# Subspace methods: Find the dimension balance between approximation to optimization problem and subproblem solving

Pengcheng Xie
xpc@lsec.cc.ac.cn
Supervised by Prof. Ya-xiang Yuan

Institute of Computational Mathematics and Scientific/Engineering Computing
Academy of Mathematics and Systems Science
Chinese Academy of Sciences, China

Group Seminar
December 8, 2020

# Outline

- Introduction:
  Why study subspace methods?

- Subspace methods with different structure:
  How to design subspace methods?

- Conclusion and future work:
  What are wanted?

# PDFO[1] and image reconstruction in CT

An inverse problem in [Chen et al. 2017]: find a best $x \in \mathbb{R}^n$ which satisfies

$$f(x) = y \Rightarrow \min_{x \in \mathbb{R}^n} \|f(x) - y\|_2^2,$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ and $y \in \mathbb{R}^n$. We can use PDFO to solve.
Notice that $x$ or $y$ represent a long vector reshaped from matrix of $512 \times 512 = 262144$.
PDFO tells us an error:
"uobyqa: problem too large for uobyqa. Try other solvers."

Figure 1: 40 KeV monochromatic images of the DE-472 lung phantoms.

**Sad: This problem can not and do not have to be solved by DFO**

**Happy: Tom's words**

[1] Powell's Derivative-Free Optimization solvers : https://www.pdfo.net

# What did Tom[2] say and Zaikun's Subspace Method

*"In DFO, n=100 is considered as a large problem, n=200 is considered as a very large problem. I read once that NEWUOA has been tested with n=1000, but this is incredibly huge."*

*"Do you have any way to reduce the size of your problem, to find some kind of space (or lower dimension) in which your variables may belong (even approximately). If yes, you may try to restrict them uphill from calling PDFO (hopefully you are able to find a space of dimension some hundreds)."*

Solve subproblem on the subspace[Zhang 2012]

$$\mathfrak{S}_k = \operatorname{span}\left\{\nabla Q_k\left(x_k\right), d_{k-1}, \bar{d}_k\right\},$$

where

$$\bar{d}_k = \sum_{y \in I_k} \frac{f(y) - f\left(x_k\right)}{\|y - x_k\|_2} \cdot \frac{y - x_k}{\|y - x_k\|_2}$$

is a approximation to $\nabla f(x_k)$, where $I_k$ is the interpolation point set.

Precondition:

$$\mathfrak{S}_k = \operatorname{span}\left\{\tilde{g}_k, A_k \tilde{g}_k, s_{k-1}\right\}.$$

[2] Tom M. Ragonneau: Ph.D. Student in PolyU. Supervised by Prof. Zaikun Zhang and co-supervised by Prof. Xiaojun Chen.

# Zaikun's Subspace Method

**Algorithm 1** NEWUOAs

1: Given $x_1$ and $I_1$, s.t. $x_1 \in I_1$, and $f(x_1) = \min_{y \in I_1} f(y)$, Given $\Delta_1$, $k := 1$.
2: Model function $Q_k$: $Q_k(y) = f(y), y \in I_k$.
3: Solve the subspace trust region problem:

$$\min_{d \in \mathfrak{S}_k} Q_k(d)$$
$$\text{s.t. } \|d\| \leq \Delta_k,$$

Then get the trial step $s_k$.
4: If $f(x_k + s_k) < f(x_k)$, then $x_{k+1} := x_k + s_k$, otherwise $x_{k+1} := x_k$.
5: Judge whether the well-poisedness of the interpolation point set is good and update $I_k$.
6: Update $\Delta_{k+1}$. $k := k+1$, go to the step 2.

NEWUOA: dimension $<$ 1000
NEWUOAs: dimension = 2000

# Optimization problem and its subproblem

Optimization problem
Find $x^*$ satisfies

$$\min_x f(x)$$
$$\text{s.t. } x \in X.$$

Subproblem
Find $x_{k+1} = x_k + d$ satisfies

$$\min_d m_k(x_k + d)$$
$$\text{s.t. } d \in D.$$

# Choose $x_{k+1}$ from $x_k$ in subproblem

Line search method

1. Generate a descent search direction $d_k$
2. Search along this direction for a step size $\alpha_k$

$$x_{k+1} = x_k + \alpha_k d_k.$$

1-dimension problem

Trust region method

1. Given trust region radius whose role is similar to the step size.
2. Compute a search direction in trust region.

$$\min_{s\in\mathbb{R}^n} Q_k(s) = g_k^\top s + \frac{1}{2} s^\top B_k d$$
$$\text{s.t. } \|s\|_2 \le \Delta_k$$

n-dimension problem

**Where is the mediant dimension problem?**
$(1 < \textbf{mediant} < n)$

# Why do we need the mediant dimension problem

You may ask: “There is no need to deliberately produce mediant dimension problem. We like 1 and $n$.”
Balance between Optimization problem and the Subproblem:
***Find the balance between Looking for direction and looking for stepsize[3].***

- Reduce the dimension.
- Gather more.
- Special problem or needs.

[Conn et al. 1994] :
We consider it important from a practical point of view to require that $\mathfrak{S}_k$ contains at least two components:

- a Gradient-related direction, such as $-g(k)$, to encourage global convergence.
- a Newton-related direction, to encourage fast asymptotic convergence, with safeguards to account for indefiniteness.

[3] Prof. Ya-xiang Yuan said on ICM 2014

# Typical scenarios to design subspace methods

[Liu, Wen and Yuan 2020][4]

Subproblem: $x_k \to x_{k+1}$:

$$\min_d m_k(x_k + d)$$
$$\text{s.t. } d \in D$$

**Find a linear combination of several known directions.**

Linear and nonlinear conjugate gradient methods[Sun and Yuan 2006; Nocedal and Wright 2006]

Nesterov's accelerated gradient method[Nesterov 2003; Nesterov 1983]

Heavy-ball method[Polyak 1964]

Momentum method[Goodfellow, Bengio, and Courville 2016]

Problem:

$$\min_x f(x)$$
$$\text{s.t. } x \in X$$

**Keep the objective function and constraints, but add an extra restriction in a certain subspace.**

OMP[Tropp and Gilbert 2008]

CoSaMP[Needell and Tropp 2010]

LOBPCG[Conjugategradi and Knyazev 2001]

LMSVD[Liu, Wen, and Zhang 2013]

Subspace refinement and multilevel methods

[4] Subspace Methods for Nonlinear Optimization: http://bicmr.pku.edu.cn/ wenzw/paper/SubOptv.pdf

# Typical scenarios to design subspace methods

Subproblem:

$x_k \to x_{k+1}$:

$$\min_d m_k(x_k + d)$$
$$\text{s.t. } d \in D$$

**Approximate the objective function but keep the constraints.**

BCD[Tseng and Yun 2009]

RBR[Wen, Goldfarb, and Scheinberg 2012]

Trust region with subspaces[Shultz, Schnabel, and Byrd 1985]

Parallel subspace correction[Fornasier 2007; Fornasier and Schönlieb 2008]

Problem:

$$\min_x f(x)$$
$$\text{s.t. } x \in X$$

**Use subspace techniques to approximate the objective functions.**

Sampling/Sketching[Goodfellow, Bengio, and Courville 2016; Mahoney 2011]

Nystrom approximation[Tropp et al. 2017]

**Approximate the objective function and design new constraints.**

Trust region with subspaces

FPC_AS[Wen et al. 2010]

# Typical scenarios to design subspace methods

Subproblem: $x_k \to x_{k+1}$:

$$\min_d m_k(x_k + d)$$
$$\text{s.t. } d \in D$$

Problem:

$$\min_x f(x)$$
$$\text{s.t. } x \in X$$

**Add a postprocess procedure after the subspace problem is solved.**

Truncated subspace method for tensor train[Zhang, Wen, and Zhang 2016]

**Integrate the optimization method and subspace update in one framework.**

Polynomial-filtered subspace method for low-rank matrix optimization
[Liu, Wen and Yuan 2020]

# Subspace Relationship

$dim(\mathfrak{S}_k) = dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \approx \mathfrak{S}_{k+1}$

Fix-dimension Subspaces:
Direction-Gradient Subspaces
One-add-one-drop Subspaces

$dim(\mathfrak{S}_k) \leq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \subseteq \mathfrak{S}_{k+1}$

Nested Subspaces

$\sum_{k=1}^{n} dim(\mathfrak{S}_k) = p$: $\mathfrak{S}_1 + \cdots + \mathfrak{S}_n = \mathbb{R}^p$

Complement Subspaces

$dim(\mathfrak{S}_k) \geq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \supseteq \mathfrak{S}_{k+1}$

Active methods

$dim(\mathfrak{S}_k) = i_k$: $\mathfrak{S}_k = I_k$

Subsampling/Sketching
Stochastic Optimization

# Subspace Relationship

$dim(\mathfrak{S}_k) = dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \approx \mathfrak{S}_{k+1}$

Fix-dimension Subspaces:
**Direction-Gradient Subspaces**
One-add-one-drop Subspaces

$dim(\mathfrak{S}_k) \leq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \subseteq \mathfrak{S}_{k+1}$

Nested Subspaces

$\sum_{k=1}^{n} dim(\mathfrak{S}_k) = p$: $\mathfrak{S}_1 + \cdots + \mathfrak{S}_n = \mathbb{R}^p$

Complement Subspaces

$dim(\mathfrak{S}_k) \geq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \supseteq \mathfrak{S}_{k+1}$

Active methods

$dim(\mathfrak{S}_k) = i_k$: $\mathfrak{S}_k = I_k$

Subsampling/Sketching
Stochastic Optimization

# Direction-Gradient Subspace Method for $x \in \mathbb{R}^n$

Linear combination of several known direction

- Conjugate gradient methods:

$$
\begin{aligned}
d_k &= -g_k + \beta_{k-1} d_{k-1}, \\
\mathfrak{S}_k &= \operatorname{span}\{g_k, d_{k-1}, x_k\}.
\end{aligned}
$$

- Nesterov's accelerated gradient methods (FISTA method)[Beck and Teboulle 2009], [Nesterov 2003]:

$$
\begin{aligned}
y_k &= x_{k-1} + \frac{k-2}{k+1}\left(x_{k-1} - x_{k-2}\right), \\
x_k &= y_k - \alpha_k \nabla f\left(y_k\right), \\
\mathfrak{S}_k &= \operatorname{span}\left\{x_{k-1}, x_{k-2}, \nabla f\left(y_k\right)\right\}.
\end{aligned}
$$

- Heavy-ball method[Polyak 1964]:

$$
\begin{aligned}
d_k &= -g_k + \beta d_{k-1}, \\
x_{k+1} &= x_k + \alpha_k d_k, \\
\mathfrak{S}_k &= \operatorname{span}\{g_k, d_{k-1}, x_k\}.
\end{aligned}
$$

# Limited memory methods for eigenvalue Computation

Finding a $p$-dimensional eigenspace associated with $p$ largest eigenvalues of $A$ is equivalent to solving problems the optimization problem:

$$\max_{X \in \mathbb{R}^{n\times p}} \operatorname{tr}\left(X^\top AX\right), \text{ s.t. } X^\top X = I. \tag{1}$$

The first-order optimality conditions of (1) are

$$AX = X\Lambda,\ X^\top X = I,$$

where $\Lambda = X^\top AX \in \mathbb{R}^{p\times p}$ is the matrix of Lagrangian multipliers.
At each iteration, the methods solve a subspace trace maximization problem

$$Y = \operatorname*{arg\,max}_{X \in \mathbb{R}^{n\times p}} \left\{ \operatorname{tr}\left(X^\top AX\right) : X^\top X = I, X \in \mathfrak{S} \right\}.$$

LOBPCG [Conjugategradi and Knyazev 2001]: $\mathfrak{S} = \operatorname{span}\{X_{i-1}, X_i, AX_i\}$.
LMSVD [Liu, Wen, and Zhang 2013]:$\mathfrak{S} = \operatorname{span}\{X_i, X_{i-1}, \cdots, X_{i-\tau}\}$.

# Truncated Subspace Method for Tensor Train

$x \in \mathbb{R}^n \to \mathbf{x} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$: [Zhang, Wen, and Zhang 2016]

$$x_{i_1 i_2 \ldots i_d} = X_1(i_1) X_2(i_2) \cdots X_d(i_d).$$

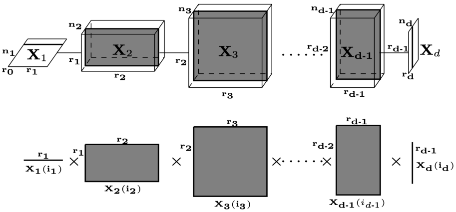


Figure 2: $x_{i_1 i_2 \ldots i_d} = X_1(i_1) X_2(i_2) \cdots X_d(i_d)$
TT format

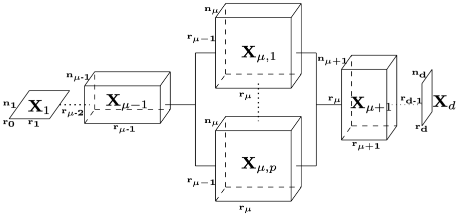


Figure 3:
$X(i_1, \ldots, i_\mu, \ldots, i_d; j) = X_1(i_1) \cdots X_{\mu,j}(i_\mu) \cdots X_d(i_d)$
$\mu$-BTT format

$$A_{i_1 i_2 \cdots i_d, j_1 j_2 \cdots j_d} = A_1(i_1, j_1) A_2(i_2, j_2) \cdots A_d(i_d, j_d).$$

where $A_\mu(i_\mu, j_\mu) \in \mathbb{R}^{r_{\mu-1} \times r_\mu}$ for $i_\mu, j_\mu \in \{1, \ldots, n_\mu\}$.

# Truncated Subspace Method for Tensor Train

Then the eigenvalue problem in the BTT format is

$$\min_{X\in\mathbb{R}^{n\times p}} \operatorname{tr}\left(\mathbf{X}^\top \mathbf{A}\mathbf{X}\right), \quad \text{s.t.} \quad \mathbf{X}^\top\mathbf{X} = I_p \text{ and } \mathbf{X} \in \mathbf{T}_{\mathbf{n},r,p}.$$

One can choose either the following subspace

$$\mathfrak{S}_k^{\mathrm{T}} = \operatorname{span}\left\{P_{\mathbf{T}}\left(\mathbf{A}\mathbf{X}_k\right), \mathbf{X}_k, \mathbf{X}_{k-1}\right\},$$

or a subspace with two truncations as

$$\mathfrak{S}_k^{\mathbf{T}} = \operatorname{span}\left\{\mathbf{X}_k, P_{\mathbf{T}}\left(\mathbf{R}_k\right), P_{\mathbf{T}}\left(\mathbf{P}_k\right)\right\}.$$

The subspace problem in the BTT format is

$$\mathbf{Y}_{k+1} := \arg\min_{X\in\mathbb{R}^{n\times p}} \operatorname{tr}\left(\mathbf{X}^\top \mathbf{A}\mathbf{X}\right), \text{ s.t. } \mathbf{X}^\top\mathbf{X} = I_p, \mathbf{X} \in \mathfrak{S}_k^{\mathrm{T}}, \tag{2}$$

which is equivalent to a generalized eigenvalue decomposition problem:

$$\min_{V\in\mathbb{R}^{q\times p}} \operatorname{tr}\left(V^\top\left(S^\top A S\right)V\right), \text{ s.t. } V^\top S^\top S V = I_p.$$

Electronic Structure Calculations

# Subspace Relationship

$dim(\mathfrak{S}_k) = dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \approx \mathfrak{S}_{k+1}$

Fix-dimension Subspaces:
Direction-Gradient Subspaces
**One-add-one-drop Subspaces**

$dim(\mathfrak{S}_k) \leq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \subseteq \mathfrak{S}_{k+1}$

Nested Subspaces

$\sum_{k=1}^{n} dim(\mathfrak{S}_k) = p$: $\mathfrak{S}_1 + \cdots + \mathfrak{S}_n = \mathbb{R}^p$

Complement Subspaces

$dim(\mathfrak{S}_k) \geq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \supseteq \mathfrak{S}_{k+1}$

Active methods

$dim(\mathfrak{S}_k) = i_k$: $\mathfrak{S}_k = I_k$

Subsampling/Sketching
Stochastic Optimization

L-BFGS matrix $B_k$ and inverse matrix $H_k$, generated from a few most recent pairs $\{s_i, y_i\}$, where $s_i = x_{i+1} - x_i,\ y_i = g_{i+1} - g_i$.[Sun and Yuan 2006], [Nocedal and Wright 2006]

Then the search direction is $d_k = -B_k^{-1} g_k = -H_k g_k$ (Both $B_k$ and $H_k$ can be written in a compact representation[Byrd, Nocedal, and Schnabel 1997].

Assume that there are $p$ pairs of vectors:

$$U_k = \left[s_{k-p}, \ldots, s_{k-1}\right] \in \mathbb{R}^{n\times p}, \quad Y_k = \left[y_{k-p}, \ldots, y_{k-1}\right] \in \mathbb{R}^{n\times p}.$$

For a given initial matrix $H_k^0$, the $H_k$ matrix is $H_k = H_k^0 + C_k P_k C_k^\top$, where

$$C_k := \left[U_k, H_k^0 Y_k\right] \in \mathbb{R}^{n\times 2p}, \quad D_k = \operatorname{diag}\left[s_{k-p}^\top y_{k-p}, \ldots, s_{k-1}^\top y_{k-1}\right],$$

$$P_k := \begin{bmatrix} R_k^{-\top}\left(D_k + Y_k^\top H_k^0 Y_k\right) R_k^{-1} & -R_k^{-\top} \\ -R_k^{-1} & 0 \end{bmatrix}, (R_k)_{i,j} = \begin{cases} s_{k-p+i-1}^\top y_{k-p+j-1}, & \text{if } i \le j, \\ 0, & \text{o.w.} \end{cases}$$

The initial matrix $H_k^0$ is $\gamma_k I$. Then

$$d_k \in \operatorname{span}\left\{g_k, s_{k-1}, \ldots, s_{k-p}, y_{k-1}, \ldots, y_{k-p}\right\}.$$

# Limited Memory Methods

Finding a $p$-dimensional eigenspace associated with $p$ largest eigenvalues of $A$ is equivalent to solving a trace maxmization problem with orthogonality constraints:

$$\max_{X \in \mathbb{R}^{n \times p}} \operatorname{tr}\left(X^\top A X\right), \text{ s.t. } X^\top X = I. \tag{3}$$

The first-order optimality conditions of (3) are $AX = X\Lambda$, $X^\top X = I$, where $\Lambda = X^\top A X \in \mathbb{R}^{p \times p}$ is the matrix of Lagrangian multipliers.

$$Y = \operatorname*{arg\,max}_{X \in \mathbb{R}^{n \times p}} \left\{ \operatorname{tr}\left(X^\top A X\right) : X^\top X = I, X \in \mathfrak{S} \right\}. \tag{4}$$

RR procedure $\Rightarrow$ the closed-form solution of (4).
LOBPCG[Conjugategradi and Knyazev 2001]: $\mathfrak{S} = \operatorname{span}\{X_{i-1}, X_i, AX_i\}$.
LMSVD[Liu, Wen, and Zhang 2013]: $\mathfrak{S} = \operatorname{span}\{X_i, X_{i-1}, \ldots, X_{i-t}\}$.

# Subspace Relationship

$dim(\mathfrak{S}_k) = dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \approx \mathfrak{S}_{k+1}$

Fix-dimension Subspaces:
Direction-Gradient Subspaces
One-add-one-drop Subspaces

$dim(\mathfrak{S}_k) \leq dim(\mathfrak{S}_{k+1})$**:** $\mathfrak{S}_k \subseteq \mathfrak{S}_{k+1}$

**Nested Subspaces**

$\sum_{k=1}^{n} dim(\mathfrak{S}_k) = p$: $\mathfrak{S}_1 + \cdots + \mathfrak{S}_n = \mathbb{R}^p$

Complement Subspaces

$dim(\mathfrak{S}_k) \geq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \supseteq \mathfrak{S}_{k+1}$

Active methods

$dim(\mathfrak{S}_k) = i_k$: $\mathfrak{S}_k = I_k$

Subsampling/Sketching
Stochastic Optimization

# Trust Region Methods with Subspace Method

The trust region subproblem (TRS) is normally

$$\begin{aligned} \min_{s\in\mathbb{R}^n} \ & Q_k(s) = g_k^\top s + \frac{1}{2} s^\top B_k d \\ \text{s.t. } & \|s\|_2 \le \Delta_k, \end{aligned} \tag{5}$$

where $B_k$ is an approximation to the Hessian and $\Delta_k$ is the trust region radius.
A subspace version of the trust region subproblem is suggested in [Shultz, Schnabel, and Byrd 1985]

$$\begin{aligned} \min_{s\in\mathbb{R}^n} \ & Q_k(s) \\ \text{s.t. } & \|s\|_2 \le \Delta_k, \quad s \in \mathfrak{S}_k. \end{aligned} \tag{6}$$

The Steihaug truncated CG method [Steihaug 1983] for solving (5) is a subspace method.
$B_k$: quasi-Newton updates SR1, PSB or the Broyden family [Sun and Yuan 2006], the TRS has subspace properties.

# Parallel Computing Experiment of trust region methods based on truncated CG method [5]

Table 1: Speedup Ratio

| dimension | np=2 | np=4 | np=6 |
|---|---|---|---|
| $100 = 10^2$ | 1.68180 | 1.94154 | 2.36451 |
| $1000 = 10^3$ | 0.920956 | 1.47545 | 1.55419 |
| $10000 = 10^4$ | 1.79342 | 2.94063 | 3.86112 |
| $50000 = 5 \times 10^4$ | 1.87369 | 3.04962 | 3.94852 |
| $100000 = 10^5$ | 1.89060 | 3.55094 | 5.17231 |
| | np=8 | np=10 | np=12 |
| $100 = 10^2$ | 2.91613 | 3.43903 | 3.67575 |
| $1000 = 10^3$ | 1.84841 | 2.43805 | 2.64320 |
| $10000 = 10^4$ | 4.49823 | 4.94911 | 5.18691 |
| $50000 = 5 \times 10^4$ | 5.10126 | 6.29814 | 6.71970 |
| $100000 = 10^5$ | 5.88022 | 6.52538 | 7.02531 |

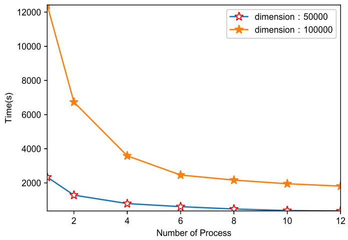


Figure 4: Time vesus number of process.

[5] Homework of Parallel Computing supervised by Prof. Tao Cui

# Trust Region Methods with Subspace Method

**Theorem**

*Suppose $B_1 = \sigma I$, with $\sigma > 0$, let $s_k$ be an optimal solution of TRS (5) and set $x_{k+1} = x_k + s_k$. Let $\mathfrak{S}_k = \text{span}\{g_1, g_2, \cdots, g_k\}$. Then for $s_k \in \mathfrak{S}_k$ and for any $z \in \mathfrak{S}_k, \mu \in \mathfrak{S}_k^{\perp}$, it holds*

$$B_k z \in \mathfrak{S}_k, \quad B_k u = \sigma u.$$

- Subspace trust region quasi-Newton method for unconstrained optimization[Wang and Yuan 2006] .

  Assume that $B$ is a limited memory quasi-Newton matrix which can be expressed as

$$B = \sigma I + PDP^{\top}, \quad P \in \mathbb{R}^{n \times l}, \|s\|_P = \max\left\{\left\|P^{\top} s\right\|_{\infty}, \left\|P_{\perp}^{\top} s\right\|_2\right\}.$$

- Line search quasi-Newton methods [Gill and Leonard 1999; Gill and Leonard 2000].
- Subspace Powell–Yuan trust region method for equality constrained optimization[Grapiglia, Yuan, and Yuan 2013].

# Augmented Rayleigh-Ritz Method for eigenvalue computation

The RR map $(Y,\Sigma) = \mathrm{RR}(A,Z)$ is equivalent to solving the trace-maximization subproblem with the subspace $\mathfrak{S} = R(Z)$, the augmentation of the subspaces in LOGPCG and LMSVD is the main reason why they generally achieve faster convergence than the classic SSI.

ARR: For some integer $t \geq 0$, design a block Krylov subspace strycture:

$$\mathfrak{S} = \mathrm{span}\left\{X, AX, A^2X, \ldots, A^tX\right\}. \tag{7}$$

Then the optimal solution of the trace maximization problem , restricted in the subspace $\mathfrak{S}$ in (7), is computed via the RR procedure using $(\hat{Y},\hat{\Sigma}) = \mathrm{RR}(A,K_t)$, where $K_t = \left[X, AX, A^2X, \ldots, A^tX\right]$. Finally, the $p$ leading Ritz pairs $(Y,\Sigma)$ is extracted from $(\hat{Y},\hat{\Sigma})$ .

The analysis of ARR in [Wen and Zhang 2017] shows that the convergence rate of SSI is improved from $\left|\rho\left(\lambda_{p+1}\right)/\rho\left(\lambda_p\right)\right|$ for $\mathrm{RR}(t=0)$ to $\left|\rho\left(\lambda_{(t+1)p+1}\right)/\rho\left(\lambda_p\right)\right|$ for ARR $(t>0)$.

# Subspace Relationship

$dim(\mathfrak{S}_k) = dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \approx \mathfrak{S}_{k+1}$

Fix-dimension Subspaces:
Direction-Gradient Subspaces
One-add-one-drop Subspaces

$dim(\mathfrak{S}_k) \leq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \subseteq \mathfrak{S}_{k+1}$

Nested Subspaces

$\sum_{k=1}^{n} dim(\mathfrak{S}_k) = p$: $\mathfrak{S}_1 + \cdots + \mathfrak{S}_n = \mathbb{R}^p$

**Complement Subspaces**

$dim(\mathfrak{S}_k) \geq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \supseteq \mathfrak{S}_{k+1}$

Active methods

$dim(\mathfrak{S}_k) = i_k$: $\mathfrak{S}_k = I_k$

Subsampling/Sketching
Stochastic Optimization

# Coordinate Descent Methods

According to [He and Buccafusca 2016]

**Algorithm 2** Coordinate Descent Algorithm

1: Input initial value $x^{(0)}$.
2: For $t = 1, 2, \ldots$.
3: Pick coordinate $i$ from $1, 2, \ldots n$,

$$x_i^{(t+1)} = \arg\min_{x_i \in \mathbb{R}} f\left(x_i, \omega_{-i}^t\right).$$

4: End.

where $\omega_{-i}^t$ represent all other coordinates.

- Convergent slowly.
- Does not require calculation of the gradient $\nabla f_k$.
- Several algorithms, such as that of Hooke and Jeeves, are based on these ideas; see [Mackworth 1987], [Ricketts 1982].

# Parallel Line Search Subspace Correction Method

In this subsection, we consider the following optimization problem

$$\min_{x\in\mathbb{R}^n} \varphi(x) := f(x) + h(x), \tag{8}$$

where $f(x)$ is differentiable convex function and $h(x)$ is a convex function that is possibly nonsmooth. Suppose that $\mathbb{R}^n$ is split into $p$ subspaces, namely,

$$\mathbb{R}^n = X_1 + X_2 + \cdots + X_p,$$

where

$$X_i = \{x \in \mathbb{R}^n \mid \operatorname{supp}(x) \subset J_i\}, \quad 1 \leq i \leq p,$$

such that $J := \{1, \ldots, n\}$ and $J = \bigcup_{i=1}^{p} J_i$.

Let $\varphi_k^{(i)}$ be a surrogate function of $\varphi$ restricted to the $i$-th subspace at $k$-th iteration. The PSC framework for solving (8) is:

$$d_k^{(i)} = \arg\min_{d^i \in X^i} \varphi_k^{(i)}\left(d^{(i)}\right), i = 1, \ldots, p, \tag{9}$$

$$x_{k+1} = x_k + \sum_{i=1}^{p} \alpha_k^{(i)} d_k^{(i)}$$

# Parallel Line Search Subspace Correction Method

The convergence can be proved if the step sizes $\alpha_k^{(i)}(1 \leq i \leq p)$ satisfy the conditions: $\sum_{i=1}^{p} \alpha_k^{(i)} \leq 1$ and $\alpha_k^{(i)} > 0(1 \leq i \leq p)$. Usually, the step size $\alpha_k^{(i)}$ is quite small under these conditions and convergence becomes slow.
A parallel subspace correction method (PSCL) is proposed in [Dong et al. 2015].
At the $k$-th iteration, The next iteration is

$$x_{k+1} = x_k + \alpha_k d_k,$$

where $\alpha_k$ satisfies the Armijo backtracking conditions. When $h(x) = 0$ and $f(x)$ is strongly convex, the surrogate function can be set to the original objective function $\varphi$. Otherwise,

$$\varphi_k^i\left(d^{(i)}\right) = \nabla f\left(x_k\right)^\top d^{(i)} + \frac{1}{2\lambda^i}\left\|d^{(i)}\right\|_2^2 + h\left(x_k + d^{(i)}\right), \text{ for } d^{(i)} \in X^i.$$

Both non-overlapping and overlapping schemes can be designed for PSCL.

# Parallel Line Search Subspace Correction Method

The directions from different subproblems can be equipped with different step sizes. Let $Z_k = \left(d_k^{(1)}, d_k^{(2)}, \ldots, d_k^{(p)}\right)$. The next iteration is set to

$$x_{k+1} = x_k + Z_k \alpha_k. \ \alpha_k = \arg\min_{\alpha \in \mathbb{R}^p} \varphi\left(x_k + Z_k \alpha\right).$$

Alternatively, we can solve the following approximation:

$$a_k \approx \arg\min_{\alpha \in \mathbb{R}^p} \nabla f\left(x_k\right)^\top Z_k \alpha + \frac{1}{2t_k} \left\|Z_k \alpha\right\|_2^2 + h\left(x_k + Z_k a\right).$$

- The global convergence of PSCL is established by following the convergence analysis of the subspace correction methods for strongly convex problem [Tai and Xu 2003].
- The active-set method for $l_1$ minimization and the BCD method for nonsmooth separable minimization [Tseng and Yun 2009].
- Specifically, linear convergence rate is proved for the strongly convex case and convergence to the solution set of problem (8) globally is obtained for the general nonsmooth case.

# Subspace Relationship

$dim(\mathfrak{S}_k) = dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \approx \mathfrak{S}_{k+1}$

Fix-dimension Subspaces:
Direction-Gradient Subspaces
One-add-one-drop Subspaces

$dim(\mathfrak{S}_k) \leq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \subseteq \mathfrak{S}_{k+1}$

Nested Subspaces

$\sum_{k=1}^{n} dim(\mathfrak{S}_k) = p$: $\mathfrak{S}_1 + \cdots + \mathfrak{S}_n = \mathbb{R}^p$

Complement Subspaces

$dim(\mathfrak{S}_k) \geq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \supseteq \mathfrak{S}_{k+1}$

**Active methods**

$dim(\mathfrak{S}_k) = i_k$: $\mathfrak{S}_k = I_k$

Subsampling/Sketching
Stochastic Optimization

# Active Set Methods

Consider the $\ell_1$-regularized minimization problem

$$\min_{x\in\mathbb{R}^n} \psi_\mu(x) := \mu\|x\|_1 + f(x), \tag{10}$$

where $\mu > 0$ and $f(x) : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. The optimality condition of (10) is that there exists a vector

$$(\nabla f(x))^i \begin{cases} = -\mu, & x_i > 0 \\ = +\mu, & x_i < 0 \\ \in [-\mu, \mu], & \text{otherwise} . \end{cases}$$

FPC_AS[Wen et al. 2010], a two-stage active set algorithm, for an initial point $x_0$

$$x_{k+1} := \arg\min_x \quad \mu\|x\|_1 + (x - x_k)^\top g_k + \frac{1}{2\alpha_k}\|x - x_k\|_2^2,$$

where $g_k := \nabla f(x_k)$ and $\alpha_k > 0$.

$$x_{k+1} = S(x_k - \alpha_k g_k, \mu\alpha_k), \tag{11}$$

where for $y \in \mathbb{R}^n$ and $\nu \in \mathbb{R}$, the shrinkage operator is defined as

$$S(y, \nu) = \arg\min_x \nu\|x\|_1 + \frac{1}{2}\|x - y\|_2^2 = \operatorname{sign}(y) \odot \max\{|y| - \nu, \mathbf{0}\}.$$

# Active Set Methods

The convergence of (11) has been studied in [Hale, Yin, and Zhang 2008] under suitable conditions on $\alpha_k$ and the Hessian $\nabla^2 f$.
Subspace optimization in the second stage. For a given vector $x \in \mathbb{R}^n$:

$$A(x) := \left\{ i \in \{1, \cdots, n\} \middle| |x^i| = 0 \right\} \text{ and } I(x) := \left\{ i \in \{1, \cdots, n\} \middle| |x^i| > 0 \right\}.$$

We require that each component $x^i$ either has the same sign as $x_k^i$ or is zero, i.e., $x$ is required to be in the set

$$\Omega(x_k) := \left\{ x \in \mathbb{R}^n : \operatorname{sign}\left(x_k^i\right) x^i \geq 0,\ i \in I(x_k) \text{ and } x^i = 0, i \in A(x_k) \right\}.$$

Then, a smooth subproblem is formulated as either an essentially unconstrained problem

$$\min_x \ \mu \operatorname{sign}\left(x_k^{I_k}\right)^\top x^{I_k} + f(x), \text{ s.t. } x^i = 0,\ i \in A(x_k), \tag{12}$$

- Problem (12) can be solved by L-BFGS-B.
- The active set strategies have also been studied in [Solntsev, Nocedal, and Byrd 2014; Keskar et al. 2015].

# Subspace Relationship

$dim(\mathfrak{S}_k) = dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \approx \mathfrak{S}_{k+1}$

Fix-dimension Subspaces:
Direction-Gradient Subspaces
One-add-one-drop Subspaces

$dim(\mathfrak{S}_k) \leq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \subseteq \mathfrak{S}_{k+1}$

Nested Subspaces

$\sum_{k=1}^{n} dim(\mathfrak{S}_k) = p$: $\mathfrak{S}_1 + \cdots + \mathfrak{S}_n = \mathbb{R}^p$

Complement Subspaces

$dim(\mathfrak{S}_k) \geq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \supseteq \mathfrak{S}_{k+1}$

Active methods

$dim(\mathfrak{S}_k) = i_k$: $\mathfrak{S}_k = I_k$

**Subsampling/Sketching**
Stochastic Optimization

# Subspace by Subsampling/Sketching

For a linear least squares problem on massive data sets:

$$\min_x \ \|Ax-b\|_2^2, \to \min_x \ \|W(Ax-b)\|_2^2. \tag{13}$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ . The sketching technique chooses a matrix $W \in \mathbb{R}^{r\times m}$ with $r \ll m$ and formulates a reduced problem

Randomly select $r$ rows from the identity matrix to form $W$ so that $WA$ is a submatrix of $A$.

Each element of $W$ is sampled from an i.i.d. normal random variable with mean zero and variance $\frac{1}{r}$[Mahoney 2011], [Woodruff 2014].

Consider the system of nonlinear equations

$$F(x) = 0, \ x \in \mathbb{R}^n \tag{14}$$

and nonlinear least squares problem $\min_{x\in\mathbb{R}^n} \|F(x)\|_2^2$, where $F(x) = (F^1(x), F^2(x), \cdots, F^m(x))^\top \in \mathbb{R}^m$.

Consider $F_i(x) = 0, \ i \in I_k$. To solve the nonlinear equations (14) is to find a $x$ at which $F$ maps to the origion[Yuan 2009].

# Eigenvalue Computation

For a given real symmetric matrix $A \in \mathbb{R}^{n\times n}$, suppose $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and $q_1, \ldots, q_n \in \mathbb{R}^n$ satifies $Aq_i = \lambda_i q_i,\ \|q_i\|_2 = 1,\ i = 1, \ldots, n$ and $q_i^\top q_j = 0$ for $i \neq j$. $A = Q_n \Lambda_n Q_n^\top$ , where, for any integer $i \in [1, n]$,

$$Q_i = [q_1, q_2, \ldots, q_i] \in \mathbb{R}^{n\times i}, \quad \Lambda_i = \mathrm{Diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_i\right) \in \mathbb{R}^{i\times i}, \tag{15}$$

For simplicity, we also write $A = Q\Lambda Q^\top$ where $Q = Q_n$ and $\Lambda = \Lambda_n$. The Rayleigh-Ritz (RR) step consists of the following four steps:

(i) Given $Z \in \mathbb{R}^{n\times m}$, orthonormalize $Z$ to obtain $U \in \mathrm{orth}(Z)$, where orth $(Z)$ is the set of orthonormal bases for the range space of $Z$.
(ii) Compute $H = U^\top AU \in \mathbb{R}^{m\times m}$, the projection of $A$ onto the range space of $U$.
(iii) Compute the eigenvalue decomposition $H = V^\top \Sigma V$, where $V^\top V = I$ and $\Sigma$ is diagonal.
(iv) Assemble the Ritz pairs $(Y, \Sigma)$ where $Y = UV \in \mathbb{R}^{n\times m}$ satisfies $Y^\top Y = I$.
The RR procedure is denoted as a map $(Y, \Sigma) = \mathrm{RR}(A, Z)$ where the output $(Y, \Sigma)$ is a Ritz pair block.

# Simple Subspace iteration method for Eigenvalue Computation

SSI: The simple (simultaneous) subspace iteration (SSI) method [Rutishauser 1969], [Rutishauser 1970], [Stewart 1976], [Stewart and Jennings 1981], starting from an initial matrix $U$,

$$\text{orthogonalization} : Z = \text{orth}(AU).$$
$$\text{RR projection} : U = \text{RR}(A, Z).$$

The convergence rates for different eigenpairs are not the same. $q$ extra vectorsare added to $U$ to accelerate convergence. Although the iteration cost is increased at the initial stage, the overall performance may be better.

- Simultaneous matrix-block multiplications have advantages over individual matrix-vector multiplications.
- Whenever there is a gap between the $p$-th and the $(p+1)$-th eigenvalues of $A$, the SSI method is ensured to converge to the largest $p$ eigenpairs from any generic starting point.
- SSI method converges slow if the eigenvalue distributions are not favorable.

Simultaneous matrix-block multiplications have advantages over individual matrix-vector multiplications.

# Subspace By Coordinate Directions

For sparsity structures. Let $g_k^i$ be the $i$-th component of the gradient $g_k$, satisfies

$$\left|g_k^{i_1}\right| \geq \left|g_k^{i_2}\right| \geq \left|g_k^{i_3}\right| \geq \cdots \geq \left|g_k^{i_n}\right|.$$

The subspace

$$\mathfrak{S}_k = \operatorname{span}\left\{e^{i_1}, e^{i_2}, \ldots, e^{i_\tau}\right\}$$

is called as the $\tau$-steepest coordinates subspace, Then, the steepest descent direction in the subspace is sufficiently descent, namely

$$\min_{d \in \mathfrak{S}_k} \frac{d^\top g_k}{\|d\|_2 \|g_k\|_2} \leq -\frac{\tau}{n}.$$

Consequently, a sequential steepest coordinates search (SSCS) technique can be designed by augmenting the steepest coordinate directions into the subspace sequentially. For example, consider minimizing a convex quadratic function

$$Q(x) = g^\top x + \frac{1}{2} x^\top B x$$

# Electronic Structure Calculations

Therefore, the total energy minimization problem can be formulated as

$$\min_{X \in \mathbb{C}^{n\times p}} E(X), \quad \text{s.t.} \quad X^*X = I_p, \tag{16}$$

where $E(X)$ is $E_{ks}(X)$ in KSDFT and $E_{\mathrm{hf}}(X) := E_{\mathrm{ks}}(X) + E_{\mathrm{f}}(X)$ in HF.

$$E_{\mathrm{ks}}(X) := \frac{1}{4}\operatorname{tr}(X^*LX) + \frac{1}{2}\operatorname{tr}(X^*V_{\mathrm{ion}}X) + \frac{1}{2}\sum_l \sum_i \zeta_l \left|x_i^* w_l\right|^2 + \frac{1}{4}\rho^\top L^\dagger \rho + \frac{1}{2} e_n^\top \varepsilon_{\mathrm{xc}}(\rho)$$

$$E_{\mathrm{f}}(X) := \frac{1}{4}\langle V(XX^*)X, X\rangle = \frac{1}{4}\langle V(XX^*), XX^*\rangle.$$

Let $Z = V\left(X_k X_k^*\right)\Omega$ where $\Omega$ is an orthogonal basis of the subspace such as

$$\operatorname{span}\{X_k\}, \operatorname{span}\{X_{k-1}, X_k\} \text{ or } \operatorname{span}\{X_{k-1}, X_k, V(X_k X_k^*)X_k\}.$$

Then the low rank approximation $\hat{V}\left(X_k X_k^*\right) := Z(Z^*\Omega)^\dagger Z^*$ is able to reduce the computational cost significantly. New subproblem is formulated as

$$\min_{X \in \mathbb{C}^{n\times p}} E_{\mathrm{ks}}(X) + \frac{1}{4}\left\langle \hat{V}(X_k X_k^*)X, X\right\rangle \quad \text{s.t.} \quad X^*X = I_p. \tag{17}$$

The subproblem (17) can be solved by the SCF iteration, the Riemannian gradient method or the modified CG method.

# Subspace Relationship

$dim(\mathfrak{S}_k) = dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \approx \mathfrak{S}_{k+1}$

Fix-dimension Subspaces:
Direction-Gradient Subspaces
One-add-one-drop Subspaces

$dim(\mathfrak{S}_k) \leq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \subseteq \mathfrak{S}_{k+1}$

Nested Subspaces

$\sum_{k=1}^{n} dim(\mathfrak{S}_k) = p$: $\mathfrak{S}_1 + \cdots + \mathfrak{S}_n = \mathbb{R}^p$

Complement Subspaces

$dim(\mathfrak{S}_k) \geq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \supseteq \mathfrak{S}_{k+1}$

Active methods

$dim(\mathfrak{S}_k) = i_k$: $\mathfrak{S}_k = I_k$

Subsampling/Sketching
**Stochastic Optimization**

**Stochastic First-order Methods**
Stochastic gradient method selects a uniformly random sample $s_k$ from $\{1, \ldots, N\}$ and updates

$$x_{k+1} = x_k - \alpha_k \nabla f_{s_k}(x_k). \tag{18}$$

A common assumption for convergence is

$$\mathbb{E}\left[\nabla f_{s_k}(x_k) \mid x_k\right] = \nabla f(x_k),$$

$$x_{k+1} = x_k - \frac{\alpha_k}{|I_k|} \sum_{s_k \in I_k} \nabla f_{s_k}(x_k).$$

The momentum method:

$$\begin{aligned} v_{k+1} &= \mu_k v_k - \alpha_k \nabla f_{s_k}(x_k), \\ x_{k+1} &= x_k + v_{k+1}. \end{aligned}$$

This new update direction $v$ is a linear combination of the previous update direction $v_k$ and the gradient $\nabla f_{s_k}(x_k)$ to obtain a new $v_{k+1}$. When $\mu_k = 0$, the algorithm degenerates to SGD.

# Stochastic Methods

The adaptive subgradient method (AdaGrad) controls the step sizes of each component separately

$$G_k = \sum_{i=1}^{k} \nabla f_{s_i}(x_i) \odot \nabla f_{s_i}(x_i),$$

where $\odot$ is the Hadamard product between two vectors. The AdaGrad method is

$$x_{k+1} = x_k - \frac{\alpha_k}{\sqrt{G_k + \varepsilon e_n}} \odot \nabla f_{s_{k+1}}(x_{k+1}),$$
$$G_{k+1} = G_k + \nabla f_{s_{k+1}}(x_{k+1}) \odot \nabla f_{s_{k+1}}(x_{k+1}),$$

where the division in $\frac{\alpha_k}{\sqrt{G_k+\varepsilon e_n}}$ is also performed elementwisely.

**Stochastic Second-Order method**

$$\left[\frac{1}{|I_k^H|}\sum_{i\in I_k^H} \nabla^2 f_i(x)\right] d_k = -\frac{1}{|I_k|}\sum_{s_k\in I_k} \nabla f_{s_k}(x_k).$$

# Optimization problem and its subproblem

Optimization problem: Find $x^*$

$$\min_x f(x)$$
$$\text{s.t. } x \in \mathscr{X}$$

Subproblem: Find $x_{k+1} = x_k + d$

$$\min_d m_k(x_k + d)$$
$$\text{s.t. } d \in D$$

$dim(\mathfrak{S}_k) = dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \approx \mathfrak{S}_{k+1}$

$dim(\mathfrak{S}_k) \leq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \subseteq \mathfrak{S}_{k+1}$

$\sum_{k=1}^n dim(\mathfrak{S}_k) = p$: $\mathfrak{S}_1 + \cdots + \mathfrak{S}_n = \mathbb{R}^p$

$dim(\mathfrak{S}_k) \geq dim(\mathfrak{S}_{k+1})$: $\mathfrak{S}_k \supseteq \mathfrak{S}_{k+1}$

$dim(\mathfrak{S}_k) = i_k$: $\mathfrak{S}_k = I_k$

# Future work

- Relationship between subspace in the iteration
- Subspace Method in Manifold Optimization
- Subspace Method in Derivative Free Optimization
- Subspace Method in Functional Optimization
- Subspace Accelerate for given algorithms

# Future work: Relationship between subspace in the iteration

Conjugate direction Method

Conjugate Subspace Method

subspace is an evolution of the direction

## Definition

$p_0, p_1, \cdots, p_l$ is conjugate with respect to the symmetric positive definite matrix $A$ if

$$p_i^T A p_j = 0, \text{for all } i \neq j.$$

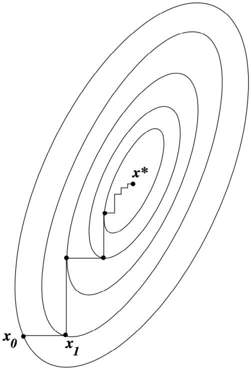


Figure 5: Coordinate search method can make slow progress.

### Theorem

*For any $x_0 \in \mathbb{R}^n$ the sequence $\{x_k\}$ generated by the conjugate direction algorithm (5.6), (5.7) converges to the solution $x^*$ of the linear system (5.1) in at most n steps.*

### Theorem (Expanding Subspace Minimization)

*Let $x_0 \in \mathbb{R}^n$ be any starting point and suppose that the sequence $\{x_k\}$ is generated by the conjugate direction algorithm* $(5.6), (5.7)$. *Then*

$$r_k^T p_i = 0, \quad \text{for } i = 0, 1, \ldots, k-1 \tag{19}$$

*and $x_k$ is the minimizer of $\phi(x) = \frac{1}{2} x^T A x - b^T x$ over the set*

$$\{x \mid x = x_0 + \operatorname{span}\{p_0, p_1, \ldots, p_{k-1}\}\} \tag{20}$$

# Future work: Subspace Method in Derivative Free Optimization

Main difference between Powell's Derivative Free Optimization and Optimization with derivative: How to get subproblem objective function $m_k(x)$.

$$\begin{cases} \alpha_0 + \alpha^\top y^1 + \frac{1}{2}\left(y^1\right)^\top H y^1 = F\left(y^1\right) \\ \alpha_0 + \alpha^\top y^2 + \frac{1}{2}\left(y^2\right)^\top H y^2 = F\left(y^2\right) \\ \qquad\qquad\qquad\cdots\cdots \\ \alpha_0 + \alpha^\top y^k + \frac{1}{2}\left(y^k\right)^\top H y^k = F\left(y^k\right) \end{cases}$$

$$\text{NEWUOA: } \min_{Q_k} \left\| \nabla^2 Q_k - \nabla^2 Q_{k-1} \right\|_F^2$$
$$\text{s.t. } Q_k(y) = F(y), y \in Y_k$$

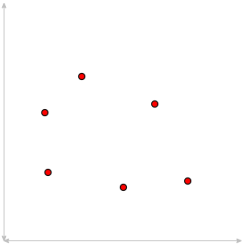

Figure 6: Model function by interpolation

# Future work: Subspace Method in Manifold Optimization

- Riemannian Steepest Descent Method[Udriste 1994]: -grad $f(x)$.

  Robust global convergence

  Slow local convergence: linear

- Riemannian Newton Method[Luenberger 1972; Gabay 1982] : -Hess $f(x)^{-1}$grad $f(x)$.

  Fast local convergence: quadratic or even cubic

  Requires additional work for global convergence

- Riemannian trust-region method[Absil, Baker, and Gallivan 2007]

  Find solution to $\eta = \underset{\eta \in T_x M, \|\eta\| \leq \Delta}{\operatorname{argmin}} \; m_x(\eta)$, $x_{\text{next}} = R_x(\eta)$,

$$\min_{\mathbf{X}} f(\mathbf{X}) := \frac{1}{2} \|\mathrm{P}_\Omega \mathbf{X} - \mathrm{P}_\Omega \mathbf{A}\|^2, \text{ s.t. } \mathbf{X} \in M_{\mathbf{r}} := \left\{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \mid \operatorname{rank}_{\mathrm{TT}}(\mathbf{X}) = \mathbf{r}\right\}$$

in Riemannian Optimization for high-dimensional tensor complement[Steinlechner 2016].

# Thank You

**Advises and guidance are needed**