Numerical Optimisation: Least squares

Department of Computer Sc
Centre for Medical Image Com
Centre for Inverse Proble
University College London

Lecture 10 & 11

Least squares is a problem where the objective function has the following special form

$$m$$

$$-$$

wher j

each of the r_j as a *residual*, and we assume t

Least squares problems are ubiquitous in applic
discrepancy between the model and the observe
minimised.

Let's assemble the individual components r_j into the *residual vector* $r : \mathbb{R}^n \to \mathbb{R}^m$

$$r(x) = (r_1(x), r_2(x), \ldots, r_m(x))^{\mathrm{T}}.$$

Using this vector, f becomes $f(x) = \frac{1}{2}\|r(x)\|_2^2$. The derivatives of f can be calculated with help of the Jacobian

$$\nabla r_1(x)^{\mathrm{T}} \quad {}^{\mathrm{T}}$$

$$\nabla f(x) = \sum_{j=1}^{m} r_j(x) \nabla r_j(x) =$$

$$\nabla^2 f(x) = \sum_{j=1}^{m} \nabla r_j(x) \nabla r_j(x)^{\mathrm{T}} + \sum_{j=1} r_j(x) \nabla^2 r_j(x)$$

$$= J(x)^{\mathrm{T}} J(x) + \sum_{j=1}^{m} r_j(x) \nabla^2 r_j(x),$$

Example

Model of concentration of drug in bloodstream

$$\phi(x;t) = x_1 + tx_2 + t^2x_3 + x_4 \exp^{-x_5 t}.$$

Find a set of parameters x so that the model best matches the data solving

$$\frac{1}{} \quad {}^{m}$$

Figure: Nocedal Wright Fig 10.1 (left), Fig 10.2 (right)

Bayesian perspective

Bayes' theorem

$$\pi(x|y) = \frac{\pi(y|x)\pi(x)}{\pi(y)} \propto \pi(y|x)\pi(x).$$

Den

at da

Assume that ϵ_js are independent and identic

variance σ^2 and probability density functio   od

of a particular set of observations y_j, j

parameter vector x is given by

$$\pi(y|x) = \prod_{j=1}^{m} g_\sigma(\epsilon_j) = \prod_{j=1}^{m} g_\sigma(\phi(x;t_j) - y_j).$$

The *maximum a posteriori probability* (MAP) estimate vs the maximum likelyhood estimate

$$x_{\mathrm{MAP}} = \max_x \pi(x|y) = \max_x \pi(y|x)\pi(x).$$

If g_ϵ is a *normal* distribution

$$\frac{1}{\quad} \qquad \frac{\epsilon^2}{\quad}$$

and ... ate reads

$$\begin{aligned} x_{\mathrm{MAP}} &= \max_x \frac{1}{(\sqrt{2\pi}\sigma)^m} \exp\left(- \frac{\quad}{\quad} \sum_{j=1} \right) \\ &= \min_x \sum_{j=1}^{m} (\phi(x;t_j) - y_j)^2. \end{aligned}$$

If $\phi(x; t)$ is linear function in x, the least squares problem becomes linear:

The re

nota

where

- the vector of measurements $y = ($
- the matrix J with rows $J_{j,:} := \phi(x;$

are both independent of x.

The linear least squares has the form

$$f(x) = \frac{1}{2}\|Jx - y\|^2.$$

The gradient and Hessian are

Note: $f(x)$ is convex i.e. the stationary point is th

minimiser $\nabla f(x^\star) = 0$

Normal equations

$$\nabla f(x^\star) = J^{\mathrm{T}}(Jx^\star - y) = 0 \quad \Leftrightarrow \quad J^{\mathrm{T}} J x^\star = J^{\mathrm{T}} y.$$

- Solve the normal equations $J^{\mathrm{T}} J x^\star = J^{\mathrm{T}} y$
  - \+ If $m \gg n$, computing $J^{\mathrm{T}} J$ explicitly results is a smaller matrix easier to store than J. This can be solved by e.g. Cholesky decomposition.
  - \- Formulating $J^{\mathrm{T}} J$ squares the condition number.

- Solve the least squares $x^\star = \arg\min$
  - \+ If J is of moderate size you can use direct m decomposition or SVD decompositio
  - \+ If J is large and sparse or given in operato x use iterative methods like CGLS or LSQR.
  - \+ Does not square the condition number.
  - \+ In particular SVD and iterative methods e.g. LSQR can easily deal with ill-conditioning.

Let

$$JP = Q\begin{bmatrix} R \\ 0 \end{bmatrix} = [\ \underbrace{Q_1}_{\in\mathbb{R}^{m\times n}} \quad \underbrace{Q_2}_{\in\mathbb{R}^{m\times(m-n)}}\ ]\begin{bmatrix} R \\ 0 \end{bmatrix} = Q_1 R,$$

where

- $n \quad n$
-
-

Recall: Multiplication with orthogonal mat

$$\|Jx - y\|_2^2 = \|Q^{\mathrm{T}}(J\overbrace{PP^{\mathrm{T}}}^{=}x - y)\|_2^2 = \|(Q\ \ JP)P\ \ x - Q\ \ y\|_2^2$$

$$= \|RP^{\mathrm{T}}x - Q_1^{\mathrm{T}}y\|_2^2 + \|Q_2^{\mathrm{T}}y\|_2^2.$$

Solution: $x^\star = PR^{-1}Q_1^{\mathrm{T}}y$. In practice we perform backsubstitution on $Rz = Q_1^{\mathrm{T}}y$ and permute for $x^\star = Pz$.

Let

$$J = U \begin{bmatrix} S \\ 0 \end{bmatrix} V^{\mathrm{T}} = [\ \underbrace{U_1}_{\in\mathbb{R}^{m\times n}} \quad \underbrace{U_2}_{\in\mathbb{R}^{m\times(m-n)}}\ ] \begin{bmatrix} S \\ 0 \end{bmatrix} V^{\mathrm{T}} = U_1 S V^{\mathrm{T}},$$

wher

-
- $_n > 0$.

$$\|Jx - y\|_2^2 = \|U^{\mathrm{T}}(J \overbrace{VV^{\mathrm{T}}}^{=I} x - y)\|_2^2$$

$$= \|SV^{\mathrm{T}}x - U_1^{\mathrm{T}}y\|_2^2 + \|U$$

Solution: $x^\star = VS^{-1}U_1^{\mathrm{T}}y = \sum_{i=1}^n \frac{u_i^{\mathrm{T}}y}{\sigma_i} v_i$. If σ_i are small, they would undue amplify the noise and can be omitted from the sum. Picard condition: $|u_i^{\mathrm{T}}y|$ should decay faster than σ_i.

LSQR applied to

$$\min_f \|Af-g\|_2^2+\tau\|f\|_2^2 = \min_f \left\| \begin{bmatrix} A \\ \sqrt{\tau} I\end{bmatrix} f - \begin{bmatrix} g\\ 0\end{bmatrix}\right\|_2^2,$$

wher
equi
avoi
the G

G-K bidiagonalisation yields the projected leas

$$\min_{y_i} \left\| \begin{bmatrix} B_i \\ \sqrt{\tau} I\end{bmatrix} y_i - \beta_1 e_1 \right\|, \qquad \text{(P-LS)}$$

which is then solved using QR decomposition yielding the approximation for the solution of the original problem, $f_i = V_i y_i$.

G-K bidiagonalization with a starting vector g for $\min_f \|g - Af\|$

$$
\begin{aligned}
U_{i+1}(\beta_1 e_1) &= g \\
AV_i &= U_{i+1} B_i \\
A^{\mathrm{T}} U_{i+1} &= V_i B^{\mathrm{T}} + \alpha_{i+1} v_{i+1} e^{\mathrm{T}} ,
\end{aligned}
$$

e_i: i … hat
$\|u_i\|$ …

$$
B_i = \begin{bmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \beta_3 & \ddots & \\ & & \ddots & \alpha_i \\ & & & \beta_{i+1} \end{bmatrix}, \quad U_i = [u_1, \ldots v_i]. \tag{1}
$$

Preconditioned LSQR

Preconditioned LSQR

$$\hat{f} = \text{argmin}\|g - AL^{-1}\hat{f}\|.$$

The c

prec

$$L^{-\mathrm{T}}A^{\mathrm{T}}AL^{-1}\hat{f} = L^{-\mathrm{T}}A^{\mathrm{T}}g \tag{2}$$

Similarly as for CG, the LSQR algorithm can be for
without the need to provide a factorization
algorithm [Arridge, B, Harhanen '14].

Initialization:
$\beta_1 u_1 = g$
$\tilde{p} = A^{\mathrm{T}} u_1$
$\tilde{v}_1 = M^{-1}\tilde{p}$, $\alpha_1 = (\tilde{v}_1, \tilde{p})^{1/2}$, $\tilde{v}_1 = \tilde{v}_1/\alpha_1$
$\tilde{w}_1 = \tilde{v}_1,\ f_0 = 0,\ \bar{\phi}_1 = \beta_1,\ \bar{\rho}_1 = \alpha_1$
f

    $\tilde{v}_{i+1} = M^{-1}\tilde{p}$, $\alpha_{i+1} = (\tilde{v}_{i+1}, \tilde{p})^{1/2}$,
    $\tilde{p} = \tilde{p}/\alpha_{i+1}$, $\tilde{v}_{i+1} = \tilde{v}_{i+1}/\alpha_{i+1}$
    Orthogonal transformation: smar
    Update:
    $f_i = f_{i-1} + (\phi_i/\rho_i)\tilde{w}_i$
    $\tilde{w}_{i+1} = \tilde{v}_{i+1} - (\theta_{i+1}/\rho_i)\tilde{w}_i$
    Break if stopping criterion satisfied
end for

- In preconditioned formulation, Tikhonov (explicit) regularization amounts to damping. For a fixed value of τ, damping can be easily incorporated in LSQR at the cost of doubling the number of Givens rotations [Paige, Saunders '82].

- Due to the shift invariance of Krylov spaces, V_i are the same

- Solving (P-LS) with a variable τ i
using singular value decomposition of the b B_i
(no efficient SVD update even though
row and a column in each iteration). Those qu
obtained at the cost $\mathcal{O}(i^2)$ at the i^{th} iteration.

- For larger i, the algorithm described in [Elden '77] for the least squares solution of (P-LS) at the cost of $\mathcal{O}(i)$ for each value of τ is the preferable option.

[Saunders Paige '82] discusses three stopping criteria:

S1: $\|\bar{r}_i\| \leq \mathrm{BTOL}\|g\| + \mathrm{ATOL}\|\bar{A}\|\|f_i\|$ (consistent systems),

S2: $\frac{\|\bar{A}^T \bar{r}_i\|}{\|\bar{A}\|\|\bar{r}_i\|} \leq \mathrm{ATOL}$ (inconsistent systems),

S3: $\mathrm{cond}(\bar{A}) \geq \mathrm{CONLIM}$ (both),

wher –

[Arri

(suitable for ill-posed problems)

S4: $\|r_i\| \leq \eta\delta, \quad \eta > 1$,

where $r_i := g - Af_i$, δ is the (estimated) n ts overregularization

+ if $\tau = 0$, $r_i = \bar{r}_i$ and the sequence $\|r_i\| = \|\bar{r}_i\|$ is monotonically decreasing. Moreover if initialised with f_0, $\|f_i\|$ is strictly monotonically growing (relevant for damped problem),
+ priorconditioning does not alter the residual.

Gauss-Newton (GN) can be viewed as a modified Newton method with line search.

Recall the specific form of gradient and Hessian of least squares problems

∇f m $^2 r_j(x)$.

Subs

$$\nabla^2 f(x_k) p_k = -$$

and using the approximation $\nabla^2 f(x_k)$

$$\underbrace{J(x_k)^{\mathrm{T}}}_{=:J_k^{\mathrm{T}}} J(x_k) p_k^{\mathrm{GN}} = -J(x_k) \underbrace{r(x_k)}_{=:r_k}.$$

Implementations of GN usually perform a line search along p_k^{GN} requiring the step length to satisfy e.g. Armijo or Wolfe conditions.

- Does not require computation of the individual Hessians $\nabla^2 r_j, j = 1, \dots, m$. If the Jacobian J_k has been computed when evaluating the gradient no other derivatives are needed.
- Frequently the first term $J_k^T J_k$ dominates the second term in the Hessian i.e. $|r_j(x)| \|\nabla^2 r_j(x)\|$ are much smaller than the eigenvalues $J^T J$. This happens if either the residual r_j or

- or f and hence suitable for line search

$$
\begin{aligned}
(p_k^{\mathrm{GN}})^{\mathrm{T}} \nabla f_k &= (p_k^{\mathrm{GN}})^{\mathrm{T}} J_k^{\mathrm{T}} r_k \\
&= -\|J_k p_k^{\mathrm{GN}}\|^2
\end{aligned}
$$

The final inequality is strict unless $J_k p_k^{\mathrm{GN}} = 0$ in which case by the GN equation and J_k being full-rank we have $0 = J_k^{\mathrm{T}} r_k = \nabla f_k$ and x_k is a stationary point.

The GN equation

$$J_k^{\mathrm{T}} J_k p_k^{\mathrm{GN}} = -J_k^{\mathrm{T}} r_k$$

is exactly the normal equation for the linear least squares problem

$$\frac{1}{} \qquad\qquad {}^{2}$$

Henc st
squa

We can view GN equation as obtained from a linear m
vector function $r(x_k + p) \approx r_k + J_k p$,

$$f(x_k + p) = \frac{1}{2}\|r_k(x_k + p)\|^2 \approx \frac{1}{2}\|J_k p + r_k\|^2$$

and $p_k^{\mathrm{GN}} = \arg\min_p \frac{1}{2}\|J_k p + r_k\|^2$.

Global convergence of GN

The global convergence is a consequence of the convergence theorem for line search methods [Zoutendijk].

To satisfy the conditions of Zoutendijk's theorem, we need to make following assumptions:

-

-

$$\|J(x)z\| \geq \gamma \|z\|,$$

Then for the iterates x_k generated by the length satisfying Wolfe conditions, we have

$$\lim_{k \to \infty} J_k^{\mathrm{T}} r_k = \nabla f(x_k) = 0.$$

Similarly as for the line search, we check that the angle $\theta_k = \angle(p_k^{\mathrm{GN}}, -\nabla f_k)$ is uniformly bounded away from $\pi/2$

$$\cos\theta_k = -\frac{(p_k^{\mathrm{GN}})^{\mathrm{T}}\nabla f_k}{\|p_k^{\mathrm{GN}}\|\|\nabla f_k\|} = \frac{\|J_k p_k^{\mathrm{GN}}\|^2}{\|p_k^{\mathrm{GN}}\|\|J_k^{\mathrm{T}} J_k p_k^{\mathrm{GN}}\|} \geq \frac{\gamma^2\|p_k^{\mathrm{GN}}\|^2}{\beta^2\|p_k^{\mathrm{GN}}\|^2} = \frac{\gamma^2}{\beta^2} > 0,$$

wher ss
of the l
r_j, j

Then from $\sum_{k\geq 0}\cos^2\theta_k\|\nabla f_k\|^2 < \infty$ follows $\nabla f_k \to 0$.

If J_k for some k is rank deficient, the matrix $_k$ $_k$ is singular and the system has infinitely many solutions, however $\cos\theta_k$ is not necessarily bounded away from 0.

The convergence of GN can be rapid if $J_k^{\mathrm{T}} J_k$ dominates the second term in the Hessian. Similarly as showing the convergence rate of Newton iteration, if x_k is sufficiently close to $x^\star$, $J(x)$ satisfies the uniform full-rank condition, we have for a unit step in GN direction

$:= J(x_k)^{\mathrm{T}} J(x_k)$

$x_k +$

$) - \nabla f(x_k) \Big] .$

Using $H(x)$ to denote the second term in the Hessi

from Taylor theorem that

$$\nabla f(x_k) - \nabla f(x^\star) = \int_0^1 J^{\mathrm{T}} J(x^\star + t(x_k - x^\star))(x_k - x^\star) dt$$
$$+ \int_0^1 H(x^\star + t(x_k - x^\star))(x_k - x^\star) dt,$$

Putting everything together and assuming Lipschitz continuity of $H(\cdot)$ near $x^\star$ and using L.c.d. of $r_j \Rightarrow$ L.c. of $J^{\mathrm{T}} r(x)$

$$\|x_k + p_k^{\mathrm{GN}} - x^\star\| \le \int_0^1 \|[J^{\mathrm{T}} J(x_k)]^{-1} H(x^\star + t(x_k - x^\star))\| \|x_k - x^\star\| dt$$

$$+ \int_0^1 \|\underbrace{[J^{\mathrm{T}} J(x_k)]^{-1}}_{\|\cdot\| \le \gamma^{-2}} \underbrace{\left(J^{\mathrm{T}} J(x^\star + t(x_k - x^\star)) - J^{\mathrm{T}} J(x_k)\right)}_{\|\cdot\| = \mathcal{O}(\|x_k - x^\star\|) \text{ would need L.c. of } J^{\mathrm{T}} J(x)} \| \|x_k - x^\star\| dt$$

$\star\|^2)$.

Henc

quic ce

is quadratic (Newton).

When the Jacobian $J(x)$ is large and sparse, t

equation can be replaced by an *inexact* solve as in inexact Newton methods but with the true Hessian $\nabla^2 f(x_k)$ replaced with $J(x_k)^{\mathrm{T}} J(x_k)$. The positive semidefinitness of $J(x_k)^{\mathrm{T}} J(x_k)$ simplifies the algorithms. Instead of (preconditioned) CG, (preconditioned) LSQR should be used.

Levenberg-Marquardt (LM) makes use of the same Hessian approximation as GN but within the framework of trust region methods. Trust region methods can cope with (nearly) rank-deficient Hessian, which is a weakness of GN.

The c

(CM-LM)

where $\Delta_k > 0$ is the trust region radius.

Note: The least squares term corresponds to q

$$m_k(p) = \frac{1}{2}\|r_k\|^2 + p^{\mathrm{T}} J_k^{\mathrm{T}} r_k + \frac{1}{2} p^{\mathrm{T}} J_k^{\mathrm{T}} J_k p.$$

The solution of the constraint model problem (CM-LM) is an immediate consequence of the general result for trust region methods [More, Sorensen]:

- If the solution p_k^{GN} of the GN equation lies strictly inside the trust region i.e. $\|p_k^{\mathrm{GN}}\| < \Delta_k$, then $p_k^{\mathrm{LM}} = p_k^{\mathrm{GN}}$ solves

- [illegible] of (CM-LM) satisfies $\|p_k \| = \Delta_k$ and

$$(J_k^{\mathrm{T}} J_k + \lambda I) p_k^{\mathrm{LM}} [illegible]$$

Note: The last equation is the normal equation [illegible] squares problem

$$\min_p \frac{1}{2} \left\| \begin{bmatrix} J_k \\ \sqrt{\lambda} I \end{bmatrix} p + \begin{bmatrix} r \\ 0 \end{bmatrix} \right\|^2,$$

which gives us a way of solving (CM-LM) without computing $J_k^{\mathrm{T}} J_k$.

Global convergence is a consequence of the corresponding trust region global convergence theorem.

To satisfy the conditions of that theorem we make the following assumptions:

- $\eta \quad (0, \frac{1}{4})$ (for strong convergence)
-
-

$$m_k(0) - m_k(p_k) \geq c_1 \|J_k^{\mathrm{T}} r_k$$

for some constant $c_1 > 0$, and in addit

some constant $\gamma \geq 1$.

We then have that

$$\lim_{k \to \infty} \nabla f_k = \lim_{k \to \infty} J_k^{\mathrm{T}} r_k = 0.$$

- As for trust region methods, there is no need to evaluate the right hand side of the decrease condition, but it is sufficient to ensure reduction of at least the Cauchy point, which can be calculated inexpensively. If the iterative CG-Steighaus

-
solution x , at which the first term $J(x\ )\ J(x\ )$ of the Hessian $\nabla^2 f(x^\star)$ dominates the secon
becomes inactive and the algorithm takes G
local convergence.

- [Paige, Saunders '82] Link to website with papers and codes `https://web.stanford.edu/group/SOL/software/lsqr/`
- [Bjorck '96] A. Bjorck, "Numerical Methods for Least Squares Problems", SIAM, 1996
- *AM*
- ill-conditioned least squares problems"
- [Arridge, B, Harhanen '14] S R Arridge, M M B Harhanen, "Iterated preconditioned L problems on unstructured grids", *Inverse Problems*, 30(7) 2014
- [Hansen '98] P. C. Hansen "Rank-Deficient and Discrete Ill-Posed Problems", SIAM, 1998