Generalized Newton Algorithms for Nonsmooth Systems with Applications to Lasso Problems

Boris Mordukhovich

aa1086@wayne.edu

Department of Mathematics

WAYNE STATE UNIVERSITY

talk given at **One World Optimization Seminar**

based on joint work with Pham Duy Khanh (HCMUE, Vietnam), Vo Thanh Phat (WSU), M. E. Sarabi (Miami Univ.) and Dat Ba Tran (WSU)

Supported by NSF and Air Force grants

February 1, 2021

CLASSICAL NEWTON METHOD

Let $\varphi\colon \boldsymbol{R}^n \to \boldsymbol{R}$ be $\mathcal{C}^2$-smooth around $\bar{x}$. The classical Newton method to solve the nonlinear gradient system $\nabla\varphi(x) = 0$ and optimization problems constructs the iterative procedure

$$x^{k+1} := x^k + d^k \ \text{ for all } k \in \boldsymbol{N} := \{1, 2, \ldots\}$$

where x^0 is a given starting point and where d^k is a solution to the linear system

$$-\nabla\varphi(x^k) = \nabla^2\varphi(x^k)d^k, \quad k = 0, 1, \ldots$$

The classical Newton algorithm is well-defined (solvable for d^k) and the sequence of its iterates $\{x^k\}$ superlinearly (even quadratically) converges to a solution $\bar{x}$ if x^0 is chosen sufficiently close to $\bar{x}$ and the Hessian $\nabla^2\varphi(\bar{x})$ is positive-definite

The are many nonsmooth extensions; see, e.g., the books by Facchinei and Pang [FP03], Izmailov and Solodov [IS14], and Klatte and Kummer [KK02]

In order to derive global convergence of Newton method, a common way is to use a line search strategy and update the sequence $\{x^k\}$ by

$$x^{k+1} := x^k + \tau_k d^k \quad \text{for all } k \in \mathbb{N} := \{1, 2, \dots\}$$

where τ_k is chosen by the Armijo rule, i.e.

$$\varphi(x^{k+1}) \leq \varphi(x^k) + \sigma\tau_k \langle \nabla\varphi(x^k), d^k \rangle$$

where $\sigma \in (0, 1/2)$. The resulting algorithm using Newton directions with the backtracking line search is known the damped Newton method

In this talk we report recent results on the following topics:

- Design and justification of locally convergent generalized Newton algorithms with superlinear convergence rates to find tilt-stable local minimizers for $\mathcal{C}^{1,1}$ optimization problems that are based on second-order subdifferentials and also on subgradient graphical derivatives
- Design and justification of such generalized Newton algorithms for minimization of extended-real-valued prox-regular functions that cover problems of constrained optimization
- Design and justification of superlinearly locally convergent algorithms to solve subgradient systems $0 \in \partial\varphi(x)$ associated with extended-real-valued prox-regular functions

MAJOR GOALS

• Design and justification of globally convergent algorithms of damped Newton type based on second-order subdifferentials to solve $\mathcal{C}^{1,1}$ optimization problems

• Design and justification of globally convergent algorithms of damped Newton type to solve convex composite optimization problems in the unconstrained form

$$\text{minimize} \quad \varphi(x) := f(x) + g(x)$$

where f is a convex quadratic function, and g is a lower semicontinuous convex function which may be extended-real-valued

• Apply the obtained results to a major class of Lasso problems

• Conduct numerical implementations and comparison with some first-order and second-order algorithms to solve the basic Lasso problem

See [M06,M18,Rock.-Wets98] for more details

Normal cone to $\Omega \subset \boldsymbol{R}^n$ at $\bar{x} \in \Omega$ is

$$N_\Omega(\bar{x}) := \Big\{ v \in \boldsymbol{R}^n \Big|\ \exists x_k \overset{\Omega}{\to} \bar{x},\ v_k \to v,\ \limsup_{x \overset{\Omega}{\to} x_k} \frac{\langle v_k, x - x_k \rangle}{\|x - x_k\|} \le 0 \Big\}$$

where $x \overset{\Omega}{\to} \bar{x}$ means that $x \to \bar{x}$ and $x \in \Omega$
Coderivative of $F \colon \boldsymbol{R}^n \rightrightarrows \boldsymbol{R}^m$ at $(\bar{x}, \bar{y}) \in \operatorname{gph} F$ is

$$D^*F(\bar{x}, \bar{y})(v) := \Big\{ u \in \boldsymbol{R}^n \Big| (u, -v) \in N_{\operatorname{gph} F}(\bar{x}, \bar{y}) \Big\},\ v \in \boldsymbol{R}^m$$

Subdifferential of $\varphi \colon \boldsymbol{R}^n \to \overline{\boldsymbol{R}} := (-\infty, \infty]$ at $\bar{x} \in \operatorname{dom} \varphi$ is

$$\partial\varphi(\bar{x}) := \Big\{ v \in \boldsymbol{R}^n \Big|\ (v, -1) \in N_{\operatorname{epi} \varphi}\big(\bar{x}, \varphi(\bar{x})\big) \Big\}$$

Second-order subdifferential/generalized Hessian [M92] of φ at $\bar{x}$ relative to $\bar{v} \in \partial\varphi(\bar{x})$ is

$$\partial^2\varphi(\bar{x},\bar{x})(u) := \big(D^*\partial\varphi\big)(\bar{x},\bar{v})(u), \quad u \in \boldsymbol{R}^n$$

If $\varphi \in \mathcal{C}^2$-smooth around $\bar{x}$, then

$$\partial^2\varphi(\bar{x},\bar{v})(u) = \big\{\nabla^2\varphi(\bar{x})u\big\}, \quad u \in \boldsymbol{R}^n$$

In general $\partial^2\varphi(\bar{x},\bar{v})(u)$ enjoys full calculus and is computed in terms of the given data for large classes of structural functions that appear in variational analysis, optimization, and control theory; see the publications by Colombo, Ding, Dontchev, Henrion, Hoang, Huy, Mordukhovich, Nam, Outrata, Poliquin, Qui, Rockafellar, Römisch, Sarabi, Son, Sun, Surowiec, Yao, Ye, Yen, Zhang, etc.

Definition [Poliquin-Rock96, Rock-Wets98]

A mapping $\varphi\colon \boldsymbol{R}^n \to \overline{\boldsymbol{R}}$ is prox-regular at $\bar{x} \in \operatorname{dom}\varphi$ for $\bar{v} \in \partial\varphi(\bar{x})$ if φ is lower semicontinuous and there are $\varepsilon > 0$ and $\rho \geq 0$ such that for all $x \in \boldsymbol{B}_\varepsilon(\bar{x})$ with $\varphi(x) \leq \varphi(\bar{x}) + \varepsilon$ we have

$$\varphi(x) \geq \varphi(u) + \langle \bar{v}, x - u\rangle - \frac{\rho}{2}\|x - u\|^2 \ \forall\, (u, v) \in (\operatorname{gph}\partial\varphi) \cap \boldsymbol{B}_\varepsilon(\bar{x}, \bar{v})$$

φ is subdifferentially continuous at $\bar{x}$ for $\bar{v}$ if the convergence $(x_k, v_k) \to (\bar{x}, \bar{v})$ with $v_k \in \partial\varphi(x_k)$ yields $\varphi(x_k) \to \varphi(\bar{x})$. If both properties hold, φ is continuously prox-regular. This is the major class in second-order variational analysis

Definition (Poliquin and Rockafellar, 1998)

Given $\varphi\colon \mathbb{R}^n \to \overline{\mathbb{R}}$, a point $\bar{x} \in \operatorname{dom}\varphi$ is said to be a tilt-stable local minimizer of φ if for some $\gamma > 0$ the argminimum mapping

$$M_\gamma\colon v \mapsto \operatorname{argmin}\Big\{\varphi(x) - \langle v, x\rangle \Big| \, x \in \mathbb{B}_\gamma(\bar{x})\Big\}$$

is single-valued and Lipschitz continuous on a neighborhood of $\bar{v} = 0$ with $M_\gamma(\bar{v}) = \{\bar{x}\}$

This notion is very well investigated and comprehensively characterized in second-order variational analysis with many applications to constrained optimization. In particular, tilt-stable local minimizers of prox-regular functions $\varphi\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ are characterized via second-order subdifferential by [Poliquin-Rock98]

$$\partial^2\varphi(\bar{x}, 0) > 0$$

There are other characterizations of tilt-stable minimizers for broad classes of structural problems in constrained optimization and optimal control. We refer to publications by Benko, Bonnans, Chieu, Drusvyatskiy, Eberhard, Gfrerer, Hien, Lewis, Mordukhovich, Ng, Nghia, Outrata, Poliquin, Qui, Rockafellar, Sarabi, Shapiro, Wachsmuth, Zhang, Zheng, Zhu, etc.

Algorithm 1 (to find tilt-stable local minimizers) [M.-Sarabi20]

Step 0: Choose a starting point x^0 and set $k = 0$

Step 1: If $\nabla\varphi(x^k) = 0$, stop the algorithm. Otherwise move to Step 2

Step 2: Choose $d^k \in \boldsymbol{R}^n$ satisfying

$$-\nabla\varphi(x^k) \in \partial^2\varphi(x^k)(d^k) = \partial\langle d^k, \nabla\varphi\rangle(x^k)$$

Step 3: Set x^{k+1} given by

$$x^{k+1} := x^k + d^k, \quad k = 0, 1, \dots$$

Step 4: Increase k by 1 and go to Step 1

Definition (Gfrerer and Outrata, 2019)

A mapping $F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is semismooth* at $(\bar{x},\bar{y}) \in \operatorname{gph} F$ if whenever $(u,v) \in \mathbb{R}^n \times \mathbb{R}^m$ we have the condition

$$\langle u^*, u\rangle = \langle v^*, v\rangle \;\text{ for all }\; (v^*, u^*) \in \operatorname{gph} D^*F\big((\bar{x},\bar{y});(u,v)\big)$$

Theorem [M.-Sarabi20]

Let φ be a $\mathcal{C}^{1,1}$ function on a neighborhood of its tilt-stable local minimizer $\bar{x}$. Then Algorithm 1 is well-defined around $\bar{x}$. If gradient mapping $\nabla\varphi$ is semismooth* at $\bar{x}$, then there exist $\delta > 0$ such that for any starting point $x_0 \in \mathbb{B}_\delta(\bar{x})$ every sequence $\{x_k\}$ constructed by Algorithm 1 converges to $\bar{x}$ and the rate of convergence is superlinear

Consider the set

$$Q(x) := \left\{ y \in \mathbb{R}^n \,\middle|\, -\nabla\varphi(x) \in (D\nabla\varphi)(x)(y) \right\}$$

Algorithm 2 [M.-Sarabi20]

Pick $x_0 \in \mathbb{R}^n$ and set $k := 0$
Step 1: If $\nabla\varphi(x_k) = 0$, then stop
Step 2: Otherwise, select a direction $d_k \in Q(x_k)$ and set $x_{k+1} := x_k - d_k$
Step 3: Let $k \leftarrow k + 1$ and then go to Step 1

Theorem

Let φ be a $\mathcal{C}^{1,1}$ function on a neighborhood of $\bar{x}$, which is a tilt-stable local minimizer φ. Then there exists a neighborhood O of $\bar{x}$ such that the set-valued mapping $Q(x)$ is nonempty and compact-valued for all x in O

The second subderivative [Rock.88] of $\varphi\colon \boldsymbol{R}^n \to \overline{\boldsymbol{R}}$ at $\bar{x}$ for $\bar{v}$ is

$$\mathrm{d}^2\varphi(\bar{x},\bar{v})(w) := \liminf_{t\downarrow 0,\, w'\to w} \Delta_t^2\varphi(\bar{x},\bar{v})(w')$$

where the second-order finite difference are

$$\Delta_t^2\varphi(\bar{x},\bar{v})(w) := \frac{\varphi(\bar{x}+tw') - \varphi(\bar{x}) - t\langle \bar{v}, w'\rangle}{\frac{1}{2}t^2}$$

φ is twice epi-differentiable at $\bar{x}$ for $\bar{v}$ if for every $w \in \boldsymbol{R}^n$ and $t_k \downarrow 0$ there is $w_k \to w$ with $\Delta_{t_k}^2\varphi(\bar{x},\bar{v})(w_k) \to \mathrm{d}^2\varphi(\bar{x},\bar{v})(w)$.

The latter class includes fully amenable functions [Rock.-Wets98], parabolically regular functions [Mohammadi-M.-Sarabi21], etc.

Subproblems for directions: At each iteration x^k with $v^k := -\nabla\varphi(x^k)$ find $w = D^k$ as a stationary point of

$$\min \varphi(x^k) + \langle v^k, w\rangle + \frac{1}{2}\mathrm{d}^2\varphi(x^k, v^k)(w)$$

Constructive implementations of **subproblems** are given, in particular, for the classes of extended linear-quadratic programs and for minimization of augmented Lagrangians.

Theorem [M.-Sarabi20]

Let $\varphi\colon \boldsymbol{R}^n \to \overline{\boldsymbol{R}}$ be a $\mathcal{C}^{1,1}$ function around $\bar{x}$, where $\bar{x}$ is its tilt-stable local minimizer, and let φ be twice epi-differentiable at x for $v = \nabla\varphi(x)$. Then for each large $k \in \boldsymbol{N}$ the subproblem admits a unique optimal solution

Theorem [M.-Sarabi20]

Let $\varphi\colon \boldsymbol{R}^n \to \overline{\boldsymbol{R}}$ be a $\mathcal{C}^{1,1}$ function on a neighborhood of its tilt-stable local minimizer $\bar{x}$, and let $\nabla\varphi$ be semismooth* at $\bar{x}$. Then there exists $\delta > 0$ such that for any starting point $x_0 \in \boldsymbol{B}_\delta(\bar{x})$ we have that every sequence $\{x_k\}$ constructed by Algorithm 2 converges to $\bar{x}$ and the rate of convergence is superlinear

Recall that Moreau envelope of $\varphi\colon \mathbb{R}^n \to \overline{\mathbb{R}}$

$$e_r\varphi(x) := \inf_w \left\{\varphi(w) + \frac{1}{2r}\|w - x\|^2\right\}, \quad r > 0$$

and the result from [Rock.-Wets88] that if φ is continuously prox-regular at $\bar{x}$ for $\bar{v}$, then its Moreau envelope for small $r > 0$ is a $\mathcal{C}^{1,1}$ function with $\nabla e_r\varphi(\bar{x} + r\bar{v}) = \bar{v}$.

Consider the unconstrained problem

$$\text{minimize} \;\; e_r\varphi(x) \;\; \text{subject to} \;\; x \in \mathbb{R}^n$$

Theorem [M.-Sarabi20]

Let $\varphi\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ be continuously prox-regular at $\bar{x}$ for $\bar{v} = 0$, where $\bar{x}$ is a tilt-stable local minimizer of φ. If $\partial\varphi$ is semismooth* at $(\bar{x}, \bar{v})$, then for any small $r > 0$ there exists $\delta > 0$ such that for each starting point $x_0 \in \mathbb{B}_\delta(\bar{x})$ both Algorithms 1 and 2 and are well-defined, and every sequence of iterates $\{x_k\}$ superlinearly converges to $\bar{x}$

Consider the constrained problem

$$\text{minimize} \quad \psi(x) \quad \text{subject to} \quad f(x) \in \Theta$$

where the functions $\psi\colon \mathbb{R}^n \to \mathbb{R}$ and $f\colon \mathbb{R}^n \to \mathbb{R}^m$ are $\mathcal{C}^2$-smooth and the set $\Theta \subset \mathbb{R}^m$ is closed and convex. Denote

$$\varphi(x) := \psi(x) + \delta_\Omega(x) \quad \text{with} \quad \Omega := \left\{x \in \mathbb{R}^n \;\middle|\; f(x) \in \Theta\right\}$$

Algorithm 3 [M.-Sarabi20]

Set $k := 0$, and pick any $r > 0$

Step 1: If $0 \in \partial\varphi(x_k)$, then stop.

Step 2: Otherwise, let $v_k = \nabla(e_r\varphi)(x_k)$, select w_k as a stationary point of the subproblem

$$\min_{w \in \mathbf{R}^n} \langle v_k, w \rangle + \frac{1}{2}\mathrm{d}^2\varphi(x_k - rv_k, v_k)(w)$$

and then set $d_k := w_k - rv_k$, $x_{k+1} := x_k + d_k$

Step 3: Let $k \leftarrow k + 1$ and then go to Step 1

In addition to the conditions of the previous type, the metric subregularity of $x \mapsto f(x) - \Theta$ is needed for superlinear convergence of Algorithm 3

The above locally convergent generalized Newton algorithms based in 2nd-order subdifferential are extended in [Khanh-M.-Phat20] to solve the subgradient inclusions

$$0 \in \partial\varphi(x) \text{ where } \varphi\colon \boldsymbol{R}^n \to \overline{\boldsymbol{R}}$$

with the usage of the proximal mapping

$$\mathrm{Prox}_\lambda\varphi(x) := \mathrm{argmin}\left\{\varphi(y) + \frac{1}{2\lambda}\|y-x\|^2 \;\middle|\; y \in \boldsymbol{R}^n\right\}$$

for prox-regular functions. Here is the main algorithm developed in [Khanh-M.-Phat20]

ALgorithm 4

Step 0: Pick any $\lambda \in (0, r^{-1})$, set $k := 0$, choose a starting point x^0 by

$$x^0 \in U_\lambda := \text{rge}(I + \lambda\partial\varphi).$$

Step 1: If $0 \in \partial\varphi(x^k)$, then stop. Otherwise compute

$$v^k := \frac{1}{\lambda}\left(x^k - \text{Prox}_\lambda\varphi(x^k)\right)$$

Step 2: Choose $d^k \in \boldsymbol{R}^n$ such that

$$-v^k \in \partial^2\varphi(x^k - \lambda v^k, v^k)(\lambda v^k + d^k)$$

Step 3: Compute x^{k+1} by

$$x^{k+1} := x^k + d^k.$$

Then increase k by 1 and go to Step 1

General conditions for well-posedness of Algorithm 4 are given in [Khanh-M.-Phat20]

Theorem [Khanh-M.-Phat20]

Let $\varphi\colon \boldsymbol{R}^n \to \overline{\boldsymbol{R}}$ be bounded from below by a quadratic function and continuously prox-regular at $\bar{x}$ for $0 \in \partial\varphi(\bar{x})$ with parameters $r > 0$. Assume that $\partial\varphi$ is semismooth* and metrically regular around $(\bar{x}, 0)$. Then there exists a neighborhood U of $\bar{x}$ such that for all starting points $x^0 \in U$ Algorithm 4 generates a sequence of iterates $\{x^k\}$, which converges superlinearly to the solution $\bar{x}$ of the subgradient inclusion $0 \in \partial\varphi(x)$

Applications to solving a Lasso problem are obtained in [Khanh-M.-Phat20]

Algorithm 5 [Khanh-M.-Phat-Tran21]

Step 0: Choose $\sigma \in (0, 1/2), \beta \in (0, 1)$, a starting point x^0 and set $k = 0$

Step 1: If $\nabla\varphi(x^k) = 0$, stop the algorithm. Otherwise move to Step 2

Step 2: Choose $d^k \in \boldsymbol{R}^n$ satisfying

$$-\nabla\varphi(x^k) \in \partial\langle d^k, \nabla\varphi\rangle(x^k)$$

Step 3: Set $\tau_k = 1$. If

$$\varphi(x^k + \tau_k d^k) > \varphi(x^k) + \sigma\tau_k\langle\nabla\varphi(x^k), d^k\rangle$$

then set $\tau_k := \beta\tau_k$.

Step 4: Set x^{k+1} given by

$$x^{k+1} := x^k + \tau_k d^k, \quad k = 0, 1, \ldots$$

Step 5: Increase k by 1 and go to Step 1

Theorem [Khanh-M.-Phat-Tran21]

Let $\varphi : \boldsymbol{R}^n \to \boldsymbol{R}$ be a $\mathcal{C}^{1,1}$ function on $\boldsymbol{R}^n$, and let $x^0 \in \boldsymbol{R}^n$. Denote

$$\Omega := \left\{ x \in \boldsymbol{R}^n \middle| \; \varphi(x) \leq \varphi(x^0) \right\}$$

Suppose that Ω is bounded and that $\partial^2\varphi(x)$ is positive-definite for all $x \in \Omega$. Then the sequence $\{x^k\}$ constructed by Algorithm 5 globally R-linearly converges to $\bar{x}$, which is a tilt-stable local minimizer of φ with some modulus $\kappa > 0$. The rate of the global convergence is at least Q-superlinear if either one of two following conditions holds:

(i) $\nabla\varphi$ is semismooth* at $\bar{x}$ and $\sigma \in (0, 1/(2\ell\kappa))$where $\ell > 0$ is a Lipschitz constant of φ around $\bar{x}$

(ii) $\nabla\varphi$ is semismooth at $\bar{x}$

Consider the following composite optimization problem

$$\text{minimize} \quad \varphi(x) := f(x) + g(x), \quad x \in \mathbb{R}^n,$$

where g is an extended-real-valued lower semicontinuous convex function, and where f is quadratic convex function given by

$$f(x) := \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + \alpha$$

with $A \in \mathbb{R}^{n\times n}$ being positive semidefinite, $b \in \mathbb{R}^n$, and $\alpha \in \mathbb{R}$

Algorithm 6 [Khanh-M.-Phat-Tran21]

Step 0: Choose $\gamma > 0$ such that $I - \gamma A$ is positive definite, calculate $Q := (I - \gamma A)^{-1}$, $c := \gamma Q b$, $P := Q - I$, and define

$$\psi(y) := \frac{1}{2}\langle Py, y\rangle + \langle c, y\rangle + \gamma e_\gamma g(y)$$

Then choose an arbitrary starting point $y^0 \in \boldsymbol{R}^n$ and set $k := 0$

Step 1: If $\nabla\psi(y^k) = 0$, then stop. Otherwise compute

$$v^k := \mathrm{Prox}_\gamma g(y^k)$$

Step 2: Choose $d^k \in \boldsymbol{R}^n$ such that

$$\frac{1}{\gamma}(-\nabla\psi(y^k) - Pd^k) \in \partial^2 g\left(v^k, \frac{1}{\gamma}(y^k - v^k)\right) Qd^k + \nabla\psi(y^k))$$

Step 3: (line search) Set $\tau_k = 1$. If

$$\psi(y^k + \tau_k d^k) > \psi(y^k) + \sigma\tau_k\langle\nabla\psi(y^k), d^k\rangle$$

then set $\tau_k := \beta\tau_k$

Step 4: Compute y^{k+1} by

$$y^{k+1} := y^k + \tau_k d^k, \quad k = 0, 1, \ldots$$

Step 5: Increase k by 1 and go to Step 1

Theorem [Khanh-M.-Phat-Tran21]

Suppose that A is positive-definite. Then we have

(i) Algorithm 6 is well-defined and the sequence of its iterates $\{y^k\}$ globally converges at least R-linearly to $\bar{y}$.

(ii) $\bar{x} := Q\bar{y} + c$ is a tilt-stable local minimizer of φ, and it is the unique solutionof this problem

The rate of convergence of $\{y^k\}$ is at least Q-superlinear if either one of two following conditions holds:

(a) ∂g is semismooth* on $\boldsymbol{R}^n$ and $\sigma \in (0, 1/(2\ell\kappa))$, where $\ell := \max\{1, \|Q\|\}$ and $\kappa := \frac{1}{\lambda_{\min}(P)}$

(b) g is twice epi-differentiable and the subgradient mapping ∂g is semismooth* on $\boldsymbol{R}^n$

The basic version of this problem, known also as the ℓ^1-regularized least square optimization problem, is formulated in [Tibshirani96] as follows

$$\text{minimize} \quad \varphi(x) := \frac{1}{2}\|Ax - b\|_2^2 + \mu\|x\|_1, \quad x \in \mathbb{R}^n$$

where A is an $m \times n$ matrix, $\mu > 0$, $b \in \mathbb{R}^m$ with the standard norms $\|\cdot\|_1$ and $\|\cdot\|_2$. This problem is of the convex composite optimization type with

$$f(x) = \frac{1}{2}\|Ax - b\|^2 \quad \text{and} \quad g(x) = \mu\|x\|_1$$

In [Khanh-M.-Phat-Tran21] we compute ∂g, $\partial^2 g$, $\mathrm{Prox}_\gamma g(x)$ entirely via the problem data and then run Algorithm 6 with providing numerical experiments

The numerical experiments to solve the Lasso problem using the generalized damped Newton algorithm (Algorithm 6), abbreviated as GDNM, are conducted in [Khanh-M.-Phat-Tran21] on a desktop with 10th Gen Intel(R) Core(TM) i5-10400 processor (6-Core, 12M Cache, 2.9GHz to 4.3GHz) and 16GB memory. All the codes are written in MATLAB 2016a. The data sets are collected from large scale regression problems taken from UCI data repository [Lichman]. The results are compared with the following

(i) Second-order method: the highly efficient semismooth Newton augmented Lagrangian method (SSNAL) from [Li-Sun-Toh18]

(ii) First-order methods:

- alternating direction methods of multipliers (ADMM) [Boyd et al., 2010]
- accelerated proximal gradient (APG) [Nesterov83]
- fast iterative shrinkage-thresholing algorithm (FISTA) [Beck-Teboulle09]

TESTING DATA [Lichman]

Test ID	Name	m	n
1	UCI-Relative location of CT slices on axial axis Data Set	53500	385
2	UCI-YearPredictionMSD	515345	90
3	UCI-Abalone	4177	6
4	Random	1024	1024
5	Random	4096	4096
6	Random	16384	16384

Test ID		1	2	3
m		53500	515345	4177
n		385	90	6
GDNM	Iter	2	3	4
	Time	0.59	0.41	0.03
	Value	1803564.0809648	82000054.9300050	10555.6018267
SSNAL	Iter	18	10	13
	Time	49.1	5.98	0.15
	Value	1803574.3685158	82000054.9300060	10555.6023802

Figure 1: GDNM with SSNAL - UCI tests

Test ID		4	5	6
m		1024	4096	16384
n		1024	4096	16384
GDNM	Iter	20	40	60
	Time	1.17	68.83	4097.58
	Value	0.6676880	1.4094035	5.5652481
SSNAL	Iter	21	55	39
	Time	6.24	660.72	30100.90
	Value	0.6856949	1.4106609	169.1900000

Figure 2: GDNM with SSNAL - random tests

Test ID		1	2	3
m		53500	515345	4177
n		385	90	6
GDNM	Iter	2	3	4
	Time	0.59	0.41	0.03
	Value	1803564.0809648	82000054.9300050	10555.6018267
ADMM	Iter	110	20	100
	Time	1.26	3.49	0.01
	Value	1803564.1578868	82000054.9300050	10555.6018267
APG	Iter	80000	10000	15
	Time	1381.72	539.18	0.03
	Value	1815607.3011418	82229781.4255074	10555.6018267
FISTA	Iter	8000	8000	2000
	Time	874.57	2163.45	0.03
	Value	1808977.9771392	82005831.1179086	10555.6018267

Figure 3: GDNM with first order methods - UCI tests

Test ID		1	2	3
m		53500	515345	4177
n		385	90	6
GDNM	Iter	2	3	4
	Time	0.59	0.41	0.03
	Value	1803564.0809648	82000054.9300050	10555.6018267
ADMM	Iter	110	20	100
	Time	1.26	3.49	0.01
	Value	1803564.1578868	82000054.9300050	10555.6018267
APG	Iter	80000	10000	15
	Time	1381.72	539.18	0.03
	Value	1815607.3011418	82229781.4255074	10555.6018267
FISTA	Iter	8000	8000	2000
	Time	874.57	2163.45	0.03
	Value	1808977.9771392	82005831.1179086	10555.6018267

Figure 4: GDNM with first order methods - random tests

Results on Test 1

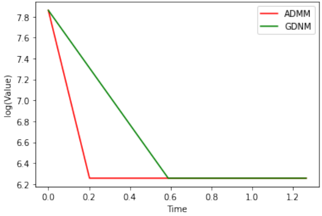

Figure 5: GDNM and ADMM

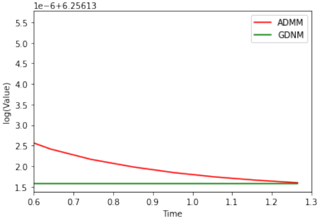

Figure 6: GDNM, ADMM from 0.6s

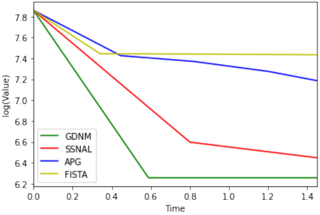

Figure 7: GDMN, SSNAL, APG, FISTA

Results on Test 2

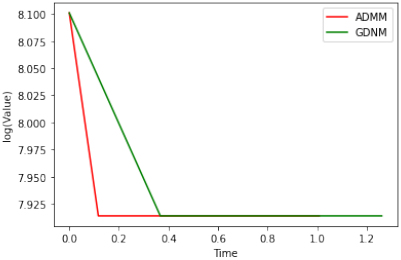

Figure 8: GDNM and ADMM

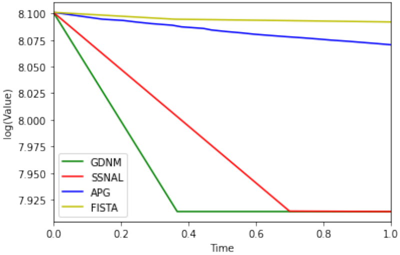

Figure 9: GDNM, SSNAL, APG, FISTA

Results on Test 4

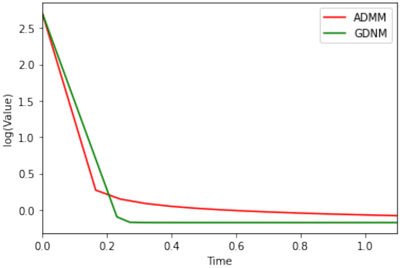

Figure 10: GDNM and ADMM

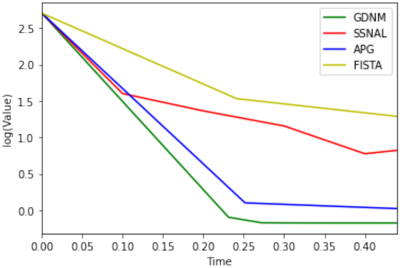

Figure 11: GDNM, SSNAL, APG, FISTA

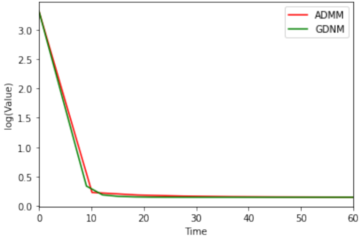

Figure 12: GDNM and ADMM

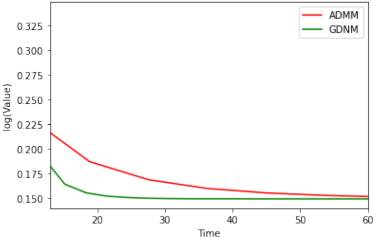

Figure 13: GDNM, ADMM 13s

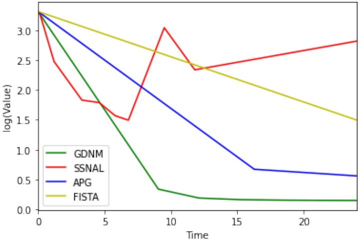

Figure 14: GDMN, SSNAL, APG, FISTA

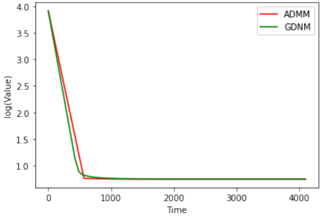

Figure 15: GDNM and ADMM

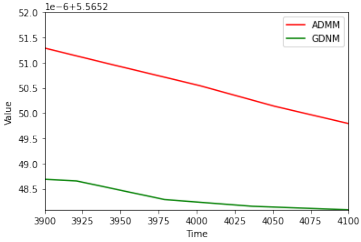

Figure 16: GDNM, ADMM from 3900s

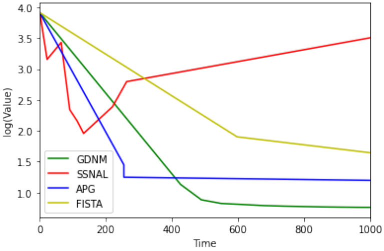

Figure 17: GDMN, SSNAL, APG, FISTA

REFERENCES

[BT09] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2, 183–202 (2009)

[Boyd10] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learning, 3, 1–122 (2010)

[FP03] F. Facchinei and J.-C. Pang, Finite-Dimensional Variational Inequalities and Complementarity Problems, Springer (2003)

[GO20] H. Gfrerer and J. V. Outrata, On a semismoothness* Newton method for solving generalized equations,to appear in SIAM J. Optim., arXiv:1904.09167 (2019)

[IS14] A. F. Izmailov and M. V. Solodov, Newton-Type Methods for Optimization and Variational Problems, Springer (2014)

[KMP20] P. D. Khanh, B. S. Mordukhovich and V. T. Phat, A generalized Newton method for subgradient systems, submitted; arXiv:2009.10551 (2020)

[KMPD21] P. D. Khanh, B. S. Mordukhovich, V. T. Phat and D. B. Tran Generalized damped Newton algorithms in nonsmooth optimization problems with applications to Lasso problems, submitted; arXiv:2101.10555 (2021)

[KK02] D. Klatte and B. Kummer, Nonsmooth Equations in Optimization. Regularity, Calculus, Methods and Applications, Kluwer (2002)

[Lichman] M. Lichman, UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets.html

[LSK18] X. Li, D. Sun and K.-C. Toh, A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems, SIAM J. Optim. 28, 433–458 (2018)

[MMS20] A. Mohammadi, B. S. Mordukhovich and M. E. Sarabi, Parabolic regularity in geometric variational analysis, Trans. Amer. Math. Soc. 374, 1711–1763 (2021)

[M18] B. S. Mordukhovich, Variational Analysis and Applications, Springer (2018)

[MS20] B. S. Mordukhovich and M. E. Sarabi, Generalized Newton algorithms for tilt-stable minimizers in nonsmooth optimization, to appear in SIAM J. Optim.; arXiv:1909.00241 (2020)

[N83] Y. E. Nesterov, A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$, Soviet Math. Dokl. 27, 372–376 (1983)

[PR96] R. A. Poliquin and R. T. Rockafellar, Prox-regular functions in variational analysis, Trans. Amer. Math. Soc. 348, 1805–1838 (1996)

[PR98] R. A. Poliquin and R. T. Rockafellar, Tilt stability of a local minimum, SIAM J. Optim. 8, 287–299 (1998)

[QS93] L. Qi and J. Sun, A nonsmooth version of Newton's method, Math. Program. 58. 353–367 (1993)

[RW98] R. T. Rockafellar and R. J-B. Wets, Variational Analysis, Springer (1998)

[T96] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. 58, 267–288 (1996)