is a descent direction.
(In particular, $-\nabla f(x)$ is a descent direction).

Proof: By Taylor's theorem, we have

$$(\bullet\bullet)\quad f(x+\alpha\rho) = f(x) + \alpha \nabla f(x)^T \rho + o(\alpha).$$

Let $\phi(\alpha) := f(x+\alpha\rho)$.

Then

$$\phi(\alpha) \overset{(\bullet\bullet)}{=} \phi(0) + \phi'(0)\cdot\alpha + o(\alpha)$$

Since $\lim_{\alpha\to 0} \frac{|o(\alpha)|}{\alpha} = 0$, there exists some $\bar{\alpha} > 0$ s.t.

$$(\bullet\bullet\bullet)\quad \frac{|o(\alpha)|}{\alpha} < |\nabla f(x)^T \rho| \quad \forall \alpha \in (0, \bar{\alpha}).$$

This, together with our assumption that $\nabla f(x)^T \rho < 0$, implies that

$$\forall \alpha \in (0, \bar{\alpha}): \quad f(x+\alpha\rho) - f(x) < 0.$$

Because:

$$f(x+\alpha\rho) - f(x) = \alpha \nabla f(x)^T \rho + o(\alpha)$$

$$\Rightarrow \quad \frac{f(x+\alpha\rho) - f(x)}{\alpha} = \underbrace{\nabla f(x)^T \rho}_{<0 \text{ per assumption}} + \underbrace{\frac{o(\alpha)}{\alpha}}_{\begin{array}{l}\text{if } o(\alpha)<0 \Rightarrow \text{ok} \\ \text{if } o(\alpha)>0 \rightarrow \text{use } (\bullet\bullet\bullet)\end{array}}$$

Hence, ρ is a descent direction. ∎

Lemma: Comide a positive definite matrix B. For any point x with $\nabla f(x) \neq 0$, the direction $-B \nabla f(x)$ is a descent direction.

Proof: We have $-\nabla f(x)^T B \nabla f(x) < 0$, by the assumption that B is positive definite. The assertion follows with the previous lemma. ∎

This suggests a general paradigm for a descent algorithm:

$$x^{k+1} = x^k - \alpha_k B_k \nabla f(x^k)$$

with B_k being positive definite.

Common choices of descent direction:

- Steepest descent: $B_k = \underbrace{I}_{\text{identity matrix}} \quad \forall k$
- Newton direction: $B_k = \left(\nabla^2 f(x^k)\right)^{-1}$ (if the Hessian is positive definite)
- Modified Newton direction:

$$B_k = \left(\nabla^2 f(x^*)\right)^{-1} \quad \forall k$$

(compute Newton direction only at the beginning; or: once every M steps). (assuming that $\nabla^2 f(x^*)$ is pos. def.)

Line Search Methods

We consider the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

with objective function $f : \mathbb{R}^n \to \mathbb{R}$.

In order to compute approximations of minimizers of this problem, let us consider iterative methods.

The general form of the iterations are as follows:

$$x^{k+1} = x^k + \alpha_k p^k,$$

where

$k \in \mathbb{N}$: index

$x^k \in \mathbb{R}^n$: current iteration point

$x^{k+1} \in \mathbb{R}^n$: next iteration point

$p^k \in \mathbb{R}^n$: direction to move along at iteration k

$\alpha_k \in \mathbb{R}_+$: step size at iteration k.

How can we choose the step length α_k?

- Constant Step Size: For all k, we choose $\alpha_k = \alpha > 0$. $\Rightarrow$ simplest rule to implement
  However: The iterations may not converge if α is chosen too large; or the iterations may not converge / converge only slowly if α is chosen too small.

  Example: $f(x) = x^2$, $\alpha = 1$ constant step size; $p^k = -f'(x^k) = 2x^k$.

  We get

$$x^0 = 1, \quad x^1 = x^0 - 1 \cdot f'(x^0)$$
$$= 1 - 2 = -1$$
$$x^2 = x^1 - 1 \cdot f'(x^1)$$
$$= -1 + 2 = 1$$

$\Rightarrow$ Thus, the iterations do not converge.

- Exact line search: We consider the problem
$$\min_{\alpha > 0} f(x^k + \alpha p^k) =: \Phi(\alpha).$$
This is a minimization problem itself, but an easier one, as it is one-dimensional. If f is convex, then the problem is also convex.
In general, however, it is too expensive to identify the minimizer α^* in this way. A more practical strategy is the following one:

- Inexact line search: Here, one identifies a step length that achieves adequate reduction in f at minimal cost.

Inexact Line Search Methods

In modern optimization: Try to determine with low effort a step size α_k s.t. the condition
$$f(x^k + \alpha p^k) < f(x^k) \qquad (1)$$
holds true.

We will now see that that this condition

is not sufficient for convergence
minimizes if the step size α is chosen
to small or too large.

Example: $\min_{x \in \mathbb{R}^2} x_1^2 + 2x_2^2 = f(x)$

(The minimizer of this problem is
$x^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$).

Choose $p^k = -\nabla f(x^k)$ and a step size α_k.

The iteration reads:

$$x^{k+1} = x^k + \alpha_k p^k$$

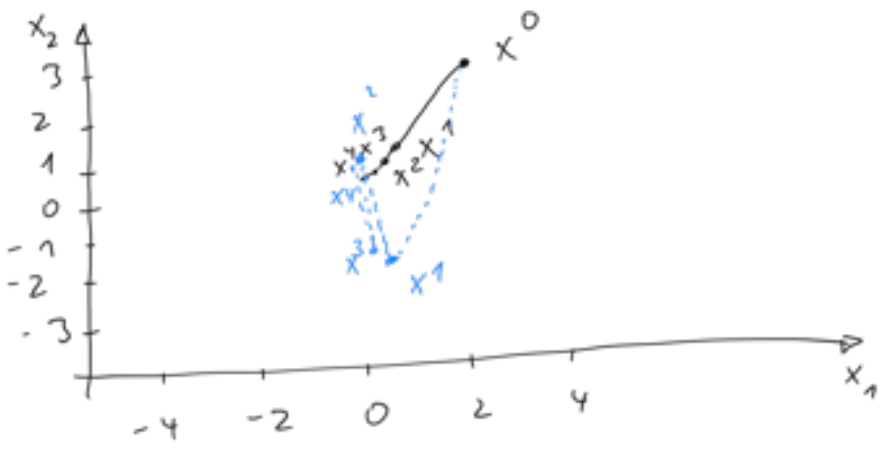
$$\Leftrightarrow \begin{pmatrix} x_1^{k+1} \\ x_2^{k+1} \end{pmatrix} = \begin{pmatrix} x_1^k \\ x_2^k \end{pmatrix} - \alpha_k \cdot \begin{pmatrix} 2x_1^k \\ 4x_2^k \end{pmatrix}$$

Starting point: $x^0 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$.

step size: $\boxed{\alpha_k = \frac{1}{2^{k+3}}}$

We get:

k	x^k	$f(x^k)$
0	$\begin{pmatrix} 2 \\ 3 \end{pmatrix}$	22
1	$\begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix}$	6.75
2	$\begin{pmatrix} 1.313 \\ 1.125 \end{pmatrix}$	4.255
3	$\begin{pmatrix} 1.23 \\ 0.9844 \end{pmatrix}$	3.451
4	$\begin{pmatrix} 1.192 \\ 0.9229 \end{pmatrix}$	3.124



$\Rightarrow$ convergence to the non-minimal point $\bar{x} = \begin{pmatrix} 1.155 \\ 0.8664 \end{pmatrix}$

Reason: The stepsizes are too small compared to the decrease of the objective value.

Stepsize: $\alpha_k = \frac{1}{2} - \frac{1}{2^{k+3}}$

We get:

k	x^k	$f(x^k)$
0	$\begin{pmatrix} 2 \\ 3 \end{pmatrix}$	22
1	$\begin{pmatrix} 0.5 \\ -1.5 \end{pmatrix}$	4.75
2	$\begin{pmatrix} 0.0625 \\ 1.125 \end{pmatrix}$	2.535
3	$\begin{pmatrix} 0.0039 \\ -0.98 \end{pmatrix}$	1.938
4	$\begin{pmatrix} 0.0001221 \\ 0.9229 \end{pmatrix}$	1.703

$\Rightarrow$ The iterations osaillate for large values of k between $\tilde{x} = \begin{pmatrix} 0 \\ 0.86 \end{pmatrix}$ and $-\tilde{x}$.

Reason: The steps are too large.

We will now develop strategies which result in an adequate step size.
Let us substitute the condition (1) by the so-called Armijo-condition:

For some $c_1 \in (0,1)$, let

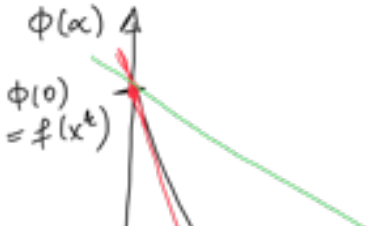
$$(2)\quad \underbrace{f(x^k + \alpha p^k)}_{=:\phi(\alpha)} \leq \underbrace{f(x^k) + c_1 \alpha \nabla f^{kT} p^k}_{=:\ell(\alpha)}$$

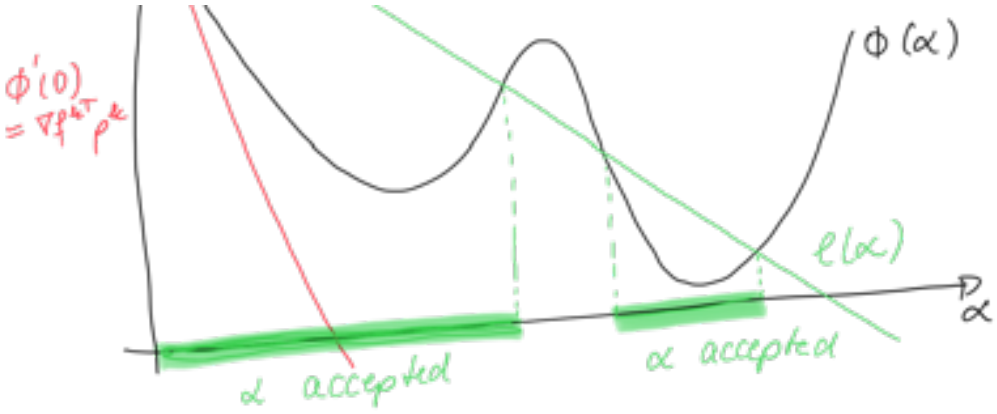
Interpretation: The reduction in f should be proportional to both the step length as well as the directional derivative $\nabla f^{kT} p^k$. This ensures sufficient decrease, and (2) is also called sufficient decrease condition.
(2) states that the step size α is acceptable if $\boxed{\phi(\alpha) \leq \ell(\alpha)}$

Note that $\ell(\alpha)$ is an affine function with neg. slope, as $\nabla f^{kT} p^k < 0$, since p^k is assumed to be a descent direction.

We have $\phi'(0) = \nabla f^{kT} p^k$, and so, the gradient of ϕ at 0 is negative.

Illustration of (2):





→ The step size α is accepted if $\phi(\alpha) \leq \ell(\alpha)$.

For small positive values of α, we have $\phi(\alpha) \leq \ell(\alpha)$, and so, (2) is fulfilled for <u>all</u> sufficiently small α.

Hence, the sufficient decrease condition (2) is not enough to ensure that the algorithm makes reasonable progress.

Therefore, we introduce the so-called <u>curvature-condition</u> for some $c_2 \in (c_1, 1)$:

$$(3) \quad \underbrace{\nabla f(x^k + \alpha_k p^k)^T p^k}_{= \phi'(\alpha_k)} \geq \underbrace{c_2 \nabla f(x^k)^T p^k}_{= c_2 \cdot \phi'(0)}$$

This means that the slope of ϕ at α_k is not lower than, the slope of ϕ at 0.
c_2 times

$\phi(\alpha)$