Approximate Derivatives for Tensor Methods

Karl Welzel and Raphael A. Hauser

Mathematical Institute

University of Oxford

International Symposium on Mathematical Programming 2024

Outline

1 Motivation
- Why tensor methods?

2 Higher-order secant updates
- Derivation
- Characterization
- Convergence

3 Experiments
- Convergence of quasi-tensor methods

Outline

1 Motivation
- Why tensor methods?

2 Higher-order secant updates
- Derivation
- Characterization
- Convergence

3 Experiments
- Convergence of quasi-tensor methods

Why tensor methods?

- Unconstrained nonconvex optimization
  - Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be sufficiently smooth. Find $\boldsymbol{x}_* = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$.
- ARp: Iteratively minimize $T_p(\boldsymbol{x}_k, \boldsymbol{s}) + \frac{\sigma_k}{p+1}\|\boldsymbol{s}\|^{p+1}$
- More derivatives = faster convergence

	$p=1$	$p=2$	$p=3$	$\ldots$
Global complexity[1]	$O(\varepsilon^{-2})$	$O(\varepsilon^{-3/2})$	$O(\varepsilon^{-4/3})$	$O(\varepsilon^{-(p+1)/p})$
Local convergence[2]	linear	quadratic	cubic	pth-order

[1] For adaptive regularization methods (ARp)

[2] Under the right assumptions

Outline

1 Motivation
- Why tensor methods?

2 Higher-order secant updates
- Derivation
- Characterization
- Convergence

3 Experiments
- Convergence of quasi-tensor methods

Quasi-Newton 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

Secant equation

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

Quasi-Newton updates

Secant equation

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

Quasi-Newton updates

Secant equation

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

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

Higher-order secant updates (HOSU)

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

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

Quasi-Newton updates

$$\boldsymbol{B}_{k+1} := \operatorname*{arg\,min}_{\boldsymbol{B}\in\mathbb{R}^{n\times n}_{\mathrm{sym}}} \|\boldsymbol{B} - \boldsymbol{B}_k\|_F \text{ s.t. } \boldsymbol{B}\boldsymbol{s}_k = \widetilde{\boldsymbol{B}}_k\boldsymbol{s}_k$$

Higher-order secant updates (HOSU)

Higher-order secant equation

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

Higher-order secant updates

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

Characterization of HOSU 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 $\boldsymbol{C}_\bullet \in \mathbb{R}^{\otimes^p n}_{\text{sym}}$, $\widetilde{\boldsymbol{C}} \in \mathbb{R}^{\otimes^p n}_{\text{sym}}$ *and a nonzero* $\boldsymbol{s} \in \mathbb{R}^n$ *be given. The following equations all have the same unique solution* $\boldsymbol{C}_+ \in \mathbb{R}^{\otimes^p n}_{\text{sym}}$*:*

- a $\boldsymbol{C}_+ = \arg\min_{\boldsymbol{C} \in \mathbb{R}^{\otimes^p n}_{\text{sym}}} \|\boldsymbol{C} - \boldsymbol{C}_\bullet\|_F$ *s.t.* $\boldsymbol{C}[\boldsymbol{s}] = \widetilde{\boldsymbol{C}}[\boldsymbol{s}]$
- b $\boldsymbol{C}_+ = \boldsymbol{C}_\bullet + \sum_{j=1}^{p} (-1)^j \binom{p}{j} \|\boldsymbol{s}\|^{-2j} P_{\text{sym}}\Big((\otimes^j \boldsymbol{s}) \otimes \Big(\boldsymbol{C}_\bullet - \widetilde{\boldsymbol{C}}\Big)[\boldsymbol{s}]^j \Big)$
- 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 $\boldsymbol{C}_+ - \widetilde{\boldsymbol{C}} = (\boldsymbol{C}_\bullet - \widetilde{\boldsymbol{C}}) \left[\boldsymbol{I} - \frac{\boldsymbol{s}\boldsymbol{s}^T}{\boldsymbol{s}^T\boldsymbol{s}} \right]^p$

Characterization of HOSU 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 $\boldsymbol{C}_\bullet \in \mathbb{R}^{\otimes^p n}_{\mathrm{sym}}$, $\widetilde{\boldsymbol{C}} \in \mathbb{R}^{\otimes^p n}_{\mathrm{sym}}$ *and a nonzero* $\boldsymbol{s} \in \mathbb{R}^n$ *be given. The following equations all have the same unique solution* $\boldsymbol{C}_+ \in \mathbb{R}^{\otimes^p n}_{\mathrm{sym}}$*:*

- a $\boldsymbol{C}_+ = \arg\min_{\boldsymbol{C} \in \mathbb{R}^{\otimes^p n}_{\mathrm{sym}}} \|\boldsymbol{C} - \boldsymbol{C}_\bullet\|_F$ *s.t.* $\boldsymbol{C}[\boldsymbol{s}] = \widetilde{\boldsymbol{C}}[\boldsymbol{s}]$
- b *Explicit formula to compute update*
- c *Update has a certain low-rank structure*
- d *Recursive formula for the approximation error*

Convergence of HOSU updates

Theorem (convergence to the true derivative)

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

Remark

In practice, only convergence up to $\sqrt{\varepsilon_{\mathrm{mach}}}$ because of cancellation errors.

Outline

1 Motivation
- Why tensor methods?

2 Higher-order secant updates
- Derivation
- Characterization
- Convergence

3 Experiments
- Convergence of quasi-tensor methods

Convergence of quasi-tensor methods

- AR3 method + HOSU = quasi-tensor method
- Custom AR3 implementation (joint work with C. Cartis, R. A. Hauser, Y. Liu, W. Zhu)
- Test problems: 34 MGH problems $(2 \le n \le 40)$ and 100-dim. Rosenbrock
- Problem solved when $\frac{f(\boldsymbol{x}_k)-f^*}{\max(1,|f^*|)} < 10^{-8}$

	$f(\boldsymbol{x}_k)$	$\nabla f(\boldsymbol{x}_k)$	$\nabla^2 f(\boldsymbol{x}_k)$	$\nabla^3 f(\boldsymbol{x}_k)$
Successful iterations	0	1	0	0
Evaluation cost	1	n	$\frac{n(n+1)}{2}$	$\frac{n(n+1)(n+2)}{6}$

Convergence of quasi-tensor methods

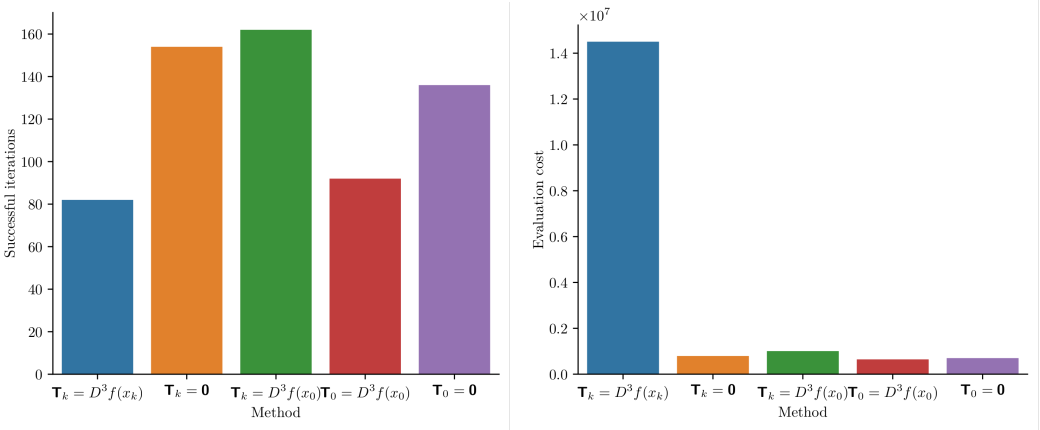

Figure: Performance on 100-dim. Rosenbrock function

Convergence of quasi-tensor methods

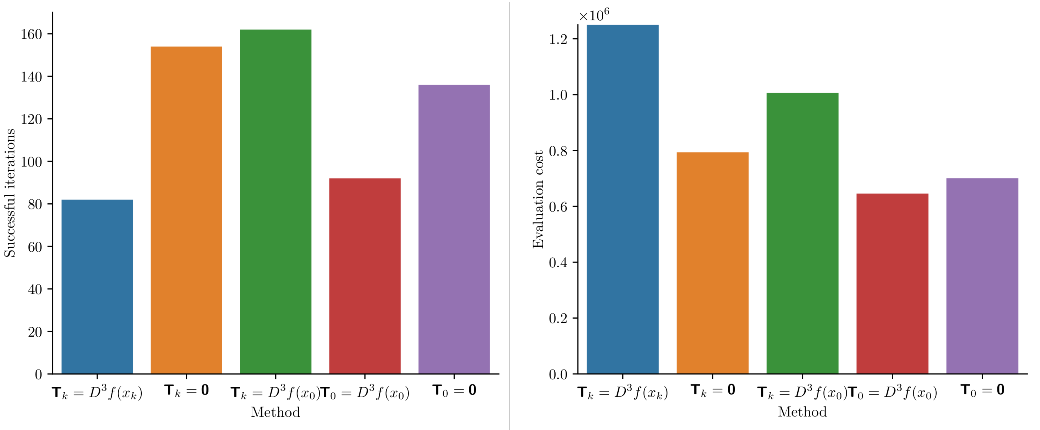

Figure: Performance on 100-dim. Rosenbrock function

Convergence of quasi-tensor methods

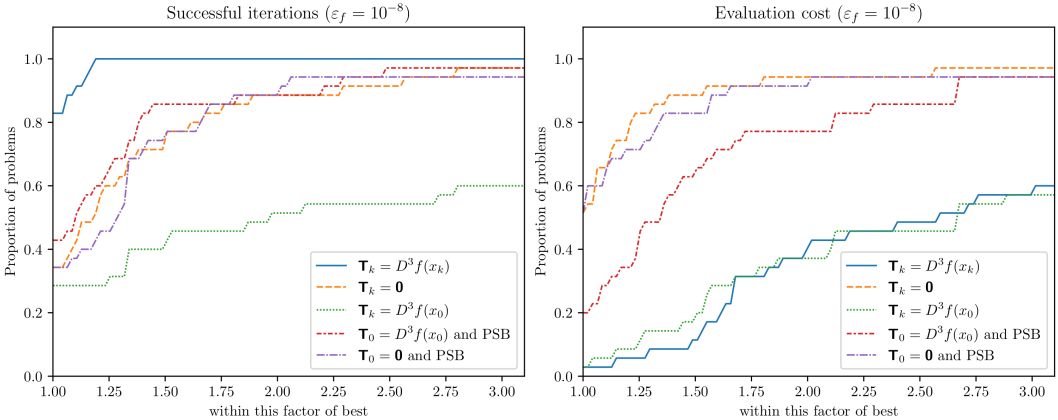

Figure: Performance profiles for the complete test set

Summary

- Quasi-Newton updates can be generalized to higher order
- HOSU provide cheap approximations of third derivatives
- Quasi-tensor methods can outperform second-order methods on certain problems

- Reference for HOSU
  - Welzel, K., & Hauser, R. A. (2024). Approximating Higher-Order Derivative Tensors Using Secant Updates. SIAM Journal on Optimization, 34(1), 893–917. https://doi.org/10.1137/23M1549687