Accelerating Inexact Successive Quadratic Approximation for Regularized Optimization Through Manifold Identification

LEE Ching-pei

Outline

Regularized Optimization Problem

Consider the following regularized optimization problem:

$$\min_{x} F(x) := f(x) + \Psi(x), \qquad \text{(REG)}$$

- $f : \mathbb{R}^n \to \mathbb{R}$: L-Lipschitz-continuously differentiable (L-smooth)
- $\Psi : \mathbb{R}^n \to \mathbb{R}$: convex, extended-valued, proper, and closed, but might be nonsmooth.
- F is lower-bounded and the solution set Ω of (REG) is non-empty.

Inexact Successive Quadratic Approximation (ISQA)

At the tth iteration, with iterate x^t, find an update direction p^t by solving

$$p^t \approx \operatorname*{argmin}_{p \in \mathbb{R}^n} \quad Q_{H_t}^{x^t}\left(p; x^t\right) := \nabla f\left(x^t\right)^\top d + \frac{1}{2} d^\top H_t d + \Psi\left(x^t + d\right) - \Psi\left(x^t\right) \quad \text{(SUBPROB)}$$

for some symmetric and positive-semidefinite H_t.

- A stepsize α_t along p^t is then decided for updating the iterate
- Many existing algorithms included in this framework: proximal Newton (PN) when $H_t = \nabla^2 f(x^t)$, proximal quasi-Newton (PQN), proximal gradient, and so on
- Subproblem has no closed-form solution when H_t is not diagonal: apply an iterative solver to obtain an approximate solution
- abbreviation: $Q_t(p) := Q_{H_t}^{x^t}(p; x^t)$

Solution Inexactness

- For PN and PQN, under suitable conditions, superlinear convergence in the number of times updating x^t can still be obtained
- Similar to the smooth case (i.e. $\Psi \equiv 0$): requires increasing solution accuracy of (SUBPROB)
- Unlike the smooth case: no closed-form or finite-termination solver (direct inverse/matrix factorization/conjugate gradient) exists for (SUBPROB)
- Superlinear convergence only in theory and in outer iterations, but not observed in real running time

Possible Remedy

- If Ψ is partly smooth around a point x^*, and the iterates converge to x^*, then after identifying the active manifold $\mathcal{M} \ni x^*$ such that $\Psi|_{\mathcal{M}}$ is smooth, we can switch to smooth optimization
- Partly smooth: function value is smooth along a manifold but changes drastically along directions leaving the manifold
- An algorithm identifies $\mathcal{M}$ if there is a neighborhood $U \ni x^*$ such that $x^t \in U$ implies $x^{t+1} \in \mathcal{M}$
- Call such an algorithm possesses the manifold identification property
- If (SUBPROB) is always solved exactly, it is known that the active manifold can be identified
- But due to the inexactness in subproblem solution, ISQA in general does not have the manifold identification property

ISQA Cannot Identify Active Manifold in General

Example 1

$$\min_{x \in \mathbb{R}^2} (x_1 - 2.5)^2 + (x_2 - 0.3)^2 + \|x\|_1,$$

- $\Psi(\cdot) = \|\cdot\|_1$, the only solution is $x^* = (2, 0)$, and $\|x\|_1$ is smooth relative to $\mathcal{M} = \{x \mid x_2 = 0\}$ around x^*.
- Consider $\{x^t\}$ with $x_1^t = 2 + f(t), x_2^t = f(t)$, for some $f(t) > 0$ with $f(t) \downarrow 0$, $H_t \equiv I, \alpha_t \equiv 1$, and $p^t = x^{t+1} - x^t$.
- The subproblem optimum is $p^{t*} = x^* - x^t$, so $\|x^t - x^*\| = O(f(t))$ and $\|p^t - p^{t*}\| = O(f(t))$.
- f is arbitrary, both the subproblem inexact solution and its corresponding objective converge to the optimum arbitrarily fast, but $x^t \notin \mathcal{M}$ for all t

- Interestingly, our numerical experience in Lee and Wright (2019); Lee et al. (2019); Li et al. (2020) suggests the opposite: ISQA can identify the active manifold in practice
- This discrepancy between theory and practice motivates this work

Our Contributions

- Prove that ISQA essentially possesses the manifold identification property either through the subproblem solver or a specific solution accuracy requirement (2nd one skipped in this talk)
- Strong convergence of the iterates under a mild growth condition (skipped in this talk)
- Propose acceleration techniques to achieve superlinear convergence in running time even without local strong convexity
- Numerical result shows that our new algorithm ISQA$^+$ greatly improves upon existing PN and PQN methods

Outline

Algorithm Details

- Choice of H_t: bounded and PD

$$\exists M, m > 0, \quad \text{such that} \quad M \succeq H_t \succeq m, \ \forall t \geq 0. \tag{BD+PD}$$

- Inexact solution: consider

$$Q_t(p^t) - \min_p Q_t(p) \leq \epsilon_t, \tag{OBJ}$$

- Step size: given $\gamma \in (0, 1)$ find α_t such that

$$F(x^t + \alpha_t p^t) \leq F(x^t) + \alpha_t \gamma Q_t(p^t) \tag{Armijo}$$

Algorithmic Framework

Algorithm 1: Framework of ISQA

input : x^0, $\gamma, \beta \in (0,1)$

for $t = 0, 1, \ldots$ **do**

  $\alpha_t \leftarrow 1$, pick $\epsilon_t \geq 0$ and H_t, and solve (SUBPROB) for p^t satisfying (OBJ)

  while (Armijo) *not satisfied* **do** $\alpha_t \leftarrow \beta \alpha_t$

  $x^{t+1} \leftarrow x^t + \alpha_t p^t$

Definition 2 (Partly smooth)

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

1. Around x^*, $\mathcal{M}$ is a $\mathcal{C}^2$-manifold and $\Psi|_{\mathcal{M}}$ is $\mathcal{C}^2$.
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 $\mathcal{M}$.

Outline

Identification from Subproblem Solver I

- Consider relative accuracy in (OBJ) for easier analysis:

$$\exists \eta \in [0,1): \quad \epsilon_t = \eta \left(Q_t(0) - \min_p Q_t(p) \right) = -\eta \min_p Q_t(p), \quad \forall t. \qquad \text{(Relative)}$$

- Easily satisfied by applying a linear-convergent solver to (SUBPROB) for a fixed number of iterations
- Define the proximal mapping: for any function g, $\tau \geq 0$, and Λ PD,

$$\operatorname{prox}_{\tau g}^{\Lambda}(x) := \operatorname*{argmin}_y \frac{1}{2} \langle x - y, \Lambda(x - y) \rangle + \tau g(y)$$

- p^{t*} denotes the optimal solution to (SUBPROB) and $Q_t^* := Q_t(p^{t*})$

Identification from Subproblem Solver II

Theorem 3

Consider a point x^* satisfying

$$0 \in \operatorname{relint}\left(\partial F(x^*)\right) = \nabla f(x^*) + \operatorname{relint}\left(\partial \Psi(x^*)\right), \qquad \text{(Nondegenerate)}$$

with Ψ partly smooth at x^* relative to some manifold $\mathcal{M}$. Assume f is locally L-smooth for $L > 0$ around x^*. If Algorithm 1 is run with (OBJ) *and* (Relative) *for some $\eta \in [0, 1)$, and the update direction p^t satisfies*

$$x^t + p^t = \operatorname{prox}_{\Psi}^{\Lambda_t}\left(y^t - \Lambda_t^{-1}\left(\nabla f\left(x^t\right) + H_t\left(y^t - x^t\right) + s^t\right)\right), \qquad \text{(Prox)}$$

where s^t satisfies $\|s^t\| \leq R\left(\|y^t - (x^t + p^{t*})\|\right)$ for some continuous and increasing R with $R(0) = 0$, Λ_t is symmetric and PD, with $M_1 \geq \|\Lambda_t\|$ for $M_1 > 0$, and y^t satisfies

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

for some $\nu > 0$ and $\eta_1 \geq 0$, then there exists $\epsilon, \delta > 0$ such that $\|x^t - x^*\| \leq \epsilon$, $|Q_t^*| \leq \delta$, and $\alpha_t = 1$ imply $x^{t+1} \in \mathcal{M}$.

Examples of Solvers Fitting (Prox)

- Proximal Gradient (PG)
- Accelerated PG
- Prox-SAGA/SVRG
- Proximal (Cyclic) Coordinate Descent (CD)
- Almost all solvers used in practice satisfy (Prox), so ISQA essentially possesses the manifold identification property

Outline

Algorithm Flow

The proposed algorithm ISQA$^+$:

- ISQA stage:
  1. Solve (SUBPROB)
  2. If (Armijo) fails then modify H_t and resolve
  3. If x^t stays within the same manifold for T iterations: switch to the smooth stage
- Smooth stage:
  1. One iteration of Newton or quasi-Newton within the current manifold
  2. One iteration of PG
  3. If the manifold changes after PG or the smooth step fails to decrease the objective, go back to the ISQA stage

Superlinear Convergence of ISQA$^+$ Without Strong Convexity

- Use $\phi_t : \mathbb{R}^m \to \mathcal{M}_{x^t} \in \mathbb{R}^n$ with $\phi_t(y^t) = x^t$ to parameterize the current manifold, then $F_{\phi_t} := F(\phi_t(\cdot))$ is smooth
- Apply a damping term to the Hessian: find q^t the update direction for y^t such that

$$H_t q^t \approx -g^t,\ g^t := \nabla F(\phi_t(y^t)),\ H_t = \nabla^2 F\left(\phi_t(y^t)\right) + \mu_t I,\ \mu_t := c\left\|g^t\right\|^\rho \quad \text{(Newton)}$$

satisfying

$$\left\|H_t q^t + g^t\right\| \leq 0.1 \min\left\{\left\|g^t\right\|, \left\|g^t\right\|^{1+\rho}\right\} \quad \text{(Tolerance)}$$

with pre-specified $c > 0$ and $\rho \in (0, 1]$.

- Apply (preconditioned) conjugate gradient to solve the problem
- Backtracking along q^t for F_{ϕ_t}

Superlinear Convergence

Theorem 4

Consider a critical point x^* *of* (REG) *satisfying* (Nondegenerate) *at which* Ψ *is partly smooth relative to* $\mathcal{M}$ *with a parameterization* ϕ *and* y^* *such that* $\phi(y^*) = x^*$. *Assume* $\nabla^2 F_\phi$ *is PSD and Lipschitz continuous within a neighborhood* U *of* y^*, Ψ *is convex, proper, closed,* f *is* L*-smooth. Then there is a neighborhood* V *of* x^* *such that if at the* t_0*th iteration of* ISQA$^+$ *for some* $t_0 > 0$ $x^{t_0} \in V$, *we have entered the smooth stage,* $\mathcal{M}$ *is correctly identified, and* $\alpha_t = 1$ *is taken in the Newton steps for all* $t \geq t_0$, *we get the following for all* $t \geq t_0$.

1. *For* $\rho \in (0, 1]$ *in* (Newton) *and* (Tolerance) *and* F_ϕ *satisfying* $\zeta^{\hat\theta}\|y - y^*\| \leq (F_\phi(y) - F(y^*))^{\hat\theta}, \quad \forall y \in U$, *with* $\hat\theta = 1/2$ *for some* $\zeta > 0$:

$$\left\|x^{t+2} - x^*\right\| = O\left(\left\|x^t - x^*\right\|^{1+\rho}\right), \left\|\nabla F_\phi\left(x^{t+2}\right)\right\| = O\left(\left\|\nabla F_\phi\left(x^t\right)\right\|^{1+\rho}\right).$$

2. *For* $\rho = 0.69$ *and* F_ϕ *satisfying the same sharpness condition for some* $\zeta > 0$ *and* $\hat\theta \geq 3/8$,

$$\left\|x^{t+2} - x^*\right\| = o\left(\left\|x^t - x^*\right\|\right).$$

Outline

Experiment Setting

- ℓ_1-regularized logistic regression: domain $\mathbb{R}^d$,

$$\Psi(x) = \lambda \|x\|_1, \; f(x) = \sum_{i=1}^{n} \log\left(1 + \exp\left(-b_i \langle a_i, x \rangle\right)\right),$$

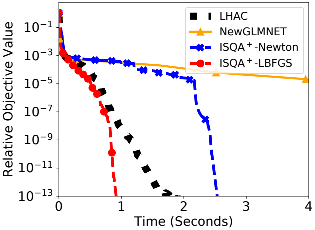
  ($\lambda = 1$ in the experiments)

- Algorithms to compare:
  - LHAC (Scheinberg and Tang, 2016): an inexact proximal L-BFGS method with CD for (SUBPROB) and a trust-region-like approach.
  - NewGLMNET (Yuan et al., 2012): a line-search PN with a CD subproblem solver.
  - ISQA$^+$-LBFGS and ISQA$^+$-Newton: our algorithm with the first stage using L-BFGS and real Hessian for H_t, respectively

Results

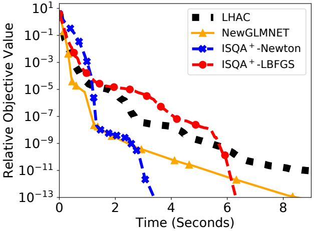
a9a

$n = 32,561, d = 123$



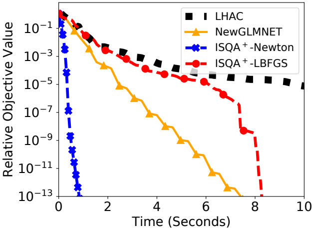
realsim

$n = 72,309, d = 20,958$



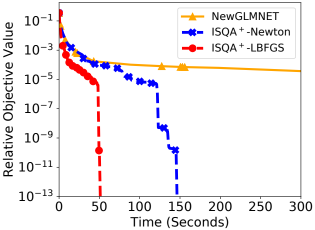
news20

$n = 19,996, d = 1,355,191$



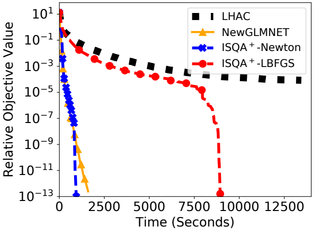
covtype.scale

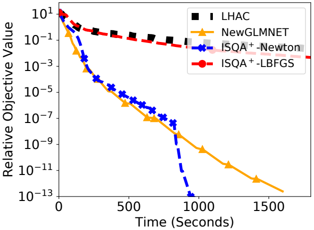
$n = 581,012, d = 54$



webspam / webspam (finer scale)

$n = 350,000, d = 16,609,143$





Experiment Results

- No clear winner among PN and PQN: depending on data
- But our acceleration improves individual performance no matter which one is better
- Although PN and PQN have superlinear convergence in terms of outer iterations, not observed in running time
- Superlinear convergence in running time clearly observed in our accelerated algorithms

Paper available at: Ching-pei Lee. Accelerating inexact successive quadratic approximation for regularized optimization through manifold identification, 2020. arXiv:2012.02522

Implementation for the experiment at: `https://github.com/leepei/ISQA_plus`

References I

Ching-pei Lee. Accelerating inexact successive quadratic approximation for regularized optimization through manifold identification, 2020. arXiv:2012.02522.

Ching-pei Lee and Stephen J. Wright. Inexact successive quadratic approximation for regularized optimization. *Computational Optimization and Applications*, 2019.

Ching-pei Lee, Cong Han Lim, and Stephen J. Wright. A distributed quasi-Newton algorithm for primal and dual regularized empirical risk minimization, 2019. arXiv:1912.06508.

Yu-Sheng Li, Wei-Lin Chiang, and Ching-pei Lee. Manifold identification for ultimately communication-efficient distributed optimization. In *Proceedings of the International Conference on Machine Learning*, 2020.

Katya Scheinberg and Xiaocheng Tang. Practical inexact proximal quasi-Newton method with global complexity analysis. *Mathematical Programming*, 160(1-2):495–529, 2016.

References II

Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. An improved GLMNET for $L1$-regularized logistic regression. *Journal of Machine Learning Research*, 13: 1999–2030, 2012.