Mathematical
Institute

Generalizing Quasi-Newton Updates to Higher-Order Derivatives

Karl Welzel and Raphael Hauser
Mathematical Institute
University of Oxford

SIAM Conference on Optimization, 3 June 2023

Oxford
Mathematics

Outline

Outline

Motivation: Why higher-order derivatives?

Nonlinear unconstrained optimization

Let $f : \mathbb{R}^n \to \mathbb{R}$. Find $\boldsymbol{x}_* = \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x})$

Number of iterations depending on derivative access

- Using first derivatives: gradient descent, many iterations
- Using second derivatives: Newton's method, few iteration
- Using third derivatives: tensor methods, even fewer iterations

Goal

Approximate third derivatives from second derivatives

Outline

Definition

A p-tensor is a multilinear map

$$\boldsymbol{T}\colon \underbrace{\mathbb{R}^n \times \cdots \times \mathbb{R}^n}_{p \text{ times}} \to \mathbb{R}$$

and the set of these tensors is denoted $\mathbb{R}^{\otimes^p n}$.

Tensor Notation

Tensor evaluation

$\boldsymbol{T}[\boldsymbol{s}_1, \ldots, \boldsymbol{s}_p], \boldsymbol{T}[\boldsymbol{s}]^p \in \mathbb{R}$ where $\boldsymbol{s}_i \in \mathbb{R}^n$ and $\boldsymbol{T}[\boldsymbol{s}_1, \ldots, \boldsymbol{s}_q] \in \mathbb{R}^{\otimes^{p-q} n}$

Tensor-matrix multiplication

$\boldsymbol{T}[\boldsymbol{W}_1, \ldots, \boldsymbol{W}_p], \boldsymbol{T}[\boldsymbol{W}]^p \in \mathbb{R}^{\otimes^p n}$ where $\boldsymbol{W}_i \in \mathbb{R}^{n \times n}$

Outer product

$$(\boldsymbol{T}_1 \otimes \boldsymbol{T}_2)[\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{p_1+p_2}] \coloneqq \boldsymbol{T}_1[\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{p_1}] \cdot \boldsymbol{T}_2[\boldsymbol{s}_{p_1+1}, \ldots, \boldsymbol{s}_{p_1+p_2}]$$

Centre for Intelligent Multidimensional Data Analysis Limited
智能多維數據分析研究中心有限公司
UNIVERSITY OF OXFORD
Mathematical Institute

Definition

For $\boldsymbol{T} \in \mathbb{R}^{\otimes^p n}$ and any permutation $\sigma \in S_p$, let

$$\sigma(\boldsymbol{T})[\boldsymbol{s}_1, \ldots, \boldsymbol{s}_p] \coloneqq \boldsymbol{T}[\boldsymbol{s}_{\sigma(1)}, \ldots, \boldsymbol{s}_{\sigma(p)}].$$

A tensor is *symmetric* if $\sigma(\boldsymbol{T}) = \boldsymbol{T}$ for all permutations $\sigma \in S_p$. The set of symmetric tensors is denoted $\mathbb{R}^{\otimes^p n}_{\text{sym}}$.

Symmetric projection

$$P_{\text{sym}}(\boldsymbol{T}) = \frac{1}{p!} \sum_{\sigma \in S_p} \sigma(\boldsymbol{T})$$

Tensor norms

p-dimensional arrays

$$t_{i_1,\ldots,i_p} = \boldsymbol{T}[\boldsymbol{e}_{i_1},\ldots,\boldsymbol{e}_{i_p}] \quad \text{and} \quad \boldsymbol{T} = \sum_{i_1,\ldots,i_p=1}^{n} t_{i_1,\ldots,i_p}\boldsymbol{e}_{i_1}\otimes\cdots\otimes\boldsymbol{e}_{i_p}$$

Frobenius inner product and norm

$$\langle \boldsymbol{T},\boldsymbol{U}\rangle_F := \sum t_{i_1,\ldots,i_p}u_{i_1,\ldots,i_p} \quad \text{and} \quad \|\boldsymbol{T}\|_F = \sqrt{\langle \boldsymbol{T},\boldsymbol{T}\rangle_F}$$

2-norm

$$\|\boldsymbol{T}\|_2 = \max_{\|\boldsymbol{s}_i\|_2=1} |\boldsymbol{T}[\boldsymbol{s}_1,\ldots,\boldsymbol{s}_p]|$$

Higher-order derivatives

pth derivative

$D^p f(\boldsymbol{x}) \in \mathbb{R}^{\otimes^p n}_{\text{sym}}$ is the pth derivative of f at $\boldsymbol{x}$

Taylor expansion

$$T_{f,p}(\boldsymbol{x}, \boldsymbol{s}) = \sum_{k=0}^{p} \frac{1}{k!} D^k f(\boldsymbol{x})[\boldsymbol{s}]^k \approx f(\boldsymbol{x} + \boldsymbol{s})$$

Outline

Centre for Intelligent Multidimensional Data Analysis Limited
智能多維數據分析研究中心有限公司
UNIVERSITY OF OXFORD
Mathematical Institute

Goal

Approximate third derivatives from second derivatives

Newton vs quasi-Newton

- Newton's method: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k (\nabla^2 f(\boldsymbol{x}_k))^{-1} \nabla f(\boldsymbol{x}_k)$
- Quasi-Newton: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \boldsymbol{B}_k^{-1} \nabla f(\boldsymbol{x}_k)$, $\boldsymbol{B}_k \approx \nabla^2 f(\boldsymbol{x}_k)$
  - $\boldsymbol{B}_k$ is obtained from gradients only (cheap)
  - Superlinear local convergence (fast)
  - Very successful in theory and practice

Centre for Intelligent Multidimensional Data Analysis Limited
智能多維數據分析研究中心有限公司
UNIVERSITY OF OXFORD
Mathematical Institute

Secant equation

$$\boldsymbol{B}_{k+1}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) = \nabla f(\boldsymbol{x}_{k+1}) - \nabla f(\boldsymbol{x}_k)$$

Secant equation

$$\boldsymbol{B}_{k+1}\boldsymbol{s}_k = \widetilde{\boldsymbol{B}}_k\boldsymbol{s}_k, \quad \boldsymbol{s}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$$

$$\widetilde{\boldsymbol{B}}_k = \int_0^1 \nabla^2 f(\boldsymbol{x}_k + t\boldsymbol{s}_k)\,\mathrm{d}t$$

Centre for Intelligent Multidimensional Data Analysis Limited
智能多維數據分析研究中心有限公司
UNIVERSITY OF OXFORD
Mathematical Institute

Secant equation

$$\boldsymbol{B}_{k+1}\boldsymbol{s}_k = \widetilde{\boldsymbol{B}}_k\boldsymbol{s}_k, \quad \boldsymbol{s}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$$

$$\widetilde{\boldsymbol{B}}_k = \int_0^1 \nabla^2 f(\boldsymbol{x}_k + t\boldsymbol{s}_k)\,\mathrm{d}t$$

Quasi-Newton updates

- PSB: $\min_{\boldsymbol{B}\in\mathbb{R}^{n\times n}_{\mathrm{sym}}} \|\boldsymbol{B} - \boldsymbol{B}_k\|_F$ s.t. $\boldsymbol{B}\boldsymbol{s}_k = \widetilde{\boldsymbol{B}}_k\boldsymbol{s}_k$
- DFP: $\min_{\boldsymbol{B}\in\mathbb{R}^{n\times n}_{\mathrm{sym}}} \|\boldsymbol{W}_k^T(\boldsymbol{B} - \boldsymbol{B}_k)\boldsymbol{W}_k\|_F$ s.t. $\boldsymbol{B}\boldsymbol{s}_k = \widetilde{\boldsymbol{B}}_k\boldsymbol{s}_k$
- BFGS: $\min_{\boldsymbol{B}\in\mathbb{R}^{n\times n}_{\mathrm{sym}}} \|\boldsymbol{W}_k^T(\boldsymbol{B}^{-1} - \boldsymbol{B}_k^{-1})\boldsymbol{W}_k\|_F$ s.t. $\boldsymbol{B}\boldsymbol{s}_k = \widetilde{\boldsymbol{B}}_k\boldsymbol{s}_k$

Generalized quasi-Newton (GQN) updates

Secant equation

$$\boldsymbol{B}_{k+1}\boldsymbol{s}_k = \widetilde{\boldsymbol{B}}_k\boldsymbol{s}_k = \nabla f(\boldsymbol{x}_{k+1}) - \nabla f(\boldsymbol{x}_k)$$

$$\widetilde{\boldsymbol{B}}_k = \int_0^1 \nabla^2 f(\boldsymbol{x}_k + t\boldsymbol{s}_k)\,\mathrm{d}t$$

Quasi-Newton updates

$$\boldsymbol{B}_{k+1} := \underset{\boldsymbol{B}\in\mathbb{R}^{n\times n}_{\mathrm{sym}}}{\arg\min} \|\boldsymbol{W}_k^T(\boldsymbol{B}-\boldsymbol{B}_k)\boldsymbol{W}_k\|_F \text{ s.t. } \boldsymbol{B}\boldsymbol{s}_k = \widetilde{\boldsymbol{B}}_k\boldsymbol{s}_k$$

Generalized quasi-Newton (GQN) updates

Centre for Intelligent Multidimensional Data Analysis Limited
智能多維數據分析研究中心有限公司

UNIVERSITY OF OXFORD
Mathematical Institute

Generalized secant equation

$$\boldsymbol{C}_{k+1}[\boldsymbol{s}_k] = \widetilde{\boldsymbol{C}}_k[\boldsymbol{s}_k] = D^{p-1} f(\boldsymbol{x}_{k+1}) - D^{p-1} f(\boldsymbol{x}_k)$$

$$\widetilde{\boldsymbol{C}}_k = \int_0^1 D^p f(\boldsymbol{x}_k + t\boldsymbol{s}_k)\,\mathrm{d}t$$

Generalized quasi-Newton updates

$$\boldsymbol{C}_{k+1} := \underset{\boldsymbol{C}\in\mathbb{R}^{\otimes^p n}_{\mathrm{sym}}}{\arg\min} \|(\boldsymbol{C} - \boldsymbol{C}_k)[\boldsymbol{W}_k]^p\|_F \text{ s.t. } \boldsymbol{C}[\boldsymbol{s}_k] = \widetilde{\boldsymbol{C}}_k[\boldsymbol{s}_k] \quad \text{(GQN)}$$

Characterization of GQN updates

Centre for Intelligent Multidimensional Data Analysis Limited
智能多維數據分析研究中心有限公司

UNIVERSITY OF OXFORD
Mathematical Institute

Theorem

Let $\boldsymbol{C}_\bullet \in \mathbb{R}^{\otimes^p n}_{\text{sym}}$, $\widetilde{\boldsymbol{C}} \in \mathbb{R}^{\otimes^p n}_{\text{sym}}$, *a nonsingular matrix* $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ *and a nonzero* $\boldsymbol{s} \in \mathbb{R}^n$ *be given. The following equations all have the same unique solution* $\boldsymbol{C}_+ \in \mathbb{R}^{\otimes^p n}_{\text{sym}}$ *using* $\boldsymbol{v} = \boldsymbol{W}^{-T}\boldsymbol{W}^{-1}\boldsymbol{s}$*:*

a. $\boldsymbol{C}_+ = \arg\min_{\boldsymbol{C}\in\mathbb{R}^{\otimes^p n}_{\text{sym}}} \|(\boldsymbol{C}-\boldsymbol{C}_\bullet)[\boldsymbol{W}]^p\|_F$ *s.t.* $\boldsymbol{C}[\boldsymbol{s}] = \widetilde{\boldsymbol{C}}[\boldsymbol{s}]$

b. $\boldsymbol{C}_+ = \boldsymbol{C}_\bullet + \sum_{j=1}^{p} \binom{p}{j} (-\boldsymbol{v}^T\boldsymbol{s})^{-j} P_{\text{sym}}\left((\otimes^j \boldsymbol{v}) \otimes \left(\boldsymbol{C}_\bullet - \widetilde{\boldsymbol{C}}\right)[\boldsymbol{s}]^j\right)$

c. $\boldsymbol{C}_+ = \boldsymbol{C}_\bullet + P_{\text{sym}}(\boldsymbol{A}\otimes\boldsymbol{v})$ *and* $\boldsymbol{A} \in \mathbb{R}^{\otimes^{p-1} n}_{\text{sym}}$ *is the unique* $(p-1)$*-tensor s.t.* $P_{\text{sym}}(\boldsymbol{A}\otimes\boldsymbol{v})[\boldsymbol{s}] = (\widetilde{\boldsymbol{C}}-\boldsymbol{C}_\bullet)[\boldsymbol{s}]$

d. $(\boldsymbol{C}_+ - \widetilde{\boldsymbol{C}})[\boldsymbol{W}]^p = (\boldsymbol{C}_\bullet - \widetilde{\boldsymbol{C}})[\boldsymbol{W}]^p \left[\boldsymbol{I} - \frac{\boldsymbol{W}^{-1}\boldsymbol{s}\boldsymbol{s}^T\boldsymbol{W}^{-T}}{\boldsymbol{s}^T\boldsymbol{W}^{-T}\boldsymbol{W}^{-1}\boldsymbol{s}}\right]^p$

Outline

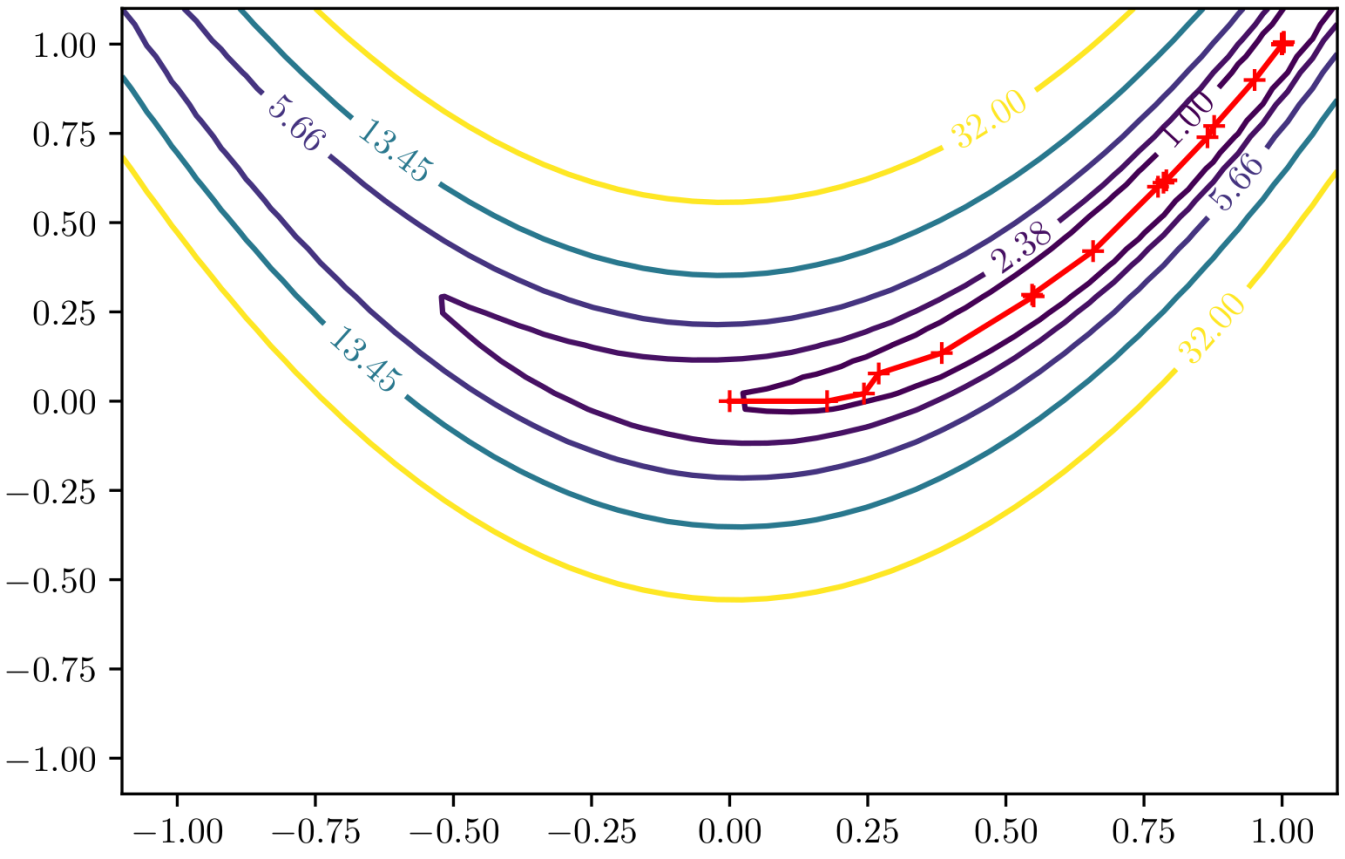
Convergence results

Theorem

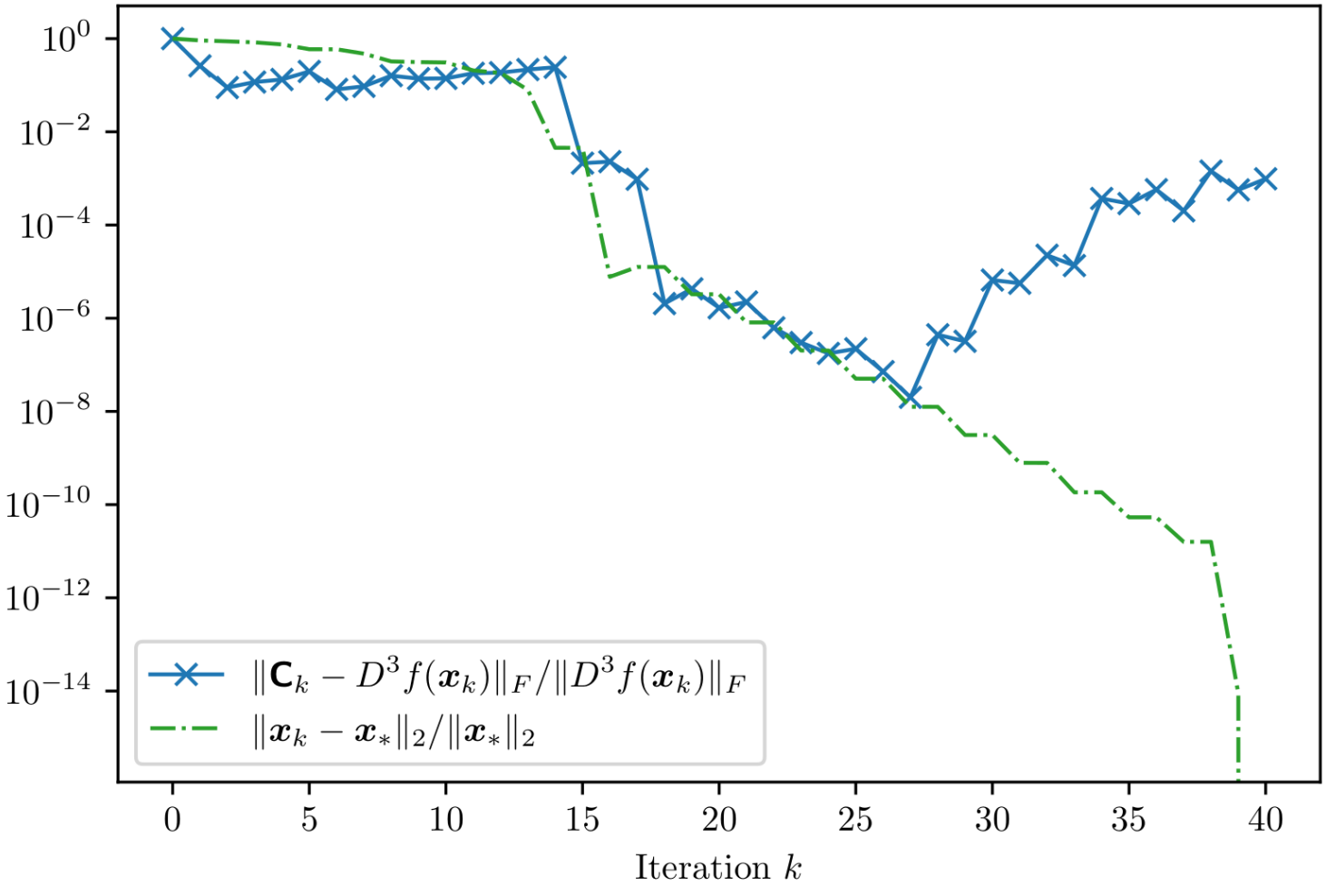
Let $\boldsymbol{C}_0 \in \mathbb{R}^{\otimes^p n}_{\mathrm{sym}}$ be given and update the approximations $\boldsymbol{C}_k$ according to (GQN). Assume $\boldsymbol{x}_k$ converge to $\boldsymbol{x}_* \in \mathbb{R}^n$, $\boldsymbol{W}_k$ converge to some nonsingular matrix $\boldsymbol{W}_* \in \mathbb{R}^{n \times n}$ and the steps are uniformly linearly independent. Then $\boldsymbol{C}_k$ converges to $\boldsymbol{C}_* := D^p f(\boldsymbol{x}_*)$.

Outline

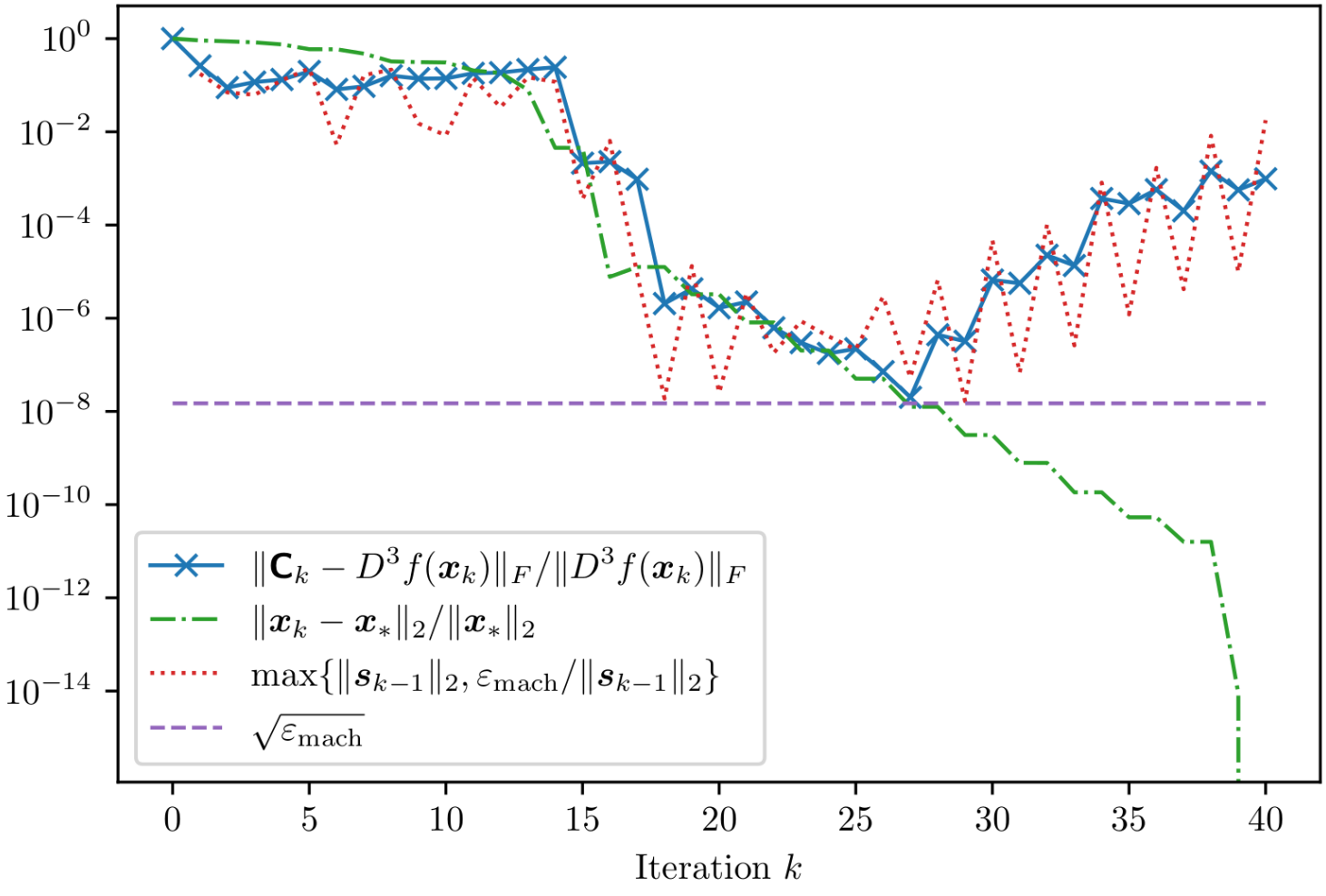
Experiments: Setup



Experiment: First-order method



Experiment: First-order method



Outline

Conclusion

Centre for Intelligent Multidimensional Data Analysis Limited
智能多維數據分析研究中心有限公司

UNIVERSITY OF OXFORD
Mathematical Institute

Results

- Explicit characterization of GQN updates
- Theoretical convergence results that mirror what is known for quasi-Newton methods
- Encouraging numerical experiments on a simple function (up to numerical limitations)

Open Questions

- How do we incorporate third order information?
- Are the generated approximations good enough?

Dennis–Moré condition

$$\lim_{k\to\infty} \frac{\|(\boldsymbol{B}_k - \nabla^2 f(\boldsymbol{x}_*))\boldsymbol{s}_k\|}{\|\boldsymbol{s}_k\|} = 0$$

Convergence results

Generalized Dennis–Moré condition

$$\lim_{k\to\infty} \frac{\|(\boldsymbol{C}_k - D^p f(\boldsymbol{x}_*))[\boldsymbol{s}_k]\|}{\|\boldsymbol{s}_k\|} = 0$$

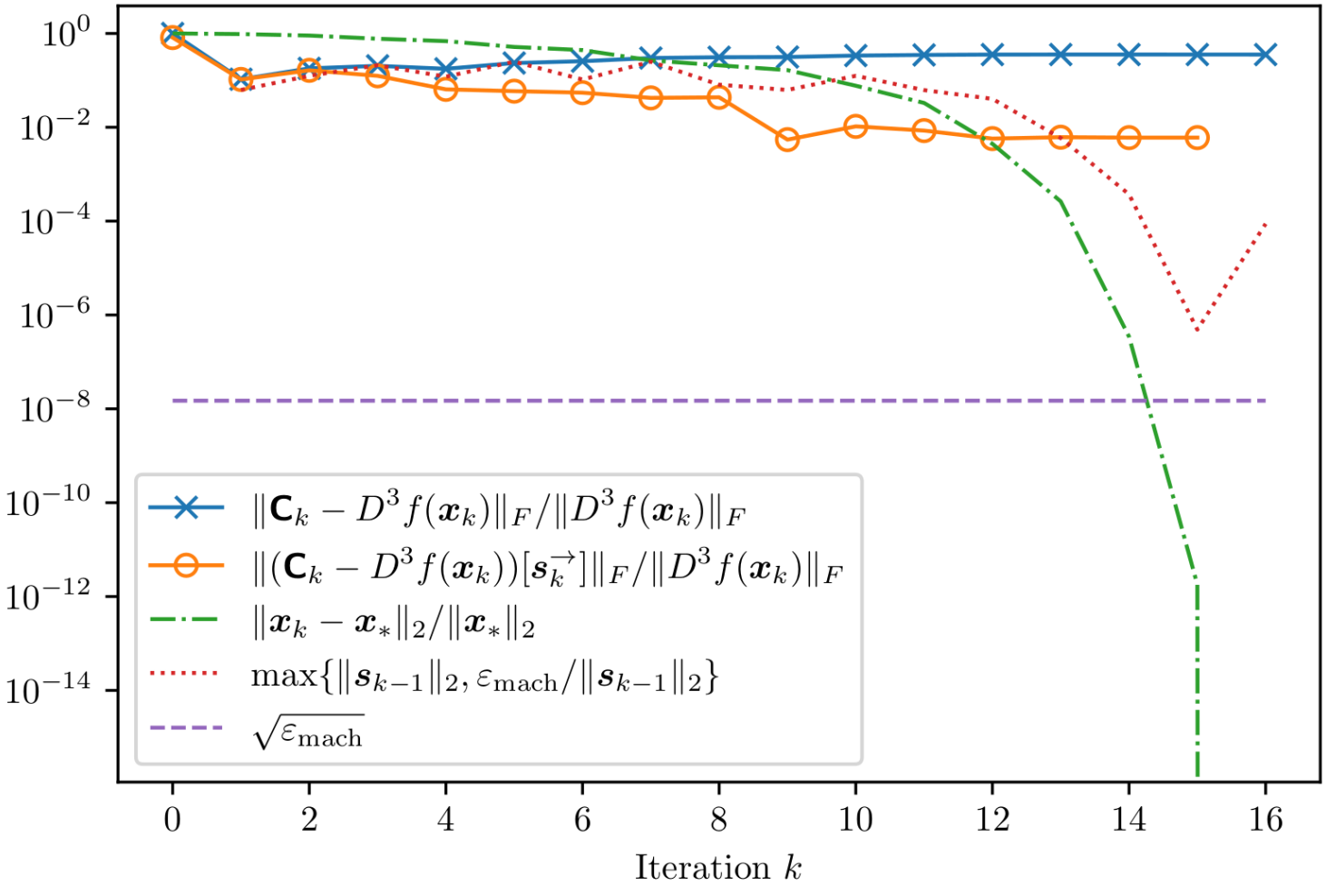
Theorem

Let $\boldsymbol{C}_0 \in \mathbb{R}^{\otimes^p n}_{\mathrm{sym}}$ be given and update the approximations $\boldsymbol{C}_k$ according to (GQN). Let $D^p f$ be Lipschitz continuous. Assume $\boldsymbol{x}_k$ converge to $\boldsymbol{x}_* \in \mathbb{R}^n$ and $\boldsymbol{W}_k$ to $\boldsymbol{W}_* \in \mathbb{R}^{n\times n}$ fast enough s.t.

$$\sum_{k\geq 0} \|\boldsymbol{x}_k - \boldsymbol{x}_*\|_2 < \infty \quad \text{and} \quad \sum_{k\geq 0} \|\boldsymbol{W}_k - \boldsymbol{W}_*\|_2 < \infty.$$

Then the generalized Dennis–Moré condition holds.

Experiment 2: Exact trust-region (subspace)



Experiment 2: Exact trust-region (subspace)