A Proximal Quasi-Newton Trust-Region Method for Nonsmooth Regularized Optimization

Robert Baraldi
with: Sasha Aravkin, Dominique Orban

SIOPT 2021

July 21^{st}, 2021

Overview I

Inverse Problem Cost Functions and Regularizers I

- Regularized cost functions:
  - Sum of two functions with exploitable characteristics; (non)smoothness, (non)convexity

$$\underset{x}{\text{minimize}} \;\; f(x) + h(x) \tag{1}$$

Inverse Problem Cost Functions and Regularizers II

- Smooth term f- contains derivative information
  - Usually data misfit
  - Nonconvex in nonlinear functions - PDE/ODE inverse problems, ML, etc

Inverse Problem Cost Functions and Regularizers III

- Nonsmooth term regularizers *h* promote sparsity in ill-conditioned problems
  - Large datasets encourage overfitting
  - Sparsity-inducing functions temper model-complexity - **lack derivatives**
  - Examples: sparse regression, matrix completion (rank), phase retrieval, TV regularization
  - In literature: usually convex approximations of nonconvex functions

Technical Focus: Quasi-Newton PG + TR Method

- Problem Statement: $\min_x \; f(x) + h(x)$
  - $f \in \mathcal{C}^1$, h proper, lsc.
- Confounding Issue: Nonsmooth TR theory can be niche in scope, often difficult to implement
- Approach: TR method where **steps** are computed by minimizing simpler nonsmooth models based on PG.
- Results:
  - Global convergence to critical points
  - $O(1/\epsilon^2)$ worst-case complexity - equivalent to smooth cases
  - Comparisons between PG and QR method
- Next/Tools: TR, Proximal Gradient (PG)

Trust-Region Sub-problem: Smooth Case

- TR methods: numerically efficient approximations of nonlinear functions.
- At x_k, TR methods compute step s_k - approximate solution of subproblem

$$\underset{s}{\text{minimize}} \;\; \varphi(s; x_k) \quad \text{subject to} \;\; \|s\| \le \Delta_k,$$

  - $\varphi(\cdot; x_k)$ is a quadratic model of f about x_k
  - $\|\cdot\|$ is a norm
  - $\Delta_k > 0$ is the trust-region radius.
  - Compare $\varphi(0; x_k) - \varphi(s_k; x_k)$ to $f(x_k) - f(x_k + s_k)$: decide if s_k is accepted
- Solved exactly Moré and Sorensen (1983) or approximately Steihaug (1983) (truncated conjugate gradient method).

Other nonsmooth results in the literature

1. Convex composite objectives Yuan (1985): $h(x) = g(c(x))$, $c \in \mathcal{C}^2$, g convex.
   - Applications to penalty functions - special case of (1).
2. $f = 0$, h Lipschitz cont. Dennis, Li, and Tapia (1995) - no approach for nonsmooth subproblem.
3. $f \in \mathcal{C}^2$ with h convex and globally L-cont. Cartis, Gould, and Toint (2011) - not generally nonsmooth, no subproblem approach.
4. Various PG accelerations & modifications Stella, Themelis, Sopasakis, and Patrinos (2017); Themelis, Stella, and Patrinos (2018); Bolte, Sabach, and Teboulle (2014b)

Nonsmooth Analysis: Proximal Operator

Definition 1

Proper lsc function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a parameter $\nu > 0$, the *Moreau envelope* $e_{\nu h}$ and the *proximal mapping* $\mathrm{prox}_{\nu h}$ are defined by

$$e_{\nu h}(x) := \inf_{w} \tfrac{1}{2\nu}\|w - x\|^2 + h(w), \tag{2a}$$

$$\operatorname*{prox}_{\nu h}(x) := \arg\min_{w} \tfrac{1}{2\nu}\|w - x\|^2 + h(w). \tag{2b}$$

- Interpretation: extension of cost function to minimizing h and near x

Proximal Gradient (PG) Descent

- Solve

$$\min_s \varphi(s) + \psi(s)$$

Choose s_0, repeat:

$$\begin{aligned} s_{j+1} &\leftarrow \operatorname*{prox}_{\nu\psi}(s_j - \nu\nabla\varphi(s_j)) \\ &= \arg\min_s \tfrac{1}{2}\nu^{-1}\|s - (s_j - \nu\nabla\varphi(s_j))\|^2 + \psi(s). \end{aligned}$$

Optimality in Nonsmooth Regimes

Proposition 2 (Rockafellar and Wets, 1998, Theorem 10.1 - Optimality)

- *If $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper and has a local minimum at $\bar{x}$, then $0 \in \partial\phi(\bar{x})$.*
- *If ϕ is convex, $\bar{x}$ is a global minimum.*
- *If $\phi = f + h$ where f is differentiable on a neighborhood of $\bar{x}$ and h is finite at $\bar{x}$, then $\partial\phi(\bar{x}) = \nabla f(\bar{x}) + \partial h(\bar{x})$.*

- Note: $\partial\phi(x)$ is the *limiting subdifferential* of ϕ at $\bar{x}$

Theoretical Contribution Approach

1. Assume that we generated s_k that optimizes $m_k(s; x_k)$ via PG. How do we extend TR theory to nsmth ncvx case?
   - What model/step assumptions? What conclusions?
2. How do we generate s_k via PG?
   - Guaranted a step? Convergence results?

Tricky parts: we have an outer/overall TR problem and an inner s_k problem!

Algorithm 1 Nonsmooth Regularized Trust-Region Algorithm.

Choose constants $0 < \eta_1 \le \eta_2 < 1, 0 < \gamma_1 \le \gamma_2 < 1 < \gamma_3 \le \gamma_4$ and $\alpha > 0,\ \beta \ge 1$.

Choose $x_0 \in \mathbb{R}^n, \Delta_0 > 0$, compute $f(x_0) + h(x_0)$.

for $k = 0, 1, \dots$ **do**

Choose $0 < \nu_k \le 1/(L(x_k) + \alpha^{-1}\Delta_k^{-1})$.

Define $m_k(s) := \varphi(s) + \psi(s)$, $m_k^\nu(s) := \varphi^\nu(s) + \psi(s)$

Compute $s_{k,1} = \arg\min_s m_k^\nu(s)$.

Compute $s_k = \arg\min_s m_k(s)$ with $\|s_k\| \le \min(\Delta_k,\ \beta\|s_{k,1}\|)$.

Calculate $\rho_k := \frac{f(x_k)+h(x_k)-(f(x_k+s_k)+h(x_k+s_k))}{m_k(0)-m_k(s_k)}$.

If $\rho_k \ge \eta_1 \quad \Rightarrow \ x_{k+1} = x_k + s_k$. Else $x_{k+1} = x_k$.

Update TR radius

$$\Delta_{k+1} \in \begin{cases} [\gamma_3\Delta_k,\ \gamma_4\Delta_k] & \text{if } \rho_k \ge \eta_2, & \text{(VSI)} \\ [\gamma_2\Delta_k,\ \Delta_k] & \text{if } \eta_1 \le \rho_k < \eta_2, & \text{(SI)} \\ [\gamma_1\Delta_k,\ \gamma_2\Delta_k] & \text{if } \rho_k < \eta_1 & \text{(UI).} \end{cases}$$

end for

Technical Setup for TR Method

- Fixed $x \in \mathbb{R}^n$; parametric problem and optimal set

$$p(\Delta; x, \nu) := \min_s \; \varphi(s; x) + \psi(s; x) + \chi(s; \Delta), \quad (3a)$$

$$P(\Delta; x, \nu) := \arg\min_s \; \varphi(s; x) + \psi(s; x) + \chi(s; \Delta), \quad (3b)$$

- Goal: WTS eventually, yield s that decreases $f(x) + h(x)$.
- Next: Model Assumptions & Properties

Assumptions on φ and ψ

- φ has same assumptions as f, ψ yields a proximal operator
  - (Baraldi, Aravkin, and Orban, 2020, Model Assumption 3.1)
  - Note: h just proper, lsc!
- Model info matches (1) at TR center
  - (Baraldi et al., 2020, Model Assumption 3.2)
  - $\varphi(0; x) = f(x)$, and $\nabla_s \varphi(0; x) = \nabla f(x)$.
  - $\nabla_s \varphi(\cdot; x)$ is L-cont with $0 \leq L(x) \leq L \ \forall \ x$.
  - $\psi(\cdot; x)$ is proper, lsc, $\psi(0; x) = h(x)$, and $\partial \psi(0; x) = \partial h(x)$.

Stationarity of $\varphi + \psi$ implies stationarity of $f + h$

- If $0 \in P(\Delta; x, \nu)$: $s = 0$ is F.O. stationary for (3a)/$p(\Delta; x, \nu)$ iff x is F.O. stationary for (1).
  - (Baraldi et al., 2020, Proposition 3.2)
- $\xi(\Delta; x, \nu) = f(x) + h(x) - \varphi(s; x) - \psi(s; x) = 0 \iff 0 \in P(\Delta; x, \nu) \implies x$ is first-order stationary for (1).
  - (Baraldi et al., 2020, Proposition 3.3)

Model Decrease Condition

Step Assumption 4.1 (Baraldi et al., 2020, Model Assumption 3.1)

There exists $\kappa_{\mathrm{m}} > 0$ *and* $\kappa_{\mathrm{mdc}} \in (0, 1)$ *such that for all* k, $\|s_k\| \le \Delta_k$ *and*

$$|f(x_k + s_k) + h(x_k + s_k) - m_k(s_k; x_k)| \le \kappa_{\mathrm{m}} \|s_k\|^2, \tag{4a}$$

$$m_k(0; x_k) - m_k(s_k; x_k) \ge \kappa_{\mathrm{mdc}} \xi(\Delta_k; x_k, \nu_k). \tag{4b}$$

- Step-size bounds model performance

Theoretical Results

Recover analogous results for classic theory Conn, Gould, and Toint (2000) in nsmth, ncvx case

- Successful step is guaranteed if radius is small enough, with

$$\Delta_{\text{succ}} := \frac{\kappa_{\text{mdc}}(1-\eta_2)}{2\kappa_{\text{m}}\alpha\beta^2} > 0. \tag{5}$$

  - Baraldi et al. (2020, Theorem 3.4)
- Algorithm 1 identifies F.O. critical point
  - Baraldi et al. (2020, Theorem 3.5)

Theoretical Results - Criticality Metrics

- Can write down $\Delta_k \geq \Delta_{\min} \, \forall \, k \in \mathbb{N}$, with

$$\Delta_{\min} := \min(\Delta_0, \gamma_1 \Delta_{\text{succ}}) > 0. \tag{6}$$

and F.O. measure prop yields $\nu_k^{-1} \xi(\Delta_{\min}; x_k, \nu_k)^{\frac{1}{2}}$.

- Smallest iteration number $k(\epsilon)$ satisfying F.O. optimality condition

$$\nu_k^{-1} \xi(\Delta_{\min}; x_k, \nu_k)^{\frac{1}{2}} \leq \epsilon \quad (0 < \epsilon < 1). \tag{7}$$

with

$$\mathcal{S} := \{k \in \mathbb{N} \mid \rho_k \geq \eta_1\}, \tag{8a}$$
$$\mathcal{S}(\epsilon) := \{k \in \mathcal{S} \mid k < k(\epsilon)\}, \tag{8b}$$
$$\mathcal{U}(\epsilon) := \{k \in \mathbb{N} \mid k \notin \mathcal{S} \text{ and } k < k(\epsilon)\}, \tag{8c}$$

Theoretical Results - Iteration Complexity

- Set of successful iterations for which (7) is attained is $O(\epsilon^{-2})$.

$$|\mathcal{S}(\epsilon)| \leq \frac{(f+h)(x_0) - (f+h)_{\text{low}}}{\eta_1 \kappa_{\text{mdc}} \nu_{\min}^2 \epsilon^2} = O(\epsilon^{-2}). \tag{9}$$

- # of unsuccessful iterations is similarly bounded

$$|\mathcal{U}(\epsilon)| \leq \log_{\gamma_2}(\Delta_{\min}/\Delta_0) + |\mathcal{S}(\epsilon)| |\log_{\gamma_2}(\gamma_4)| = O(\epsilon^{-2}). \tag{10}$$

- Baraldi et al. (2020, Lemmas 3.6 & 3.7)

Theorem 3 (Baraldi et al., 2020, Theorem 3.8)

$$|\mathcal{S}(\epsilon)| + |\mathcal{U}(\epsilon)| = O(\epsilon^{-2}). \tag{11}$$

Theoretical Results - Criticality

Theorem 4 (Baraldi et al., 2020, Theorem 3.11)

Let Step Assumption 4.1 be satisfied. If there are infinitely many successful iterations,

$$\lim_{k\to\infty} f(x_k) + h(x_k) \to -\infty \textit{ or } \lim_{k\to\infty} \nu_k^{-1}\xi(\Delta_{\min}; x_k, \nu_k)^{\frac{1}{2}} = 0.$$

- Without extra assumptions, every limit point of $\{x_k, \nu_k\}$ is an F.O. critical point.

PG for descent steps

- How do we find subproblem solutions?
- To produce an s, we need to solve

$$\underset{s}{\text{minimize}} \;\; \varphi(s) + \psi(s) + \chi(s), \tag{12}$$

- <u>Tool:</u> Proximal gradient updates

Every PG-step Decreases Surrogate Models I

- Generic descent for every inner step Bolte, Sabach, and Teboulle (2014a)
- Special case since φ is quadratic:

$$(\varphi + \psi)(s_{j+1}) \leq (\varphi + \psi)(s_j) - \frac{\theta}{2\nu_k}\|s_{j+1} - s_j\|^2, \quad j \geq 0.$$

for $0 < \nu_k \leq (1-\theta)\|B_k\|^{-1}$ and $\theta \in (0, 1)$.

Every PG-step Decreases Surrogate Models II

- $v_k = (B_k - \nu_k^{-1} I)(s_{j+1} - s_j) \in \nabla\varphi(s_{j+1}) + \partial\psi(s_{j+1})$
- Complexity for $v_k \in \partial(\varphi + \psi)(s)$

$$\min_{j=0,\dots,N-1} \|v_{j+1}\| \leq \sqrt{\frac{2}{N\theta\nu_k}((\varphi+\psi)(s_0) - (\varphi+\psi)^*)}$$

- Results: We eventually arrive at $s_k \in P(\Delta; x)$ - i.e. stationary point of the surrogate model at sublinear rate

Proximal Operators

Need prox-operator for $\psi + \chi$.

1. ψ cvx & separable, $\chi = \ell_\infty$, $q = s_j - \nu \nabla \varphi(s_j)$
2. $\psi = \ell_1,\ \chi = \ell_2$
3. When h is nonconvex, greater variety of cases:
   - $h(x) = \lambda \|x\|_0$, a global solution is one of the bounds, or 1 of 2 local mins. Course: evaluate the objective at four points and choose lowest.

Extension to Quadratic Regularization Schemes

- Can use this methodology to provide complexity results to F.O. method with adaptive stepsize
  - Related to PG
- Benefits: nonsmooth, nonconvex models, without Lipschitz estimation

Theoretical Conclusions; Numerical Comparisons

- Theory:
  - Outer/TR Method: s_k created by nonsmooth means (PG) still converges to critical point of $f + h$
  - Inner/PG Method: PG will create an s_k, eventually reaches critical point of $\varphi + \psi + \chi$
  - QR: Complexity result for PG-type method, can also use as inner solver
- Next:
  - Test case on BPDN example
  - Perform model reduction on nonlinear inverse problem
  - Compare against two similar methods: PANOC and ZeroFPR

BPDN

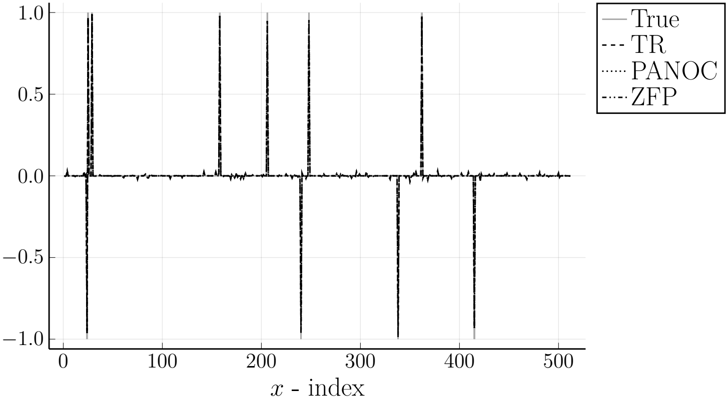
- $b = Ax_0 + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, 0.1)$.
- For $p = 0, 1$

$$\underset{x}{\text{minimize}} \, \|Ax - b\|_2^2 + \lambda \|x\|_p. \tag{13}$$

- Compare to ZeroFPR, PANOC

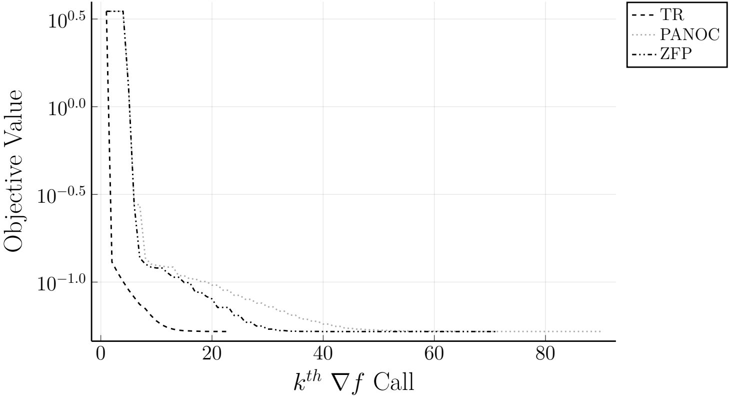
Results I

Figure 1: (13) for $p = 0, 1$ with full Hessian, with final objective values and objective function decrease for ℓ_∞-norm TR.



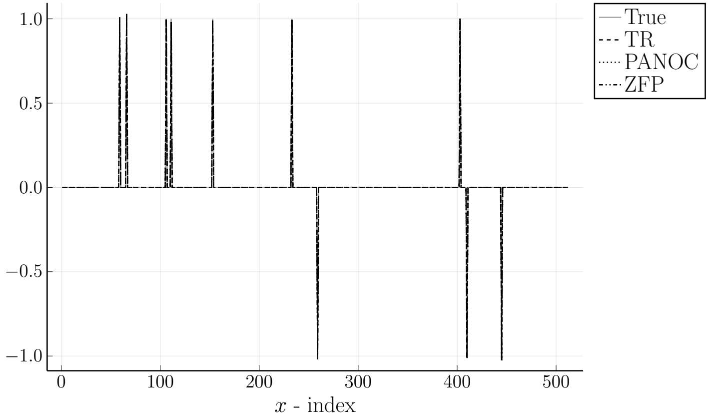
a ℓ_1 - Basis

Results II



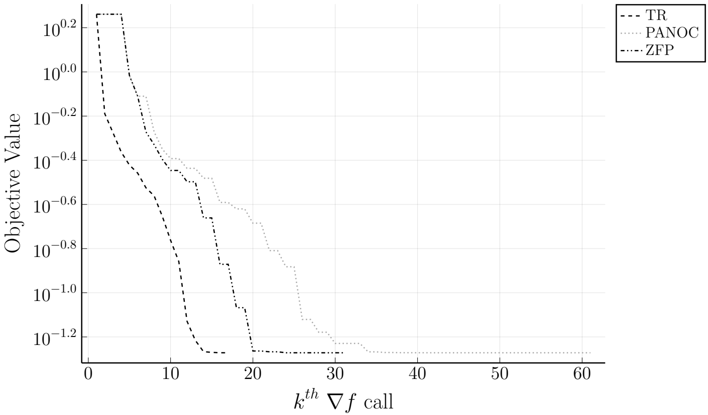
b ℓ_1 - Objective history

Results III



c ℓ_0 - Basis

Results IV



d ℓ_0 - Objective history

Classical ODE Inverse Problem

We would like to solve

$$\min_{x} \|F(x) - b\|_2^2 + h(x). \tag{14}$$

where nonlinear $F(x)$ is the solution of a system of ODEs.

Fitzhugh-Nagumo Model

The Fitzhugh-Nagumo model for neuron activation is given by

$$\frac{dV}{dt} = (V - V^3/3 - W + x_1)x_2^{-1} \tag{15a}$$

$$\frac{dW}{dt} = x_2(x_3 V - x_4 W + x_5). \tag{15b}$$

For $x_1 = x_4 = x_5 = 0$, it becomes the Van-der-Pol oscillator

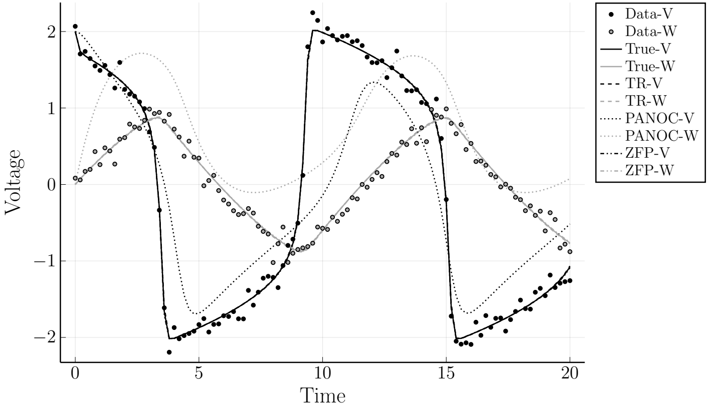
$$\frac{dV}{dt} = (V - V^3/3 - W)x_2^{-1} \tag{16a}$$

$$\frac{dW}{dt} = x_2(x_3 V). \tag{16b}$$

- Highly nonlinear and ill-conditioned
- LBFGS for f, $h(x) = \lambda\|x\|_0$, and an ℓ_∞-norm TR ball
- Goal: Fit data, exactly enforce $x_1 = x_4 = x_5 = 0$

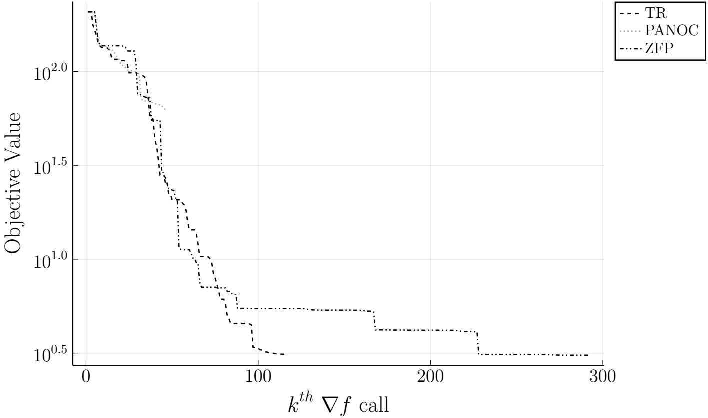
TR Results I

Figure 2: Fitzhugh-Nagumo solution ((15a), (15b)) for $h(x) = \lambda\|x\|_0$ in (14) with ℓ_∞-norm TR and LBFGS approximation.



a Solution Comparisons

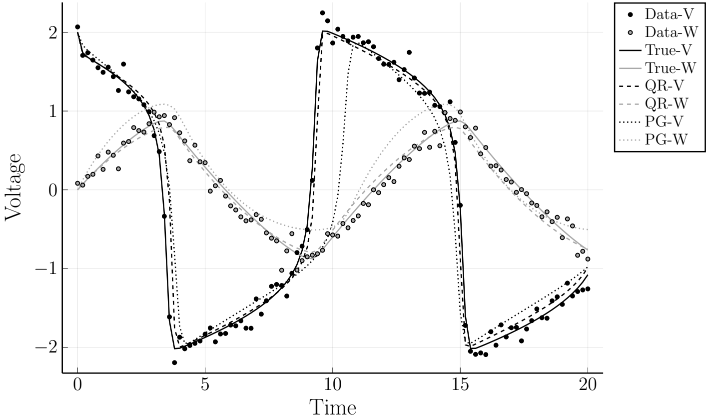
TR Results II



b Objective Descent

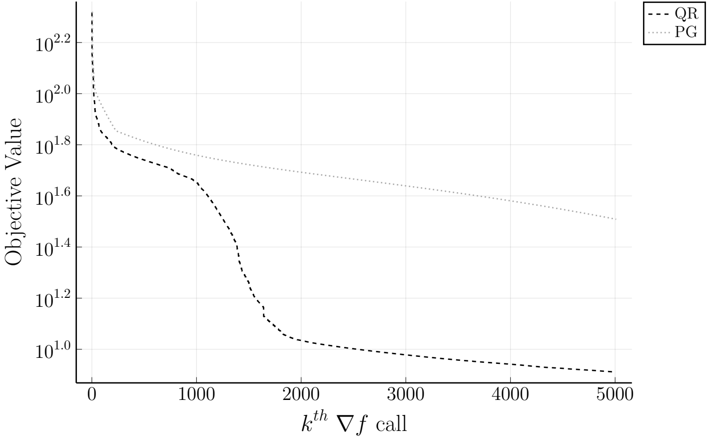
QR Results I

Figure 3: Fitzhugh-Nagumo solution ((15a), (15b)) for $h(x) = \lambda\|x\|_0$ in (14) with QR. 5000 Max-Iter



a Solution Comparisons

QR Results II



b Objective Descent

Conclusions & Current Work

- Theoretical
  - General Prox Operator computation?
  - Extension to penalty methods
  - Different B_k operators - LBFGS, LSR1, Gauss-Newton
- Practical
  - Finalize numerical Julia packages/tests (https://github.com/UW-AMO/TRNC) - extensions to C++
  - Add in constraints/barrier methods
  - Implementation for harder PDE examples

Nonlinear Least Squares

- Extend framework to NLS examples with Gauss-Newton Hessian operators:

$$\underset{x}{\text{minimize}} \|F(x)\|^2 + h(x) \tag{17}$$

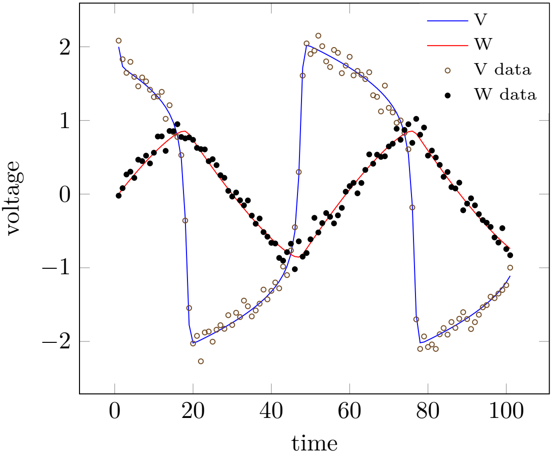
- New "model" takes the form of:

$$m_k(s; x) := \tfrac{1}{2}\|J(x)s + F(x)\|^2 + \psi(s; x) \tag{18}$$

with $J(x)$ the Jacobian of $F(x)$; we propose QR extension as well.

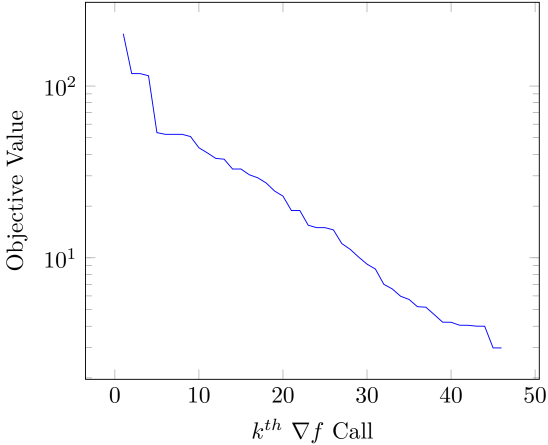
Preliminary Numerical Results I

Figure 4: Fitzhugh-Nagumo solution ((15a), (15b)) for $h(x) = \lambda \|x\|_0$ in (14) with ℓ_∞-norm, LMTR.



a Solution for LMTR

Preliminary Numerical Results II



b Objective Descent for LMTR

Future Directions

- Inexact methods for PDE-constrained optimization
  - Imprecise gradient, subgradients
  - Inexact prox solution for incomputable proxes
  - Semismooth regularizer specifics
- Fast linear algebra for ν_k computation
- Fidelity-tuning for numerical simulations
- Applications to PDE-constrained inversion in CFD, earth/climate modeling, ... huge host of national lab resources
- Numerical software/HPC implementation - Trilinos/ROL, Dakota, GPU compatibility

Thank you!

- Questions?
- Acknowledgments - Sasha and Dominique
- Support - DOE CSGF

References I

Robert Baraldi, Aleksandr Aravkin, and Dominique Orban. A proximal quasi-newton trust-region method for nonsmooth regularized optimization. *SIAM Journal of Optimization - in revision*, 2020. URL https://eartharxiv.org/repository/view/2267/.

J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. 1(146):459—-494, 2014a. DOI: 10.1007/s10107-013-0701-9.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014b.

Coralia Cartis, Nicholas I. M. Gould, and Ph. L. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. 21(4): 1721–1739, 2011. DOI: 10.1137/11082381X.

References II

A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. Number 1 in MOS-SIAM Series on Optimization. 2000. DOI: 10.1137/1.9780898719857.

J.E. Dennis, SB.B. Li, and R.A. Tapia. A unified approach to global convergence of trust region methods for nonsmooth optimization. 1(68):319—-346, 1995. DOI: 10.1007/BF01585770.

Jorge J. Moré and D. C. Sorensen. Computing a trust region step. 4(3):553–572, 1983. DOI: 10.1137/0904038.

R.T. Rockafellar and R.J.B. Wets. *Variational Analysis*, volume 317. 1998. DOI: 10.1007/978-3-642-02431-3.

T. Steihaug. The conjugate gradient method and trust regions in large scale optimization. 20(3):626–637, 1983. DOI: 10.1137/0720042.

L. Stella, A. Themelis, P. Sopasakis, and P. Patrinos. A simple and efficient algorithm for nonlinear model predictive control. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1939–1944, 2017. DOI: 10.1109/CDC.2017.8263933.

References III

Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. 28(3):2274–2303, 2018. DOI: 10.1137/16M1080240.

Y.-X. Yuan. Conditions for convergence of trust region algorithms for nonsmooth optimization. 1(31):220—228, 1985. DOI: 10.1007/BF02591750.