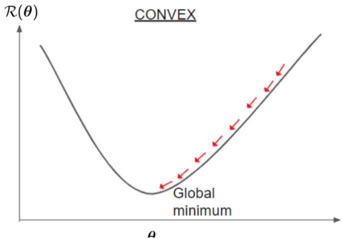
Optimization in Machine Learning

Deep dive: Gradient descent and optimality

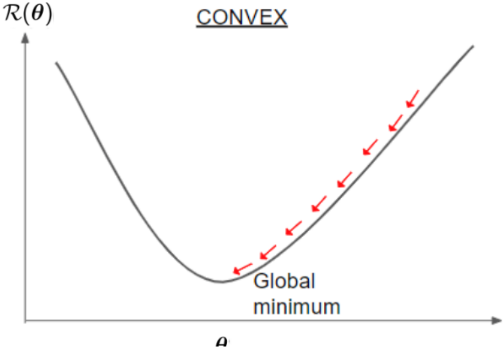


Learning goals

- Convergence of GD

GRADIENT DESCENT AND OPTIMALITY

- GD is a greedy algorithm: In every iteration, it makes locally optimal moves.
- If $\mathcal{R}(\theta)$ is **convex** and **differentiable**, and its gradient is Lipschitz continuous, GD is guaranteed to converge to the global minimum for small enough step-size.



GRADIENT DESCENT AND OPTIMALITY

Assume $f : \mathbb{R}^d \rightarrow \mathbb{R}$ convex and differentiable and assume that global minimum $\mathbf{x}^*$ exists. Assume ∇f is Lipschitz continuous with $L > 0$:

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

(i.e., gradient can't change arbitrarily fast).

Convergence of GD: GD with k iterations with starting point $\mathbf{x}^{[0]}$ and fixed step-size $\alpha \leq 1/L$ will yield a solution $f(\mathbf{x}^{[k]})$, which satisfies

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

This means, that GD converges with rate $\mathcal{O}(1/k)$.

GRADIENT DESCENT AND OPTIMALITY

Proof: From ∇f Lipschitz it follows that $\nabla^2 f(\mathbf{x}) \preccurlyeq L \cdot \mathbf{I}$ for all $\mathbf{x}$.

NB: The generalized inequality $\nabla^2 f(\mathbf{x}) \preccurlyeq LI$ means that $L \cdot \mathbf{I} - \nabla^2 f(\mathbf{x})$ is positive semidefinite. This means that $\boldsymbol{v}^\top \nabla^2 f(\boldsymbol{u})\boldsymbol{v} \leq L||\boldsymbol{v}||^2$ for any $\boldsymbol{u}$ and $\boldsymbol{v}$.

We perform a quadratic expansion of f around $\tilde{\mathbf{x}}$:

$$
\begin{aligned}
f(\mathbf{x}) &\approx f(\tilde{\mathbf{x}}) + \nabla f(\tilde{\mathbf{x}})^\top(\mathbf{x} - \tilde{\mathbf{x}}) + 0.5(\mathbf{x} - \tilde{\mathbf{x}})^\top \nabla^2 f(\tilde{\mathbf{x}})(\mathbf{x} - \tilde{\mathbf{x}}) \\
&\leq f(\tilde{\mathbf{x}}) + \nabla f(\tilde{\mathbf{x}})^\top(\mathbf{x} - \tilde{\mathbf{x}}) + 0.5L||\mathbf{x} - \tilde{\mathbf{x}}||^2 \text{ (descent lemma)},
\end{aligned}
$$

as the blue term is at most $0.5 \cdot L \cdot ||\mathbf{x} - \tilde{\mathbf{x}}||^2$.

Now, do one GD update with step size $\alpha \leq 1/L$:

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

and plug this in the descent lemma.

GRADIENT DESCENT AND OPTIMALITY

We get

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

where we used $\alpha \leq 1/L$ and therefore $-(1 - \frac{1}{2} L \alpha) \leq \frac{1}{2} L \frac{1}{L} - 1 = -\frac{1}{2}$.
Since $\frac{1}{2} \alpha ||\nabla f(\mathbf{x}^{[t]})||^2$ is always positive unless $\nabla f(x) = 0$, it implies that f strictly decreases with each iteration of GD until the optimal value is reached. So, it is a bound on guaranteed progress if $\alpha \leq 1/L$.

GRADIENT DESCENT AND OPTIMALITY

Now, we bound $f(\mathbf{x})$ in terms of $f(\mathbf{x}^*)$ using that f is convex:

$$f(\mathbf{x}) \leq f(\mathbf{x}^*) + \nabla f(\mathbf{x})^\top (\mathbf{x} - \mathbf{x}^*)$$

When we combine this and the bound derived before, we get

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

This holds for every iteration of GD.

GRADIENT DESCENT AND OPTIMALITY

Summing over iterations, we get:

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

where we used that the LHS is a telescoping sum. In addition, we know that f decreases on every iteration, so we can conclude that

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