LECTURE 3

PRACTICAL NEWTON METHODS

Modified Hessian

(Based on Identity Shifts)

fixed positive scalar

$$M_k := \begin{cases} \nabla^2 f_k & \lambda_{min}(\nabla^2 f_k) \geq \delta \\ \nabla^2 f_k + \tau_k I & \text{otherwise} \end{cases}$$

where $\tau_k := \delta - \lambda_{min}(\nabla^2 f_k)$

$$M_k := Q_k \begin{bmatrix} \hat{\lambda}_1 & & 0 \\ & \hat{\lambda}_2 & \\ 0 & & \ddots \\ & & & \hat{\lambda}_n \end{bmatrix} Q_k^T$$

(Based on Orthogonal Eigenvalue Decomposition)

given orthogonal eigenvalue decomposition

$$\nabla^2 f_k = Q_k \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \\ & & & \lambda_n \end{bmatrix} Q_k^T$$

and $\hat{\lambda}_k := \max(\delta, \lambda_k)$.

THM

The modified Hessian by identity shifts satisfies

$$\| M_k - \nabla^2 f_k \|_2 \leq \| S - \nabla^2 f_k \|_2$$

for every S such that $S^T = S$ and $\lambda_{min}(S) \geq \delta$

THM

The modified Hessian by orthogonal eigenvalue decomposition satisfies

$$\|M_k - \nabla^2 f_k\|_F \leq \|S - \nabla^2 f_k\|_F$$

for every S such that $S^T = S$ and $\lambda_{min}(S) \geq \delta$.

A complete framework

Compute ∇f_0, $\nabla^2 f_0$ and $k \leftarrow 0$

While $\|\nabla f_k\| > \epsilon$

Form M_k from $\nabla^2 f_k$

Solve $M_k p_k = -\nabla f_k$

Determine an α_k satisfying Wolfe conditions

$x^{(k+1)} \leftarrow x^{(k)} + \alpha_k p_k$

Compute $\nabla f_{k+1} := \nabla f(x^{(k+1)})$, $\nabla^2 f_{k+1} := \nabla^2 f(x^{(k+1)})$

$k \leftarrow k+1$

end

Guaranteed to converge if $\|\nabla^2 f(x)\|_2$ is bounded for all $x \in \mathbb{R}^n$.

Practical Variant becomes the pure Newton's method for large k provided

$\lim_{k\to\infty} x^{(k)} = x_*$ such that

$$\lambda_{min}\left(\nabla^2 f(x_*)\right) > \delta.$$

Specifically, for large k

(i) $M_k = \nabla^2 f_k$

(ii) $\alpha_k = 1$ satisfies Wolfe conditions (if $c_1 < \frac{1}{2}$).

To see (ii), noting $p_k = -\left[\nabla^2 f_k\right]^{-1} \nabla f_k$ for large k, by Taylor's thm

satisfaction of sufficient decrease for $\alpha_k = 1$

$$\begin{aligned} f(x^{(k)} + p_k) &= f(x^{(k)}) + \nabla f(x^{(k)})^T p_k \\ &\quad + \frac{1}{2} p_k^T \nabla^2 f(x^{(k)} + t p_k) p_k \\ &= f(x^{(k)}) + \nabla f(x^{(k)})^T p_k + \frac{1}{2} p_k^T \nabla^2 f(x^{(k)}) p_k \\ &\quad + O\left(\|\nabla f_k\|^3\right) \\ &\qquad \left(\text{or } O\left(\|p_k\|^3\right)\right) \\ &= f(x^{(k)}) + \frac{1}{2} \nabla f(x^{(k)})^T p_k + O\left(\|\nabla f_k\|^3\right) \\ &< f(x^{(k)}) + c_1 \nabla f(x^{(k)})^T p_k \end{aligned}$$

By Taylor's thm with integral remainder

satisfaction of sufficient curvature by $\alpha_k = 1$

$$\left| \nabla f(x^{(k)} + p_k)^T p_k \right| =$$

$$\left| \left[\nabla f(x^{(k)}) + \int_0^1 \nabla^2 f(x^{(k)} + t p_k) p_k \, dt \right]^T p_k \right| =$$

$$\left| \nabla f(x^{(k)})^T p_k + \int_0^1 p_k^T \nabla^2 f(x^{(k)} + t p_k) p_k \, dt \right| =$$

$$O(\| \nabla f_k \|^3) \leq c_2 \left| \nabla f(x^{(k)})^T p_k \right|$$

$$\Longrightarrow$$

$$\nabla f(x^{(k)} + p_k)^T p_k \geq c_2 \nabla f(x^{(k)})^T p_k .$$

THM (Local Convergence of Newton's Method)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a twice-continuously differentiable function with Lipschitz-continuous Hessian, that is there exists a $\gamma > 0$ such that

$$\| \nabla^2 f(x) - \nabla^2 f(y) \|_2 \leq \gamma \| x - y \|_2 \quad \text{for all } x, y \in \mathbb{R}^n.$$

Suppose also $x_* \in \mathbb{R}^n$ is a point such that $\nabla f(x_*) = 0$ and $\nabla^2 f(x_*)$ is invertible.

There exists a ball $B(x_*, r)$ such that if $x^{(k)} \in B(x_*, r)$ (recall $\{x^{(k)}\}$ is the sequence generated by pure Newton's method), then

the following hold:

(i) $x^{(k)} \in B(x_*, r)$ for each $k \geq K$;

(ii) $\lim_{k \to \infty} x^{(k)} = x_*$;

(iii) $$\frac{\| x^{(k+1)} - x_* \|}{\| x^{(k)} - x_* \|^2} \leq \frac{M r}{2} \quad \text{for each } k \geq K$$

and for each $M > \| [\nabla^2 f(x_*)]^{-1} \|_2$.

PROOF

By continuity of second derivatives, there exists a ball $B(x_*, r_1)$ such that

$$\| [\nabla^2 f(x)]^{-1} \|_2 \leq M.$$

Let $r := \min(r_1, r_2)$ for any $r_2 < \frac{2}{(Mr)}$

Suppose $x^{(k)} \in B(x_*, r)$. By Taylor's thm

$$\nabla f(x_*) = \nabla f_k + \int_0^1 \nabla^2 f\left(x^{(k)} + t(x_* - x^{(k)})\right)(x_* - x^{(k)})\, dt$$

$$0 = \nabla f_k + \nabla^2 f_k (x_* - x^{(k)}) + \int_0^1 \left[\nabla^2 f\left(x^{(k)} + t(x_* - x^{(k)})\right) - \nabla^2 f_k\right] \times (x_* - x^{(k)})\, dt$$

$$0 = \underbrace{\left[\nabla^2 f_k\right]^{-1} \nabla f_k}_{-p_k = x^{(k)} - x^{(k+1)}} + (x_* - x^{(k)}) + \left[\nabla^2 f_k\right]^{-1} \int_0^1 \left[\nabla^2 f\left(x^{(k)} + t(x_* - x^{(k)})\right) - \nabla^2 f_k\right] \times (x_* - x^{(k)})\, dt$$

$$\Longrightarrow \quad x^{(k+1)} - x_* = \left[\nabla^2 f_k\right]^{-1} \int_0^1 \left[\nabla^2 f\left(x^{(k)} + t\left(x_* - x^{(k)}\right)\right) - \nabla^2 f_k\right] \times \left(x_* - x^{(k)}\right) dt$$

Taking the norms of both sides

$$(R) \quad \left\| x^{(k+1)} - x_* \right\| \leq \left\| \left[\nabla^2 f_k\right]^{-1} \right\| \int_0^1 \left\| \nabla^2 f\left(x^{(k)} + t\left(x_* - x^{(k)}\right)\right) - \nabla^2 f_k \right\| \times \left\| x^{(k)} - x_* \right\| dt$$

$$\leq M \int_0^1 \gamma t \left\| x^{(k)} - x_* \right\|^2 dt$$

$$= \frac{M\gamma}{2} \left\| x^{(k)} - x_* \right\|^2.$$

Since $x^{(k)} \in B(x_*, \rho)$, we have

$$\left\| x^{(k)} - x_* \right\| < 2/(M\gamma)$$

$$\Longrightarrow \frac{M\gamma}{2} \left\| x^{(k)} - x_* \right\| < 1.$$

It follows from (R) that

$$\left\| x^{(k+1)} - x_* \right\| < \left\| x^{(k)} - x_* \right\|,$$

so $x^{(k+1)} \in B(x_*, \rho)$ proving (i).

Moreover defining $\theta := \frac{M\gamma}{2} \left\| x^{(k)} - x_* \right\| < 1$

$$\left\| x^{(k+1)} - x_* \right\| \leq \left[\frac{M\gamma}{2} \left\| x^{(k)} - x_* \right\|\right] \left\| x^{(k)} - x_* \right\|$$

$$\leq \underbrace{\left[\frac{M\gamma}{2} \left\| x^{(k)} - x_* \right\|\right]}_{\theta} \left\| x^{(k)} - x_* \right\|.$$

Consequently,

$$\| x^{(k)} - x_* \| \leq \rho^{k-\kappa} \| x^{(\kappa)} - x_* \|$$

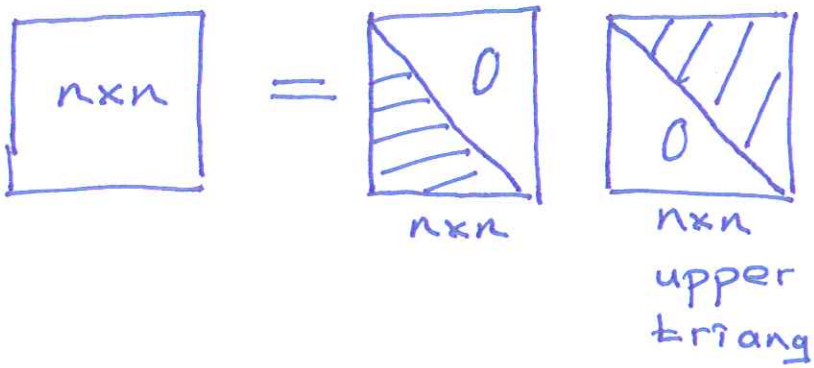
Taking the limit as $k \to \infty$ yields (ii).

Finally (iii) immediately follows from (R). □

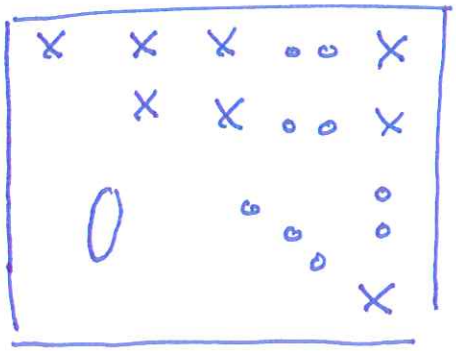
Hessian modifications based on Cholesky factorization

Suppose A is PD, then it has a Cholesky factorization of form

$$A = R^T R$$



Computation of R



row-by-row

top to bottom

To compute r_{ij}, $i \le j$, exploit

$$a_{ij} = r_i^T r_j$$

$$= \sum_{k=1}^{i} r_{ki} r_{kj}$$

$\underline{i = j}$

$$a_{ii} = \sum_{k=1}^{i} r_{ki}^2$$

$$\Longrightarrow r_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2}$$

rows $1, \ldots, i-1$

$j = i+1, i+2, \ldots, n$

$$r_{ij} = \left[a_{ij} - \sum_{k=1}^{i-1} r_{ki}\, r_{kj} \right] / r_{ii}$$

rows $1, \ldots, i-1$

Pseudocode

$R \leftarrow 0$

for $i = 1, \ldots, n$

$r_{ii} \leftarrow \sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2}$ $\quad c_{ii} :=$

for $j = i+1, \ldots, n$

$c_{ij} \leftarrow \left[a_{ij} - \sum_{k=1}^{i-1} r_{ki}\, r_{kj} \right]$

$r_{ij} \leftarrow c_{ij} / r_{ii}$

end

end

The only way this can break-down

$c_{ii} \leq 0$ for some i

Modify so that

$$c_{ii} \geq \delta$$

and

$$|r_{ij}| \leq \beta \qquad j = i+1, \ldots, n.$$

given positive scalars

Setting

$$r_{ii} := \sqrt{\max\left(c_{ii}, \delta, \left(\frac{\theta}{\beta}\right)^2\right)}$$

where $\theta := \max_{i+1 \leq j \leq n} |c_{ij}|$

achieves both. In particular,

$$|r_{ij}| \leq \frac{|c_{ij}|}{(\theta/\beta)} \leq \beta.$$

Modified Pseudocode

$R \leftarrow 0$

for $i = 1, \ldots, n$

$$c_{ii} \leftarrow a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2$$

for $j = i+1, \ldots, n$

$$c_{ij} \leftarrow a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj}$$

end

~~$r_{ii} \leftarrow$~~

$$\theta \leftarrow \max_{i+1 \leq j \leq n} |c_{ij}|$$

$r_{ii} \leftarrow \sqrt{\max\left(c_{ii}, \delta, (\theta/\beta)^2\right)}$

for $j = i+1, \ldots, n$

$r_{ij} \leftarrow c_{ij} / r_{ii}$

end

end

Modified Hessian

$$M_k := R^T R$$

where R is given by the modified pseudocode above.

Remark

M_k is PD, i.e.,

$$M_k = \begin{bmatrix} 1 & & & \\ \frac{r_{12}}{\sqrt{r_{11}}} & 1 & & \\ \vdots & \vdots & \ddots & \\ \frac{r_{1n}}{\sqrt{r_{11}}} & \frac{r_{2n}}{\sqrt{r_{22}}} & & 1 \end{bmatrix} \underbrace{\begin{bmatrix} r_{11} & & & \\ & r_{22} & & \\ & & \ddots & \\ & & & r_{nn} \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} 1 & \frac{r_{12}}{\sqrt{r_{11}}} & \cdots & \frac{r_{1n}}{\sqrt{r_{11}}} \\ & 1 & \cdots & \frac{r_{2n}}{\sqrt{r_{22}}} \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix}}_{\widetilde{R}}$$

for any nonzero $x \in \mathbb{R}^n$

$$\begin{aligned} x^T M_k x &= x^T \widetilde{R}^T D \underbrace{\widetilde{R} x}_{y \text{ (nonzero)}} \\ &= y^T D y \\ &= r_{11} y_1^2 + r_{22} y_2^2 + \cdots + r_{nn} y_n^2 > 0. \end{aligned}$$