LECTURE 4

PRACTICAL BFGS

BFGS update rule

$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k}$$

where $s_k := x^{(k+1)} - x^{(k)}$

$y_k := \nabla f_{k+1} - \nabla f_k$

Let $H_{k+1} := B_{k+1}^{-1}$ and $H_k := B_k^{-1}$.

Update rule relating H_{k+1}, H_k is desirable.

THM (Sherman-Morrison-Woodbury)

Let $A, \underline{A} \in \mathbb{R}^{n \times n}$ be invertible matrices such that

$$\underline{A} = A + uv^T$$

for some $u, v \in \mathbb{R}^n$. Then

$$\underline{A}^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}.$$

Letting

(1) $\underline{B}_k = B_k + \frac{y_k y_k^T}{(y_k^T s_k)}$

we have

$$② \quad B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k}$$

Applying the Sherman-Morrison-Woodbury formula first to ①, then ② results in

$$H_{k+1} = (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T$$

where $\rho_k := 1 / s_k^T y_k$

THM

If α_k is chosen so as to satisfy Wolfe conditions and H_k is PD, then H_{k+1} is PD.

PROOF

Sufficient curvature condition implies

$$\nabla f_{k+1}^T p_k \geq c_2 \nabla f_k^T p_k > \nabla f_k^T p_k$$

$$\Rightarrow \underbrace{(\nabla f_{k+1} - \nabla f_k)^T}_{y_k^T} \underbrace{(\alpha_k p_k)}_{s_k} > 0$$

$$\Rightarrow \rho_k > 0.$$

Consider for each nonzero $z \in \mathbb{R}^n$

$$z^T H_{k+1} z = \underbrace{z^T (I - \rho_k s_k y_k^T)}_{w^T} H_k \underbrace{(I - \rho_k y_k s_k^T) z}_{w} + \rho_k z^T s_k s_k^T z$$

$$= w^T H_k w + \rho_k (z^T s_k)^2$$

If $z^T s_k = 0$, then $w = z$ and

$$z^T H_{k+1} z = z^T H_k z > 0$$

On the other hand, if $z^T s_k \neq 0$, then

$$z^T H_{k+1} z = \underbrace{w^T H_k w}_{\geq 0} + \underbrace{\rho_k (z^T s_k)^2}_{> 0} > 0. \quad \square$$

A complete framework

Choose $x^{(0)}$, $k \leftarrow 0$

$H_0 \leftarrow$ a PD matrix (e.g. $H_0 \leftarrow I_n$)

Calculate f_0 and ∇f_0

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

$p_k \leftarrow -H_k \nabla f_k$

Choose a step-length α_k satisfying Wolfe conditions

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

Calculate f_{k+1} and ∇f_{k+1}

$H_{k+1} \leftarrow (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T$

$k \leftarrow k+1$

end

Inverse Hessian estimate H_{k+1} solves the following problem:

$$\begin{aligned} \text{minimize} \quad & \| H - H_k \|_W \\ & H \in \mathbb{R}^{n \times n} \\ & H^T = H \\ & H y_k = s_k \end{aligned}$$

where

$$\| A \|_W := \| W A W \|_F$$

$$= \sqrt{\text{Trace}\left((WAW)^T (WAW) \right)}$$

$W \in \mathbb{R}^{n \times n}$ is any symmetric matrix such that $W^2 s_k = y_k$.

PROOF - based on Karush-Kuhn-Tucker (KKT) conditions, i.e., optimality conditions for constrained optimization. (To be discussed soon.)

Global Convergence

If there exists $M > 0$ such that

$$\| B_k \|_2 \cdot \| B_k^{-1} \|_2 \leq M \quad \text{for all } k,$$

by Zoutendijk's thm

$$\lim_{k \to \infty} \nabla f_k = 0.$$

Local Convergence

Suppose

(i) $\nabla^2 f(x)$ is Lipschitz continuous,

(ii) x_* is a local minimizer, and

(iii) the pure BFGS sequence $\{x^{(k)}\}$ is such that $\lim_{k \to \infty} x^{(k)} = x_*$ and there exists $m > 0$ such that

$$\sum_{l=1}^{k} \| x^{(l)} - x_* \| \leq m \quad \text{for every } k.$$

Then

$$\lim_{k \to \infty} \frac{\| x^{(k+1)} - x_* \|}{\| x^{(k)} - x_* \|} = 0. \quad \text{(i.e. superlinear convergence)}$$

Determination of a step-length satisfying Wolfe conditions

$x^{(k)}$, $p_k \in \mathbb{R}^n$ given

p_k descent direction

$\phi(\alpha) := f(x^{(k)} + \alpha p_k)$

Wolfe conditions

$$\phi(\alpha_k) \leqslant \phi(0) + c_1 \alpha_k \phi'(0)$$

$$\phi'(\alpha_k) \geqslant c_2 \phi'(0)$$

where $0 < c_1 < c_2 < 1$.

Following pseudocode keeps α_{lo}, α_{hi} such that

(i) α_{lo} satisfies (SD) strictly and $\phi(\alpha_{lo})$ is the smallest $\phi(\alpha_k)$ among all α_k satisfying (SD),

(ii) α_{hi} satisfies $\phi'(\alpha_{lo})(\alpha_{hi} - \alpha_{lo}) < 0$.

Pseudocode

$\alpha_k \leftarrow (\alpha_{lo} + \alpha_{hi})/2$

OR

$\alpha_k \leftarrow$ local minimizer of a cubic Hermite interpolating polynomial at α_{lo}, α_{hi}.

if $\left[\phi(\alpha_k) < \phi(0) + c_1 \alpha_k \phi'(0) \right.$ and $\left. \phi(\alpha_k) < \phi(\alpha_{lo}) \right]$

  if $\phi'(\alpha_k) \geqslant +c_2 \phi'(0)$

    return α_k

  end

  if $\phi'(\alpha_k)\,[\alpha_{hi} - \alpha_k] \geqslant 0$

    $\alpha_{hi} \leftarrow \alpha_{lo}$

  end

  $\alpha_{lo} \leftarrow \alpha_k$

else

  $\alpha_{hi} \leftarrow \alpha_k$

end

Initially,

$\alpha_{lo} = 0$

$\alpha_{hi} > 0$, a point violating (SD).
(e.g. choosing α_{hi} large enough)

Interval I with end points α_{lo} and α_{hi} contains step-lengths satisfying Wolfe conditions.

Case 1: $\phi'(\alpha_{lo}) < 0$

In this case $\alpha_{hi} > \alpha_{lo}$.

(1.a) α_{hi} violates (SD).

$\phi(\alpha_{lo}) \leq l(\alpha_{lo})$ and $\phi(\alpha_{hi}) > l(\alpha_{hi})$

$\Longrightarrow \exists \underline{\alpha} \in I \quad \phi(\underline{\alpha}) = l(\underline{\alpha})$

Indeed choose $\underline{\alpha}$ as above and as close to α_{lo} as possible so that (SD) holds for all $\alpha \in (\alpha_{lo}, \underline{\alpha})$.

By the mean value thm

$$c_1 \phi'(0) = \frac{\phi(\underline{\alpha}) - \phi(0)}{\underline{\alpha}} \leq \frac{\phi(\underline{\alpha}) - \phi(\alpha_{lo})}{\underline{\alpha} - \alpha_{lo}} = \phi'(\underline{\underline{\alpha}})$$

for some $\underline{\underline{\alpha}} \in (\alpha_{lo}, \underline{\alpha})$. Indeed,

$$\phi'(\underline{\underline{\alpha}}) \geq c_1 \phi'(0) > c_2 \phi'(0),$$

so $\underline{\underline{\alpha}}$ satisfies Wolfe conditions. By continuity Wolfe conditions also hold in a neighborhood containing $\underline{\underline{\alpha}}$.

(1.b) α_{hi} satisfies (SD) and $\phi(\alpha_{hi}) > \phi(\alpha_{lo})$.

By mean value thm, there exists $\underline{\alpha} \in (\alpha_{lo}, \alpha_{hi})$ such that $\phi'(\underline{\alpha}) > 0$. Due to continuity of $\phi'(\alpha)$, there exist $\underline{\underline{\alpha}} \in (\alpha_{lo}, \underline{\alpha})$ such that

$\phi'(\underline{\underline{\alpha}}) = 0$. It follows that

$$\phi'(\underline{\underline{\alpha}}) \geq c_2 \phi'(0).$$

If $\underline{\underline{\alpha}}$ satisfies (SD), it satisfies Wolfe conditions. Otherwise, argument in [1.a] applies to show the existence of step-lengths in $(\alpha_{lo}, \underline{\underline{\alpha}})$ satisfying Wolfe conditions.

Case 2: $\phi'(\alpha_{lo}) > 0$

In this case $\alpha_{lo} > \alpha_{hi}$.

The existence of step-Lengths in $(\alpha_{hi}, \alpha_{lo})$ satisfying Wolfe conditions follows from the observation $\phi(\alpha_{hi}) > \phi(\alpha_{lo})$. (Details exercise)

Cubic Hermite Interpolating Polynomial

$P_3 : \mathbb{R} \to \mathbb{R}$

polynomial (unique) of degree 3 such that

$$P_3(\alpha_{lo}) = \phi(\alpha_{lo})$$
$$P_3(\alpha_{hi}) = \phi(\alpha_{hi})$$
$$P_3'(\alpha_{lo}) = \phi'(\alpha_{lo})$$
$$P_3'(\alpha_{hi}) = \phi'(\alpha_{hi})$$

If a minimizer of P_3 is in $(\alpha_{lo}, \alpha_{hi})$ (or $(\alpha_{hi}, \alpha_{lo})$), it is given by

$$\alpha_k = \alpha_{hi} - (\alpha_{hi} - \alpha_{lo}) \left[\frac{\phi'(\alpha_{hi}) + d_2 - d_1}{\phi'(\alpha_{hi}) - \phi'(\alpha_{lo}) + 2d_2} \right]$$

where

$$d_1 = \phi'(\alpha_{lo}) + \phi'(\alpha_{hi}) - 3 \frac{\phi(\alpha_{lo}) - \phi(\alpha_{hi})}{\alpha_{lo} - \alpha_{hi}}$$

$$d_2 = \sqrt{\left[d_1^2 - \phi'(\alpha_{lo}) \phi'(\alpha_{hi}) \right]}$$