First-Order Methods for Nonsmooth Nonconvex Functional Constrained Optimization with or without Slater Points

Zhichao Jia

School of Industrial and Systems Engineering
Georgia Institute of Technology
Joint work with Benjamin Grimmer

ISMP 2024
Montreal, Canada
July 22, 2024

Problem of Interest

Consider

$$\begin{cases} \min_{x \in X} & f(x) \\ \text{s.t.} & g_i(x) \leq 0, \qquad i = 1, \ldots, m. \end{cases}$$

- x: decision variable
- $X \subseteq \mathbb{R}^n$: convex and closed set (not necessarily bounded)
- $f(x), g_i(x)$: continuous and weakly convex (nonsmooth)

Weak Convexity of f and g_i

μ-Strongly Convex

$\forall x, x' \in X$ and $\forall \zeta \in \partial h(x)$, $\exists \mu > 0$, $h - \frac{\mu}{2}\|\cdot\|^2$ is convex, or equivalently

$$h(x') \geq h(x) + \zeta^T(x' - x) + \frac{\mu}{2}\|x' - x\|^2.$$

ρ-Weakly Convex

$\forall x, x' \in X$ and $\forall \zeta \in \partial h(x)$, $\exists \rho > 0$, $h + \frac{\rho}{2}\|\cdot\|^2$ is convex, or equivalently

$$h(x') \geq h(x) + \zeta^T(x' - x) - \frac{\rho}{2}\|x' - x\|^2.$$

Some Practical Scenarios

Nonsmooth and Weakly Convex Examples:

- Objective function
  - Phase retrieval: $f(x) = \frac{1}{m}\sum_{i=1}^{m} |(a_i^T x)^2 - b_i^2|$
  - Blind deconvolution: $f(x, y) = \frac{1}{m}\sum_{i=1}^{m} |(u_i^T x)(v_i^T y) - b_i|$
  - Robust PCA: $f(X, Y) = \|XY^T - M\|_1$
- Constraint function
  - Smoothly Clipped Absolute Deviation (SCAD) regularizer
- Classification Problems
  - Multi-class Neyman-Pearson classification
  - Classification with fairness constraints

Inexact Proximal Point Methods

For an unconstrained problem

$$x_{k+1} \approx \operatorname{argmin}_{x \in X} \left\{ f(x) + \frac{1}{2\alpha} \|x - x_k\|^2 \right\}$$

Extension to a constrained problem

$$x_{k+1} \approx \operatorname{argmin}_{x \in X} \left\{ f(x) + \frac{1}{2\alpha} \|x - x_k\|^2 \mid g_i(x) + \frac{1}{2\alpha} \|x - x_k\|^2 \leq \tau \right\}$$

- α: Stepsize which depends on the weak convexity
- τ: Feasibility tolerance, $\|x_{k+1} - x_k\| \geq \sqrt{2\alpha\tau}$ implies $g_i(x_{k+1}) \leq 0$

Fritz-John (FJ) Stationarity

FJ Stationarity

A feasible x^* is an FJ point if $\exists \gamma_0^* \in \mathbb{R}_+, \gamma^* = (\gamma_1^*, ..., \gamma_m^*)^T \in \mathbb{R}_+^m$, $\zeta_f \in \partial f(x^*), \zeta_{gi} \in \partial g_i(x^*)$, $(\gamma_0^*, \gamma^*) \neq 0$

$$\gamma_0^* \zeta_f + \sum_{i=1}^m \gamma_i^* \zeta_{gi} \in -N_X(x^*), \quad \gamma_i^* g_i(x^*) = 0, \ \forall i = 1, ..., m.$$

Approximate FJ Stationarity

(a) A feasible x is an ϵ-FJ point if $\exists \gamma_0 \in \mathbb{R}_+, (\gamma_1, ..., \gamma_m)^T \in \mathbb{R}_+^m$, $\zeta_f \in \partial f(x), \zeta_{gi} \in \partial g_i(x), \sum_{i=0}^m \gamma_i = 1$

$$\mathrm{dist}(\gamma_0 \zeta_f + \sum_{i=1}^m \gamma_i \zeta_g, -N_X(x)) \leq \epsilon, \quad |\gamma_i g_i(x)| \leq \epsilon^2 \ \forall i = 1, ..., m.$$

(b) A feasible point x is an (ϵ, η)-FJ point if there exists an ϵ-FJ point x' such that $\|x - x'\| \leq \eta$.

Constraint Qualification (CQ)

Mangasarian-Fromovitz Constraint Qualification (MFCQ)

Let $A(x) = \{i \mid g_i(x) = 0, i = 1, ..., m\}$. MFCQ holds at x^* if

$$\exists v \in -N_X^*(x^*) \text{ s.t. } \zeta_{gi}^T v < 0 \ \forall i \in A(x^*), \forall \zeta_{gi} \in \partial g_i(x^*).$$

- MFCQ indicates the existence of a Slater point

σ-strong MFCQ

σ-strong MFCQ holds at x if $\exists \sigma > 0$

$$\exists v \in -N_X^*(x), \|v\| = 1 \text{ s.t. } \zeta_{gi}^T v \leq -\sigma \ \forall i \in A(x), \forall \zeta_{gi} \in \partial g_i(x).$$

Karush-Kuhn-Tucker (KKT) Stationarity

KKT Stationarity

A feasible x^* is a KKT point if $\exists \lambda^* = (\lambda_1^*, ..., \lambda_m^*)^T \in \mathbb{R}_+^m$, $\zeta_f \in \partial f(x^*), \zeta_{gi} \in \partial g_i(x^*)$

$$\zeta_f + \sum_{i=1}^m \lambda_i^* \zeta_{gi} \in -N_X(x^*), \quad \lambda_i^* g_i(x^*) = 0, \ \forall i = 1, ..., m.$$

Approximate KKT Stationarity

(a) A feasible x is an ϵ-KKT point if $\exists \lambda^* = (\lambda_1^*, ..., \lambda_m^*)^T \in \mathbb{R}_+^m$, $\zeta_f \in \partial f(x^*), \zeta_{gi} \in \partial g_i(x^*)$

$$\mathrm{dist}(\zeta_f + \sum_{i=1}^m \lambda_i \zeta_{gi}, -N_X(x)) \le \epsilon, \quad |\lambda_i g_i(x)| \le \epsilon^2 \ \forall i = 1, ..., m.$$

(b) A feasible point x is an (ϵ, η)-KKT point if there exists an ϵ-KKT point x' such that $\|x - x'\| \le \eta$.

Related Work and Outline

Related work:

- Inexact proximal point methods in [Ma, Lin, Yang 2020] and [Boob, Deng, Lan 2023]

Outline of our developments:

- Attain KKT of FJ stationarity with or without CQ
- Always feasible iterates
- Not requiring boundedness of domain

Double-loop Algorithm Structure

$$\begin{cases} \min_{x \in X} & f(x) \\ \text{s.t.} & g(x) := \max_{i=1,\dots,m} g_i(x) \le 0. \end{cases}$$

- $f(x)$ is ρ-weakly convex
- $g(x)$ is ρ-weakly convex if each $g_i(x)$ is ρ-weakly convex
- $f_{lb} = \inf_{x \in X} f(x) > -\infty$ and $g_{lb} = \inf_{x \in X} g(x) > -\infty$
- For any $x \in X$, $\zeta_f \in \partial f(x)$, $\zeta_g \in \partial g(x)$: $\|\zeta_f\|, \|\zeta_g\| \le M$

In each outer loop, we solve a $(\hat{\rho} - \rho)$-strongly convex functional constrained subproblem

$$\begin{cases} \min_{x \in X} & F_k(x) := f(x) + \frac{\hat{\rho}}{2}\|x - x_k\|^2 \\ \text{s.t.} & G_k(x) := g(x) + \frac{\hat{\rho}}{2}\|x - x_k\|^2 \le 0. \end{cases}$$

Inner Layer: Switching Subgradient Method (SSM)

$$\begin{cases} \min_{z\in Z} & F(z) \\ \text{s.t.} & G(z) \leq 0. \end{cases}$$

Definition

z is a (δ, τ)-optimal solution if $F(z) - F(z^*) \leq \delta$ and $G(z) \leq \tau$.

Algorithm The Switching Subgradient Method (SSM)

Input: $\tau > 0$, $T > 0$, $z_0 \in Z$, $\{\alpha_t\}_{t=0}^{T-1}$
Set $I = \emptyset$, $J = \emptyset$
for $t = 0, 1, ..., T-1$ **do**
If $G(z_t) \leq \tau$: $z_{t+1} = \text{proj}_Z(z_t - \alpha_t \zeta_{Ft})$, $\zeta_{Ft} \in \partial F(z_t)$, $I = I \cup \{t\}$; else: $z_{t+1} = \text{proj}_Z(z_t - \alpha_t \zeta_{Gt})$, $\zeta_{Gt} \in \partial G(z_t)$, $J = J \cup \{t\}$
end for
Output: $\bar{z}_T = \frac{\sum_{t\in I}(t+1)z_t}{\sum_{t\in I}(t+1)}$

NonLipschitz Conditions

- Previous convergence analyses on SSM [Lan and Zhou 2020, Ma, Lin, Yang 2020] assume uniform Lipschitz continuity for both $F(z)$ and $G(z)$
- Uniform Lipschitz continuity NOT hold for our subproblems!
    - Z can be unbounded
    - $F(z), G(z)$ are strongly convex

Consider the weaker nonLipschitz conditions [Grimmer 2019]: $\forall \tau > 0, \exists L_0, L_1 \geq 0$ such that $\forall z_1 \in \{z \mid G(z) \leq \tau\}, z_2 \in \{z \mid G(z) > \tau\}, \zeta_F \in \partial F(z_1),\ \zeta_G \in \partial G(z_2)$

$$\|\zeta_F\|^2 \leq L_0^2 + L_1(F(z_1) - F(z^*)),$$
$$\|\zeta_G\|^2 \leq L_0^2 + L_1(G(z_2) - G(z^*)).$$

- Subgradients bounded by suboptimality/infeasibility
- $F(z)$ and $G(z)$ are L_0-Lipschitz when $L_1 = 0$
- Allows $F(z)$ and $G(z)$ to grow quadratically when $L_1 > 0$

Convergence Results for SSM

Theorem

Given $\alpha_t = \frac{2}{\mu(t+2)+L_1^2/[\mu(t+1)]}$, $\tau > 0$, z_0 with $G(z_0) \le \tau$, we attain a (τ,τ)-optimal solution when $T \ge \max\left\{\frac{8L_0^2}{\mu\tau}, \sqrt{\frac{2L_1^2\|z_0 - z^*\|^2}{\mu\tau}}\right\}$.

Lemma

For any feasible $x_k \in X$, the non-Lipschitz condition holds for the subproblems with $L_0^2 = 9M^2 - 6\hat{\rho}g_{lb}$ and $L_1 = 6\hat{\rho}$.

Corollary

With $z_0 = x_k$, $\mu = \hat{\rho} - \rho$, $\alpha_t = \frac{2}{(\hat{\rho}-\rho)(t+2)+36\hat{\rho}^2/[(\hat{\rho}-\rho)(t+1)]}$, $\tau > 0$, we attain a (τ,τ)-optimal solution when

$$T \ge \max\left\{24(3M^2 - 2\hat{\rho}g_{lb})/(\mu\tau), \sqrt{72\hat{\rho}^2 D^2/(\mu\tau)}\right\}.$$

Outer Layer: Proximally Guided Switching Subgradient Method

Algorithm The Proximally Guided Switching Subgradient Method

Input: $\hat{\rho} > \max\{\rho, 1\}$, $\tau > 0$, T_{inner}, $x_0 \in X$ with $g(x_0) \leq 0$.
Set $\mu = \hat{\rho} - \rho$ and $\alpha_t = \frac{2}{(\hat{\rho}-\rho)(t+2)+\frac{36\hat{\rho}^2}{(\hat{\rho}-\rho)(t+1)}}$
for $k = 0, 1, ...,$ **do**
Set x_{k+1} as the output of $SSM(\tau, T_{inner}, x_k, \{\alpha_t\}_{t=0}^{T_{inner}})$ for the subproblem
end for

Boundedness of the Lagrange Multiplier

Assumption 1

The σ-strong MFCQ condition is satisfied for any subproblem.

Define $D := \sqrt{-8g_{lb}/(\hat{\rho} - \rho)}$ as an upper bound for the diameter of $\{x \mid G_k(x) \leq 0\}$

Lemma

Under Assumption 1, the optimal Lagrange multipliers for the subproblems are uniformly upper bounded by

$$B := \frac{M + \hat{\rho} D}{\sigma}.$$

Feasible Sequence of Iterates

$$\begin{cases} \tau_{FJ} = \frac{(\hat{\rho}-\rho)\epsilon^2}{4\hat{\rho}(2\hat{\rho}-\rho)} \\ T_{FJ} = \max\left\{\frac{96\hat{\rho}(2\hat{\rho}-\rho)(3M^2-2\hat{\rho}g_{lb})}{(\hat{\rho}-\rho)^2\epsilon^2}, \sqrt{\frac{288\hat{\rho}^3(2\hat{\rho}-\rho)D^2}{(\hat{\rho}-\rho)^2\epsilon^2}}\right\} \end{cases},$$

$$\begin{cases} \tau_{KKT} = \frac{(\hat{\rho}-\rho)\epsilon^2}{4(1+B)^2\hat{\rho}(2\hat{\rho}-\rho)} \\ T_{KKT} = \max\left\{\frac{96(1+B)^2\hat{\rho}(2\hat{\rho}-\rho)(3M^2-2\hat{\rho}g_{lb})}{(\hat{\rho}-\rho)^2\epsilon^2}, \sqrt{\frac{288(1+B)^2\hat{\rho}^3(2\hat{\rho}-\rho)D^2}{(\hat{\rho}-\rho)^2\epsilon^2}}\right\} \end{cases}.$$

Lemma

(a) For any $\hat{\rho} > \max\{\rho, 1\}$ with τ_{FJ} and T_{FJ}, we guarantee $g(x_k) \leq 0$ before x_k is an (ϵ, ϵ)-FJ point;

(b) Under Assumption 1, for any $\hat{\rho} > \max\{\rho, 1\}$ with τ_{KKT} and T_{KKT}, we guarantee $g(x_k) \leq 0$ before x_k is an (ϵ, ϵ)-KKT point.

Overall Convergence Results

Theorem: FJ Stationarity

For any $\hat{\rho} > \max\{\rho, 1\}$ with $\tau_{FJ} \in \mathcal{O}(\epsilon^2)$ and $T_{FJ} \in \mathcal{O}(1/\epsilon^2)$, we attain an (ϵ, ϵ)-FJ point for the original problem in $K \in \mathcal{O}(1/\epsilon^2)$ outer iterations.

Theorem: KKT Stationarity

Under Assumption 1, for any $\hat{\rho} > \max\{\rho, 1\}$ with $\tau_{KKT} \in \mathcal{O}(\epsilon^2)$ and $T_{KKT} \in \mathcal{O}(1/\epsilon^2)$, we attain an (ϵ, ϵ)-KKT point for the original problem in $K \in \mathcal{O}(1/\epsilon^2)$ outer iterations.

With or without Slater points, we attain the approximate FJ or KKT stationarity in $\mathcal{O}(1/\epsilon^4)$ subgradient evaluations.

Sparse Phase Retrieval under SCAD Constraints

Consider the experimental example:

$$\begin{cases} \min_{x\in X} & f(x) = \frac{1}{m}\sum_{i=1}^{m} |(a_i^T x)^2 - b_i^2| \\ \text{s.t.} & g(x) = \sum_{i=1}^{n} s(x_i) - p \le 0\ . \end{cases}$$

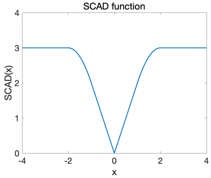
The SCAD function is defined as

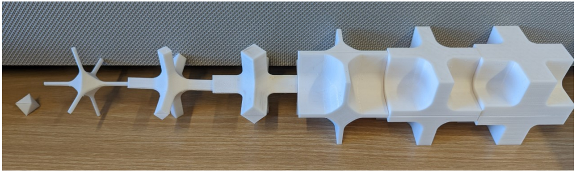
$$s(x_i) := \begin{cases} 2|x_i| & 0 \le |x_i| \le 1, \\ -x_i^2 + 4|x_i| - 1 & 1 < |x_i| \le 2, \\ 3 & |x_i| > 2. \end{cases}$$

- $X = [-10, 10]^n$, $m = 120$, $n = 120$
- $A \in \mathbb{R}^{m\times n}$, $x^* \in \mathbb{R}^n$, noise $\eta \in \mathbb{R}^m$ randomly sampled
- $b^2 = (Ax^*)^2 + \eta$
- $p \in [0, 3n)$ controls the sparsity

SCAD Function and Constraint Qualification (CQ)

1D SCAD function and seven 3D SCAD level sets with $p \in \{2.5, 3.5, 4.5, 5.5, 6.5, 7.5, 8.5\}$:

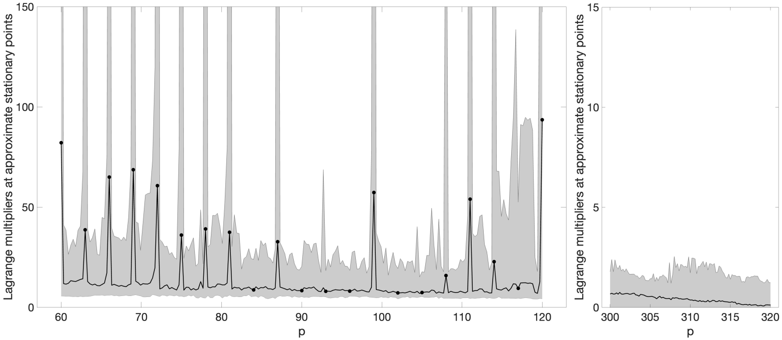




- A Slater point exists for all the subproblems if and only if p is a multiple of three!

SCAD Function and Constraint Qualification (CQ)

Small and large values of p:



- Over 30 independent replicates, approximately 5% has the Lagrange multipliers diverge when p is a multiple of three

Evaluate Stationarity and the Lagrange Multipliers

Evaluate the Stationarity:

- $\|\gamma_{k0}\zeta_{fk} + \gamma_k\zeta_{gk}\| = \hat{\rho}\|\hat{x}_{k+1} - x_k\| \approx \hat{\rho}\|x_{k+1} - x_k\|$
- $\|\zeta_{fk} + \lambda_k\zeta_{gk}\| = (1+\lambda_k)\hat{\rho}\|\hat{x}_{k+1} - x_k\| \approx (1+\lambda_k)\hat{\rho}\|x_{k+1} - x_k\|$

Evaluate the Lagrange multipliers:

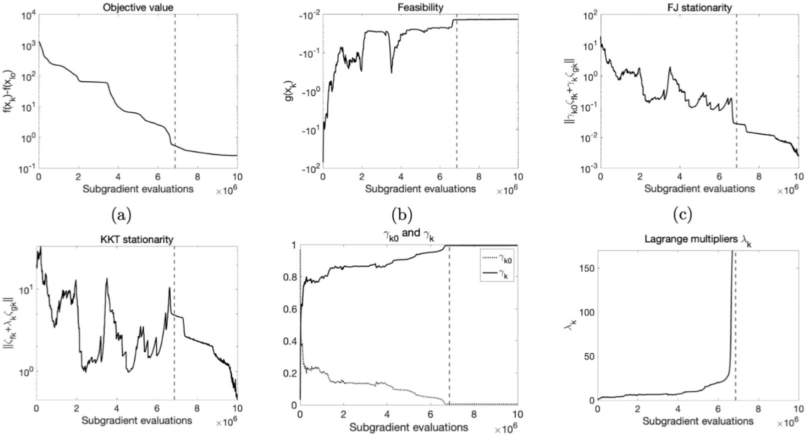
$$\frac{\sum_{t\in I_F}\alpha_t\zeta_{ft} + \sum_{t\notin I_F}\alpha_t\zeta_{gt}}{\sum_t \alpha_t} \approx \frac{\sum_{t\in I_F}\alpha_t\zeta_{Ft} + \sum_{t\notin I_F}\alpha_t\zeta_{Gt}}{\sum_t \alpha_t}$$

$$= \frac{x_k - x_{k+1}}{\sum_t \alpha_t} \approx \frac{x_k - \hat{x}_{k+1}}{\sum_t \alpha_t} \approx 0$$

- $\gamma_{k0} \approx \frac{\sum_{t\in I_F}\alpha_t}{\sum_t \alpha_t}, \quad \gamma_k \approx \frac{\sum_{t\notin I_F}\alpha_t}{\sum_t \alpha_t}, \quad \lambda_k \approx \frac{\sum_{t\notin I_F}\alpha_t}{\sum_{t\in I_F}\alpha_t}$

Experimental Results

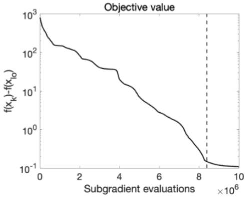
$K = 10^3, T = 10^4, p = 90$:



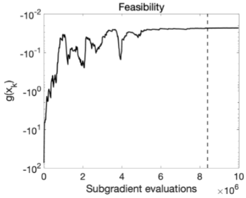
(a) (b) (c) (d) (e) (f)

Experimental Results

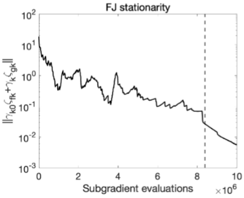
$K = 10^3, T = 10^4, p = 91$:



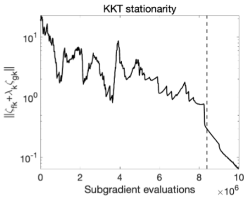
(a)



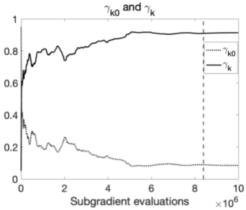
(b)



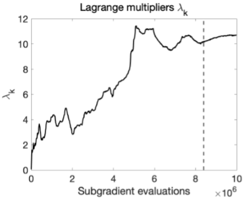
(c)



(d)



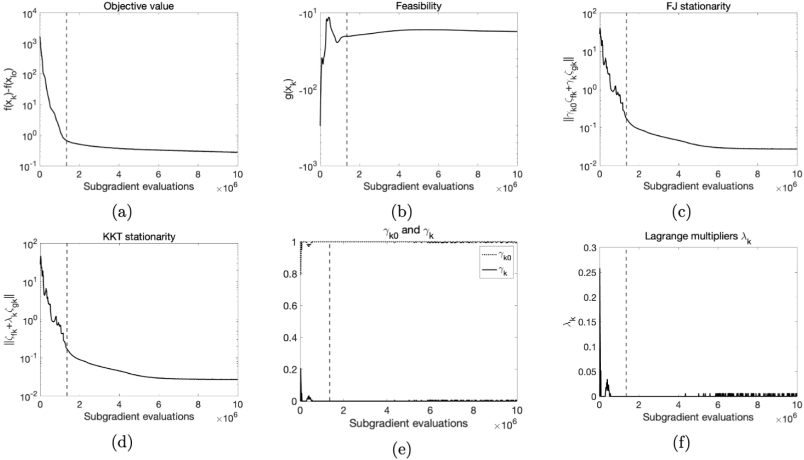
(e)



(f)

Experimental Results

$K = 10^3, T = 10^4, p = 320$:



Summary

- Convergence guarantee for KKT or FJ stationarity with or without Constraint Qualification
- Always feasible iterates
- Convergence guarantee without requiring domain compactness

Reference: Z. Jia and B. Grimmer, First-Order Methods for Nonsmooth Nonconvex Functional Constrained Optimization with or without Slater Points, arXiv Pre-print 2212.00927.