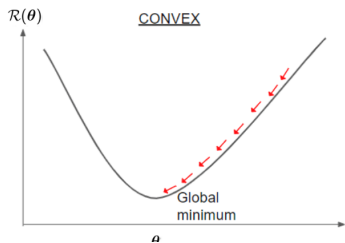
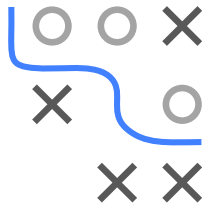


# Optimization in Machine Learning

## Deep dive

## Gradient descent and optimality



### Learning goals

- Convergence of GD
- Proof strategy and tools
- Descent lemma



## SETTING / 2

### Assumptions:

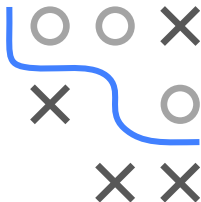
- $f$  convex and differentiable
- Global minimum  $\mathbf{x}^*$  exists
- $f$  has Lipschitz gradient ( $\nabla f$  does not change too fast)

$$\|\nabla f(\mathbf{x}) - \nabla f(\tilde{\mathbf{x}})\| \leq L\|\mathbf{x} - \tilde{\mathbf{x}}\| \quad \text{for all } \mathbf{x}, \tilde{\mathbf{x}}$$

**Theorem** (Convergence of GD). GD with step size  $\alpha \leq 1/L$  yields

$$f(\mathbf{x}^{[k]}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{[0]} - \mathbf{x}^*\|^2}{2\alpha k}.$$

In other words: GD converges with rate  $\mathcal{O}(1/k)$ .



# PROOF STRATEGY

- 1 Show that  $f(\mathbf{x}^{[t]})$  **strictly decreases** with each iteration  $t$

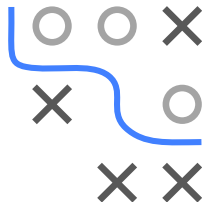
**Descent lemma:**

$$f(\mathbf{x}^{[t+1]}) \leq f(\mathbf{x}^{[t]}) - \frac{\alpha}{2} \|\nabla f(\mathbf{x}^{[t]})\|^2$$

- 2 Bound **error of one step**

$$f(\mathbf{x}^{[t+1]}) - f(\mathbf{x}^*) \leq \frac{1}{2\alpha} \left( \|\mathbf{x}^{[t]} - \mathbf{x}^*\|^2 - \|\mathbf{x}^{[t+1]} - \mathbf{x}^*\|^2 \right)$$

- 3 Finalize by **telescoping** argument

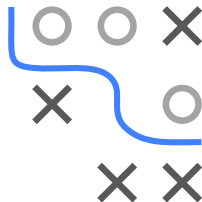
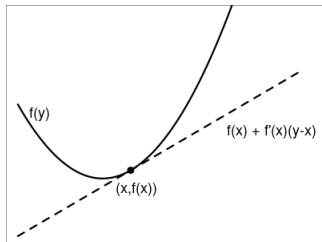


# MAIN TOOL

**Recall:** First order condition of convexity

Every tangent line of  $f$  is always below  $f$ .

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$



# DESCENT LEMMA

**Recall:**  $\nabla f$  Lipschitz  $\implies \nabla^2 f(\mathbf{x}) \preceq L \cdot \mathbf{I}$  for all  $\mathbf{x}$

This gives convexity of  $g(\mathbf{x}) := \frac{L}{2}\|\mathbf{x}\|^2 - f(\mathbf{x})$  since

$$\nabla^2 g(\mathbf{x}) = L \cdot \mathbf{I} - \nabla^2 f(\mathbf{x}) \succeq 0.$$

First order condition of convexity of  $g$  yields

$$g(\mathbf{x}) \geq g(\mathbf{x}^{[t]}) + \nabla g(\mathbf{x}^{[t]})^\top (\mathbf{x} - \mathbf{x}^{[t]})$$

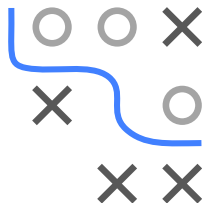
$$\Leftrightarrow \frac{L}{2}\|\mathbf{x}\|^2 - f(\mathbf{x}) \geq \frac{L}{2}\|\mathbf{x}^{[t]}\|^2 - f(\mathbf{x}^{[t]}) + (L\mathbf{x}^{[t]} - \nabla f(\mathbf{x}^{[t]}))^\top (\mathbf{x} - \mathbf{x}^{[t]})$$

$$\Leftrightarrow \quad \quad \quad \vdots$$

$$\Leftrightarrow f(\mathbf{x}) \leq f(\mathbf{x}^{[t]}) + \nabla f(\mathbf{x}^{[t]})^\top (\mathbf{x} - \mathbf{x}^{[t]}) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^{[t]}\|^2$$

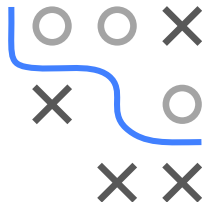
**Now:** One GD step with step size  $\alpha \leq 1/L$ :

$$\mathbf{x} \leftarrow \mathbf{x}^{[t+1]} = \mathbf{x}^{[t]} - \alpha \nabla f(\mathbf{x}^{[t]})$$



## DESCENT LEMMA / 2

$$\begin{aligned}f(\mathbf{x}^{[t+1]}) &\leq f(\mathbf{x}^{[t]}) + \nabla f(\mathbf{x}^{[t]})^\top (\mathbf{x}^{[t+1]} - \mathbf{x}^{[t]}) + \frac{L}{2} \|\mathbf{x}^{[t+1]} - \mathbf{x}^{[t]}\|^2 \\&= f(\mathbf{x}^{[t]}) + \nabla f(\mathbf{x}^{[t]})^\top (\mathbf{x}^{[t]} - \alpha \nabla f(\mathbf{x}^{[t]}) - \mathbf{x}^{[t]}) \\&\quad + \frac{L}{2} \|\mathbf{x}^{[t]} - \alpha \nabla f(\mathbf{x}^{[t]}) - \mathbf{x}^{[t]}\|^2 \\&= f(\mathbf{x}^{[t]}) - \nabla f(\mathbf{x}^{[t]})^\top \alpha \nabla f(\mathbf{x}^{[t]}) + \frac{L}{2} \|\alpha \nabla f(\mathbf{x}^{[t]})\|^2 \\&= f(\mathbf{x}^{[t]}) - \alpha \|\nabla f(\mathbf{x}^{[t]})\|^2 + \frac{L\alpha^2}{2} \|\nabla f(\mathbf{x}^{[t]})\|^2 \\&\leq f(\mathbf{x}^{[t]}) - \frac{\alpha}{2} \|\nabla f(\mathbf{x}^{[t]})\|^2\end{aligned}$$



**Note:**  $\alpha \leq 1/L$  yields  $L\alpha^2 \leq \alpha$

- $\|\nabla f(\mathbf{x}^{[t]})\|^2 > 0$  unless  $\nabla f(\mathbf{x}) = \mathbf{0}$
- $f$  **strictly decreases** with each GD iteration until optimum reached
- Descent lemma yields bound on **guaranteed progress** if  $\alpha \leq 1/L$   
(explains why GD may diverge if step sizes too large)

# ONE STEP ERROR BOUND

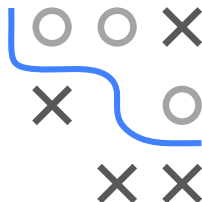
Again, first order condition of convexity gives

$$f(\mathbf{x}^{[t]}) - f(\mathbf{x}^*) \leq \nabla f(\mathbf{x}^{[t]})^\top (\mathbf{x}^{[t]} - \mathbf{x}^*).$$

This and the descent lemma yields

$$\begin{aligned} f(\mathbf{x}^{[t+1]}) - f(\mathbf{x}^*) &\leq f(\mathbf{x}^{[t]}) - \frac{\alpha}{2} \|\nabla f(\mathbf{x}^{[t]})\|^2 - f(\mathbf{x}^*) \\ &= f(\mathbf{x}^{[t]}) - f(\mathbf{x}^*) - \frac{\alpha}{2} \|\nabla f(\mathbf{x}^{[t]})\|^2 \\ &\leq \nabla f(\mathbf{x}^{[t]})^\top (\mathbf{x}^{[t]} - \mathbf{x}^*) - \frac{\alpha}{2} \|\nabla f(\mathbf{x}^{[t]})\|^2 \\ &= \frac{1}{2\alpha} \left( \|\mathbf{x}^{[t]} - \mathbf{x}^*\|^2 - \|\mathbf{x}^{[t]} - \mathbf{x}^* - \alpha \nabla f(\mathbf{x}^{[t]})\|^2 \right) \\ &= \frac{1}{2\alpha} \left( \|\mathbf{x}^{[t]} - \mathbf{x}^*\|^2 - \|\mathbf{x}^{[t+1]} - \mathbf{x}^*\|^2 \right) \end{aligned}$$

**Note:** Line 3  $\rightarrow$  4 is hard to see (just expand line 4).





# FINALIZATION

Summing over iterations yields

$$\begin{aligned}k(f(\mathbf{x}^{[k]}) - f(\mathbf{x}^*)) &\leq \sum_{t=1}^k [f(\mathbf{x}^{[t]}) - f(\mathbf{x}^*)] \\ &\leq \sum_{t=1}^k \frac{1}{2\alpha} \left[ \|\mathbf{x}^{[t-1]} - \mathbf{x}^*\|^2 - \|\mathbf{x}^{[t]} - \mathbf{x}^*\|^2 \right] \\ &= \frac{1}{2\alpha} \left( \|\mathbf{x}^{[0]} - \mathbf{x}^*\|^2 - \|\mathbf{x}^{[k]} - \mathbf{x}^*\|^2 \right) \\ &\leq \frac{1}{2\alpha} \left( \|\mathbf{x}^{[0]} - \mathbf{x}^*\|^2 \right).\end{aligned}$$



**Arguments:** Descent lemma (line 1). Telescoping sum (line 2  $\rightarrow$  3).

$$f(\mathbf{x}^{[t+1]}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{[0]} - \mathbf{x}^*\|^2}{2\alpha k}$$