


# Gradient Descent Convergence

**Yingzhen Li**

Department of Computing  
Imperial College London

@liyzhen2  
yingzhen.li@imperial.ac.uk

October 24, 2022

# Gradient descent

We can use gradient descent to find the solution of

$$\theta^* = \arg \min L(\theta)$$

But when does gradient descent converge to a (local) optimum?

Skills you will learn from this lecture:

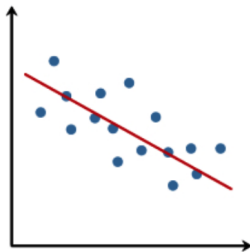
- Analysing linear regression models
- Applying eigendecomposition techniques

# Gradient descent for linear regression

Fitting linear regression models:

- Dataset:  $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$ ,  
 $\mathbf{X} = [x_1, \dots, x_N]^\top \in \mathbb{R}^{N \times D}$ ,  
 $\mathbf{y} = [y_1, \dots, y_N]^\top \in \mathbb{R}^{N \times 1}$
- Goal: find  $\boldsymbol{\theta} \in \mathbb{R}^{D \times 1}$  such that

$$\mathbf{y} \approx \mathbf{X}\boldsymbol{\theta}$$



# Gradient descent for linear regression

A typical linear regression model:

- $\mathbf{x} \in \mathbb{R}^{D \times 1}$ : input features;  $y \in \mathbb{R}$ : output value
- Model and loss:

$$f(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}^\top \boldsymbol{\theta}, \quad y = f(\mathbf{x}, \boldsymbol{\theta}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$L(\boldsymbol{\theta}) = \frac{1}{2\sigma^2} \sum_n (f(\mathbf{x}_n, \boldsymbol{\theta}) - y_n)^2$$

# Gradient descent for linear regression

A typical linear regression model:

- $\mathbf{x} \in \mathbb{R}^{D \times 1}$ : input features;  $y \in \mathbb{R}$ : output value
- Model and loss:

$$f(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}^\top \boldsymbol{\theta}, \quad y = f(\mathbf{x}, \boldsymbol{\theta}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$L(\boldsymbol{\theta}) = \frac{1}{2\sigma^2} \sum_n (f(\mathbf{x}_n, \boldsymbol{\theta}) - y_n)^2$$

- Rewriting the loss in matrix form:

$$L(\boldsymbol{\theta}) = \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$$

# Gradient descent for linear regression

Gradient descent to find  $\theta^*$ :

Assume constant step-sizes  $\gamma_t = \gamma$ :

1. Define **starting point**  $\theta_0$ , set  $t \leftarrow 0$
2. Set  $\theta_{t+1} = \theta_t - \gamma_t \nabla_{\theta} L(\theta_t)$ ,  $t \leftarrow t + 1$

$$\begin{aligned}\theta_{t+1} &= \theta_t - \gamma_t \nabla_{\theta} L(\theta_t) \\ &= \theta_t - \gamma \frac{1}{\sigma^2} \mathbf{X}^{\top} (\mathbf{X} \theta_t - \mathbf{y})\end{aligned}$$

# Gradient descent for linear regression

Gradient descent to find  $\theta^*$ :

Assume constant step-sizes  $\gamma_t = \gamma$ :

1. Define **starting point**  $\theta_0$ , set  $t \leftarrow 0$
2. Set  $\theta_{t+1} = \theta_t - \gamma_t \nabla_{\theta} L(\theta_t)$ ,  $t \leftarrow t + 1$

$$\begin{aligned}\theta_{t+1} &= \theta_t - \gamma_t \nabla_{\theta} L(\theta_t) \\ &= \theta_t - \gamma \frac{1}{\sigma^2} \mathbf{X}^{\top} (\mathbf{X} \theta_t - \mathbf{y}) \\ &= (\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^{\top} \mathbf{X}) \theta_t + \frac{\gamma}{\sigma^2} \mathbf{X}^{\top} \mathbf{y}\end{aligned}$$

3. Repeat 1 until stopping criterion.

# Gradient descent for linear regression

Gradient descent to find  $\theta^*$ :

Assume constant step-sizes  $\gamma_t = \gamma$ :

- GD returns the following iterative updates:

$$\theta_{t+1} = (\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X}) \theta_t + \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{y}$$

- We would like to figure out  $\theta_t$  as a function of  $\theta_0$  and  $\gamma$ !  
(and also other hyper-parameters & data)



# Gradient descent for linear regression

Arithmetico-geometric sequence:

If a sequence  $\{\theta_0, \theta_1, \dots, \theta_T\}$  is defined by

$$\theta_{t+1} = \mathbf{B}\theta_t + \mathbf{c}, \quad t \geq 0,$$

Then we have

$$\theta_{t+1} = \mathbf{A}(\theta_t + \beta) - \beta, \quad \text{for some } \mathbf{A}, \beta.$$

# Gradient descent for linear regression

Arithmetico-geometric sequence:

If a sequence  $\{\theta_0, \theta_1, \dots, \theta_T\}$  is defined by

$$\theta_{t+1} = \mathbf{B}\theta_t + \mathbf{c}, \quad t \geq 0,$$

Then we have

$$\theta_{t+1} = \mathbf{A}(\theta_t + \beta) - \beta, \quad \text{for some } \mathbf{A}, \beta.$$

Let's work out what are  $\mathbf{A}$  and  $\beta$ :

$$\begin{aligned}\theta_{t+1} &= \mathbf{A}(\theta_t + \beta) - \beta = \mathbf{B}\theta_t + \mathbf{c} \\ \Leftrightarrow \quad \mathbf{A}\theta_t + (\mathbf{A} - \mathbf{I})\beta &= \mathbf{B}\theta_t + \mathbf{c} \\ \Leftrightarrow \quad \mathbf{A} &= \mathbf{B}, \quad \beta = (\mathbf{B} - \mathbf{I})^{-1}\mathbf{c}\end{aligned}$$

# Gradient descent for linear regression

Arithmetico-geometric sequence:

If a sequence  $\{\theta_0, \theta_1, \dots, \theta_T\}$  is defined by

$$\theta_{t+1} = \mathbf{B}\theta_t + \mathbf{c}, \quad t \geq 0,$$

Then we have

$$\theta_{t+1} = \mathbf{A}(\theta_t + \beta) - \beta, \quad \text{for some } \mathbf{A}, \beta.$$

Let's work out what are  $\mathbf{A}$  and  $\beta$ :

$$\theta_{t+1} = \mathbf{A}(\theta_t + \beta) - \beta = \mathbf{B}\theta_t + \mathbf{c}$$

$$\Leftrightarrow \mathbf{A}\theta_t + (\mathbf{A} - \mathbf{I})\beta = \mathbf{B}\theta_t + \mathbf{c}$$

$$\Leftrightarrow \mathbf{A} = \mathbf{B}, \quad \beta = (\mathbf{B} - \mathbf{I})^{-1}\mathbf{c}$$

$$\Rightarrow \theta_{t+1} = \mathbf{B}(\theta_t + (\mathbf{B} - \mathbf{I})^{-1}\mathbf{c}) - (\mathbf{B} - \mathbf{I})^{-1}\mathbf{c}$$

$$\Rightarrow \theta_t = \mathbf{B}^t(\theta_0 + (\mathbf{B} - \mathbf{I})^{-1}\mathbf{c}) - (\mathbf{B} - \mathbf{I})^{-1}\mathbf{c}$$

# Gradient descent for linear regression

Gradient descent to find  $\theta^*$ :

Assume constant step-sizes  $\gamma_t = \gamma$ :

- GD returns the following iterative updates:

$$\theta_{t+1} = (\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X}) \theta_t + \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{y}$$

- Solving this iterative update returns:

# Gradient descent for linear regression

Gradient descent to find  $\theta^*$ :

Assume constant step-sizes  $\gamma_t = \gamma$ :

- GD returns the following iterative updates:

$$\theta_{t+1} = (\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X}) \theta_t + \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{y}$$

- Solving this iterative update returns:

$$\theta_t = (\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^t (\theta_0 - \theta^*) + \theta^*, \quad \theta^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- GD converges ( $\theta_t \rightarrow \theta^*$ ) if  $(\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^t (\theta_0 - \theta^*) \rightarrow \mathbf{0}$

# Convergence of GD for linear regression

Gradient descent with constant step-size to find  $\theta^*$ :

$$\theta_t = (\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^t (\theta_0 - \theta^*) + \theta^*, \quad \theta^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- ▶ The  $\ell_2$  distance between  $\theta_t$  and  $\theta^*$ :

$$\begin{aligned} \|\theta_t - \theta^*\|_2^2 &= \|(\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^t (\theta_0 - \theta^*)\|_2^2 \\ &= |(\theta_0 - \theta^*)^\top (\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^{2t} (\theta_0 - \theta^*)| \end{aligned}$$

# Convergence of GD for linear regression

Gradient descent with constant step-size to find  $\theta^*$ :

$$\theta_t = (\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^t (\theta_0 - \theta^*) + \theta^*, \quad \theta^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- ▶ The  $\ell_2$  distance between  $\theta_t$  and  $\theta^*$ :

$$\begin{aligned} \|\theta_t - \theta^*\|_2^2 &= \|(\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^t (\theta_0 - \theta^*)\|_2^2 \\ &= |(\theta_0 - \theta^*)^\top (\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^{2t} (\theta_0 - \theta^*)| \end{aligned}$$

- ▶ Fact:  $\lambda_{\min}(\mathbf{A})\|x\|_2^2 \leq x^\top \mathbf{A}x \leq \lambda_{\max}(\mathbf{A})\|x\|_2^2$ :

$$\|\theta_t - \theta^*\|_2^2 \geq \lambda_{\min}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^{2t}) \|\theta_0 - \theta^*\|_2^2$$

$$\|\theta_t - \theta^*\|_2^2 \leq \lambda_{\max}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^{2t}) \|\theta_0 - \theta^*\|_2^2$$

# Convergence of GD for linear regression

Gradient descent with constant step-size to find  $\theta^*$ :

$$\theta_t = (\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^t (\theta_0 - \theta^*) + \theta^*, \quad \theta^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- ▶ The  $\ell_2$  distance between  $\theta_t$  and  $\theta^*$ :

$$\begin{aligned} \|\theta_t - \theta^*\|_2^2 &= \|(\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^t (\theta_0 - \theta^*)\|_2^2 \\ &= |(\theta_0 - \theta^*)^\top (\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^{2t} (\theta_0 - \theta^*)| \end{aligned}$$

- ▶ Fact:  $\lambda_{\min}(\mathbf{A})\|x\|_2^2 \leq x^\top \mathbf{A}x \leq \lambda_{\max}(\mathbf{A})\|x\|_2^2$ :

$$\|\theta_t - \theta^*\|_2^2 \geq \lambda_{\min}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^{2t}) \|\theta_0 - \theta^*\|_2^2$$

$$\|\theta_t - \theta^*\|_2^2 \leq \lambda_{\max}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^{2t}) \|\theta_0 - \theta^*\|_2^2$$



# Convergence of GD for linear regression

Gradient descent with constant step-size to find  $\theta^*$ :

$$\lambda_{min}^t \|\theta_0 - \theta^*\|_2^2 \leq \|\theta_t - \theta^*\|_2^2 \leq \lambda_{max}^t \|\theta_0 - \theta^*\|_2^2$$

$$\lambda_{min} := \lambda_{min}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2) \geq 0, \quad \lambda_{max} := \lambda_{max}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2)$$

# Convergence of GD for linear regression

Gradient descent with constant step-size to find  $\theta^*$ :

$$\lambda_{min}^t \|\theta_0 - \theta^*\|_2^2 \leq \|\theta_t - \theta^*\|_2^2 \leq \lambda_{max}^t \|\theta_0 - \theta^*\|_2^2$$

$$\lambda_{min} := \lambda_{min}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2) \geq 0, \quad \lambda_{max} := \lambda_{max}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2)$$

Convergence properties in difference cases:

1.  $\lambda_{max} < 1$ : always converge
2.  $\lambda_{min} \geq 1$ : always diverge
3.  $\lambda_{min} < 1$  but  $\lambda_{max} \geq 1$ : convergence depending on  $\theta_0$

# Convergence of GD for linear regression

Gradient descent with constant step-size to find  $\theta^*$ :

$$\lambda_{min}^t \|\theta_0 - \theta^*\|_2^2 \leq \|\theta_t - \theta^*\|_2^2 \leq \lambda_{max}^t \|\theta_0 - \theta^*\|_2^2$$

$$\lambda_{min} := \lambda_{min}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2) \geq 0, \quad \lambda_{max} := \lambda_{max}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2)$$

Deriving the eigenvalues  $\lambda_{min}, \lambda_{max}$ :

# Convergence of GD for linear regression

Gradient descent with constant step-size to find  $\theta^*$ :

$$\lambda_{min}^t \|\theta_0 - \theta^*\|_2^2 \leq \|\theta_t - \theta^*\|_2^2 \leq \lambda_{max}^t \|\theta_0 - \theta^*\|_2^2$$

$$\lambda_{min} := \lambda_{min}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2) \geq 0, \quad \lambda_{max} := \lambda_{max}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2)$$

Deriving the eigenvalues  $\lambda_{min}, \lambda_{max}$ :

- If  $\lambda$  is an eigenvalue of  $\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X}$ ,  
then  $\lambda^2$  is an eigenvalue of  $(\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2$

# Convergence of GD for linear regression

Gradient descent with constant step-size to find  $\theta^*$ :

$$\lambda_{min}^t \|\theta_0 - \theta^*\|_2^2 \leq \|\theta_t - \theta^*\|_2^2 \leq \lambda_{max}^t \|\theta_0 - \theta^*\|_2^2$$

$$\lambda_{min} := \lambda_{min}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2) \geq 0, \quad \lambda_{max} := \lambda_{max}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2)$$

Deriving the eigenvalues  $\lambda_{min}, \lambda_{max}$ :

- If  $\lambda$  is an eigenvalue of  $\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X}$ ,  
then  $\lambda^2$  is an eigenvalue of  $(\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2$
- If  $\lambda$  is an eigenvalue of  $\mathbf{X}^\top \mathbf{X}$ ,  
then  $1 - \frac{\gamma\lambda}{\sigma^2}$  is an eigenvalue of  $\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X}$ :

$$\mathbf{X}^\top \mathbf{X} \mathbf{q} = \lambda \mathbf{q} \quad \Leftrightarrow \quad (\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X}) \mathbf{q} = (1 - \frac{\gamma\lambda}{\sigma^2}) \mathbf{q}$$

# Convergence of GD for linear regression

Gradient descent with constant step-size to find  $\theta^*$ :

$$\lambda_{min}^t \|\theta_0 - \theta^*\|_2^2 \leq \|\theta_t - \theta^*\|_2^2 \leq \lambda_{max}^t \|\theta_0 - \theta^*\|_2^2$$

$$\lambda_{min} := \lambda_{min}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2) \geq 0, \quad \lambda_{max} := \lambda_{max}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2)$$

- ▶ If  $\lambda$  is an eigenvalue of  $\mathbf{X}^\top \mathbf{X}$ ,  
then  $(1 - \frac{\gamma\lambda}{\sigma^2})^2$  is an eigenvalue of  $(\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2$

# Convergence of GD for linear regression

Gradient descent with constant step-size to find  $\theta^*$ :

$$\lambda_{min}^t \|\theta_0 - \theta^*\|_2^2 \leq \|\theta_t - \theta^*\|_2^2 \leq \lambda_{max}^t \|\theta_0 - \theta^*\|_2^2$$

$$\lambda_{min} := \lambda_{min}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2) \geq 0, \quad \lambda_{max} := \lambda_{max}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2)$$

- If  $\lambda$  is an eigenvalue of  $\mathbf{X}^\top \mathbf{X}$ ,  
then  $(1 - \frac{\gamma\lambda}{\sigma^2})^2$  is an eigenvalue of  $(\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2$
- $\mathbf{X}^\top \mathbf{X}$  is positive semi-definite  $\Rightarrow \lambda \geq 0$

# Convergence of GD for linear regression

Gradient descent with constant step-size to find  $\theta^*$ :

$$\lambda_{min}^t \|\theta_0 - \theta^*\|_2^2 \leq \|\theta_t - \theta^*\|_2^2 \leq \lambda_{max}^t \|\theta_0 - \theta^*\|_2^2$$

$$\lambda_{min} := \lambda_{min}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2) \geq 0, \quad \lambda_{max} := \lambda_{max}((\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2)$$

- If  $\lambda$  is an eigenvalue of  $\mathbf{X}^\top \mathbf{X}$ ,  
then  $(1 - \frac{\gamma\lambda}{\sigma^2})^2$  is an eigenvalue of  $(\mathbf{I} - \frac{\gamma}{\sigma^2} \mathbf{X}^\top \mathbf{X})^2$
- $\mathbf{X}^\top \mathbf{X}$  is positive semi-definite  $\Rightarrow \lambda \geq 0$
- Ensuring convergence: we want  $\lambda_{max} = \max(1 - \frac{\gamma\lambda}{\sigma^2})^2 < 1$

$$\Rightarrow \gamma < \frac{2\sigma^2}{\lambda_{max}(\mathbf{X}^\top \mathbf{X})}$$



# Choosing step-size for linear regression

To ensure convergence at **any** initialisation:  $\gamma < 2\sigma^2/\lambda_{\max}(\mathbf{X}^\top \mathbf{X})$

**Q:** Can we use larger step-sizes?

# Choosing step-size for linear regression

To ensure convergence at **any** initialisation:  $\gamma < 2\sigma^2/\lambda_{\max}(\mathbf{X}^\top \mathbf{X})$

**Q:** Can we use larger step-sizes?

**A:** Yes and No.

1. You choose a step-size  $\gamma \geq 2\sigma^2/\lambda_{\min}(\mathbf{X}^\top \mathbf{X}) \Rightarrow \text{diverge}$

# Choosing step-size for linear regression

To ensure convergence at **any** initialisation:  $\gamma < 2\sigma^2/\lambda_{\max}(\mathbf{X}^\top \mathbf{X})$

**Q:** Can we use larger step-sizes?

**A:** Yes and No.

1. You choose a step-size  $\gamma \geq 2\sigma^2/\lambda_{\min}(\mathbf{X}^\top \mathbf{X}) \Rightarrow$  diverge
2. You choose a step-size  $\gamma \in [\frac{2\sigma^2}{\lambda_{\max}(\mathbf{X}^\top \mathbf{X})}, \frac{2\sigma^2}{\lambda_{\min}(\mathbf{X}^\top \mathbf{X})}) \Rightarrow$  good luck
  - Convergence result may be sensitive to initialisation  $\theta_0$

# Choosing step-size for linear regression

To ensure convergence at **any** initialisation:  $\gamma < 2\sigma^2 / \lambda_{\max}(\mathbf{X}^\top \mathbf{X})$

If you want to test your luck: choose  $\gamma \in [\frac{2\sigma^2}{\lambda_{\max}(\mathbf{X}^\top \mathbf{X})}, \frac{2\sigma^2}{\lambda_{\min}(\mathbf{X}^\top \mathbf{X})})$

Is my choice of  $\gamma$  robust to initialisation of  $\boldsymbol{\theta}_0$ ?

# Choosing step-size for linear regression

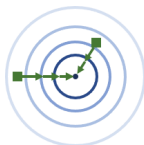
To ensure convergence at **any** initialisation:  $\gamma < 2\sigma^2/\lambda_{\max}(\mathbf{X}^\top \mathbf{X})$

If you want to test your luck: choose  $\gamma \in [\frac{2\sigma^2}{\lambda_{\max}(\mathbf{X}^\top \mathbf{X})}, \frac{2\sigma^2}{\lambda_{\min}(\mathbf{X}^\top \mathbf{X})})$

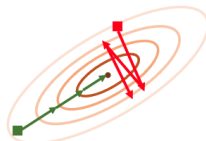
Is my choice of  $\gamma$  robust to initialisation of  $\theta_0$ ?

- Depending on the **condition number**:

$$\kappa(\mathbf{X}^\top \mathbf{X}) := \frac{\lambda_{\max}(\mathbf{X}^\top \mathbf{X})}{\lambda_{\min}(\mathbf{X}^\top \mathbf{X})}$$



well conditioned  
 $\kappa(\mathbf{X}^\top \mathbf{X}) \approx 1$



ill conditioned  
 $\kappa(\mathbf{X}^\top \mathbf{X}) \gg 1$

- Need careful choice of step-sizes if the loss is “very stretched”

# Choosing step-size for linear regression

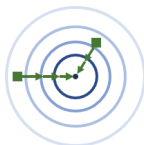
To ensure convergence at **any** initialisation:  $\gamma < 2\sigma^2 / \lambda_{\max}(\mathbf{X}^\top \mathbf{X})$

If you want to test your luck: choose  $\gamma \in [\frac{2\sigma^2}{\lambda_{\max}(\mathbf{X}^\top \mathbf{X})}, \frac{2\sigma^2}{\lambda_{\min}(\mathbf{X}^\top \mathbf{X})})$

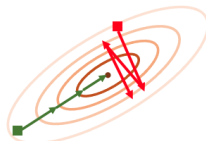
Is my choice of  $\gamma$  robust to initialisation of  $\theta_0$ ?

- Depending on the **condition number**:

$$\kappa(\mathbf{X}^\top \mathbf{X}) := \frac{\lambda_{\max}(\mathbf{X}^\top \mathbf{X})}{\lambda_{\min}(\mathbf{X}^\top \mathbf{X})}$$



well conditioned  
 $\kappa(\mathbf{X}^\top \mathbf{X}) \approx 1$

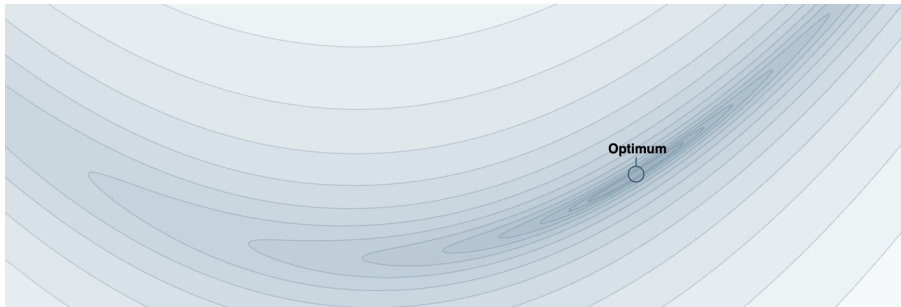


ill conditioned  
 $\kappa(\mathbf{X}^\top \mathbf{X}) \gg 1$

- Need careful choice of step-sizes if the loss is “very stretched”
- Note:  $\kappa(\mathbf{X}^\top \mathbf{X}) = \kappa(\mathbf{X})^2 = \frac{\sigma_{\max}(\mathbf{X})}{\sigma_{\min}(\mathbf{X})}$

# Choosing step-size: general case

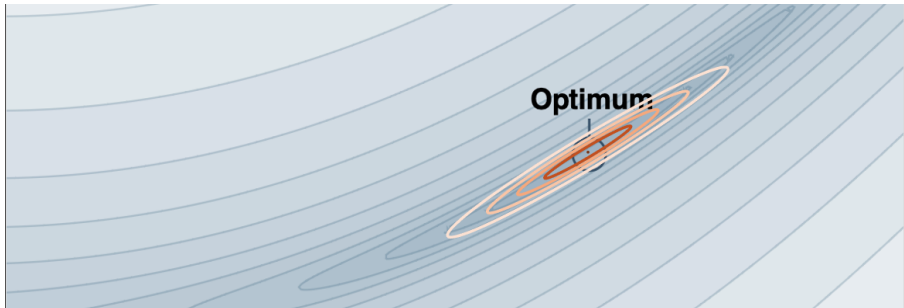
In general the loss function is non-quadratic nor convex:



<https://distill.pub/2017/momentum/>

# Choosing step-size: general case

In general the loss function is non-quadratic nor convex:



Local quadratic approximation when  $\theta_t \approx \theta^*$ :

- locally approximate  $L(\theta_t) \approx L(\theta^*) + \frac{1}{2}(\theta_t - \theta^*)^\top \nabla^2 L(\theta^*)(\theta_t - \theta^*)$   
(in linear regression  $\nabla^2 L(\theta) \propto \mathbf{X}^\top \mathbf{X}$ )
- $\kappa(\nabla^2 L)$  can tell whether the loss is “locally stretched”

<https://distill.pub/2017/momentum/>



# Choosing step-size: general case

Let's see what happens for different step-sizes.

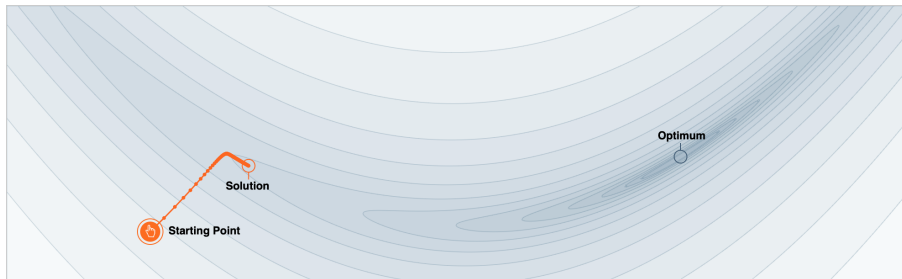


Image shows:

- Path of  $\theta_t$  from Gradient Descent
- Constant step size  $\gamma_t = \gamma$

<https://distill.pub/2017/momentum/>

# Choosing step-size: general case

Let's see what happens for different step-sizes.

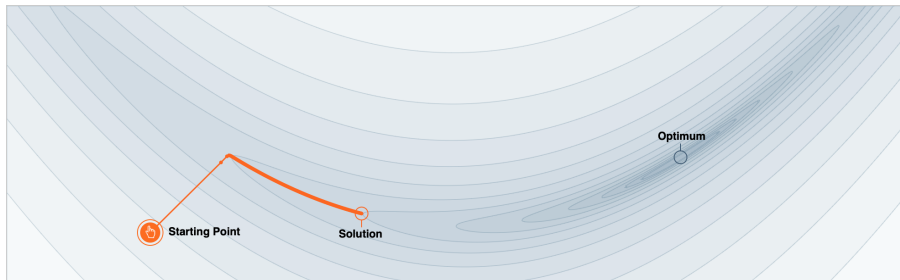


Image shows:

- Path of  $\theta_t$  from Gradient Descent
- Constant step size  $\gamma_t = \gamma$

<https://distill.pub/2017/momentum/>

# Choosing step-size: general case

Let's see what happens for different step-sizes.

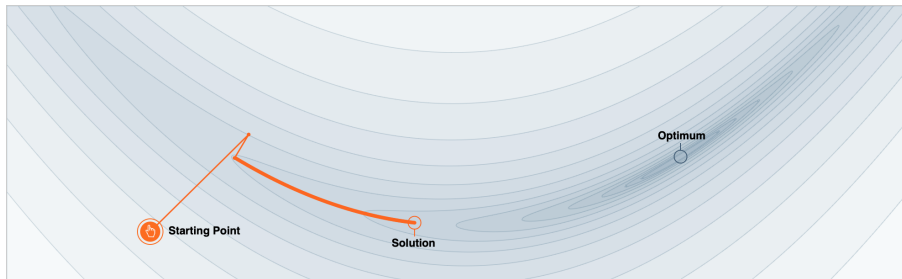


Image shows:

- Path of  $\theta_t$  from Gradient Descent
- Constant step size  $\gamma_t = \gamma$

<https://distill.pub/2017/momentum/>

# Choosing step-size: general case

Let's see what happens for different step-sizes.

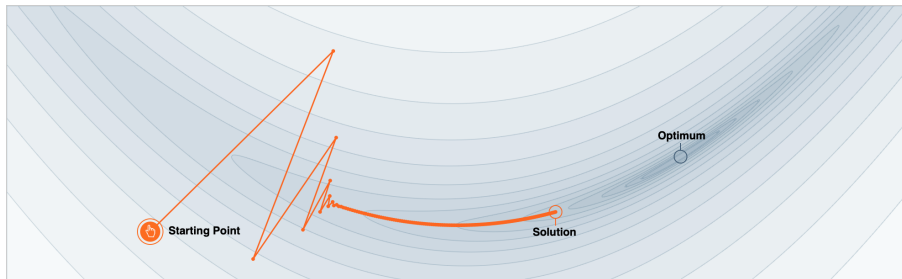


Image shows:

- Path of  $\theta_t$  from Gradient Descent
- Constant step size  $\gamma_t = \gamma$

<https://distill.pub/2017/momentum/>

# Choosing step-size: general case

Let's see what happens for different step-sizes.

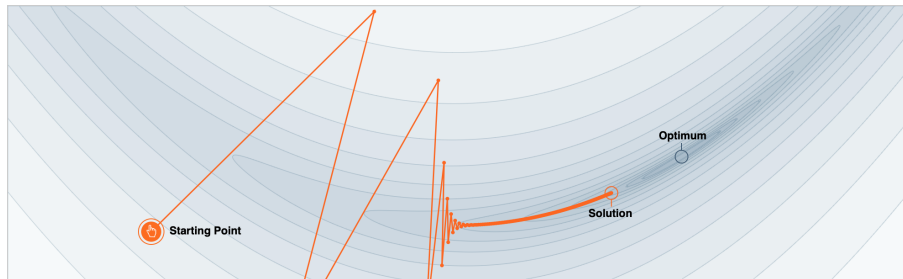


Image shows:

- Path of  $\theta_t$  from Gradient Descent
- Constant step size  $\gamma_t = \gamma$

<https://distill.pub/2017/momentum/>

# Choosing step-size: general case

Let's see what happens for different step-sizes.

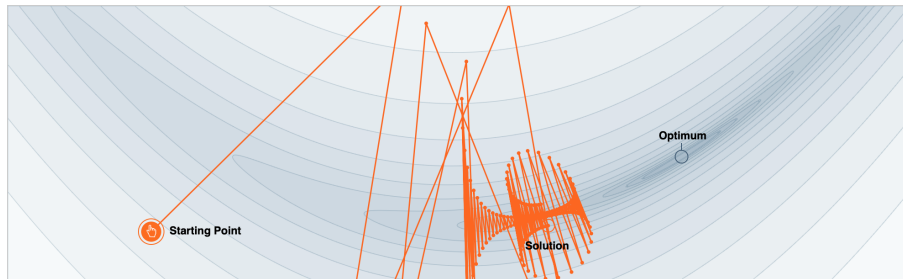


Image shows:

- Path of  $\theta_t$  from Gradient Descent
- Constant step size  $\gamma_t = \gamma$

<https://distill.pub/2017/momentum/>

# Choosing step-size: summary

Summary on choosing step size:

- too small: slow convergence
- too large: divergence
- just right: depends on problem (often: trial and error)

# Choosing step-size: summary

Summary on choosing step size:

- too small: slow convergence
- too large: divergence
- just right: depends on problem (often: trial and error)

Rule of thumb:

Start from a relatively large step size,  
decrease step size as getting closer to a (local) optimum.



# Exercises

Finish relevant exercises in the exercise sheet

- You should be able to analyse more advanced gradient-based optimisation methods for linear regression

Next lecture: multivariate probability

Pre-requisite knowledge: univariate probability