ICML 2022

A Single-Loop Gradient Descent and Perturbed Ascent Algorithm for Nonconvex Functional Constrained Optimization

Songtao Lu

IBM Thomas J. Watson Research Center

Nonconvex Functional Constrained Problems

- Consider the following problem

$$\mathbb{P}1 \quad \min_{\mathbf{x}\in\mathcal{X}\subseteq\mathbb{R}^d} \quad f(\mathbf{x}) \quad \text{subject to} \quad g(\mathbf{x}) \leq 0 \tag{1}$$

- $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ and $g(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^m$ are smooth (possibly) nonconvex.
  - $\mathcal{X}$: convex feasible set
  - m: number of constraints

- Applications
  - Multi-class Neyman-Pearson classification (mNPC)
  - Constrained Markov decision processes (CMDP)
  - Deep neural networks training under energy budget

Gradient Descent and Perturbed Ascent

- Find a stationary (quasi-Nash equilibrium) point of the following problem

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\boldsymbol{\lambda}\geq 0} F_\beta(\mathbf{x},\boldsymbol{\lambda}) \tag{2}$$

- Perturbed augmented Lagrangian function

$$F_\beta(\mathbf{x},\boldsymbol{\lambda}) \triangleq f(\mathbf{x}) + \frac{\beta}{2}\left\|\left[g(\mathbf{x})+\frac{(1-\tau)\boldsymbol{\lambda}}{\beta}\right]_+\right\|^2 - \frac{\|(1-\tau)\boldsymbol{\lambda}\|^2}{2\beta} \tag{3}$$

  - $\boldsymbol{\lambda}\in\mathbb{R}_+^m$: dual variable (Lagrangian multiplier)
  - $[\mathbf{x}]_+$: component-wise nonnegative part of vector $\mathbf{x}$
  - $\tau\in(0,1)$: perturbation term
  - $\beta>0$

- Gradient descent and perturbed ascent (GDPA)

$$\mathbf{x}_{r+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\langle \nabla_{\mathbf{x}} F_{\beta_r}(\mathbf{x}_r,\boldsymbol{\lambda}_r), \mathbf{x}-\mathbf{x}_r\right\rangle + \frac{1}{2\alpha_r}\|\mathbf{x}-\mathbf{x}_r\|^2 \tag{4a}$$

$$\boldsymbol{\lambda}_{r+1} = \arg\max_{\boldsymbol{\lambda}\geq 0} \left\langle \frac{1}{1-\tau}\nabla_{\boldsymbol{\lambda}} F_{\beta_r}(\mathbf{x}_{r+1},\boldsymbol{\lambda}_r), \boldsymbol{\lambda}-\boldsymbol{\lambda}_r\right\rangle - \frac{1-\tau}{2\beta_r}\|\boldsymbol{\lambda}-\boldsymbol{\lambda}_r\|^2 - \frac{\gamma_r}{2}\|\boldsymbol{\lambda}\|^2 \tag{4b}$$

  - $\alpha_r,\beta_r,\gamma_r>0$: dynamic sequences (learning rates)

Gradient Descent and Perturbed Ascent

- Substituting the perturbed augmented Lagrangian function into (4) yields

$$\mathbf{x}_{r+1} = \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}_r - \alpha_r\left(\nabla f(\mathbf{x}_r) + J^T(\mathbf{x}_r)\left[(1-\tau)\boldsymbol{\lambda}_r + \beta_r g(\mathbf{x}_r)\right]_+\right)\right) \quad (5a)$$

$$[\boldsymbol{\lambda}_{r+1}]_i = \begin{cases} \mathcal{P}_{\geq 0}\left((1-\tau)[\boldsymbol{\lambda}_r]_i + \beta_r g_i(\mathbf{x}_{r+1})\right), i \in \mathcal{S}_r \\ 0, i \in \overline{\mathcal{S}}_r \end{cases} \quad (5b)$$

  - $J(\mathbf{x})$: Jacobian matrix of the constraints at point $\mathbf{x}$
  - $g_i(\mathbf{x})$: the ith constraint
  - $[\mathbf{x}]_i$: the ith entry of vector $\mathbf{x}$
  -

$$\mathcal{S}_r \triangleq \left\{ i | g_i(\mathbf{x}_r) + \frac{(1-\tau)[\boldsymbol{\lambda}_r]_i}{\beta_r} > 0 \right\} \quad (6)$$

  - $g_i(\mathbf{x}_r) \leq 0, i \in \overline{\mathcal{S}}_r$
  - $\mathcal{P}_{\mathcal{X}}$: the projection of iterates to the feasible set
  - $\mathcal{P}_{\geq 0} \triangleq []_+$: the component-wise nonnegative projection operator

Theoretical Guarantees

Assumptions

A1 Lipschitz continuity of function $f(\mathbf{x})$.
A2 Lipschitz continuity of function $g(\mathbf{x})$.
A3 Lower boundedness of function $f(\mathbf{x})$
A4 Upper boundedness of function $g_+(\mathbf{x})$
A5 Regularity condition (constraint qualification)

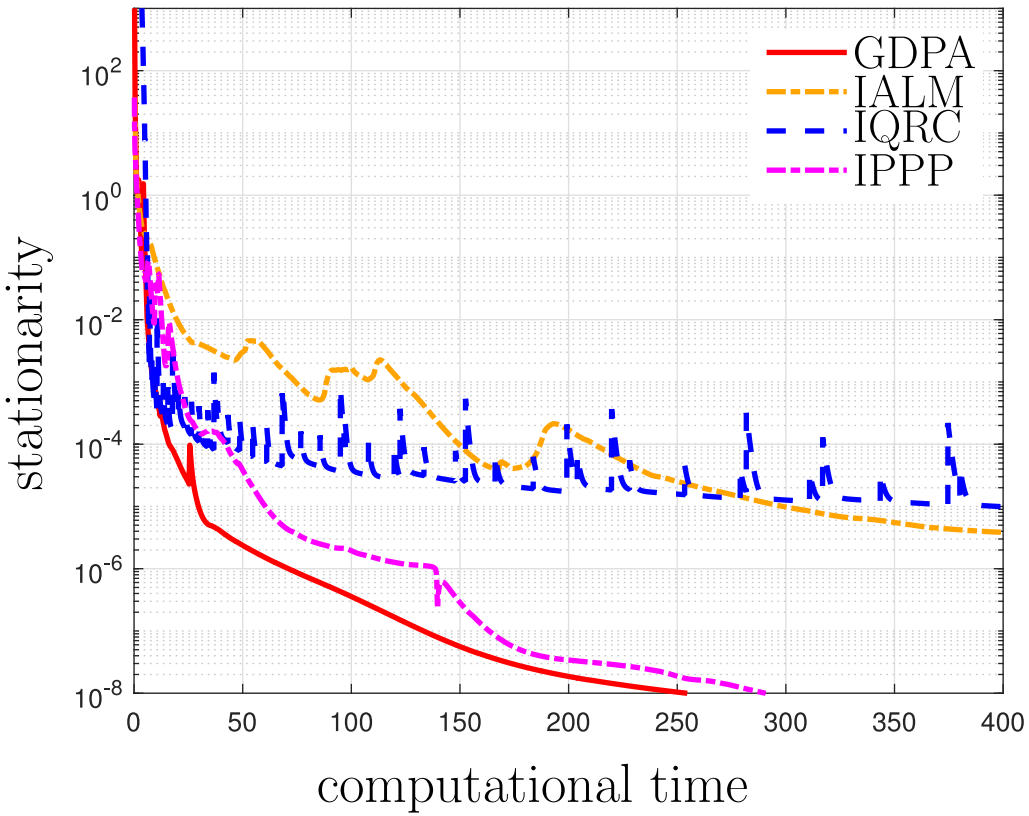
Theorem 1: Under assumptions A1-A5. When the step-sizes are chosen as $\alpha_r \sim 1/\beta_r \sim \mathcal{O}(1/r^{1/3})$ and $\gamma_r\beta_r = \tau > 1 - \sigma/\sqrt{66U_J^2 + \sigma^2}$, then the outputs of GDPA $\bar{\mathbf{x}}_{T(\epsilon)}, \bar{\boldsymbol{\lambda}}_{T(\epsilon)}$ converge to an ϵ-approximate KKT point satisfying

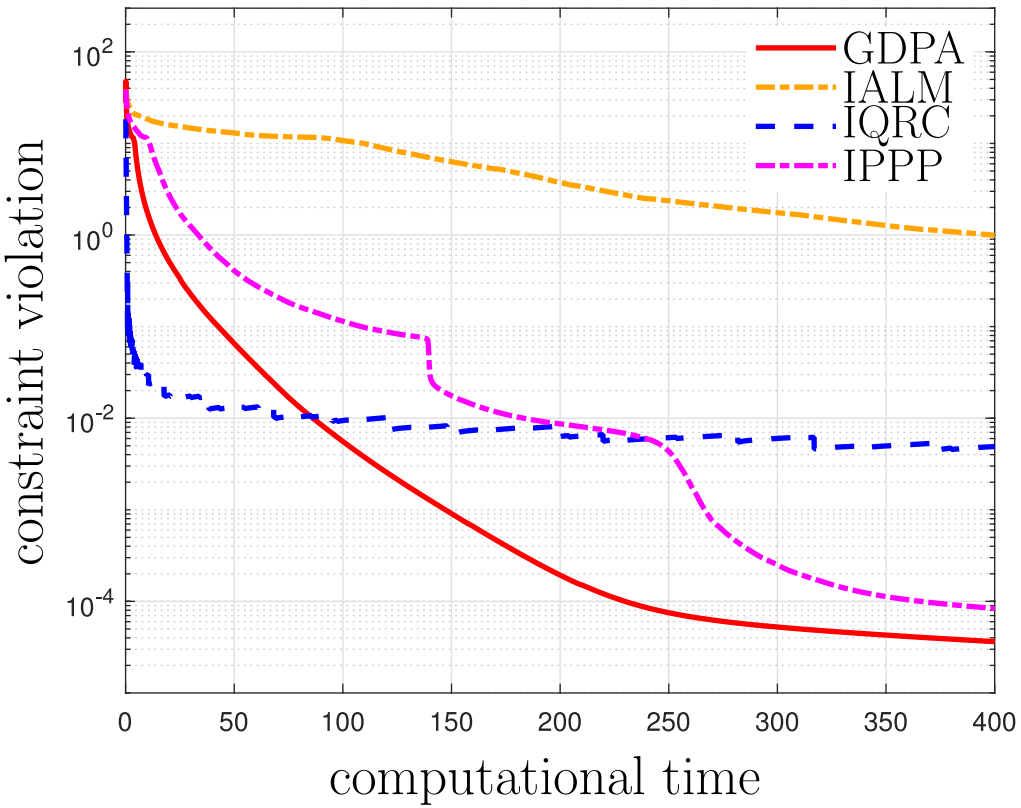
$$\text{dist}\left(\nabla f(\bar{\mathbf{x}}_{T(\epsilon)}) + \sum_{i=1}^{m} [\bar{\boldsymbol{\lambda}}_{T(\epsilon)}]_i \nabla g_i(\bar{\mathbf{x}}_{T(\epsilon)}), -\mathcal{N}_{\mathcal{X}}(\mathbf{x})\right) \leq \epsilon,$$

$$\|g_+(\bar{\mathbf{x}}_{T(\epsilon)})\| \leq \epsilon, \quad \sum_{i=1}^{m} |[\boldsymbol{\lambda}_{T(\epsilon)}]_i g_i(\bar{\mathbf{x}}_{T(\epsilon)})| \leq \epsilon, \tag{7}$$

in the number of $\mathcal{O}(1/\epsilon^3)$ iterations.

Numerical Results: mNPC problem





- Compared algorithms
  - inexact augmented lagrangian method (IALM) (Sahin et al., 2019; Li et al., 2021) ($\mathcal{O}(1/\epsilon^3)$, double-loop)
  - inexact quadratically regularized constrained (IQRC) methods (Ma et al., 2020) ($\mathcal{O}(1/\epsilon^3)$, double-loop)
  - inexact proximal-point penalty (IPPP) method (Lin et al., 2022) ($\mathcal{O}(1/\epsilon^3)$, triple-loop)

Thank You!