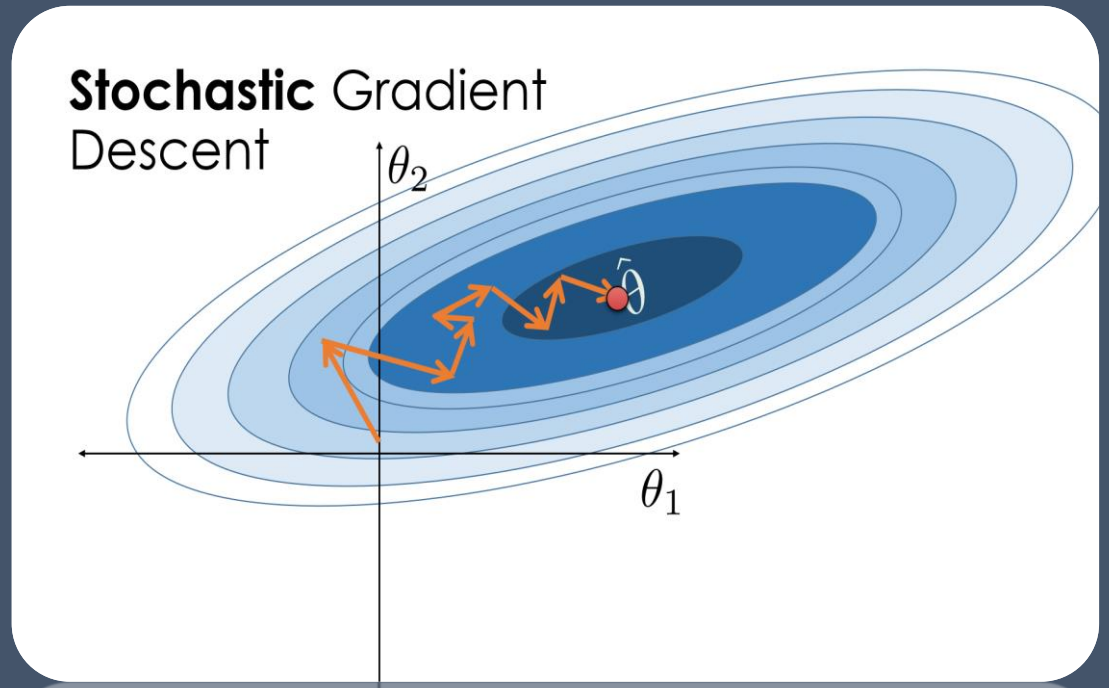
دوره‌ی آموزشی «علم داده»
Data Science Course



جلسه سی‌ام (بخش دوم)

گرادیان کاهشی تصادفی و تکانه
Stochastic Gradient Descent & Momentum

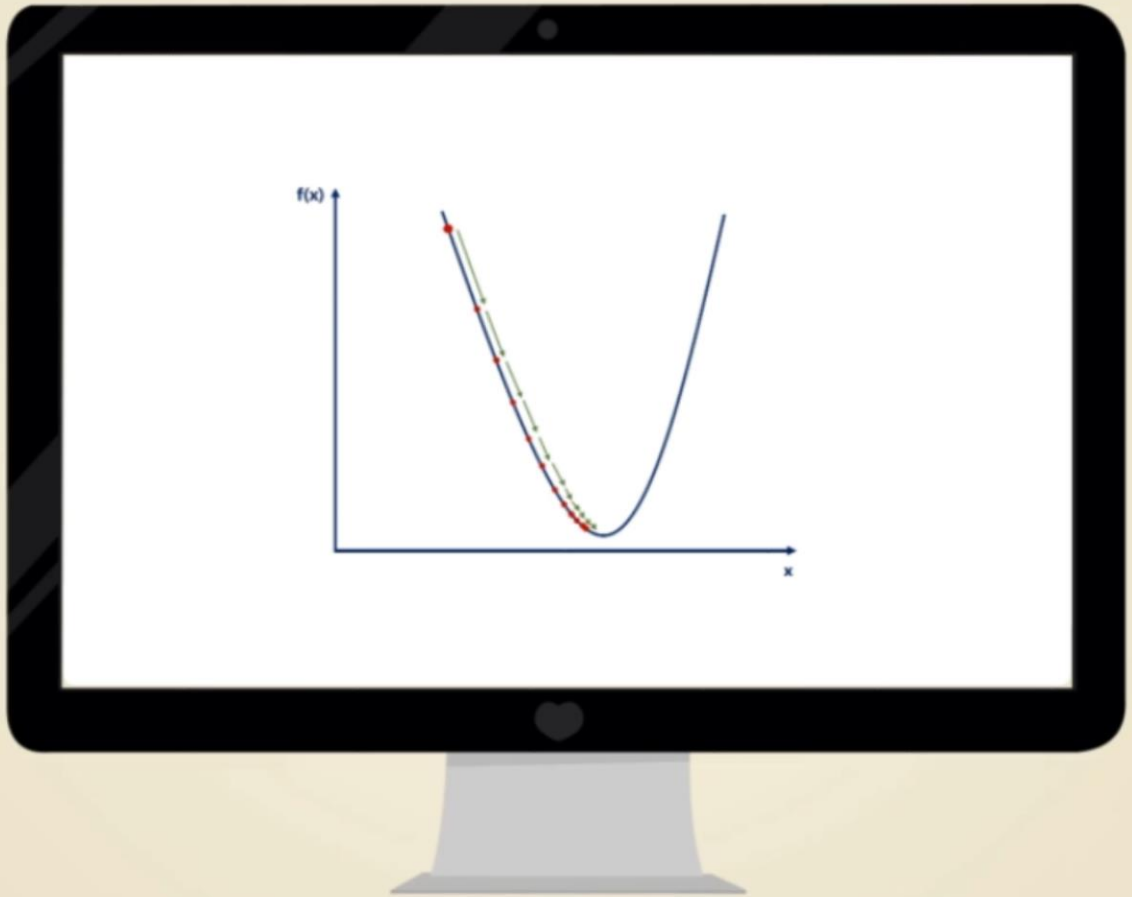
مدرس: محمد فزونی
عضو هیئت علمی دانشگاه گنبدکاووس

OPTIMIZATION

الگوریتم‌هایی که بکار می‌گیریم تا پارامترهای مدلمون رو در جهت بهبود عملکرد، تغییر بدیم

Algorithms we will use to vary our model's parameters

OPTIMIZATION

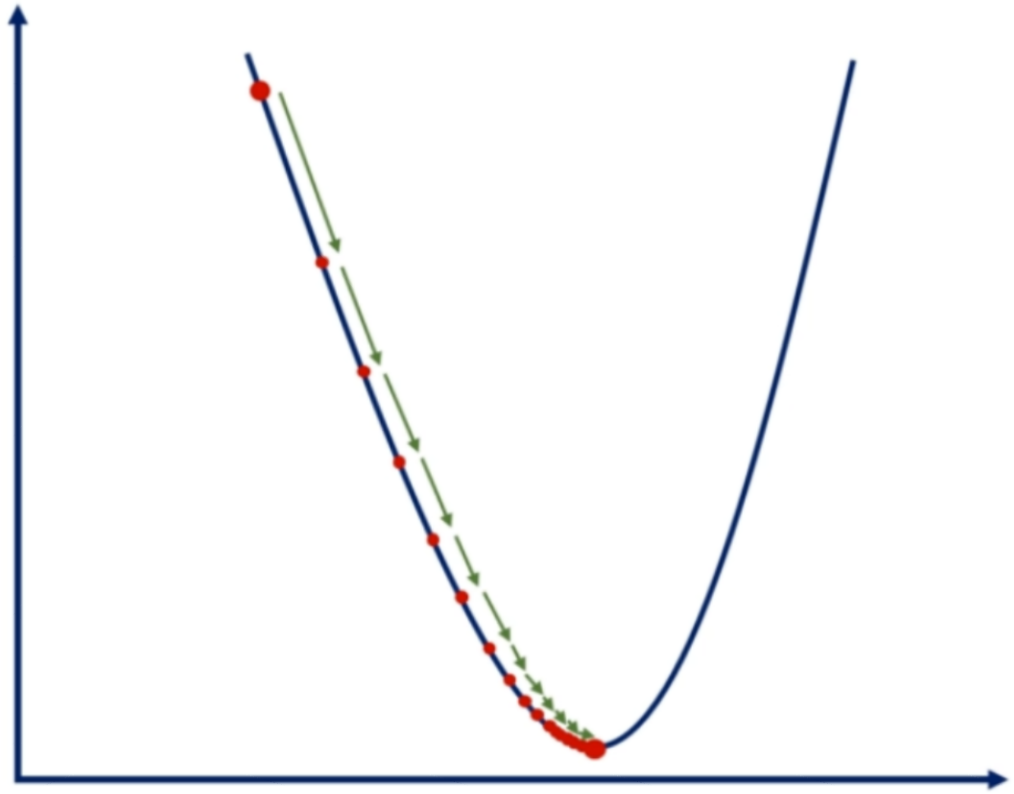


OPTIMIZATION

IN

گرادیان
کاهشی،
روی کاغذ،
نامبر وان
هست، ولی
در هنگام
بکارگیری
کمی دست و
پا شکسته
عمل می‌کنه

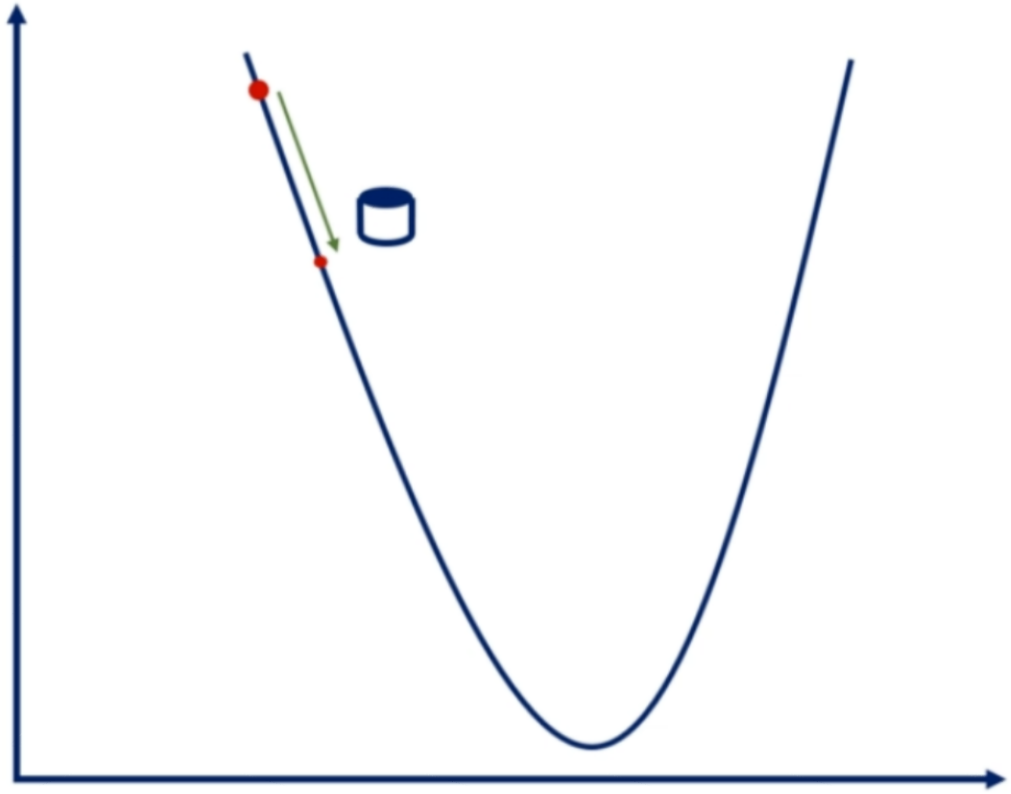
Gradient descent



x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	x_{n6}	x_{n7}	x_{n8}

n x k

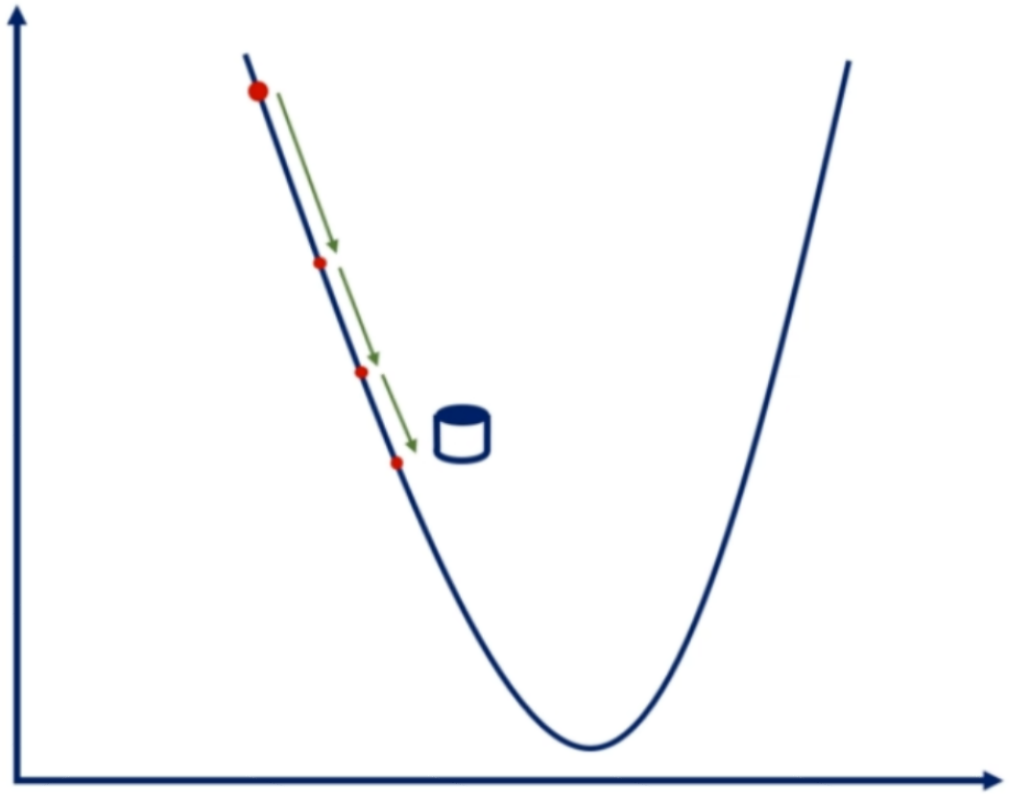
Gradient descent



x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	x_{n6}	x_{n7}	x_{n8}

n x k

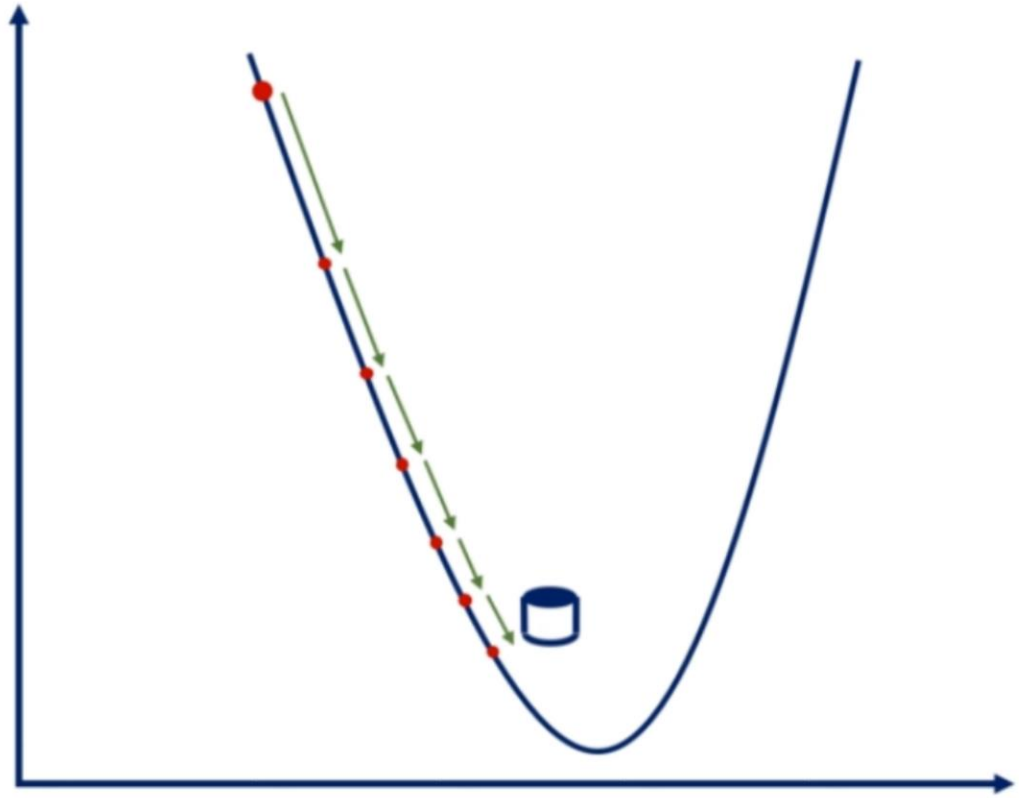
Gradient descent



x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	x_{n6}	x_{n7}	x_{n8}

n x k

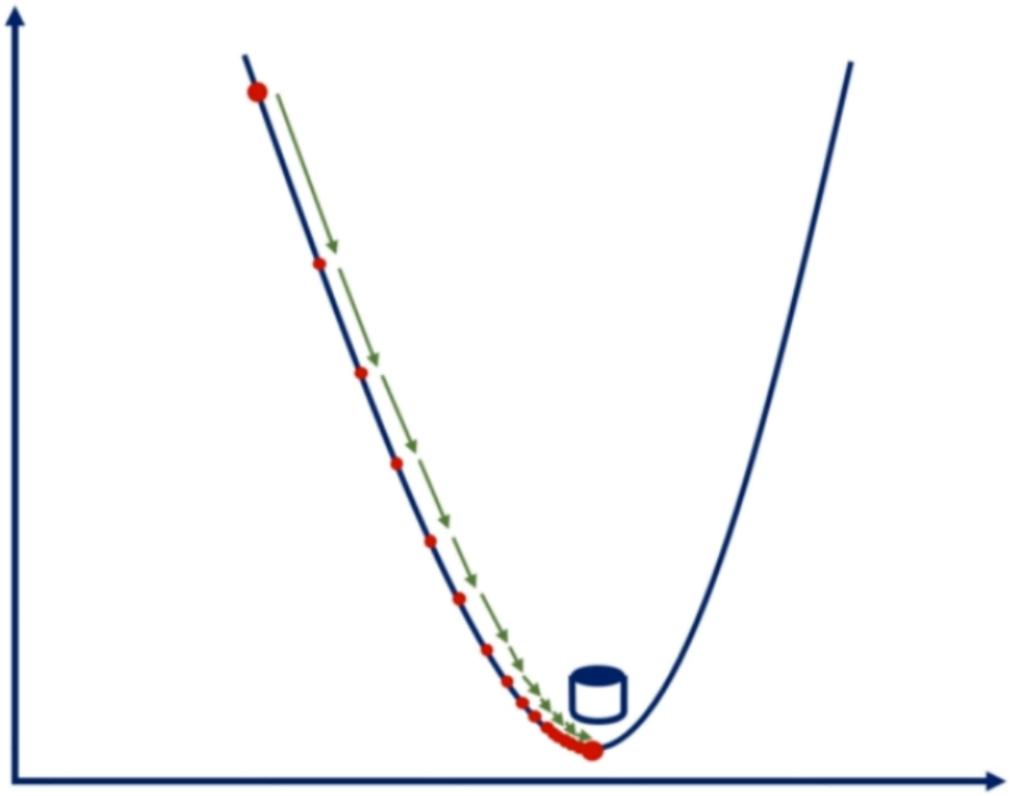
Gradient descent



x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	x_{n6}	x_{n7}	x_{n8}

n x k

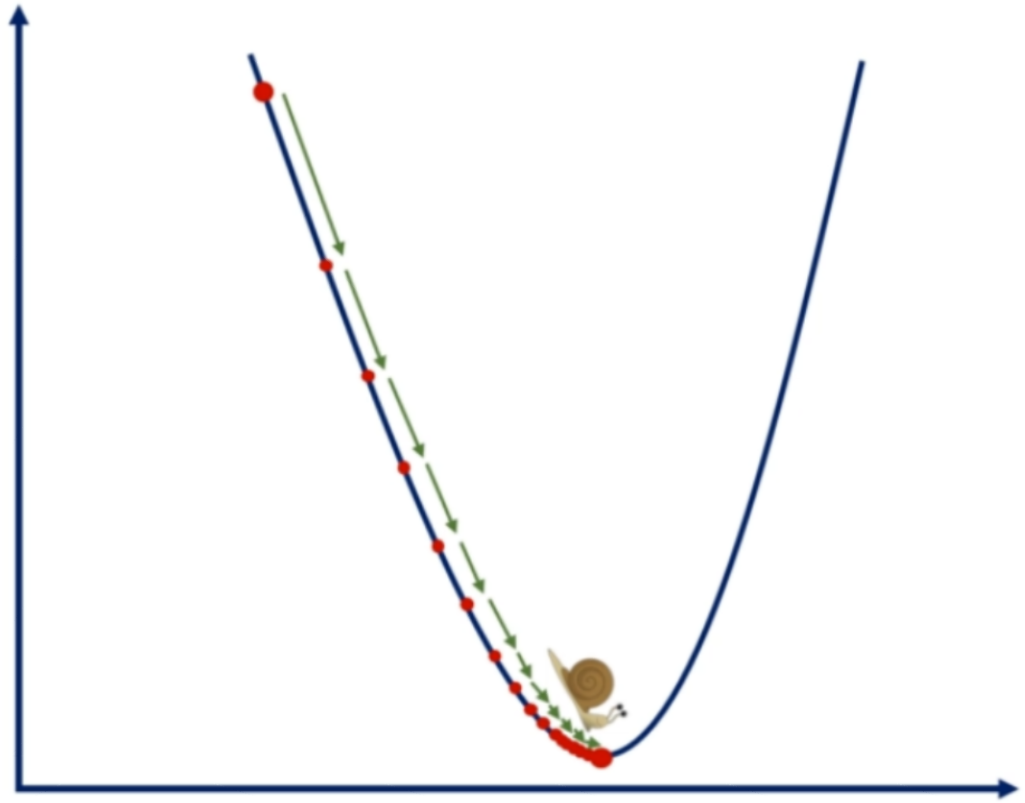
Gradient descent



x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	x_{n6}	x_{n7}	x_{n8}

n x k

Gradient descent

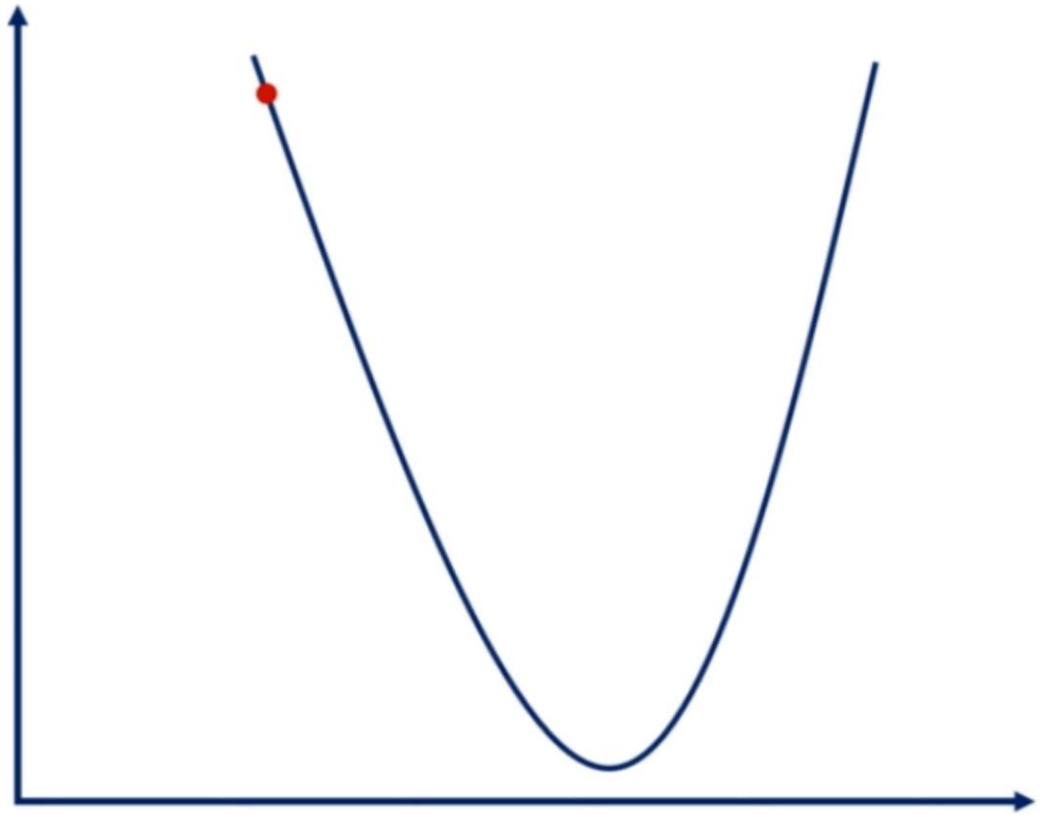


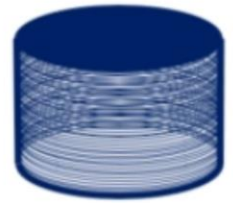
snail pace

x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	x_{n6}	x_{n7}	x_{n8}

n x k

Stochastic gradient descent





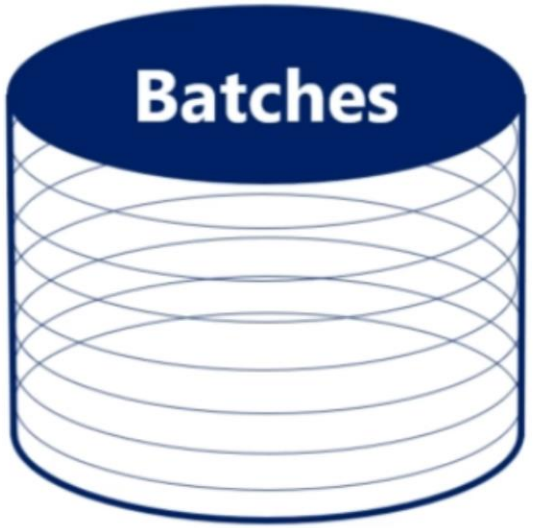
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	x_{n6}	x_{n7}	x_{n8}

n x k

It works in the same way, but updates the weights many times inside a single epoch

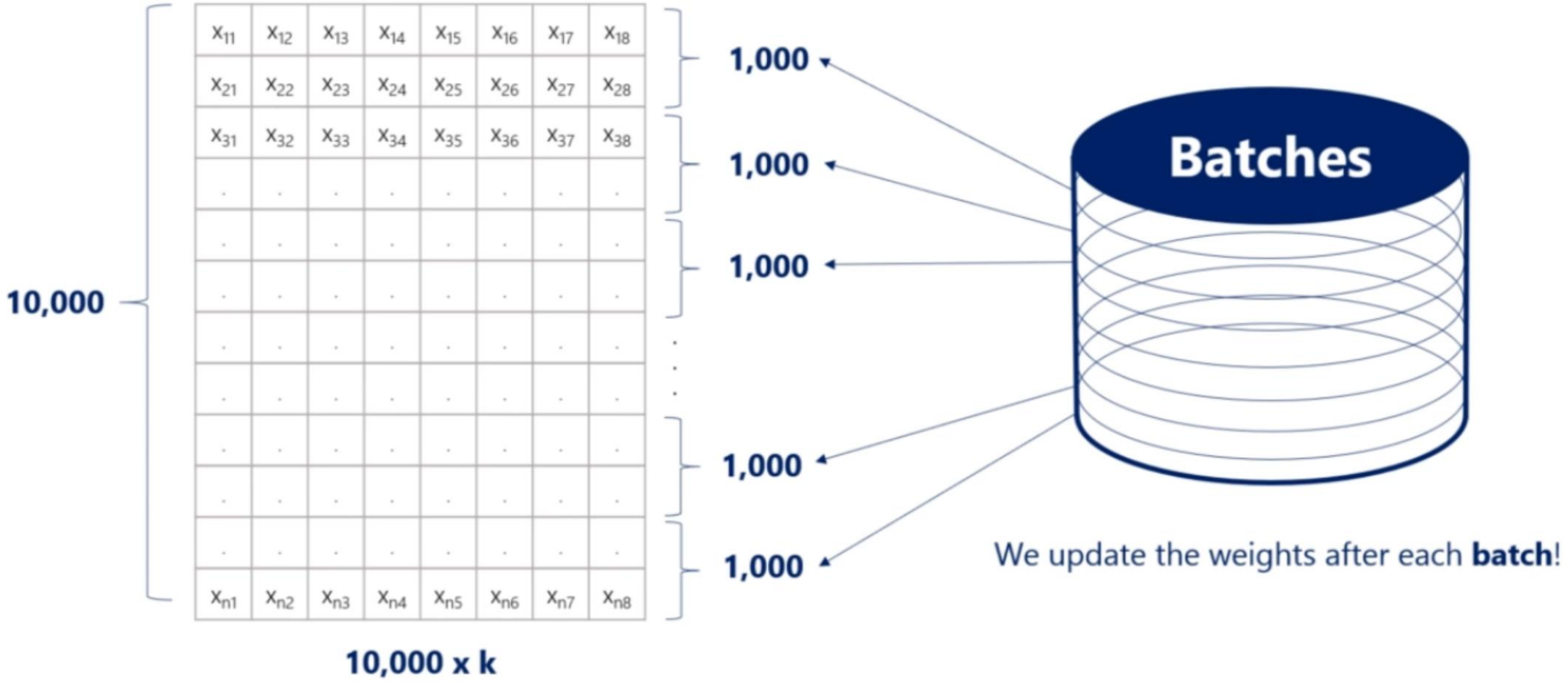
Batching

x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	x_{n6}	x_{n7}	x_{n8}



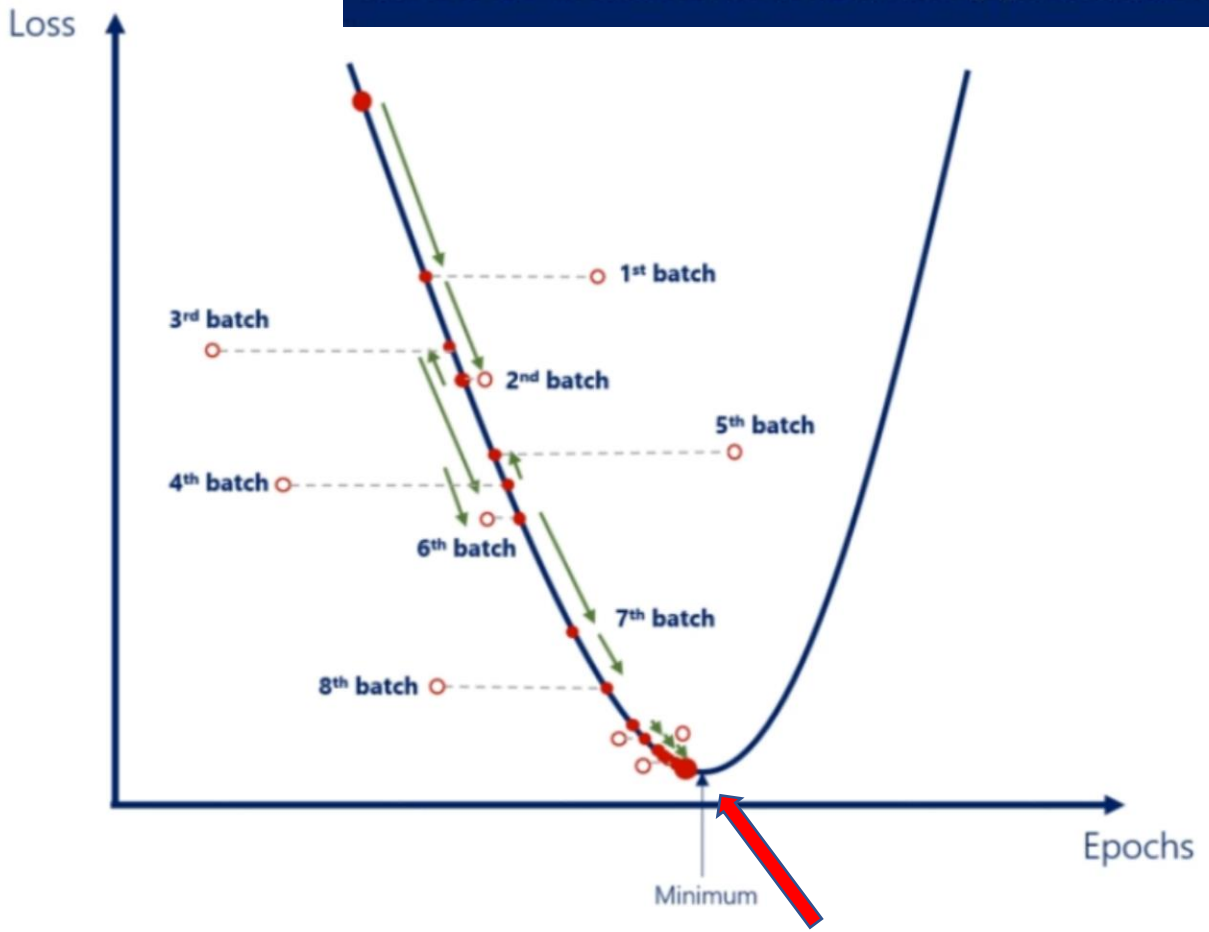
Batching: the process of splitting the dataset in **n** batches (mini-batches)

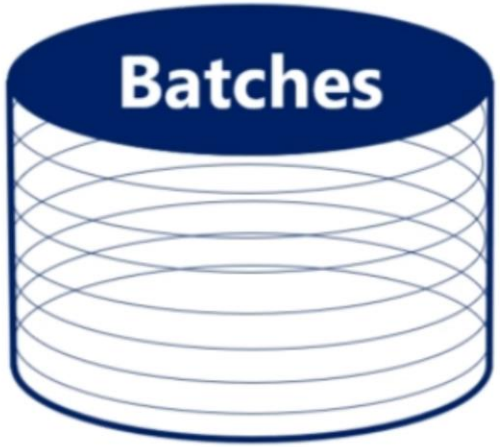
Batching


10,000
10,000 x k
1,000
1,000
1,000
1,000
1,000
Batches
We update the weights after each **batch**!

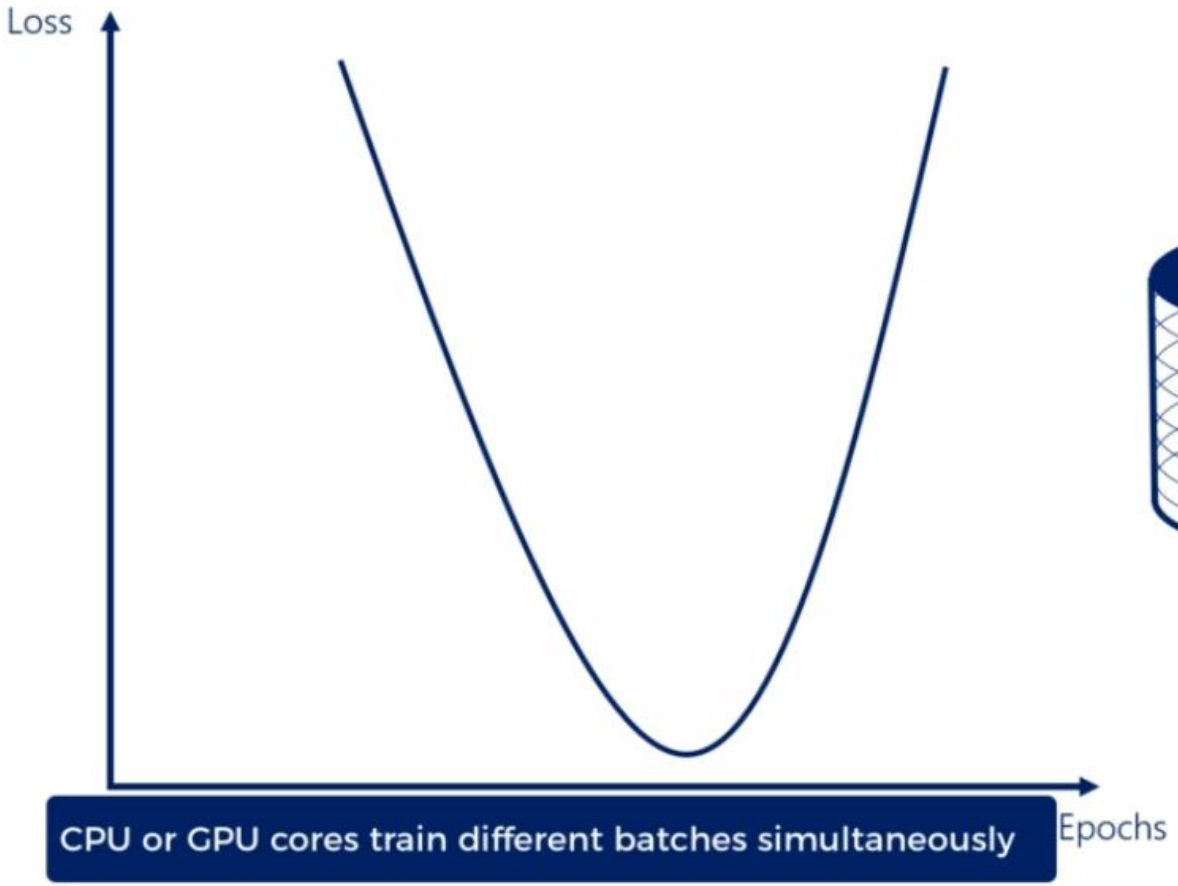
Batching

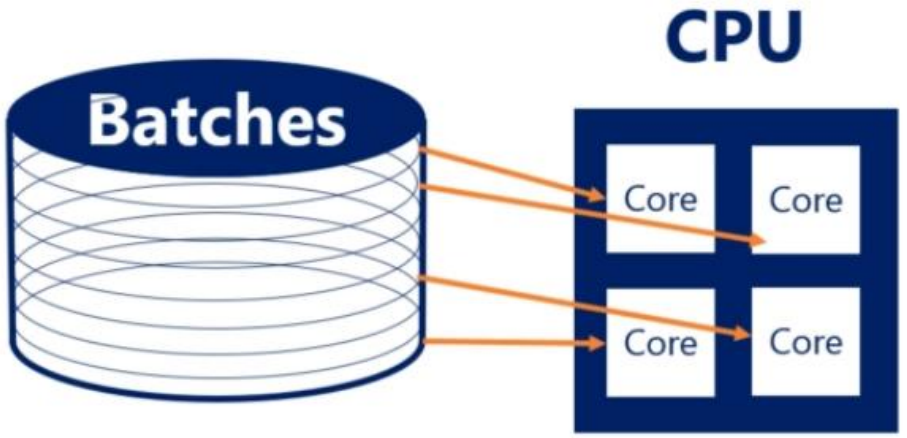
The SGD comes at a cost: it approximates things a bit





Batching


Loss
Epochs
CPU or GPU cores train different batches simultaneously

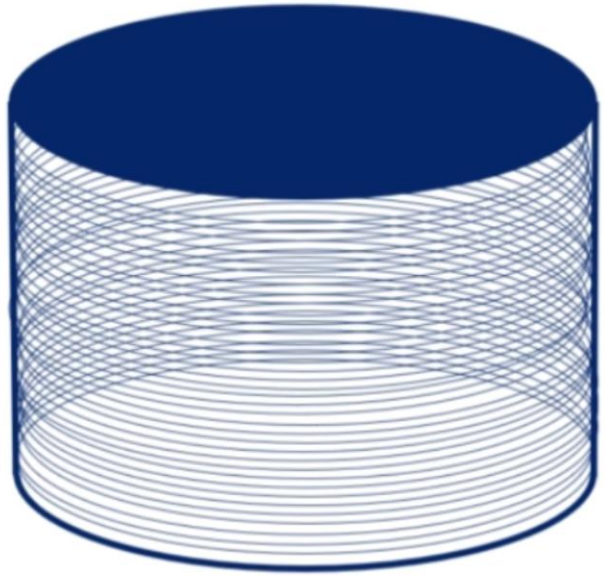

CPU
Batches
Core
Core
Core
Core

Stochastic gradient descent

	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	1
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	1
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}	1
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	x_{n6}	x_{n7}	x_{n8}	1

10,000

10,000 x k



We update the weights after each **sample**!

Mini-batch GD

10,000

x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	x_{n6}	x_{n7}	x_{n8}

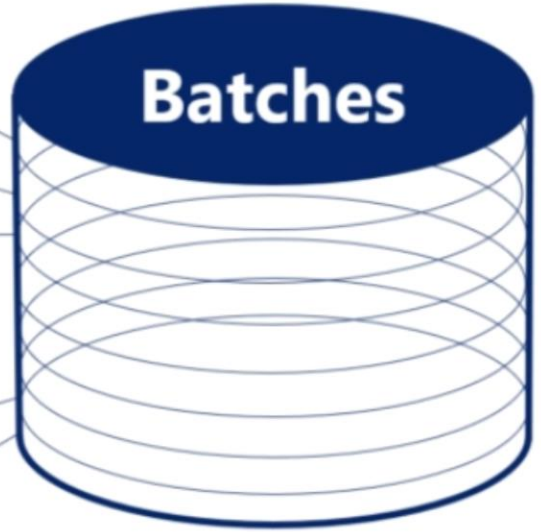
1,000

1,000

1,000

1,000

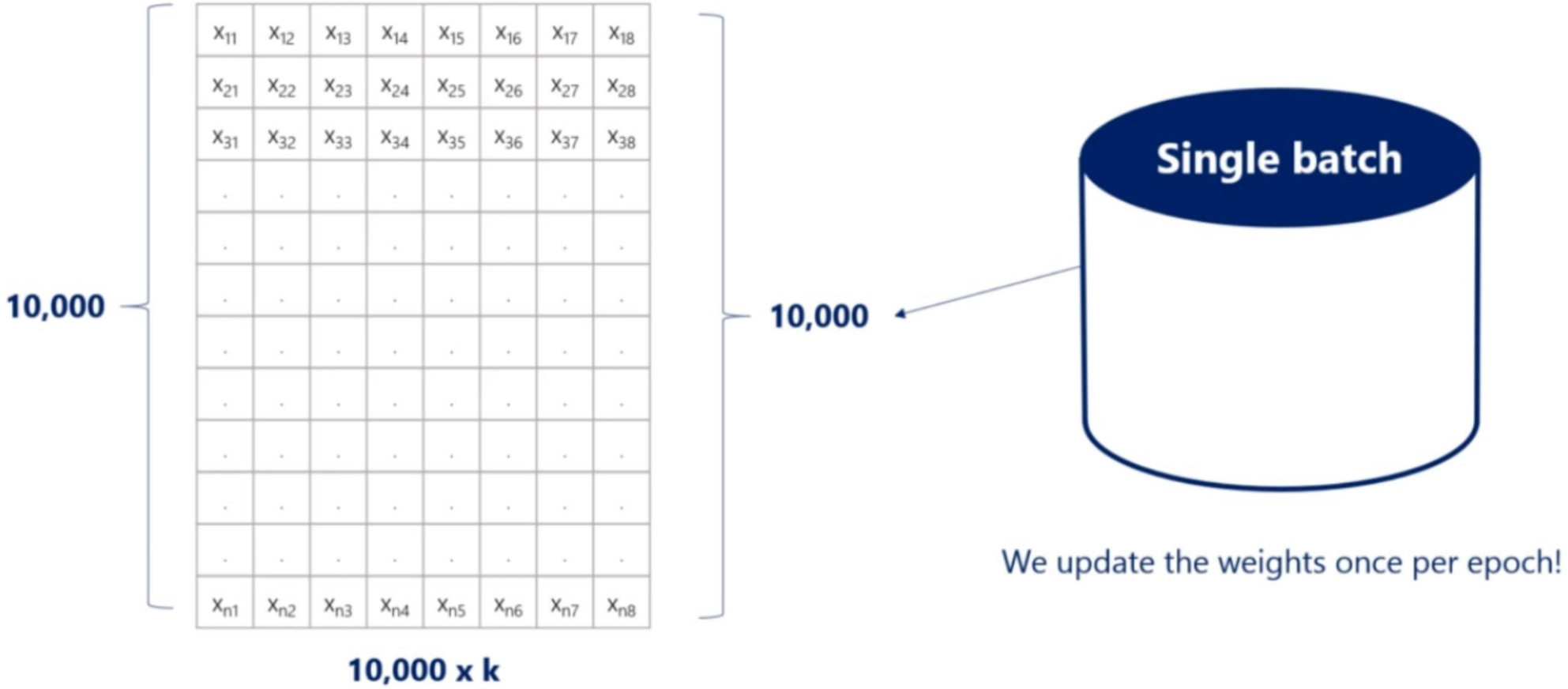
1,000



We update the weights after each **batch**!

Practitioners refer to the mini-batch GD as SGD

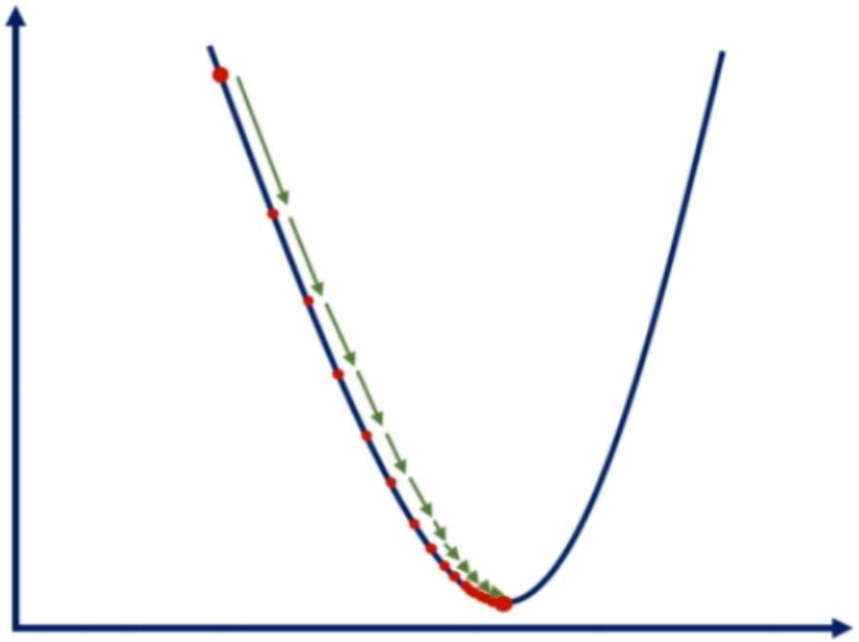
Batch GD


10,000
10,000 x k
10,000
Single batch
We update the weights once per epoch!

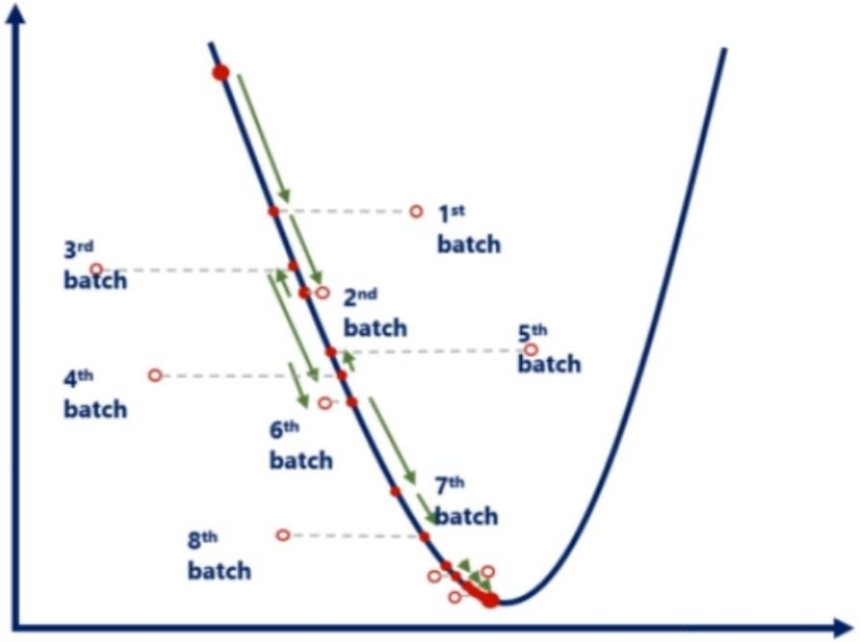
معایب گرادیان کاهشی و
نحوه‌ی رفع آنها

Gradient descent pitfalls

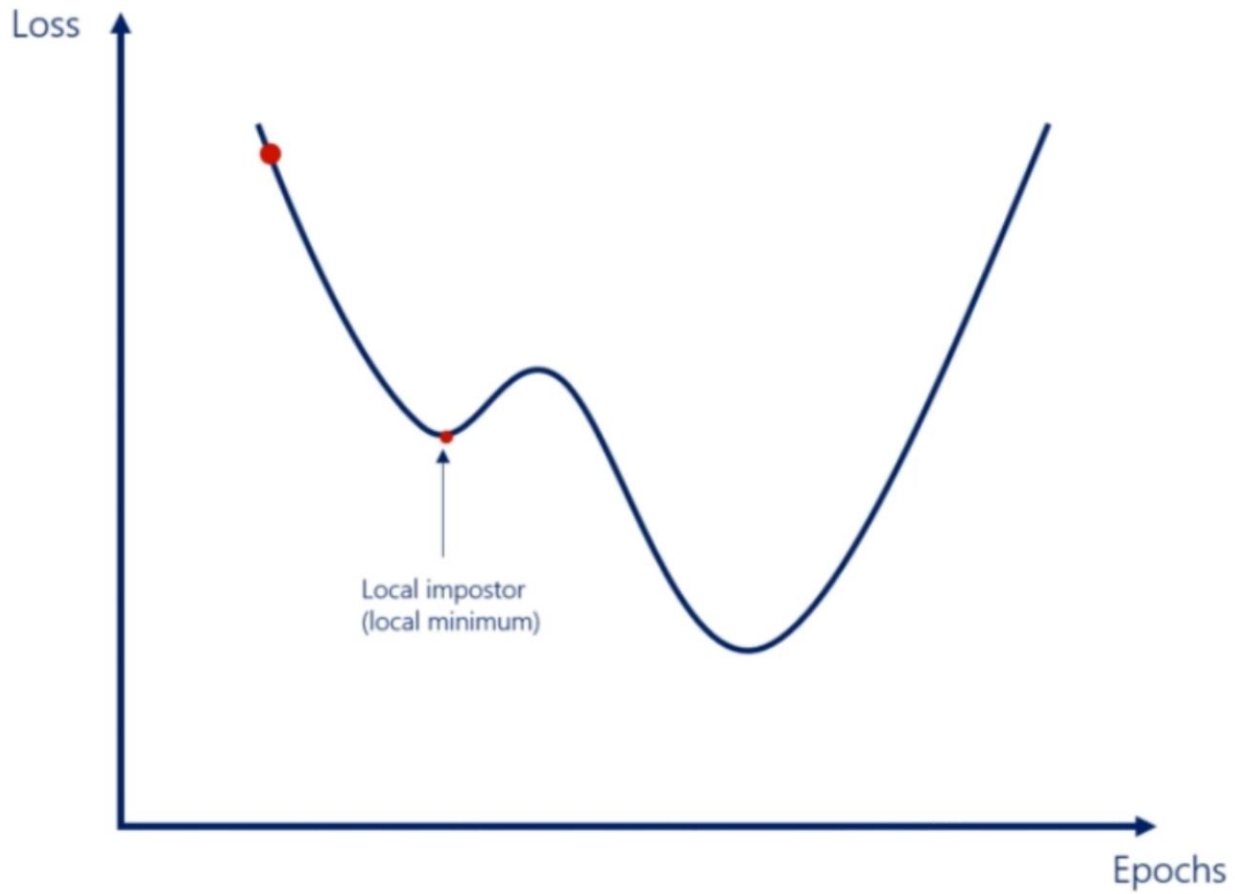
GD



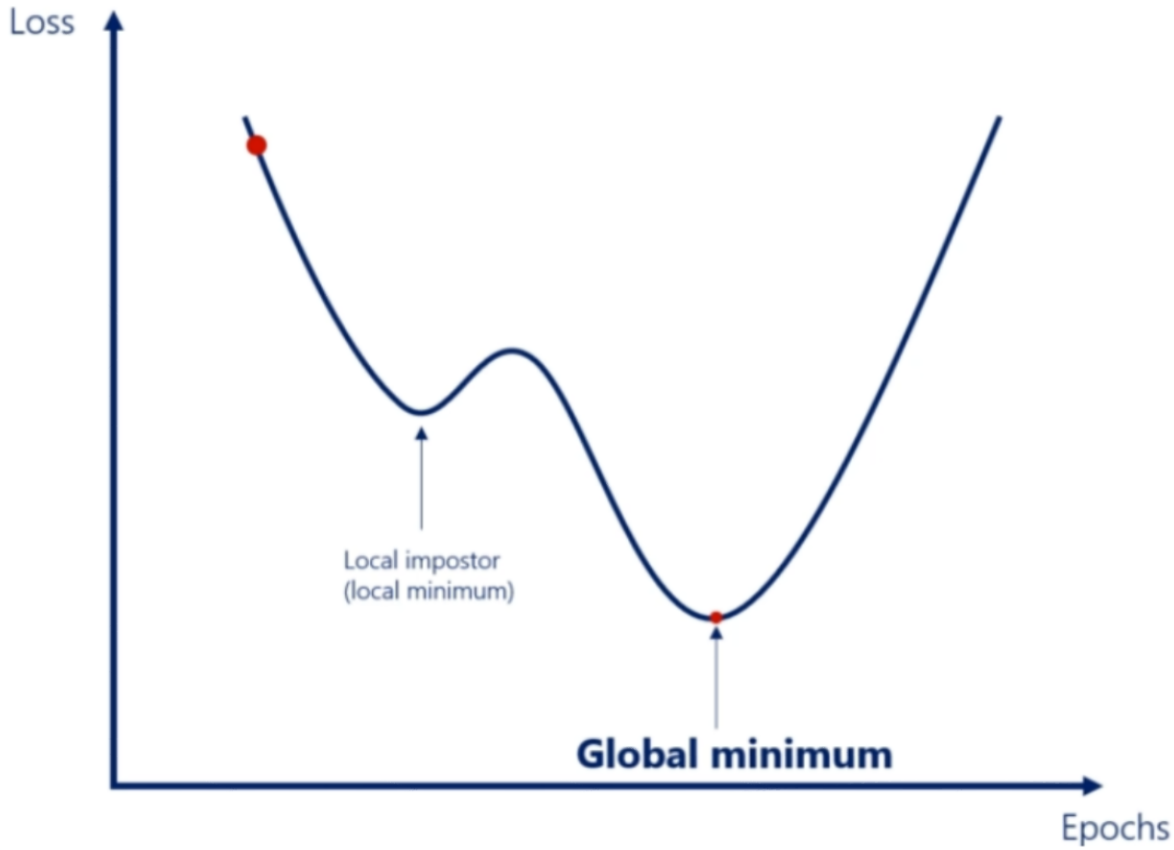
SGD



Gradient descent pitfalls

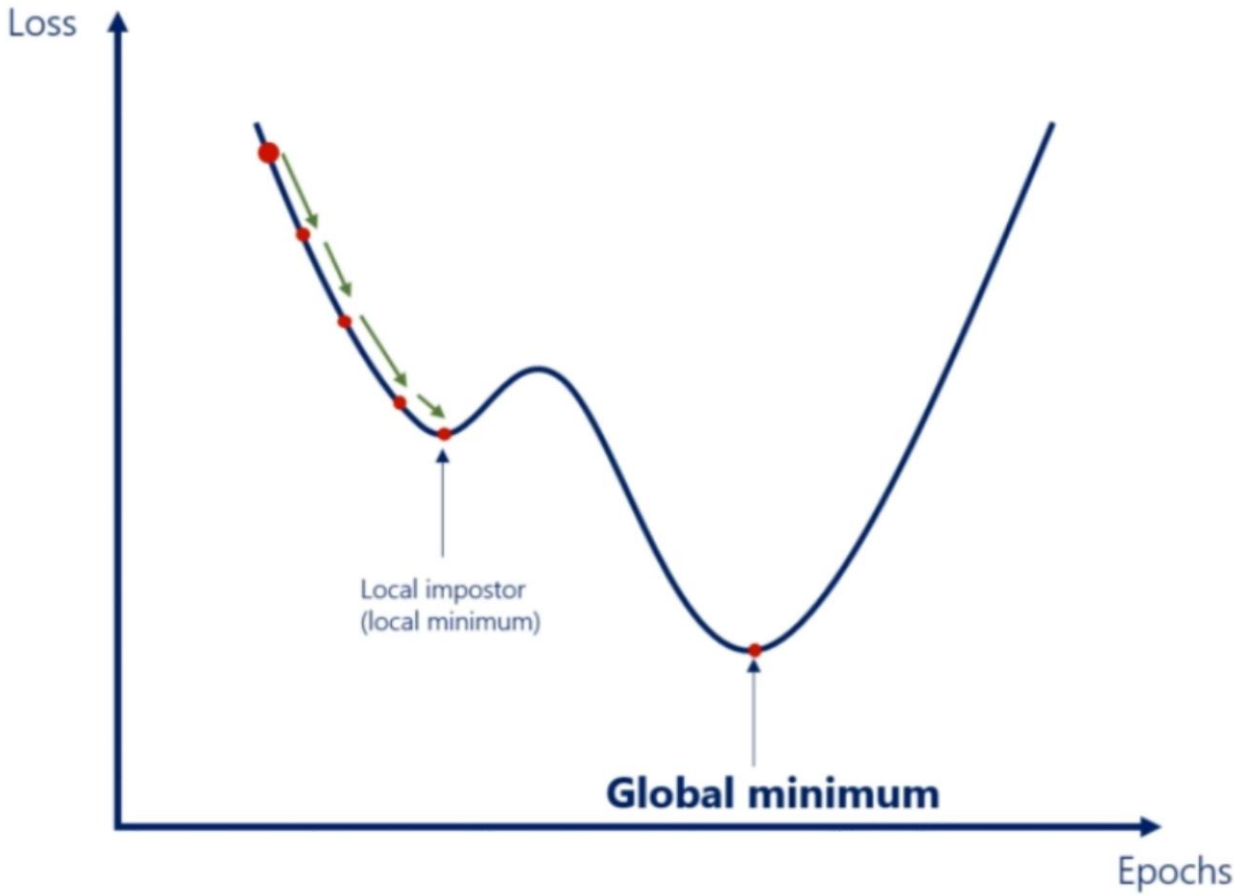


Gradient descent pitfalls

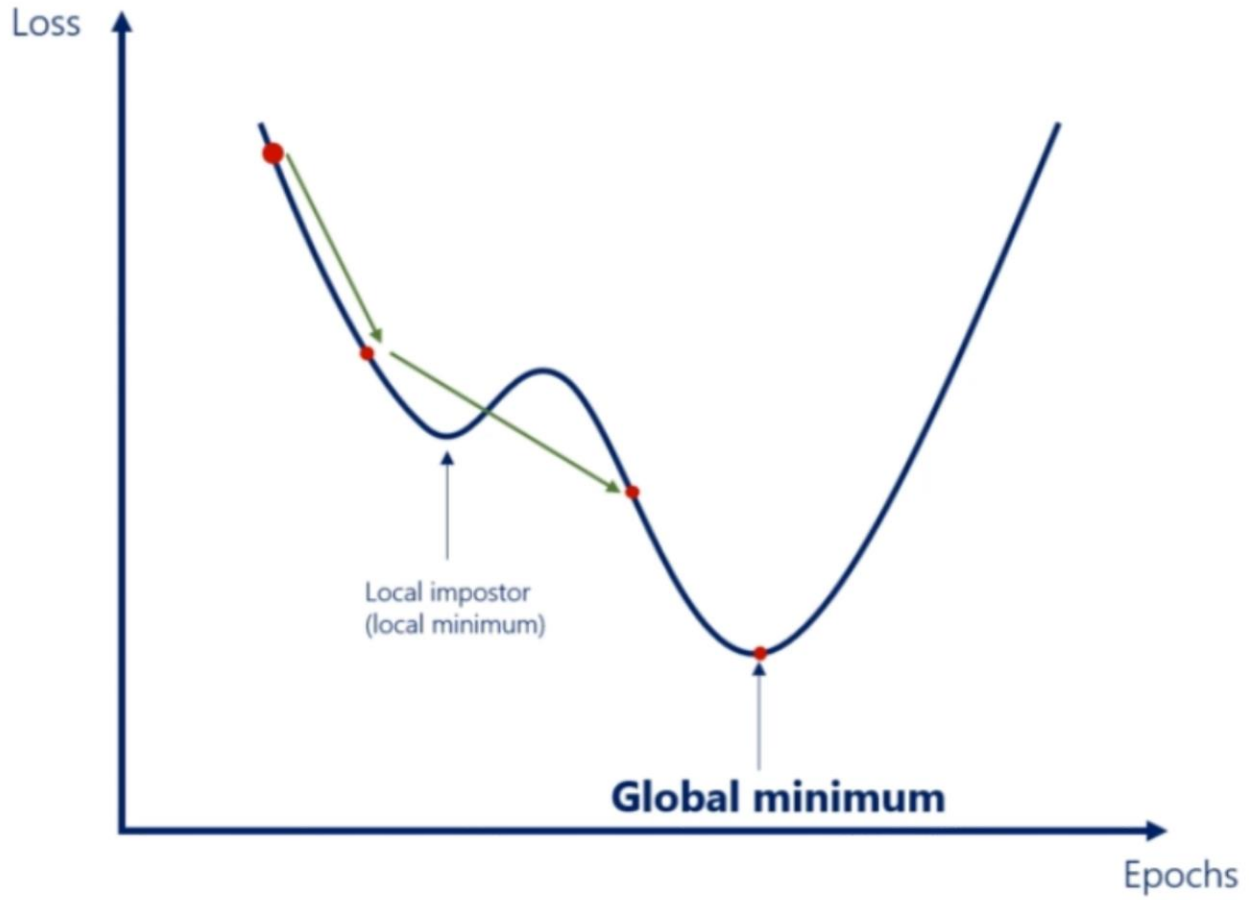


Each local minimum is a suboptimal solution to the optimization problem

Gradient descent pitfalls

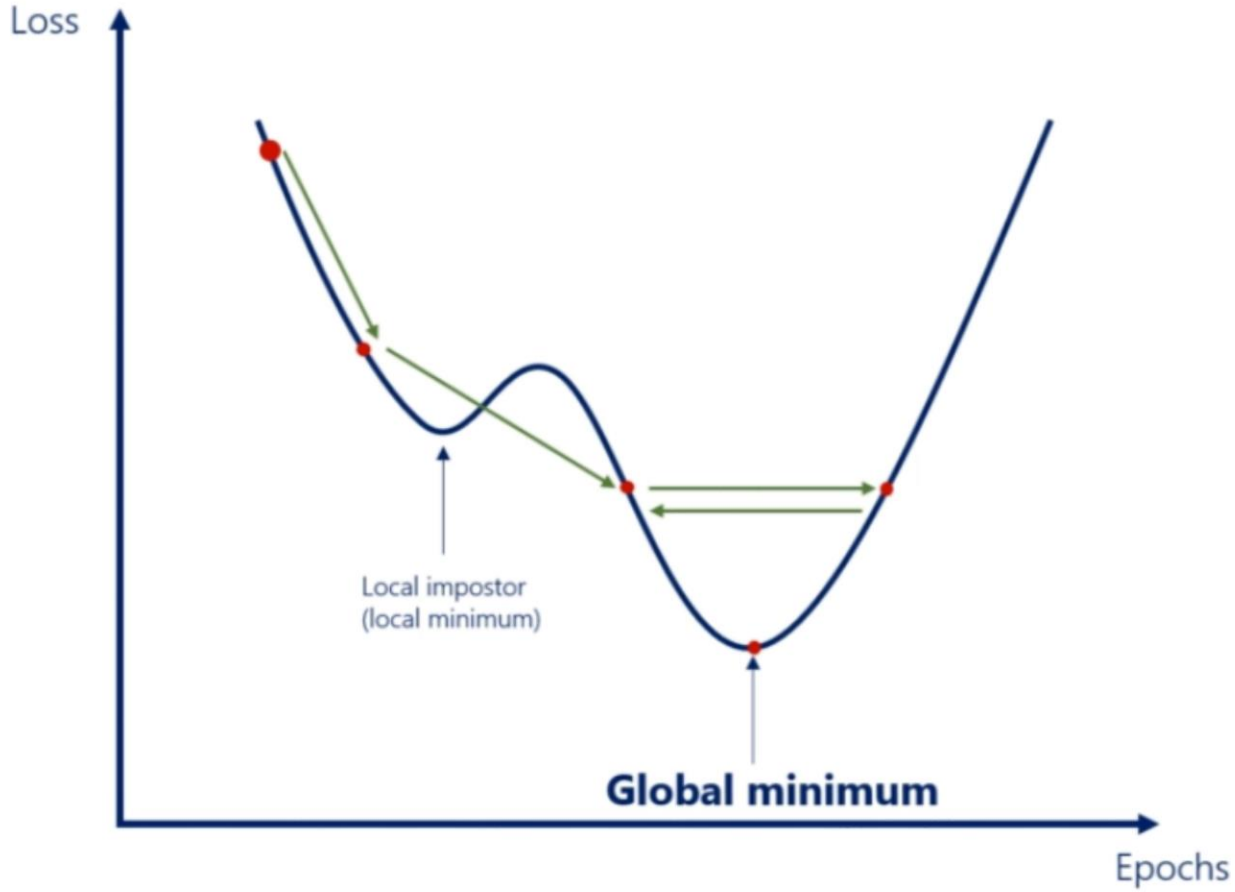


Gradient descent pitfalls



ممکن است با افزایش نرخ یادگیری، به سمت نقطه‌ی مینیمم حرکت کنیم

Gradient descent pitfalls



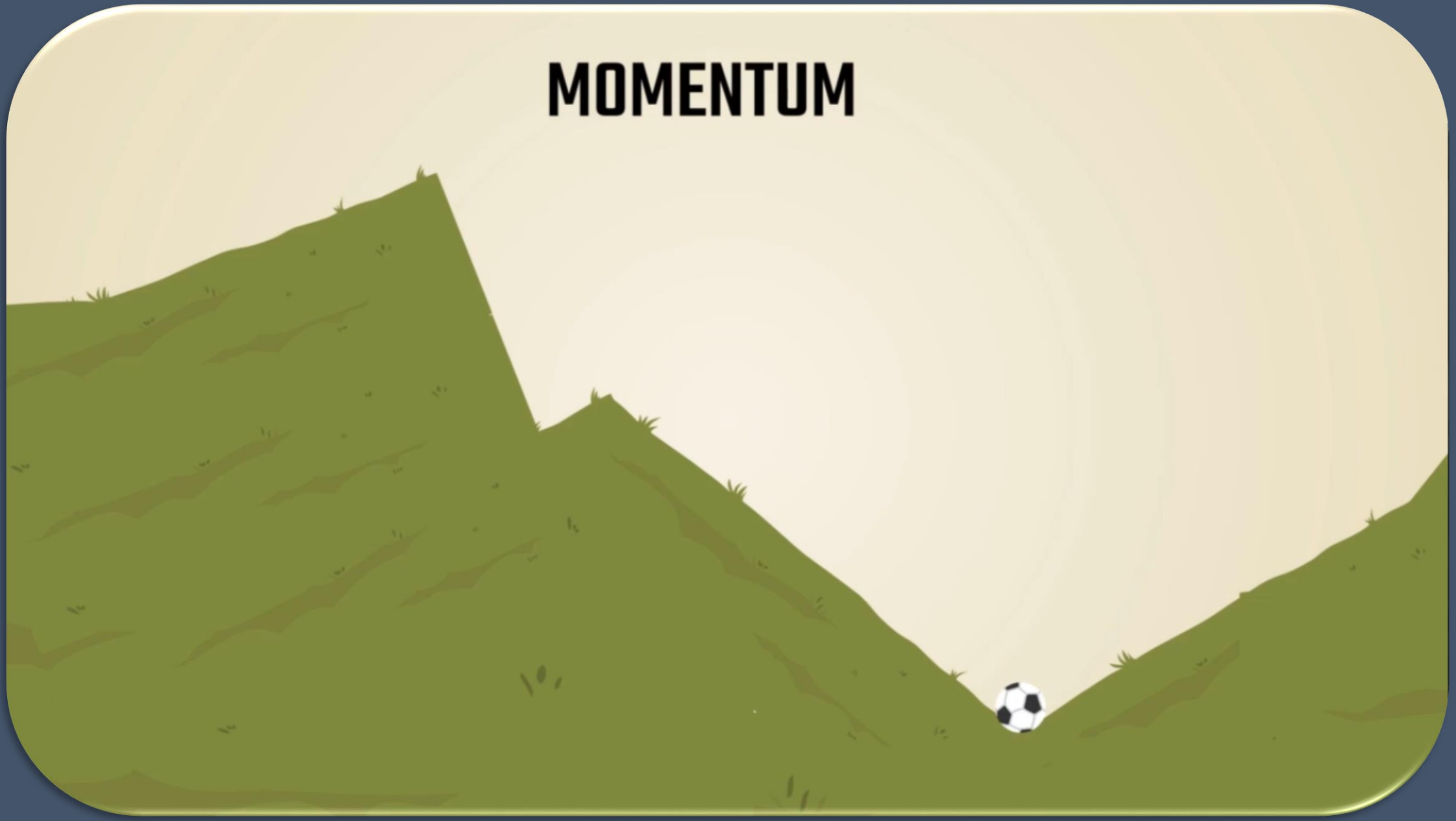
اما بزرگ انتخاب کردن نرخ یادگیری، خطر در نوسان بودن را ایجاد خواهد کرد

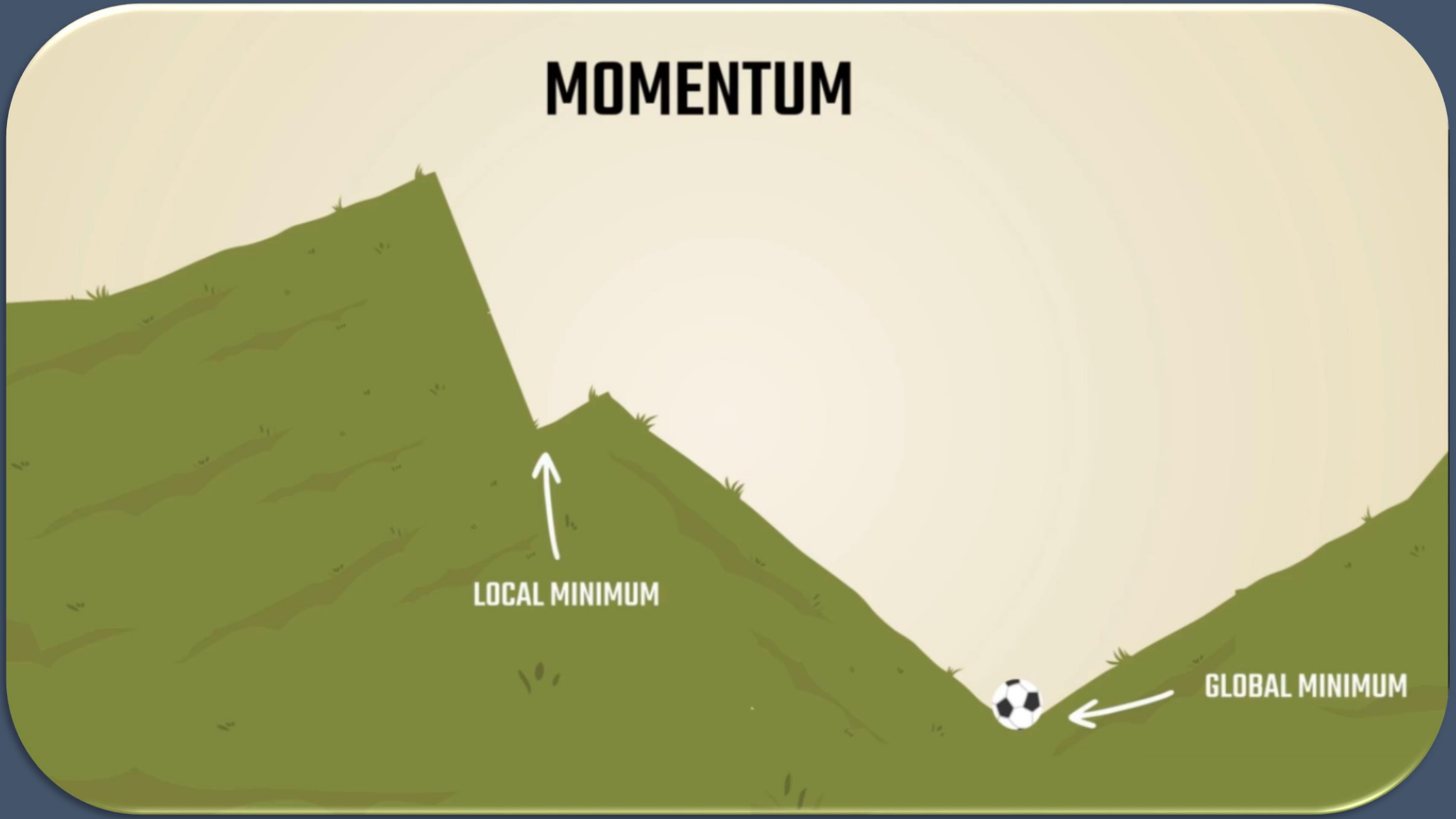
What is the remedy?

MOMENTUM

MOMENTUM

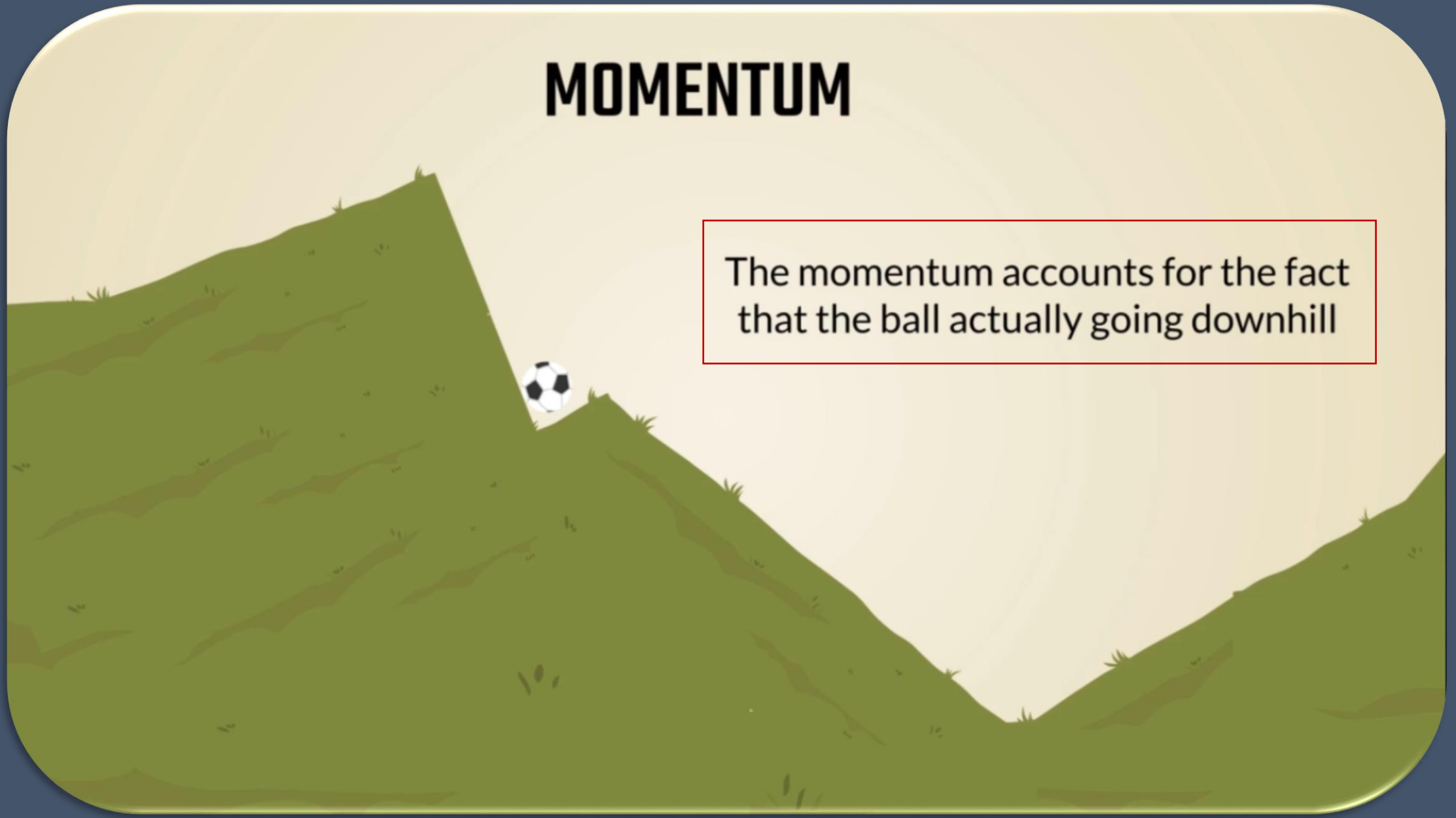
MOMENTUM



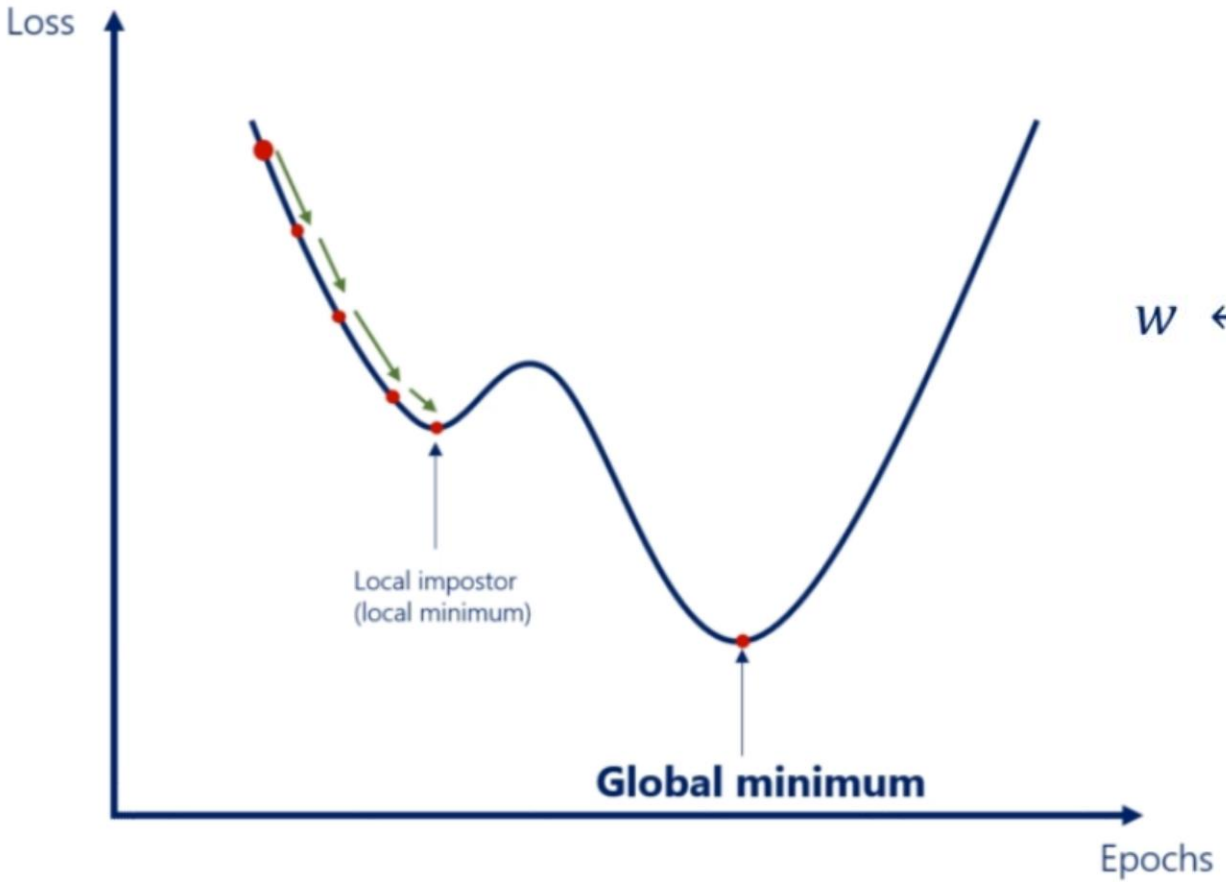

MOMENTUM
LOCAL MINIMUM
GLOBAL MINIMUM

MOMENTUM

MOMENTUM


The momentum accounts for the fact that the ball actually going downhill

Momentum



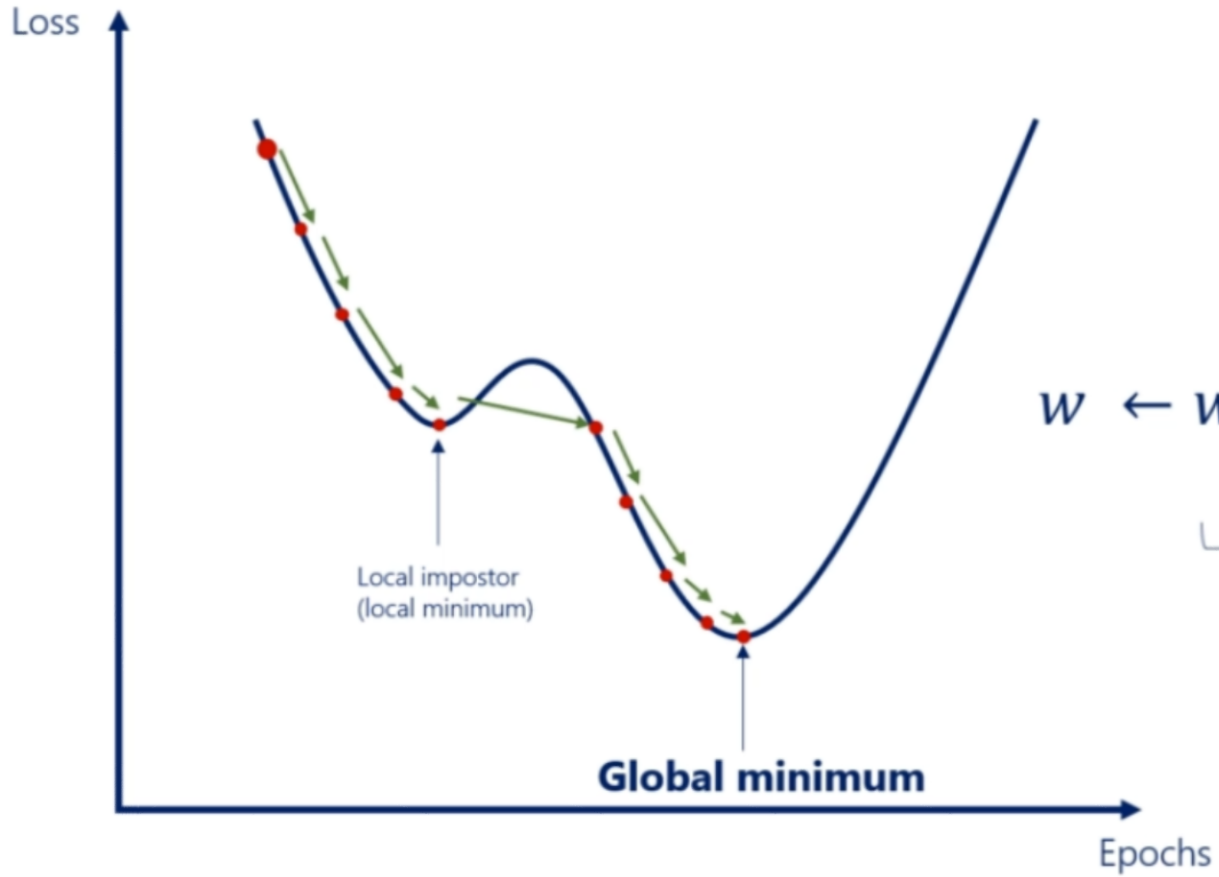
$$w \leftarrow w - \eta \frac{\partial L}{\partial w}$$

$$w \leftarrow w(t) - \eta \frac{\partial L}{\partial w}(t) - \eta \frac{\partial L}{\partial w}(t-1)$$

Current update | **Update a moment ago**

The best way to check how fast the ball rolls, is to check how fast it rolled a **moment** ago

Momentum

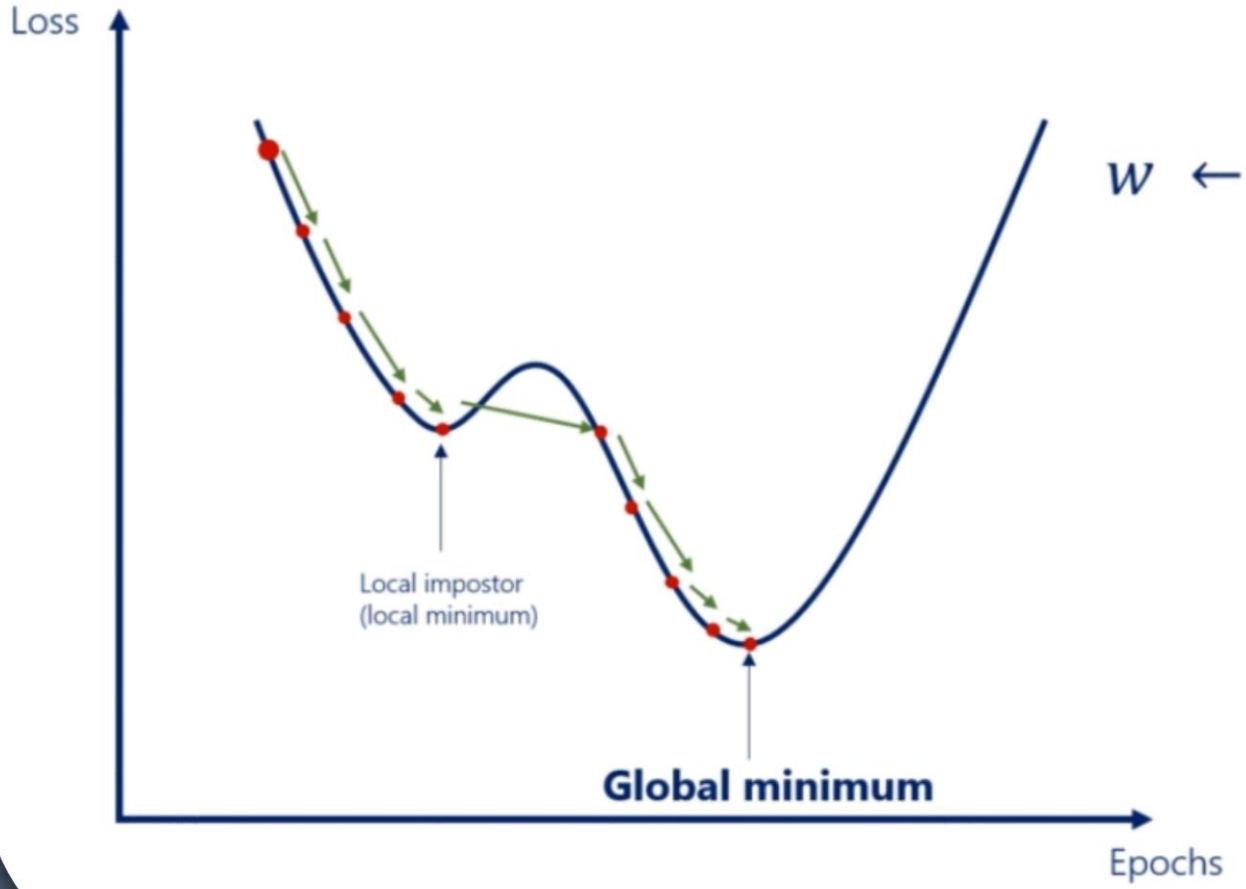


$$w \leftarrow w(t) - \eta \frac{\partial L}{\partial w}(t) - \eta \frac{\partial L}{\partial w}(t-1)$$

Current update

Update a moment ago

Momentum



$$w \leftarrow w(t) - \eta \frac{\partial L}{\partial w}(t) - \alpha \eta \frac{\partial L}{\partial w}(t-1)$$

Current update

Update a moment ago

$\alpha = 0.9$ is conventional

α is a hyperparameter

در ویدیوی بعدی نشون می‌دهیم که چطور
نرخ یادگیری بهینه‌ای رو انتخاب کنیم تا مدل
طبق میل و خواسته‌ی ما جلو بره

Just Stay Tuned