A PROXIMAL MODIFIED QUASI-NEWTON METHOD FOR NONSMOOTH REGULARIZED OPTIMIZATION

Youssef Diouane [1] Mohamed L. Habiboullah [1] Dominique Orban [1]
[1]GERAD and Polytechnique Montréal

dominique.orban@gerad.ca

November 8th, 2024
Midwest Optimization Meeting
WATERLOO

General Problem

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

- $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}^1$;
- $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is proper lsc;
- both may be nonconvex;
- there could be bound constraints and/or h could model hidden constraints.

Previous Work

Aravkin, Baraldi, and Orban (2022) and Aravkin, Baraldi, and Orban (2024):

1. R2 (quadratic regularization): similar to proximal gradient in flavor but does not require knowledge of any Lipschitz constant

$$\underset{s}{\text{minimize}} \;\; f(x_k) + \nabla f(x_k)^T s + \tfrac{1}{2}\sigma_k \|s\|^2 + h(x_k + s)$$

2. TR (trust region): exact or inexact Hessian quadratic model; step computed with R2

$$\underset{s}{\text{minimize}} \;\; f(x_k) + \nabla f(x_k)^T s + \tfrac{1}{2} s^T B_k s + h(x_k + s) \quad \text{subject to} \;\; s \in \Delta_k \mathbb{B}$$

3. LM and LMTR: variants for nonlinear least squares.

All have worst-case evaluation complexity $O(\epsilon^{-2})$ to bring a stationarity measure below $\epsilon > 0$ if $\{B_k\}$ is bounded.

Overview of Literature

Existing literature makes restrictive assumptions, e.g.,

- Cartis, Gould, and Toint (2011): $f \in \mathcal{C}^2$, h globally Lipschitz and convex;
- Kanzow and Lechner (2021): h convex, $\{B_k\}$ bounded;
- Li and Lin (2015): $f + h$ coercive.

Research on nonconvex regularized problems includes

- Bolte, Sabach, and Teboulle (2014): alternating minimization;
- Bo, Csetnek, and László (2016): linesearch with inertia;
- Stella et al. (2017): PANOC; LBFGS with linesearch;
- Themelis, Stella, and Patrinos (2018): ZeroFPR; nonmonotone LBFGS.

Complexity

AFAIK, all complexity analyses require $\{B_k\}$ bounded, e.g., Cartis, Gould and Toint's book.

Except:

1. Leconte and Orban (2023b), who assume $\|B_k\| = O(|\mathcal{S}_k|^p)$, $0 \leqslant p < 1$, where $\mathcal{S}_k$ is the set of *successful* iterations of a trust-region method up to iteration k;
2. Diouane, Habiboullah, and Orban (2024), who generalize their result to $p = 1$ for smooth optimization.

Objectives

1. Devise an efficient proximal-type method to drive a stationarity measure $\to 0$;
2. compute steps via a quadratic model of f + a model of h;
3. both models may be nonconvex;
4. allow *unbounded* model Hessians;
5. determine worst-case complexity in the presence of unbounded model Hessians;
6. provide an alternative to the trust-region method of Aravkin, Baraldi, and Orban (2022) where proximal operators must take the trust-region constraint into account.

Basics

First-order optimality conditions: $0 \in \nabla f(x) + \partial h(x)$.

Proximal operator: $\text{prox}_{\nu h}(q) := \text{argmin}_x \ \frac{1}{2}\nu^{-1}\|x - q\|_2^2 + h(x)$

Proximal gradient iteration: $x_{k+1} \in \text{prox}_{\nu_k h}(x_k - \nu_k \nabla f(x_k))$

More instructively: $x_{k+1} = x_k + s_k$ with

$$s_k \in \underset{s}{\text{argmin}} \ \underbrace{f(x_k) + \nabla f(x_k)^T s}_{\approx f(x_k+s)} + \frac{1}{2}\nu_k^{-1}\|s\|_2^2 + h(x_k + s)$$

If ∇f is L-Lipschitz and $\nu_k < 1/L$, we obtain decrease in f.

Models

For $\nu > 0$, $\sigma \geqslant 0$, $x \in \mathbb{R}^n$, and $B(x) = B(x)^T \in \mathbb{R}^{n\times n}$,

$$\begin{aligned} \varphi_{\text{cp}}(s;x) &:= f(x) + \nabla f(x)^T s \\ \psi(s;x) &\approx h(x+s) \\ m_{\text{cp}}(s;x,\nu^{-1}) &:= \varphi_{\text{cp}}(s;x) + \tfrac{1}{2}\nu^{-1}\|s\|^2 + \psi(s;x) \end{aligned}$$

and

$$\begin{aligned} \varphi(s;x) &:= f(x) + \nabla f(x)^T s + \tfrac{1}{2} s^T B(x) s \\ m(s;x,\sigma) &:= \varphi(s;x) + \tfrac{1}{2}\sigma\|s\|^2 + \psi(s;x). \end{aligned}$$

Assume:

1. $\psi(\cdot;x)$ proper, lsc, $\psi(0;x) = h(x)$, and $\partial\psi(0;x) = \partial h(x)$;
2. $\psi(\cdot;x)$ uniformly prox-bounded over all x.

Stationarity Measure

Let $s_{\text{cp}} := s_{\text{cp}}(x, \nu^{-1}) \in \operatorname{argmin}_s \; m_{\text{cp}}(s; x, \nu^{-1})$.

Definition: Cauchy Decrease.

$$\xi_{\text{cp}}(x, \nu^{-1}) := f(x) + h(x) - (\varphi_{\text{cp}}(s_{\text{cp}}; x) + \psi(s_{\text{cp}}; x)).$$

$$h = 0 \quad \Longrightarrow \quad \xi_{\text{cp}}(x, \nu^{-1}) = \tfrac{1}{2}\nu \|\nabla f(x)\|^2.$$

Definition: Stationarity Measure.

$$\nu^{-1/2} \xi_{\text{cp}}(x, \nu^{-1})^{1/2}.$$

Lemma 1

$\xi_{\text{cp}}(x, \nu^{-1}) = 0 \quad \Longrightarrow \quad 0 \in \nabla f(x) + \partial h(x).$

Algorithm: Main Ideas

1. $s_k \approx \operatorname{argmin}_s \; m(s; x_k, \sigma_k)$;
2. s_k must satisfy *sufficient decrease*:

$$\varphi(0; x_k) + \psi(0; x_k) - (\varphi(s_k; x_k) + \psi(s_k; x_k)) \geqslant (1 - \theta_1)\ \xi_{\text{cp}}(x_k, \nu_k^{-1})$$

   for fixed $0 < \theta_1 < 1$ and well-chosen ν_k;
3. accept/reject s_k by assessing decrease in $f + h$;
4. update σ_k accordingly.

Lemma 2
Let s_{cp} be computed with $\nu := \theta_1/(\|B(x)\| + \sigma)$.
If $m(s; x, \sigma) \leqslant m(s_{\text{cp}}; x, \sigma)$, s satisfies sufficient decrease.

Algorithm R2N

1: Choose x_0, $\sigma_0 > 0$, $0 < \theta_1 < 1 < \theta_2$, $0 < \eta_1 \leqslant \eta_2 < 1$ and $0 < \gamma_3 \leqslant 1 < \gamma_1 \leqslant \gamma_2$.
2: **for** $k = 0, 1, \ldots$
3: Choose $B_k = B_k^T$ and set $\nu_k := \theta_1/(\|B_k\| + \sigma_k)$.
4: Compute $s_{k,\text{cp}}$ and $\xi_{\text{cp}}(x_k, \nu_k^{-1})$. *// one prox*
5: Compute s_k such that $m(s_k; x_k, \sigma_k) \leqslant m(s_{k,\text{cp}}; x_k, \sigma_k)$.
6: If $\|s_k\| > \theta_2 \|s_{k,\text{cp}}\|$, reset $s_k = s_{k,\text{cp}}$.
7: Compute

$$\rho_k := \frac{f(x_k) + h(x_k) - (f(x_k + s_k) + h(x_k + s_k))}{\varphi(0; x_k) + \psi(0; x_k) - (\varphi(s_k; x_k) + \psi(s_k; x_k))}.$$

8: If $\rho_k \geqslant \eta_1$, set $x_{k+1} = x_k + s_k$. Otherwise, set $x_{k+1} = x_k$.
9: Update the regularization parameter according to

$$\sigma_{k+1} \in \begin{cases} [\gamma_3\sigma_k,\ \sigma_k] & \text{if } \rho_k \geqslant \eta_2, \quad \text{very successful iteration} \\ [\sigma_k,\ \gamma_1\sigma_k] & \text{if } \eta_1 \leqslant \rho_k < \eta_2, \quad \text{successful iteration} \\ [\gamma_1\sigma_k,\ \gamma_2\sigma_k] & \text{if } \rho_k < \eta_1. \quad \text{unsuccessful iteration} \end{cases}$$

Convergence: Assumptions

$$|h(x_k + s_k) - \psi(s_k; x_k)| = o(\|s_k\|) \quad \text{as} \quad s_k \to 0.$$

Satisfied if

- $\psi(s; x) := h(x + s)$;
- $h(x) = g(c(x))$, $\psi(s; x_k) = g(c(x_k) + \nabla c(x_k)s)$ with g Lipschitz, ∇c Hölder.

$$\sum_{k \in \mathbb{N}} \frac{1}{\max_{0 \leqslant j \leqslant k} \|B_j\| + 1} = +\infty.$$

Conn, Gould, and Toint (2000): BFGS and SR1 approximations are $B_k = O(k)$.
Powell (2010): same for PSB.

Convergence

Let $\mathcal{S}$ be the set of successful iterations.

Theorem 3
If $|\mathcal{S}| < \infty$, $x_k = x^*$ *for all sufficiently large* k, $\liminf \nu_k^{-\frac{1}{2}} \xi_{\text{cp}}(x_k, \nu_k^{-1})^{\frac{1}{2}} = 0$.

Theorem 4
If $|\mathcal{S}| = \infty$ *and* $(f + h)(x_k) \geqslant (f + h)_{\text{low}}$ *for all* $k \in \mathbb{N}$. *Then,*
$\liminf \nu_k^{-1/2} \xi_{\text{cp}}(x_k, \nu_k^{-1})^{1/2} = 0$.

Complexity: Assumptions

$$|(f+h)(x_k+s_k) - (\varphi+\psi)(s_k;x_k)| \leqslant \kappa_{\mathrm{m}}(1+\|B_k\|)\|s_k\|^2.$$

Satisfied if ∇f Lipschitz.

There are $\mu > 0$ and $0 \leqslant p \leqslant 1$ such that

$$\|B_k\| \leqslant \mu(1+|\mathcal{S}_k|^p) \quad \text{for all } k.$$

Results are similar if $\|B_k\| \leqslant \mu(1+k^p)$.

Complexity: Results

Theorem 5
If $|\mathcal{S}| < \infty$, $x_k = x^*$ for all sufficiently large k where x^* is stationary.

Complexity: Results

Let $\epsilon > 0$ and k_ϵ the first iteration such that $\nu_k^{-1/2} \xi_{\text{cp}}(x_k, \nu_k^{-1})^{1/2} \leqslant \epsilon$.
Let $\mathcal{S}(\epsilon) = \{k \in \mathcal{S} \mid k < k_\epsilon\}$.

Theorem 6

If $|\mathcal{S}| = \infty$,

1. *If* $0 \leqslant p < 1$,

$$|\mathcal{S}(\epsilon)| \leqslant ((1-p)\kappa_1 \epsilon^{-2} + 1)^{1/(1-p)} - 1 = O(\epsilon^{-2/(1-p)}),$$

where

$$\kappa_1 = \frac{((f+h)(x_0) - (f+h)_{\text{low}})(b_{\max} + 2\mu(1+b_{\max}))}{\eta_1 \theta_1 (1-\theta_1)}, \quad b_{\max} \sim \kappa_{\text{m}}.$$

2. *If* $p = 1$,

$$|\mathcal{S}(\epsilon)| \leqslant \exp(\kappa_1 \epsilon^{-2}) - 1.$$

Total Number of Iterations

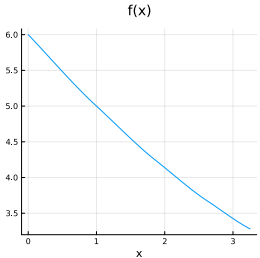
Because

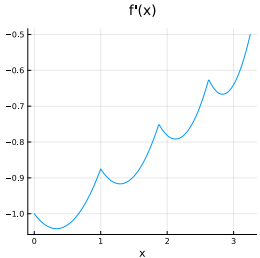
$$|\mathcal{U}(\epsilon)| \leqslant |\log_{\gamma_1}(\gamma_3)||\mathcal{S}(\epsilon)| + \log_{\gamma_1}(1 + \mu(1 + |\mathcal{S}(\epsilon)|^p)) + \frac{\log(b_{\max}/\sigma_0)}{\log(\gamma_1)},$$

the total number of iterations satisfies a bound similar to that on $|\mathcal{S}(\epsilon)|$.

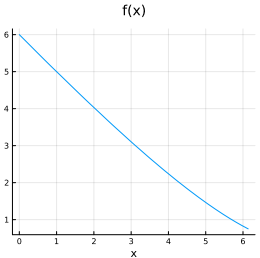
The above improves the constant in the bound of Leconte and Orban (2023b) for $0 \leqslant p < 1$.

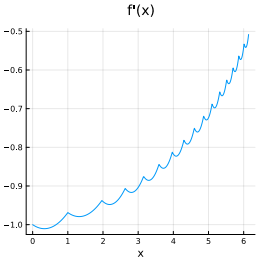
Bounds are Sharp



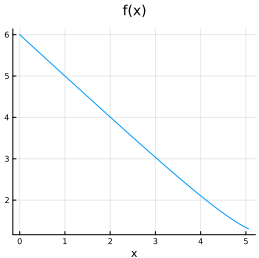


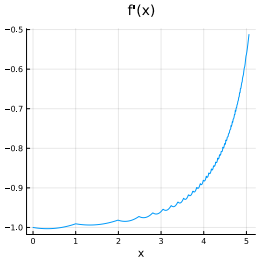
$p = 0$





$p = 0.5$





$p = 1$

Algorithmic Variants

- Non-monotone variant;
- diagonal Hessian variant R2DH: with B_k diagonal and h separable, the subproblem can sometimes be solved analytically (e.g., $\|\cdot\|_0$, $\|\cdot\|_1$, etc.) (Leconte and Orban, 2023a). This variant can be used as standalone solver or subproblem solver in place of R2;
- the non-monotone spectral-gradient approximation $B_k = (s_k^T y_k / s_k^T s_k) I$ tends to perform best.

Implementation

R2N, R2, TR, LM and LMTR are all part of the RegularizedOptimization.jl Julia module (Baraldi, Leconte, and Orban, 2024):

`https://github.com/JuliaSmoothOptimizers/RegularizedOptimization.jl`.

Parameters: $\theta_1 \approx 0.999$, $\theta_2 \approx 10^{15}$, $\eta_1 \approx 10^{-4}$, $\eta_2 = 0.9$, and $\sigma_0 \approx 10^{-6}$

Stopping condition:

$$\nu_k^{-1/2}\xi_{\text{cp}}(x_k, \nu_k)^{1/2} < \epsilon_a + \epsilon_r \nu_0^{-1/2}\xi_{\text{cp}}(x_0, \nu_0)^{1/2}, \quad (\epsilon_a = \epsilon_r \approx 10^{-5}).$$

Subsolver stopping condition decreases progressively.

Basis Pursuit Denoise

$$\underset{x}{\text{minimize}} \ \tfrac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_0, \quad \lambda = 0.1\,\|A^Tb\|_\infty,$$

A is 2000 $\times$5120, $AA^T = I$, $b = Ax_{\text{true}} + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, 10^{-2})$, $\|x_{\text{true}}\|_0 = 100$.

Solver	f	h/λ	$\Delta(f+h)$	$\sqrt{\xi/\nu}$	$\#f$	$\#\nabla f$	**#prox**	$t(s)$
R2	9.22e−02	100	8.74e−07	8.0e−04	366	362	366	7.37
R2DH-Spec-NM	9.22e−02	100	5.05e−07	5.9e−04	57	56	56	0.89
R2DH-Spec	9.22e−02	100	0.00e+00	3.6e−04	86	57	85	0.97
R2DH-Andrei	3.42e+00	2926	1.58e+02	8.6e−01	1001	988	1991	20.61
R2DH-PSB	3.69e−06	3400	1.81e+02	6.7e−04	592	591	1194	11.85
R2DH-DBFGS	9.22e−02	100	7.26e−07	6.4e−04	300	153	299	5.53

Matrix Completion: Rank

$$\underset{X}{\text{minimize}}\ \tfrac{1}{2}\|P_\Omega(X-M)\|_F^2 + \lambda h(X), \quad \lambda = 0.1,\ h(X) = \text{rank}(X) \text{ or } \|X\|_\star,$$

$$M = (1-c)(X_r + \mathcal{N}(0,\sigma_A^2)) + c(X_r + \mathcal{N}(0,\sigma_B^2)),$$

with X_r of size 120 and rank 40.

Solver	f	h/λ	$\Delta(f+h)$	$\sqrt{\xi/\nu}$	$\#f$	$\#J$	**#prox**	$t(s)$
R2	2.78e−07	111	1.60e+00	2.3e−04	37	31	37	0.20
R2DH	1.34e−11	95	0.00e+00	3.7e−06	28	15	27	0.17
LM-R2DH	2.16e−08	111	1.60e+00	1.7e−04	2	115	48	0.29
LM-R2	1.67e−12	111	1.60e+00	6.4e−07	3	183	61	0.33

Matrix Completion: Nuclear Norm

Solver	f	h/λ	$\Delta(f+h)$	$\sqrt{\xi/\nu}$	$\#f$	$\#J$	**#prox**	$t(s)$
R2	1.00e−02	7.5e+00	1.30e−05	3.6e−04	82	52	82	0.43
R2DH	1.00e−02	7.5e+00	7.28e−07	2.3e−04	43	19	42	0.21
LM-R2DH	1.00e−02	7.5e+00	1.98e−10	3.1e−06	3	246	108	0.63
LM-R2	1.00e−02	7.5e+00	0.00e+00	2.0e−06	3	340	144	0.86

Binary Classifier

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \ \tfrac{1}{2}\|\mathbf{1} - \tanh(b \odot \langle A, x\rangle)\|^2 + \lambda\|x\|_0,$$

- Distinguish between "1" and "7" in m MNIST images;
- $\lambda = 0.1$;
- A is $m \times n$ with $m = 13,007$ and $n = 784 = 28^2$.

Solver	f	h/λ	$\Delta(f+h)$	$\sqrt{\xi/\nu}$	$\#f$	$\#\nabla f$	**#prox**	$t(s)$
R2	1.94e+01	175	5.33e+00	1.6e−02	1002	775	1002	9.47
R2DH	1.59e+01	157	0.00e+00	1.8e−03	781	441	780	4.97
R2N-R2	1.85e+01	233	1.02e+01	2.3e−03	59	59	10404	1.21
R2N-R2DH	1.67e+01	237	8.84e+00	2.5e−03	71	71	11284	1.29

Takeaway Points

1. Weak assumptions for convergence: $f \in C^1$, h proper lsc prox-bounded;
2. ∇f Lipschitz for complexity;
3. $\epsilon^{-2/(1-p)}$ sharp complexity if $\|B_k\| \approx |\mathcal{S}_k^p$ with $0 \leqslant p < 1$;
4. $\exp(\epsilon^{-2})$ sharp complexity if $\|B_k\| \approx |\mathcal{S}_k|$;
5. worse sharp complexity if B_k grows faster;
6. efficient Julia implementations of R2, R2N, TR, LM, LMTR.

Wish List

- Inexact prox evaluations (ongoing work);
- an efficient solver to minimize $\frac{1}{2}\|Ax - b\|^2 + h(x)$;
- weaker assumptions on models of h;
- better characterization of limit points.

Thank You!

dominique.orban@gerad.ca

dpo.github.io

jso.dev

References

Aravkin, A. Y., R. Baraldi, and D. Orban (2022). "A Proximal Quasi-Newton Trust-Region Method for Nonsmooth Regularized Optimization". In: *SIAM J. Optim.* 32.2, pp. 900–929. DOI: 10.1137/21M1409536.

Aravkin, Aleksandr Y., Robert Baraldi, and Dominique Orban (2024). "A LevenbergMarquardt Method for Nonsmooth Regularized Least Squares". In: *SIAM J. Sci. Comput.* 46.4, A2557–A2581. DOI: 10.1137/22M1538971.

Baraldi, R., G. Leconte, and D. Orban (2024). *RegularizedOptimization.jl: Algorithms for Regularized Optimization*. DOI: 10.5281/zenodo.6940313. URL: https://github.com/JuliaSmoothOptimizers/RegularizedOptimization.jl.

Bolte, J., S. Sabach, and M. Teboulle (2014). "Proximal alternating linearized minimization for nonconvex and nonsmooth problems". In: *Math. Program.* 146, pp. 459–494. DOI: 10.1007/s10107-013-0701-9.

Bo, R. I., E. R. Csetnek, and S.C. László (2016). "An inertial forwardbackward algorithm for the minimization of the sum of two nonconvex functions". In: *EURO J. Comput. Optim.* 4, pp. 3–25. DOI: 10.1007/s13675-015-0045-8.

Cartis, Coralia, Nicholas I. M. Gould, and Ph. L. Toint (2011). "On the Evaluation Complexity of Composite Function Minimization with Applications to Nonconvex Nonlinear Programming". In: *SIAM J. Optim.* 21.4, pp. 1721–1739. DOI: 10.1137/11082381X.

Conn, A. R., N. I. M. Gould, and Ph. L. Toint (2000). *Trust-region methods*. MOS-SIAM Series on Optimization 1. Philadelphia, USA: SIAM. DOI: 10.1137/1.9780898719857.

Diouane, Youssef, Mohamed Laghdaf Habiboullah, and Dominique Orban (2024). *Complexity of trust-region methods in the presence of unbounded Hessian approximations*. Cahier G-2024-43. Montréal, Canada: GERAD. DOI: 10.48550/arXiv.2408.06243. URL: https://www.gerad.ca/fr/papers/G-2024-43.

Kanzow, C and T Lechner (2021). "Globalized inexact proximal Newton-type methods for nonconvex composite functions". In: *Comput. Optim. Appl.* 78.2, pp. 377–410. DOI: 10.1007/s10589-020-00243-6.

Leconte, G. and D. Orban (2023a). *The Indefinite Proximal Gradient Method*. Cahier G-2023-37. Montréal, QC, Canada: GERAD. DOI: 10.13140/RG.2.2.11836.41606.

Leconte, Geoffroy and Dominique Orban (2023b). *Complexity of trust-region methods with unbounded Hessian approximations for smooth and nonsmooth optimization*. Cahier G-2023-65. Montréal, QC, Canada: GERAD. URL: https://www.gerad.ca/fr/papers/G-2023-65.

Li, Huan and Zhouchen Lin (2015). "Accelerated Proximal Gradient Methods for Nonconvex Programming". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, pp. 379–387. URL: http://irc.cs.sdu.edu.cn/973project/result/download/2015/28.AcceleratedProximal.pdf.

Powell, M. J. D. (2010). "On the convergence of a wide range of trust region methods for unconstrained optimization". In: *IMA J. Numer. Anal.* 30.1, pp. 289–301. DOI: 10.1093/imanum/drp021.

Stella, L. et al. (2017). "A simple and efficient algorithm for nonlinear model predictive control". In: *2017 IEEE 56th Annual*