Global Secant Methods and Matrix Structures

Stefano Fanelli

Padova, February 2007

1. Global Optimisation Quasi-Newton(QN) algorithms

2. Matrix structures in a global minimization scheme

Global Optimisation Quasi-Newton(QN) algorithms

Given an approssimation B_k of $\nabla^2 E(\mathbf{w}_k)$, let us define the matrix $\mathcal{L}_{B_k}$:

$$\|\mathcal{L}_{B_k} - B_k\|_F = \min_{X \in \mathcal{L}} \|X - B_k\|_F, \ \|\cdot\|_F = \mathit{Frob.norm}$$

where $\mathcal{L} = \mathit{algebra} \subset \mathbb{C}^{n \times n}$,

Global Optimisation Quasi-Newton(QN) algorithms

Given an approssimation B_k of $\nabla^2 E(\mathbf{w}_k)$, let us define the matrix $\mathcal{L}_{B_k}$:

$$\|\mathcal{L}_{B_k} - B_k\|_F = \min_{X \in \mathcal{L}} \|X - B_k\|_F, \quad \|\cdot\|_F = Frob.norm$$

where $\mathcal{L} = algebra \subset \mathbb{C}^{n \times n}$, one can define descent methods $\mathcal{L}QN$ [DFLZ]:

$$\begin{array}{l} \mathbf{w}_0 \in R^n, \quad \mathbf{d}_0 = -\mathbf{g}_0 \\ For \quad k = 0, 1, \ldots \\ \left\{ \begin{array}{l} \mathbf{w}_{k+1} = \mathbf{w}_k + \lambda_k \mathbf{d}_k \qquad \lambda_k > 0 \\ B_{k+1} = \varphi(\mathcal{L}_{B_k}, \underbrace{\mathbf{w}_{k+1} - \mathbf{w}_k}_{\mathbf{s}_k}, \underbrace{\mathbf{g}_{k+1} - \mathbf{g}_k}_{\mathbf{y}_k}), \quad \mathbf{g}_k = \nabla E(\mathbf{w}_k) \\ \mathbf{d}_{k+1} = -B_{k+1}^{-1} \mathbf{g}_{k+1} \end{array} \right. \end{array}$$

Global Optimisation Quasi-Newton(QN) algorithms

Given an approssimation B_k of $\nabla^2 E(\mathbf{w}_k)$, let us define the matrix $\mathcal{L}_{B_k}$:

$$\|\mathcal{L}_{B_k} - B_k\|_F = \min_{X \in \mathcal{L}} \|X - B_k\|_F, \quad \|\cdot\|_F = Frob.norm$$

where $\mathcal{L} = algebra \subset \mathbb{C}^{n\times n}$, one can define descent methods $\mathcal{L}QN$ [DFLZ]:

$$\begin{array}{l} \mathbf{w}_0 \in R^n, \quad \mathbf{d}_0 = -\mathbf{g}_0 \\ For \quad k = 0, 1, \dots \\ \left\{ \begin{array}{l} \mathbf{w}_{k+1} = \mathbf{w}_k + \lambda_k \mathbf{d}_k \qquad \lambda_k > 0 \\ B_{k+1} = \varphi(\, \mathcal{L}_{B_k}\,,\ \underbrace{\mathbf{w}_{k+1} - \mathbf{w}_k}_{\mathbf{s}_k},\ \underbrace{\mathbf{g}_{k+1} - \mathbf{g}_k}_{\mathbf{y}_k}), \quad \mathbf{g}_k = \nabla E(\mathbf{w}_k) \\ \mathbf{d}_{k+1} = -B_{k+1}^{-1}\mathbf{g}_{k+1} \end{array} \right. \end{array}$$

Remark

The classical BFGS method [NW] and the more recent minimization methods introduced in [BDFZ], [DFZ2], [DFZ3] are examples of $\mathcal{L}QN$ *algorithms, (being* $\mathcal{L} = \mathbb{C}^{n\times n}$, $\mathcal{L} = \{\alpha I\}, \{Hartley - type\}, \mathcal{L}^k$*)*

The *step* λ_k is determined such that:

$$\lambda_k \mid \mathbf{s}_k^T \mathbf{y}_k > 0 \quad \& \quad E(\mathbf{w}_{k+1}) < E(\mathbf{w}_k)$$

The *updating function* φ in $B_{k+1} = \varphi\left(\mathcal{L}_{B_k}, \mathbf{s}_k, \mathbf{y}_k\right)$ is

$$\varphi\left(\Box, \mathbf{s}, \mathbf{y}\right) = \Box + \frac{1}{\mathbf{y}^T\mathbf{s}}\mathbf{y}\mathbf{y}^T - \frac{1}{\mathbf{s}^T\Box\mathbf{s}}\Box\mathbf{s}\mathbf{s}^T\Box.$$

The *step* λ_k is determined such that:

$$\lambda_k \mid \mathbf{s}_k^T \mathbf{y}_k > 0 \quad \& \quad E(\mathbf{w}_{k+1}) < E(\mathbf{w}_k)$$

The *updating function* φ in $B_{k+1} = \varphi\left(\mathcal{L}_{B_k}, \mathbf{s}_k, \mathbf{y}_k\right)$ is

$$\varphi\left(\Box, \mathbf{s}, \mathbf{y}\right) = \Box + \frac{1}{\mathbf{y}^T\mathbf{s}}\mathbf{y}\mathbf{y}^T - \frac{1}{\mathbf{s}^T\Box\mathbf{s}}\Box\mathbf{s}\mathbf{s}^T\Box.$$

The choice of λ_k and the properties of φ and $\mathcal{L}_{B_k}$ imply:

- *B_{k+1} inherites positive definiteness from B_k*
- $B_{k+1}(\mathbf{w}_{k+1} - \mathbf{w}_k) = \mathbf{g}_{k+1} - \mathbf{g}_k, \Rightarrow$ *$\mathcal{L}QN$ secant* algorithms

The *step* λ_k is determined such that:

$$\lambda_k \mid \mathbf{s}_k^T \mathbf{y}_k > 0 \quad \& \quad E(\mathbf{w}_{k+1}) < E(\mathbf{w}_k)$$

The *updating function* φ in $B_{k+1} = \varphi\left(\mathcal{L}_{B_k}, \mathbf{s}_k, \mathbf{y}_k\right)$ is

$$\varphi\left(\square, \mathbf{s}, \mathbf{y}\right) = \square + \frac{1}{\mathbf{y}^T\mathbf{s}}\mathbf{y}\mathbf{y}^T - \frac{1}{\mathbf{s}^T\square\mathbf{s}}\square\mathbf{s}\mathbf{s}^T\square.$$

The choice of λ_k and the properties of φ and $\mathcal{L}_{B_k}$ imply:

- B_{k+1} *inherites positive definiteness from* B_k
- $B_{k+1}(\mathbf{w}_{k+1} - \mathbf{w}_k) = \mathbf{g}_{k+1} - \mathbf{g}_k$, $\Rightarrow$ $\mathcal{L}QN$ *secant* algorithms

The structured space $\mathcal{L} \Rightarrow \mathcal{L}QN$ of low complexity

One can prove the following global convergence theorem [DFZ4]:

Theorem 1

Given $E \in C^2$, let E_{min} be the value of its global minimum.
Assume that:

$$\forall \epsilon_a \in \Re^+, \exists \epsilon_s \in \Re^+ : \|\nabla E(\mathbf{w}_k)\| > \epsilon_s \text{ apart from } k : \; E(\mathbf{w}_k) - E_{min} < \epsilon_a$$

One can prove the following global convergence theorem [DFZ4]:

Theorem 1

Given $E \in C^2$, let E_{min} be the value of its global minimum.
Assume that:

$$\forall \epsilon_a \in \Re^+, \exists \epsilon_s \in \Re^+ : \|\nabla E(\mathbf{w}_k)\| > \epsilon_s \ \text{ apart from } k : \ E(\mathbf{w}_k) - E_{min} < \epsilon_a$$

If in an iterative scheme *of BFGS-type* $\mathbf{w}_{k+1} = \mathbf{w}_k - \lambda_k A_k^{-1} \nabla E(\mathbf{w}_k)$, $\left(A_k = \varphi(\tilde{A}_{k-1}, \ldots), \forall k\right)$ the following conditions are satisfied:

One can prove the following global convergence theorem [DFZ4]:

Theorem 1
Given $E \in C^2$, let E_{min} be the value of its global minimum.
Assume that:

$$\forall \epsilon_a \in \Re^+, \exists \epsilon_s \in \Re^+ : \|\nabla E(\mathbf{w}_k)\| > \epsilon_s \text{ apart from } k : \; E(\mathbf{w}_k) - E_{min} < \epsilon_a$$

If in an iterative scheme *of BFGS-type* $\mathbf{w}_{k+1} = \mathbf{w}_k - \lambda_k A_k^{-1} \nabla E(\mathbf{w}_k)$, $\left(A_k = \varphi(\tilde{A}_{k-1}, \ldots), \forall k \right)$ the following conditions are satisfied:

$$\frac{\|\nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k)\|^2}{(\nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k))^T \lambda_k \mathbf{d}_k} = \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T \mathbf{s}_k} \leq M,$$

One can prove the following global convergence theorem [DFZ4]:

Theorem 1
Given $E \in C^2$, let E_{min} be the value of its global minimum.
Assume that:

$$\forall \epsilon_a \in \Re^+, \exists \epsilon_s \in \Re^+ : \|\nabla E(\mathbf{w}_k)\| > \epsilon_s \text{ apart from } k : \; E(\mathbf{w}_k) - E_{min} < \epsilon_a$$

If in an iterative scheme *of BFGS-type* $\mathbf{w}_{k+1} = \mathbf{w}_k - \lambda_k A_k^{-1} \nabla E(\mathbf{w}_k)$, $\left(A_k = \varphi(\tilde{A}_{k-1}, \ldots), \, \forall k \right)$ the following conditions are satisfied:

$$\frac{\|\nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k)\|^2}{(\nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k))^T \lambda_k \mathbf{d}_k} = \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T \mathbf{s}_k} \leq M,$$

$$\|A_k\| \|A_k^{-1}\| \leq N.$$

One can prove the following global convergence theorem [DFZ4]:

Theorem 1

Given $E \in C^2$, let E_{min} be the value of its global minimum.
Assume that:

$$\forall \epsilon_a \in \Re^+, \exists \epsilon_s \in \Re^+ : \|\nabla E(\mathbf{w}_k)\| > \epsilon_s \ \text{ apart from } k : \ E(\mathbf{w}_k) - E_{min} < \epsilon_a$$

If in an iterative scheme *of BFGS-type* $\mathbf{w}_{k+1} = \mathbf{w}_k - \lambda_k A_k^{-1} \nabla E(\mathbf{w}_k)$, $\left(A_k = \varphi(\tilde{A}_{k-1}, \ldots), \forall k\right)$ the following conditions are satisfied:

$$\frac{\|\nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k)\|^2}{(\nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k))^T \lambda_k \mathbf{d}_k} = \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T \mathbf{s}_k} \leq M,$$

$$\|A_k\| \|A_k^{-1}\| \leq N.$$

Then, $\forall \epsilon_a \in \Re^+$, $\exists k^{**} : \forall k > k^{**}$:

$$E(\mathbf{w}_k) - E_{min} < \epsilon_a$$

Theorem 1 allows the definition of the following
Non-Suspiciousness Conditions for a BFGS-type method:

Theorem 1 allows the definition of the following Non-Suspiciousness Conditions for a BFGS-type method:

1. $\forall \varepsilon_a > 0, \exists \varepsilon_s$: $||\nabla E(\mathbf{w}_k)|| > \varepsilon_s$ during the BFGS-type descent algorithm, apart for $k : E(\mathbf{w}_k) - E_{min} < \varepsilon_a$;

Theorem 1 allows the definition of the following Non-Suspiciousness Conditions for a BFGS-type method:

1. $\forall \varepsilon_a > 0,\ \exists \varepsilon_s\colon\ \|\nabla E(\mathbf{w}_k)\| > \varepsilon_s$ during the BFGS-type descent algorithm, apart for $k:\ E(\mathbf{w}_k) - E_{min} < \varepsilon_a$;
2. $\lambda_k \leq \varepsilon_a / \|\nabla E(\mathbf{w}_k)\|^2$;

Theorem 1 allows the definition of the following Non-Suspiciousness Conditions for a BFGS-type method:

1. $\forall \varepsilon_a > 0,\ \exists \varepsilon_s:\ ||\nabla E(\mathbf{w}_k)|| > \varepsilon_s$ during the BFGS-type descent algorithm, apart for $k:\ E(\mathbf{w}_k) - E_{min} < \varepsilon_a$;
2. $\lambda_k \leq \varepsilon_a / \|\nabla E(\mathbf{w}_k)\|^2$;
3. $\dfrac{\|\nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k)\|^2}{(\nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k))^T \lambda_k \mathbf{d}_k} = \dfrac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T \mathbf{s}_k} \leq M,$
4. $\|A_k\| \|A_k^{-1}\| \leq N$

Theorem 1 allows the definition of the following Non-Suspiciousness Conditions for a BFGS-type method:

1. $\forall \varepsilon_a > 0,\ \exists \varepsilon_s$: $||\nabla E(\mathbf{w}_k)|| > \varepsilon_s$ during the BFGS-type descent algorithm, apart for $k :\ E(\mathbf{w}_k) - E_{min} < \varepsilon_a$;
2. $\lambda_k \leq \varepsilon_a / \|\nabla E(\mathbf{w}_k)\|^2$;
3. $$\frac{\|\nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k)\|^2}{(\nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k))^T \lambda_k \mathbf{d}_k} = \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T \mathbf{s}_k} \leq M,$$
4. $\|A_k\| \|A_k^{-1}\| \leq N$

N.B.
The second condition derives from terminal attractors theory [DFZ0].

Theorem 1 allows the definition of the following Non-Suspiciousness Conditions for a BFGS-type method:

1. $\forall \varepsilon_a > 0$, $\exists \varepsilon_s$: $||\nabla E(\mathbf{w}_k)|| > \varepsilon_s$ during the BFGS-type descent algorithm, apart for k : $E(\mathbf{w}_k) - E_{min} < \varepsilon_a$;
2. $\lambda_k \leq \varepsilon_a / \|\nabla E(\mathbf{w}_k)\|^2$;
3. $$\frac{\|\nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k)\|^2}{(\nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k))^T \lambda_k \mathbf{d}_k} = \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T \mathbf{s}_k} \leq M,$$
4. $\|A_k\| \|A_k^{-1}\| \leq N$

N.B.
The second condition derives from terminal attractors theory [DFZ0].
The third condition is a sort of weak discrete convexity assumption [P].

Theorem 1 allows the definition of the following Non-Suspiciousness Conditions for a BFGS-type method:

1. $\forall \varepsilon_a > 0,\ \exists \varepsilon_s\colon\ \|\nabla E(\mathbf{w}_k)\| > \varepsilon_s$ during the BFGS-type descent algorithm, apart for $k:\ E(\mathbf{w}_k) - E_{min} < \varepsilon_a$;
2. $\lambda_k \leq \varepsilon_a / \|\nabla E(\mathbf{w}_k)\|^2$;
3. $$\frac{\|\nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k)\|^2}{(\nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k))^T \lambda_k \mathbf{d}_k} = \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T \mathbf{s}_k} \leq M,$$
4. $\|A_k\| \|A_k^{-1}\| \leq N$

N.B.
The second condition derives from terminal attractors theory [DFZ0].
The third condition is a sort of weak discrete convexity assumption [P].

Since every descent direction is associated to a p.d. matrix with a well defined spectral structure, the fourth condition may be satisfied, by suitably modifying the matrices A_k during the optimization process.

Matrix structures in a global minimization scheme

In order to define a global minimization scheme, we must satisfy Theorem 1 assumptions from an operational point of view.

Matrix structures in a global minimization scheme

In order to define a global minimization scheme, we must satisfy Theorem 1 assumptions from an operational point of view.

This leads to compute a repeller matrix A_{rep} for each local minimization. The basic idea is ([T]) to approximate A_{rep} by the following expression:

$$A_{rep} \approx \lambda_{rep} I + (I/\mu + R)^{-1}, \quad rank(R) \leq 4$$

Matrix structures in a global minimization scheme

In order to define a global minimization scheme, we must satisfy Theorem 1 assumptions from an operational point of view.

This leads to compute a repeller matrix A_{rep} for each local minimization. The basic idea is ([T]) to approximate A_{rep} by the following expression:

$$A_{rep} \approx \lambda_{rep} I + (I/\mu + R)^{-1}, \quad rank(R) \leq 4$$

being, by Condition (2), λ_{rep} the *maximal scalar repeller*, i.e.:

$$\lambda_{rep} = \frac{\epsilon_a}{\|\nabla E(\mathbf{w}_r)\|^2}, \quad \|\nabla E(\mathbf{w}_r)\| << \sqrt{\epsilon_a}, \quad E(\mathbf{w}_r) >> E_{min}$$

Matrix structures in a global minimization scheme

In order to define a global minimization scheme, we must satisfy Theorem 1 assumptions from an operational point of view.

This leads to compute a repeller matrix A_{rep} for each local minimization. The basic idea is ([T]) to approximate A_{rep} by the following expression:

$$A_{rep} \approx \lambda_{rep} I + (I/\mu + R)^{-1}, \quad rank(R) \leq 4$$

being, by Condition (2), λ_{rep} the *maximal scalar repeller*, i.e.:

$$\lambda_{rep} = \frac{\epsilon_a}{\|\nabla E(\mathbf{w}_r)\|^2}, \quad \|\nabla E(\mathbf{w}_r)\| << \sqrt{\epsilon_a}, \quad E(\mathbf{w}_r) >> E_{min}$$

and R with the following structure:

$$R = \mu_1 \mathbf{p}\mathbf{p}^T + \mu_2 \mathbf{q}\mathbf{q}^T + \mu_3 \mathbf{q}\mathbf{p}^T + \mu_4 \mathbf{p}\mathbf{q}^T, \quad \mathbf{p} \;\; \text{e} \; \mathbf{q} \;\; \text{suitable vectors}$$

In the first iterations of every local minimization, it is sufficient to verify Condition (1) with a “large” ϵ_s.

In the first iterations of every local minimization, it is sufficient to verify Condition (1) with a "large" ϵ_s.

Since $\mathbf{w}_k$ is such that $E(\mathbf{w}_{k+1}) < E(\mathbf{w}_k)$,

for $\epsilon_a > E(\mathbf{w}_1) - E_{\min}$, (1) is satisfied if

$$\epsilon_s = \frac{1}{2}\|\nabla E(\mathbf{w}_0)\|$$

In the first iterations of every local minimization, it is sufficient to verify Condition (1) with a "large" ϵ_s.

Since $\mathbf{w}_k$ is such that $E(\mathbf{w}_{k+1}) < E(\mathbf{w}_k)$,

for $\epsilon_a > E(\mathbf{w}_1) - E_{\min}$, (1) is satisfied if

$$\epsilon_s = \frac{1}{2}\|\nabla E(\mathbf{w}_0)\|$$

... for $\epsilon_a > E(\mathbf{w}_r) - E_{\min}$, (1) is satisfied if

$$\epsilon_s = \frac{1}{2}\min_{k=0,\ldots,r}\|\nabla E(\mathbf{w}_k)\|$$

In the first iterations of every local minimization, it is sufficient to verify Condition (1) with a "large" ϵ_s.

Since $\mathbf{w}_k$ is such that $E(\mathbf{w}_{k+1}) < E(\mathbf{w}_k)$,

for $\epsilon_a > E(\mathbf{w}_1) - E_{\min}$, (1) is satisfied if

$$\epsilon_s = \frac{1}{2}\|\nabla E(\mathbf{w}_0)\|$$

... for $\epsilon_a > E(\mathbf{w}_r) - E_{\min}$, (1) is satisfied if

$$\epsilon_s = \frac{1}{2}\min_{k=0,\ldots,r}\|\nabla E(\mathbf{w}_k)\|$$

When ϵ_s is becoming "small" and $E(\mathbf{w}_r) >> E_{\min} = E(\mathbf{w}^*)$, then the sequence is converging to a stationary point $\hat{\mathbf{w}}$ which cannot correspond to the global minimum $E_{\min}$.

Setting

$$\begin{aligned} M_r &= \max_{k=0,\ldots,r} \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T \mathbf{s}_k} \\ N_r &= \max_{k=0,\ldots,r} \|A_k\| \|A_k^{-1}\| \end{aligned}$$

Setting

$$M_r = \max_{k=0,\ldots,r} \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T \mathbf{s}_k}$$
$$N_r = \max_{k=0,\ldots,r} \|A_k\| \|A_k^{-1}\|$$

$K = \max\{M_r, N_r\}$. It follows from Condition (3):

$$\forall k \leq r, \qquad \lambda_r \geq \frac{\|\mathbf{y}_k\|^2}{K \mathbf{y}_k^T \mathbf{d}_k}.$$

Setting

$$M_r = \max_{k=0,\dots,r} \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^T \mathbf{s}_k}$$
$$N_r = \max_{k=0,\dots,r} \|A_k\| \|A_k^{-1}\|$$

$K = \max\{M_r, N_r\}$. It follows from Condition (3):

$$\forall k \leq r, \qquad \lambda_r \geq \frac{\|\mathbf{y}_k\|^2}{K \mathbf{y}_k^T \mathbf{d}_k}.$$

Purpose:

Define $\mathbf{w}_{r+1}$ such that, by using the latter point as the new starting vector, the algorithm $\mathcal{L}QN$ is convergent to a stationary point $\hat{\hat{\mathbf{w}}}$, with $E(\hat{\hat{\mathbf{w}}}) < E(\hat{\mathbf{w}})$.

SOME PRELIMINARY IDEAS:

Suggestions for a proper *tunneling phase*.

SOME PRELIMINARY IDEAS:

Suggestions for a proper *tunneling phase*.

1) Compute $\mathbf{w}_{r+1}^{(1)} = \mathbf{w}_r - \lambda_{rep} \nabla E(\mathbf{w}_r)$, $\lambda_{rep} = \frac{\epsilon_a}{\|\nabla E(\mathbf{w}_r)\|^2}$

SOME PRELIMINARY IDEAS:

Suggestions for a proper *tunneling phase*.

1) Compute $\mathbf{w}_{r+1}^{(1)} = \mathbf{w}_r - \lambda_{rep} \nabla E(\mathbf{w}_r)$, $\lambda_{rep} = \frac{\epsilon_a}{\|\nabla E(\mathbf{w}_r)\|^2}$
2) Define $R_{r+1}(\mu)$: $rank(R_{r+1}(\mu)) = 2$, $\mu > 0$

$$R_{r+1}(\mu) = \frac{\mathbf{q}_r \mathbf{q}_r^T}{\mathbf{q}_r^T \mathbf{p}_r} - (I/\mu) \frac{\mathbf{p}_r \mathbf{p}_r^T}{\mathbf{p}_r^T \mathbf{p}_r}$$

being: $\mathbf{p}_r = \mathbf{w}_{r+1}^{(1)} - \mathbf{w}_r \quad \mathbf{q}_r = \nabla E(\mathbf{w}_{r+1}^{(1)}) - \nabla E(\mathbf{w}_r)$

SOME PRELIMINARY IDEAS:

Suggestions for a proper *tunneling phase*.

1) Compute $\mathbf{w}_{r+1}^{(1)} = \mathbf{w}_r - \lambda_{rep} \nabla E(\mathbf{w}_r)$, $\lambda_{rep} = \frac{\epsilon_a}{\|\nabla E(\mathbf{w}_r)\|^2}$
2) Define $R_{r+1}(\mu) : rank(R_{r+1}(\mu)) = 2, \quad \mu > 0$

$$R_{r+1}(\mu) = \frac{\mathbf{q}_r \mathbf{q}_r^T}{\mathbf{q}_r^T \mathbf{p}_r} - (I/\mu) \frac{\mathbf{p}_r \mathbf{p}_r^T}{\mathbf{p}_r^T \mathbf{p}_r}$$

being: $\mathbf{p}_r = \mathbf{w}_{r+1}^{(1)} - \mathbf{w}_r \quad \mathbf{q}_r = \nabla E(\mathbf{w}_{r+1}^{(1)}) - \nabla E(\mathbf{w}_r)$

By applying Sherman-Morrison-Woodbury formula, it follows:

SOME PRELIMINARY IDEAS:

Suggestions for a proper *tunneling phase*.

1) Compute $\mathbf{w}_{r+1}^{(1)} = \mathbf{w}_r - \lambda_{rep} \nabla E(\mathbf{w}_r)$, $\lambda_{rep} = \frac{\epsilon_a}{\|\nabla E(\mathbf{w}_r)\|^2}$
2) Define $R_{r+1}(\mu)$: $rank(R_{r+1}(\mu)) = 2$, $\mu > 0$

$$R_{r+1}(\mu) = \frac{\mathbf{q}_r \mathbf{q}_r^T}{\mathbf{q}_r^T \mathbf{p}_r} - (I/\mu)\frac{\mathbf{p}_r \mathbf{p}_r^T}{\mathbf{p}_r^T \mathbf{p}_r}$$

being: $\mathbf{p}_r = \mathbf{w}_{r+1}^{(1)} - \mathbf{w}_r \quad \mathbf{q}_r = \nabla E(\mathbf{w}_{r+1}^{(1)}) - \nabla E(\mathbf{w}_r)$

By applying Sherman-Morrison-Woodbury formula, it follows:

$$\left(I/\mu + R_{r+1}(\mu)\right)^{-1} = \mu I + (1 + \mu \frac{\mathbf{q}_r^T \mathbf{q}_r}{\mathbf{q}_r^T \mathbf{p}_r})\frac{\mathbf{p}_r \mathbf{p}_r^T}{\mathbf{q}_r^T \mathbf{p}_r} - \mu(\frac{\mathbf{p}_r \mathbf{q}_r^T + \mathbf{q}_r \mathbf{p}_r^T}{\mathbf{q}_r^T \mathbf{p}_r}), \quad \mu > 0$$

Thus, we obtain a **memoryless updating formula**.

3) Define $\mathbf{w}_{r+1}^{(2)}$:

$$\mathbf{w}_{r+1}^{(2)} = \mathbf{w}_{r+1}^{(1)} - \left(I/\mu_0 + R_{r+1}(\mu_0)\right)^{-1} \nabla E(\mathbf{w}_r) :$$

3) Define $\mathbf{w}_{r+1}^{(2)}$:

$$\mathbf{w}_{r+1}^{(2)} = \mathbf{w}_{r+1}^{(1)} - \left(I/\mu_0 + R_{r+1}(\mu_0) \right)^{-1} \nabla E(\mathbf{w}_r) :$$

$$E(\mathbf{w}_{r+1}^{(2)}) = \min_{\mu} \; E\left[\mathbf{w}_{r+1}^{(1)} - \left(I/\mu + R_{r+1}(\mu) \right)^{-1} \nabla E(\mathbf{w}_r) \right]$$

Therefore:

$$\mathbf{w}_{r+1}^{(2)} = \mathbf{w}_r - \left[\lambda_{rep} I + \left(I/\mu_0 + R_{r+1}(\mu_0) \right)^{-1} \right] \nabla E(\mathbf{w}_r)$$

3) Define $\mathbf{w}_{r+1}^{(2)}$:

$$\mathbf{w}_{r+1}^{(2)} = \mathbf{w}_{r+1}^{(1)} - \left(I/\mu_0 + R_{r+1}(\mu_0)\right)^{-1} \nabla E(\mathbf{w}_r) :$$

$$E(\mathbf{w}_{r+1}^{(2)}) = \min_{\mu} \; E\left[\mathbf{w}_{r+1}^{(1)} - \left(I/\mu + R_{r+1}(\mu)\right)^{-1} \nabla E(\mathbf{w}_r)\right]$$

Therefore:

$$\mathbf{w}_{r+1}^{(2)} = \mathbf{w}_r - \left[\lambda_{rep} I + \left(I/\mu_0 + R_{r+1}(\mu_0)\right)^{-1}\right] \nabla E(\mathbf{w}_r)$$

4) Evaluate:

$$E(\mathbf{w}_{r+1}^{(2)}) - E(\mathbf{w}_r)$$

if:

$$\begin{cases} E(\mathbf{w}_{r+1}^{(2)}) < E(\mathbf{w}_r) \quad \text{or} \\ E(\mathbf{w}_{r+1}^{(2)}) - E(\mathbf{w}_r) < \quad c\Big(E(\mathbf{w}_{r+1}^{(2)}) - E_{\min}\Big) \end{cases} \qquad (EC)$$

being $c(.)$ a suitable function

if:

$$\begin{cases} E(\mathbf{w}_{r+1}^{(2)}) < E(\mathbf{w}_r) \quad \text{or} \\ E(\mathbf{w}_{r+1}^{(2)}) - E(\mathbf{w}_r) < \quad c\Big(E(\mathbf{w}_{r+1}^{(2)}) - E_{\min}\Big) \end{cases} \qquad (EC)$$

being $c(.)$ a suitable function

Define $\mathbf{w}_{r+1} = \mathbf{w}_{r+1}^{(2)}$ and start a new local search.

Else

5) Set: $\mathbf{p}_r = \mathbf{w}_{r+1}^{(2)} - \mathbf{w}_r \quad \mathbf{q}_r = \nabla E(\mathbf{w}_{r+1}^{(2)}) - \nabla E(\mathbf{w}_r)$.

if:

$$\begin{cases} E(\mathbf{w}_{r+1}^{(2)}) < E(\mathbf{w}_r) \quad \text{or} \\ E(\mathbf{w}_{r+1}^{(2)}) - E(\mathbf{w}_r) < \quad c\Big(E(\mathbf{w}_{r+1}^{(2)}) - E_{\min}\Big) \end{cases} \qquad (EC)$$

being $c(.)$ a suitable function

Define $\mathbf{w}_{r+1} = \mathbf{w}_{r+1}^{(2)}$ and start a new local search.

Else

5) Set: $\mathbf{p}_r = \mathbf{w}_{r+1}^{(2)} - \mathbf{w}_r \quad \mathbf{q}_r = \nabla E(\mathbf{w}_{r+1}^{(2)}) - \nabla E(\mathbf{w}_r)$.

Solve the new minimum problem associated to the corresponding

$$\Big(I/\mu + R_{r+1}(\mu)\Big)^{-1}$$

If one of conditions (EC) is fulfilled, define:

$$\mathbf{w}_{r+1} = \mathbf{w}_{r+1}^{(3)}$$

and start a new local search.

....

If one of conditions (EC) is fulfilled, define:

$$\mathbf{w}_{r+1} = \mathbf{w}_{r+1}^{(3)}$$

and start a new local search.

....

Else:

6) Redefine λ_r by Condition 3 i.e.:

$$\frac{\|\mathbf{y}_r\|^2}{K\mathbf{y}_r^T\mathbf{d}_r} \leq \lambda_r \leq \frac{\epsilon_a}{\|\nabla E(\mathbf{w}_r)\|^2}$$

If one of conditions (EC) is fulfilled, define:

$$\mathbf{w}_{r+1} = \mathbf{w}_{r+1}^{(3)}$$

and start a new local search.

....

Else:

6) Redefine λ_r by Condition 3 i.e.:

$$\frac{\|\mathbf{y}_r\|^2}{K\mathbf{y}_r^T\mathbf{d}_r} \leq \lambda_r \leq \frac{\epsilon_a}{\|\nabla E(\mathbf{w}_r)\|^2}$$

and then assume:

$$\lambda_r < \frac{\epsilon_a}{\|\nabla E(\mathbf{w}_r)\|^2}$$

in the indicated admissible interval, thereby iterating the procedure.

Final Remarks:

- Every application of Shermann-Morrison-Woodbury formula has in our case a cost $O(n)$

Final Remarks:

- Every application of Shermann-Morrison-Woodbury formula has in our case a cost $O(n)$
- The one-dimensional optimal search of μ_0 can be efficiently performed by applying Armijo-Goldstein method

Final Remarks:

- Every application of Shermann-Morrison-Woodbury formula has in our case a cost $O(n)$
- The one-dimensional optimal search of μ_0 can be efficiently performed by applying Armijo-Goldstein method
- A satisfactory application of Theorem 1 $\Longrightarrow$ the fulfillment of Non–Suspiciousness Conditions depending on the operational values M, N and $\rightarrow$ K

Final Remarks:

- Every application of Shermann-Morrison-Woodbury formula has in our case a cost $O(n)$
- The one-dimensional optimal search of μ_0 can be efficiently performed by applying Armijo-Goldstein method
- A satisfactory application of Theorem 1 $\Longrightarrow$ the fulfillment of Non–Suspiciousness Conditions depending on the operational values M, N and $\rightarrow$ K , i.e.:

$$\begin{cases} \longrightarrow \text{the degree of } \textit{weak discrete convexity} \text{ of E} \\ \longrightarrow \text{the condition number of } \textit{repeller matrix approximations} \end{cases}$$

FUTURE RESEARCH:

- Define: R_{r+1}, $\quad rank(R) = 3, 4$, with a suitable structure

FUTURE RESEARCH:

- Define: R_{r+1}, $rank(R) = 3, 4$, with a suitable structure
- Apply a "Black dot algorithm" ([OT])
  Determine R_{r+1}'s rows and columns
  (Hints: From a rank-p matrix R, construct the R *skeleton decomposition*, by using the *black dot meta-arithmetic*)

FUTURE RESEARCH:

- Define: R_{r+1}, $rank(R) = 3, 4$, with a suitable structure
- Apply a "Black dot algorithm" ([OT])
  Determine R_{r+1}'s rows and columns
  (Hints: From a rank-p matrix R, construct the R *skeleton decomposition*, by using the *black dot meta-arithmetic*)
- Use a structured approximation of $\left(I/\mu + R_{r+1}(\mu)\right)^{-1}$

APPENDIX

Sherman-Morrison-Woodbury Formula in the general case

Given a square nonsingular matrix $A \in R^{n \times n}$, let U and V be matrices $\in R^{n \times p}$, $1 < p < n$. Define:

$$\hat{A} = A + UV^T$$

then:

$$\hat{A}^{-1} = A^{-1} - A^{-1}U(I + V^T A^{-1} U)^{-1} V^T A^{-1}$$

Remark

Inversion of $(I + V^T A^{-1} U)$ in inexpensive if p=3,4.

References

[DFZ0] C.Di Fiore, S.Fanelli, P.Zellini, Optimisation strategies for nonconvex functions and applications to neural networks, *ICONIP 2001*, Shanghai, 1, pp.453–458, 2001.

[DFZ1] C.Di Fiore, S.Fanelli, P.Zellini, Computational experiences of a novel algorithm for optimal learning in MLP-networks, *ICONIP 2002*, Singapore, 1, pp.317–321, 2002.

[DFLZ] C.Di Fiore, S. Fanelli, F. Lepore, P. Zellini, Matrix algebras in Quasi-Newton methods for unconstrained optimization, *Numerische Mathematik*, 94, pp. 479–500, 2003.

[BDFZ] A.Bortoletti, C.Di Fiore, S.Fanelli, P.Zellini, A new class of quasi-newtonian methods for optimal learning in MLP-networks, *IEEE Transactions on Neural Networks*, 14, pp. 263–273, 2003.

[DFZ2] C.Di Fiore, S.Fanelli, P.Zellini, An efficient generalization of Battiti-Shanno's Quasi-Newton Algorithm for learning in MLP-networks, *ICONIP'04*, Calcutta, pp.483–488, 2004.

[DFZ3] C. Di Fiore, S. Fanelli, P. Zellini, Low complexity minimization algorithms, *Numerical Linear Algebra with Applications*, 12, pp.755–768, 2005.

[DFZ4] C. Di Fiore, S. Fanelli, P. Zellini, Low complexity secant quasi-Newton minimization algorithms for non convex functions, *Journal of Computational and Applied Mathematics*, to appear.

[NW] J.Nocedal, S.J.Wright, *Numerical Optimization*,Springer-Verlag, 1999.

[P] M.J.D.Powell, Some global convergence properties of a variable metric algorithm for minimization without exact line search, *Nonlinear Programming, SIAM- AMS Proc.*, 9, pp. 53–72, 1976.

[OT] I. Oseledets. E.Tyrtyshnikov, A unifying approach to the construction of circulant preconditioners, *Linear Algebra and its Applications*, 418, pp. 435-449, 2006.

[T] E.Tyrtyshnikov, private communication