since $\alpha < \left(f(x^{(k)}) - f^\star \right) / G^2$,

$$\|x^{(m+1)} - x^{(m)}\|_2 = \alpha \|g^{(m)}\|_2 \leq aG < 2 \left(f(x^{(k)}) - f^\star \right) / G$$

By the triangle inequality,

$$\begin{aligned} \operatorname{dist}\left(x^{(m+1)}, X^\star\right) &\leq \operatorname{dist}\left(x^{(m+1)}, x^{(m)}\right) + \operatorname{dist}\left(x^{(m)}, X^\star\right) \\ &< 2\left(f(x^{(k)}) - f^\star\right)/G + H_k \end{aligned}$$

Inductive step:

Case 1: $x^{(k)} \in S_k$

Since $\alpha < \left(f(x^{(k)}) - f^\star\right)/G^2$,

$$\|x^{(m+1)} - x^{(m)}\|_2 = \alpha \|g^{(m)}\|_2 \leq aG < 2\left(f(x^{(k)}) - f^\star\right)/G$$

By the triangle inequality,

$$\begin{aligned}\operatorname{dist}\left(x^{(m+1)}, X^\star\right) &\leq \operatorname{dist}\left(x^{(m+1)}, x^{(m)}\right) + \operatorname{dist}\left(x^{(m)}, X^\star\right) \\ &< 2\left(f(x^{(k)}) - f^\star\right)/G + H_k\end{aligned}$$

Case 2: $x^{(k)} \notin S_k$.

Inductive step:

Case 1: $x^{(k)} \in S_k$

Since $\alpha < \left(f(x^{(k)}) - f^\star\right)/G^2$,

$$\|x^{(m+1)} - x^{(m)}\|_2 = \alpha \|g^{(m)}\|_2 \leq aG < 2\left(f(x^{(k)}) - f^\star\right)/G$$

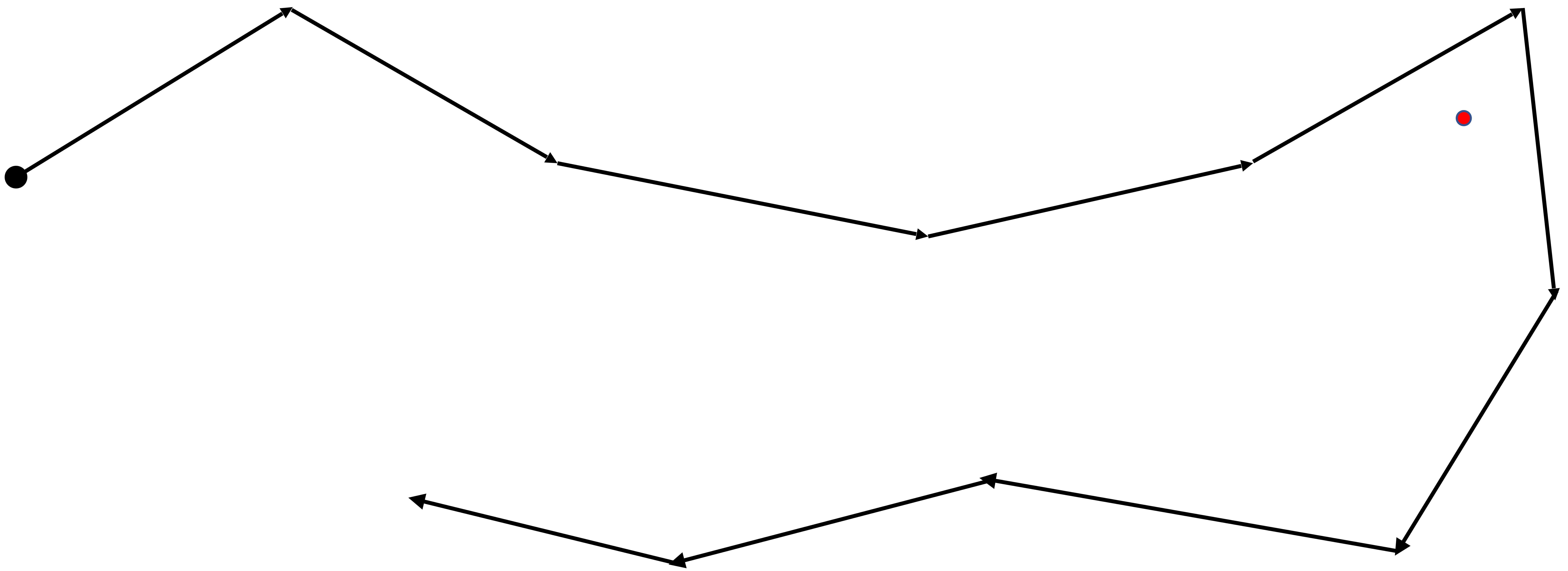
By the triangle inequality,

$$\begin{aligned} \operatorname{dist}\left(x^{(m+1)}, X^\star\right) &\leq \operatorname{dist}\left(x^{(m+1)}, x^{(m)}\right) + \operatorname{dist}\left(x^{(m)}, X^\star\right) \\ &< 2\left(f(x^{(k)}) - f^\star\right)/G + H_k \end{aligned}$$

Case 2: $x^{(k)} \notin S_k$.

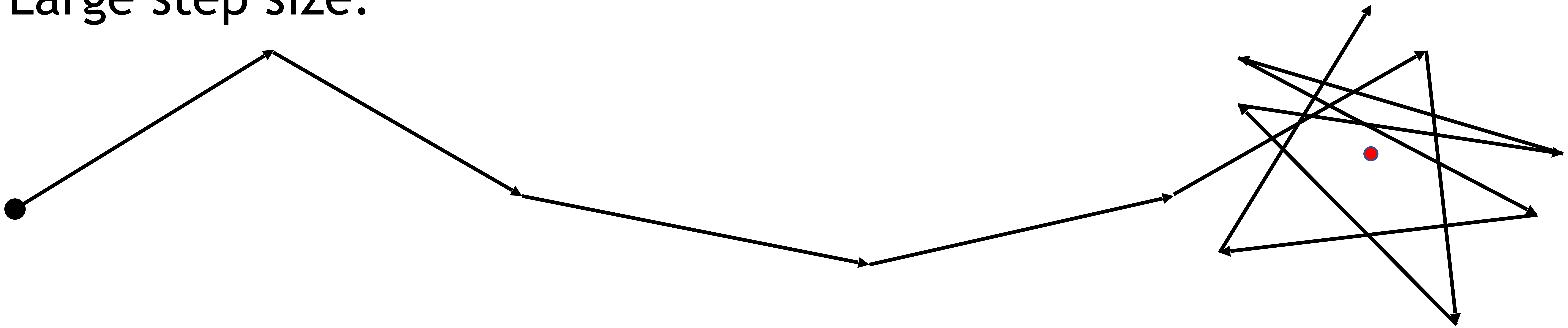
Then we move closer to the optimal set (as before). ■

Theorem: the trajectory always stays within a bounded distance of H_k.

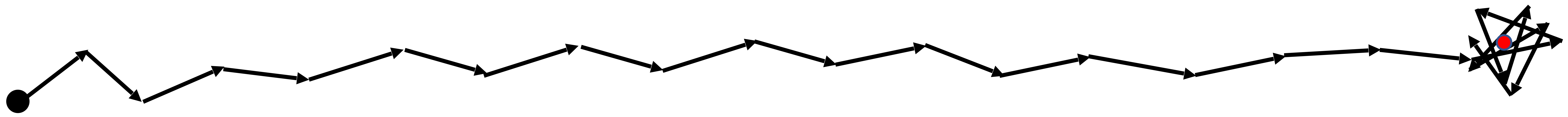


Theorem: the trajectory always stays within a bounded distance of H_k.

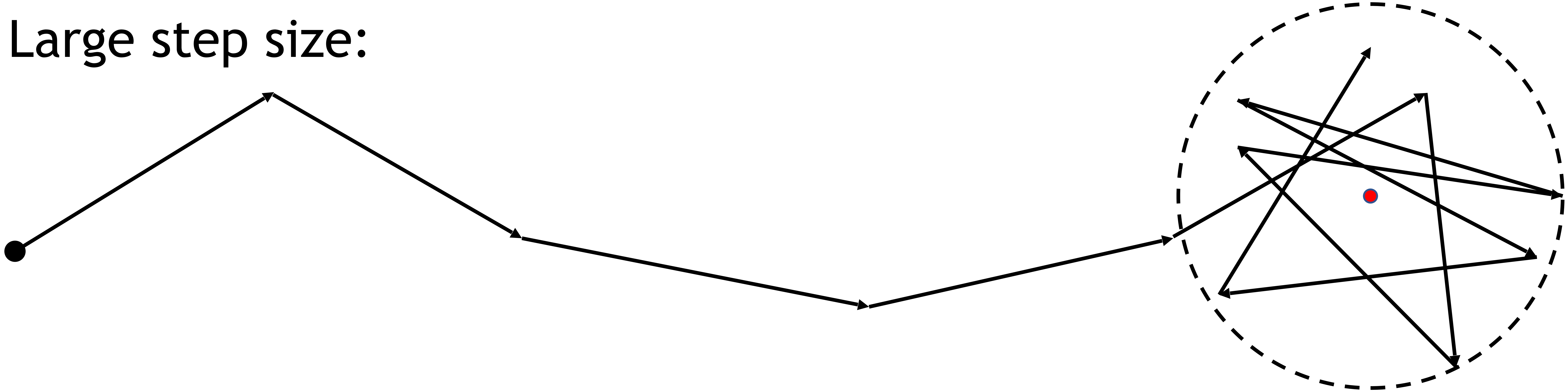
Large step size:



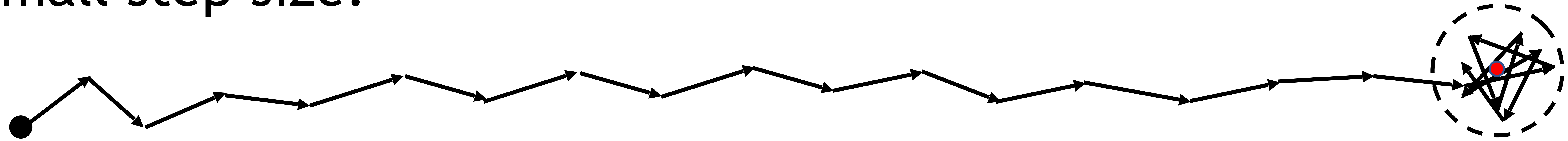
Small step size:



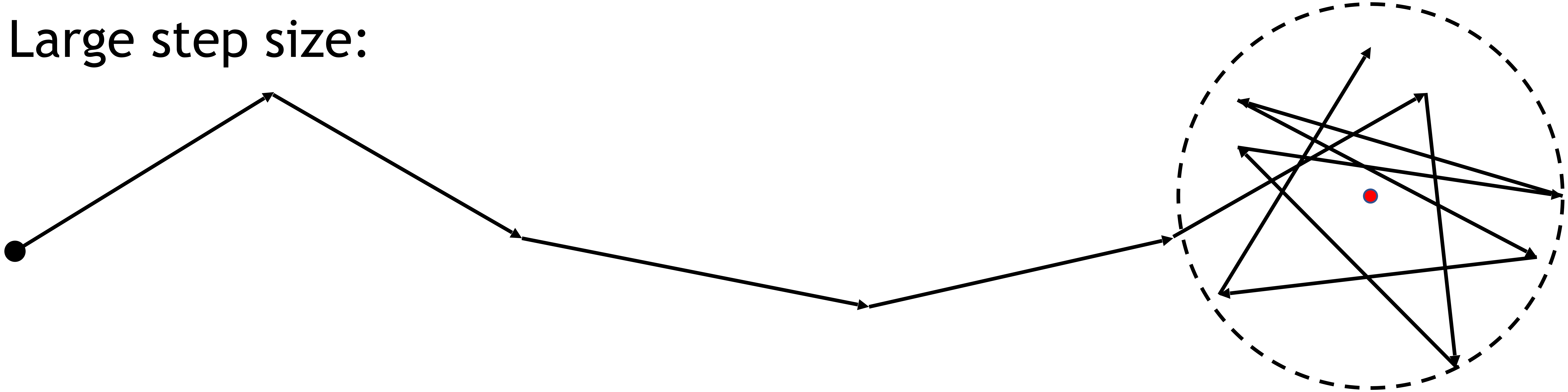
Large step size:



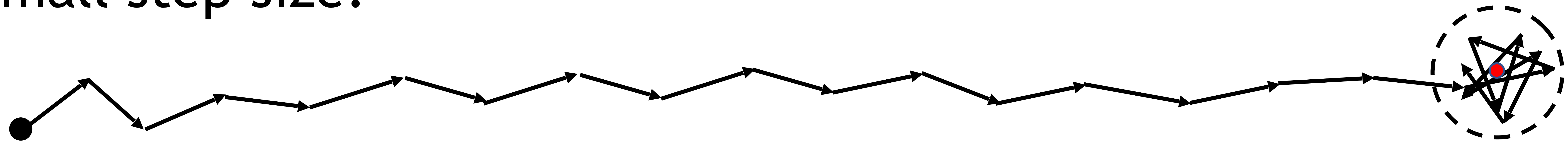
Small step size:



Large step size:



Small step size:

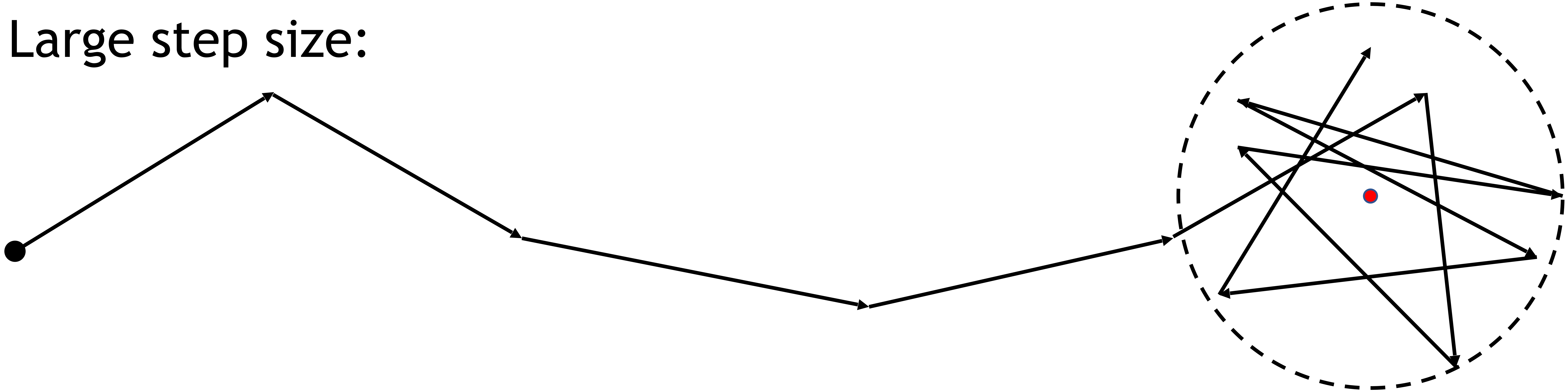


Two possibilities

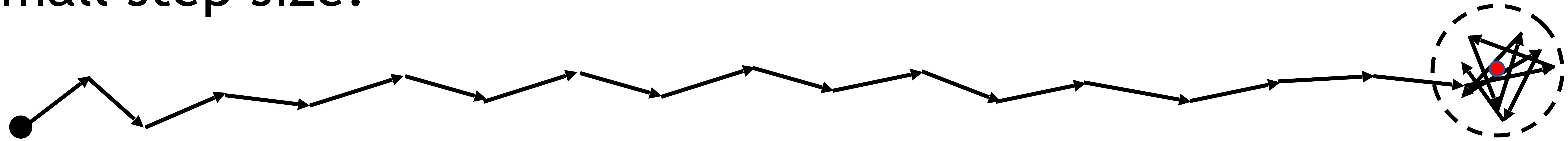
1) We're converging to a solution

2) We're bouncing around the solution

Large step size:

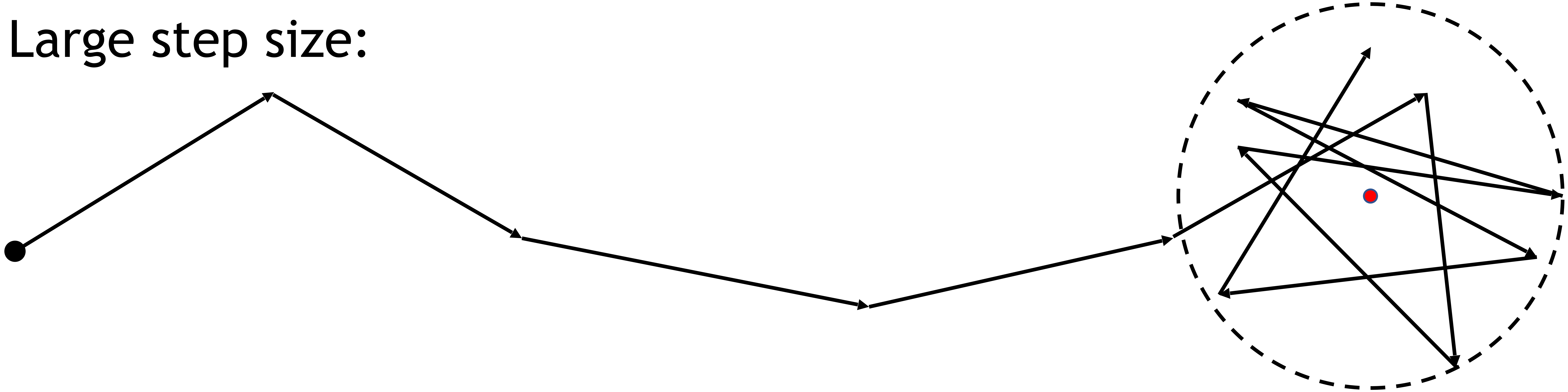


Small step size:

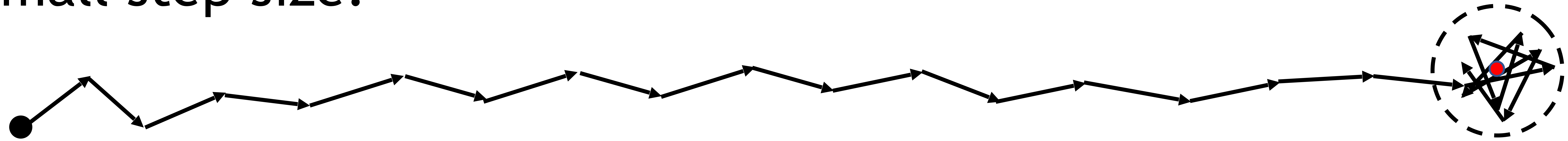


Can we detect when the trajectory is bouncing around the optimal set?

Large step size:



Small step size:

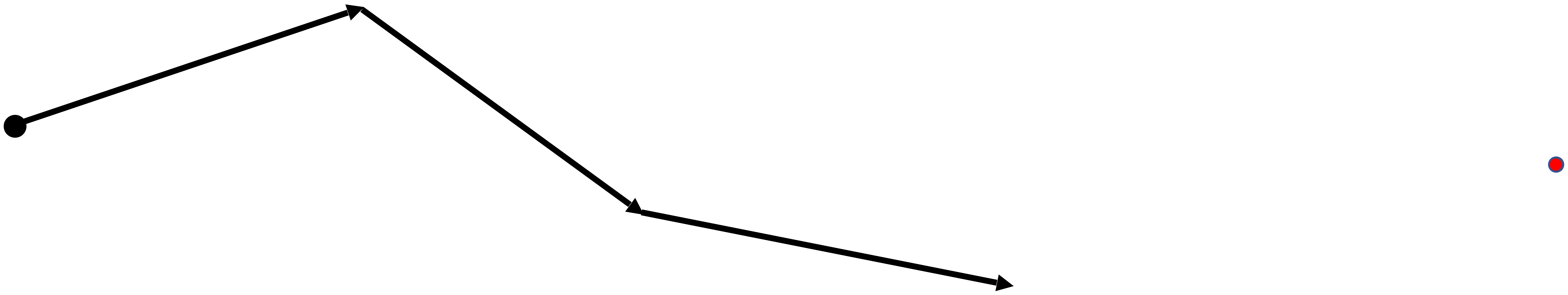


Can we detect when the trajectory is bouncing around the optimal set?

If so, can we make use of this knowledge?

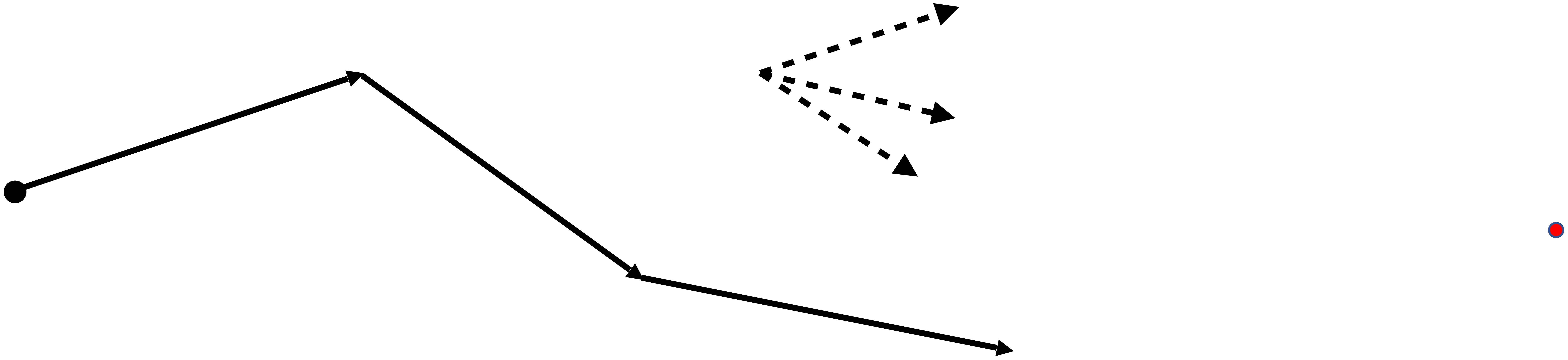
Metric of bouncing: the average normalized subgradient vector

$$u_k = \frac{-g_k}{\|g_k\|_2} \qquad m = \left\| \frac{\sum u_k}{n} \right\|_2$$



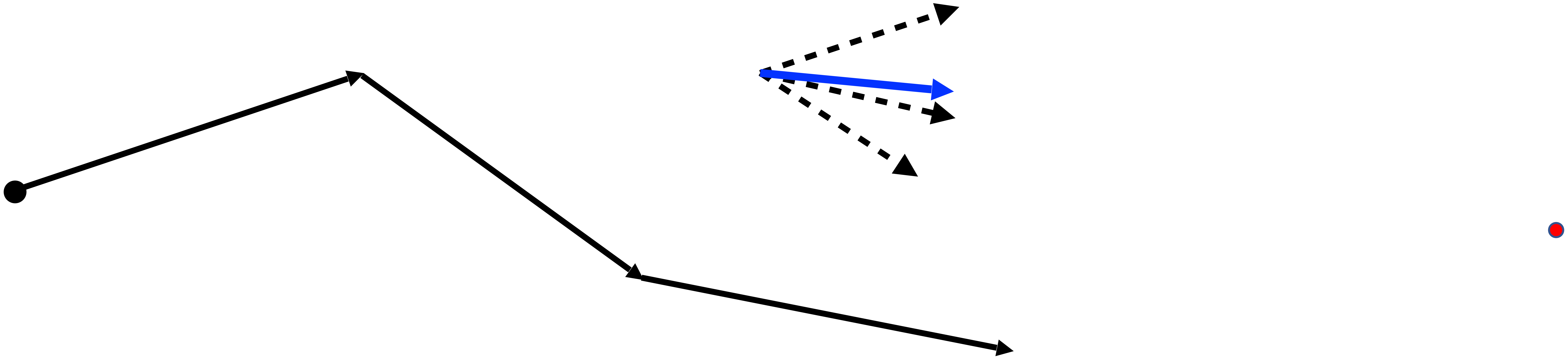
Metric of bouncing: the average normalized subgradient vector

$$u_k = \frac{-g_k}{\|g_k\|_2} \qquad\qquad m = \left\| \frac{\sum u_k}{n} \right\|_2$$



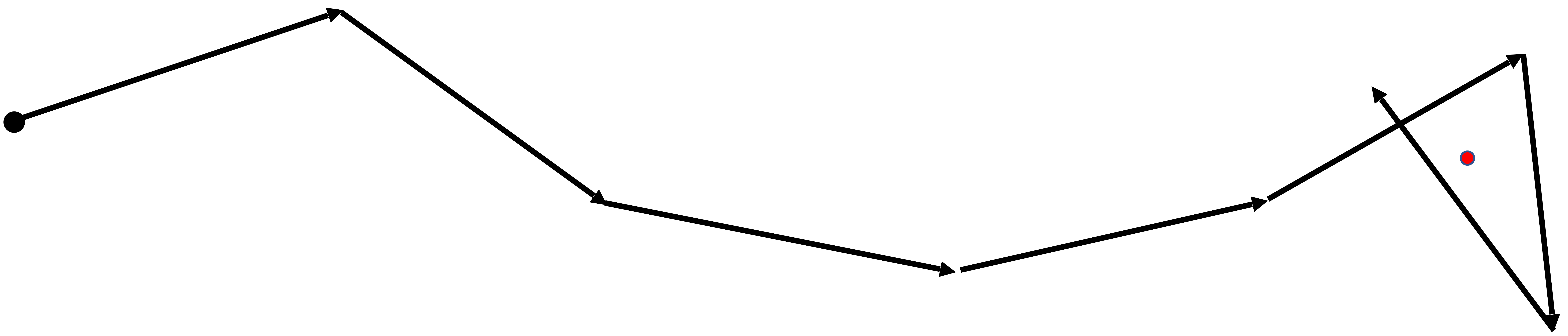
Metric of bouncing: the average normalized subgradient vector

$$u_k = \frac{-g_k}{\|g_k\|_2} \qquad m = \left\| \frac{\sum u_k}{n} \right\|_2$$



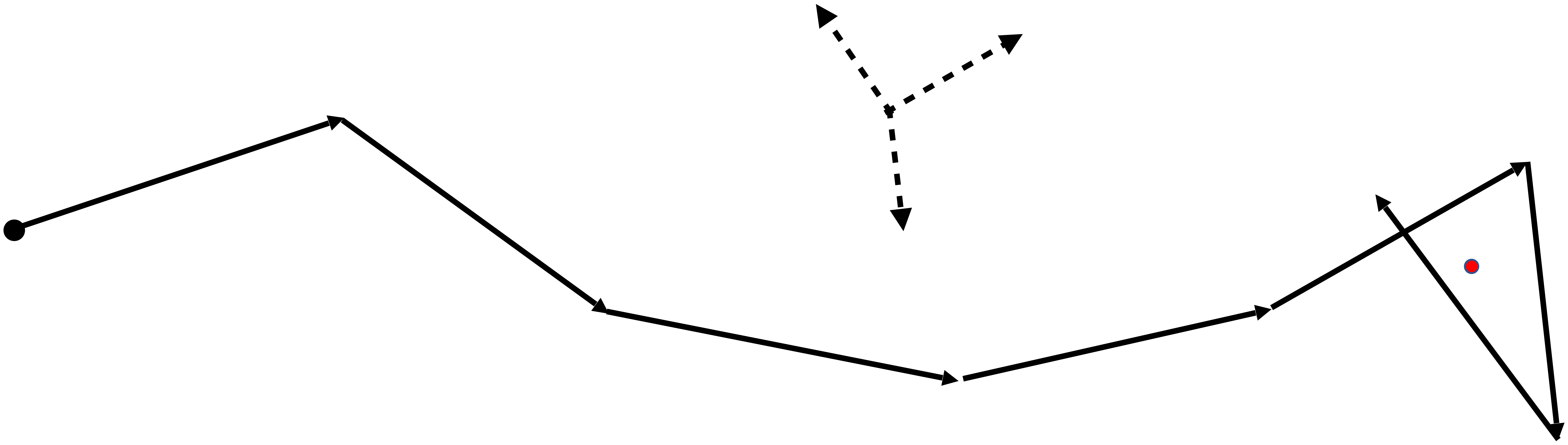
Metric of bouncing: the average normalized subgradient vector

$$u_k = \frac{-g_k}{\|g_k\|_2} \qquad\qquad m = \left\| \frac{\sum u_k}{n} \right\|_2$$



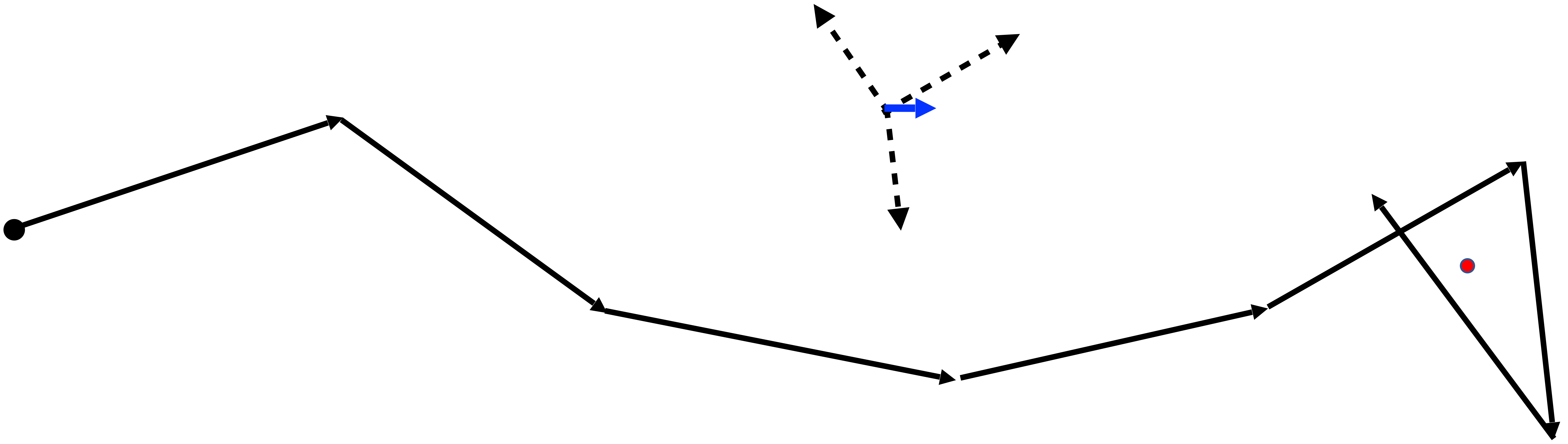
Metric of bouncing: the average normalized subgradient vector

$$u_k = \frac{-g_k}{\|g_k\|_2} \qquad\qquad m = \left\| \frac{\sum u_k}{n} \right\|_2$$



Metric of bouncing: the average normalized subgradient vector

$$u_k = \frac{-g_k}{\|g_k\|_2} \qquad\qquad m = \left\| \frac{\sum u_k}{n} \right\|_2$$



Recap

Recap

We know that if the step size is small enough, then the next trajectory point is closer to the solution than the current point.

Recap

We know that if the step size is small enough, then the next trajectory point is closer to the solution than the current point.

We know that the trajectory will bounce around the solution if the step size is too large.

Recap

We know that if the step size is small enough, then the next trajectory point is closer to the solution than the current point.

We know that the trajectory will bounce around the solution if the step size is too large.

We know that we can detect when the trajectory is bouncing around the solution.

Adaptive Bounding Method

Adaptive Bounding Method

If the bouncing metric is below a threshold

Reduce the step size: $\alpha_k = r\,\alpha_{k-1} \quad 0 < r < 1$

Adaptive Bounding Method

If the bouncing metric is below a threshold

Reduce the step size: $\alpha_k = r\,\alpha_{k-1} \qquad 0 < r < 1$

Perform a subgradient descent update

$$x^{(k+1)} = x^{(k)} - \alpha_k\, g\left(x^{(k)}\right)$$

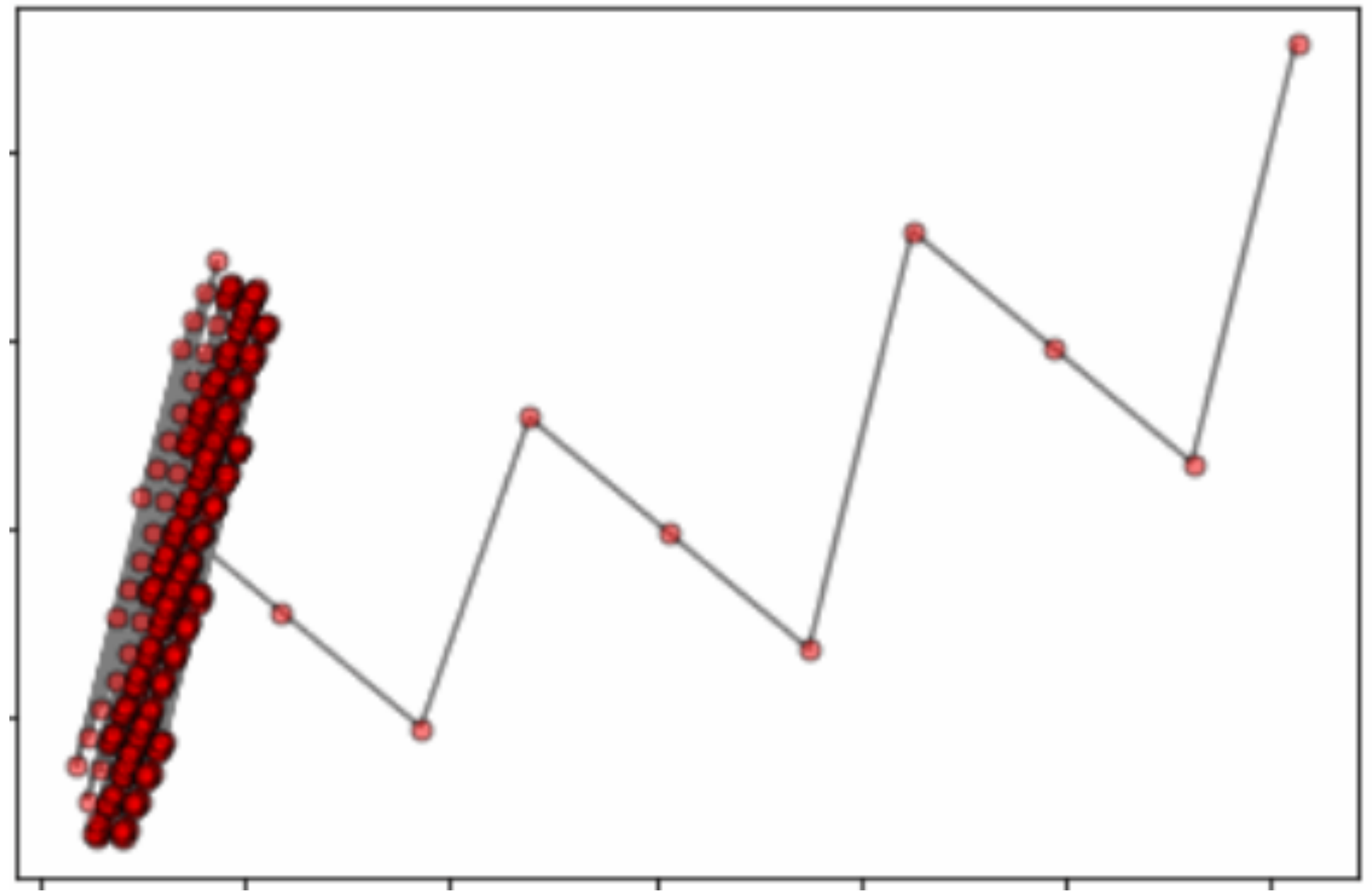
Minimize the maximum of a set of affine functions

$$f(x) = \max\left(A_i\, x + b_i\right) \qquad A \in \mathbb{R}^{500\times 2} \qquad b \in \mathbb{R}^{500}$$

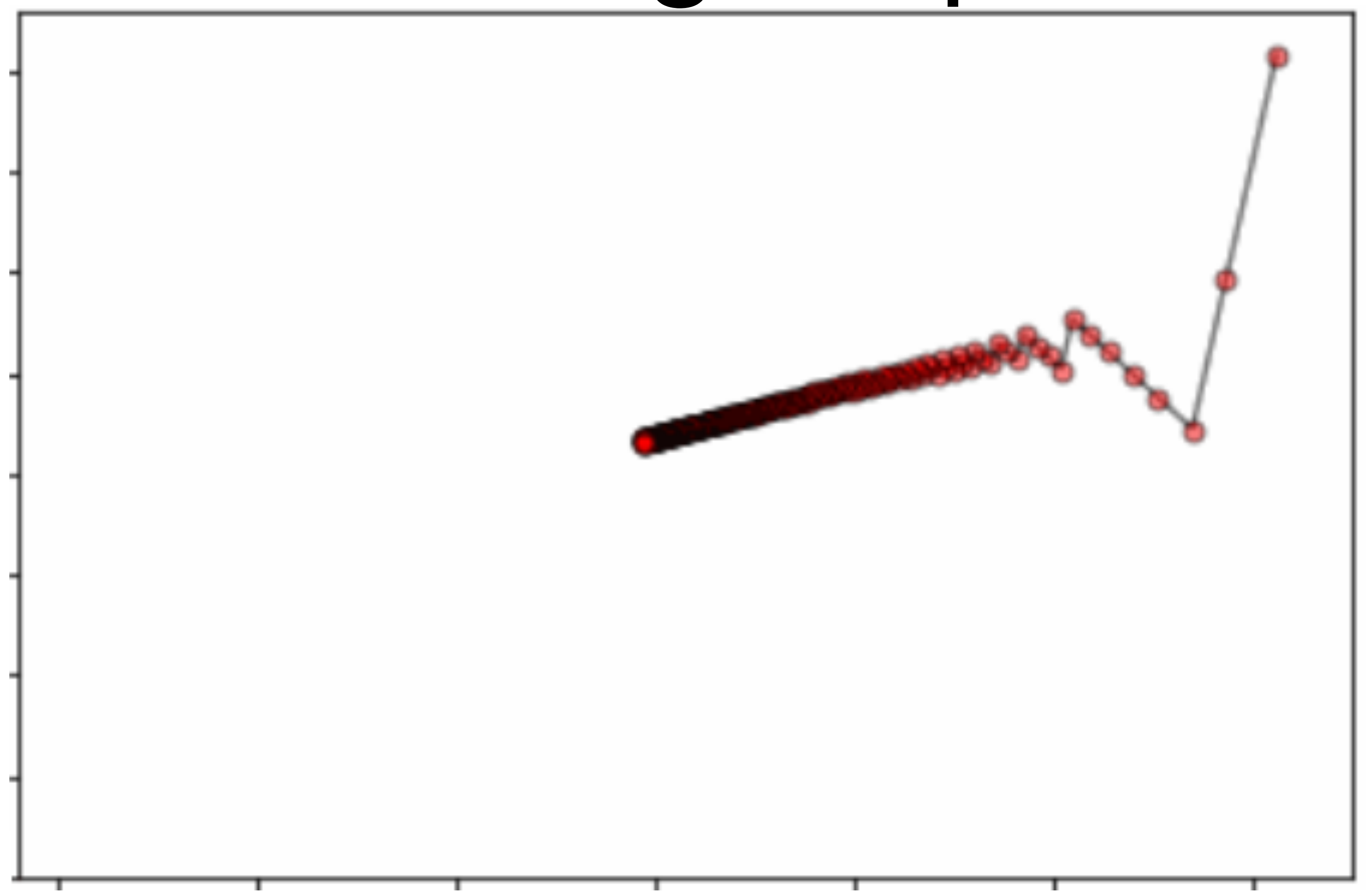
Minimize the maximum of a set of affine functions

$$f(x) = \max\left(A_i\, x + b_i\right) \qquad A \in \mathbb{R}^{500\times 2} \qquad b \in \mathbb{R}^{500}$$

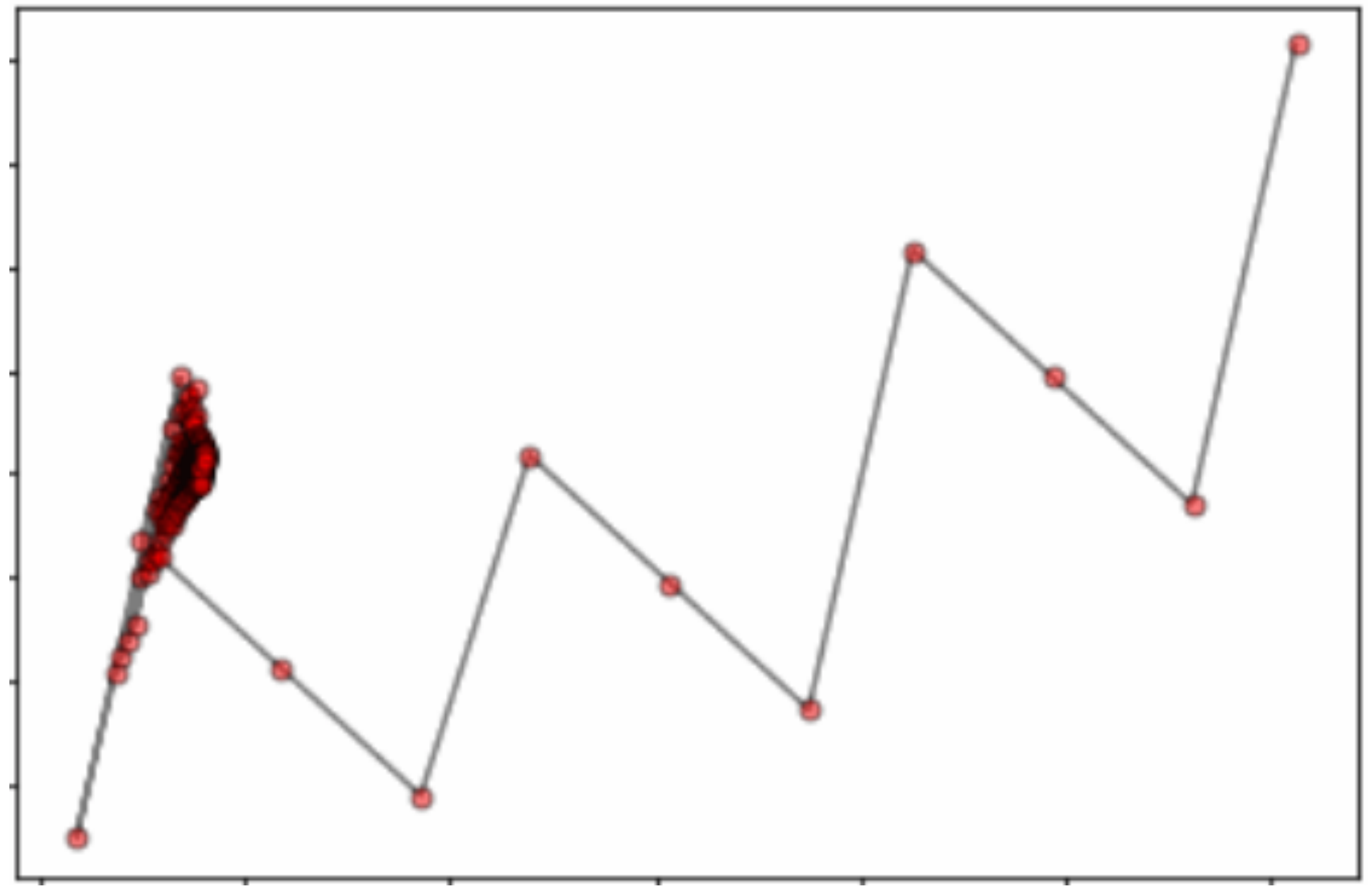
Constant Step Size



Decreasing Step Size

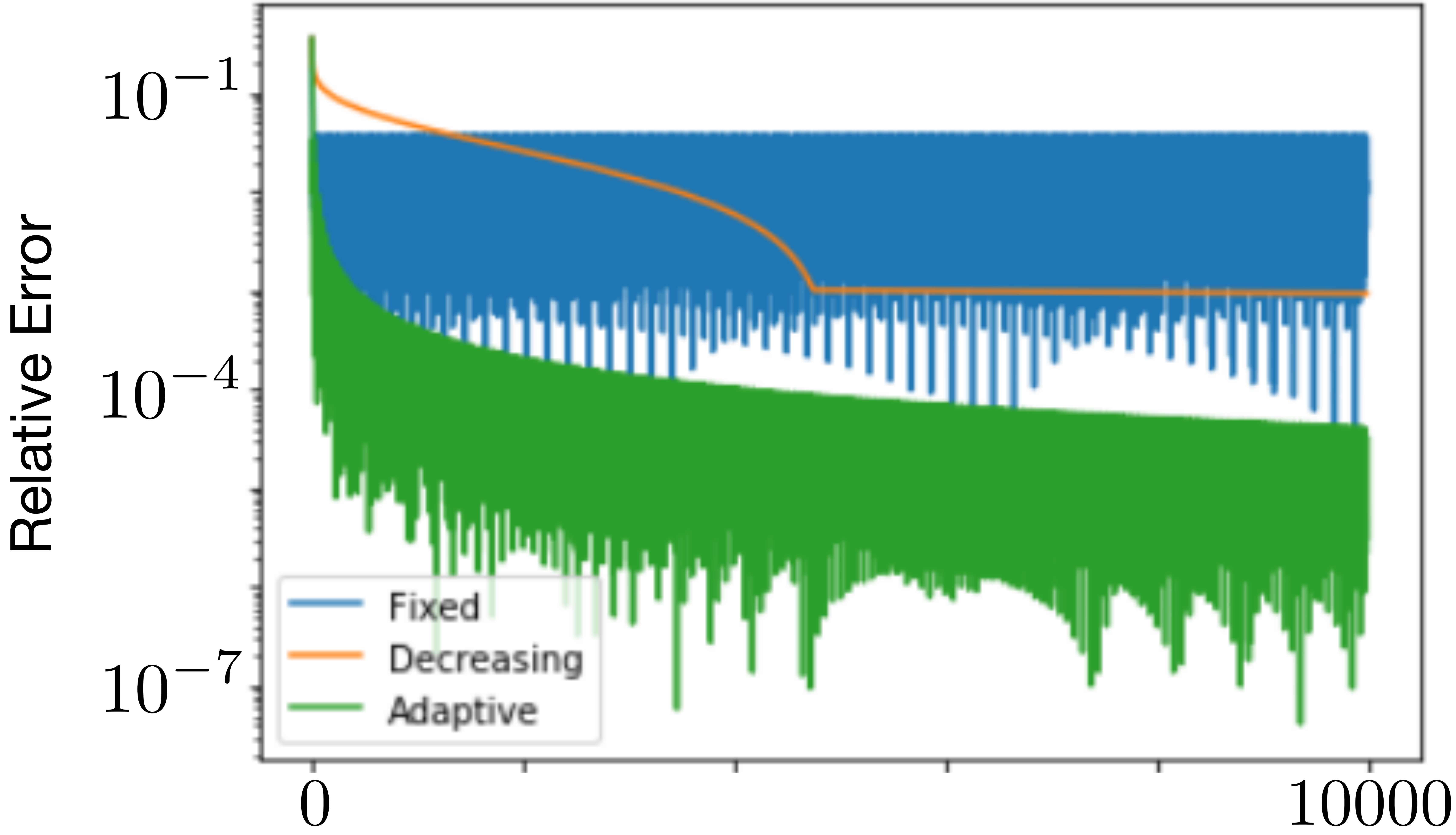


Adaptive Bounding



Minimize the maximum of a set of affine functions

$f(x) = \max\left(A_i\, x + b_i\right) \qquad A \in \mathbb{R}^{500\times 2} \qquad b \in \mathbb{R}^{500}$

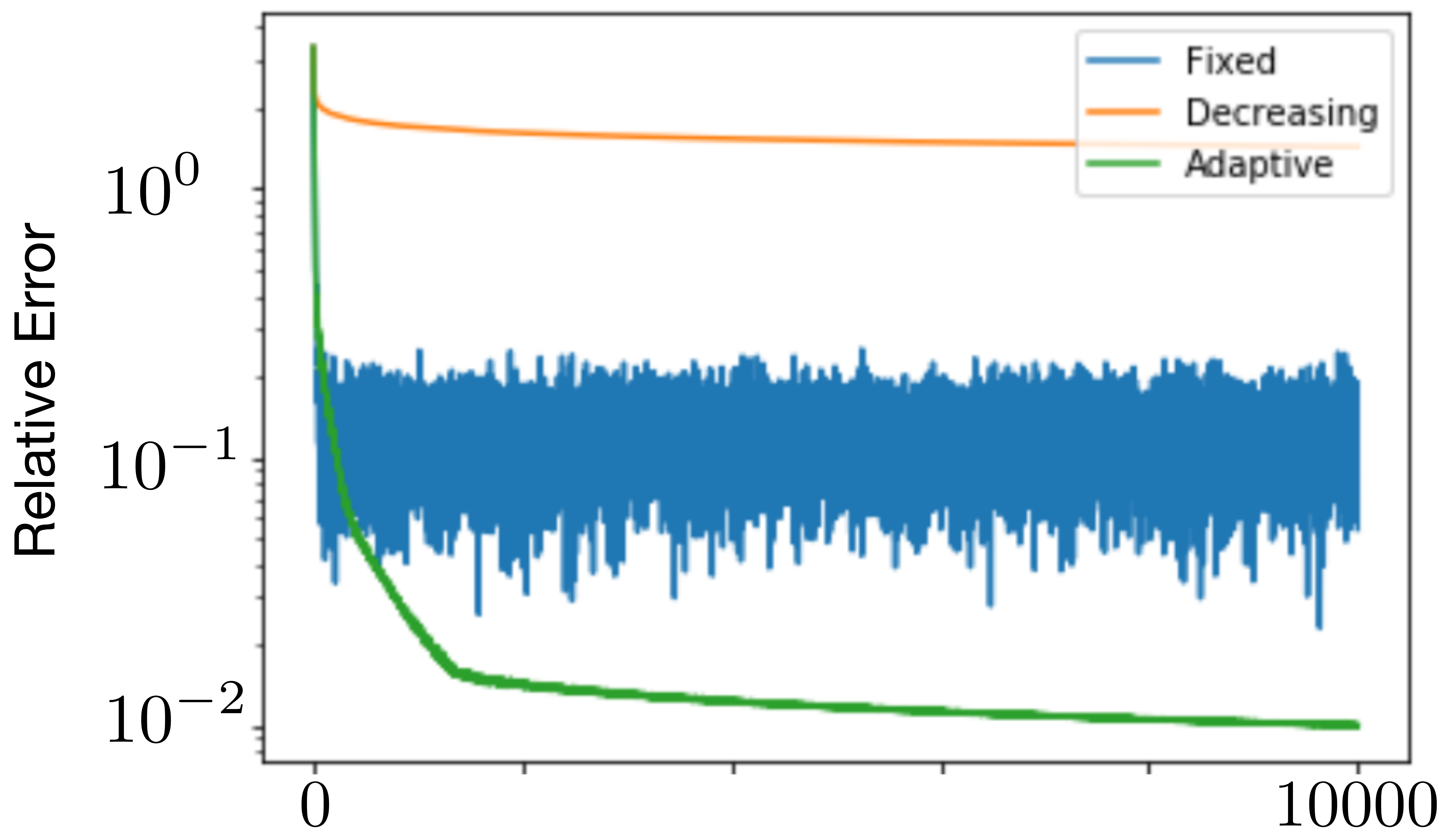


Minimize the maximum of a set of affine functions

$$f(x) = \max\left(A_i\, x + b_i\right) \qquad A \in \mathbb{R}^{200\times 10} \quad b \in \mathbb{R}^{200}$$

Minimize the maximum of a set of affine functions (larger problem)

$$f(x) = \max\left(A_i\, x + b_i\right) \qquad A \in \mathbb{R}^{200\times 10} \quad b \in \mathbb{R}^{200}$$

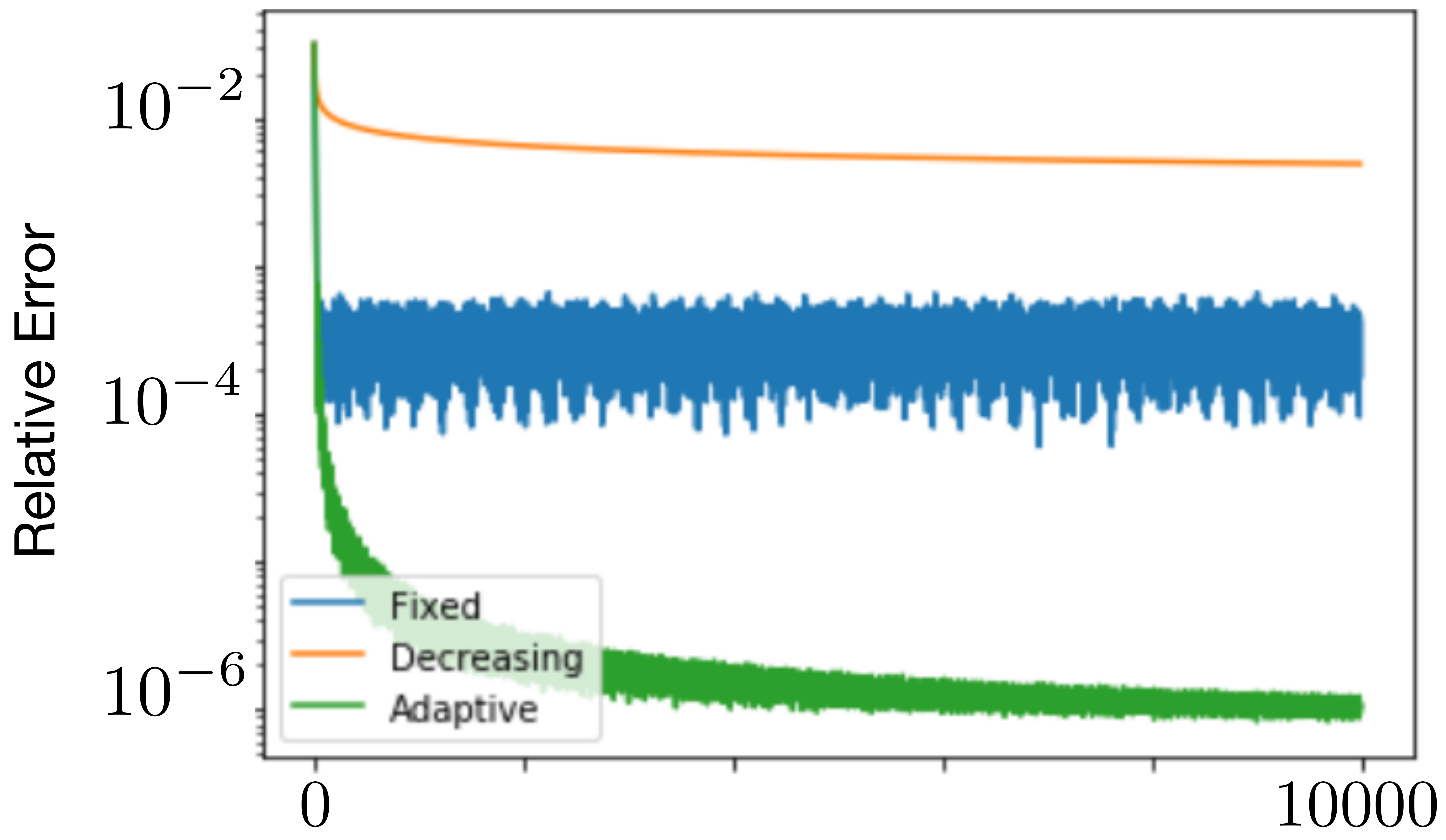


Lasso Problem

$$f(x) = \frac{1}{2}\|A\,x - b\|_2^2 + \lambda\|x\|_1 \qquad A \in \mathbb{R}^{200\times 20} \qquad b \in \mathbb{R}^{200}$$

Lasso Problem

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 \qquad A \in \mathbb{R}^{200\times 20} \qquad b \in \mathbb{R}^{200}$$

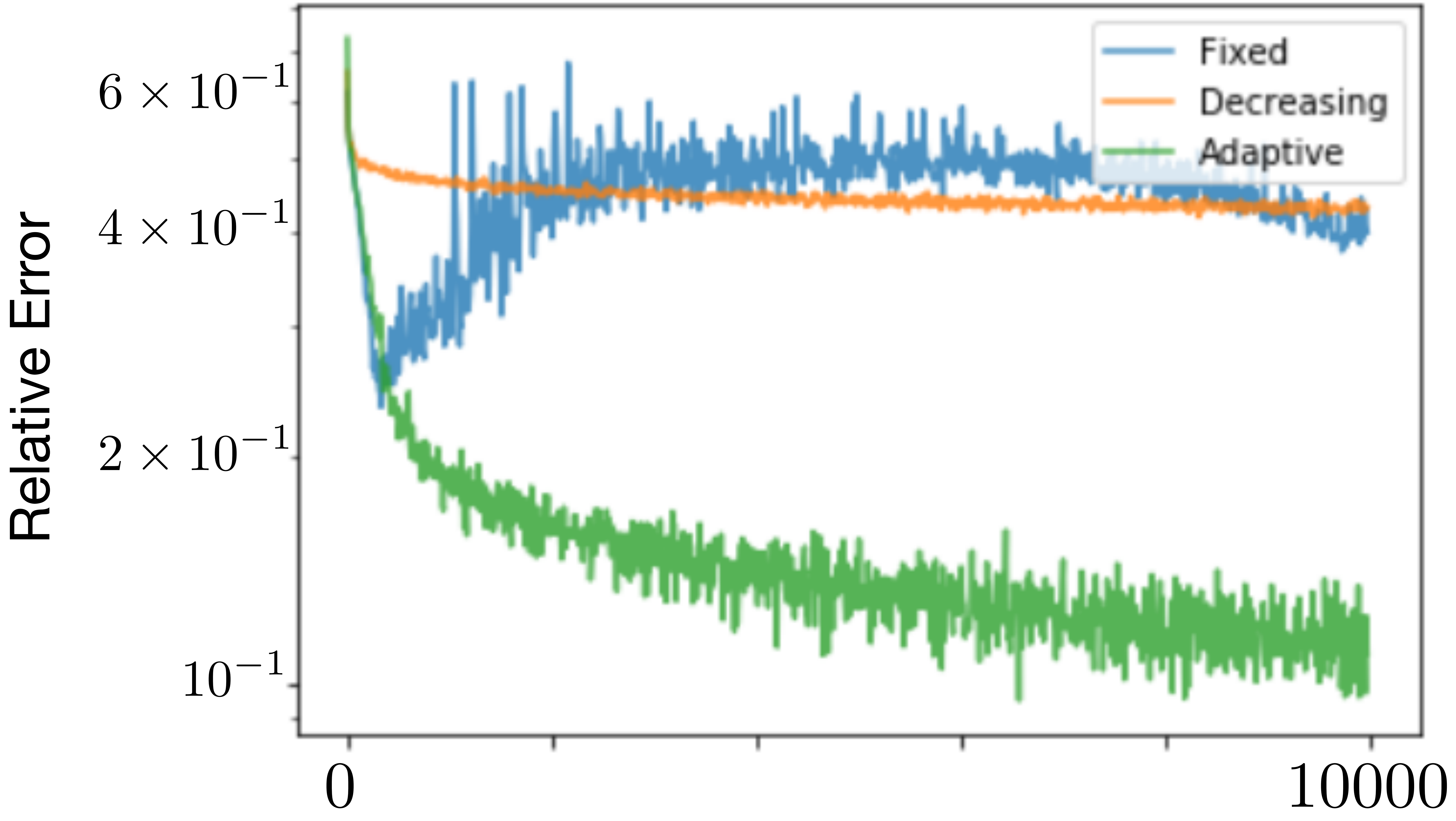


Neural Network

RESNET on CIFAR-10 dataset with L1 norm regularization

Neural Network

RESNET on CIFAR-10 dataset with L1 norm regularization



Conclusion

We have developed an apaptive method for determining the step size with subgradient descent.

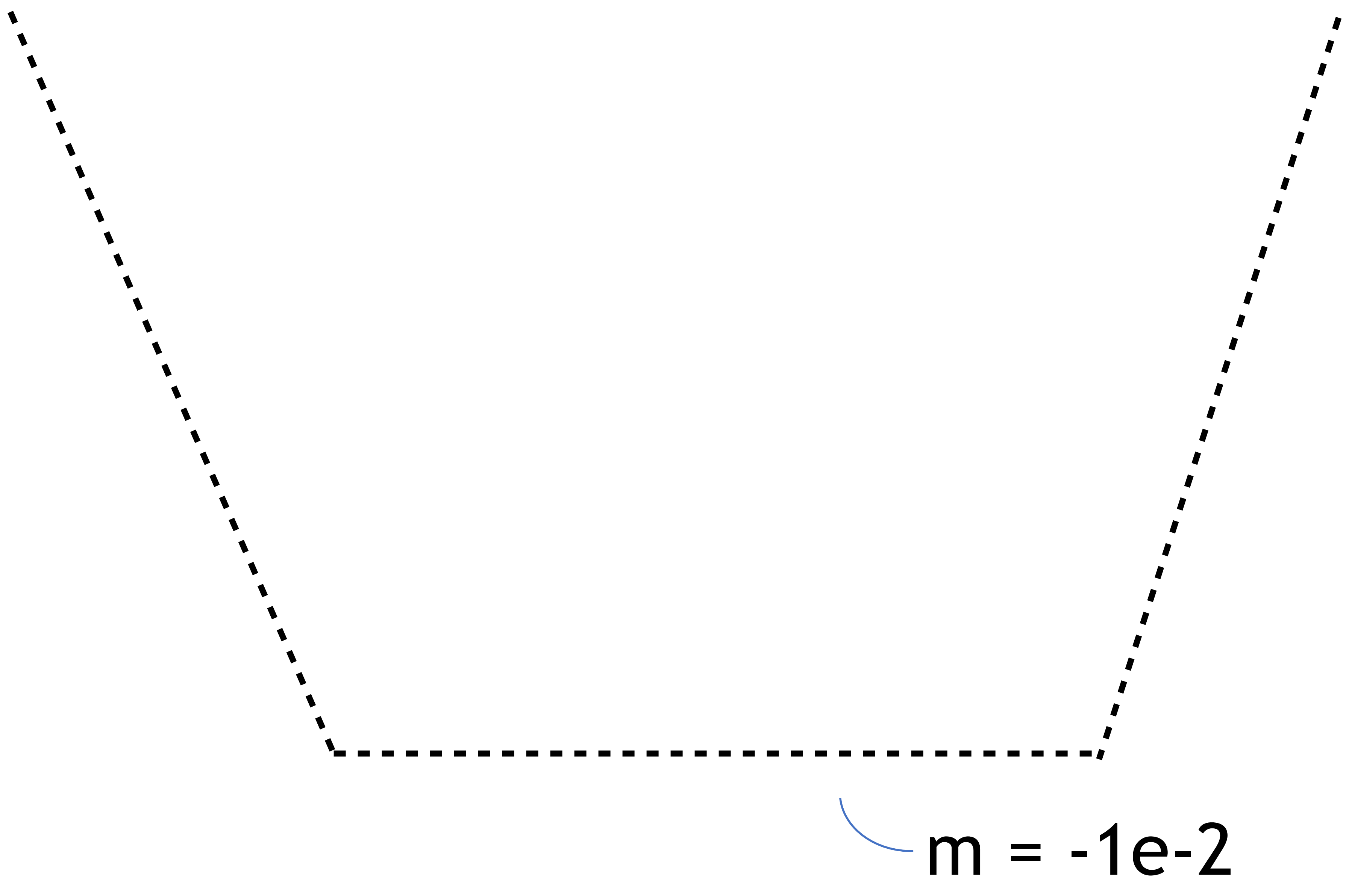
This increases the rate of convergence of the subgradient optimization algorithm.

Thank You

Backup

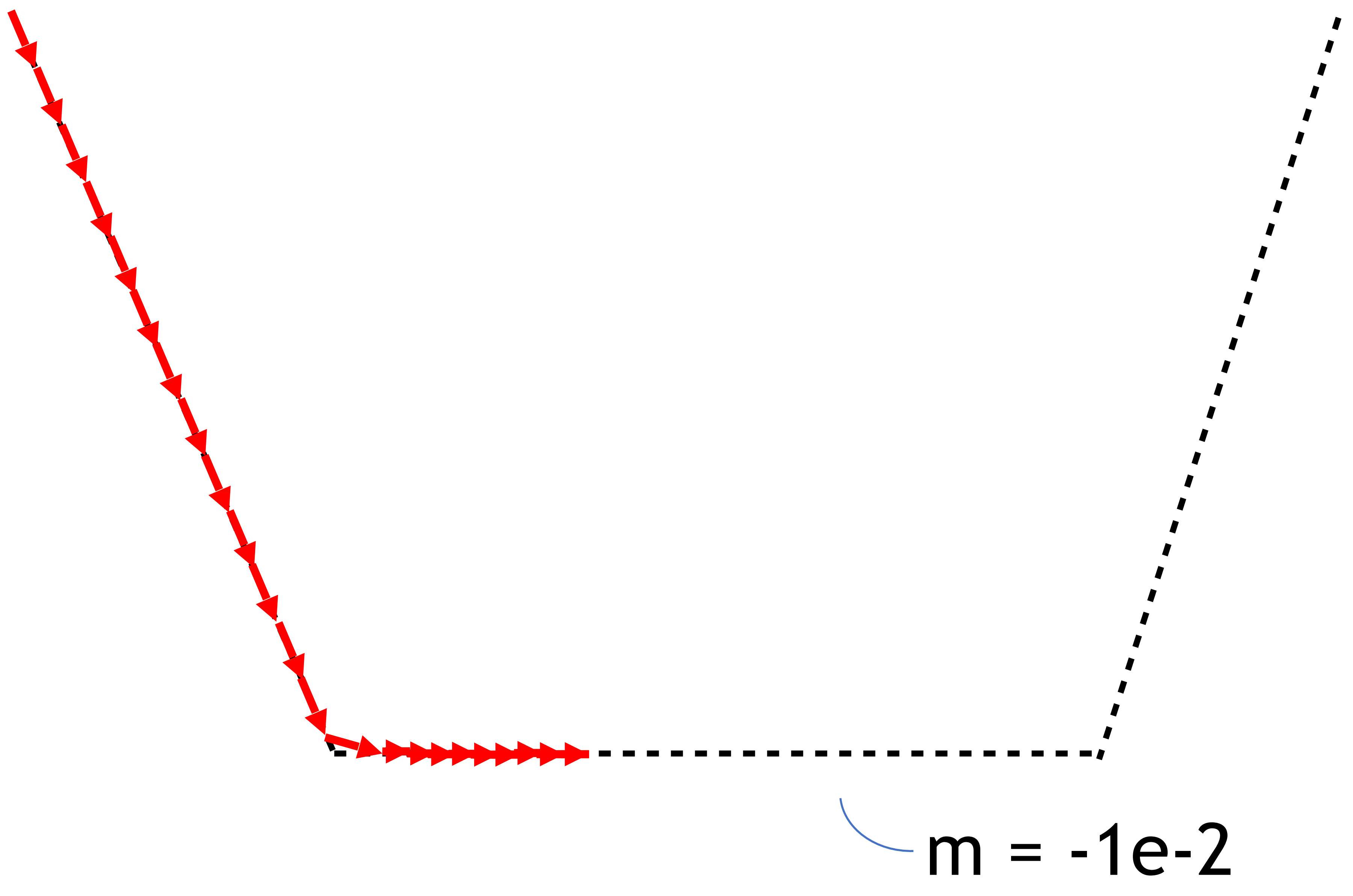
Existing Adaptive Method

Decrease the step size when the rate of change of the objective value becomes low



Existing Adaptive Method

Decrease the step size when the rate of change of the objective value becomes low



Existing Adaptive Method

Decrease the step size when the rate of change of the objective value becomes low