# Numerical Optimization

## Lecture Notes #5
## Line Search Methods; Rate of Convergence

Peter Blomgren,
⟨blomgren.peter@gmail.com⟩

Department of Mathematics and Statistics
Dynamical Systems Group
Computational Sciences Research Center
San Diego State University
San Diego, CA 92182-7720

**http://terminus.sdsu.edu/**

Fall 2018





## Outline







## Quick Recap: Some Recent Discussion

- Discussion on ``Sufficient Decrease'' for line search:
  - Wolfe Conditions
  - Strong Wolfe Conditions
- `Algorithm: Backtracking Line Search`
  - For the **Steepest Descent Direction**
  - For the **Newton Direction**
    - Example #1: $f(\bar{\mathbf{x}}) = (x_1 + x_2^2)^2$
    - Example #2: $f(\bar{\mathbf{x}}) = (x_1 + x_2^2)^2 + 0.5(x_1^2 + x_2^2)$
- `The Zoutendijk Condition`
  - Smooth and bounded (below) $f \Rightarrow$ Steepest Descent direction and Wolfe conditions will give globally convergent method:

$$\lim_{k\to\infty} \|\nabla f(\bar{\mathbf{x}}_k)\| = 0.$$

  Also true for Newton direction if Hessian $\nabla^2 f(\bar{\mathbf{x}})$ is positive definite and the condition number is uniformly bounded.





## Are We Done?

We know the following (for nice enough $f$)

- **If** we ensure that $\bar{\mathbf{p}}_k \not\perp \nabla f(\bar{\mathbf{x}}_k) \Leftrightarrow \cos\theta_k \geq \delta > 0$
  - Compute $\cos\theta_k$ in each iteration, and turn $\bar{\mathbf{p}}_k$ in the steepest descent direction if needed ($\cos\theta_k < \delta$).
- **and** $\alpha_k$ satisfies the Wolfe conditions
  - *E.g.* backtracking line search.
- **then** we have a globally convergent algorithm.

**Therefore** optimization is easy???

## No, We Are Not Done! 

We can perform angle tests ($|\cos\theta_k| > \delta$); and subsequently "turn" $\bar{\mathbf{p}}_k$ to ensure global convergence, **however**

- This may (will!) slow down the convergence rate...
  - When the Hessian is ill-conditioned (close to singular), the appropriate search direction may be almost orthogonal to the gradient, and an unlucky choice of $\delta$ may prevent this.
- They break Quasi-Newton methods (which are very important for LARGE problems)





## No, We Are Not Done! 

Algorithmic strategies for **rapid** convergence is often in direct conflict with the theoretical requirements for **global** convergence.

$\Rightarrow$ **Steepest descent** is the "model citizen" for global convergence, but it is quite slow (we'll show this today).

$\Rightarrow$ **Newton iteration** converges fast for good initial guesses, but the Newton direction may not even be a descent direction "far away" from the solution.

**The Goal:** the best of both worlds — **rapid global convergence**.





## Preview: Coming Up Next...

We take a more careful look at the **rates of convergence** for Steepest Descent, Newton, and Quasi-Newton methods.

We show, using a simple model case, that the resulting rates of convergence are:

— Linear (Steepest Descent)
— Quadratic (Newton)
— Super-Linear (Quasi-Newton)

Finally, we discuss **coordinate descent** methods.

**Note:** The convergence analysis builds on Taylor's theorem, and linear algebra results applied to quadratic forms

$$f(\bar{\mathbf{x}}) = \frac{1}{2}\bar{\mathbf{x}}^T Q\bar{\mathbf{x}} - \bar{\mathbf{b}}^T\bar{\mathbf{x}}, \quad \text{where } Q \text{ is sym. pos. def.}$$





## Convergence Analysis: Steepest Descent 

We apply the steepest descent method to the simple quadratic model objective

$$f(\bar{\mathbf{x}}) = \frac{1}{2}\bar{\mathbf{x}}^T Q\bar{\mathbf{x}} - \bar{\mathbf{b}}^T\bar{\mathbf{x}},$$

where $Q$ is an $n \times n$ symmetric positive matrix. Further, we idealize the method by using **exact line searches**.

**The gradient** is

$$\nabla f(\bar{\mathbf{x}}) = Q\bar{\mathbf{x}} - \bar{\mathbf{b}}, \quad \text{and hence} \quad \bar{\mathbf{x}}^* = Q^{-1}\bar{\mathbf{b}} \quad \text{is unique.}$$

**The step length $\alpha_k$:** Let $\bar{\mathbf{g}}_k = \nabla f(\bar{\mathbf{x}}_k)$, then $\alpha_k$ is the $\alpha$ which minimizes

$$f(\bar{\mathbf{x}}_k - \alpha\bar{\mathbf{g}}_k) = \frac{1}{2}(\bar{\mathbf{x}}_k - \alpha\bar{\mathbf{g}}_k)^T Q(\bar{\mathbf{x}}_k - \alpha\bar{\mathbf{g}}_k) - \bar{\mathbf{b}}^T(\bar{\mathbf{x}}_k - \alpha\bar{\mathbf{g}}_k).$$

## Convergence Analysis: Steepest Descent



Expanding the expression we have:

$$f(\bar{\mathbf{x}}_k - \alpha\bar{\mathbf{g}}_k) = \frac{1}{2}\bar{\mathbf{x}}_k^T Q\bar{\mathbf{x}}_k + \frac{1}{2}\alpha^2\bar{\mathbf{g}}_k^T Q\bar{\mathbf{g}}_k - \frac{1}{2}\alpha\bar{\mathbf{g}}_k^T Q\bar{\mathbf{x}}_k - \frac{1}{2}\alpha\bar{\mathbf{x}}_k^T Q\bar{\mathbf{g}}_k - \bar{\mathbf{b}}^T\bar{\mathbf{x}}_k + \alpha\bar{\mathbf{b}}^T\bar{\mathbf{g}}_k.$$

Then, we differentiate with respect to $\alpha$ and set equal to zero

$$0 = \alpha\,\bar{\mathbf{g}}_k^T Q\bar{\mathbf{g}}_k + \bar{\mathbf{g}}_k^T \underbrace{(-Q\bar{\mathbf{x}}_k + \bar{\mathbf{b}})}_{-\nabla f(\bar{\mathbf{x}}_k)}.$$

Hence,

$$\alpha_k = \frac{\bar{\mathbf{g}}_k^T\bar{\mathbf{g}}_k}{\bar{\mathbf{g}}_k^T Q\bar{\mathbf{g}}_k} = \frac{\nabla f(\bar{\mathbf{x}}_k)^T\nabla f(\bar{\mathbf{x}}_k)}{\nabla f(\bar{\mathbf{x}}_k)^T Q\nabla f(\bar{\mathbf{x}}_k)}.$$

Steepest descent iteration (with exact linesearch)

$$\bar{\mathbf{x}}_{k+1} = \bar{\mathbf{x}}_k - \left[\frac{\nabla f(\bar{\mathbf{x}}_k)^T\nabla f(\bar{\mathbf{x}}_k)}{\nabla f(\bar{\mathbf{x}}_k)^T Q\nabla f(\bar{\mathbf{x}}_k)}\right]\nabla f(\bar{\mathbf{x}}_k)$$





## Convergence Analysis: Steepest Descent



For the model $\nabla f(\bar{\mathbf{x}}_k) = Q\bar{\mathbf{x}}_k - \bar{\mathbf{b}}$ we now have a complete closed form expression for the iterations.

The figure on the next slide shows a typical convergence pattern for steepest descent methods — a zig-zagged approach to the optimum.

In this example the model is

$$f(\bar{\mathbf{x}}) = \frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

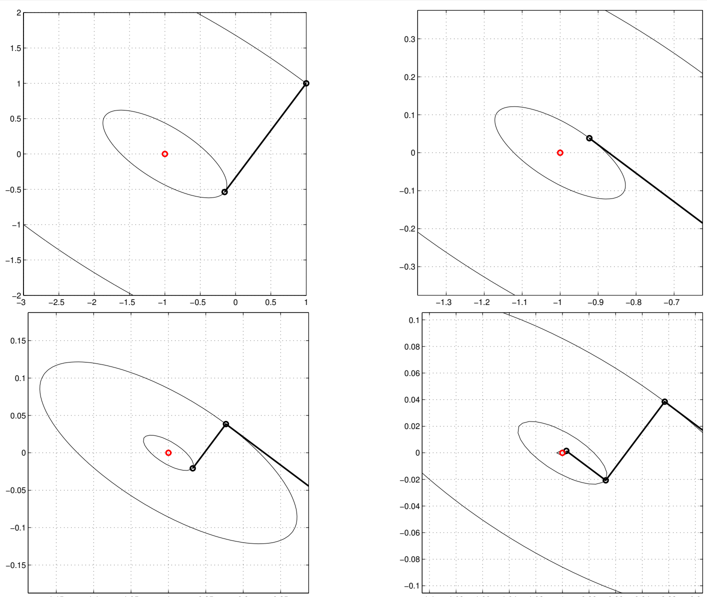
and

$$\bar{\mathbf{x}}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$





## Illustration: Steepest Descent Convergence Pattern







## Convergence Analysis: Steepest Descent



In order to measure the rate of convergence we introduce the weighed $Q$-norm

$$\|\bar{\mathbf{x}}\|_Q^2 = \bar{\mathbf{x}}^T Q\bar{\mathbf{x}}.$$

Since $Q\bar{\mathbf{x}}^* = \bar{\mathbf{b}}$, we have

$$\frac{1}{2}\|\bar{\mathbf{x}} - \bar{\mathbf{x}}^*\|_Q^2 = f(\bar{\mathbf{x}}) - f(\bar{\mathbf{x}}^*),$$

and since $\nabla f(\bar{\mathbf{x}}^*) = 0$, we note that $\nabla f(\bar{\mathbf{x}}_k) = Q(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}^*)$. We can now express the iteration in terms of the $Q$-norm:

$$\|\bar{\mathbf{x}}_{k+1} - \bar{\mathbf{x}}^*\|_Q^2 = \left[1 - \frac{\nabla f(\bar{\mathbf{x}}_k)^T\nabla f(\bar{\mathbf{x}}_k)}{(\nabla f(\bar{\mathbf{x}}_k)^T Q\nabla f(\bar{\mathbf{x}}_k))(\nabla f(\bar{\mathbf{x}}_k)^T Q^{-1}\nabla f(\bar{\mathbf{x}}_k))}\right]\|\bar{\mathbf{x}}_k - \bar{\mathbf{x}}^*\|_Q^2.$$

The details are outlined in exercise 3.7 (NW$^{1st}$ p.62, NW$^{2nd}$ p.64).

This is an exact expression of the decrease in each iteration. It is however, quite cumbersome to work with, a more useful bound can be found...

## Convergence Analysis: Steepest Descent



**Theorem**

*When the steepest descent method with exact line searches is applied to the convex quadratic function*

$$f(\bar{\mathbf{x}}) = \frac{1}{2}\bar{\mathbf{x}}^T Q \bar{\mathbf{x}} - \bar{\mathbf{b}}^T \bar{\mathbf{x}},$$

*the error norm*

$$\frac{1}{2}\|\bar{\mathbf{x}} - \bar{\mathbf{x}}^*\|_Q^2 = f(\bar{\mathbf{x}}) - f(\bar{\mathbf{x}}^*),$$

*satisfies*

$$\|\bar{\mathbf{x}}_{\mathbf{k+1}} - \bar{\mathbf{x}}^*\|_{\mathbf{Q}}^{\mathbf{2}} \leq \left[\frac{\lambda_{\mathbf{n}} - \lambda_{\mathbf{1}}}{\lambda_{\mathbf{n}} + \lambda_{\mathbf{1}}}\right]^{\mathbf{2}} \|\bar{\mathbf{x}}_{\mathbf{k}} - \bar{\mathbf{x}}^*\|_{\mathbf{Q}}^{\mathbf{2}},$$

*where $0 < \lambda_1 \leq \cdots \leq \lambda_n$ are the eigenvalues of $Q$.*





## Convergence Analysis: Steepest Descent



The theorem shows a **linear rate of convergence**

$$\|\bar{\mathbf{x}}_{\mathbf{k+1}} - \bar{\mathbf{x}}^*\|_{\mathbf{Q}} \leq \left|\frac{\lambda_{\mathbf{n}} - \lambda_{\mathbf{1}}}{\lambda_{\mathbf{n}} + \lambda_{\mathbf{1}}}\right| \|\bar{\mathbf{x}}_{\mathbf{k}} - \bar{\mathbf{x}}^*\|_{\mathbf{Q}}.$$

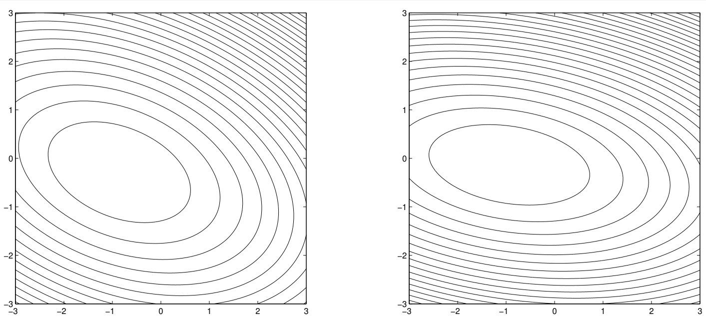
If $\lambda_n = \lambda_1$ then only one iteration is needed — in this case $Q$ is a multiple of the identity matrix, and the contours are concentric circles which means that the steepest descent direction points straight at the solution.

As the **condition number** $\kappa(Q) = \lambda_n/\lambda_1$ increases the contours (in the $\bar{\mathbf{e}}_n \times \bar{\mathbf{e}}_1$ plane) become more elongated, which increases the amount of zig-zagging. The ratio $\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}$ approaches one, which shows a significant slow-down in the convergence.





## Illustration: Contours



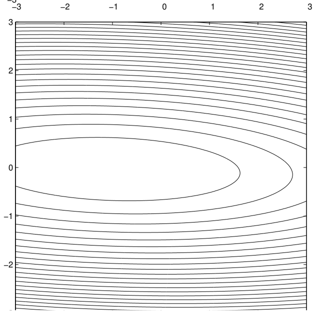


**Figure:** The contour plots for $\lambda_n/\lambda_1 = 2$, $\lambda_n/\lambda_1 = 4$, and $\lambda_n/\lambda_1 = 16$. As the contours get more stretched the steepest descent method will increase the amount of zigzagging.





## Convergence Analysis: Steepest Descent



**Theorem (Generalization to general nonlinear objective functions)**

*Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, and that the iterates generated by the steepest-descent method with exact line searches converge to a point $\bar{\mathbf{x}}^*$ where the Hessian matrix $\nabla^2 f(\bar{\mathbf{x}}^*)$ is positive definite. Let $r$ be any scalar satisfying*

$$r \in \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}, 1\right),$$

*where $\lambda_1 \leq \cdots \leq \lambda_n$ are the eigenvalues of $\nabla^2 f(\bar{\mathbf{x}}^*)$. Then for all $k$ sufficiently large, we have*

$$f(\bar{\mathbf{x}}_{k+1}) - f(\bar{\mathbf{x}}^*) \leq r^2 \Big[ f(\bar{\mathbf{x}}_k) - f(\bar{\mathbf{x}}^*) \Big].$$

## Convergence Analysis: Steepest Descent — Notes

The statement of the theorem is different from the statement in Nocedal-Wright (1st edition, $\leq$ 3rd printing); it has been updated according to Nocedal's posted errata:
(http://www.ece.northwestern.edu/∼nocedal/book/errata.html)

**Inexact line searches** will not improve the convergence rate; hence the theorem shows that convergence can be quite slow even when $\kappa(\nabla^2 f(\bar{\mathbf{x}}^*))$ is quite small.

*E.g.* when $\kappa(\nabla^2 f(\bar{\mathbf{x}}^*)) = 100$, 100 iterations will reduce the error by a multiplicative factor of 0.0183... or roughly 57 iterations per digit.





## Illustration: Slow Convergence for Steepest Descent

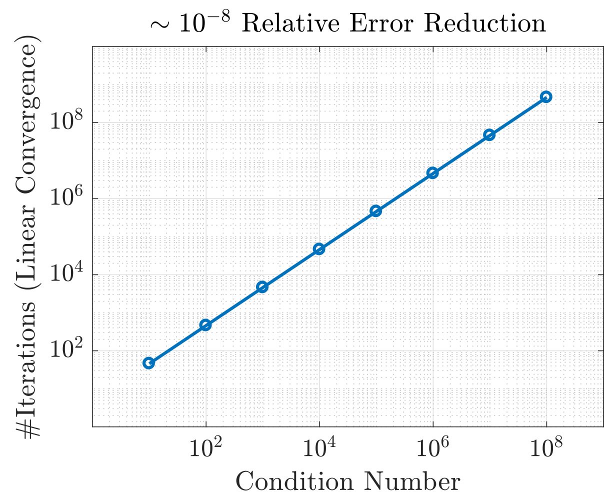



**Figure:** LINEAR CONVERGENCE — Estimated number of iterations needed to reduce the error by a factor of $10^{-8}$.





## Convergence: Newton — 1 of 3

We now look at the Newton search direction

$$\bar{\mathbf{p}}_k^N = -\left[\nabla^2 f(\bar{\mathbf{x}}_k)\right]^{-1} \nabla f(\bar{\mathbf{x}}_k).$$

This may not always be a descent direction since the Hessian matrix $\nabla^2 f(\bar{\mathbf{x}}_k)$ may not always be positive definite.

We delay the discussion of how to deal with non-descent Newton directions until a later time.

For now we focus on the local result, *i.e.* we start the iteration close enough to the optimum $\bar{\mathbf{x}}^*$ that all Hessians along the solution trajectory are positive definite.





## Convergence: Newton — 2 of 3

**Theorem**

*Suppose that $f$ is twice differentiable and that the Hessian $\nabla^2 f(\bar{\mathbf{x}})$ is Lipschitz continuous in a neighborhood of the solution $\bar{\mathbf{x}}^*$ at which the sufficient conditions $\nabla f(\bar{\mathbf{x}}^*) = 0$ and $\nabla^2 f(\bar{\mathbf{x}}^*)$ is positive definite are satisfied. Consider the Newton method, $\bar{\mathbf{x}}_{k+1} = \bar{\mathbf{x}}_k + \bar{\mathbf{p}}_k^N$. Then,*

1. *if the starting point $\bar{\mathbf{x}}_0$ is sufficiently close to $\bar{\mathbf{x}}^*$, the sequence of iterates converges to $\bar{\mathbf{x}}^*$.*
2. *the rate of convergence of $\{\bar{\mathbf{x}}_k\}$ is quadratic.*
3. *the sequence of gradient norms $\{\|\nabla f(\bar{\mathbf{x}}_k)\|\}$ converges quadratically to zero.*

The proof is in (NW[1st] pp.52–53, NW[2nd] pp.45.)

## Convergence: Newton — Illustration — 3 of 3

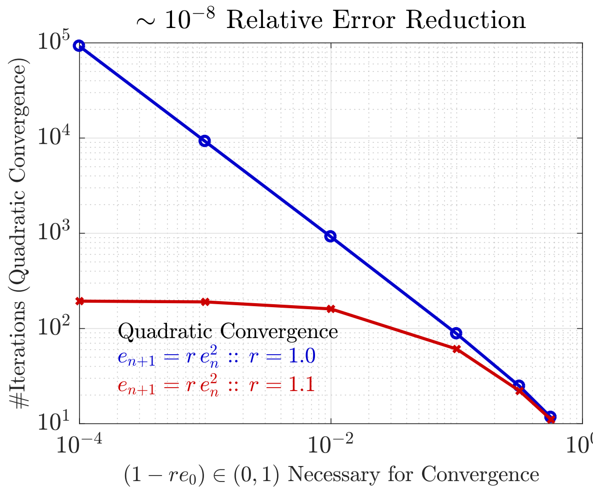



**Figure:** Quadratic Convergence — Estimated number of iterations needed to reduce the error by a factor of $10^{-8}$. At a minimum we must have $|re_0| < 1$ in order to get convergence.





## Convergence: Quasi-Newton — 1 of 4

We now turn our attention to the "middle case," where we have a search direction of the form

$$\bar{\mathbf{p}}_k = -H_k^{-1} \nabla f(\bar{\mathbf{x}}_k)$$

where $H_k$ is symmetric positive definite — and a clever approximation to the Hessian $\nabla^2 f(\bar{\mathbf{x}}_k)$ — the discussion on how to build and update $H_k$ is postponed until later.

Further, we assume that the iteration is based on an inexact line search algorithm, where $\alpha_k$ satisfies the Wolfe conditions$^{(*)}$. We also assume that the **line search algorithm always tries the step** $\alpha = 1$ **first.**

$^{(*)}$ This rules out the use of the backtracking algorithm, which may occasionally violate the 2nd Wolfe condition.





## Convergence: Quasi-Newton — 2 of 4

Theorem

*Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. Consider the iteration $\bar{\mathbf{x}}_{k+1} = \bar{\mathbf{x}}_k + \alpha_k \bar{\mathbf{p}}_k$, where $\bar{\mathbf{p}}_k$ is a descent direction and $\alpha_k$ satisfies the Wolfe conditions with $c_1 \leq 1/2$. If the sequence $\{\bar{\mathbf{x}}_k\}$ converges to a point $\bar{\mathbf{x}}^*$ such that $\nabla f(\bar{\mathbf{x}}^*) = 0$ and $\nabla^2 f(\bar{\mathbf{x}}^*)$ is positive define, and if the search direction satisfies*

$$\lim_{k \to \infty} \frac{\|\nabla f(\bar{\mathbf{x}}_k) + \nabla^2 f(\bar{\mathbf{x}}_k) \bar{\mathbf{p}}_k\|}{\|\bar{\mathbf{p}}_k\|} = 0,$$

*then*

1. *the step length $\alpha_k = 1$ is admissible for all $k > k_0$, and*
2. *$\{\bar{\mathbf{x}}_k\}$ converges to $\bar{\mathbf{x}}^*$* **super-linearly**.

**Note:** Statement of theorem updated according to errata.





## Convergence: Quasi-Newton — 3 of 4

If $c_1 > 1/2$ the line search would exclude the minimizer of a quadratic, forcing $\alpha_k < 1$.

If $\bar{\mathbf{p}}_k$ is a quasi-Newton search direction, then the limit in the theorem is equivalent to

$$\lim_{k \to \infty} \frac{\|(H_k - \nabla^2 f(\bar{\mathbf{x}}^*)) \bar{\mathbf{p}}_k\|}{\|\bar{\mathbf{p}}_k\|} = 0.$$

This shows that we can achieve super-linear convergence even if

$$\lim_{k \to \infty} H_k \neq \nabla^2 f(\bar{\mathbf{x}}^*),$$

*i.e.* it is sufficient that $H_k$ converges to the Hessian **in the search directions $\bar{\mathbf{p}}_k$.**

It turns out that this condition is both necessary and sufficient.

## Convergence: Quasi-Newton



**Theorem**

*Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. Consider the iteration $\bar{\mathbf{x}}_{k+1} = \bar{\mathbf{x}}_k + \bar{\mathbf{p}}_k$ (i.e. $\alpha_k \equiv 1$) and that $\bar{\mathbf{p}}_k$ is given by $\bar{\mathbf{p}}_k = -H_k^{-1}\nabla f(\bar{\mathbf{x}}_k)$.* **If** *$\{\bar{\mathbf{x}}_k\}$ converges to a point $\bar{\mathbf{x}}^*$ such that $\nabla f(\bar{\mathbf{x}}^*) = 0$ and $\nabla^2 f(\bar{\mathbf{x}}^*)$ is positive define,* **then** *$\{\bar{\mathbf{x}}_k\}$ converges super-linearly if and only if*

$$\lim_{k\to\infty} \frac{\|(H_k - \nabla^2 f(\bar{\mathbf{x}}_k))\bar{\mathbf{p}}_k\|}{\|\bar{\mathbf{p}}_k\|} = 0$$

*holds.*

When we return to the construction of quasi-Newton methods, we will see that satisfying the condition of this theorem is normally not a problem, hence super-linearly convergent quasi-Newton methods are readily available for most objective functions.

**Note:** Statement of theorem updated according to errata.





## Coordinate Descent Methods



Instead of **computing** the search direction, why not just cycle through the coordinates? — *i.e.*

$$\bar{\mathbf{p}}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}^T, \quad \bar{\mathbf{p}}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}^T, \quad \bar{\mathbf{p}}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}^T, \quad \cdots$$

Once we reach the final direction $\bar{\mathbf{p}}_n$, we start over from $\bar{\mathbf{p}}_1$.

Unfortunately, this scheme is quite inefficient in practice, and can in fact iterate infinitely$^{(*)}$ without reaching a stationary point.





## Coordinate Descent Methods



A cyclic search along **any** set of linearly independent directions can run into this problem of non-convergence.

The gradient $\nabla f(\bar{\mathbf{x}}_k)$ may become more and more perpendicular to the coordinate search direction, so that $\cos\theta_k$ approaches zeros rapidly enough that the Zoutendijk condition

$$\sum_{k=0}^{\infty} \cos^2\theta_k \|\nabla f(\bar{\mathbf{x}}_k)\|^2 < \infty,$$

is satisfied even though $\|\nabla f(\bar{\mathbf{x}}_k)\| \not\to 0$.





## Coordinate Descent Methods (CDMs)



Even when coordinate descent methods converge, the rate of convergence is slower than that of the steepest descent method.

The slowness increases as the number of variables increases.

There are however situations in which coordinate descent may be useful since

1. no calculation of $\nabla f(\bar{\mathbf{x}}_k)$ is required.
2. convergence can be acceptably fast if the variables are loosely coupled — the stronger the coupling, the worse the convergence.
3. it is embarrassingly easy$^{(!)}$ to parallelize CDMs.

# Index