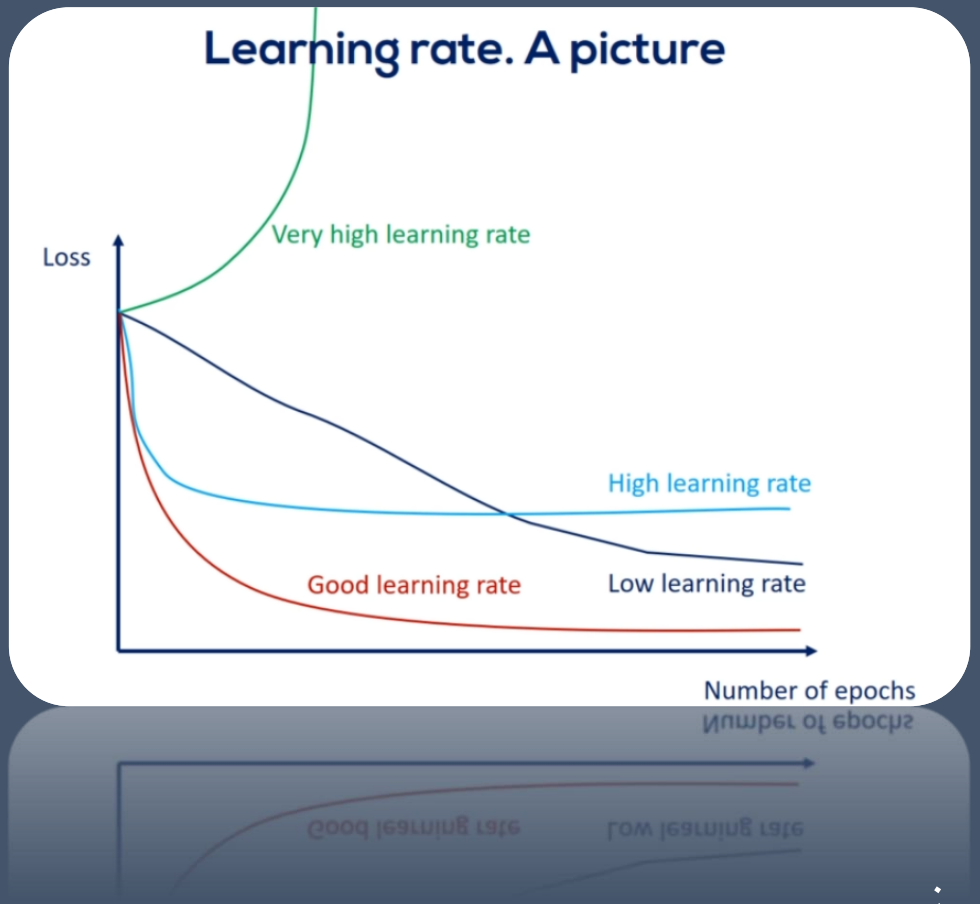
دوره‌ی آموزشی «علم داده»
Data Science Course



جلسه سی‌ام (بخش سوم)

نکاتی تکمیلی در خصوص نرخ یادگیری

مدرس: محمد فزونی
عضو هیئت علمی دانشگاه گنبدکاووس

Hyperparameters pre-set by us	vs	Parameters found by optimizing
Width		Weights (w)
Depth		Biases (b)
Learning rate (η)		

Learning rate (η)

به اندازه‌ی کافی بزرگ و به اندازه‌ی کافی کوچک. دقیقاً یعنی چی؟
چه خاکی باید تو سر بریزیم؟ 🙂

Small enough so we gently descend, instead of oscillating or diverging

Big enough so we reach it in a rational amount of time

Learning rate schedules

Small enough **Big enough**

1. We start from a high initial learning rate
2. At some point we lower the rate to avoid oscillation
3. Around the end we pick a very small rate to get a precise answer

Learning rate schedules

The simplest one is to set a predetermined **piecewise learning rate**

1. We start from a high initial learning rate	First 5 epochs $\eta = 0.1$
2. At some point we lower the rate to avoid oscillation	Next 5 epochs $\eta = 0.01$
3. Around the end we pick a very small rate to get a precise answer	Until the end $\eta = 0.001$

A learning rate schedule causes the loss to converge much faster

Learning rate schedules

1. We start from a high initial learning rate

2. At some point we lower the rate to avoid oscillation

3. Around the end we pick a very small rate to get a precise answer

$\eta_0 = 0.1$

current epoch

$$\eta = \eta_0 e^{-n/c}$$

some constant

Exponential schedule. Still simple, but much better as it smoothly **decays** the learning rate

Learning rate schedules

1. We start from a high initial learning rate

2. At some point we lower the rate to avoid oscillation

3. Around the end we pick a very small rate to get a precise answer

$\eta_0 = 0.1$
$\eta_1 = 0.0967$
$\eta_2 = 0.0905$
$\eta_3 = 0.0819$
$\eta_4 = 0.0717$
$\eta_5 = 0.0607$
$\eta_6 = 0.0497$
$\eta_7 = 0.0393$
$\eta_8 = 0.0301$

current epoch

$$\eta = \eta_0 e^{-n/20}$$

some constant

آیا برای انتخاب c قانونی وجود دارد؟

No set rule, but same order of magnitude

e.g. if we need:

100 epochs, $50 < c < 500$
1000 epochs, $500 < c < 5000$

The exact value is not important. **The presence of the learning schedule** does

Hyperparameters pre-set by us	vs	Parameters found by optimizing
Width		
Depth		Weights (w)
Learning rate (η)		
Batch size		Biases (b)
Momentum coefficient (α)		
Decay coefficient (c)		

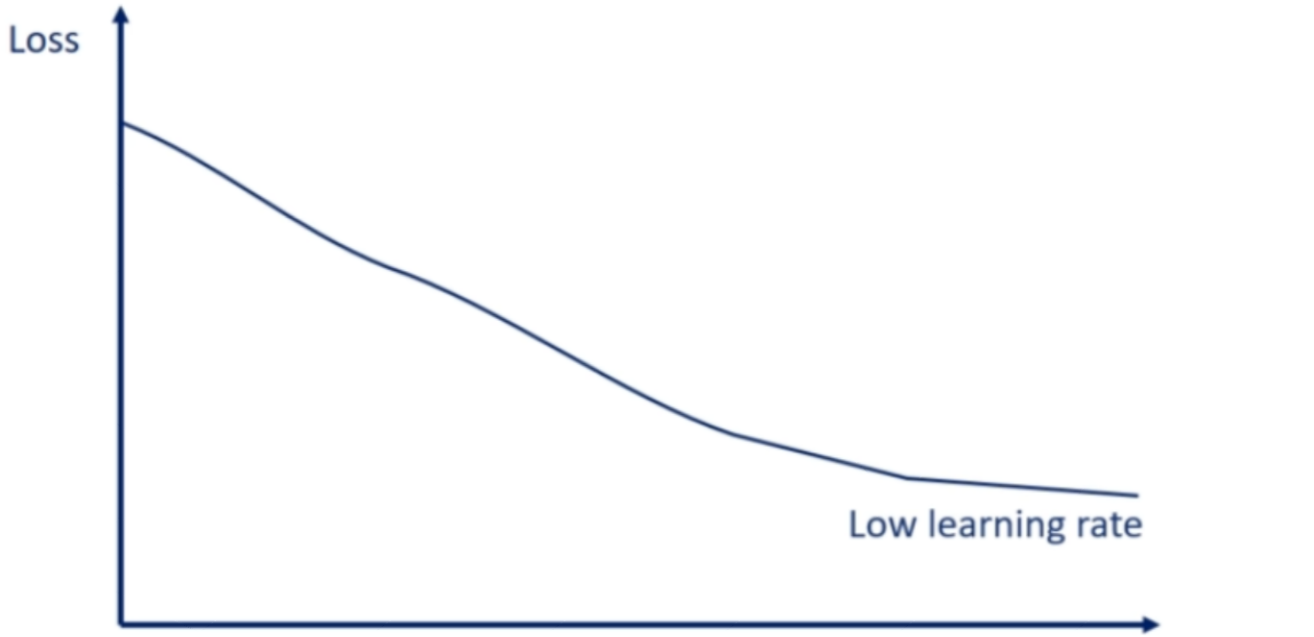
Our cost?

We increase the number of hyperparameters we must pick values for

مشاهده‌ی نرخ یادگیری روی نمودار بهترین ابزار برای انتخاب

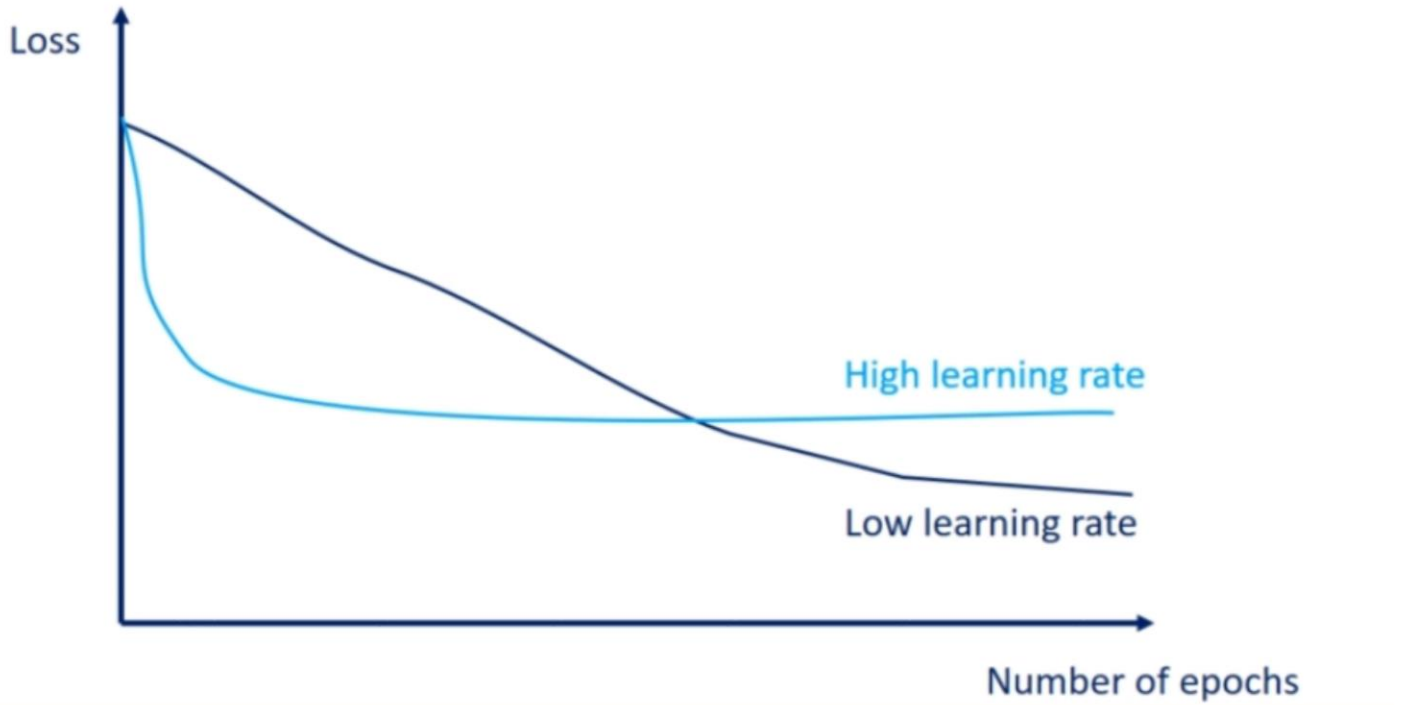
یک عکس بهتر از ساعت‌ها حرف است!

Learning rate. A picture



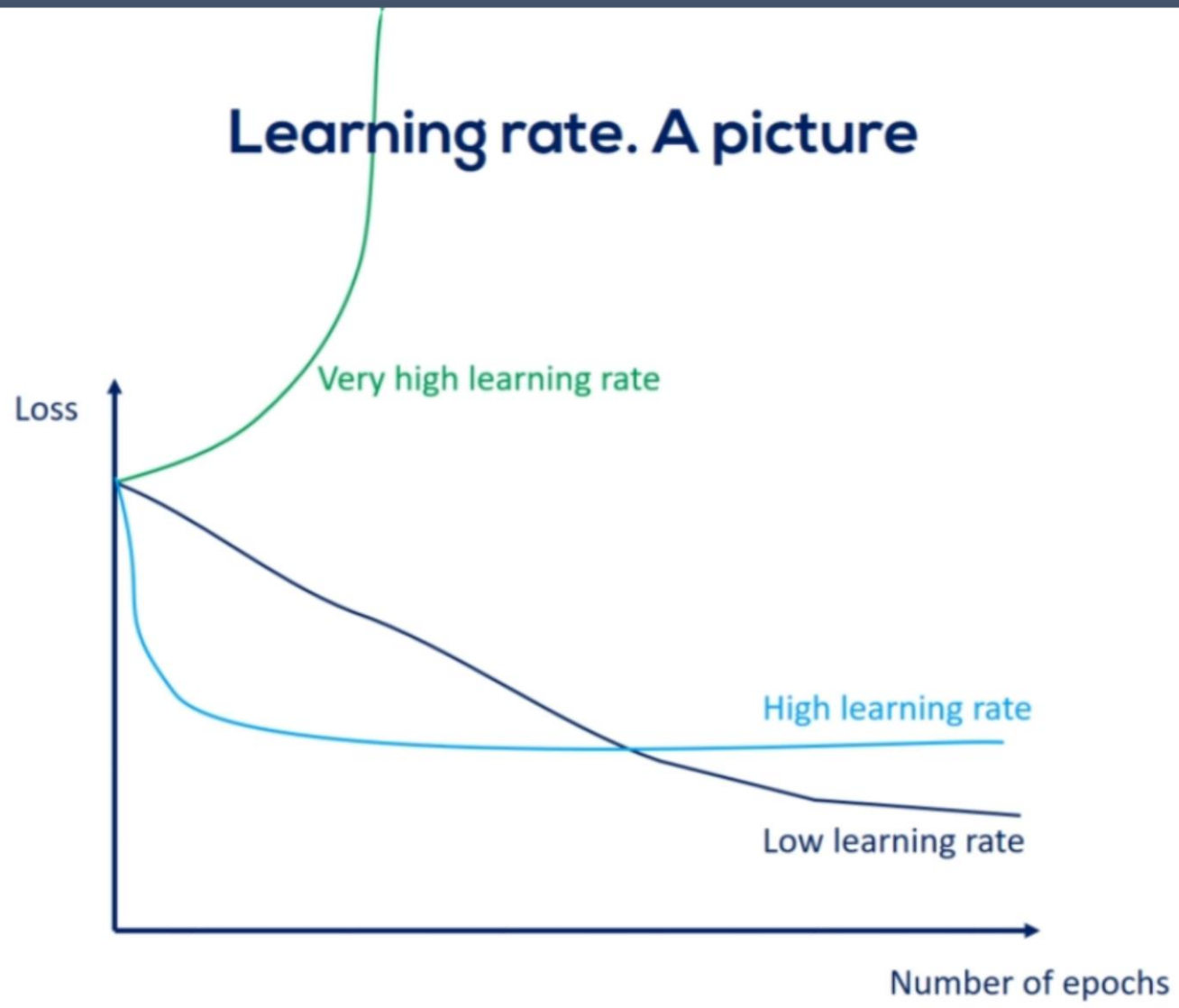
A low learning rate would minimize the loss but quite slowly

Learning rate. A picture



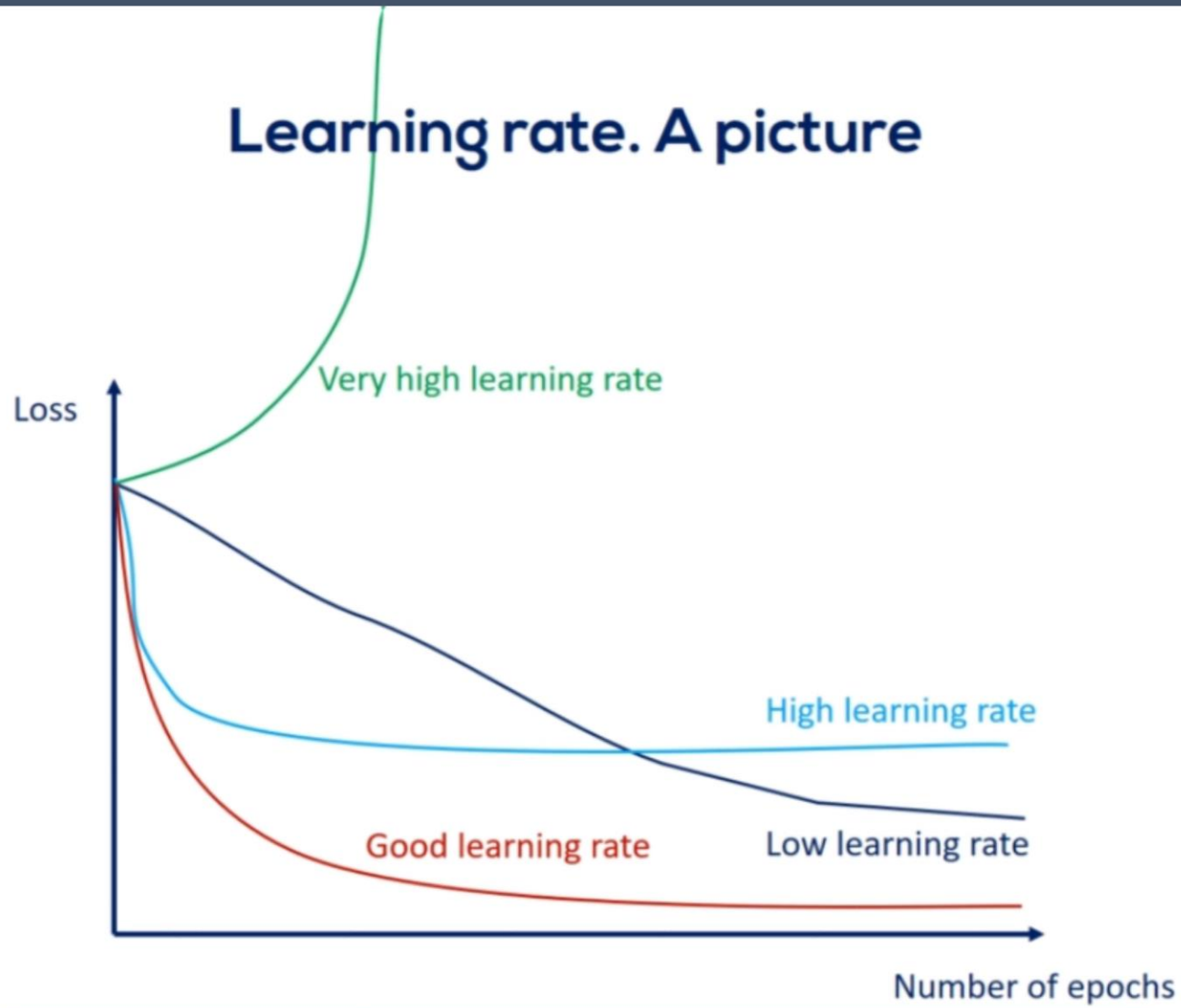
With a high learning rate, the loss is minimized faster but only to a certain extent

Learning rate. A picture



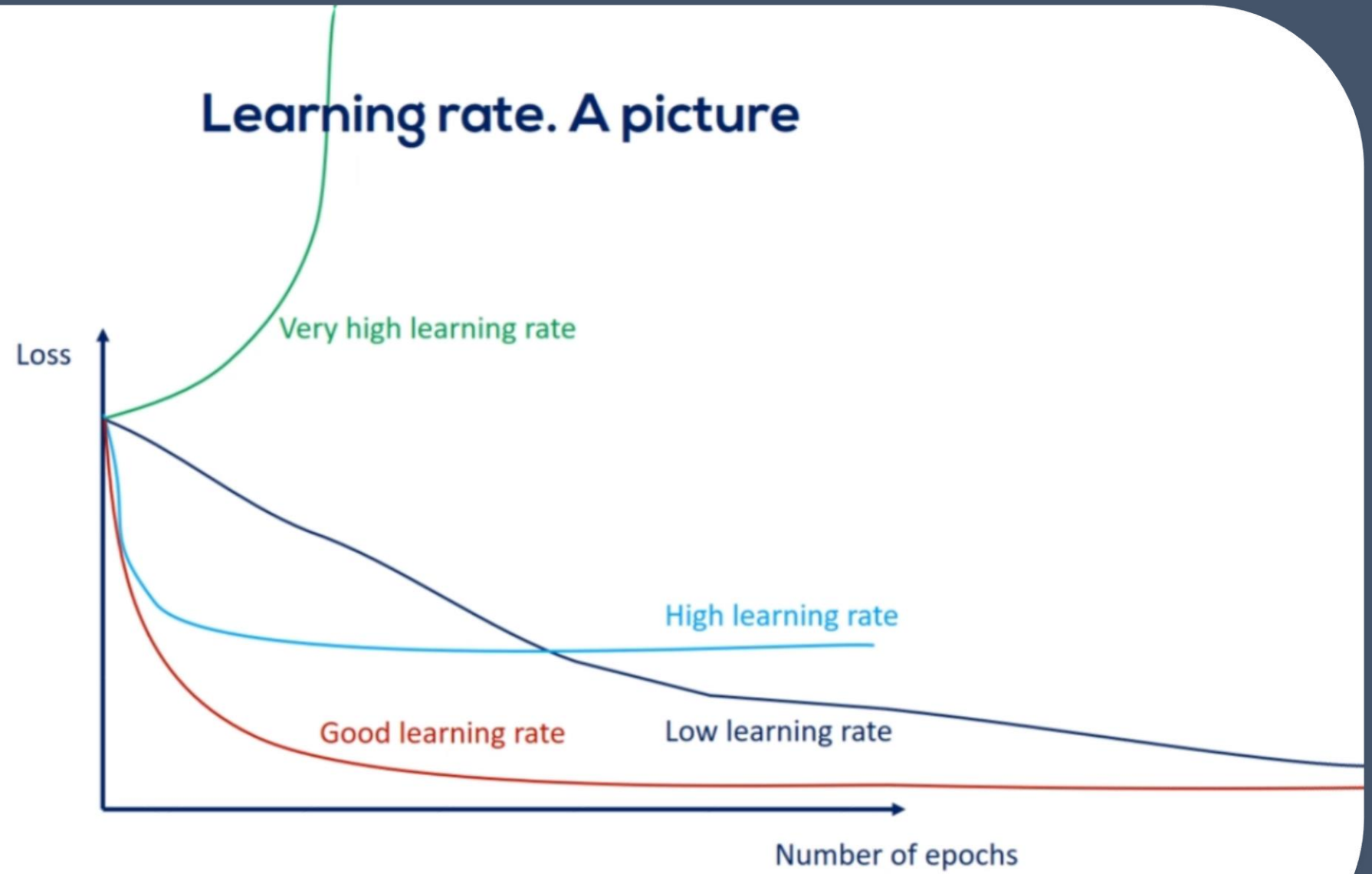
A **very** high learning rate would not even minimize the loss!

Learning rate. A picture



A learning rate following a schedule would minimize the loss faster than a low learning rate, and more accurately than a high learning rate

Learning rate. A picture



A low learning rate eventually converges with the good learning rate

در ویدیوی بعدی راجع به سه تا الگوریتم بسیار عالی
صحبت خواهیم کرد که در واقع برای بهینه‌سازی هستند، ولی خودشون نرخ یادگیری رو به خوبی مدیریت می‌کنند تا بهترین نتیجه بدست بیاد

AdaGrad, RMSProp & Adam