日本オペレーションズ・リサーチ学会
2023 年春季研究発表会

Accelerating Inexact Successive Quadratic Approximation for Regularized Optimization Through Manifold Identification

Academia Sinica　LEE Ching-pei

1. Introduction

Consider the following regularized optimization problem:

$$\min_{x\in\mathcal{E}} \quad F(x) \coloneqq f(x) + \Psi(x), \tag{1}$$

where $\mathcal{E}$ is a Euclidean space with an inner product $\langle\cdot,\,\cdot\rangle$ and its induced norm $\|\cdot\|$, the regularizer Ψ is extended-valued, convex, proper, and lower-semicontinuous, f is continuously differentiable with Lipschitz-continuous gradients, and the solution set Ω is non-empty.

One widely-used method for (1) is inexact successive quadratic approximation (ISQA). At the tth iteration with iterate x^t, ISQA obtains the update direction p^t by approximately solving

$$p^t \approx \arg\min_{p\in\mathcal{E}} \quad Q_{H_t}\left(p; x^t\right), \tag{2}$$

$$\begin{aligned} Q_{H_t}\left(p; x^t\right) \coloneqq \left\langle \nabla f\left(x^t\right),\, p\right\rangle + \frac{1}{2}\langle p,\, H_t p\rangle \\ + \Psi\left(x^t + p\right) - \Psi\left(x^t\right), \end{aligned} \tag{3}$$

where H_t is a self-adjoint positive-semidefinite linear endomorphism of $\mathcal{E}$. The iterate is then updated along p^t with a step size $\alpha_t > 0$.

ISQA is among the most efficient for (1). Its variants differ in the choice of H_t and α_t, and how accurately (2) is solved. In this class, proximal Newton (PN) and proximal quasi-Newton (PQN) are popular for their fast convergence in iterations. Regrettably, their subproblem has no closed-form solution as H_t is non-diagonal, so one needs to use an iterative solver for (2) and the running time to reach the accuracy requirement can hence be lengthy. For example, to attain the same superlinear convergence as truncated Newton for smooth optimization, PN that takes $H_t = \nabla^2 f$ requires increasing accuracies for the subproblem solution, implying a growing and unbounded number of inner iterations of the subproblem solver. Its superlinear convergence thus gives little practical advantage in running time. On the contrary, in smooth optimization, one can solve (2) with a bounded cost by either conjugate gradient (CG) or matrix factorizations since $\Psi \equiv 0$. The advantage of second-order methods over the first-order ones in regularized optimization is therefore not as significant as that in smooth optimization.

A possible remedy when Ψ is partly smooth [2] is to switch to smooth optimization after identifying an active manifold $\mathcal{M}$ that contains a solution $\hat{x}$ to (1) and makes Ψ confined to it smooth. We say an algorithm can identify $\mathcal{M}$ if there is a neighborhood U of $\hat{x}$ such that $x^t \in U$ implies $x^{t+1} \in \mathcal{M}$, and call it possesses the manifold identification property. Unfortunately, for ISQA, this property in general only holds when (2) is always solved exactly.

Interestingly, in our numerical experience, ISQA with approximate subproblem solutions, even without an increasing solution precision and on problems that are not strongly convex, often identifies the active manifold rapidly. We thus aim to provide theoretical support for such a phenomenon and utilize it to devise more efficient and practical methods that trace the superior performance of second-order methods in smooth optimization.

In this work, we show that ISQA essentially possesses the manifold identification property, by giving a sufficient condition for inexact solutions of (2) in ISQA to identify the active manifold that is satisfied by the output of most of the widely-used subproblem solvers even if (2) is solved arbitrarily roughly. We also show convergence of the iterates under a sharpness condition widely-seen in real-world problems that generalizes the quadratic growth condition and the weak sharp minima. Based on these results, we propose an improved, practical algorithm ISQA$^+$ that switches to smooth optimization after the active

manifold is presumably identified. We show that ISQA$^+$ is superior to existing PN-type methods as it possesses the same superlinear and even quadratic rates in iterations but has bounded per-iteration cost. ISQA$^+$ hence also converges superlinearly in running time, which, to our best knowledge, is the first of the kind. Numerical results also confirm ISQA$^+$'s much improved efficiency over PN and PQN.

2. Main Theoretical Results

We assume our regularizer Ψ is partly smooth at a solution x^*, and the definition is as follows.

Definition 1 (Partly smooth) *A convex function Ψ is partly smooth at a point x^* relative to a set $\mathcal{M}$ containing x^* if $\partial\Psi(x^*) \neq \emptyset$ and:*

1. *$\mathcal{M}$ is a $\mathcal{C}^2$-manifold and $\Psi|_{\mathcal{M}}$ is $\mathcal{C}^2$ around x^*.*
2. *The affine span of $\partial\Psi(x)$ is a translate of the normal space to $\mathcal{M}$ at x^*.*
3. *$\partial\Psi$ is continuous at x^* relative to $\mathcal{M}$.*

Intuitively, it means Ψ is smooth in $\mathcal{M}$ near x^* but changes sharply along directions leaving $\mathcal{M}$.

For (2), we denote its optimal solution by p^{t*}. When there is no ambiguity, we abbreviate $Q_{H_t}(\cdot; x^t)$ to $Q_t(\cdot)$, $Q_t(p^t)$ to $\hat{Q}_t$, and $Q_t(p^{t*})$ to Q_t^*. Our main theoretical result is as follows.

Theorem 1 *Consider a point x^* satisfying the nondegenerate condition*

$$0 \in \mathrm{relint}\partial F(x^*) = \nabla f(x^*) + \mathrm{relint}\left(\partial\Psi(x^*)\right) \quad (4)$$

with Ψ convex, proper, closed, and partly smooth at x^* relative to some manifold $\mathcal{M}$. Assume f is locally L-smooth for $L > 0$ around x^*. If the subproblem (2) is approximately solved such that

$$\hat{Q}_t - Q_t^* \leq \eta\left(Q_t(0) - Q_t^*\right) = -\eta Q_t^*. \quad (5)$$

for some $\eta \in [0, 1)$, there are $M, m > 0$ such that

$$M \succeq H_t \succeq m,\ \forall t \geq 0, \quad (6)$$

and the update direction p^t satisfies

$$x^t + p^t = \mathrm{prox}_{\Psi}^{\Lambda_t}\left(y^t - \Lambda_t^{-1}\left(\nabla f\left(x^t\right) + H_t\left(y^t - x^t\right) + s^t\right)\right), \quad (7)$$

where s^t satisfies $\|s^t\| \leq R\left(\left\|y^t - (x^t + p^{t*})\right\|\right)$ for some $R : [0, \infty) \to [0, \infty)$ continuous in its domain with $R(0) = 0$, Λ_t is symmetric and positive definite with $M_1 \succeq \Lambda_t$ for $M_1 > 0$, and y^t satisfies the following for some $\nu > 0$ and $\eta_1 \geq 0$,

$$\left\|\left(y^t - x^t\right) - p^{t*}\right\| \leq \eta_1\left(Q_t(0) - Q_t^*\right)^{\nu}. \quad (8)$$

Then there exists $\epsilon, \delta > 0$ such that $\|x^t - x^*\| \leq \delta, |Q_t^*| \leq \epsilon$, and $\alpha_t = 1$ imply $x^{t+1} \in \mathcal{M}$.

The mild condition in Theorem 1 includes almost all practical first-order solvers for (2), including

- Proximal gradient (PG)
- Accelerated proximal gradient (APG)
- Prox-SAGA/SVRG
- Block-coordinate descent (when the regularizer and the manifold are block-separable)

3. Efficient ISQA with Superlinear Convergence in Running Time

Now that we know ISQA identifies the active manifold $\mathcal{M}$, we utilize the fact that the optimization problem reduces to a smooth one after $\mathcal{M}$ is identified to devise a more efficient method ISQA$^+$, which has the following two stages.

- ISQA stage:
  1. Solve (2) and adjust H_t until $x^t + p^t$ provides sufficient objective decrease.
  2. If x^t stays within the same manifold for T iterations then switch to the smooth stage.
- Smooth stage:
  1. Conduct one iteration of regularized Newton within the current manifold.
  2. Conduct one iteration of PG.
  3. If the manifold changes after PG or the Newton step fails to decrease the objective, go back to the ISQA stage.

As long as $\nabla^2 F\,|_{\mathcal{M}}$ is positive definite around the point of convergence x^*, ISQA$^+$ achieves superlinear convergence in both number of iterations (from the Newton step) and in the running time (as the smooth Newton step can be computed with an upper-bounded cost).

4. Numerical Results

Numerical results will be presented in the talk.

References

[1] Ching-pei Lee. Accelerating inexact successive quadratic approximation for regularized optimization through manifold identification. *Mathematical Programming*, 2023. In press.

[2] Adrian S. Lewis. Active sets, nonsmoothness, and sensitivity. 13(3):702–725, 2002.