

GLOBALLY CONVERGENT METHODS

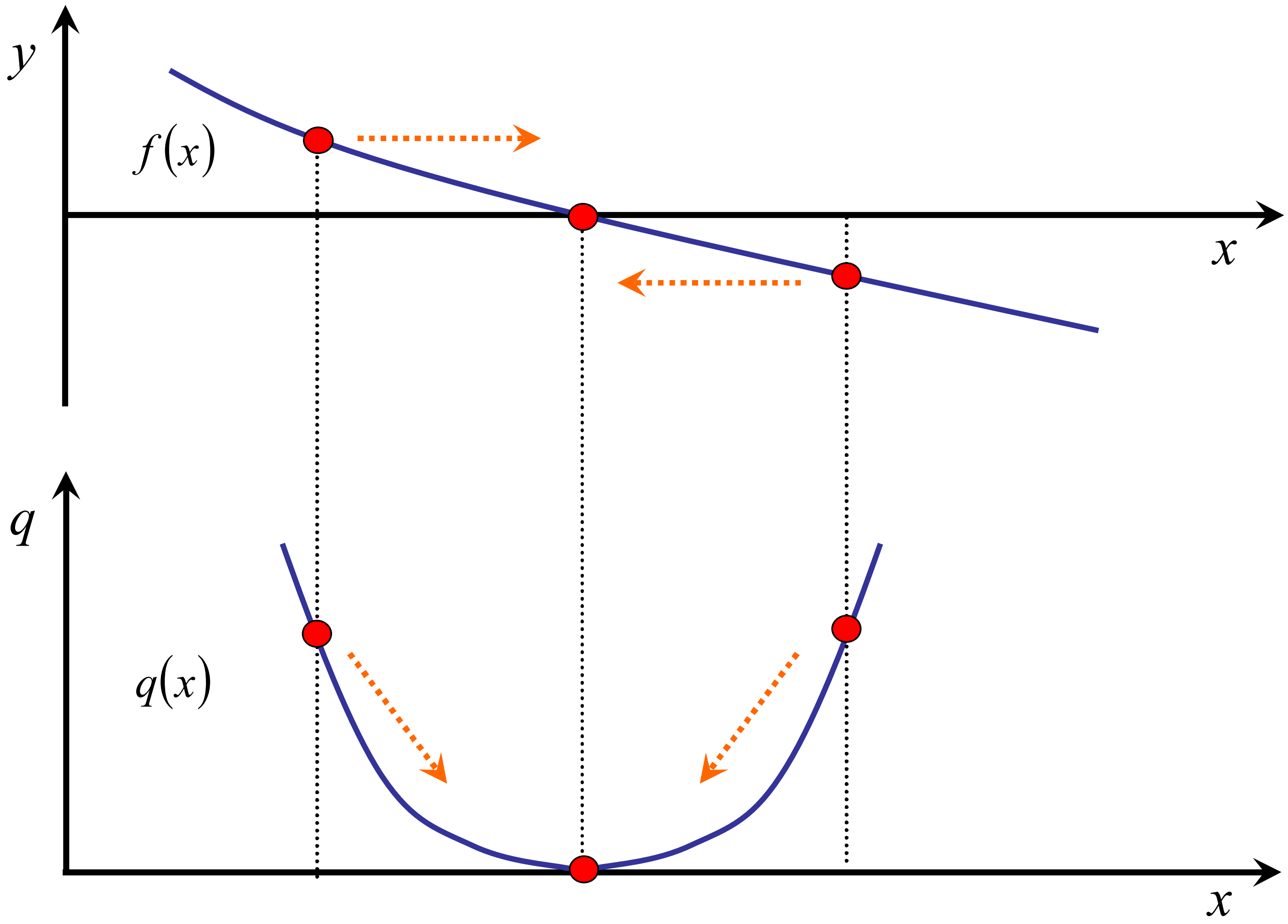
- **Topic:** How to solve a systems of non-linear equations when a good initial guess is not available, or the problem is particularly ill-behaved …
  - This kind of situations are quite frequent in real-world applications.
- Some extensions have been developed to make Newton's method globally convergent.
- Two broad families: *line search methods* and *trust region methods*.
- The same methods can be applied to guarantee global convergence of optimization algorithms

Line search methods

- A solution to $F(x)=0$ is necessarily a solution to:

$$\min_{x \in X} q(x) \equiv \| F(x) \|_2 = \sqrt{F(x)' F(x)}$$

- The converse is clearly false: a solution to this minimization problem is not *necessarily* a solution to $F(x)=0$.
- However, we may intuitively conclude that any iterative method designed to solve $F(x)=0$ should steadily move towards *"descent" directions*, i.e. directions that make q decrease.



y
f(x)
x
q
q(x)
x

- The Newton step is a *descent direction*:

$$d_k = -J(x_k)^{-1} F(x_k)$$

Going from x_k to x_k+d_k decreases, **at least initially**, the value of q, since:

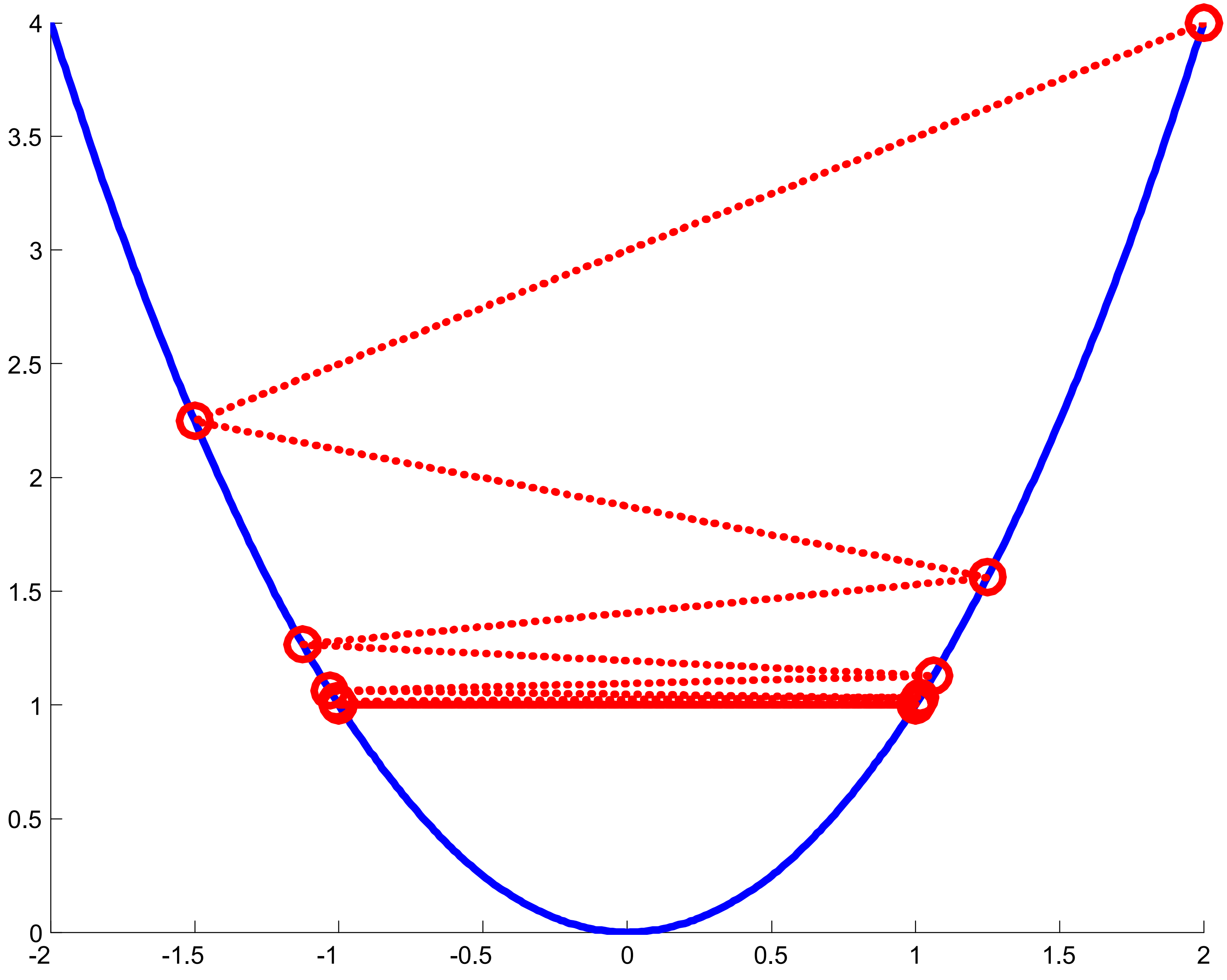
$$\nabla q(x_k) d_k = -\frac{F(x_k)' J(x_k)}{q(x_k)} J(x_k)^{-1} F(x_k) = -q(x_k) < 0$$

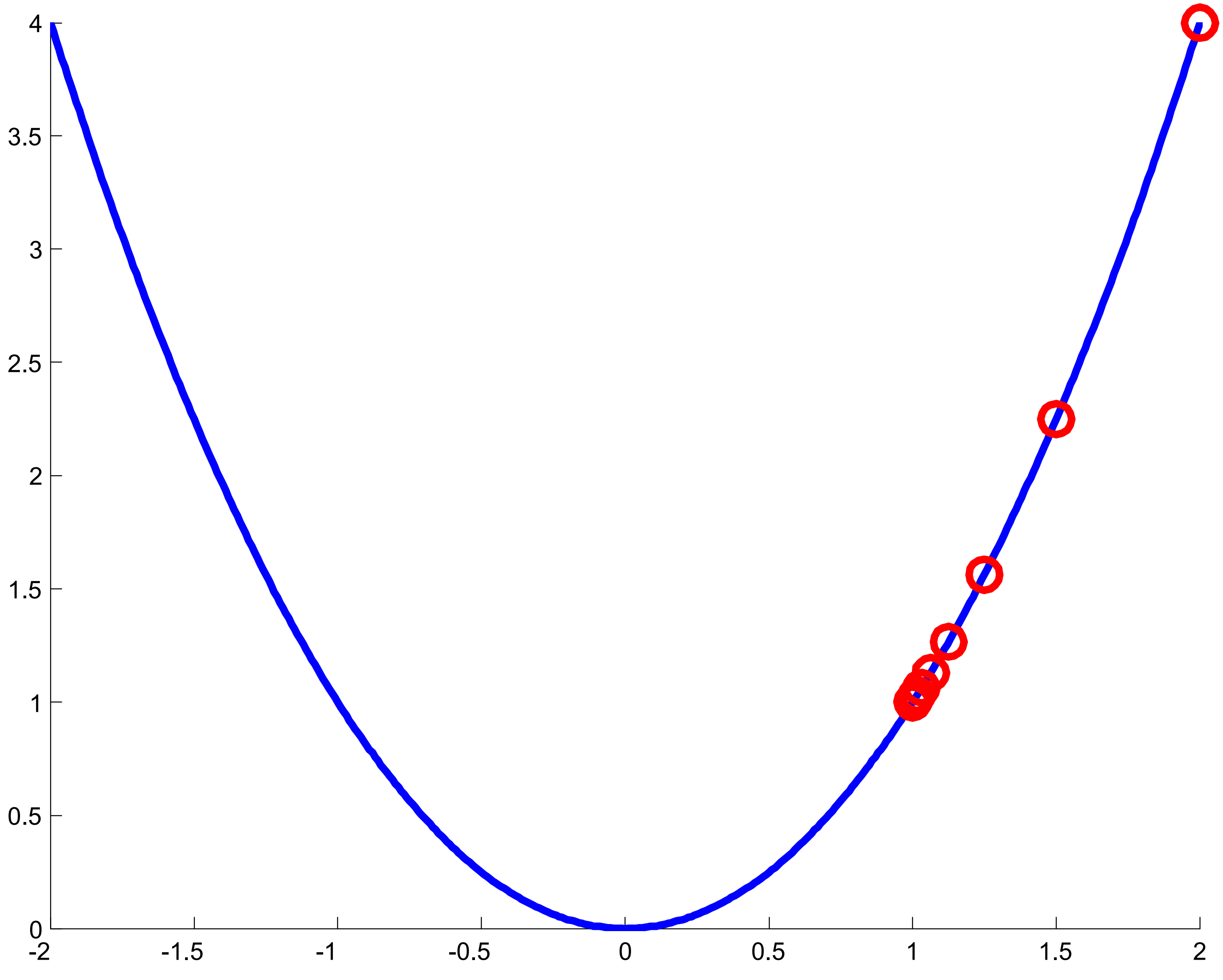
- However, nothing guarantees that: $q(x_{k+1})<q(x_k)$
- If this is not the case, the Newton step is "going to far.”

- Line search methods initially compute the standard Newton step and check whether a "sufficient" decrease – still to be defined - in q takes place or not.
- If the answer is yes, the algorithms update the guess and starts another iteration.
- Otherwise, an alternative step $\lambda_k d_k$ for some $\lambda_k > 0$ that yields a sufficient decrease is found and used to update the current guess.

The Armijo-Goldstein-Wolfe rules

- It turns out that the condition $q(x_{k+1})<q(x_k)$ is actually **too weak** to guarantee global convergence.
- It can be shown that two serious problems may arise:
  - the decreases in q may be too small relative to the lengths of the steps;
  - the steps may be too small relative to the initial rate of decrease of q.
- We can easily construct examples of these two pathologies.


4
3.5
3
2.5
2
1.5
1
0.5
0
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2


4
3.5
3
2.5
2
1.5
1
0.5
0
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

To fix the first problem, we have to impose that the average rate of decrease from $q(x_k)$ to $q(x_{k+1})$ is at least some given fraction of the initial rate of decrease in that direction:

$$q(x_k + \lambda d_k) - q(x_k) \leq \alpha\lambda\nabla q(x_k)d_k$$

where $\alpha \in (0,1)$.

This condition, known as the *(Armijo) sufficient decrease condition*, can be more compactly rewritten as:

$$\phi(\lambda) - \phi(0) \leq \alpha\lambda\phi'(0)$$

where $\phi(z) \equiv q(x_k + zd_k) : R_+ \to R$.

To fix the second problem, we have to impose that the rate of decrease of q at x_{k+1} in the direction d_k is larger of a give fraction of the rate of decrease at x_k in the same direction:

$$\phi'(\lambda) \geq \beta\phi'(0)$$

where $\beta \in (0,1)$ and $\phi'(0) < 0$.

This condition is known as the *curvature condition*. A stronger version is sometimes used:

$$|\phi'(\lambda)| \leq \beta|\phi'(0)|$$

If $\beta > \alpha$, both conditions can be simultaneously satisfied.

Theorem (**Wolfe**) *Let* $q : R^n \to R$ *be* C^1 *function, and let* $d_k \in R^n$ *be a descent direction for* q *in* $x_k \in R^n$ *(i.e. let* $\nabla q(x_k)d_k < 0$*). Suppose that* $\phi(\lambda)$ *is bounded below for all* $\lambda > 0$*. Then there exist two bounds* $\lambda_U > \lambda_L > 0$ *such that* $x_{k+1} = x_k + \lambda d_k$ *satisfies the AGW conditions for all* $\lambda \in (\lambda_L, \lambda_U)$*.*

Theorem (**Wolfe**) *Let* $q : R^n \to R$ *be a* C^1 *function bounded below on* R^n*, and let the gradient* $\nabla q(x)$ *be Lipschitz continuos in the Euclidean norm. Then for any* $x_0 \in R^n$ *there is a sequence* $\{x_k\}_{k=0}^{\infty} \in R^n$ *that satisfies the AGW conditions and either* $\nabla q(x_k)s_k < 0$ *or* $\nabla q(x_k) = 0$ *and* $s_k = 0$ *for each* $k \geq 0$*, where* $s_k \equiv x_{k+1} - x_k$*; furthermore, for any such sequence, either* $\nabla q(x_k) = 0$ *for some* $k \geq 0$*, or:*

$$\lim_{k \to \infty} \frac{\nabla q(x_k)s_k}{\|s_k\|_2} = 0$$

- In other words, line search algorithms based on the Newton step and the AGW rules converge to a zero of F if:
  - ∇q is Lipschitz continuous;
  - $\kappa(J_k)$ is bounded for all $k \geq 0$, i.e. J_k remains "sufficiently" nonsingular;
  - the algorithm does not converge to a local minimizer of q that is not a zero of $F(x)$.
- This is very powerful result: if some mild assumptions on the continuity of F hold, and if q has no "wrong" local minima, line search methods are globally convergent.

Trust-region methods

- Consider the *merit function q(x)* defined as:

$$q(x) = \frac{1}{2} \| f(x) \|_2^2$$

- We construct a model function m_k whose behavior near the current x_k is similar to that of q, i.e. a quadratic approximation of q (using $J'J$ as the approx. Hessian):

$$m_k(p) = \frac{1}{2} \| f(x_k) + J(x_k)p \|_2^2 =$$
$$f_k + p'J_k'f_k + \frac{1}{2} p'J_k'J_k p$$

- We restrict the search for a minimizer of m_k to some region around x_k.

- We find the candidate step p_k by **approximately** solve the following sub-problem:

$$\min_p m_k(p)$$
$$s.t. \|p\| \leq \Delta_k$$

- If J_k has full rank, the **unconstrained** minimizer of m_k is unique, and corresponds to the standard Newton's step:

$$p_k^J = -J_k^{-1} f_k$$

- If the constraint is **binding**, then:

$$p_k = -(J_k' J_k + \mu_c I)^{-1} J_k' f_k$$

for some μ_c such that $\|p_k\|_2 \cong \delta_k$

- If the candidate solution does not produce a sufficient decrease in q, we shrink the *trust region* and solve again.
- If the decrease is more than sufficient, we enlarge the trust region for the next iteration.
- If the decrease is just sufficient, we leave the region as it is.

- This "sufficiency" is evaluated focusing on the ratio between the *actual reduction* and the *predicted reduction*:

$$\rho_k = \frac{q(x_k) - q(x_k + p_k)}{m_k(0) - m_k(p_k)}$$

```
if ρ_k < 1/4
    Δ_{k+1} = 1/4Δ_k
else
    if ρ_k > 3/4 and ||p_k|| = Δ_k
        Δ_{k+1} = min(2Δ_k, Δ̂)
    else
        Δ_{k+1} = Δ_k
    end if
end if
```


- The approximate solution to the previous sub-problem can be computed using different algorithms:
  - The **Dogleg method.**
  - Two-dimensional subspace minimization.
  - The **CG-Steihaug method.**
  - Nearly exact solutions (Moré and Sorensen).
- Trust region algorithms satisfy the AGW conditions, and are therefore **globally convergent**, if the approximated solution obtains at least as much decrease (actually, a fixed factor suffices) in *m* as the **Cauchy point**.

The Cauchy point

- Find the vector that solves a linear version of m_k:

$$p_k^s = \arg\min_{p \in R^n} f_k + p' J_k' f_k$$
$$s.t \ \ \|p\| \leq \Delta_k$$

- The solution to the previous problem is:

$$p_k^s = -\Delta_k \frac{J_k' f_k}{\|J_k' f_k\|}$$

- This vector corresponds to the constrained **steepest descent** direction

- Then, find the scalar τ_k that solves:

$$\tau_k = \arg\min_{\tau>0} m_k(\tau p_k^s)$$
$$s.t \; \|\tau p_k^s\| \leq \Delta_k$$

- The solution is:

$$\tau_k = \min\left\{1, \frac{\|J_k' f_k\|^3}{\Delta_k f_k' J_k (J_k' J_k) J_k' f_k}\right\}$$

- The Cauchy step is defined as: $p_k^c = \tau_k p_k^s$

- In other words, the Cauchy point is the minimizer of m_k in the (constrained) steepest direction

The Dogleg step

- Construct a piece-wise linear function connecting the origin, the Cauchy point, and the unconstrained Newton step.
- Then, choose x_{k+1} on this polygonal arc such that:

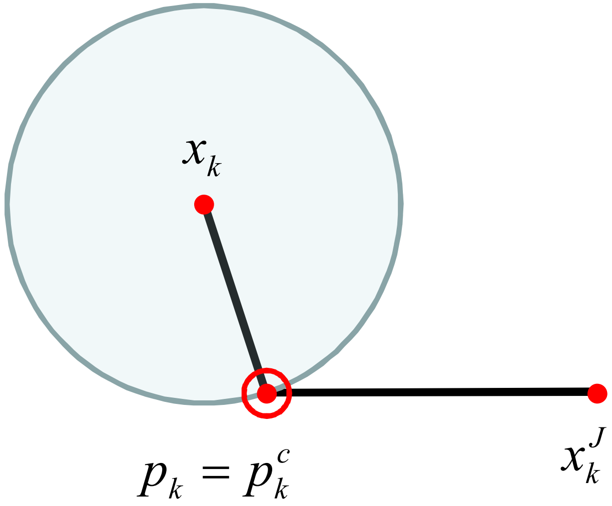
$$\|x_{k+1} - x_k\|_2 = \delta_k$$

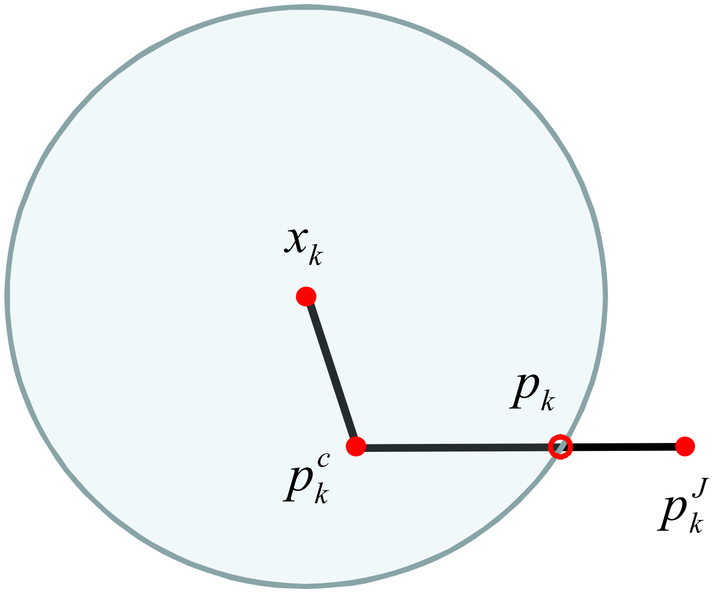
unless:

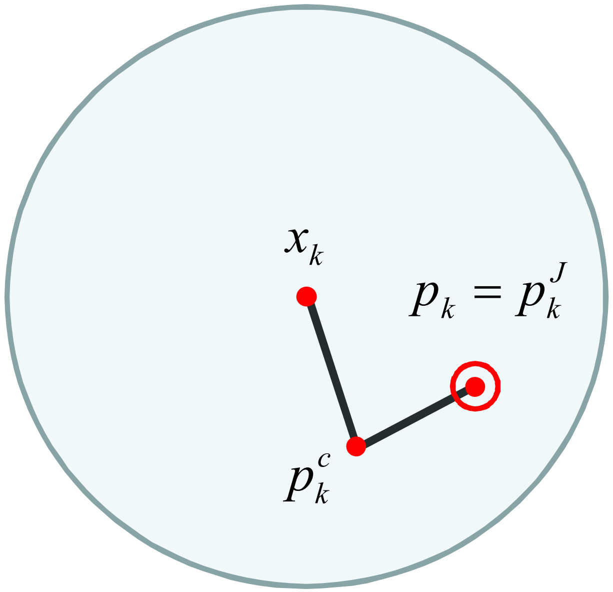
$$\|p_k^J\|_2 \leq \delta_k$$

In this case, use the Newton step.

- It can be shown that m_k decreases monotonically along the dogleg: this guarantees that each step obtains at least the same decrease in m_k than the Cauchy point


x_k
$p_k = p_k^c$
x_k^J


x_k
p_k
p_k^c
p_k^J


x_k
$p_k = p_k^J$
p_k^c