Math 273a: Optimization
Netwon's methods

Instructor: Wotao Yin
Department of Mathematics, UCLA

Fall 2015

some material taken from Chong-Zak, 4th Ed.

Main features of Newton's method

- Uses both first derivatives (gradients) and second derivatives (Hessian)
- Based on local quadratic approximations to the objective function
- Requires a positive definite Hessian to work
- Converges very quickly near the solution (under conditions)
- Require a lot of work at each iteration:
  - forming the Hessian
  - inverting or factorizing the (approximate) Hessian

Basic idea

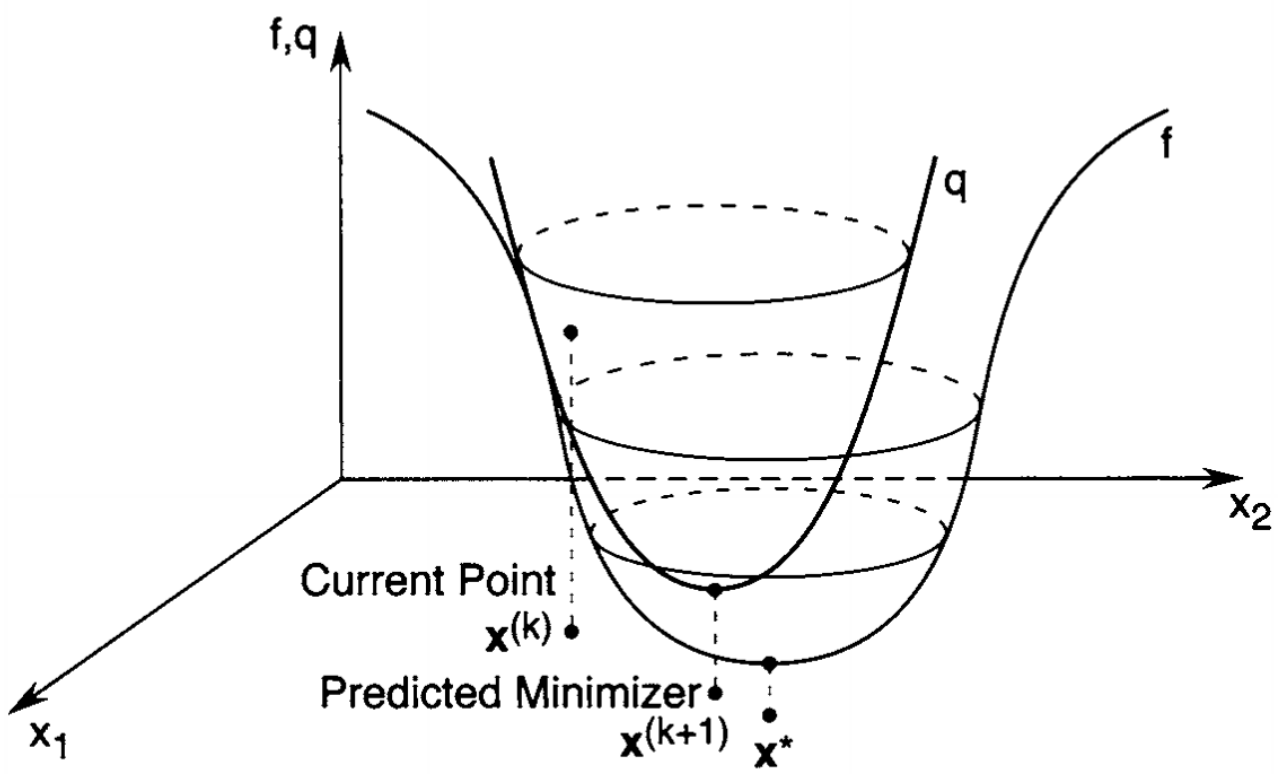
Given the current point $\boldsymbol{x}^{(k)}$

- construct a quadratic function (known as the quadratic approximation) to the objective function that matches the value and both the first and second derivatives at $\boldsymbol{x}^{(k)}$
- minimize the quadratic function instead of the original objective function
- set the minimizer as $\boldsymbol{x}^{(k+1)}$

Note: a new quadratic approximation will be constructed at $\boldsymbol{x}^{(k+1)}$

Special case: the objective is quadratic

- the approximation is exact and the method returns a solution in one step


f,q
q
f
x_2
x_1
Current Point
$\mathbf{x}^{(k)}$
Predicted Minimizer
$\mathbf{x}^{(k+1)}$
$\mathbf{x}^*$

Quadratic approximation

- Assumption: function $f \in \mathcal{C}^2$, i.e., twice continuously differentiable
- Apply Taylor's expansion, keep first three terms, drop terms of order ≥ 3

  $$f(\boldsymbol{x}) \approx q(\boldsymbol{x}) := f(\boldsymbol{x}^{(k)}) + \boldsymbol{g}^{(k)T}(\boldsymbol{x} - \boldsymbol{x}^{(k)}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^{(k)})^T \boldsymbol{F}(\boldsymbol{x}^{(k)})(\boldsymbol{x} - \boldsymbol{x}^{(k)})$$

  where
  - $\boldsymbol{g}^{(k)} := \nabla \boldsymbol{f}(\boldsymbol{x}^{(k)})$ is the gradient at $\boldsymbol{x}^{(k)}$
  - $\boldsymbol{F}(\boldsymbol{x}^{(k)}) := \nabla^2 \boldsymbol{f}(\boldsymbol{x}^{(k)})$ is the Hessian at $\boldsymbol{x}^{(k)}$

Generating the next point

- Minimizing $q(\boldsymbol{x})$ by apply the first-order necessary condition:

$$\mathbf{0} = \nabla q(\boldsymbol{x}) = \boldsymbol{g}^{(k)} + \boldsymbol{F}(\boldsymbol{x}^{(k)})(\boldsymbol{x} - \boldsymbol{x}^{(k)}).$$

- If $\boldsymbol{F}(\boldsymbol{x}^{(k)}) \succ 0$ (positive definite), then q achieves its unique minimizer at

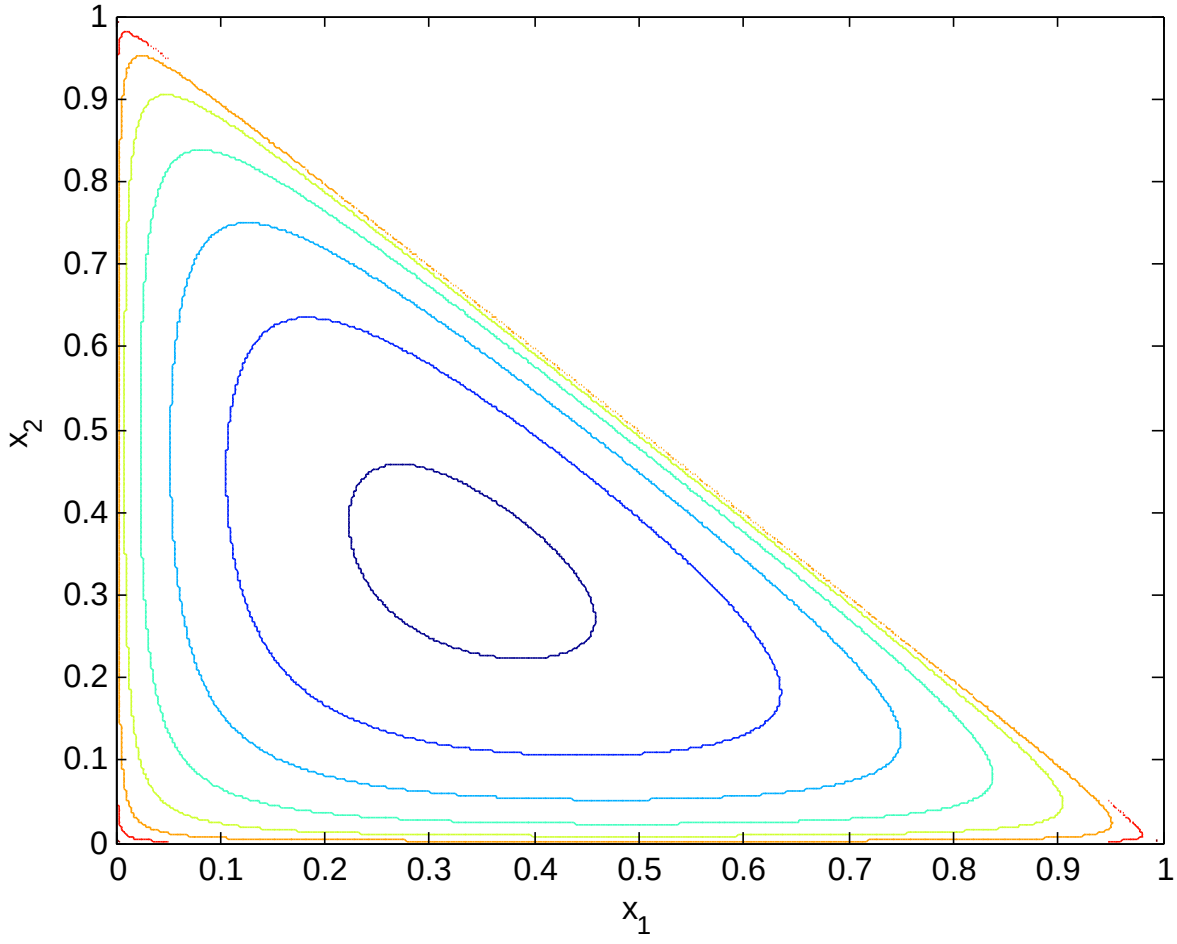
$$\boldsymbol{x}^{(k+1)} := \boldsymbol{x}^{(k)} - \boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}^{(k)}.$$

  We have $\mathbf{0} = \nabla q(\boldsymbol{x}^{(k+1)})$

- Can be viewed an iteration for solving $\boldsymbol{g}(\boldsymbol{x}) = \mathbf{0}$ using its Jacobian $\boldsymbol{F}(\mathbf{x})$.

Example

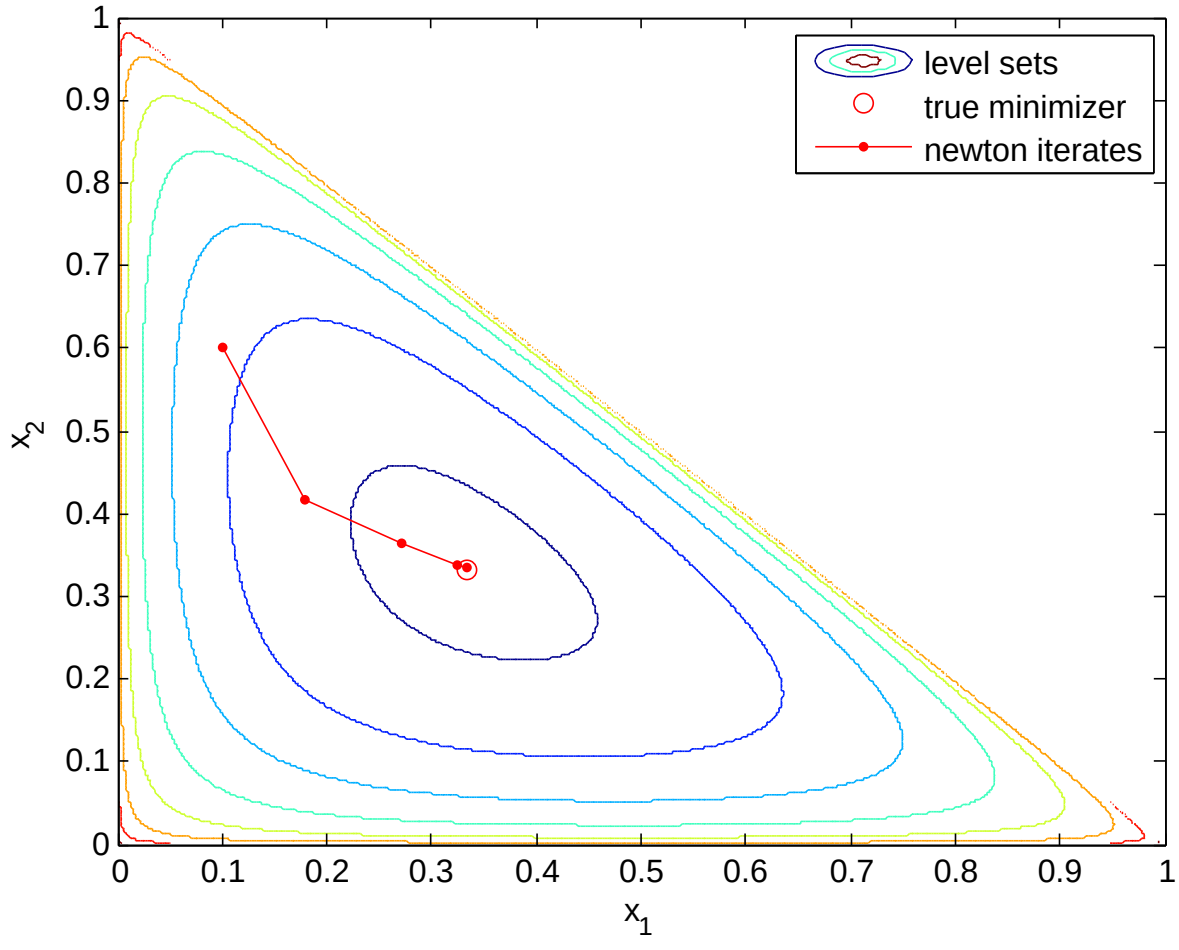
$$f(x_1, x_2) = -\log(1 - x_1 - x_2) - \log(x_1) - \log(x_2)$$



a

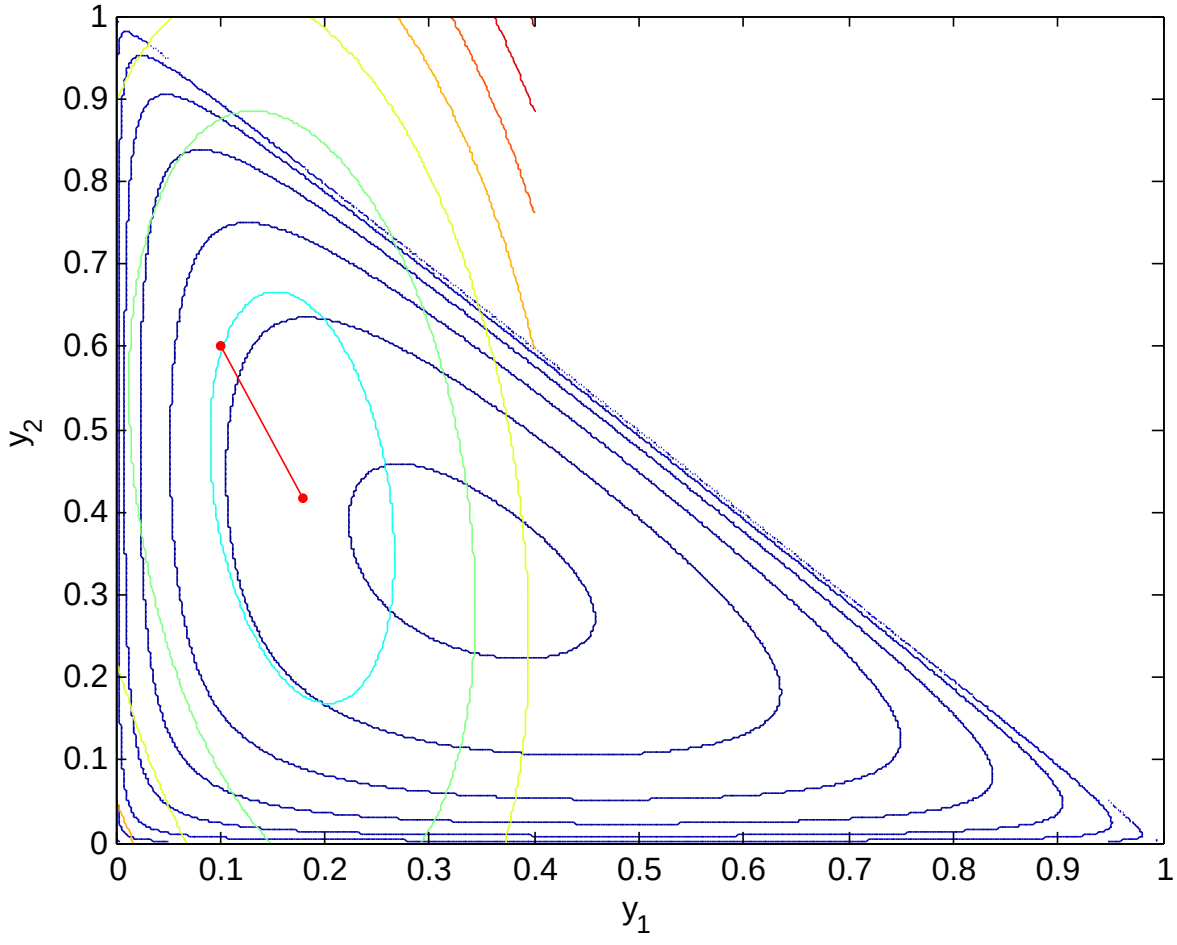
Example

Start Newton's method from $[\frac{1}{10}; \frac{6}{10}]$



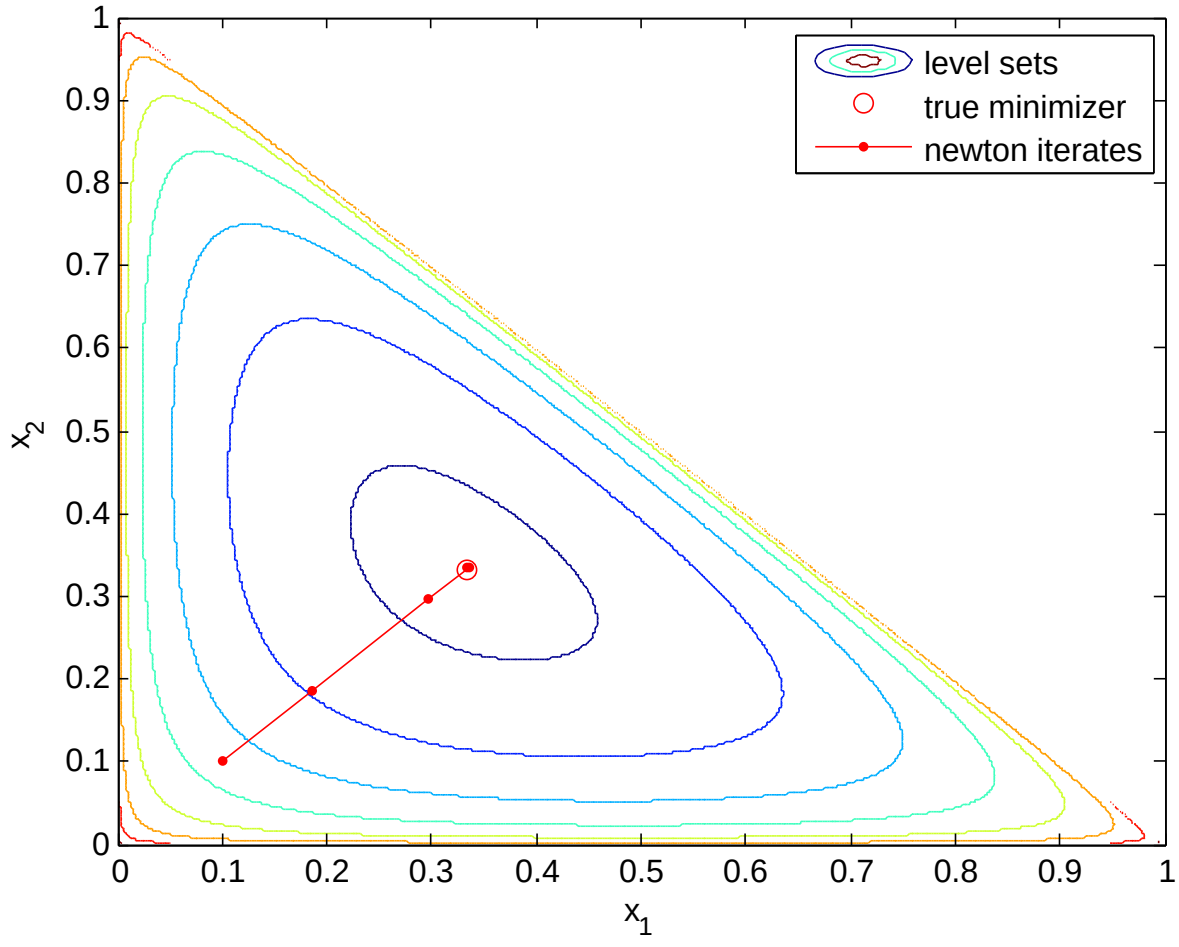
Example

$f(x_1, x_2)$ and its quadratic approximation $q(x_1, x_2)$ at $[\frac{1}{10}; \frac{6}{10}]$ share the same value, gradient and Hessian at $[\frac{1}{10}; \frac{6}{10}]$. The new point minimizes $q(x_1, x_2)$



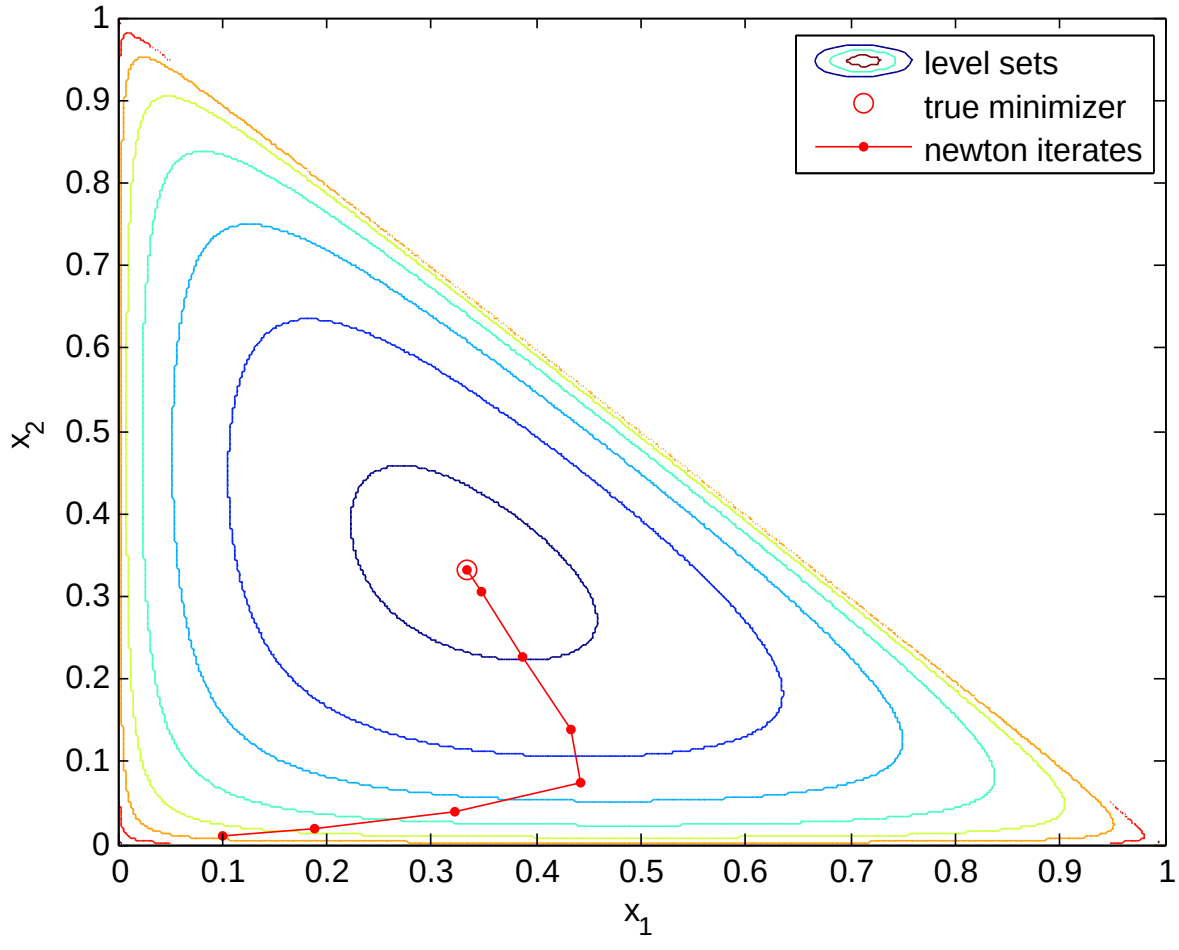
Example

Start Newton's method from $[\frac{1}{10}; \frac{1}{10}]$



Example

Start Newton's method from $[\frac{1}{10}; \frac{1}{100}]$



Aware of the drawbacks

Unlike the move along $-\nabla f(\boldsymbol{x}^{(k)})$, where a sufficiently small step size guarantees the objective decrease, Newton's method jumps to a potentially distant point. This makes it vulnerable.

- recall in 1D, $f'' < 0$ can cause divergence
- even if $\boldsymbol{F}(\boldsymbol{x}^{(k)}) \succ 0$ (positive definite), objective descent is *not* guaranteed

Nonetheless, Newton's method has superior performance when starting near the solution.

Analysis: quadratic function minimization

- The objective function

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

  Assumption: $\boldsymbol{Q}$ is symmetric and invertible

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}$$
$$\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{Q}.$$

- First-order optimality condition $\boldsymbol{g}(\boldsymbol{x}^*) = \boldsymbol{Q}\boldsymbol{x}^* - \boldsymbol{b} = \boldsymbol{0}$. So, $\boldsymbol{x}^* = \boldsymbol{Q}^{-1}\boldsymbol{b}$.
- Given any initial point $\boldsymbol{x}^{(0)}$, by Newton's method

$$\begin{aligned}\boldsymbol{x}^{(1)} &= \boldsymbol{x}^{(0)} - \boldsymbol{F}(\boldsymbol{x}^{(0)})^{-1}\boldsymbol{g}^{(0)} \\ &= \boldsymbol{x}^{(0)} - \boldsymbol{Q}^{-1}(\boldsymbol{Q}\boldsymbol{x}^{(0)} - \boldsymbol{b}) \\ &= \boldsymbol{Q}^{-1}\boldsymbol{b} \\ &= \boldsymbol{x}^*.\end{aligned}$$

  The solution is obtained in one step.

Analysis: how fast is Newton's method? (assumption: Lipschitz continuous Hessian near x^*)

- Let $\boldsymbol{e}^{(k)} := \boldsymbol{x}^{(k)} - \boldsymbol{x}^*$

Theorem

Suppose $f \in \mathcal{C}^2$, $\boldsymbol{F}$ is Lipschitz continuous near $\boldsymbol{x}^*$, and $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$. If $\boldsymbol{x}^{(k)}$ is sufficiently close to $\boldsymbol{x}^*$ and $\boldsymbol{F}(\boldsymbol{x}^*) \succ 0$, then $\exists C > 0$ such that

$$\|\boldsymbol{e}^{(j+1)}\| \leq C\|\boldsymbol{e}^{(j)}\|^2, \quad j = k, k+1, \ldots.$$

Just a sketch proof.

Since $\boldsymbol{F}$ is Lipschitz around $\boldsymbol{x}^*$ and $\boldsymbol{F}(\boldsymbol{x}^*) \succ 0$, we can have

- $(\boldsymbol{F}(\boldsymbol{x}) - c\boldsymbol{I}) \succ 0$ for some $c > 0$ for all $\boldsymbol{x}$ in a small neighborhood of $\boldsymbol{x}^*$.
- Thus, $\|\boldsymbol{F}(\boldsymbol{x})^{-1}\| < c^{-1}$ for all $\boldsymbol{x}$ in the neighborhood.

$\boldsymbol{e}^{(j+1)} = \boldsymbol{e}^{(j)} + \boldsymbol{d}^{(j)}$ where $\boldsymbol{d}^{(j)} = -\boldsymbol{F}(\boldsymbol{x}^{(j)})^{-1}\boldsymbol{g}^{(j)}$.

Taylor expansion near $\boldsymbol{x}^{(j)}$: $\boldsymbol{0} = \boldsymbol{g}(\boldsymbol{x}^*) = \boldsymbol{g}(\boldsymbol{x}^{(j)}) - \boldsymbol{F}(\boldsymbol{x}^{(j)})\boldsymbol{e}^{(j)} + O(\|\boldsymbol{e}^{(j)}\|^2)$.

Thus $\boldsymbol{e}^{(j+1)} = \boldsymbol{e}^{(j)} + \boldsymbol{d}^{(j)} = \boldsymbol{F}(\boldsymbol{x}^{(j)})^{-1}O(\|\boldsymbol{e}^{(j)}\|^2) = O(\|\boldsymbol{e}^{(j)}\|^2)$. Argue that $\boldsymbol{x}^{(j+1)}, \boldsymbol{x}^{(j+2)}, \ldots$ stay in the neighborhood. □

Asymptotic rates of convergence

Suppose sequence $\{\boldsymbol{x}^k\}$ converges to $\bar{\boldsymbol{x}}$. Perform the ratio test

$$\lim_{k\to\infty} \frac{\|\boldsymbol{x}^{k+1} - \bar{\boldsymbol{x}}\|}{\|\boldsymbol{x}^k - \bar{\boldsymbol{x}}\|} = \mu.$$

- if $\mu = 1$, then $\{\boldsymbol{x}^k\}$ converges **sublinearly**.
- if $\mu \in (0, 1)$, then $\{\boldsymbol{x}^k\}$ converges **linearly**;
- if $\mu = 0$, then $\{\boldsymbol{x}^k\}$ converges **superlinearly**;

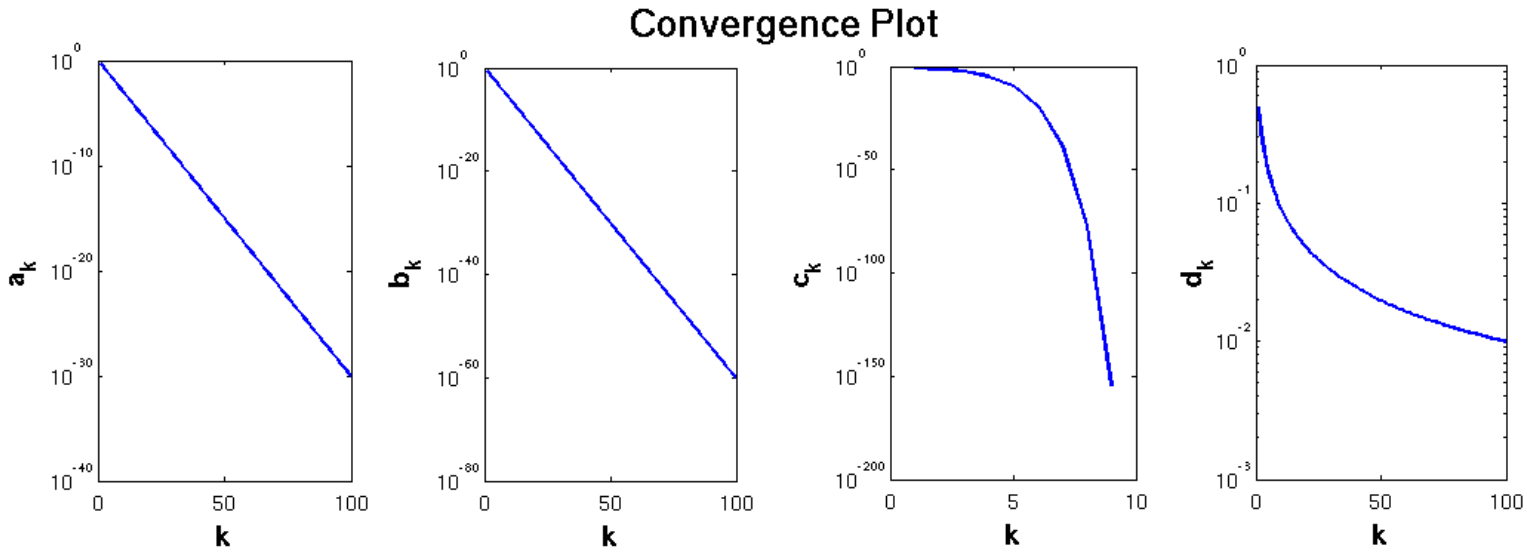
To distinguish superlinear rates of convergence, we check

$$\lim_{k\to\infty} \frac{\|\boldsymbol{x}^{k+1} - \bar{\boldsymbol{x}}\|}{\|\boldsymbol{x}^k - \bar{\boldsymbol{x}}\|^q} = \mu > 0$$

- if $q = 2$, it is **quadratic convergence**;
- if $q = 3$, it is **cubic convergence**;
- q can be non-integer, e.g., 1.618 for the secant method ...

Example

- $a_k = 1/2^k$
- $b_k = 1/4^{\lfloor k/2 \rfloor}$
- $c_k = 1/2^{2^k}$
- $d_k = 1/(k+1)$



"semilog*y*" plots (wikipedia)

Another example

Let $C = 100$.

k	$(1/k^2)$	Ce^{-k}	$Ce^{-k^{1.618}}$	Ce^{-k^2}
1	1.0e0	3.7e2	3.7e2	3.7e2
3	3.3e-1	5.0e1	2.7e0	1.2e-1
5	2.0e-1	6.7e-0	1.3e-3	1.4e-8
7	1.4e-1	9.1e-1	7.6e-8	5.2e-19
9	1.1e-1	1.2e-1	6.4e-13	6.6e-33

Comments:

- the constant C is not important in superlinear convergence
- even with a big C, higher-order convergence will quickly catch up
- the constant C is more important in lower-order convergence
- superlinear convergence is shockingly fast!

Analysis: descent direction

Theorem

If the Hessian $\boldsymbol{F}(\boldsymbol{x}^{(k)}) \succ 0$ *(positive definite) and* $\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) \neq \boldsymbol{0}$*, then the search direction*

$$\boldsymbol{d}^{(k)} = -\boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}^{(k)}$$

is a descent direction, that is, there exists $\bar{\alpha} > 0$ *such that*

$$f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}) < f(\boldsymbol{x}^{(k)}), \quad \forall \alpha \in (0, \bar{\alpha}).$$

Proof.

Let $\phi(\alpha) := f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)})$. Then $\phi'(\alpha) := \nabla f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)})^T \boldsymbol{d}^{(k)}$. Since $\boldsymbol{F}(\boldsymbol{x}^{(k)}) \succ 0$ and $\boldsymbol{g}^{(k)} \neq \boldsymbol{0}$, we have $\boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1} \succ 0$

$$\phi'(0) := \nabla f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)})^T \boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)T} \boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1} \boldsymbol{g}^{(k)} < 0.$$

Finally, apply first-order Taylor expansion to $\phi(\alpha)$ to get the result. □

Two more issues with Newton's method

Hessian evaluation:

- When the dimension n is large, obtain $\boldsymbol{F}(\boldsymbol{x}^{(k)})$ can be computationally expensive
- We will study quasi-Newton methods to alleviate this difficulty (in a future lecture)

Indefinite Hessian:

- When the Hessian is not positive definite, the direction is not necessarily descending.
- There are simple modifications.

Modified Newton's method

Strategy:

- use $\boldsymbol{F}(\boldsymbol{x}^{(k)})$ if $\boldsymbol{F}(\boldsymbol{x}^{(k)}) \succ \mathbf{0}$ and $\lambda_{\min}(\boldsymbol{F}(\boldsymbol{x}^{(k)})) > \epsilon$; *otherwise*,
- use $\hat{\boldsymbol{F}}(\boldsymbol{x}^{(k)}) = \boldsymbol{F}(\boldsymbol{x}^{(k)}) + E$ so that $\hat{\boldsymbol{F}}(\boldsymbol{x}^{(k)}) \succ \mathbf{0}$ and $\lambda_{\min}(\hat{\boldsymbol{F}}(\boldsymbol{x}^{(k)})) > \epsilon$.

Method 1 (Greenstadt): replace any tiny or negative eigenvalues by

$$\delta = \max\{\epsilon_{\text{machine}}, \epsilon_{\text{machine}} \|H\|_\infty\}$$

where $\|H\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |h_{ij}|$. This is computationally expensive.

Method 2 (Levenberg-Marquardt): Was proposed for least-squares but works here. Replace

$$\hat{\boldsymbol{F}} \leftarrow \boldsymbol{F} + \gamma I.$$

It shifts *every eigenvalue* of $\boldsymbol{F}$ up by γ.

Modified Newton's method

Method 3 (advanced topic: modified Cholesky / Gill-Murray):

Any symmetric matrix $\boldsymbol{A} \succ 0$ can be factored as

$$\boldsymbol{A} = \bar{\boldsymbol{L}}\bar{\boldsymbol{L}}^T \quad \text{or} \quad \boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T,$$

where $\boldsymbol{L}$ and $\bar{\boldsymbol{L}}$ are lower triangular, $\boldsymbol{D}$ is positive diagonal, and $\boldsymbol{L}$ has ones on its main diagonal.

Properties of the Cholesky factorization:

- Very useful in solving linear systems of equations. Reduces a system to two backsolves.
- If $\boldsymbol{A} \not\succ 0$ (indefinite but still symmetric), $\boldsymbol{D}$ has zero or negative element(s) on its diagonal.

Modified Newton's method

- The factorization is **stable** if $\boldsymbol{A}$ is positive definite. (Small errors in $\boldsymbol{A}$ will not cause large errors in $\boldsymbol{L}$ or $\boldsymbol{D}$.) Example:

$$\begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix} = \begin{bmatrix} 1 & \\ a & 1 \end{bmatrix} \begin{bmatrix} 1 & \\ & 1-a^2 \end{bmatrix} \begin{bmatrix} 1 & a \\ & 1 \end{bmatrix}.$$

- If $\boldsymbol{A} \leftarrow \boldsymbol{A} + \mathbf{v}\mathbf{v}^T$, the factorization can be updated to a product form (avoiding the factorization from scratch, which is more expensive).
- If $\boldsymbol{A}$ is sparse, Cholesky with pivots keeps $\boldsymbol{L}$ sparse with moderately more zeros
- The cost is $n^3/6 + O(n^2)$, roughly half of Gaussian elimination.

Modified Newton's method

Forsgren, Gill, Murray: perform *pivoted* Cholesky factorization. That is, permute the matrix at each step to pull the largest remaining diagonal element to the pivot position

The effect: postpone the modification and keeps it as small as possible.

When no acceptable element remains

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_{11} & \\ \boldsymbol{L}_{21} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{D}_1 & \\ & \boldsymbol{D}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{L}_{11}^T & \boldsymbol{L}_{21}^T \\ & \boldsymbol{I} \end{bmatrix},$$

replace $\boldsymbol{D}_2$ (not necessarily diagonal!) by a *positive definition* matrix and complete the factorization.

No extra work if the Cholesky factorization is taken in the outer-product form.

The Cholesky factorization also tells if the current point is a minimizer or a saddle point.

Modified Newton's method for saddle point

Suppose we are at a saddle point $\bar{\boldsymbol{x}}$. Then $\bar{\boldsymbol{g}} = \nabla f(\bar{\boldsymbol{x}}) = \boldsymbol{0}$ and 2nd-order approximation

$$f(\bar{\boldsymbol{x}} + \boldsymbol{d}) \approx q(\boldsymbol{d}) := f(\bar{\boldsymbol{x}}) + \underbrace{\bar{\boldsymbol{g}}^T \boldsymbol{d}}_{=0} + \frac{1}{2} \boldsymbol{d}^T \boldsymbol{F}(\bar{\boldsymbol{x}}) \boldsymbol{d}.$$

How do we descend?

Greenstadt: pick $\boldsymbol{d} = \sum_{i:\lambda_i<0} \alpha_i \boldsymbol{u}_i$, where $\alpha_i > 0$ and $(\lambda_i, \boldsymbol{u}_i)$ are eigen-pairs of $\boldsymbol{F}(\boldsymbol{x})$. $\boldsymbol{d}$ is a positive linear combination of the *negative curvature* directions. Then, $\boldsymbol{d}^T \boldsymbol{F}(\bar{\boldsymbol{x}}) \boldsymbol{d} < 0$ and $q(\boldsymbol{d}) < f(\bar{\boldsymbol{x}})$.

Cholesky method: recall $\boldsymbol{D}_2$ correspond to the negative curvatures. If entry d_{ij} of $\boldsymbol{D}_2$ has the largest absolute value among all entries of $\boldsymbol{D}_2$, pick

$$\boldsymbol{L}^T \boldsymbol{d} = \boldsymbol{e}_i - \operatorname{sign}(d_{ij}) \boldsymbol{e}_j.$$

Then, $\boldsymbol{d}^T \boldsymbol{F}(\boldsymbol{x}) \boldsymbol{d} = \boldsymbol{d}^T \boldsymbol{L} \boldsymbol{D} \boldsymbol{L}^T \boldsymbol{d} < 0$.

Overview of the Gauss-Newton method

- A modification to Newton's method, solves nonlinear least squares, very popular
- Pros: second derivatives are no longer computed
- Cons: does not apply to general problems
- Can be improved by line search, Levenberg-Marquardt, etc.

Nonlinear least squares

- Given functions $r_i : \mathbb{R}^n \to \mathbb{R},\ i = 1, \ldots, m$
- The goal is to find $\boldsymbol{x}^*$ so that $r_i(\boldsymbol{x}) = 0$ or $r_i(\boldsymbol{x}) \approx 0$ for all i.
- Consider the nonlinear least-squares problem

$$\underset{\boldsymbol{x}}{\text{minimize}} \ \frac{1}{2} \sum_{i=1}^{m} \left(r_i(\boldsymbol{x})\right)^2 .$$

- Define $\boldsymbol{r} = [r_1, \ldots, r_m]^T$. Then we have

$$\underset{\boldsymbol{x}}{\text{minimize}}\ f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{r}(\boldsymbol{x})^T\boldsymbol{r}(\boldsymbol{x}).$$

- The gradient $\nabla f(\boldsymbol{x})$ is formed by components

$$(\nabla f(\boldsymbol{x}))_j = \frac{\partial f}{\partial x_j}(\boldsymbol{x}) = \sum_{i=1}^{m} r_i(\boldsymbol{x})\frac{\partial r_i}{\partial x_j}(\boldsymbol{x})$$

- Define the Jacobian of $\boldsymbol{r}$

$$\boldsymbol{J}(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial r_1}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial r_i}{\partial x_n}(\boldsymbol{x}) \\ & \cdots & \\ \frac{\partial r_m}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial r_m}{\partial x_n}(\boldsymbol{x}) \end{bmatrix}$$

  Then, we have

$$\nabla f(\boldsymbol{x}) = \boldsymbol{J}(\boldsymbol{x})^T\boldsymbol{r}(\boldsymbol{x})$$

- The Hessian $\boldsymbol{F}(\boldsymbol{x})$ is symmetric matrix. Its (k, j)th component is

$$\begin{aligned}\frac{\partial^2 f}{\partial x_k \partial x_j} &= \frac{\partial}{\partial x_k}\left(\sum_{i=1}^{m} r_i(\boldsymbol{x})\frac{\partial r_i}{\partial x_j}(\boldsymbol{x})\right) \\ &= \sum_{i=1}^{m}\left(\frac{\partial r_i}{\partial x_k}(\boldsymbol{x})\frac{\partial r_i}{\partial x_j}(\boldsymbol{x}) + r_i(\boldsymbol{x})\frac{\partial^2 r_i}{\partial x_k \partial x_j}(\boldsymbol{x})\right)\end{aligned}$$

- Let $\boldsymbol{S}(\boldsymbol{x})$ be formed by (k, j)th components

$$\sum_{i=1}^{m} r_i(\boldsymbol{x})\frac{\partial^2 r_i}{\partial x_k \partial x_j}(\boldsymbol{x})$$

- Then, we have $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{J}(\boldsymbol{x})^T\boldsymbol{J}(\boldsymbol{x}) + \boldsymbol{S}(\boldsymbol{x})$

- Therefore, Newton's method has the iteration

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \underbrace{(\boldsymbol{J}(\boldsymbol{x})^T\boldsymbol{J}(\boldsymbol{x}) + \boldsymbol{S}(\boldsymbol{x}))^{-1}}_{\boldsymbol{F}(\boldsymbol{x})^{-1}}\underbrace{\boldsymbol{J}(\boldsymbol{x})^T\boldsymbol{r}(\boldsymbol{x})}_{\nabla f(\boldsymbol{x})}$$

The Gauss-Newton method

- When the matrix $\boldsymbol{S}(\boldsymbol{x})$ is ignored in some applications to save computation, we arrive at the Gauss-Newton method

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \underbrace{(\boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{J}(\boldsymbol{x}))^{-1}}_{(\boldsymbol{F}(\boldsymbol{x})-\boldsymbol{S}(\boldsymbol{x}))^{-1}} \underbrace{\boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{r}(\boldsymbol{x})}_{\nabla f(\boldsymbol{x})}$$

- A potential problem is that $\boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{J}(\boldsymbol{x}) \not\succ \boldsymbol{0}$ and $f(\boldsymbol{x}^{(k+1)}) \geq f(\boldsymbol{x}^{(k)})$.

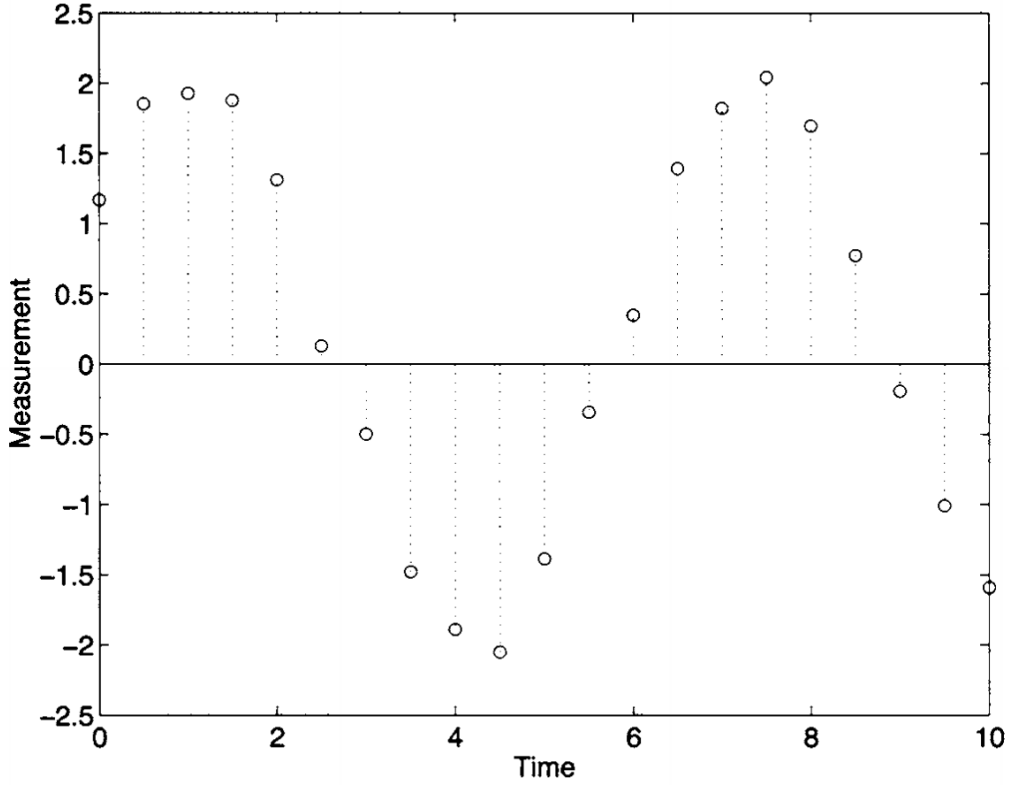
  Fixes: line search, Levenberg-Marquardt, and Cholesky/Gill-Murray.

Example: nonlinear data-fitting

- Given a sinusoid

$$y = A\sin(\omega t + \phi)$$

- Determine parameters A, ω, and ϕ so that the sinusoid best fits the observed points: (t_i, y_i), $i = 1, \ldots, 21$.



- Let $\boldsymbol{x} := [A, \omega, \phi]^T$ and

$$r_i(\boldsymbol{x}) := y_i - A\sin(\omega t_i + \phi)$$

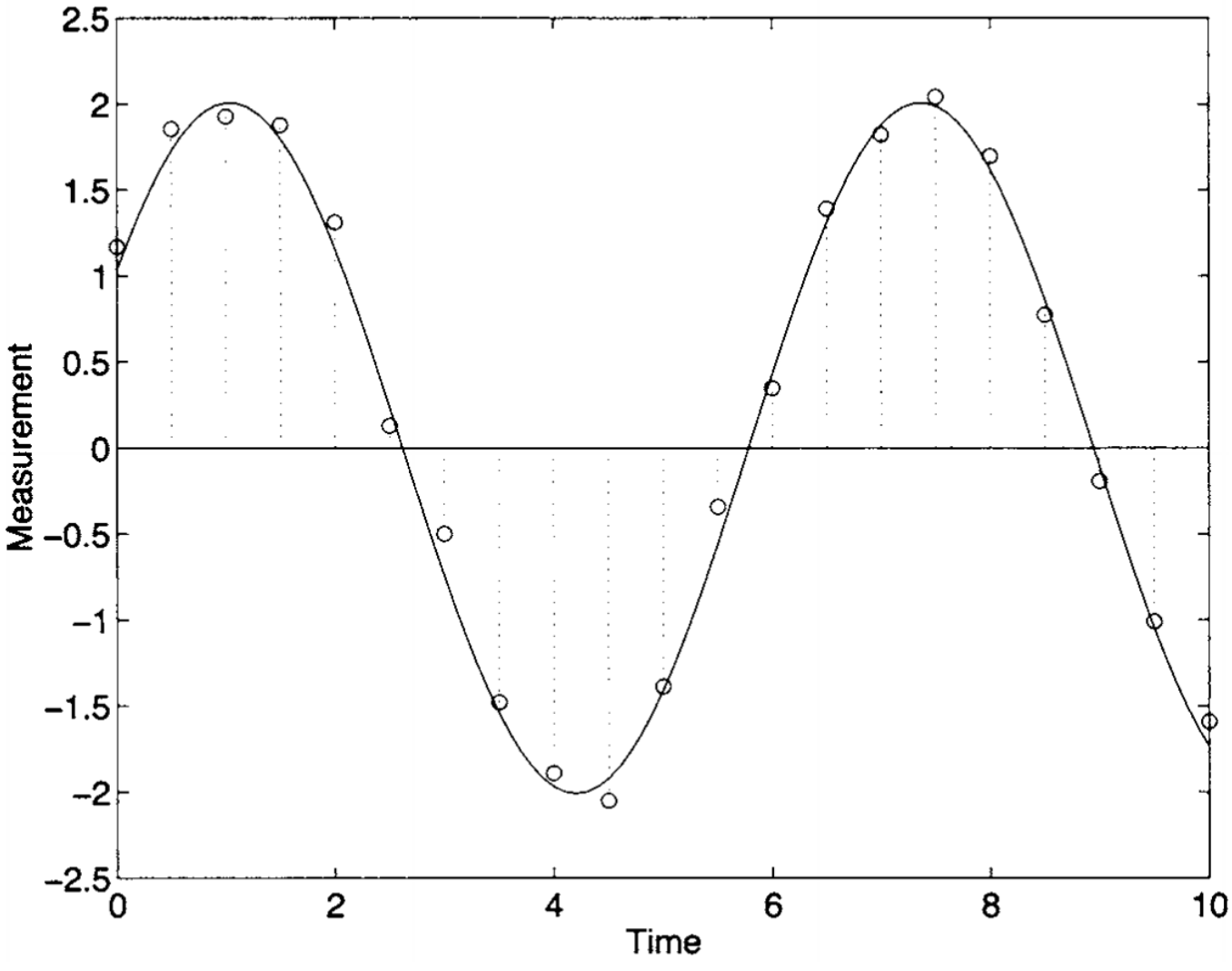
- Problem

$$\text{minimize} \sum_{i=1}^{21} (\underbrace{y_i - A\sin(\omega t_i + \phi)}_{r_i(\boldsymbol{x})})^2$$

- Derive $\boldsymbol{J}(\boldsymbol{x}) \in \mathbb{R}^{21\times 3}$ and apply the Gauss-Newton iteration

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - (\boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{J}(\boldsymbol{x}))^{-1} \boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{r}(\boldsymbol{x})$$

- Results: $A = 2.01,\ \omega = 0.992,\ \phi = 0.541.$


2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
Measurement
0
2
4
6
8
10
Time

Conclusions

Although Newton's method has many issues, such as

- the direction can be ascending if $\boldsymbol{F}(\boldsymbol{x}^{(k)}) \not\succ 0$
- may not ensure descent in general
- must start close to the solution,

Newton's method has the following strong properties:

- one-step solution for quadratic objective with an invertible $\boldsymbol{Q}$
- second-order convergence rate near the solution if $\boldsymbol{F}$ is Lipschtiz
- a number of modifications that address the issues.