Gradient Descent

Yingzhen Li

Department of Computing
Imperial College London

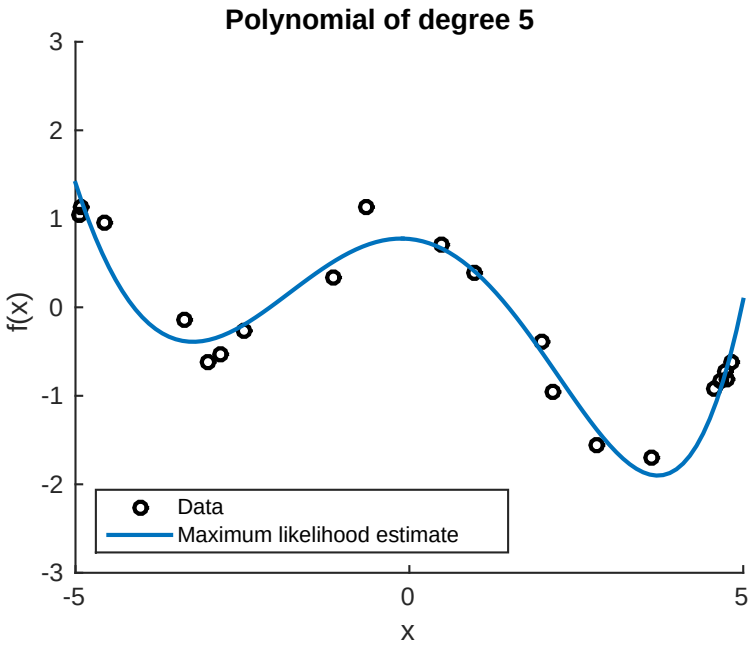
@liyzhen2
yingzhen.li@imperial.ac.uk

October 26, 2021

Reading for this week

Read MML book: Sections 4.1 - 4.4, 7.1, Chapter 9 up to 9.2.3
Do MML book exercises: Exercises 4.1 - 4.7
An extra exercise will be uploaded to course materials.

Optimisation Problems



Remember our problem:

- Find a curve that predicts well **even for unseen inputs**
- Start by minimising loss on **training points**:

$$L(\boldsymbol{\theta}) = \sum_n (f(\boldsymbol{x}_n; \boldsymbol{\theta}) - y_n)^2 \tag{1}$$

Formulating optimisation problems

- Define **objective function** $L : \mathbb{R}^D \to \mathbb{R}$
- Unconstrained minimisation, state which variable you want to optimise over

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \tag{2}$$

Gradient-based optimisation

How do we solve optimisation? We need an **algorithm**.

Gradient-based optimisation

How do we solve optimisation? We need an **algorithm**.

Gradient-based optimisation is a **class** of methods which all

Gradient-based optimisation

How do we solve optimisation? We need an **algorithm**.

Gradient-based optimisation is a **class** of methods which all

- Pick some starting point $\boldsymbol{\theta}_0$.

Gradient-based optimisation

How do we solve optimisation? We need an **algorithm**.

Gradient-based optimisation is a **class** of methods which all

- Pick some starting point $\boldsymbol{\theta}_0$.
- Iteratively update the parameters, resulting in a **sequence** of solutions $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T$.

Gradient-based optimisation

How do we solve optimisation? We need an **algorithm**.

Gradient-based optimisation is a **class** of methods which all

- Pick some starting point $\boldsymbol{\theta}_0$.
- Iteratively update the parameters, resulting in a **sequence** of solutions $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T$.
- Choose the update of the parameter by computing the **gradient** $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)$.

Gradient-based optimisation

How do we solve optimisation? We need an **algorithm**.

Gradient-based optimisation is a **class** of methods which all

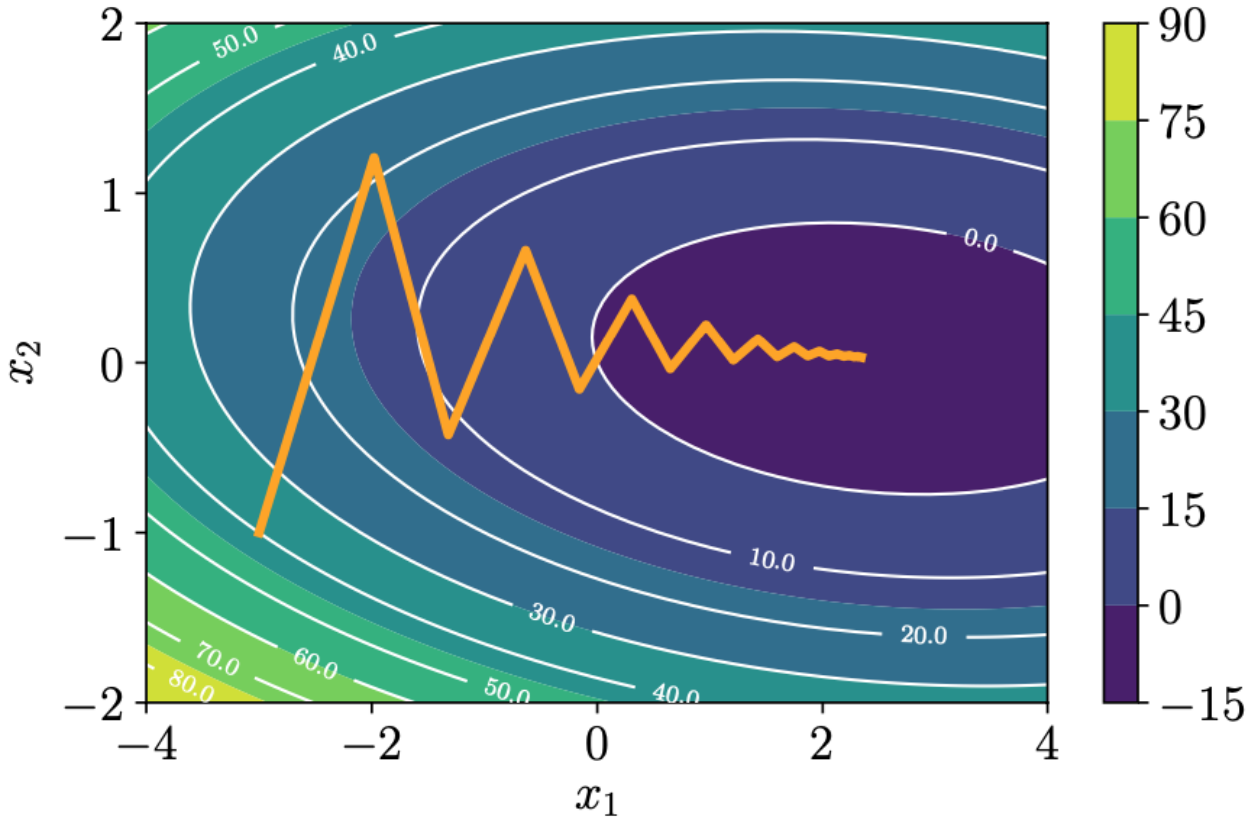
- Pick some starting point $\boldsymbol{\theta}_0$.
- Iteratively update the parameters, resulting in a **sequence** of solutions $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T$.
- Choose the update of the parameter by computing the **gradient** $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)$.
- Usually a **stopping criterion** (e.g. iteration budget, time budget, gradient size, ...)

Gradient descent

Algorithm: Gradient Descent

Define **starting point** $\boldsymbol{\theta}_0$, sequence of **step sizes** γ_t, set $t \leftarrow 0$.

1. Set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_t \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)$, $t \leftarrow t + 1$
2. Repeat 1 until stopping criterion.

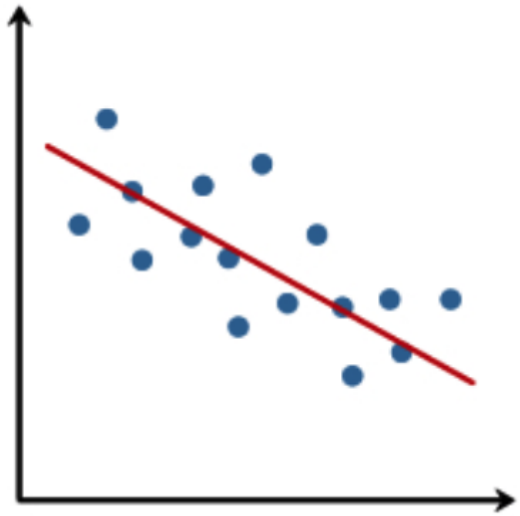


Gradient descent for linear regression

Fitting linear regression models:

- Dataset: $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$,
  $\mathbf{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times D}$,
  $\mathbf{y} = [y_1, ..., y_N]^\top \in \mathbb{R}^{N \times 1}$
- 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:

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

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

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

Gradient descent for linear regression

A typical linear regression model:

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

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

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

- 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

Optimal solution of $\boldsymbol{\theta}$:

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

- Gradient of the loss $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$:

Gradient descent for linear regression

Optimal solution of $\boldsymbol{\theta}$:

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

- Gradient of the loss $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$:

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

- Setting $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = 0$:

Gradient descent for linear regression

Optimal solution of $\boldsymbol{\theta}$:

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

- Gradient of the loss $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$:

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

- Setting $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = 0$:

$$\Rightarrow \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta}^* = \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{y}$$

$$\Rightarrow \boldsymbol{\theta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Gradient descent for linear regression

Gradient descent to find $\boldsymbol{\theta}^*$:

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

1. Define **starting point** $\boldsymbol{\theta}_0$, set $t \leftarrow 0$
2. Set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_t \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)$, $t \leftarrow t+1$

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

Gradient descent for linear regression

Gradient descent to find $\boldsymbol{\theta}^*$:

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

1. Define **starting point** $\boldsymbol{\theta}_0$, set $t \leftarrow 0$
2. Set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_t \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)$, $t \leftarrow t+1$

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

3. Repeat 1 until stopping criterion.

Gradient descent for linear regression

Gradient descent to find $\boldsymbol{\theta}^*$:

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

- GD returns the following iterative updates:

$$\boldsymbol{\theta}_{t+1} = (\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})\boldsymbol{\theta}_t + \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{y}$$

- Solving this iterative update returns:

Gradient descent for linear regression

Gradient descent to find $\boldsymbol{\theta}^*$:

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

- GD returns the following iterative updates:

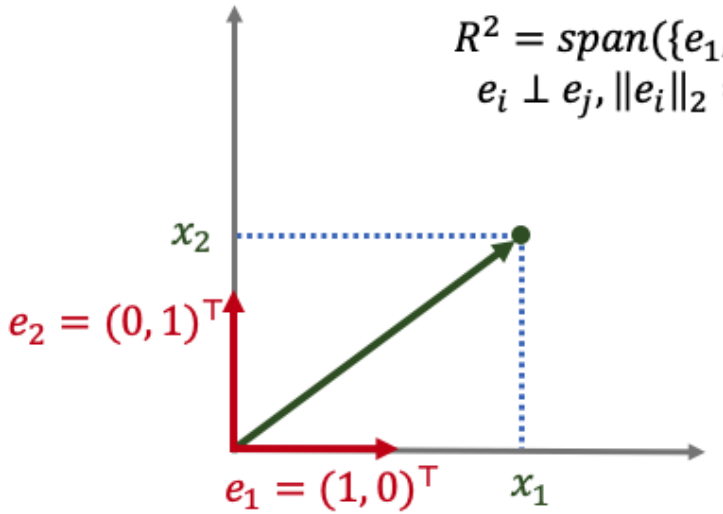
$$\boldsymbol{\theta}_{t+1} = (\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})\boldsymbol{\theta}_t + \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{y}$$

- Solving this iterative update returns:

$$\boldsymbol{\theta}_t = (\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*) + \boldsymbol{\theta}^*, \quad \boldsymbol{\theta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

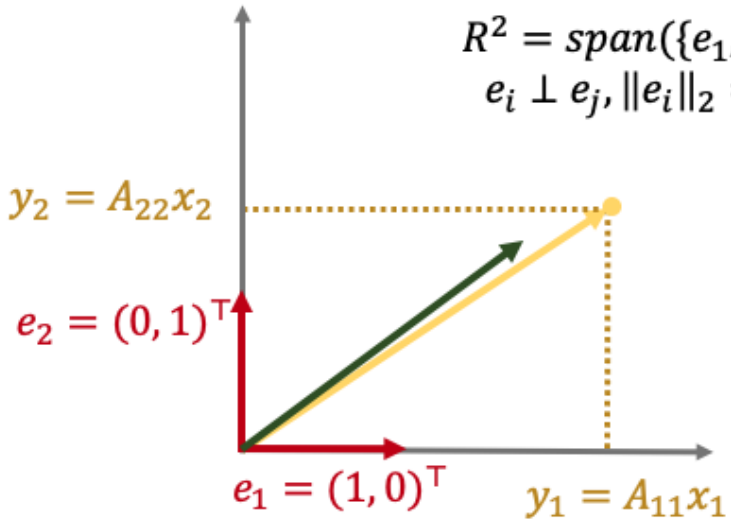
- GD converges ($\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*$) if $(\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*) \to \mathbf{0}$

Visualising some matrix multiplications



$$x = (x_1, x_2)^\top$$
$$x^\top e_i = x_i, i = 1, 2 \quad \Rightarrow \quad x = x_1 e_1 + x_2 e_2$$

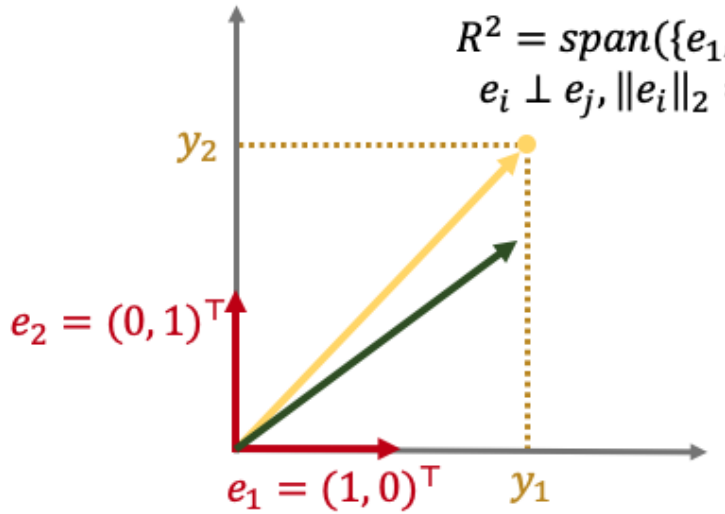
Visualising some matrix multiplications



$$x = (x_1, x_2)^\top$$
$$x^\top e_i = x_i, i = 1, 2$$
$$\Rightarrow \quad x = x_1 e_1 + x_2 e_2$$

$$y = Ax$$
$$A = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}$$
$$\Rightarrow \quad y = y_1 e_1 + y_2 e_2$$
$$y_1 = A_{11}x_1, y_2 = A_{22}x_2$$

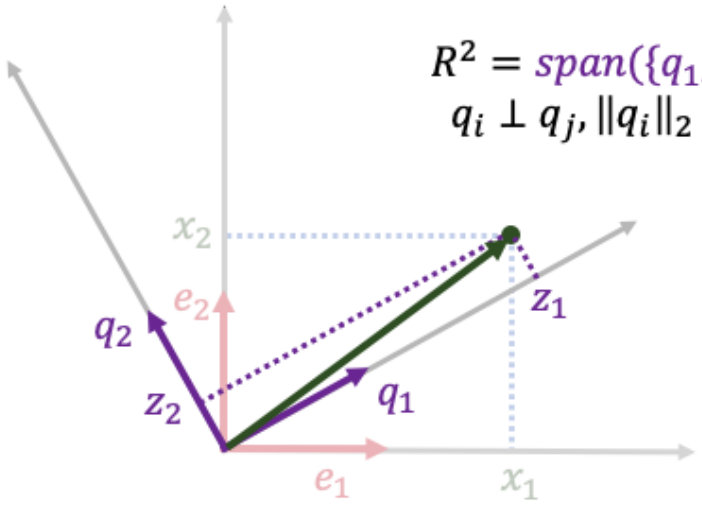
Visualising some matrix multiplications



$$x = (x_1, x_2)^\top \quad \Rightarrow \quad x = x_1 e_1 + x_2 e_2$$
$$x^\top e_i = x_i, i = 1, 2$$

$$y = Ax \quad \Rightarrow \quad y = (y_1, y_2)^\top$$
$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \qquad \begin{aligned} y_1 &= A_{11}x_1 + A_{12}x_2 \\ y_2 &= A_{21}x_1 + A_{22}x_2 \end{aligned}$$

Visualising some matrix multiplications

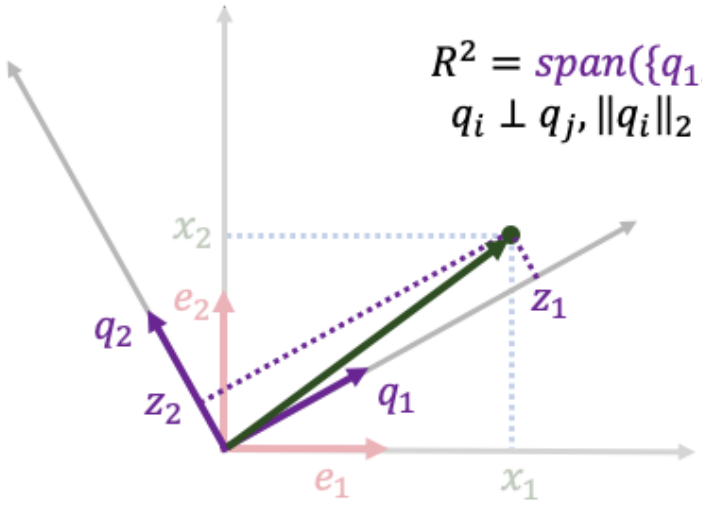


$$z = Q^{-1}x \quad \Rightarrow \quad \begin{matrix} z_1 = q_1^\top x \\ z_2 = q_2^\top x \end{matrix}$$

$$Q = \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix} = (q_1, q_2) \qquad Q^{-1} = \begin{pmatrix} q_{11} & q_{21} \\ q_{12} & q_{22} \end{pmatrix} = \begin{pmatrix} q_1^\top \\ q_2^\top \end{pmatrix}$$

Left multiply Q^{-1} on x:
Change basis from $\{e_1, e_2\}$ to $\{q_1, q_2\}$

Visualising some matrix multiplications

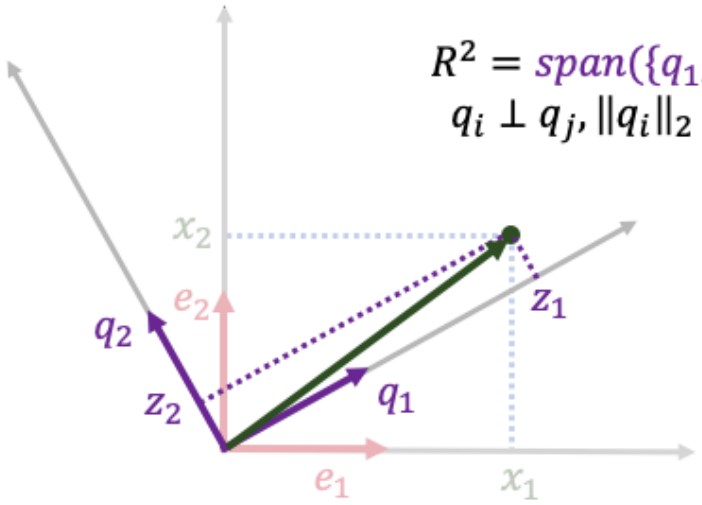


$Qz = x$ $\Rightarrow$ Find vector x such that $z = Q^{-1}x$

$$Q = \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix} = (q_1, q_2) \qquad Q^{-1} = \begin{pmatrix} q_{11} & q_{21} \\ q_{12} & q_{22} \end{pmatrix} = \begin{pmatrix} q_1^\top \\ q_2^\top \end{pmatrix}$$

Left multiply Q on z:
Change basis from $\{q_1, q_2\}$ to $\{e_1, e_2\}$

Visualising some matrix multiplications



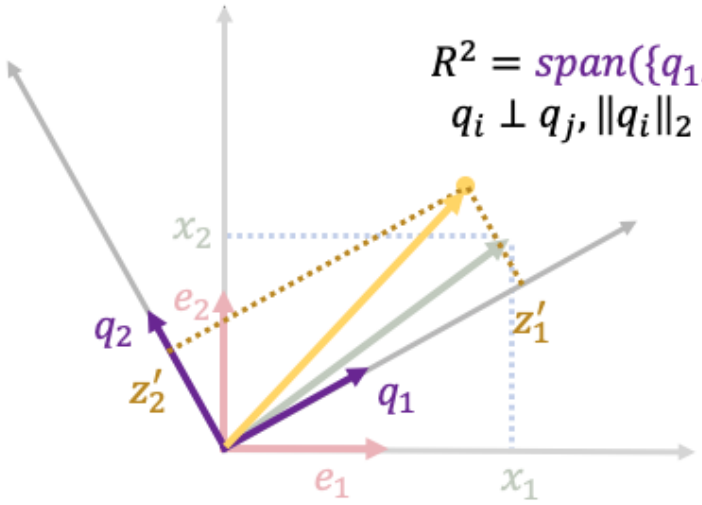
$Qz = x$ $\Rightarrow$ Find vector x such that $z = Q^\top x$

$$Q = \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix} = (q_1, q_2)$$

$$x = z_1 q_1 + z_2 q_2$$

(for orthonormal basis, $Q^{-1} = Q^\top$)

Visualising some matrix multiplications

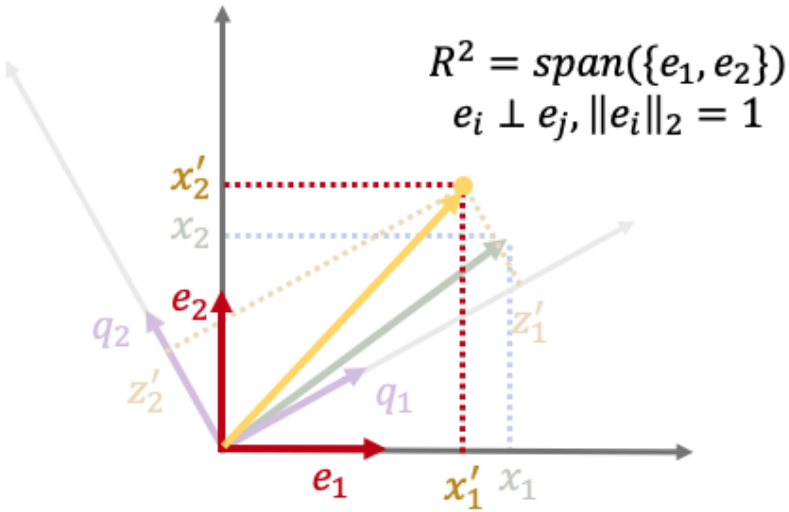


$$z' = \Lambda Q^{-1} x \quad \Rightarrow \quad \begin{aligned} z_1' &= \lambda_1 q_1^\top x \\ z_2' &= \lambda_2 q_2^\top x \end{aligned}$$

$$Q^{-1} = \begin{pmatrix} q_{11} & q_{21} \\ q_{12} & q_{22} \end{pmatrix} = \begin{pmatrix} q_1^\top \\ q_2^\top \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

Visualising some matrix multiplications



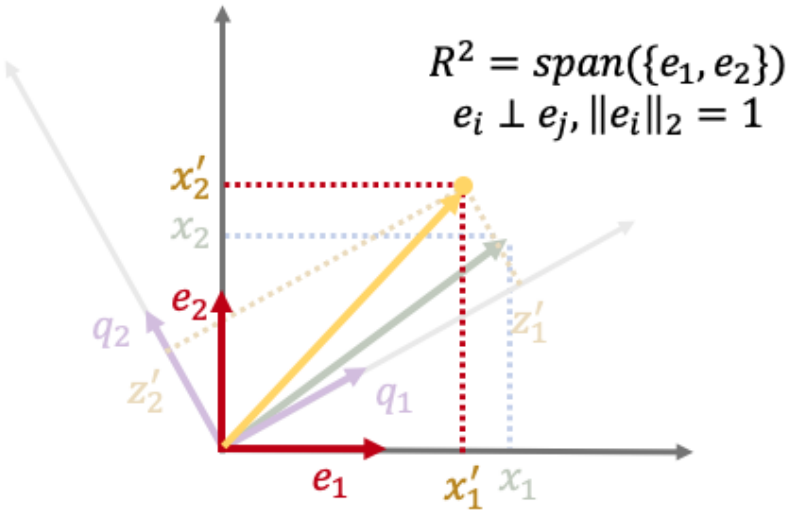
$$x' = Q\Lambda Q^{-1}x = Qz' \quad \Rightarrow \quad x' = z_1'q_1 + z_2'q_2$$

$$Q^{-1} = \begin{pmatrix} q_{11} & q_{21} \\ q_{12} & q_{22} \end{pmatrix} = \begin{pmatrix} q_1^\top \\ q_2^\top \end{pmatrix} \qquad \begin{aligned} z_1' &= \lambda_1 q_1^\top x \\ z_2' &= \lambda_2 q_2^\top x \end{aligned}$$

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

(for orthonormal basis, $Q^{-1} = Q^\top$)

Visualising some matrix multiplications



$$x' = Ax, A = Q\Lambda Q^{-1} \quad \Rightarrow \quad x' = z_1' q_1 + z_2' q_2$$

$$Q^{-1} = \begin{pmatrix} q_{11} & q_{21} \\ q_{12} & q_{22} \end{pmatrix} = \begin{pmatrix} q_1^\top \\ q_2^\top \end{pmatrix} \qquad \begin{aligned} z_1' &= \lambda_1 q_1^\top x \\ z_2' &= \lambda_2 q_2^\top x \end{aligned}$$

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

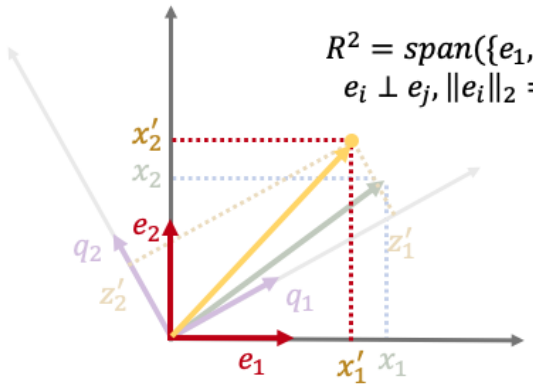
(for orthonormal basis, $Q^{-1} = Q^\top$)

Eigen decomposition

Consider $\boldsymbol{x}' = \mathbf{A}\boldsymbol{x}$ for $\mathbf{A} \in \mathbb{R}^{D\times D}$:

- Can we find $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$ (and how)?
- Key idea: find $\boldsymbol{x}$ such that $\boldsymbol{x}' = \mathbf{A}\boldsymbol{x}$ and $\boldsymbol{x}$ align on the same line.

  i.e. $\boldsymbol{x}' = \lambda\boldsymbol{x}$ for some $\lambda \neq 0$

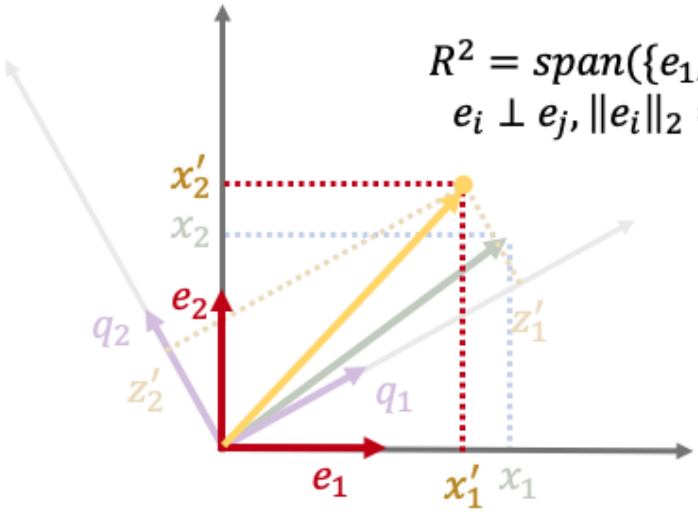


$$x' = Ax, A = Q\Lambda Q^{-1} \quad \Rightarrow \quad x' = z_1'q_1 + z_2'q_2$$

$$Q^{-1} = \begin{pmatrix} q_{11} & q_{21} \\ q_{12} & q_{22} \end{pmatrix} = \begin{pmatrix} q_1^\top \\ q_2^\top \end{pmatrix} \qquad z_1' = \lambda_1 q_1^\top x \quad z_2' = \lambda_2 q_2^\top x$$

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$ (for orthonormal basis, $Q^{-1} = Q^\top$)

Eigen decomposition



$$x' = Ax, A = Q\Lambda Q^{-1} \quad \Rightarrow \quad x' = z_1' q_1 + z_2' q_2$$

$$Q^{-1} = \begin{pmatrix} q_{11} & q_{21} \\ q_{12} & q_{22} \end{pmatrix} = \begin{pmatrix} q_1^\top \\ q_2^\top \end{pmatrix} \qquad \begin{aligned} z_1' &= \lambda_1 q_1^\top x \\ z_2' &= \lambda_2 q_2^\top x \end{aligned}$$

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$ (for orthonormal basis, $Q^{-1} = Q^\top$)

For $i = 1, \ldots, D$, if $\boldsymbol{x} = \boldsymbol{q}_i$, then $\boldsymbol{x}' = \mathbf{A}\boldsymbol{x} = \lambda_i \boldsymbol{q}_i$.

Eigen decomposition

Eigenvalue decomposition for a matrix $\mathbf{A} \in \mathbb{R}^{D \times D}$:
Find scalar λ and vector $\boldsymbol{q}$ such that

$$\mathbf{A}\boldsymbol{q} = \lambda \boldsymbol{q}$$

If solutions exist: $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$, $\Lambda = \text{diag}(\lambda_1, ..., \lambda_D)$

Eigen decomposition

Eigenvalue decomposition for a matrix $\mathbf{A} \in \mathbb{R}^{D \times D}$:
Find scalar λ and vector $\boldsymbol{q}$ such that

$$\mathbf{A}\boldsymbol{q} = \lambda \boldsymbol{q}$$

If solutions exist: $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$, $\Lambda = \text{diag}(\lambda_1, ..., \lambda_D)$

If $\mathbf{A}$ is symmetric:

- there are D pairs of solutions $(\lambda_d, \boldsymbol{q}_d)$ such that
  - Column vectors in $\mathbf{Q} = (\boldsymbol{q}_1, ..., \boldsymbol{q}_D)$ form an **orthonormal basis** (so $\mathbf{Q}^{-1} = \mathbf{Q}^\top$)
  - $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_D$
  - If $\mathbf{A}$ is positive semi-definite: $\lambda_D \geqslant 0$

Eigen decomposition

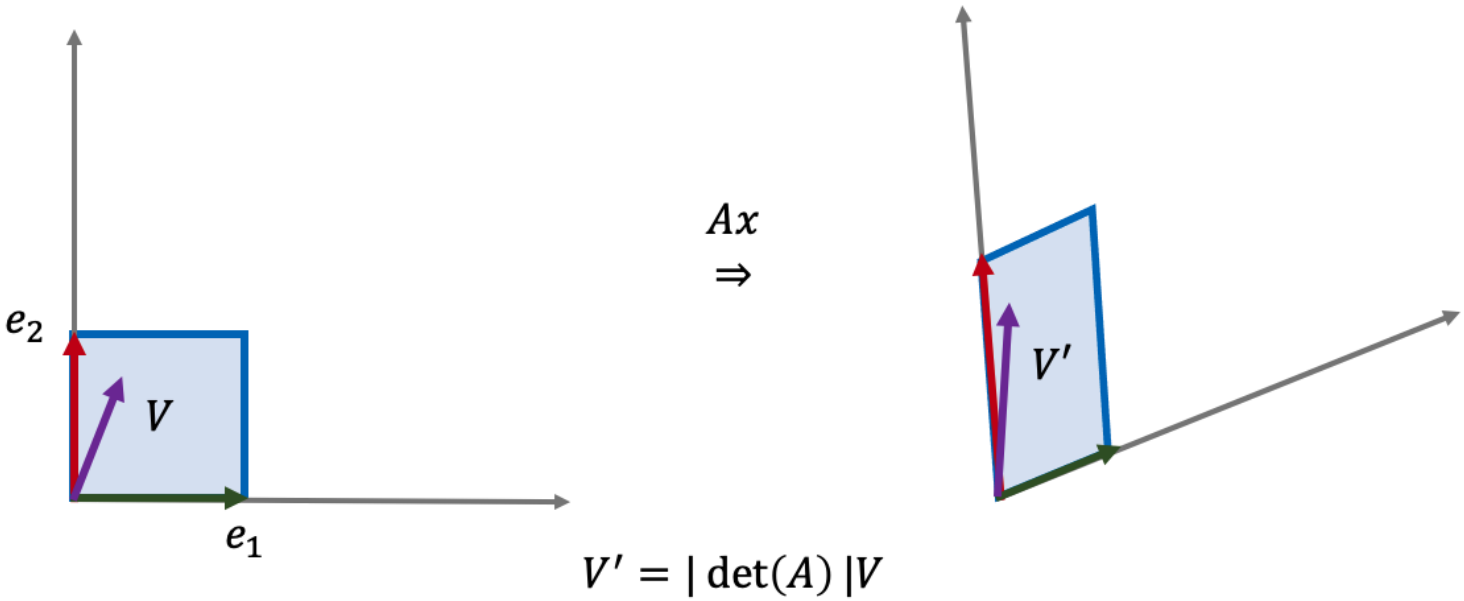
How to find $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$:

- $\mathbf{A}\boldsymbol{q} = \lambda\boldsymbol{q} \quad \Rightarrow \quad (\mathbf{A} - \lambda\mathbf{I})\boldsymbol{q} = \mathbf{0}$
- Assume $\boldsymbol{q} \neq \mathbf{0}$: find λ such that

$$det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

- Once λ is solved, plugging it back and solve for $\boldsymbol{q}$

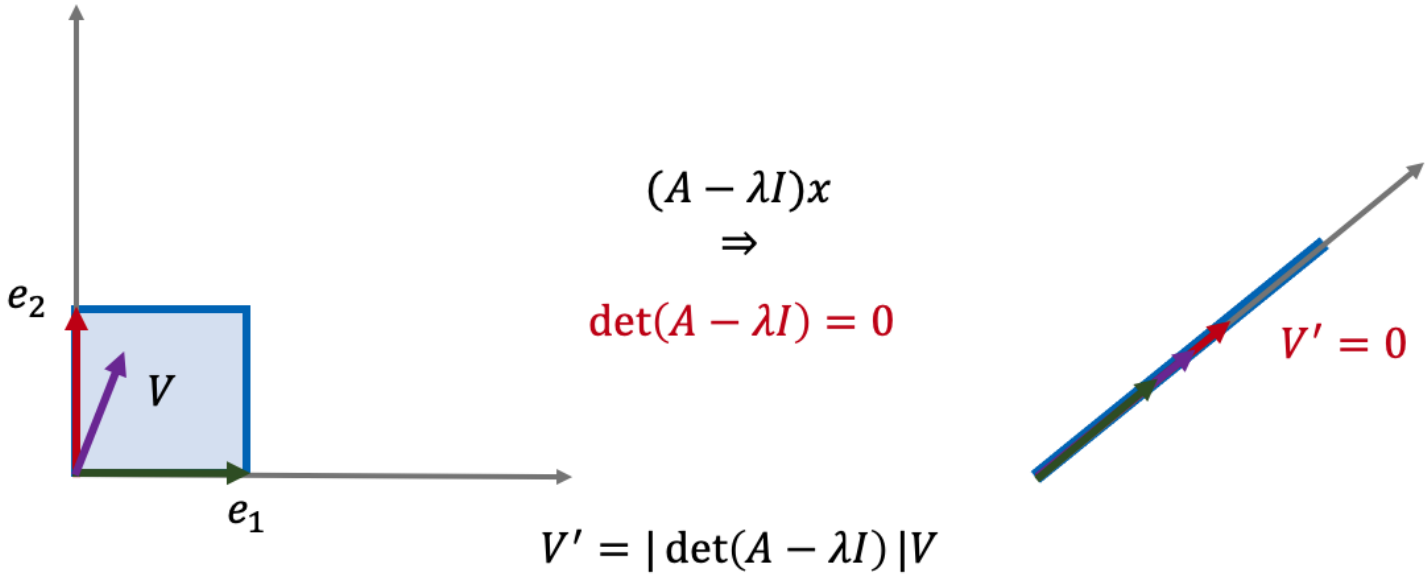
Eigen decomposition



Why solving $det(\mathbf{A} - \lambda\mathbf{I})$ w.r.t. λ:

- $|det(\mathbf{A})|$ tells how volume is scaled by linear transform

Eigen decomposition



Why solving $det(\mathbf{A} - \lambda\mathbf{I})$ w.r.t. λ:

- $|det(\mathbf{A})|$ tells how volume is scaled by linear transform
- If $det(\mathbf{A} - \lambda\mathbf{I}) = 0$: some subspace in $\mathbb{R}^D$ is squashed to $\{\mathbf{0}\}$
  $\Rightarrow$ there exists $\boldsymbol{q} \neq \mathbf{0}$ such that $(\mathbf{A} - \lambda\mathbf{I})\boldsymbol{q} = \mathbf{0}$

Eigen decomposition: Indications

Indications of $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$:

- If λ is an eigenvalue of $\mathbf{A}$, then λ^t is an eigenvalue of $\mathbf{A}^t$:

Eigen decomposition: Indications

Indications of $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$:

- If λ is an eigenvalue of $\mathbf{A}$, then λ^t is an eigenvalue of $\mathbf{A}^t$:

$$\mathbf{A}^t = \mathbf{Q}\Lambda\mathbf{Q}^{-1}\mathbf{Q}\Lambda\mathbf{Q}^{-1}\cdots\mathbf{Q}^{-1}\Lambda\mathbf{Q}^{-1} = \mathbf{Q}\Lambda^t\mathbf{Q}^{-1},$$

$$\Lambda^t = \text{diag}(\lambda_1^t, ..., \lambda_D^t)$$

Eigen decomposition: Indications

Indications of $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$:

- If λ is an eigenvalue of $\mathbf{A}$, then λ^t is an eigenvalue of $\mathbf{A}^t$:

$$\mathbf{A}^t = \mathbf{Q}\Lambda\mathbf{Q}^{-1}\mathbf{Q}\Lambda\mathbf{Q}^{-1}\cdots\mathbf{Q}^{-1}\Lambda\mathbf{Q}^{-1} = \mathbf{Q}\Lambda^t\mathbf{Q}^{-1},$$

$$\Lambda^t = \text{diag}(\lambda_1^t, ..., \lambda_D^t)$$

- If λ is an eigenvalue of $\mathbf{A}$, then $\lambda + \alpha$ is an eigenvalue of $\mathbf{A} + \alpha\mathbf{I}$:

Eigen decomposition: Indications

Indications of $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$:

- If λ is an eigenvalue of $\mathbf{A}$, then λ^t is an eigenvalue of $\mathbf{A}^t$:

$$\mathbf{A}^t = \mathbf{Q}\Lambda\mathbf{Q}^{-1}\mathbf{Q}\Lambda\mathbf{Q}^{-1}\cdots\mathbf{Q}^{-1}\Lambda\mathbf{Q}^{-1} = \mathbf{Q}\Lambda^t\mathbf{Q}^{-1},$$

$$\Lambda^t = \mathrm{diag}(\lambda_1^t, ..., \lambda_D^t)$$

- If λ is an eigenvalue of $\mathbf{A}$, then $\lambda + \alpha$ is an eigenvalue of $\mathbf{A} + \alpha\mathbf{I}$:

$$\mathbf{A} + \alpha\mathbf{I} = \mathbf{Q}\Lambda\mathbf{Q}^{-1} + \mathbf{Q}\alpha\mathbf{I}\mathbf{Q}^{-1} = \mathbf{Q}(\Lambda + \alpha\mathbf{I})\mathbf{Q}^{-1},$$

$$\Lambda + \alpha\mathbf{I} = \mathrm{diag}(\lambda_1 + \alpha, ..., \lambda_D + \alpha)$$

Eigen decomposition: Indications

Indications of $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$:

- If λ is an eigenvalue of $\mathbf{A}$, then λ^t is an eigenvalue of $\mathbf{A}^t$:

$$\mathbf{A}^t = \mathbf{Q}\Lambda\mathbf{Q}^{-1}\mathbf{Q}\Lambda\mathbf{Q}^{-1}\cdots\mathbf{Q}^{-1}\Lambda\mathbf{Q}^{-1} = \mathbf{Q}\Lambda^t\mathbf{Q}^{-1},$$

$$\Lambda^t = \mathrm{diag}(\lambda_1^t, ..., \lambda_D^t)$$

- If λ is an eigenvalue of $\mathbf{A}$, then $\lambda + \alpha$ is an eigenvalue of $\mathbf{A} + \alpha\mathbf{I}$:

$$\mathbf{A} + \alpha\mathbf{I} = \mathbf{Q}\Lambda\mathbf{Q}^{-1} + \mathbf{Q}\alpha\mathbf{I}\mathbf{Q}^{-1} = \mathbf{Q}(\Lambda + \alpha\mathbf{I})\mathbf{Q}^{-1},$$

$$\Lambda + \alpha\mathbf{I} = \mathrm{diag}(\lambda_1 + \alpha, ..., \lambda_D + \alpha)$$

- Combine: If λ is an eigenvalue of $\mathbf{A}$,
  then $(\lambda + \alpha)^t$ is an eigenvalue of $(\mathbf{A} + \alpha\mathbf{I})^t$

Eigen decomposition: Indications

Indications of $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$: Assume $\mathbf{A}$ is symmetric

- Consider the following **Rayleigh quotient**

$$R(\mathbf{A}, \boldsymbol{x}) = \frac{\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x}}{||\boldsymbol{x}||_2^2}, \quad ||\boldsymbol{x}||_2^2 = \boldsymbol{x}^\top \boldsymbol{x}$$

- We can show that

$$\lambda_{min}(\mathbf{A}) \leqslant R(\mathbf{A}, \boldsymbol{x}) \leqslant \lambda_{max}(\mathbf{A})$$

$$\Rightarrow \quad \lambda_{min}(\mathbf{A})||\boldsymbol{x}||_2^2 \leqslant \boldsymbol{x}^\top \mathbf{A}\boldsymbol{x} \leqslant \lambda_{max}(\mathbf{A})||\boldsymbol{x}||_2^2$$

(break)

Eigen decomposition: Indications

Indications of $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$: Assume $\mathbf{A}$ is symmetric

- Consider the following **Rayleigh quotient**

$$R(\mathbf{A}, \boldsymbol{x}) = \frac{\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x}}{||\boldsymbol{x}||_2^2}, \quad ||\boldsymbol{x}||_2^2 = \boldsymbol{x}^\top \boldsymbol{x}$$

- We can show: $\lambda_{min}(\mathbf{A}) \leqslant R(\mathbf{A}, \boldsymbol{x}) \leqslant \lambda_{max}(\mathbf{A})$

  $\Rightarrow \quad \lambda_{min}(\mathbf{A})||\boldsymbol{x}||_2^2 \leqslant \boldsymbol{x}^\top \mathbf{A}\boldsymbol{x} \leqslant \lambda_{max}(\mathbf{A})||\boldsymbol{x}||_2^2$

Convergence of GD for linear regression

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

$$\boldsymbol{\theta}_t = (\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*) + \boldsymbol{\theta}^*, \quad \boldsymbol{\theta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

- The ℓ_2 distance between $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}^*$:

$$\begin{aligned}||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 &= ||(\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)||_2^2 \\ &= |(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)^\top(\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^{2t}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)|\end{aligned}$$

Convergence of GD for linear regression

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

$$\boldsymbol{\theta}_t = (\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*) + \boldsymbol{\theta}^*, \quad \boldsymbol{\theta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

- The ℓ_2 distance between $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}^*$:

$$\begin{aligned}||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 &= ||(\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)||_2^2 \\ &= |(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)^\top(\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^{2t}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)|\end{aligned}$$

- Bounded distance by setting $\boldsymbol{x} = \boldsymbol{\theta}_0 - \boldsymbol{\theta}^*$, $\mathbf{A} = (\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^{2t}$:

$$||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 \geqslant \lambda_{min}((\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^{2t})||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2$$

$$||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 \leqslant \lambda_{max}((\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^{2t})||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2$$

Convergence of GD for linear regression

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

$$\boldsymbol{\theta}_t = (\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*) + \boldsymbol{\theta}^*, \quad \boldsymbol{\theta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

- The ℓ_2 distance between $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}^*$:

$$\begin{aligned}||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 &= ||(\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)||_2^2 \\ &= |(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)^\top(\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^{2t}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)|\end{aligned}$$

- Bounded distance by setting $\boldsymbol{x} = \boldsymbol{\theta}_0 - \boldsymbol{\theta}^*$, $\mathbf{A} = (\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^{2t}$:

$$||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 \geqslant \lambda_{min}((\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^2)^t||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2$$

$$||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 \leqslant \lambda_{max}((\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^2)^t||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2$$

Convergence of GD for linear regression

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

$$\lambda_{min}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2 \leqslant ||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 \leqslant \lambda_{max}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2$$

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

Convergence of GD for linear regression

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

$$\lambda_{min}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2 \leqslant ||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 \leqslant \lambda_{max}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2$$

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

Convergence properties in difference cases:

1. $\lambda_{max} < 1$: always converge
2. $\lambda_{min} \geqslant 1$: always diverge
3. $\lambda_{min} < 1$ but $\lambda_{max} \geqslant 1$: convergence depending on $\boldsymbol{\theta}_0$

Convergence of GD for linear regression

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

$$\lambda_{min}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2 \leqslant ||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 \leqslant \lambda_{max}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2$$

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

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

Convergence of GD for linear regression

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

$$\lambda_{min}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2 \leqslant ||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 \leqslant \lambda_{max}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2$$

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

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

- If λ is an eigenvalue of $\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X}$,
  then λ^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 $\boldsymbol{\theta}^*$:

$$\lambda_{min}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2 \leqslant ||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 \leqslant \lambda_{max}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2$$

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

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

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

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

Convergence of GD for linear regression

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

$$\lambda_{min}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2 \leqslant ||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 \leqslant \lambda_{max}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2$$

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

- If λ 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 $\boldsymbol{\theta}^*$:

$$\lambda_{min}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2 \leqslant ||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 \leqslant \lambda_{max}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2$$

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

- If λ 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 $\Rightarrow \lambda \geqslant 0$

Convergence of GD for linear regression

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

$$\lambda_{min}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2 \leqslant ||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 \leqslant \lambda_{max}^t ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||_2^2$$

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

- If λ 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 $\Rightarrow \lambda \geqslant 0$
- 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 if you are lucky, and no if you are unlucky.

1. You choose a step-size $\gamma > 2\sigma^2/\lambda_{max}(\mathbf{X}^\top\mathbf{X})$

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?

A: Yes if you are lucky, and no if you are unlucky.

1. You choose a step-size $\gamma > 2\sigma^2/\lambda_{max}(\mathbf{X}^\top\mathbf{X})$
2. You choose an initialisation $\boldsymbol{\theta}_0$
   - Consider eigen decomposition of $\mathbf{X}^\top\mathbf{X}$ as $\{(\lambda_d, \boldsymbol{q}_d)\}$ with $\lambda_1 \geqslant ... \geqslant \lambda_D \geqslant 0$ (therefore $\lambda_{max}(\mathbf{X}^\top\mathbf{X}) = \lambda_1$)

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 if you are lucky, and no if you are unlucky.

1. You choose a step-size $\gamma > 2\sigma^2/\lambda_{max}(\mathbf{X}^\top\mathbf{X})$
2. You choose an initialisation $\boldsymbol{\theta}_0$
   - Consider eigen decomposition of $\mathbf{X}^\top\mathbf{X}$ as $\{(\lambda_d, \boldsymbol{q}_d)\}$ with $\lambda_1 \geqslant ... \geqslant \lambda_D \geqslant 0$ (therefore $\lambda_{max}(\mathbf{X}^\top\mathbf{X}) = \lambda_1$)
3. **Lucky case**:
   - now assume it happens to be that $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^* + \alpha\boldsymbol{q}_d, d > 1$

$$||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 = \alpha^2|\boldsymbol{q}_d^\top(\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^{2t}\boldsymbol{q}_d| = \alpha^2(1 - \frac{\gamma\lambda_d}{\sigma^2})^{2t}$$

   - $\gamma < 2\sigma^2/\lambda_d \quad \Rightarrow \quad ||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 \to 0$

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 if you are lucky, and no if you are unlucky.

1. You choose a step-size $\gamma > 2\sigma^2/\lambda_{max}(\mathbf{X}^\top\mathbf{X})$
2. You choose an initialisation $\boldsymbol{\theta}_0$
   - Consider eigen decomposition of $\mathbf{X}^\top\mathbf{X}$ as $\{(\lambda_d, \boldsymbol{q}_d)\}$ with $\lambda_1 \geqslant ... \geqslant \lambda_D \geqslant 0$ (therefore $\lambda_{max}(\mathbf{X}^\top\mathbf{X}) = \lambda_1$)
3. **Lucky case**:
   - now assume it happens to be that $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^* + \alpha\boldsymbol{q}_d, d > 1$

$$||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 = \alpha^2|\boldsymbol{q}_d^\top(\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^{2t}\boldsymbol{q}_d| = \alpha^2(1 - \frac{\gamma\lambda_d}{\sigma^2})^{2t}$$

   - $\gamma < 2\sigma^2/\lambda_d \quad \Rightarrow \quad ||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 \to 0$
4. **Unlucky case**: $\gamma \geqslant 2\sigma^2/\lambda_d \quad \Rightarrow \quad$ divergence

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 if you are lucky, and no if you are unlucky.

1. You choose a step-size $\gamma > 2\sigma^2/\lambda_{max}(\mathbf{X}^\top\mathbf{X})$
2. You choose an initialisation $\boldsymbol{\theta}_0$
   - Consider eigen decomposition of $\mathbf{X}^\top\mathbf{X}$ as $\{(\lambda_d, \boldsymbol{q}_d)\}$ with $\lambda_1 \geqslant ... \geqslant \lambda_D \geqslant 0$ (therefore $\lambda_{max}(\mathbf{X}^\top\mathbf{X}) = \lambda_1$)
3. **Lucky case**:
   - now assume it happens to be that $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^* + \alpha\boldsymbol{q}_d, d > 1$

$$||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 = \alpha^2|\boldsymbol{q}_d^\top(\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^{2t}\boldsymbol{q}_d| = \alpha^2(1 - \frac{\gamma\lambda_d}{\sigma^2})^{2t}$$

   - $\gamma < 2\sigma^2/\lambda_d \quad \Rightarrow \quad ||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 \to 0$
4. **Unlucky case**: $\gamma \geqslant 2\sigma^2/\lambda_d \quad \Rightarrow \quad$ divergence

Constant step-size GD diverges if $\gamma \geqslant 2\sigma^2/\lambda_{min}(\mathbf{X}^\top\mathbf{X})$

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?

A: Yes if you are lucky, and no if you are unlucky.

Caveat: you are unlikely to be that lucky...

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 if you are lucky, and no if you are unlucky.
Caveat: you are unlikely to be that lucky...

1. You choose a step-size $\gamma > 2\sigma^2/\lambda_{max}(\mathbf{X}^\top\mathbf{X})$
2. You **randomly initialise** the parameter $\boldsymbol{\theta}_0$
   - write $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^* + \sum_{d=1}^{D} \alpha_d \boldsymbol{q}_d$

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 if you are lucky, and no if you are unlucky.
Caveat: you are unlikely to be that lucky...

1. You choose a step-size $\gamma > 2\sigma^2/\lambda_{max}(\mathbf{X}^\top\mathbf{X})$
2. You **randomly initialise** the parameter $\boldsymbol{\theta}_0$
   - write $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^* + \sum_{d=1}^{D} \alpha_d \boldsymbol{q}_d$
3. Evolution of the iterative update: $\boldsymbol{\theta}_t - \boldsymbol{\theta}^* = \sum_{d=1}^{D} \alpha_d (1 - \frac{\gamma\lambda_d}{\sigma^2})^t \boldsymbol{q}_d$

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 if you are lucky, and no if you are unlucky.
Caveat: you are unlikely to be that lucky...

1. You choose a step-size $\gamma > 2\sigma^2/\lambda_{max}(\mathbf{X}^\top\mathbf{X})$
2. You **randomly initialise** the parameter $\boldsymbol{\theta}_0$
   - write $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^* + \sum_{d=1}^{D} \alpha_d \boldsymbol{q}_d$
3. Evolution of the iterative update: $\boldsymbol{\theta}_t - \boldsymbol{\theta}^* = \sum_{d=1}^{D} \alpha_d (1 - \frac{\gamma\lambda_d}{\sigma^2})^t \boldsymbol{q}_d$
   - For direction $\boldsymbol{q}_d$ with $\gamma < 2\sigma^2/\lambda_d$: $\quad (1 - \frac{\gamma\lambda_d}{\sigma^2})^t \boldsymbol{q}_d \to \mathbf{0}$

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 if you are lucky, and no if you are unlucky.
Caveat: you are unlikely to be that lucky...

1. You choose a step-size $\gamma > 2\sigma^2/\lambda_{max}(\mathbf{X}^\top\mathbf{X})$
2. You **randomly initialise** the parameter $\boldsymbol{\theta}_0$
   - write $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^* + \sum_{d=1}^{D} \alpha_d \boldsymbol{q}_d$
3. Evolution of the iterative update: $\boldsymbol{\theta}_t - \boldsymbol{\theta}^* = \sum_{d=1}^{D} \alpha_d (1 - \frac{\gamma\lambda_d}{\sigma^2})^t \boldsymbol{q}_d$
   - For direction $\boldsymbol{q}_d$ with $\gamma < 2\sigma^2/\lambda_d$: $(1 - \frac{\gamma\lambda_d}{\sigma^2})^t \boldsymbol{q}_d \to \mathbf{0}$
   - For other directions $\gamma \geqslant 2\sigma^2/\lambda_d$: $(1 - \frac{\gamma\lambda_d}{\sigma^2})^t \boldsymbol{q}_d$ diverges
   - GD diverges unless $\alpha_d = 0$ for those d with $\gamma \geqslant 2\sigma^2/\lambda_d$.

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 if you are lucky, and no if you are unlucky.
Caveat: you are unlikely to be that lucky...

1. You choose a step-size $\gamma > 2\sigma^2/\lambda_{max}(\mathbf{X}^\top\mathbf{X})$
2. You **randomly initialise** the parameter $\boldsymbol{\theta}_0$
   - write $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^* + \sum_{d=1}^{D} \alpha_d \boldsymbol{q}_d$
3. Evolution of the iterative update: $\boldsymbol{\theta}_t - \boldsymbol{\theta}^* = \sum_{d=1}^{D} \alpha_d (1 - \frac{\gamma\lambda_d}{\sigma^2})^t \boldsymbol{q}_d$
   - For direction $\boldsymbol{q}_d$ with $\gamma < 2\sigma^2/\lambda_d$: $(1 - \frac{\gamma\lambda_d}{\sigma^2})^t \boldsymbol{q}_d \to \mathbf{0}$
   - For other directions $\gamma \geqslant 2\sigma^2/\lambda_d$: $(1 - \frac{\gamma\lambda_d}{\sigma^2})^t \boldsymbol{q}_d$ diverges
   - GD diverges unless $\alpha_d = 0$ for those d with $\gamma \geqslant 2\sigma^2/\lambda_d$.
4. For $d > 1$, $\text{span}\{\boldsymbol{q}_d, ..., \boldsymbol{q}_D\}$ has measure 0 in $\mathbb{R}^D$
   - unlikely to make $\alpha_d = 0$ for some d with random initialisation

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 γ robust to initialisation of $\boldsymbol{\theta}_0$?

Choosing step-size for linear regression

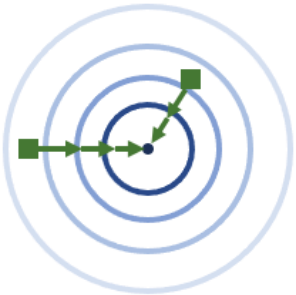
To ensure convergence at **any** initialisation: $\gamma < 2\sigma^2/\lambda_{max}(\mathbf{X}^\top\mathbf{X})$

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

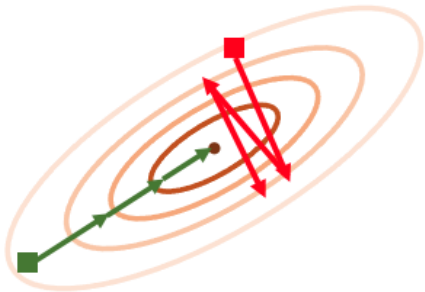
Is my choice of γ robust to initialisation of $\boldsymbol{\theta}_0$?

- Depending on the **condition number**:

$$\kappa(\mathbf{X}^\top\mathbf{X}) := \frac{\lambda_{max}(\mathbf{X}^\top\mathbf{X})}{\lambda_{min}(\mathbf{X}^\top\mathbf{X})} = \sqrt{\frac{\max_q R(\mathbf{X}^\top\mathbf{X}, q)}{\min_q R(\mathbf{X}^\top\mathbf{X}, q)}}$$



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



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

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

Choosing step-size for linear regression

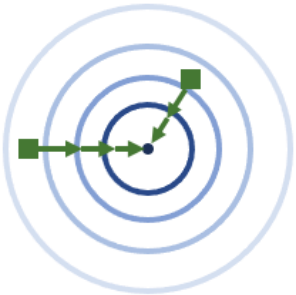
To ensure convergence at **any** initialisation: $\gamma < 2\sigma^2/\lambda_{max}(\mathbf{X}^\top\mathbf{X})$

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

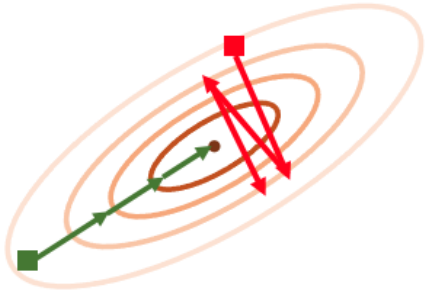
Is my choice of γ robust to initialisation of $\boldsymbol{\theta}_0$?

- Depending on the **condition number**:

$$\kappa(\mathbf{X}^\top\mathbf{X}) := \frac{\lambda_{max}(\mathbf{X}^\top\mathbf{X})}{\lambda_{min}(\mathbf{X}^\top\mathbf{X})} = \sqrt{\frac{\max_q R(\mathbf{X}^\top\mathbf{X}, q)}{\min_q R(\mathbf{X}^\top\mathbf{X}, q)}}$$



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

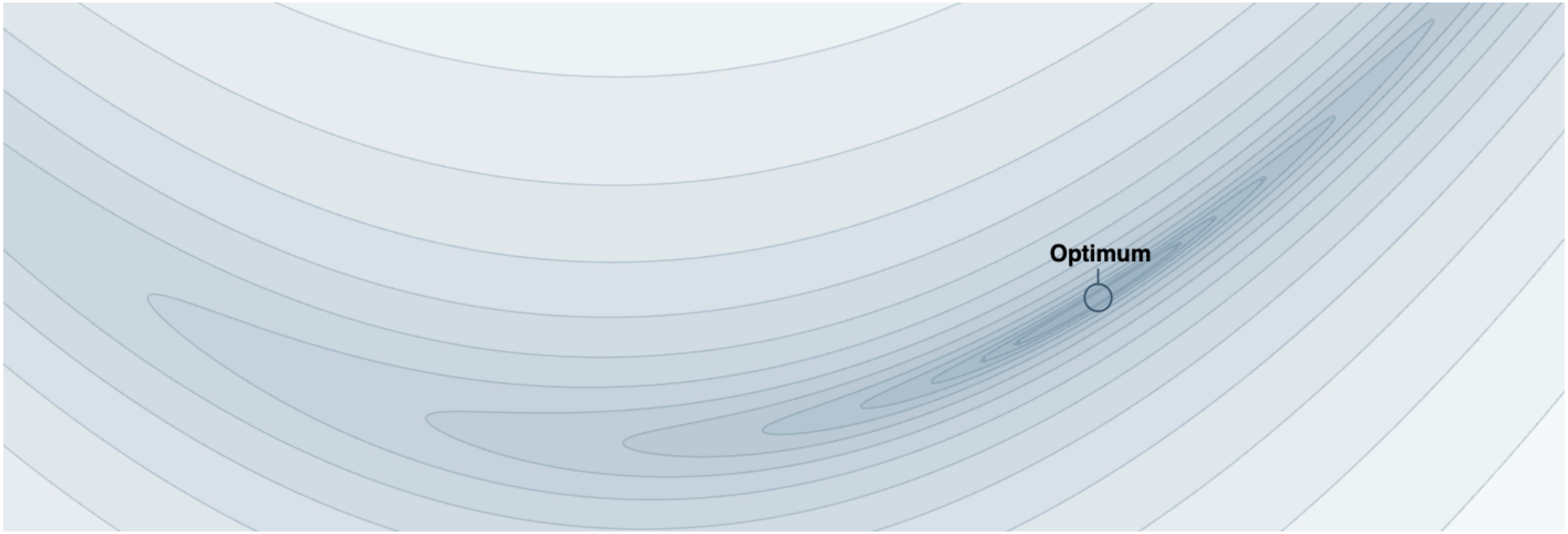


ill conditioned
$\kappa(\mathrm{X}^\top\mathrm{X}) \gg 1$

- 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

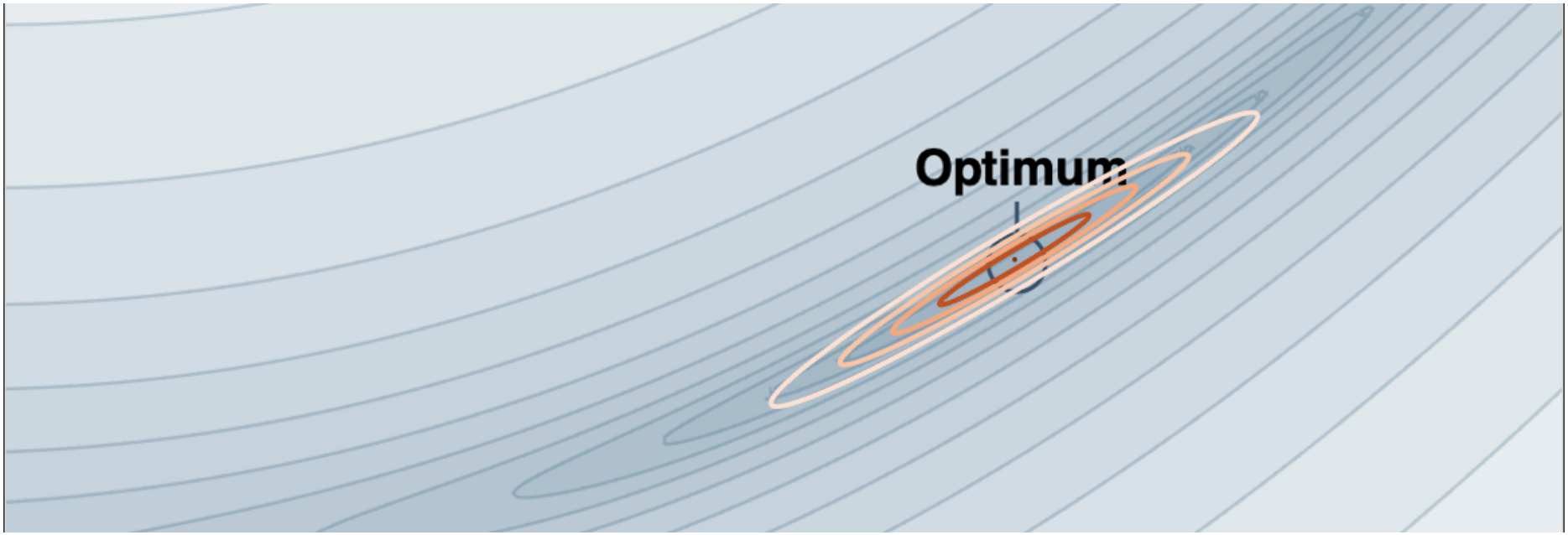
In general the loss function is non-quadratic nor convex:



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

Choosing step-size: general case

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



Local quadratic approximation when $\boldsymbol{\theta}_t \approx \boldsymbol{\theta}^*$:

- locally approximate $L(\boldsymbol{\theta}_t) \approx L(\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^\top \nabla^2 L(\boldsymbol{\theta}^*)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)$ (in linear regression $\nabla^2 L(\boldsymbol{\theta}) \propto \mathbf{X}^\top \mathbf{X}$)
- $\kappa(\nabla^2 L)$ 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.

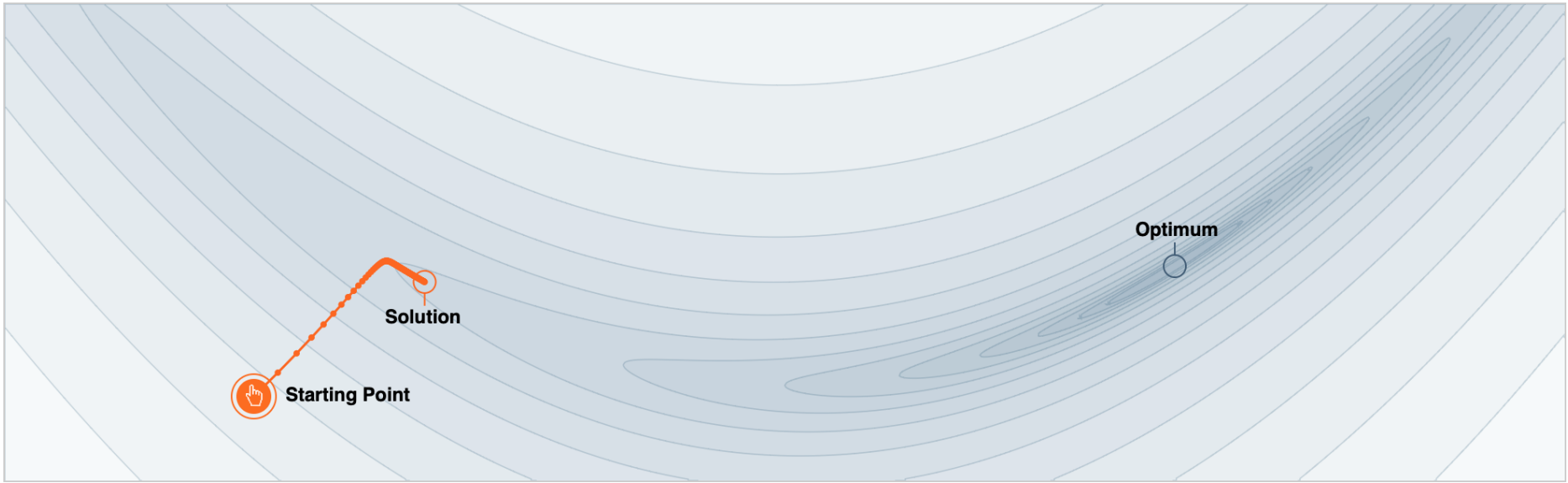


Image shows:

- Path of $\boldsymbol{\theta}_t$ 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.

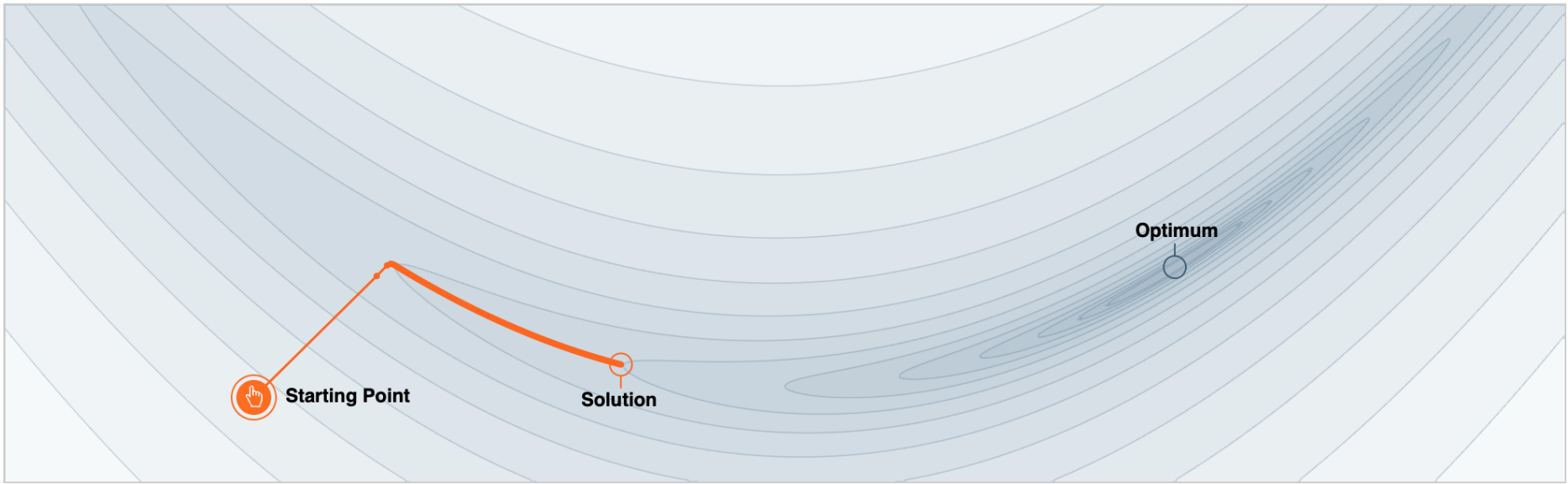


Image shows:

- Path of $\boldsymbol{\theta}_t$ 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.

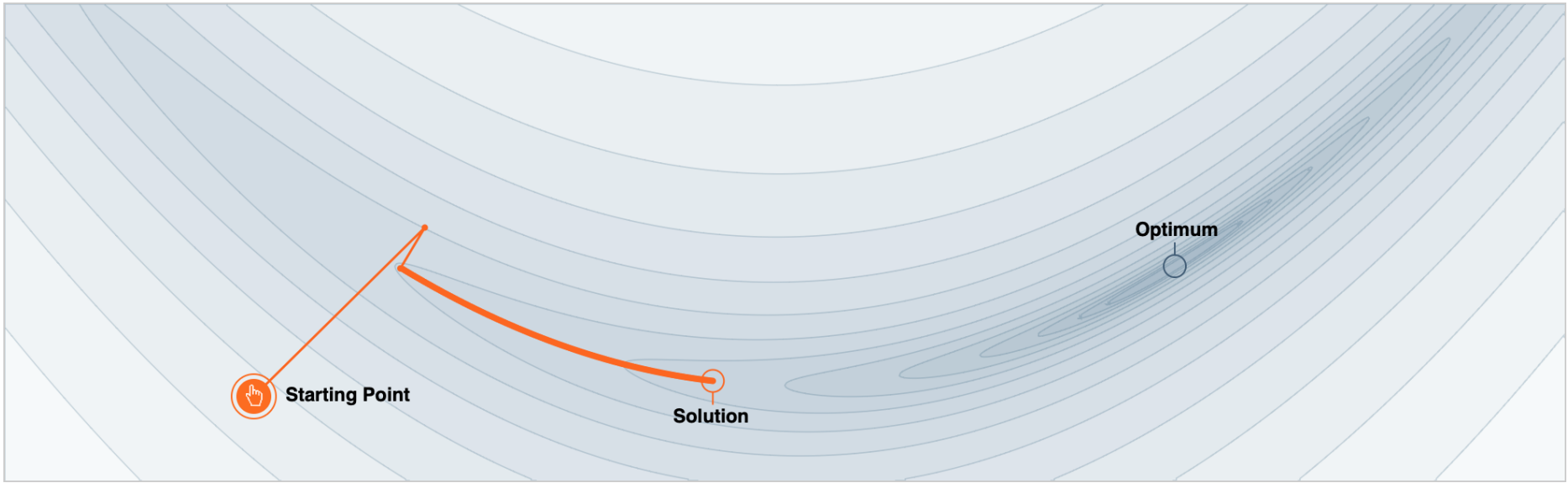


Image shows:

- Path of $\boldsymbol{\theta}_t$ 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.

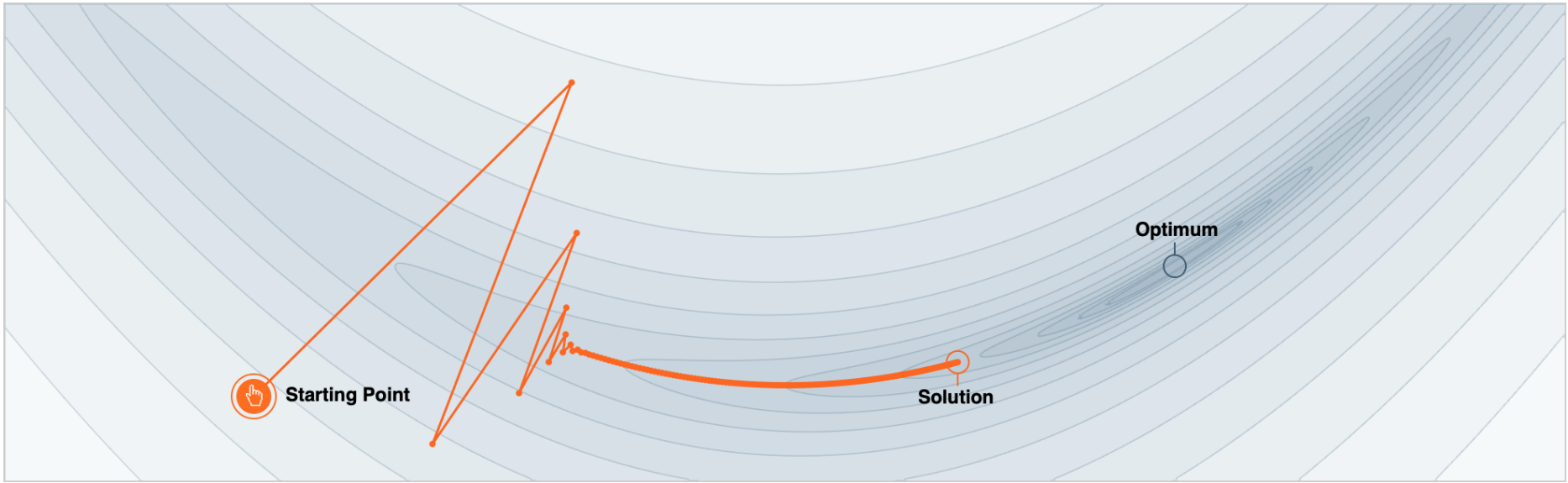


Image shows:

- Path of $\boldsymbol{\theta}_t$ 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.

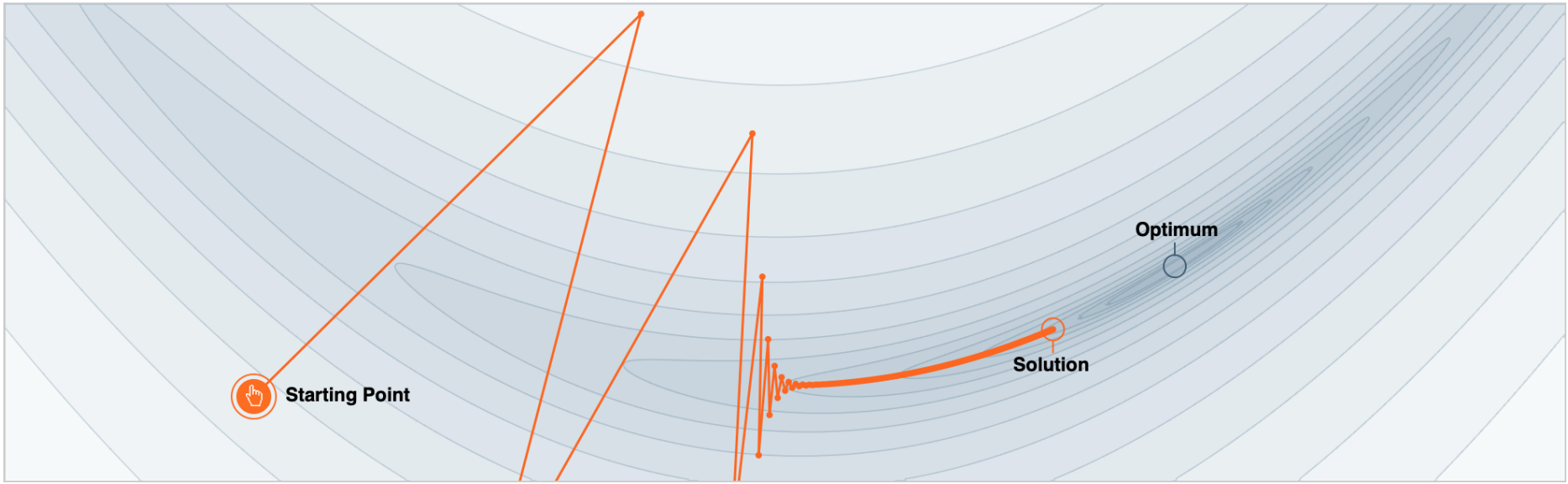


Image shows:

- Path of θ_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.

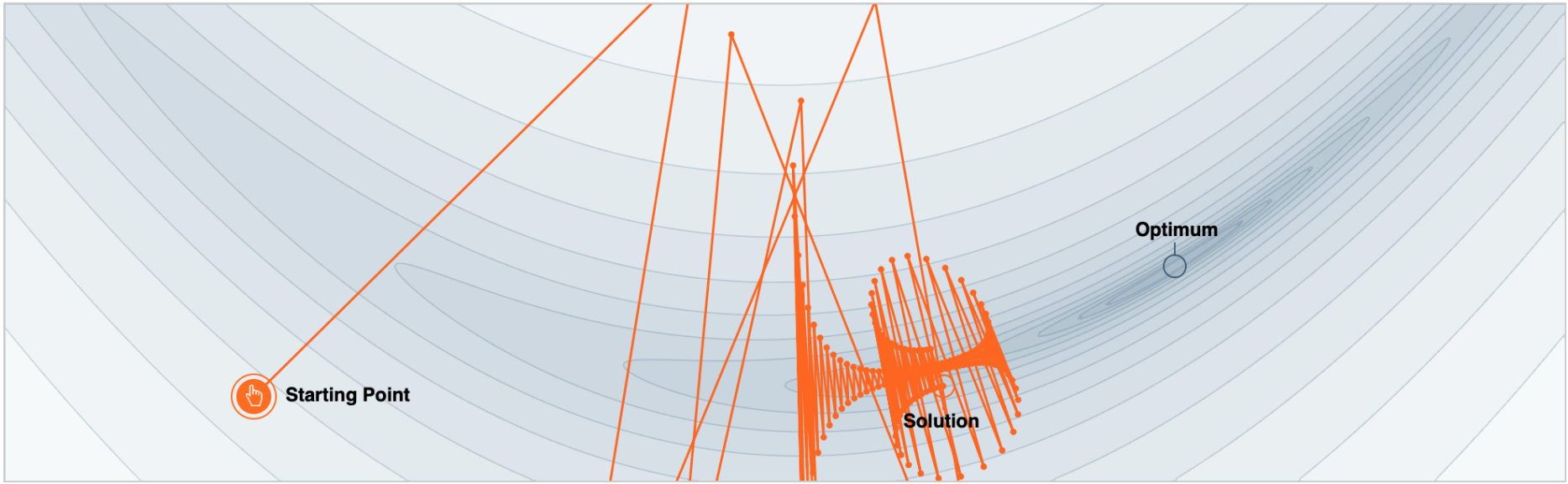


Image shows:

- Path of θ_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.