COMP 414/514: Optimization – Algorithms, Complexity and Approximations

Lecture 6

Overview

- In the last lecture, we:
  - Talked about a bit of second-order methods and their approximations
  - In theory, they break lower bounds of gradient descent
  - They come with a computational cost + often do not work in all cases

(open problem: generalizability of second order methods in NNs)

Overview

- In the last lecture, we:
  - Talked about a bit of second-order methods and their approximations
  - In theory, they break lower bounds of gradient descent
  - They come with a computational cost + often do not work in all cases

    (open problem: generalizability of second order methods in NNs)
- In this lecture, we will:
  - Discuss gradient descent versions that somehow **accelerate convergence**
  - Discuss techniques that do not accelerate in analytical complexity but help in iteration complexity

From previous lecture: lower bounds

- For objectives with Lipschitz continuous gradients:

$$f(x_t) - f(x^\star) \geq \frac{3L\|x_0 - x^\star\|_2^2}{32(t+1)^2}$$

(Under these assumptions, and using only gradients, we cannot achieve better than $O\left(\frac{1}{t^2}\right)$)

- In addition, for objectives that are strongly convex:

$$\|x_t - x^\star\|_2^2 \geq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t} \|x_0 - x^\star\|_2^2 \qquad \kappa := \frac{L}{\mu}$$

(The case we described has near optimal exponent, but does not involve the square root of κ)

From previous lecture: lower bounds

- For objectives with Lipschitz continuous gradients:

$$f(x_t) - f(x^\star) \geq \frac{3L\|x_0 - x^\star\|_2^2}{32(t+1)^2}$$

(Under these assumptions, and using only gradients, we cannot achieve better than $O\left(\frac{1}{t^2}\right)$)

- In addition, for objectives that are strongly convex:

$$\|x_t - x^\star\|_2^2 \geq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t} \|x_0 - x^\star\|_2^2 \qquad \kappa := \frac{L}{\mu}$$

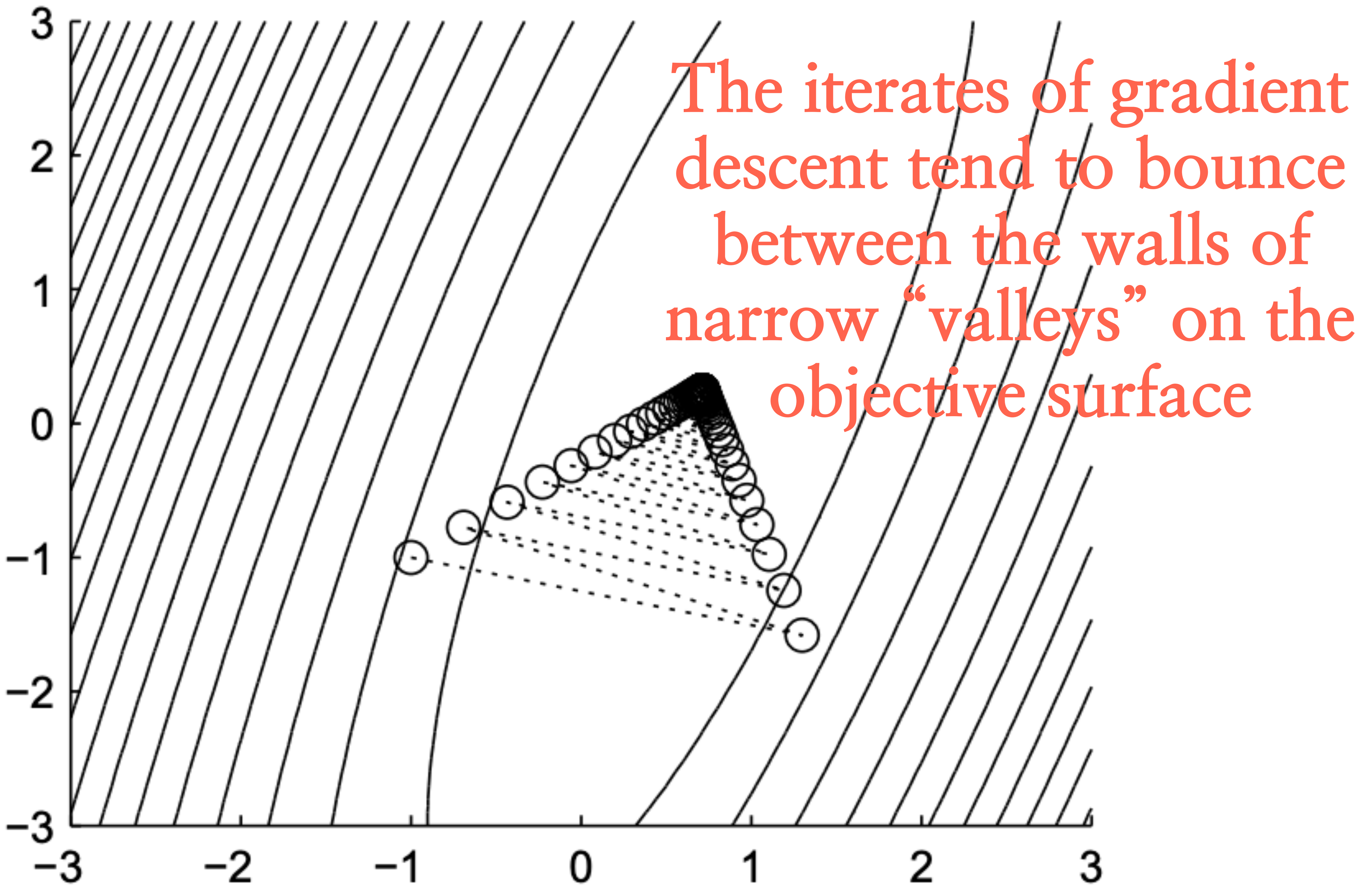
(The case we described has near optimal exponent, but does not involve the square root of κ)

Can we do better if we use more information?

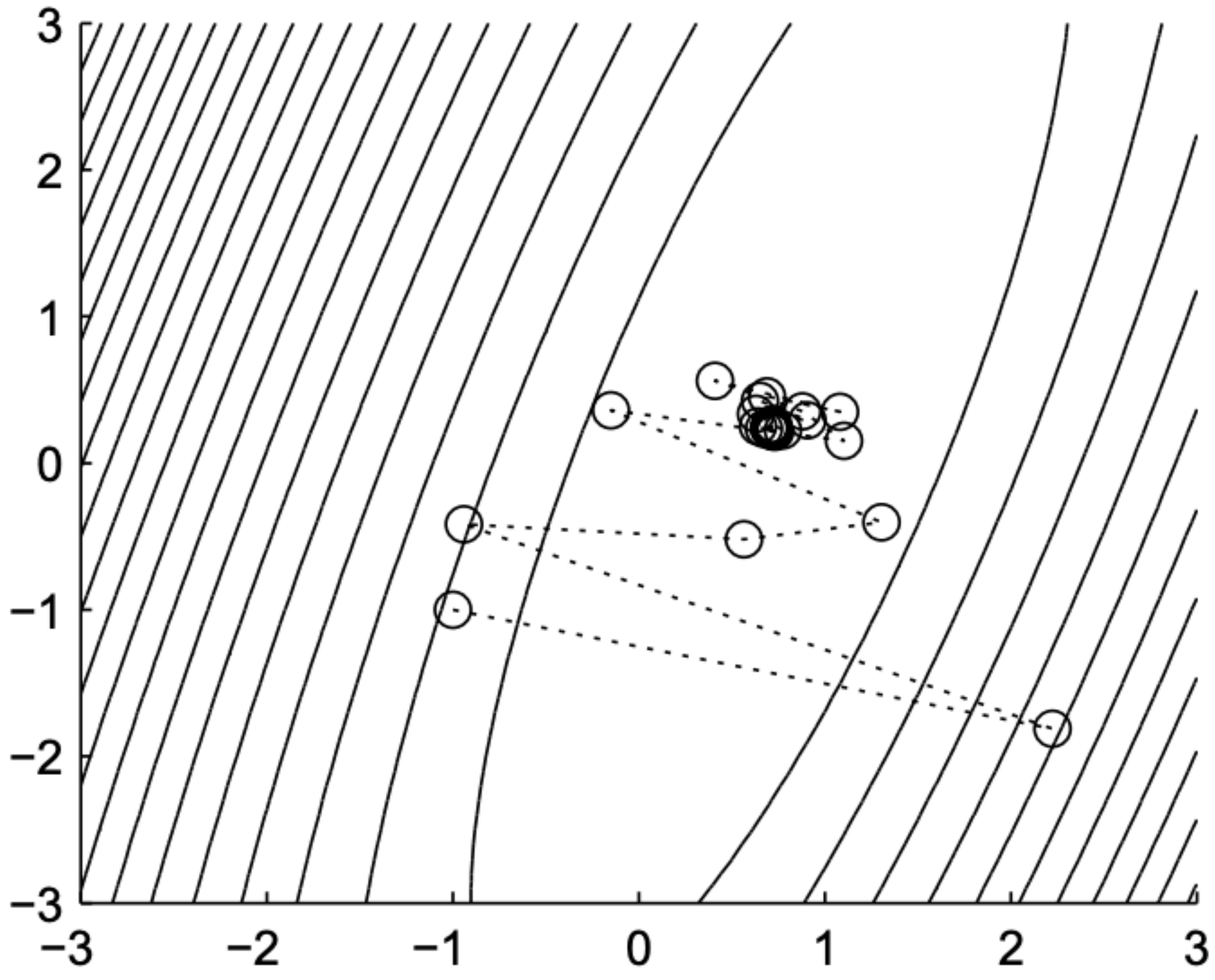
"Can we accelerate having as our basis the standard gradient descent?"

Acceleration #1: Momentum acceleration

– Heavy ball method



Gradient descent



Extrapolating previous directions

Acceleration #1: Momentum acceleration

Tribute to
Piotr Skalski

Acceleration #1: Momentum acceleration

Tribute to
Piotr Skalski

Acceleration #1: Momentum acceleration

- Heavy ball method

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

Acceleration #1: Momentum acceleration

- Heavy ball method

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

Standard gradient step

Acceleration #1: Momentum acceleration

- Heavy ball method

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

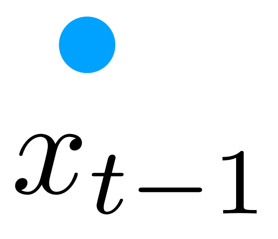
Standard gradient step — Momentum step

Acceleration #1: Momentum acceleration

- Heavy ball method

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

Standard gradient step — Momentum step

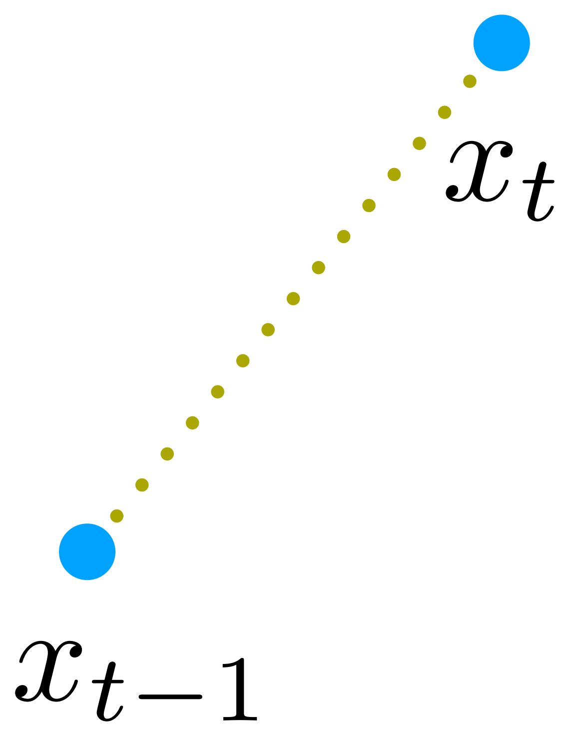


Acceleration #1: Momentum acceleration

- Heavy ball method

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

Standard gradient step   Momentum step

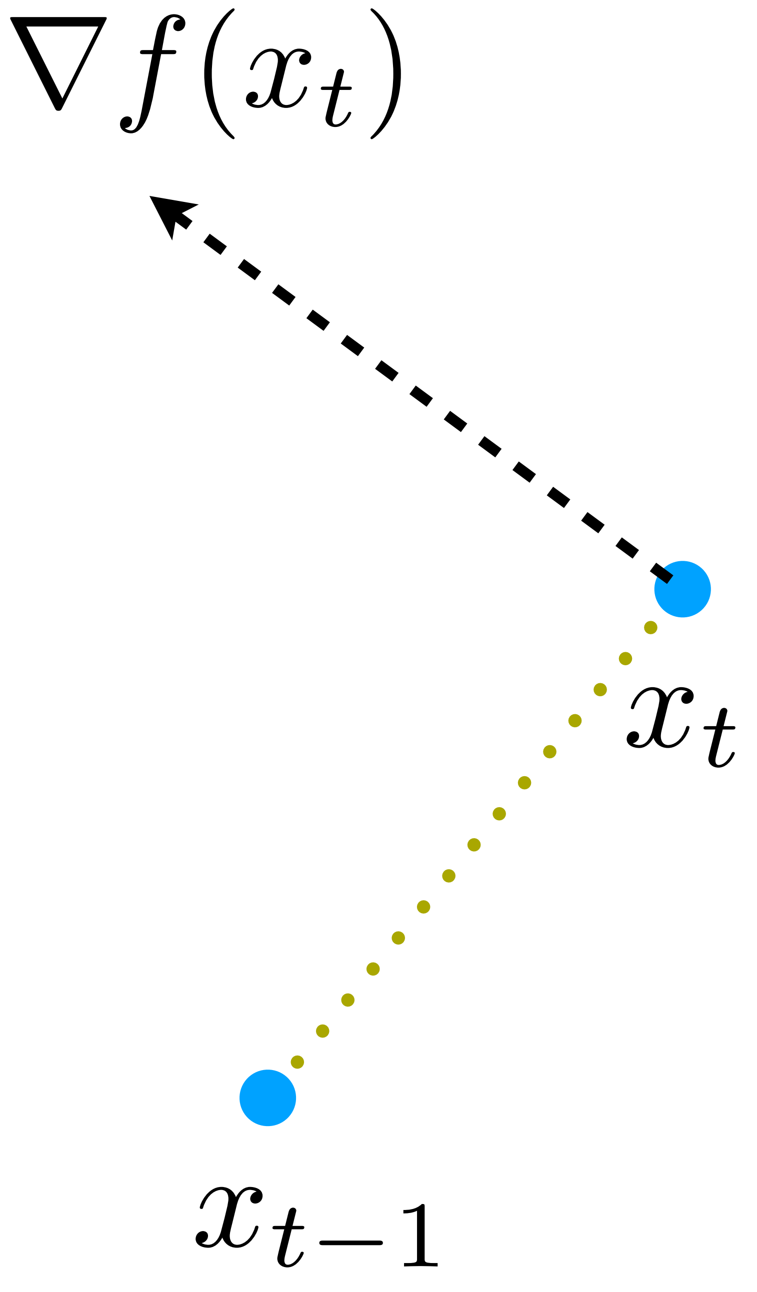


Acceleration #1: Momentum acceleration

– Heavy ball method

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

Standard gradient step — Momentum step

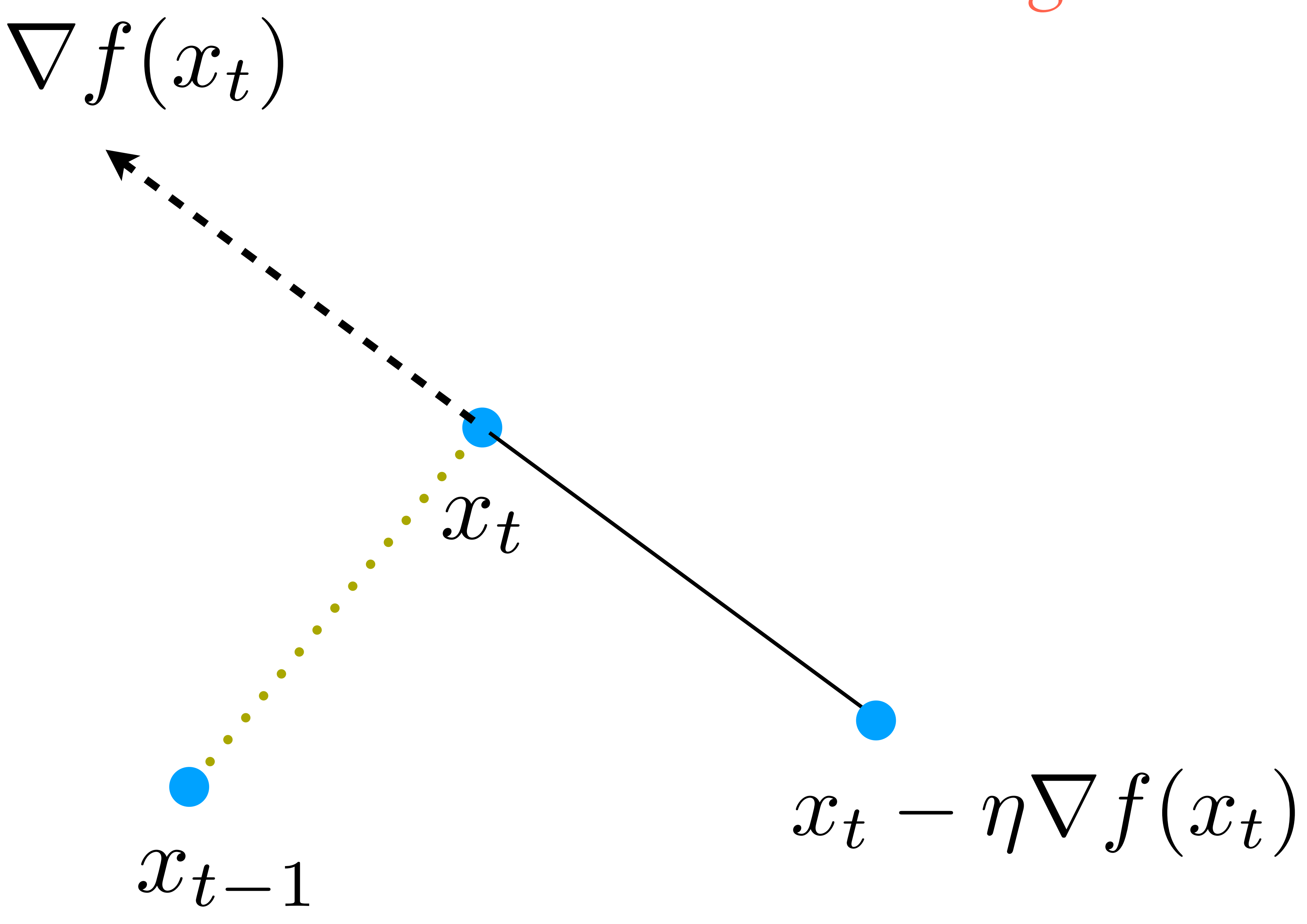


Acceleration #1: Momentum acceleration

- Heavy ball method

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

Standard gradient step   Momentum step



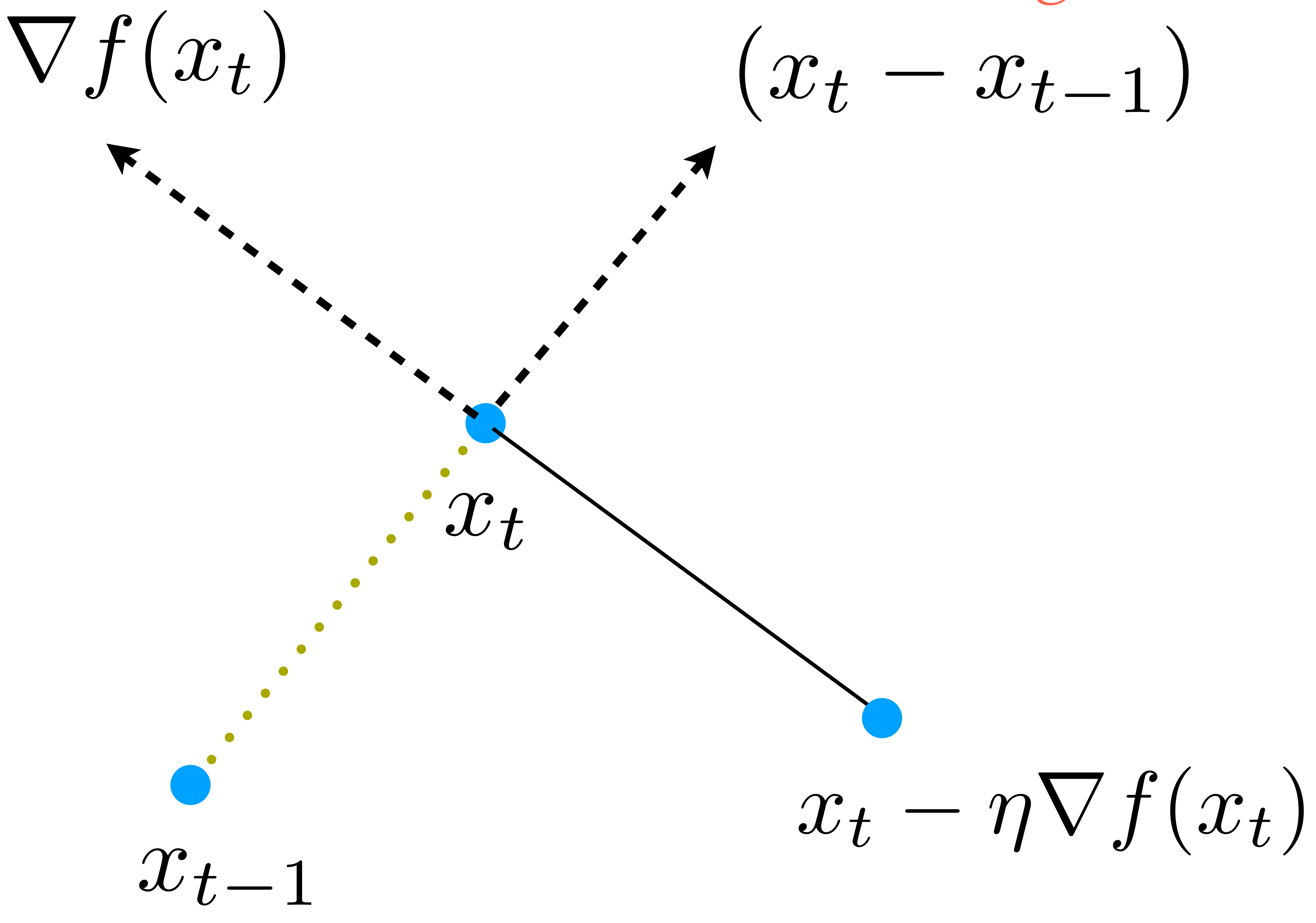
Acceleration #1: Momentum acceleration

- Heavy ball method

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

Standard gradient step

Momentum step

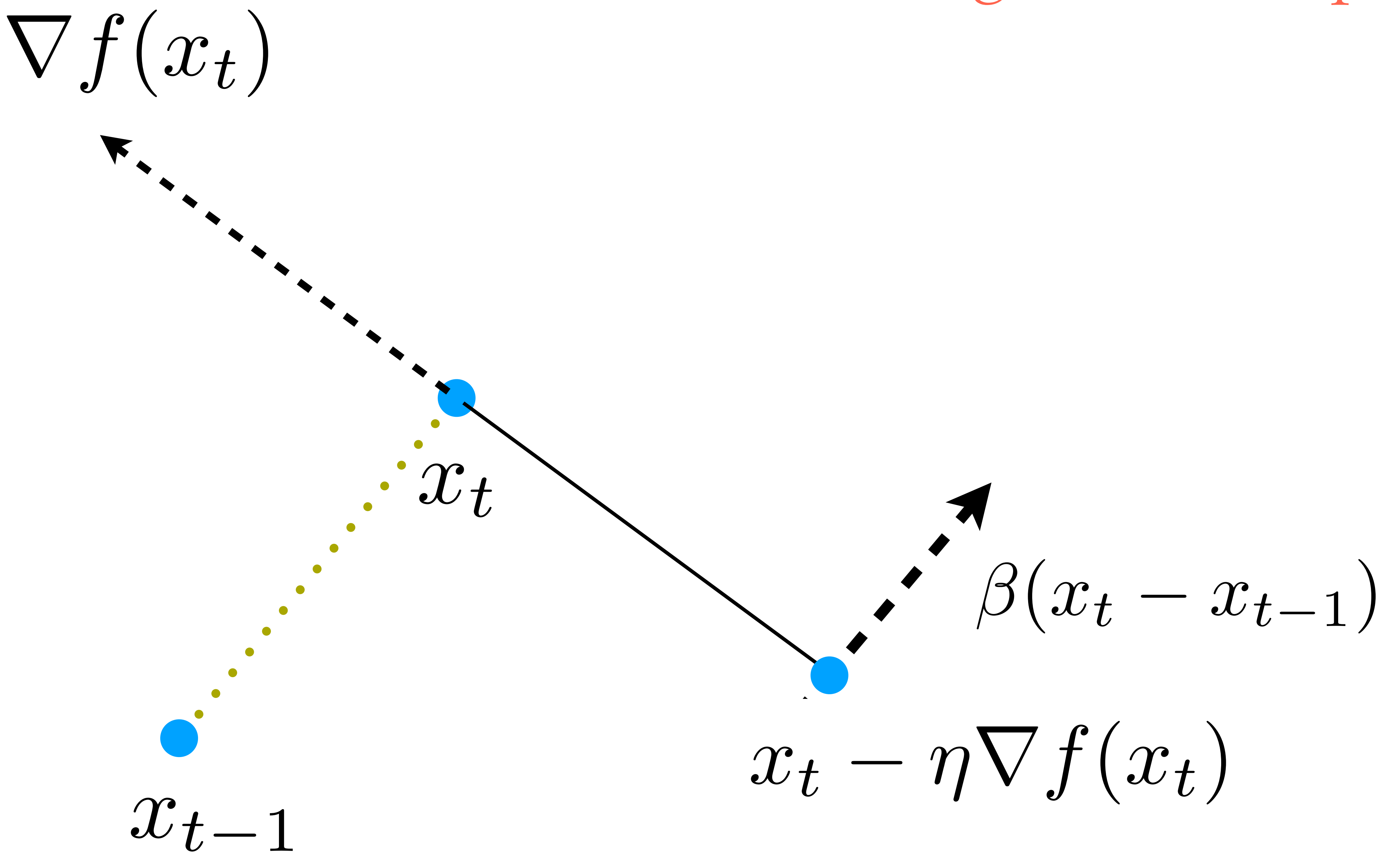


Acceleration #1: Momentum acceleration

- Heavy ball method

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

Standard gradient step   Momentum step

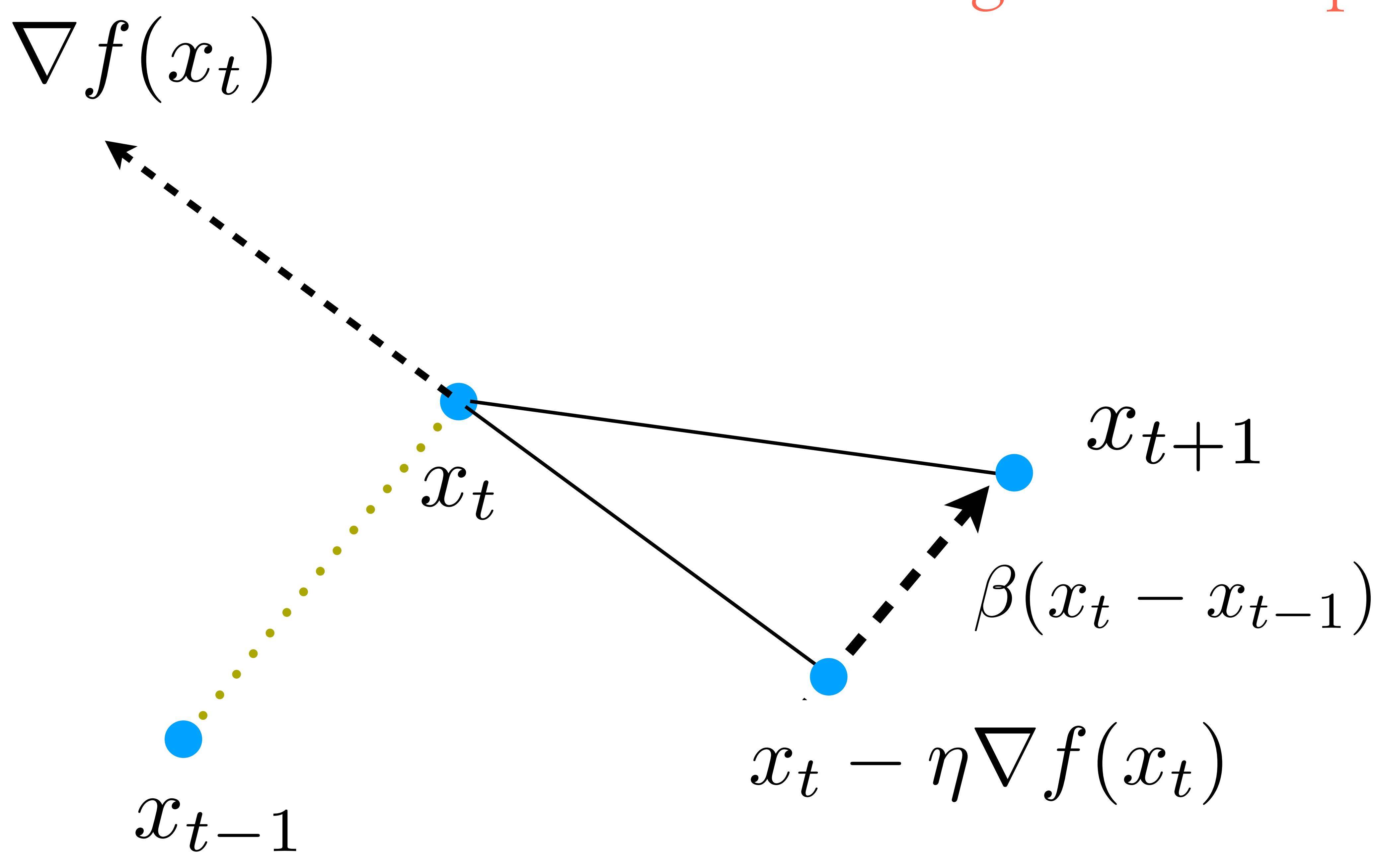


Acceleration #1: Momentum acceleration

- Heavy ball method

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

Standard gradient step — Momentum step

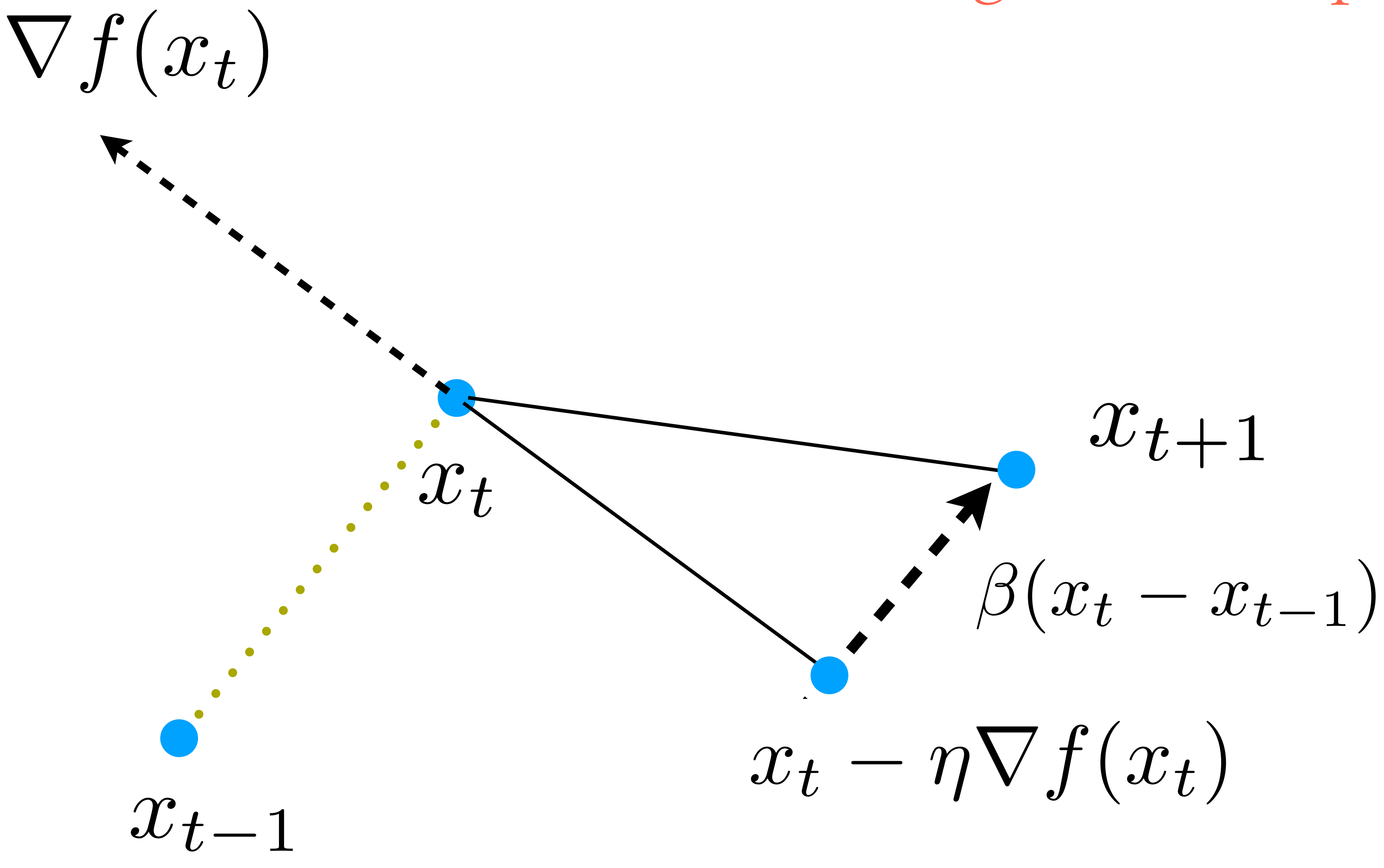


Acceleration #1: Momentum acceleration

– Heavy ball method

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

Standard gradient step   Momentum step



Any analogy in the physical world?

– If current gradient step is in same direction as previous step, then move a little further in that direction

Guarantees of Heavy Ball method

$$\min_{x \in \mathbb{R}^p} f(x)$$

"Assume the objective is has Lipschitz continuous gradients, and it is strongly convex. Then:

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

for $\eta = \frac{4}{\sqrt{L} + \sqrt{\mu}}$ *and* $\beta = \max\{|1 - \sqrt{\eta \mu}|,\ |1 - \sqrt{\eta L}|\}^2$

converges linearly according to:

$$\|x_{t+1} - x^\star\|_2 \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^t \|x_0 - x^\star\|_2$$ *"*

Guarantees of Heavy Ball method

Whiteboard

Guarantees of Heavy Ball method

- It achieves the lower bound for strongly convex cases!

“ $$\|x_t - x^\star\|_2^2 \geq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t} \|x_0 - x^\star\|_2^2 \qquad \kappa := \frac{L}{\mu}$$ “

Guarantees of Heavy Ball method

– It achieves the lower bound for strongly convex cases!

“ $$\|x_t - x^\star\|_2^2 \geq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t} \|x_0 - x^\star\|_2^2 \qquad \kappa := \frac{L}{\mu}$$ “

– In comparison with simple gradient descent:

$$O\left(\kappa \log \tfrac{1}{\varepsilon}\right) \quad \text{vs} \quad O\left(\sqrt{\kappa} \log \tfrac{1}{\varepsilon}\right)$$

Performance of Heavy Ball method

Demo

Acceleration #1: Momentum acceleration

– Nesterov's work: a collection of acceleration methods

Constant Step Scheme, I

0. Choose $x_0 \in R^n$ and $\gamma_0 > 0$. Set $v_0 = x_0$.

1. kth iteration ($k \geq 0$).

a). Compute $\alpha_k \in (0,1)$ from the equation

$$L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k \mu.$$

Set $\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k \mu$.

b). Choose $y_k = \frac{\alpha_k \gamma_k v_k + \gamma_{k+1} x_k}{\gamma_k + \alpha_k \mu}$.

Compute $f(y_k)$ and $f'(y_k)$.

c). Set $x_{k+1} = y_k - \frac{1}{L} f'(y_k)$ and

$$v_{k+1} = \frac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k v_k + \alpha_k \mu y_k - \alpha_k f'(y_k)].$$

Constant Step Scheme, II

0. Choose $x_0 \in R^n$ and $\alpha_0 \in (0,1)$.
Set $y_0 = x_0$ and $q = \frac{\mu}{L}$.

1. kth iteration ($k \geq 0$).

a). Compute $f(y_k)$ and $f'(y_k)$. Set

$$x_{k+1} = y_k - \frac{1}{L} f'(y_k).$$

b). Compute $\alpha_{k+1} \in (0,1)$ from equation

$$\alpha_{k+1}^2 = (1-\alpha_{k+1})\alpha_k^2 + q\alpha_{k+}$$

and set $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$,

$$y_{k+1} = x_{k+1} + \beta_k(x_{k+1} - x_k)$$

Constant step scheme, III

0. Choose $y_0 = x_0 \in R^n$.

1. kth iteration ($k \geq 0$).

$$x_{k+1} = y_k - \frac{1}{L} f'(y_k),$$

$$y_{k+1} = x_{k+1} + \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}(x_{k+1} - x_k).$$

Acceleration #1: Momentum acceleration

- Nesterov's work: a collection of acceleration methods

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

Acceleration #1: Momentum acceleration

- Nesterov's work: a collection of acceleration methods

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

$$\widetilde{x} = x_t - \eta \nabla f(x_t)$$
$$x_{t+1} = \widetilde{x} + \beta(x_t - x_{t-1})$$

Acceleration #1: Momentum acceleration

– Nesterov's work: a collection of acceleration methods

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta (x_t - x_{t-1})$$

$$\widetilde{x} = x_t - \eta \nabla f(x_t)$$
$$x_{t+1} = \widetilde{x} + \beta (x_t - x_{t-1})$$

Evaluate gradient at current point

Acceleration #1: Momentum acceleration

– Nesterov's work: a collection of acceleration methods

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

Evaluate gradient at current point

$$\widetilde{x} = x_t - \eta \nabla f(x_t)$$

$$x_{t+1} = \widetilde{x} + \beta(x_t - x_{t-1})$$

What if we evaluate the gradient at the point we end up?

$$\widetilde{x} = x_t - \eta \nabla f(x_t + \beta(x_t - x_{t-1}))$$

$$x_{t+1} = \widetilde{x} + \beta(x_t - x_{t-1})$$

Acceleration #1: Momentum acceleration

- Nesterov's work: a collection of acceleration methods

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1})$$

Evaluate gradient at current point

$$\widetilde{x} = x_t - \eta \nabla f(x_t)$$
$$x_{t+1} = \widetilde{x} + \beta(x_t - x_{t-1})$$

What if we evaluate the gradient at the point we end up?

Nesterov's acceleration (1/2)

$$\widetilde{x} = x_t - \eta \nabla f(x_t + \beta(x_t - x_{t-1}))$$
$$x_{t+1} = \widetilde{x} + \beta(x_t - x_{t-1})$$

Acceleration #1: Momentum acceleration

- Nesterov's work: most famous version

$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$

Acceleration #1: Momentum acceleration

– Nesterov's work: most famous version

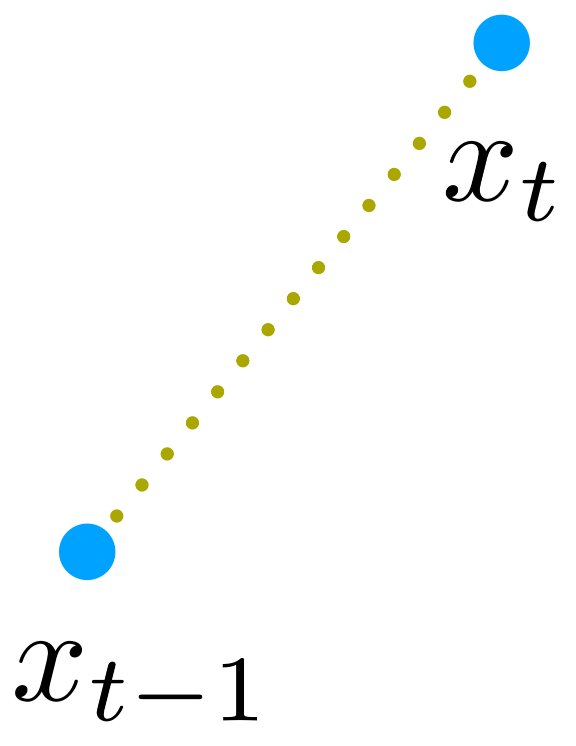
$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$

x_{t-1}

Acceleration #1: Momentum acceleration

– Nesterov's work: most famous version

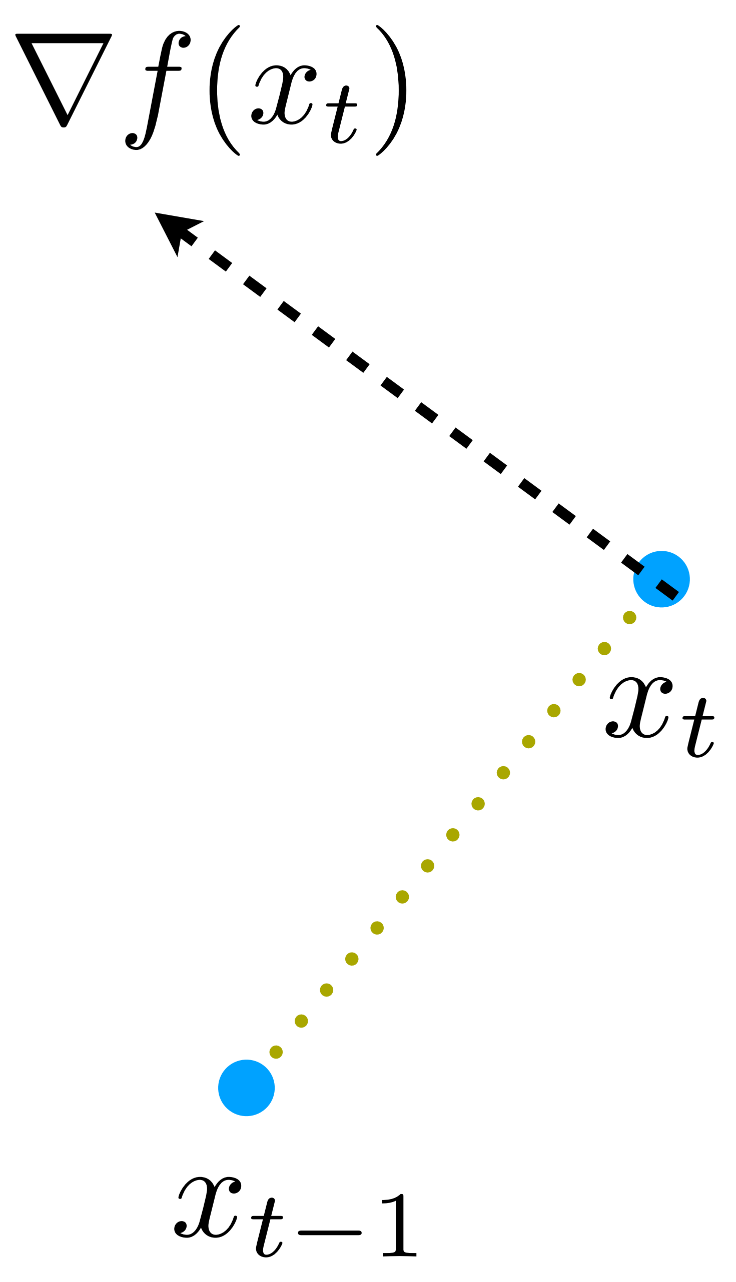
$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$



Acceleration #1: Momentum acceleration

– Nesterov's work: most famous version

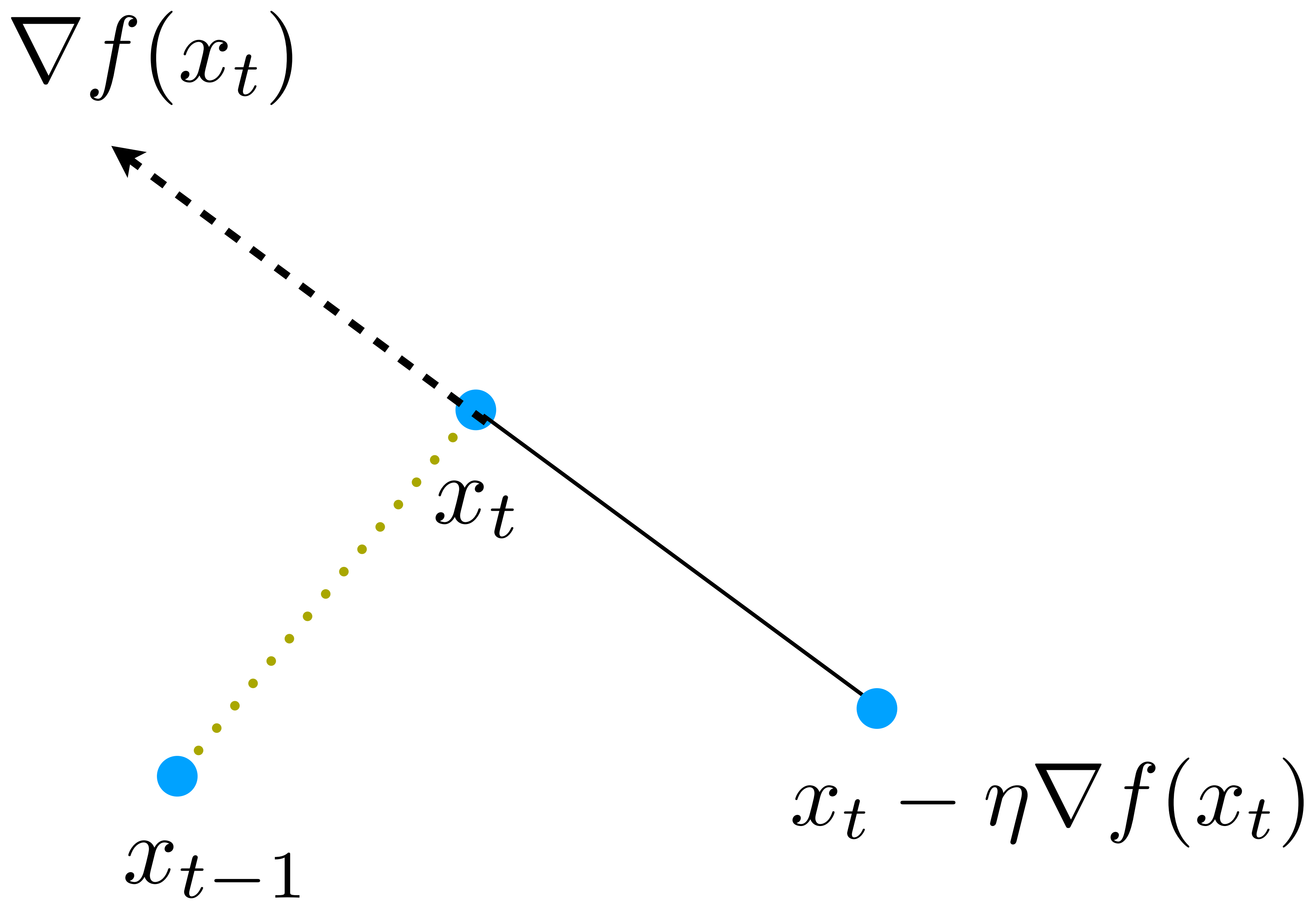
$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$



Acceleration #1: Momentum acceleration

- Nesterov's work: most famous version

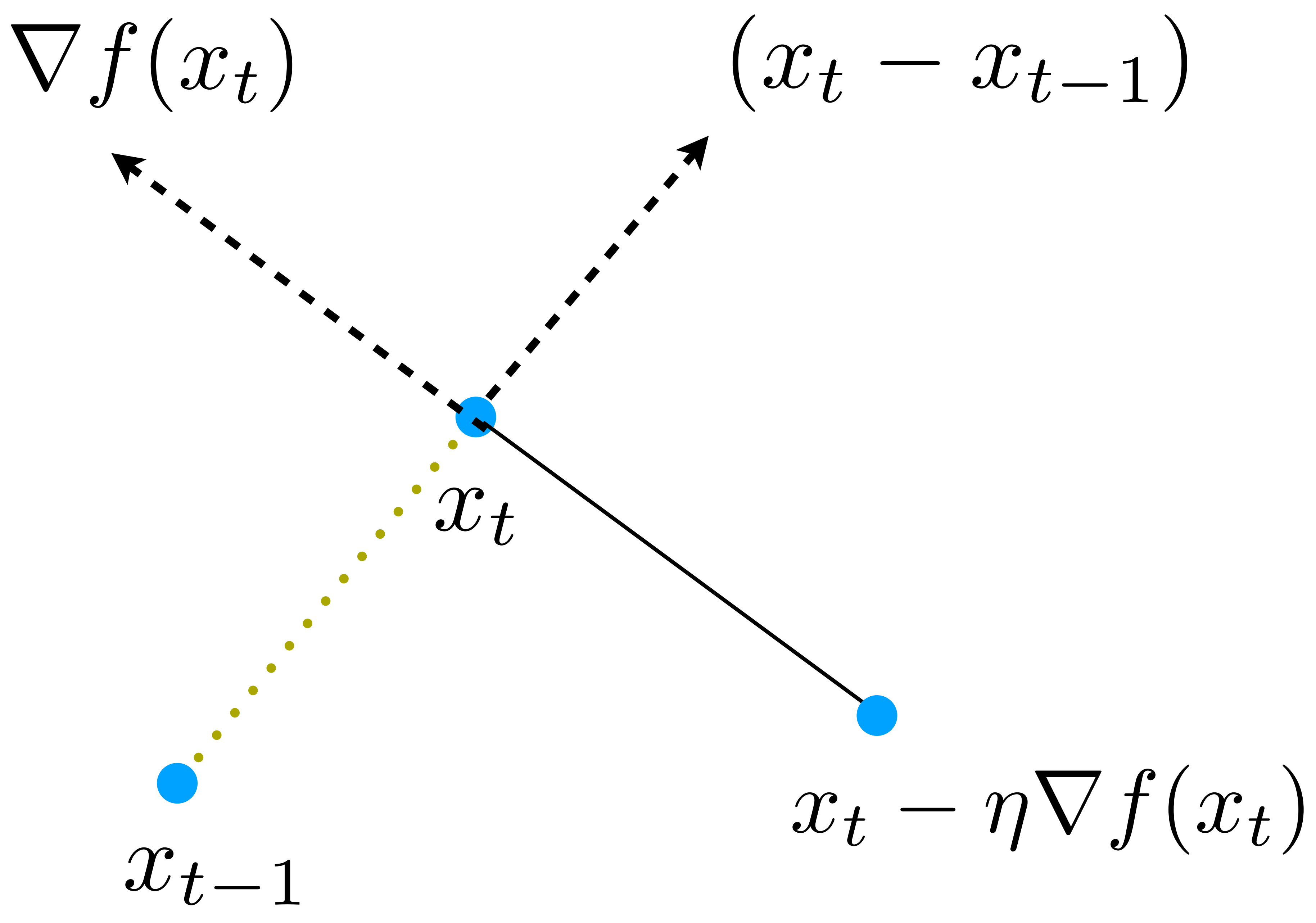
$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$



Acceleration #1: Momentum acceleration

- Nesterov's work: most famous version

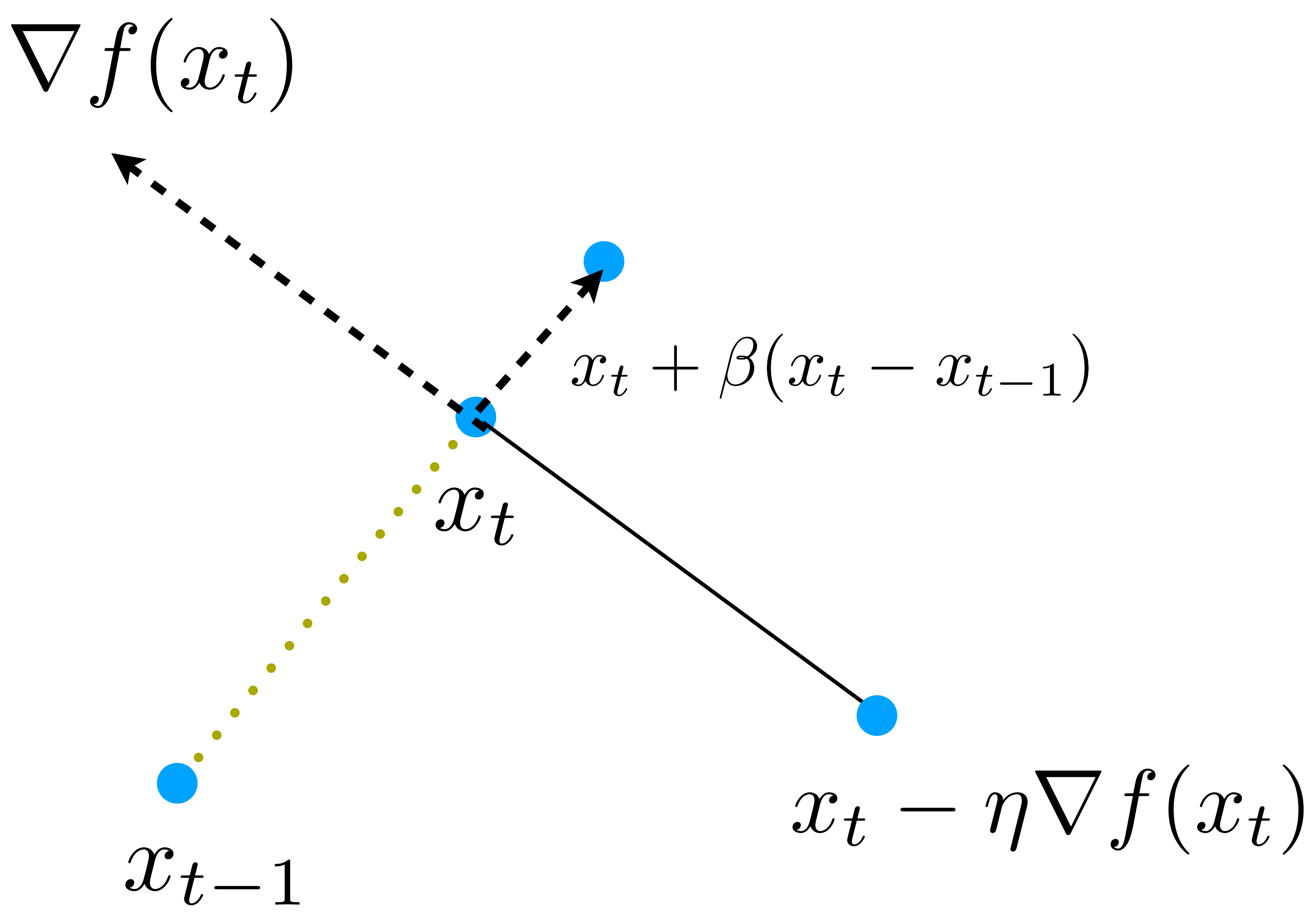
$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$



Acceleration #1: Momentum acceleration

- Nesterov's work: most famous version

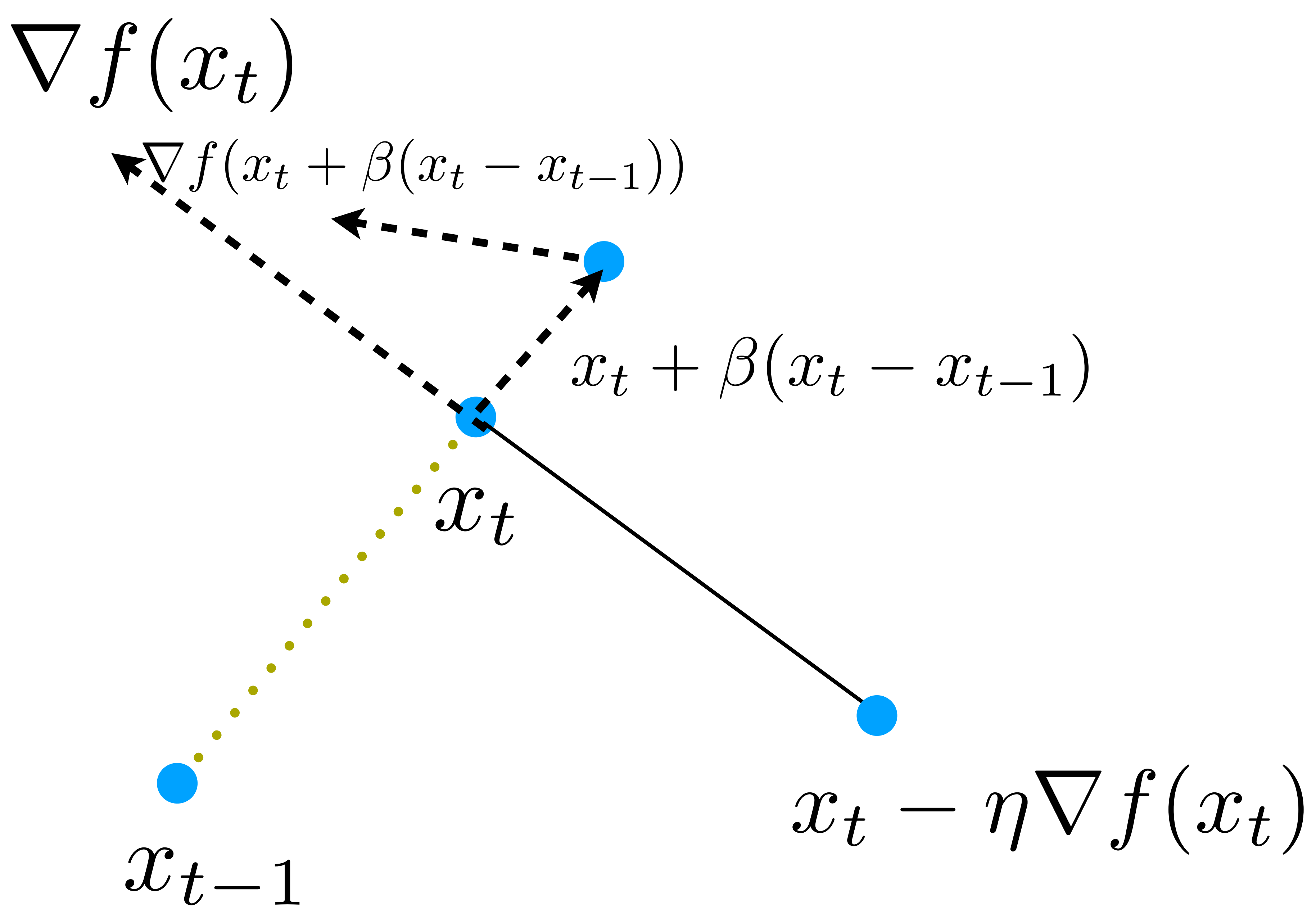
$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$



Acceleration #1: Momentum acceleration

– Nesterov's work: most famous version

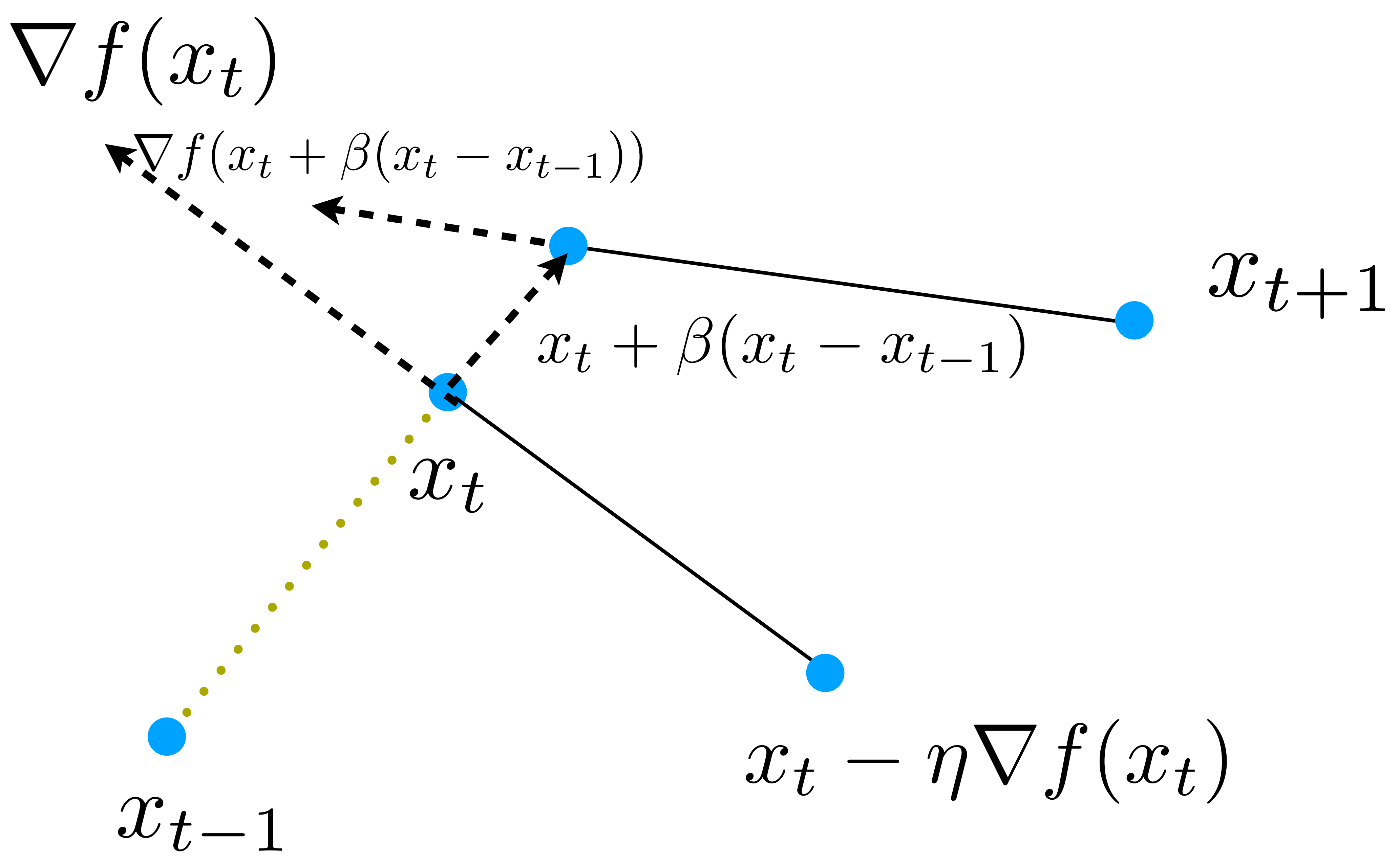
$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$



Acceleration #1: Momentum acceleration

– Nesterov's work: most famous version

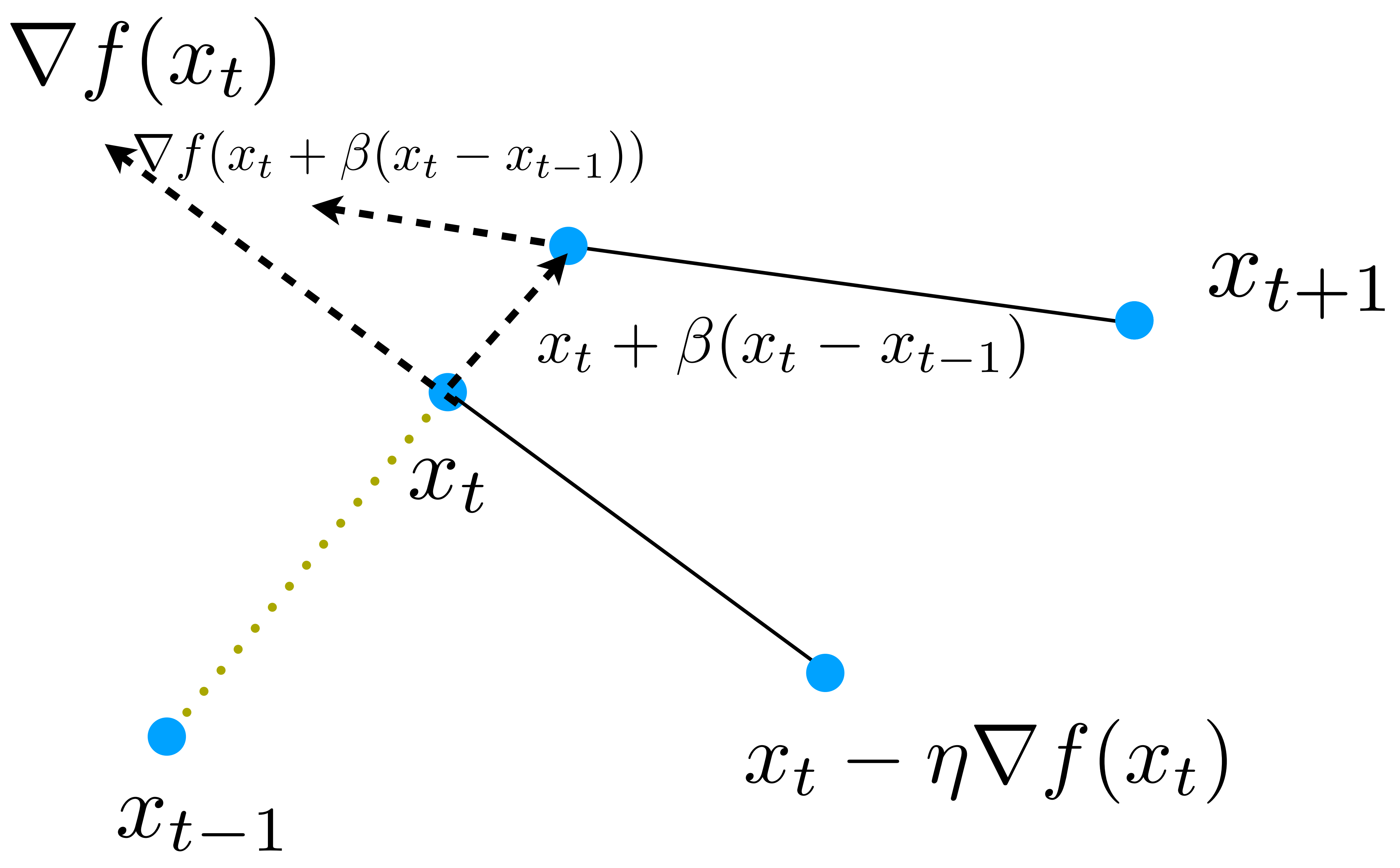
$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$



Acceleration #1: Momentum acceleration

- Nesterov's work: most famous version

$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$

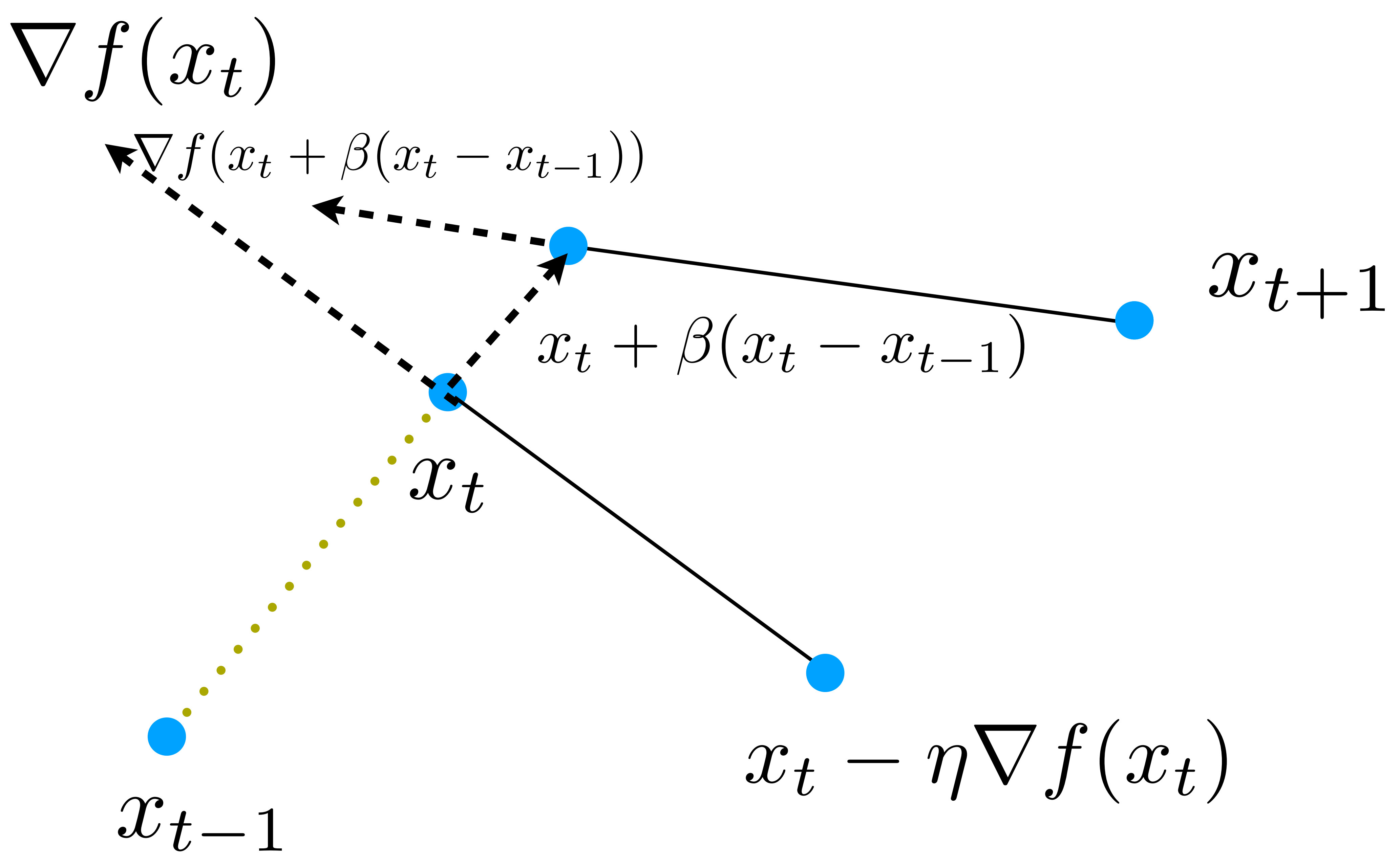


- Main difference: the point that we are calculating the gradient at.

Acceleration #1: Momentum acceleration

- Nesterov's work: most famous version

$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$



- Main difference: the point that we are calculating the gradient at.

- Heavy ball can fail converging in cases where Nesterov's scheme still succeeds

Acceleration #1: Momentum acceleration

- Nesterov's work: how do we set up the momentum parameter?

$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$

Acceleration #1: Momentum acceleration

– Nesterov's work: how do we set up the momentum parameter?

$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$

1. $\beta = \dfrac{\theta_t - 1}{\theta_{t+1}}$ where $\theta_0 = 1, \quad \theta_{t+1} = \dfrac{1 + \sqrt{1 + 4\theta_t^2}}{2}$

Acceleration #1: Momentum acceleration

– Nesterov's work: how do we set up the momentum parameter?

$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$

1. $\beta = \dfrac{\theta_t - 1}{\theta_{t+1}}$ where $\theta_0 = 1, \quad \theta_{t+1} = \dfrac{1 + \sqrt{1 + 4\theta_t^2}}{2}$

2. $\beta = \dfrac{t}{t+3}$

Acceleration #1: Momentum acceleration

– Nesterov's work: how do we set up the momentum parameter?

$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$

1. $\beta = \dfrac{\theta_t - 1}{\theta_{t+1}}$ where $\theta_0 = 1, \quad \theta_{t+1} = \dfrac{1 + \sqrt{1 + 4\theta_t^2}}{2}$

2. $\beta = \dfrac{t}{t+3}$

3. $\beta = 0.9$

Acceleration #1: Momentum acceleration

– Nesterov's work: how do we set up the momentum parameter?

$$x_{t+1} = y_t - \eta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \beta(x_{t+1} - x_t)$$

1. $\beta = \frac{\theta_t - 1}{\theta_{t+1}}$ where $\theta_0 = 1, \quad \theta_{t+1} = \frac{1 + \sqrt{1 + 4\theta_t^2}}{2}$

2. $\beta = \frac{t}{t+3}$

3. $\beta = 0.9$

One of the mysteries of optimization..

Performance of Nesterov’s acceleration

Demo

Guarantees of Nesterov's acceleration

(No theory but willing to provide links for whoever is interested)

- Gradient descent in the absence of strong convexity

$$f(x_t) - f(x^\star) \leq \frac{2L\|x_0 - x^\star\|_2^2}{t+4}$$

Guarantees of Nesterov's acceleration

(No theory but willing to provide links for whoever is interested)

- Gradient descent in the absence of strong convexity

$$f(x_t) - f(x^\star) \leq \frac{2L\|x_0 - x^\star\|_2^2}{t+4}$$

- Nesterov's acceleration (with momentum similarly set up as in previous slide)

$$f(x_t) - f(x^\star) \leq \frac{4L\|x_0 - x^\star\|_2^2}{(t+2)^2}$$

Guarantees of Nesterov's acceleration

(No theory but willing to provide links for whoever is interested)

- Gradient descent in the absence of strong convexity

$$f(x_t) - f(x^\star) \leq \frac{2L\|x_0 - x^\star\|_2^2}{t+4}$$

- Nesterov's acceleration (with momentum similarly set up as in previous slide)

$$f(x_t) - f(x^\star) \leq \frac{4L\|x_0 - x^\star\|_2^2}{(t+2)^2}$$

- Reminder of lower bounds for Lipschitz continuous gradients:

$$f(x_t) - f(x^\star) \geq \frac{3L\|x_0 - x^\star\|_2^2}{32(t+1)^2}$$

Guarantees of Nesterov's acceleration

(No theory but willing to provide links for whoever is interested)

– Gradient descent in the absence of strong convexity

$$f(x_t) - f(x^\star) \leq \frac{2L\|x_0 - x^\star\|_2^2}{t+4}$$

– Nesterov's acceleration (with momentum similarly set up as in previous slide)

$$f(x_t) - f(x^\star) \leq \frac{4L\|x_0 - x^\star\|_2^2}{(t+2)^2}$$

Optimal!

– Reminder of lower bounds for Lipschitz continuous gradients:

$$f(x_t) - f(x^\star) \geq \frac{3L\|x_0 - x^\star\|_2^2}{32(t+1)^2}$$

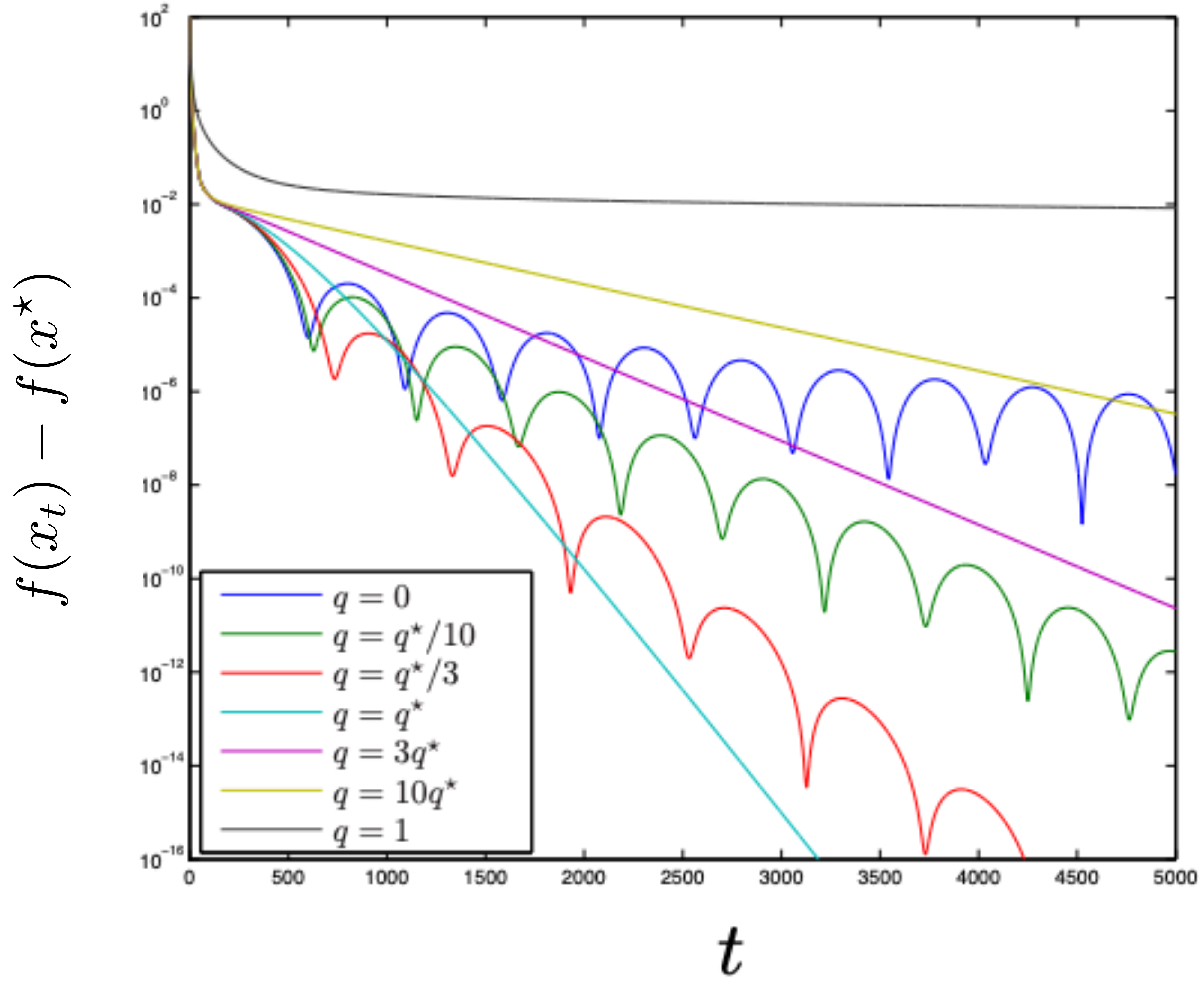
Notes on Nesterov's acceleration

- The original paper of 1983 does not converge linearly for strongly convex functions, but there is a fix to this

Notes on Nesterov's acceleration

- The original paper of 1983 does not converge linearly for strongly convex functions, but there is a fix to this

- It is a common observation to see ripples



Notes on Nesterov's acceleration

- The original paper of 1983 does not converge linearly for strongly convex functions, but there is a fix to this
- It is a common observation to see ripples
- There are heuristics for resetting the momentum term to zero that improves the convergence rate.
- Often used even in cases where it is not guaranteed to work: deep learning