Inexact Proximal Augmented Lagrangian Method for weakly-Convex Problems with Convex Constraints

Hari Dahal, *Wei Liu*, Yangyang Xu
Mathematical Sciences, Rensselaer Polytechnic Institute

SIAM-NNP

October 20, 2023

Partly supported by NSF award #2053493 and ONR award N00014-22-1-2573.

Outline

1. Problem formulation and applications
2. Existing first-order methods
3. Proposed method with complexity and numerical results
4. Conclusions

Problem formulation

$$\begin{aligned} &\min_{\mathbf{x}} \ F(\mathbf{x}) := f(\mathbf{x}) + h(\mathbf{x}), \\ &\text{s.t.} \ \ \mathbf{A}\mathbf{x} = \mathbf{b}, \ \mathbf{g}(\mathbf{x}) := [g_1(\mathbf{x}), \ldots, g_m(\mathbf{x})] \leq \mathbf{0} \end{aligned} \tag{P}$$

- f is ρ-weakly convex, i.e., $f(\cdot) + \frac{\rho}{2}\|\cdot\|^2$ is convex for some $\rho > 0$
- h closed convex and proximable
- $\mathbf{A}$ and $\mathbf{b}$ are given matrix and vector; each g_i is convex and smooth

Two cases of f to be explored

1. f is smooth
2. $f = l \circ \mathbf{c}$ with a closed convex l and a smooth vector function $\mathbf{c}$

Remark: Case 1 is a special case of Case 2, but they will be handled differently

Many applications

- all linear (equality or inequality) constrained and/or quadratically-constrained problems with a smooth objective
- linear constrained nonlinear least squares [Orban-Siqueira'20]
- constrained machine learning: Neyman-Pearson classification [Scott'07], logical neural network [Riegel al et.'20]
- robust phase retrieval [Davis-Drusvyatskiy'19] and linear-constrained variant

Existing first-order methods and their complexity

(almost all based on penalty or augmented Lagrangian method or SQP)

Proximal augmented Lagrangian method [Kong-Melo-Monteiro'23]

$$\mathcal{L}_\beta(\mathbf{x};\mathbf{y},\mathbf{z}) = F(\mathbf{x}) + \mathbf{y}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) + \tfrac{\beta}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 + \tfrac{\beta}{2}\left\|[g(\mathbf{x}) + \tfrac{\mathbf{z}}{\beta}]_+\right\|^2 - \tfrac{\|\mathbf{z}\|^2}{2\beta}$$

Consider a convex cone-constrained problem. When the cone $\mathcal{K} = \{\mathbf{0}\} \oplus \mathbb{R}^m_-$, reduces to (P) with a smooth f, and the main updates become

For $k = 0, 1, \ldots$

$$\mathbf{x}^{k+1} \approx \arg\min_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{y}^k, \mathbf{z}^k) + \tfrac{1}{2\lambda}\|\mathbf{x}-\mathbf{x}^k\|^2,$$

$$\mathbf{y}^{k+1} = \mathbf{y}^k + \beta_k(\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}), \quad \mathbf{z}^{k+1} = [\mathbf{z}^k + \beta_k \mathbf{g}(\mathbf{x}^{k+1})]_+$$

- $\lambda > 0$ is chosen to have strongly-convex $\mathbf{x}$-subproblems, which can then be solved inexactly by an accelerated proximal gradient method
- β_k will be doubled once a certain condition about $\mathbf{x}$ holds and increase to at most $\Theta(1/\varepsilon^2)$, in order to produce an ε-KKT point
- $\tilde{O}(\varepsilon^{-3})$ total complexity to produce an ε-KKT point

---

Kong-Melo-Monteiro, Iteration Complexity of a Proximal Augmented Lagrangian Method for Solving Nonconvex Composite Optimization Problems with Nonlinear Convex Constraints. MOR, 2023

Smoothing prox-linear gradient method [Drusvyatskiy-Paquette'19]

Consider compositional problem: $\min_{\mathbf{x}} l \circ \mathbf{c}(\mathbf{x}) + h(\mathbf{x})$, where $\mathbf{c}$ is smooth, and l and h closed convex. Perform the prox-linear update:

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} h(\mathbf{x}) + l\left(\mathbf{c}(\mathbf{x}^k) + J_{\mathbf{c}}^{\top}(\mathbf{x}^k)(\mathbf{x} - \mathbf{x}^k)\right) + \tfrac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}^k\|^2$$

- When l is smooth, subproblems solved by an APG to a certain accuracy
- Otherwise, smooth l by its Moreau envelope

Observation: the smoothing method applied to (P) becomes a prox-linear method on a quadratic penalty problem, by noting

- If l is the indicator function on $\mathcal{Y} = \{\mathbf{0}\}$, its Moreau envelope
$$l_\nu(\mathbf{y}) = \min_{\mathbf{y}' \in \mathcal{Y}} \tfrac{1}{2\nu}\|\mathbf{y}' - \mathbf{y}\|^2 = \tfrac{1}{2\nu}\|\mathbf{y}\|^2$$
- If l is the indicator function on $\mathcal{Y} = \mathbb{R}^m_-$, its Moreau envelope
$$l_\nu(\mathbf{y}) = \min_{\mathbf{y}' \in \mathcal{Y}} \tfrac{1}{2\nu}\|\mathbf{y}' - \mathbf{y}\|^2 = \tfrac{1}{2\nu}\|[\mathbf{y}]_+\|^2$$

Drusvyatskiy-Paquette, Efficiency of minimizing compositions of convex functions and smoothmaps. Math Programming, Series A, 2019.

Many more for nonconvex problems with constraints

- On smooth/composite problems with linear constraints
  [Kon-Melo-Monteiro'19'20'23; Zhang-Luo'20'22; Sun-Hong'19; ...]
- On composite problems with nonlinear convex constraints [Li-X.'21; Lin-Ma-X.'22; Kon-Melo-Monteiro'23; Lan-Zhou'20; Zhang-Pu-Luo'22; ...]
- On smooth problems with nonlinear nonconvex constraints
  [Curtis-Overton'12; Berahas et al'21; Xie-Wright'21; Curtis-Robinson-Zhou'23; Jin-Wang'22; ...]
- On composite problems with nonlinear nonconvex constraints
  [Cartis-Gould-Toint'11; Lin-Ma-X.'22; Li et al'21; Sahin et al'19; Bob-Deng-Lan'22; ...]

Proposed method:

inexact (smoothed) proximal augmented Lagrangian

Framework of inexact proximal augmented Lagrangian

The augmented Lagrangian function is

$$\mathcal{L}_\beta(\mathbf{x};\mathbf{y},\mathbf{z}) = F(\mathbf{x}) + \mathbf{y}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) + \tfrac{\beta}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 + \tfrac{\beta}{2}\left\|[g(\mathbf{x})+\tfrac{\mathbf{z}}{\beta}]_+\right\|^2 - \tfrac{\|\mathbf{z}\|^2}{2\beta}$$

Main updates: for $k = 0, 1, \ldots$

$$\mathbf{x}^{k+1} \approx \arg\min_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{y}^k, \mathbf{z}^k) + \rho\|\mathbf{x}-\mathbf{x}^k\|^2$$

$$\mathbf{y}^{k+1} = \mathbf{y}^k + \alpha_k(\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}), \quad \mathbf{z}^{k+1} = \mathbf{z}^k + \gamma_k \max\{-\tfrac{\mathbf{z}^k}{\beta_k}, \mathbf{g}(\mathbf{x}^{k+1})\}$$

- The proximal parameter set to ρ for convenience; good as long as each $\mathbf{x}$-subproblem is strongly convex
- α_k and γ_k adaptive to constraint violation with summable $\{v_k\}$ and $\{w_k\}$

$$\alpha_k := \min\left\{\beta_k, \tfrac{v_k}{\|\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}\|}\right\}, \quad \gamma_k := \min\left\{\beta_k, \tfrac{w_k}{\|[g(\mathbf{x}^{k+1})]_+\|}\right\}$$

- β_k increases slowly and at most to $O(1/\varepsilon)$ for ε-stationarity

Iteration complexity

Key assumptions: boundedness of $\mathrm{dom}(F)$, continuity of F, smoothness of $\mathbf{g}$, and Slater's condition

- Primal Infeasibility: $\|\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}\| + \|[g(\mathbf{x}^{k+1})]_+\| = O(1/\beta_k)$
- Complementarity: $\sum_{i=1}^{m} \left|[z_i^k + \beta_k g_i(\mathbf{x}^{k+1})]_+ g_i(\mathbf{x}^{k+1})\right| = O(1/\beta_k)$
- Dual Infeasibility: dependent on how accurate $\mathbf{x}$-subproblems are solved

How to solve subproblems: Case 1

When f is smooth, apply an accelerated proximal gradient (APG) to

$$\min_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{y}^k, \mathbf{z}^k) + \rho\|\mathbf{x} - \mathbf{x}^k\|^2$$

- $O(\sqrt{\beta_k/\rho}\ln\frac{1}{\varepsilon_k})$ APG iterations to obtain an ε_k-stationary point, $\varepsilon_k = O(\varepsilon)$

Key step for $\mathbf{x}$-subproblem solving: we find a near-stationary point of each prox-AL function, which will imply,

$$\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}; \mathbf{y}^k, \mathbf{z}^k) + \rho\|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2 \leq \mathcal{L}_{\beta_k}(\mathbf{x}^k; \mathbf{y}^k, \mathbf{z}^k) + \delta_k$$

- δ_k depends on the stationarity violation of each $\mathbf{x}$-subproblem
- $\sum_{k=0}^{K}\|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2 = O\big(1 + \sum_{k=0}^{K}\delta_k\big), \forall\, K \geq 0$
- If $\|\mathbf{x}^k - \mathbf{x}^{k+1}\|$ is small, $\mathbf{x}^{k+1}$ will nearly satisfy the dual feasibility in KKT system with multipliers $\mathbf{y}^k + \beta_k(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b})$ and $[\mathbf{z}^k + \beta_k\mathbf{g}(\mathbf{x}^{k+1})]_+$

How to solve subproblems: Case 2

When $f = l \circ \mathbf{c}$, apply an APG to

$$\min_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{y}^k, \mathbf{z}^k) + \rho\|\mathbf{x} - \mathbf{x}^k\|^2 - f(\mathbf{x}) + l^{\nu_k}\left(\mathbf{c}(\mathbf{x}^k) + J_{\mathbf{c}}^\top(\mathbf{x}^k)(\mathbf{x} - \mathbf{x}^k)\right)$$

where $l^\nu(\cdot)$ is the Moreau envelope of l, $\widehat{\mathbf{x}}^k$ is its global minimizer

- l^ν is $\frac{\|\nabla \mathbf{c}\|}{\nu}$-smooth
- $O(\sqrt{(\beta_k + \frac{1}{\nu_k})/\rho}\ln\frac{1}{\varepsilon_k})$ APG iterations to obtain an ε_k-stationary point
- $\nu_k \sim \varepsilon^2$ eventually to produce a near ε-KKT point of (P)

Key step for $\mathbf{x}$-subproblem solving: we find a near-stationary point of each smoothed prox-AL function, which will imply,

$$\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}; \mathbf{y}^k, \mathbf{z}^k) + \frac{\rho}{2}\|\widehat{\mathbf{x}}^k - \mathbf{x}^{k+1}\|^2 \le \mathcal{L}_{\beta_k}(\mathbf{x}^k; \mathbf{y}^k, \mathbf{z}^k) + \delta_k$$

- δ_k depends on the stationarity violation and smoothing parameter ν_k
- $\sum_{k=0}^K \|\widehat{\mathbf{x}}^k - \mathbf{x}^{k+1}\|^2 = O\left(1 + \sum_{k=0}^K \delta_k\right), \forall\, K \ge 0$
- If $\|\widehat{\mathbf{x}}^k - \mathbf{x}^{k+1}\|$ is small, $\widehat{\mathbf{x}}^k$ will nearly satisfy the dual feasibility in KKT system with multipliers $\mathbf{y}^k + \beta_k(\mathbf{A}\widehat{\mathbf{x}}^k - \mathbf{b})$ and $[\mathbf{z}^k + \beta_k \mathbf{g}(\widehat{\mathbf{x}}^k)]_+$

Total iteration complexity

Parameter setting: $K = \Theta(1/\varepsilon^2)$ for a given $\varepsilon > 0$, $\beta_k = \Theta(\sqrt{k})$ and $\varepsilon_k = O(\epsilon)$ for any k

- An ε-KKT point can be produced after K outer iterations
- When f is smooth, $O\left(\sum_{k=1}^{K} \sqrt{\beta_k/\rho} \ln \frac{1}{\varepsilon_k}\right) = \tilde{O}(1/\varepsilon^{2.5})$ total APG iterations
- When $f = l \circ \mathbf{c}$, set $\nu_k \sim \varepsilon^2$. Then $O(\sum_{k=1}^{K} \sqrt{(\beta_k + \frac{1}{\nu_k})/\rho} \ln \frac{1}{\varepsilon_k}) = \tilde{O}(1/\varepsilon^3)$ total APG iterations

Comparison to existing methods

- For composite nonconvex optimization, convex constraints V.S. only linear constraints in [Zeng-Yin-Zhou'22]

- For composite nonconvex optimization with convex constraints, $\tilde{O}(1/\varepsilon^{2.5})$ V.S. $\tilde{O}(1/\varepsilon^{3})$ in [Kong-Melo-Monteiro'23]
  - different step sizes for dual update, different rate to increase β_k

- For compositional optimization, with convex constraints V.S. without constraints in [Drusvyatskiy-Paquette'19],
  - same-order complexity, better empirical performance (shown next)

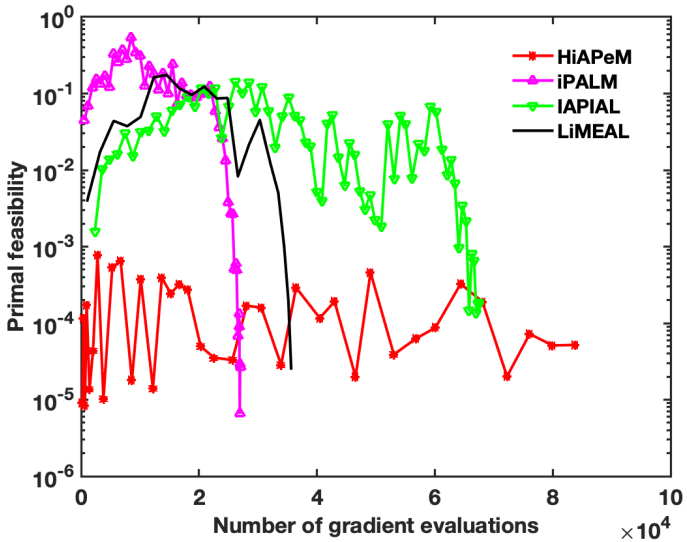
Numerical results

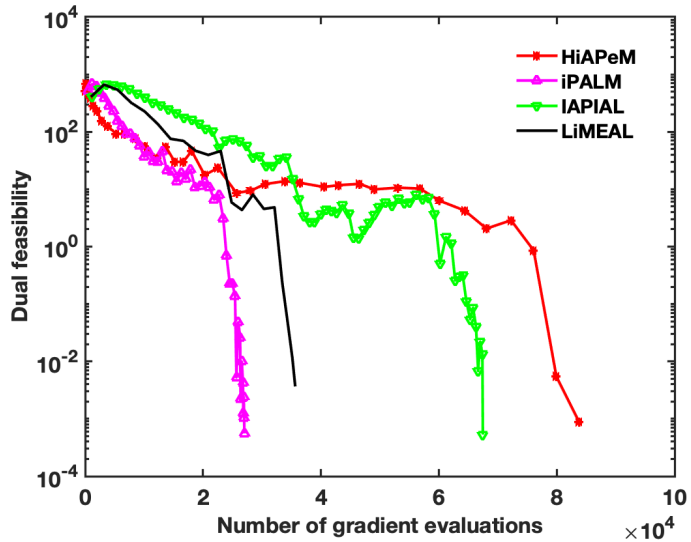
(on three different problems)

Nonconvex linearly-constrained quadratic program

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\mathbf{x}^\top \mathbf{Q}_0 \mathbf{x} + \mathbf{c}_0^\top \mathbf{x}, \text{ s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}, x_i \in [-5, 5], \forall i = 1, \ldots, n.$$

- $\mathbf{Q}_0$ symmetric and has minimum eigenvalue $-\rho < 0$



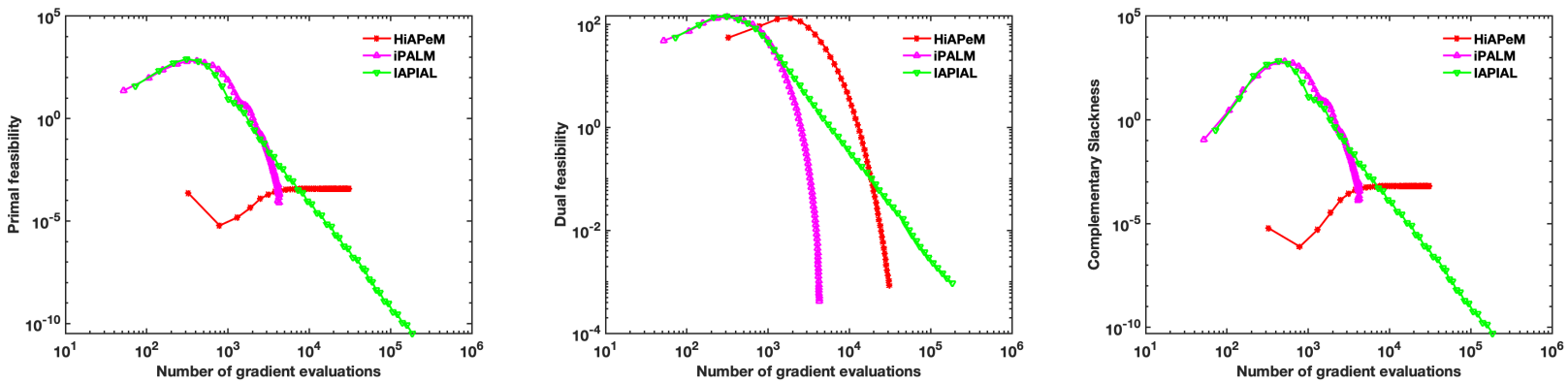


- iPALM (proposed); HiAPeM [Li-X.'21]; IAPIAL [Kong-Melo-Monteiro'23]; LiMEAL [Zeng-Yin-Zhou'22]
- $\rho = 10$; more results in paper

Nonconvex quadratically-constrained quadratic program

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ \tfrac{1}{2}\mathbf{x}^\top \mathbf{Q}_0 x + \mathbf{c}_0^\top \mathbf{x},$$
$$\text{s.t. } \tfrac{1}{2}\mathbf{x}^\top \mathbf{Q}_j \mathbf{x} + \mathbf{c}_j^\top \mathbf{x} + d_j \le 0, \forall j \in [m]; x_i \in [-5,5], \forall i \in [n]$$

- $\mathbf{Q}_0$ symmetric and has minimum eigenvalue $-\rho < 0$

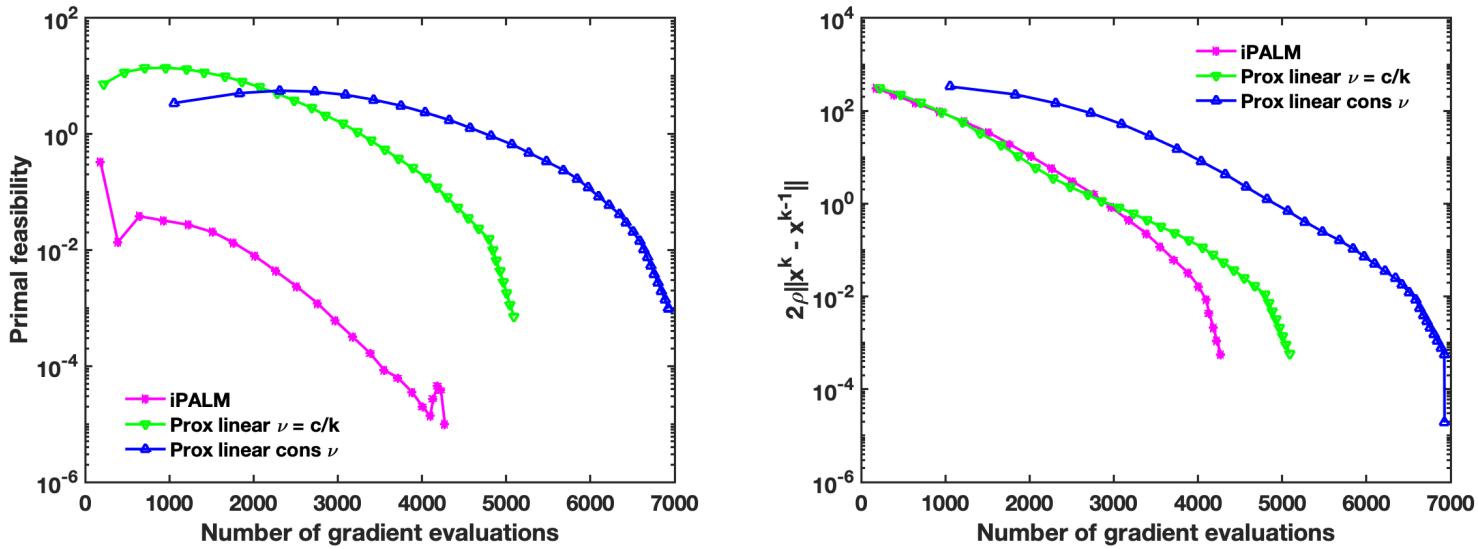


- iPALM (proposed); HiAPeM [Li-X.'21]; IAPIAL [Kong-Melo-Monteiro'23]
- $\rho = 10$; more results in paper

Linear constrained robust nonlinear least square

$$\min_{\mathbf{x}} \|\mathbf{f}(\mathbf{x})\|_1, \text{ s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}$$

- $\mathbf{f}$ quadratic vector function



- iPALM (proposed); Prox linear [Drusvyatskiy-Paquette'19]

Conclusions

- Presented a framework of inexact proximal augmented Lagrangian method (PALM) for weakly-convex optimization with linear and convex nonlinear functional constraints
  - $O(1/\varepsilon^2)$ outer iteration complexity to produce an ε-KKT point
- Explored two cases of weakly-convex objective
  - smooth composite case: accelerated proximal gradient (APG) to PALM subproblems and $\tilde{O}(1/\varepsilon^{2.5})$ total APG iterations
  - compositional case: APG to Moreau envelope smoothed PALM subproblems and $\tilde{O}(1/\varepsilon^3)$ total APG iterations
- Conducted numerical experiments on a few classes of problems and observed better performance over existing methods
  - nonconvex linearly-constrained quadratic program
  - nonconvex quadratically-constrained quadratic program
  - linear constrained robust nonlinear least square

Publication

H. Dahal, W. Liu and Y. Xu. Inexact proximal augmented Lagrangian method for weakly-convex problems with convex constraints.

Thank you!!!