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

$$\boldsymbol{\theta}^* = \arg\min L(\boldsymbol{\theta}^*)$$

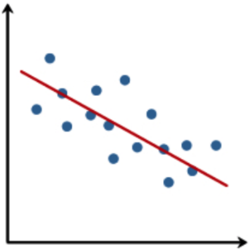
But when does gradient descent converge to a (local) optimum?

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

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:

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

- We would like to figure out $\boldsymbol{\theta}_t$ as a function of $\boldsymbol{\theta}_0$ and γ!
  (and also other hyper-parameters & data)

Gradient descent for linear regression

Arithmetico-geometric sequence:
If a sequence $\{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_T\}$ is defined by

$$\boldsymbol{\theta}_{t+1} = \mathbf{B}\boldsymbol{\theta}_t + \boldsymbol{c}, \quad t \geqslant 0,$$

Then we have

$$\boldsymbol{\theta}_{t+1} = \mathbf{A}(\boldsymbol{\theta}_t + \boldsymbol{\beta}) - \boldsymbol{\beta}, \quad \text{for some } \mathbf{A}, \boldsymbol{\beta}.$$

Gradient descent for linear regression

Arithmetico-geometric sequence:
If a sequence $\{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_T\}$ is defined by

$$\boldsymbol{\theta}_{t+1} = \mathbf{B}\boldsymbol{\theta}_t + \boldsymbol{c}, \quad t \geqslant 0,$$

Then we have

$$\boldsymbol{\theta}_{t+1} = \mathbf{A}(\boldsymbol{\theta}_t + \boldsymbol{\beta}) - \boldsymbol{\beta}, \quad \text{for some } \mathbf{A}, \boldsymbol{\beta}.$$

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

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

Gradient descent for linear regression

Arithmetico-geometric sequence:
If a sequence $\{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_T\}$ is defined by

$$\boldsymbol{\theta}_{t+1} = \mathbf{B}\boldsymbol{\theta}_t + \boldsymbol{c}, \quad t \geqslant 0,$$

Then we have

$$\boldsymbol{\theta}_{t+1} = \mathbf{A}(\boldsymbol{\theta}_t + \boldsymbol{\beta}) - \boldsymbol{\beta}, \quad \text{for some } \mathbf{A}, \boldsymbol{\beta}.$$

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

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

$$\begin{aligned} \Rightarrow \quad \boldsymbol{\theta}_{t+1} &= \mathbf{B}(\boldsymbol{\theta}_t + (\mathbf{B} - \mathbf{I})^{-1}\boldsymbol{c}) - (\mathbf{B} - \mathbf{I})^{-1}\boldsymbol{c} \\ \Rightarrow \quad \boldsymbol{\theta}_t &= \mathbf{B}^t(\boldsymbol{\theta}_0 + (\mathbf{B} - \mathbf{I})^{-1}\boldsymbol{c}) - (\mathbf{B} - \mathbf{I})^{-1}\boldsymbol{c} \end{aligned}$$

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

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 $\boldsymbol{\theta}^*$:

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

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

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

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

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 $\boldsymbol{\theta}^*$:

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

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

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

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

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 and No.

1. You choose a step-size $\gamma \geqslant 2\sigma^2/\lambda_{min}(\mathbf{X}^\top\mathbf{X}) \quad \Rightarrow$ 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 $\gamma \geqslant 2\sigma^2/\lambda_{min}(\mathbf{X}^\top\mathbf{X}) \quad \Rightarrow$ diverge
2. You choose a step-size $\gamma \in [\frac{2\sigma^2}{\lambda_{max}(\mathbf{X}^\top\mathbf{X})}, \frac{2\sigma^2}{\lambda_{min}(\mathbf{X}^\top\mathbf{X})}) \quad \Rightarrow$ good luck
   - Convergence result may be sensitive to initialisation $\boldsymbol{\theta}_0$

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 $\boldsymbol{\theta}_0$?

Choosing step-size for linear regression

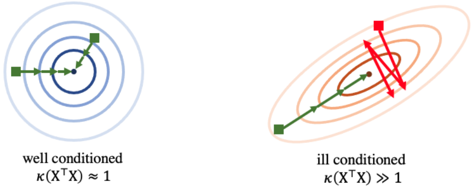
To ensure convergence at **any** initialisation: $\gamma < 2\sigma^2/\lambda_{max}(\mathbf{X}^\top \mathbf{X})$

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**:

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



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

Choosing step-size for linear regression

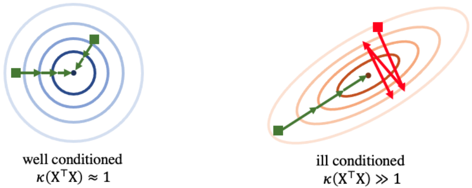
To ensure convergence at **any** initialisation: $\gamma < 2\sigma^2/\lambda_{max}(\mathbf{X}^\top\mathbf{X})$

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**:

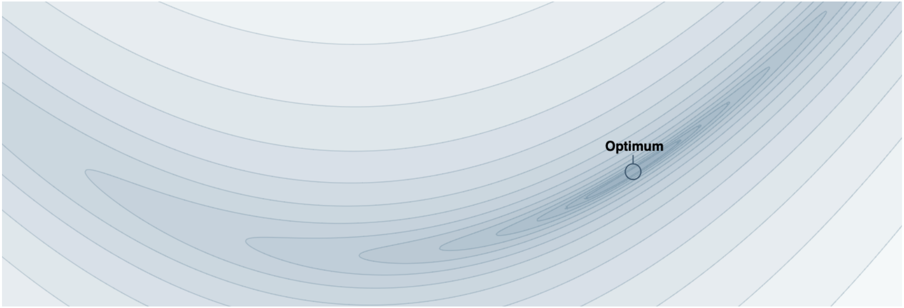
$$\kappa(\mathbf{X}^\top\mathbf{X}) := \frac{\lambda_{max}(\mathbf{X}^\top\mathbf{X})}{\lambda_{min}(\mathbf{X}^\top\mathbf{X})}$$



- 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

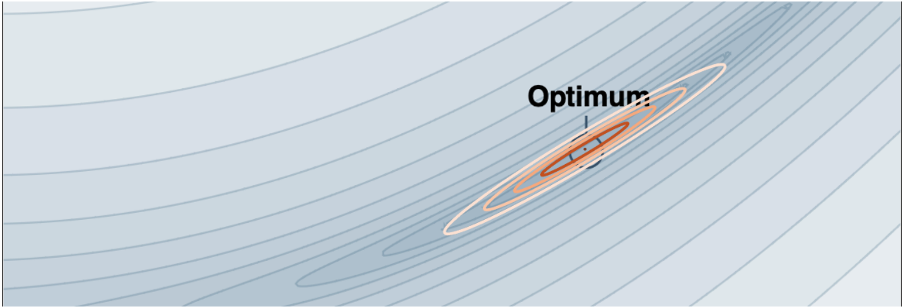
In general the loss function is non-quadratic nor convex:



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.

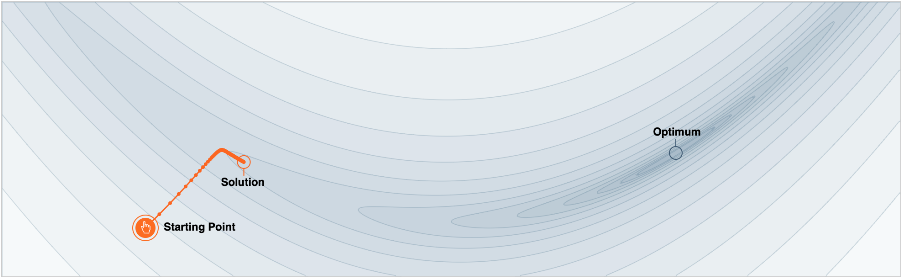


Image shows:

- 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.

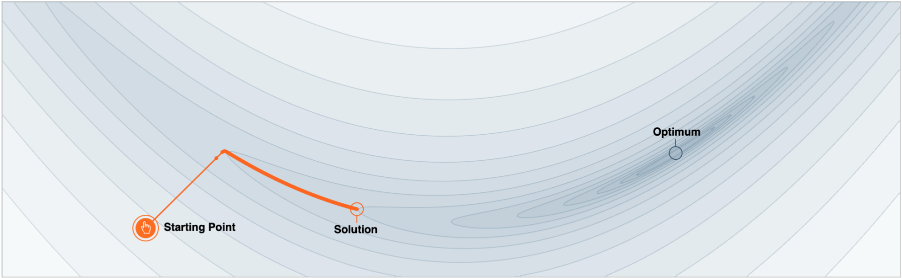


Image shows:

- 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.

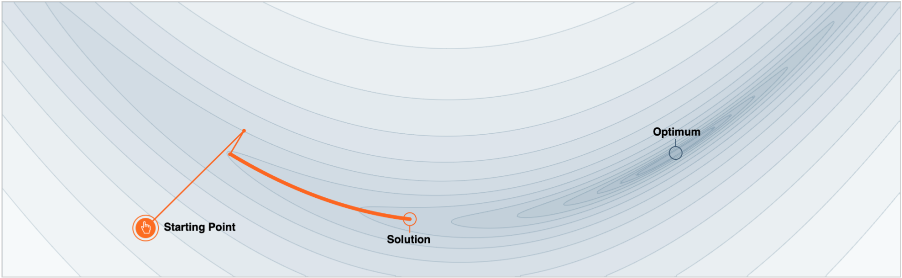


Image shows:

- 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.

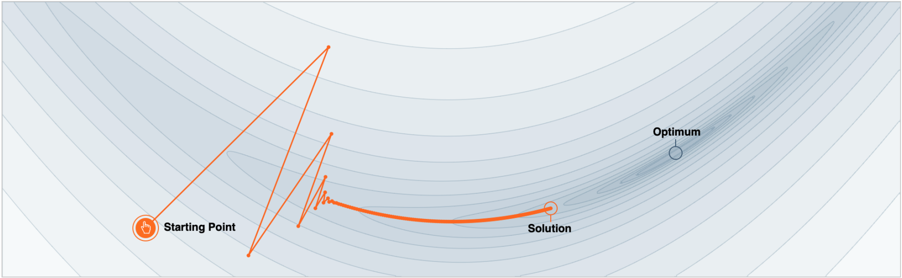


Image shows:

- 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.

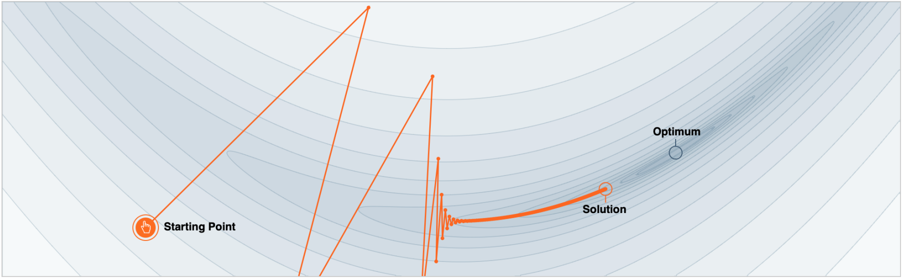


Image shows:

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

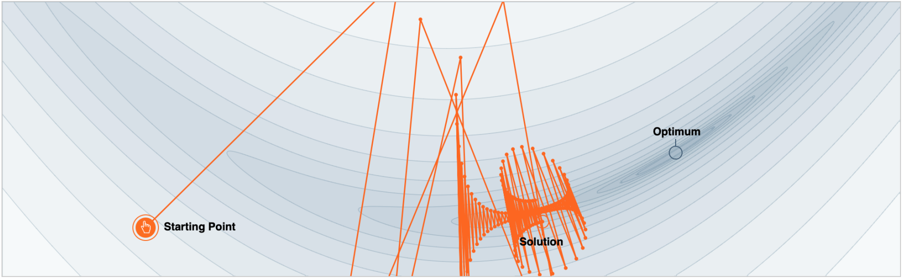


Image shows:

- Path of $\boldsymbol{\theta}_t$ from Gradient Descent
- Constant step size $\gamma_t = \gamma$

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

Choosing step-size: 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.