Optimization

First order methods: Step size and Optimality

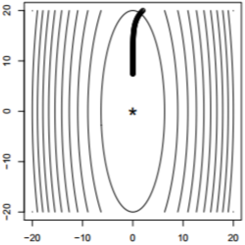
Learning goals

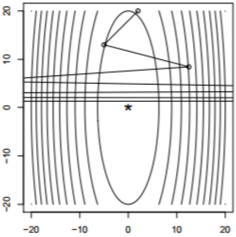
- LEARNING GOAL 1
- LEARNING GOAL 2

CONTROLLING STEP SIZE

In every iteration t, we need to choose not only a descent direction $\mathbf{d}^{[t]}$, but also a step size $\alpha^{[t]}$:

- If $\alpha^{[t]}$ is too small, the procedure may converge very slowly (left).
- If $\alpha^{[t]}$ is too large, the procedure may not converge, because we "jump" around the optimum (right).

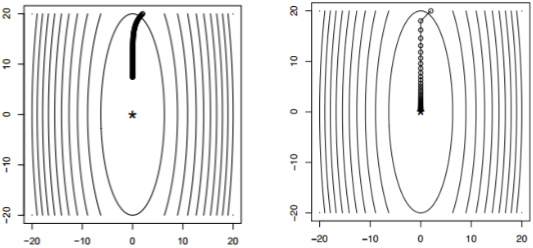




STEP SIZE CONTROL: FIXED STEP SIZE

Use fixed step size α in each iteration:

$$\alpha^{[t]} = \alpha$$

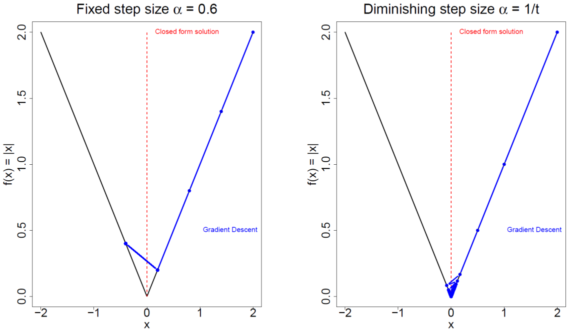


Steps of a line search for $f(\boldsymbol{x}) = 10x_1^2 + 0.5x_2^2$, left 100 steps with fixed step size, right only 40 steps with adaptively selected step size.

Problem: Difficult to determine the optimal step size and depending on the problem the optimal step size has different values at different times.

STEP SIZE CONTROL: DIMINISHING STEP SIZE

- A natural way of selecting α is to decrease its value over time



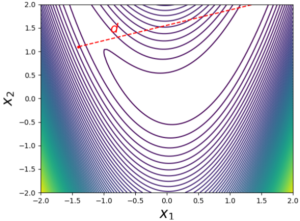
Example: GD on $f(x) = |x|$ with diminishing step size $\alpha^{[t]} = \frac{1}{t}$, with t being the iteration of GD. In this case a diminishing step length is absolutely necessary in order to reach a point close to the minimum.

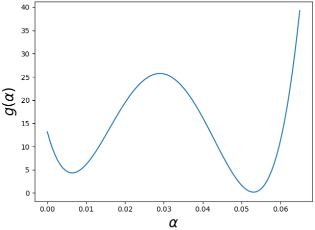
STEP SIZE CONTROL: EXACT LINE-SEARCH

Use the **optimal** step size in each iteration:

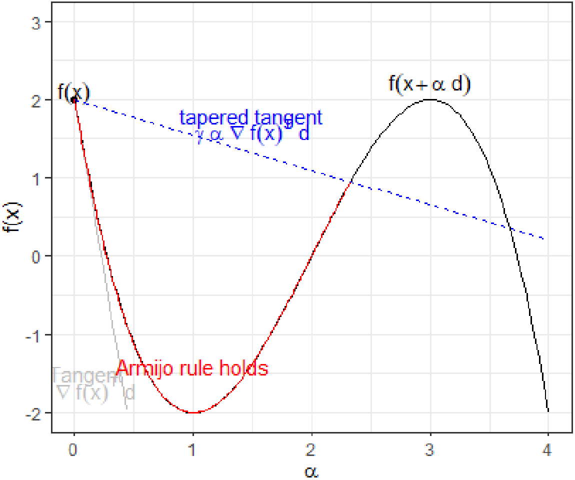
$$\alpha^{[t]} = \arg\min_{\alpha \in \mathbb{R}_{\geq 0}} g(\alpha) = \arg\min_{\alpha \in \mathbb{R}_{\geq 0}} f(\mathbf{x}^{[t]} + \alpha \mathbf{d}^{[t]})$$

In each iteration an **univariate optimization problem** $\arg\min g(\alpha)$ must be solved with methods of univariate optimization (e.g. golden ratio). However, exact line-search is often too expensive for practical purposes and prone to poorly conditioned problems.



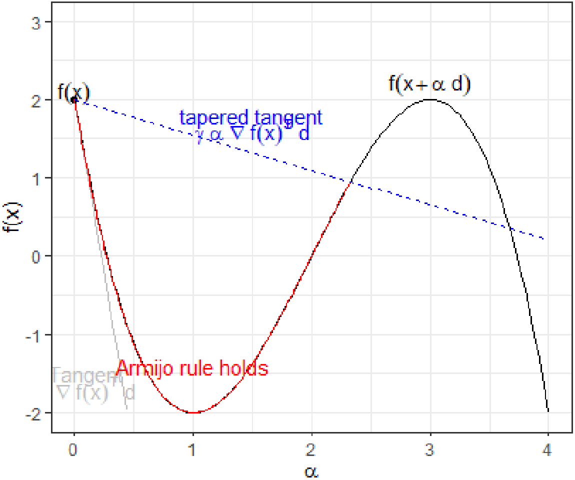


ARMIJO RULE



Inexact line search are efficient procedures of computing a step size that minimizes the objective "sufficiently", without computing the optimal step size exactly. A common condition that ensures that the objective decreases "sufficiently" is the **Armijo rule**.

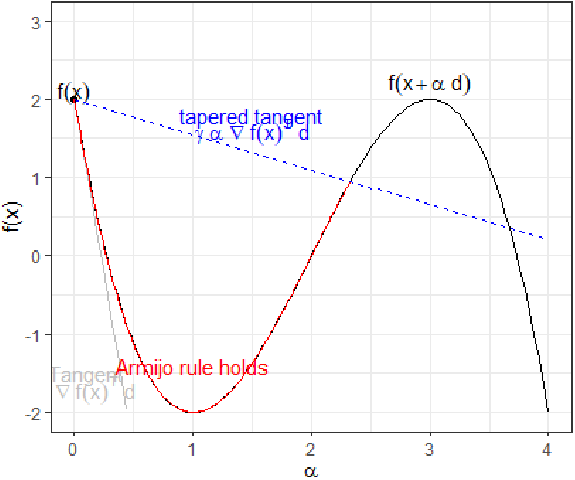
ARMIJO RULE



A step size α is said to satisfy the **Armijo rule** in $\boldsymbol{x}$ for the descent direction $\boldsymbol{d}$ if for a fixed $\gamma \in (0,1)$ the following applies:

$$f(\mathbf{x} + \alpha \boldsymbol{d}) \leq f(\mathbf{x}) + \gamma \alpha \nabla f(\mathbf{x})^\top \boldsymbol{d}.$$

ARMIJO RULE



If $\boldsymbol{d}$ is a descent direction, then for each $\gamma \in (0, 1)$ there exists a step size α, which fulfills the Armijo rule (feasibility).
In many cases, the Armijo rule guarantees local convergence of line searches and is therefore frequently used.

BACKTRACKING LINE SEARCH

Backtracking line search is based on the Armijo rule.

Idea: Decrease α until the Armijo rule is met.

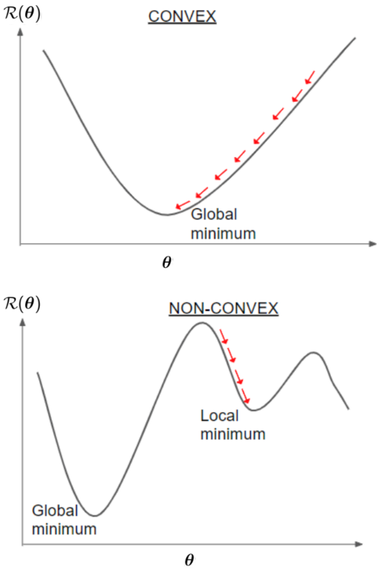
Algorithm Backtracking line search

1: Choose initial step size $\alpha = \alpha^{[0]}$, $0 < \gamma < 1$ and $0 < \tau < 1$
2: **while** $f(\boldsymbol{x} + \alpha \boldsymbol{d}) > f(\boldsymbol{x}) + \gamma \alpha \nabla f(\boldsymbol{x})^\top \boldsymbol{d}$ **do**
3:     Decrease α: $\alpha \leftarrow \tau \cdot \alpha$
4: **end while**

The procedure is simple and shows good performance in practice.

GRADIENT DESCENT AND OPTIMALITY

- GD is a greedy algorithm: In every iteration, it makes locally optimal moves.
- If $\mathcal{R}(\boldsymbol{\theta})$ is **convex** and **differentiable**, and its gradient is Lipschitz continuous, GD is guaranteed to converge to the global minimum (for small enough step-size).
- However, if $\mathcal{R}(\boldsymbol{\theta})$ has multiple local optima and/or saddle points, GD might only converge to a stationary point (other than the global optimum), depending on the starting point.



GRADIENT DESCENT AND OPTIMALITY

We assume that the gradient of the convex and differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz continuous with $L > 0$:

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

This means that the gradient can't change arbitrarily fast.

Now we have a look at the convergence of gradient descent with a fixed step size $\alpha \leq 1/L$.

Convergence: Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and have L-Lipschitz continuous gradients and assuming that the global minimum x^* exists. Then gradient descent with k iterations with a fixed step-size $\alpha \leq 1/L$ will yield a solution $f(x^k)$, which satisfies

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

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

GRADIENT DESCENT AND OPTIMALITY

Proof: The assumption that ∇f is Lipschitz continuous implies that $\nabla^2 f(x) \preccurlyeq LI$ for all x. The generalized inequality $\nabla^2 f(x) \preccurlyeq LI$ means that $LI - \nabla^2 f(x)$ is positive semidefinite. This means that $v^\top \nabla^2 f(u) v \leq L||v||^2$ for any u and v.
Therefore, we can perform a quadratic expansion of f around $\tilde{x}$ obtaining the following inequality:

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

as the blue term is at most $0.5L||x - \tilde{x}||^2$. This is called the descent lemma.
Now, we are doing one update via gradient descent with a step size $\alpha \leq 1/L$:

$$\tilde{x} = x^{t+1} = x^t - \alpha \nabla f(x^t)$$

and plug this in the descent lemma.

GRADIENT DESCENT AND OPTIMALITY

We get

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

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 $\frac{1}{2}\alpha||\nabla f(x^t)||^2$ 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, when $\alpha \leq 1/L$.

GRADIENT DESCENT AND OPTIMALITY

Now, we bound $f(x)$ in terms of $f(x^*)$ and use that f is convex:

$$f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$$

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

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

This holds for every iteration of GD.

GRADIENT DESCENT AND OPTIMALITY

Summing over iterations, we get:

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

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

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