Complexity of a quadratic penalty accelerated inexact proximal point method

W. Kong[1] J.G. Melo[2] R.D.C. Monteiro[1]

[1]School of Industrial and SystemsEngineering
Georgia Institute of Technology

[2]Institute of Mathematics and Statistics
Federal University of Goias

ICERM 2019 - April 30th , Providence

The main problem:

$$(P) \qquad \phi^* := \min\left\{\phi(z) := f(z) + h(z) : Az = b,\ z \in \mathbb{R}^n\right\}$$

where

- $A : \mathbb{R}^n \to \mathbb{R}^l$ is linear and $b \in \mathbb{R}^l$
- $h : \mathbb{R}^n \to (-\infty, \infty]$ closed proper convex with bounded domain;
- f is differentiable (not necessarily convex) on $\mathrm{dom}\, h$ and, for some $L_f > 0$,

$$\|\nabla f(z) - \nabla f(z')\| \leq L_f \|z - z'\|, \quad \forall z, z' \in \mathrm{dom}\, h$$

The main problem (continued):

$$(P) \qquad \phi^* := \min\{\phi(z) := f(z) + h(z) : Az = b,\ z \in \mathbb{R}^n\}$$

Our goal: Given $(\bar{\rho}, \bar{\eta}) > 0$, find a $(\bar{\rho}, \bar{\eta})$-approximate solution of (P), i.e., a triple $(\bar{z}, \bar{w}; \bar{v})$ such that

$$\bar{v} \in \nabla f(\bar{z}) + \partial h(\bar{z}) + A^* \bar{w}, \quad \|\bar{v}\| \le \bar{\rho}, \quad \|A\bar{z} - b\| \le \bar{\eta}$$

It will be achieved via a penalty approach.

For $c > 0$, consider

$$(P_c) \quad \phi_c^* := \min_z \phi_c(z) := f_c(z) + h(z)$$

where

$$f_c(z) := f(z) + \frac{c}{2}\|Az - b\|^2$$

Quadratic Penalty Approach:

0. choose initial $c > 0$
1. obtain a $\bar{\rho}$-approximate solution $(\bar{z}; \bar{v})$ of (P_c), i.e., satisfying

$$\bar{v} \in \nabla f_c(\bar{z}) + \partial h(\bar{z}), \quad \|\bar{v}\| \le \bar{\rho}$$

2. if $\|A\bar{z} - b\| \le \bar{\eta}$ then stop and output $\bar{z}$; otherwise, set $c \leftarrow 2c$ and go to step 1

For $c > 0$, consider

$$(P_c) \quad \phi_c^* := \min_z \phi_c(z) := f_c(z) + h(z)$$

where

$$f_c(z) := f(z) + \frac{c}{2}\|Az - b\|^2$$

Quadratic Penalty Approach:

0. choose initial $c > 0$
1. obtain a $\bar{\rho}$-approximate solution $(\bar{z}; \bar{v})$ of (P_c), i.e., satisfying

$$\bar{v} \in \nabla f_c(\bar{z}) + \partial h(\bar{z}), \quad \|\bar{v}\| \leq \bar{\rho}$$

2. if $\|A\bar{z} - b\| \leq \bar{\eta}$ then stop and output $\bar{z}$; otherwise, set $c \leftarrow 2c$ and go to step 1

Theorem

Let $(\bar{\rho}, \bar{\eta}) > 0$ be given. Assume that $(\bar{z}; \bar{v})$ is a $\bar{\rho}$-approximate solution of (P_c) and define

$$\bar{w} := c(A\bar{z} - b), \quad R := 2\Delta_\phi^* + 2\bar{\rho} D_h + L_f D_h^2$$

where

$$D_h := \sup\{\|z - z'\| : z, z' \in \operatorname{dom} h\},$$
$$\Delta_\phi^* := \phi^* - \phi_*, \quad \phi_* := \inf_z\{(f + h)(z) : z \in \mathbb{R}^n\}$$

Then, $(\bar{z}, \bar{w}; \bar{v})$ is $(\bar{\rho}, \bar{\eta})$-approximate solution of (P) whenever

$$c \geq \frac{R}{\bar{\eta}^2}$$

Recall that the objective function of (P_c) is $\phi_c = f_c + h$ where

$$f_c(z) := f(z) + c\|Az - b\|^2/2$$

For every $z, z' \in \operatorname{dom} h$,

$$-m \leq \frac{f_c(z') - [f_c(z) + \langle \nabla f_c(z), z' - z \rangle]}{\|z' - z\|^2/2} \leq M_c$$

where

$$m := L_f, \quad M_c := L_f + c\|A\|^2$$

The complexity of the composite gradient meth for solving (P_c) is

$$\mathcal{O}\left(M_c \frac{mD_h^2}{\bar{\rho}^2}\right)$$

which is high for large c, or when $M_c >> m$.

Recall that the objective function of (P_c) is $\phi_c = f_c + h$ where

$$f_c(z) := f(z) + c\|Az - b\|^2/2$$

For every $z, z' \in \operatorname{dom} h$,

$$-m \le \frac{f_c(z') - [f_c(z) + \langle \nabla f_c(z), z' - z\rangle]}{\|z' - z\|^2/2} \le M_c$$

where

$$m := L_f, \quad M_c := L_f + c\|A\|^2$$

The complexity of the composite gradient meth for solving (P_c) is

$$\mathcal{O}\left(M_c \frac{m D_h^2}{\bar{\rho}^2}\right)$$

which is high for large c, or when $M_c >> m$.

- S. Ghadimi and G. Lan "Accelerated gradient methods for nonconvex nonlinear and stochastic programming", published 2016

  Complexity:

  $$\mathcal{O}\left(\frac{M_c m D_h^2}{\bar{\rho}^2} + \left(\frac{M_c d_0}{\bar{\rho}}\right)^{2/3}\right)$$

  The dominant term (i.e., the blue one) is $\mathcal{O}(M_c)$.

- Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford "Accelerated methods for non-convex optimization", arXiv 2017

  obtained a $\mathcal{O}(\sqrt{M_c}\log M_c)$ complexity bound under the assumption that $h = 0$.

Our AIPP approach removes the $\log M_c$ from the above bound and the assumption that $h = 0$

AIPP for solving (P_c) is based on an IPP scheme whose k-th iteration is as follows. Given z_{k-1}, it chooses $\lambda_k > 0$ and approximately solves the 'prox' subproblem

$$(P_c^k) \quad \min \left\{ \lambda_k (f_c + h)(z) + \frac{1}{2}\|z - z_{k-1}\|^2 \right\}$$

i.e., for some $\sigma \in (0,1)$, it computes a point z_k and a residual pair $(v_k, \varepsilon_k) \in \mathbb{R}^n \times \mathbb{R}_+$ such that

$$v_k \in \partial_{\varepsilon_k} \left(\lambda_k (f_c + h) + \frac{1}{2}\| \cdot - z_{k-1}\|^2 \right)(z_k)$$

$$\|v_k\|^2 + 2\varepsilon_k \leq \sigma \|z_{k-1} - z_k + v_k\|^2$$

AIPP for solving (P_c) is based on an IPP scheme whose k-th iteration is as follows. Given z_{k-1}, it chooses $\lambda_k > 0$ and approximately solves the 'prox' subproblem

$$(P_c^k) \quad \min \left\{ \lambda_k (f_c + h)(z) + \frac{1}{2}\|z - z_{k-1}\|^2 \right\}$$

i.e., for some $\sigma \in (0,1)$, it computes a point z_k and a residual pair $(v_k, \varepsilon_k) \in \mathbb{R}^n \times \mathbb{R}_+$ such that

$$v_k \in \partial_{\varepsilon_k} \left(\lambda_k (f_c + h) + \frac{1}{2}\| \cdot - z_{k-1}\|^2 \right)(z_k)$$

$$\|v_k\|^2 + 2\varepsilon_k \le \sigma \|z_{k-1} - z_k + v_k\|^2$$

AIPP method: It is an accelerated instance of the above IPP scheme in which for all k:

- $\lambda_k = 1/(2m)$, and hence (P_c^k) is a strongly convex problem
- z_k and (v_k, ε_k) are computed by an accelerated composite gradient (ACG) method applied to (P_c^k) in at most

$$\mathcal{O}\left(\left\lceil \sqrt{\frac{M_c}{m}} \right\rceil\right) \quad \text{iterations}$$

Obs: Each ACG iteration requires one or two evaluations of the resolvent of h, i.e., exact solution of

$$\min\{a^T z + h(z) + \theta\|z\|^2\}$$

(0) (beginning of phase I) Let $c > 0$, $z_0 \in \operatorname{dom} h$, $\sigma \in (0,1)$ and $\bar{\rho} > 0$ be given, and set $\lambda = 1/(2m)$ and $k = 1$

(1) call an ACG variant started from z_{k-1} to approximately solve (P_c^k), i.e., to obtain z_k and (v_k, ε_k) such that

$$v_k \in \partial_{\varepsilon_k}\left(\lambda(f_c + h) + \frac{1}{2}\|\cdot - z_{k-1}\|^2\right)(z_k)$$

$$\|v_k\|^2 + 2\varepsilon_k \leq \sigma\|z_{k-1} - z_k + v_k\|^2$$

(2) if $\|z_{k-1} - z_k + v_k\| > \lambda\bar{\rho}/10$, then $k \leftarrow k+1$ and go to (1); otherwise, go to (3) (end of phase I)

(3) (phase II) restart the last call to the ACG variant in step 1 to find $\tilde{z}$ and $(\tilde{v}, \tilde{\varepsilon})$ satisfying

$$\|z_{k-1} - \tilde{z} + \tilde{v}\| \leq \frac{\lambda\bar{\rho}}{2}, \quad \tilde{\varepsilon} \leq \lambda\frac{\bar{\rho}^2}{32(M_c + 2m)}$$

and then refine $(\tilde{z}; \tilde{v}, \tilde{\varepsilon})$ to obtain a $\bar{\rho}$-approximate solution $(\bar{z}; \bar{v})$ for (P_c).

(0) (beginning of phase I) Let $c > 0$, $z_0 \in \operatorname{dom} h$, $\sigma \in (0,1)$ and $\bar{\rho} > 0$ be given, and set $\lambda = 1/(2m)$ and $k = 1$

(1) call an ACG variant started from z_{k-1} to approximately solve (P_c^k), i.e., to obtain z_k and (v_k, ε_k) such that

$$v_k \in \partial_{\varepsilon_k}\left(\lambda(f_c + h) + \frac{1}{2}\|\cdot - z_{k-1}\|^2\right)(z_k)$$
$$\|v_k\|^2 + 2\varepsilon_k \leq \sigma\|z_{k-1} - z_k + v_k\|^2$$

(2) if $\|z_{k-1} - z_k + v_k\| > \lambda\bar{\rho}/10$, then $k \leftarrow k+1$ and go to (1); otherwise, go to (3) (end of phase I)

(3) (phase II) restart the last call to the ACG variant in step 1 to find $\tilde{z}$ and $(\tilde{v}, \tilde{\varepsilon})$ satisfying

$$\|z_{k-1} - \tilde{z} + \tilde{v}\| \leq \frac{\lambda\bar{\rho}}{2}, \quad \tilde{\varepsilon} \leq \lambda\frac{\bar{\rho}^2}{32(M_c + 2m)}$$

and then refine $(\tilde{z}; \tilde{v}, \tilde{\varepsilon})$ to obtain a $\bar{\rho}$-approximate solution $(\bar{z}; \bar{v})$ for (P_c).

Theorem

The total number of ACG iterations is

$$\mathcal{O}\left(\frac{\sqrt{M_c m}}{\bar{\rho}^2}\min\left\{\Delta_0^*(c), mD_h^2\right\} + \sqrt{\frac{M_c}{m}}\log\left(\max\left\{1, \frac{M_c}{m\sqrt{m}}\right\}\right)\right)$$

where D_h is the diameter of $\operatorname{dom} h$ *and* $\Delta_0^*(c) = \phi_c(z_0) - \phi_c^*$

Hence, the complexity of the AIPP method is

$$\mathcal{O}\left(\sqrt{M_c m}\,\frac{mD_h^2}{\bar{\rho}^2}\right)$$

while that of the CG or Ghadimi-Lan's AG is

$$\mathcal{O}\left(M_c\,\frac{mD_h^2}{\bar{\rho}^2}\right)$$

Theorem

The total number of ACG iterations is

$$\mathcal{O}\left(\frac{\sqrt{M_c m}}{\bar{\rho}^2}\min\left\{\Delta_0^*(c), mD_h^2\right\} + \sqrt{\frac{M_c}{m}}\log\left(\max\left\{1, \frac{M_c}{m\sqrt{m}}\right\}\right)\right)$$

where D_h is the diameter of $\mathrm{dom}\, h$ *and* $\Delta_0^*(c) = \phi_c(z_0) - \phi_c^*$

Hence, the complexity of the AIPP method is

$$\mathcal{O}\left(\sqrt{M_c m}\,\frac{mD_h^2}{\bar{\rho}^2}\right)$$

while that of the CG or Ghadimi-Lan's AG is

$$\mathcal{O}\left(M_c\,\frac{mD_h^2}{\bar{\rho}^2}\right)$$

Complexity of the quadratic penalty AIPP: Recall that a sufficient condition for attaining $\|A\bar{z} - b\| \leq \bar{\eta}$ is that $c \geq R/(\bar{\eta})^2$ where

$$R := 2\Delta_\phi^* + 2\bar{\rho} D_h + L_f D_h^2$$

Theorem

The quadratic penalty AIPP method performs a total of at most

$$\mathcal{O}\left(\frac{\sqrt{R}\|A\| L_f^{3/2} D_h^2}{\bar{\rho}^2 \bar{\eta}} + \frac{L_f^2 D_h^2}{\bar{\rho}^2} \right)$$

ACG iterations to find a $(\bar{\rho}, \bar{\eta})$-approximate solution of (P).

Hence, the complexity of the penalty AIPP is $\mathcal{O}\left(1/(\bar{\rho}^2 \bar{\eta})\right)$

Complexity of the quadratic penalty AIPP: Recall that a sufficient condition for attaining $\|A\bar{z} - b\| \leq \bar{\eta}$ is that $c \geq R/(\bar{\eta})^2$ where

$$R := 2\Delta_\phi^* + 2\bar{\rho}D_h + L_f D_h^2$$

Theorem

The quadratic penalty AIPP method performs a total of at most

$$\mathcal{O}\left(\frac{\sqrt{R}\|A\| L_f^{3/2} D_h^2}{\bar{\rho}^2 \bar{\eta}} + \frac{L_f^2 D_h^2}{\bar{\rho}^2}\right)$$

ACG iterations to find a $(\bar{\rho}, \bar{\eta})$-approximate solution of (P).

Hence, the complexity of the penalty AIPP is $\mathcal{O}\left(1/(\bar{\rho}^2\bar{\eta})\right)$

Computational Results

- AIPP was benchmarked against Ghadimi-Lan's AG method
- The nonconvex optimization problem tested was

$$\min_{z \in S^n_+} \left\{ f(z) := -\frac{\xi}{2}\|D\mathcal{B}(z)\|^2 + \frac{\tau}{2}\|\mathcal{A}(z) - b\|^2 : z \in P_n \right\}$$

  where P_n is the unit spectraplex, i.e.,

$$P_n := \{z \in S^n_+ : \text{tr}(z) = 1\}$$

  $\mathcal{A} : \mathcal{S}^n \to \mathbb{R}^n$, $\mathcal{B} : \mathcal{S}^n \to \mathbb{R}^l$ are linear operators, D is a positive diagonal matrix, $b \in \mathbb{R}^n$
- Values in A, B and b were sampled from the $\mathcal{U}[0, 1]$ distribution at sparsity level d and values for D were sampled from $\mathcal{U}[0, 1000]$ distribution

Results for composite unconstrained problems
($l = 50,\ n = 200,\ d = 0.025,\ \bar{\rho} = 10^{-7}$)

Size		$\bar{f}$	Iteration Count		Runtime	
M	m		AG	AIPP	AG	AIPP
1000000	1	3.84E+01	7039	**1760**	517.72	**92.68**
100000	1	3.82E+00	7041	**1564**	512.92	**83.85**
10000	1	3.67E-01	7064	**2770**	511.87	**142.52**
1000	1	2.05E-02	7305	**3087**	532.94	**159.03**
100	1	-1.74E-02	8670	**2258**	807.36	**146.33**
10	1	-3.65E-02	5790	**1561**	793.71	**141.38**

Results for composite unconstrained problems ($l = 50, n = 1000, d = 0.001, \bar{\rho} = 10^{-7}$)						
Size		$\bar{f}$	Iteration Count		Runtime	
M	m		AG	AIPP	AG	AIPP
1000000	1	2.98E+03	2351	**883**	3625.82	**923.69**
100000	1	2.98E+02	2351	**668**	3820.18	**713.07**
10000	1	2.97E+01	2347	**608**	3793.74	**660.79**
1000	1	2.91E+00	2312	**588**	3625.51	**626.42**
100	1	2.28E-01	1969	**582**	3076.48	**618.78**
10	1	-6.80E-02	603	**179**	1034.78	**204.82**

- QP-AIPP was benchmarked against a penalty version of G-L's AG method
- The linearly constrained nonconvex optimization problem tested was

$$\min_{z \in S^n_+} \left\{ f(z) = -\frac{\xi}{2}\|D\mathcal{B}(z)\|^2 : z \in P_n,\ \mathcal{A}(z) = b \right\}$$

where $\mathcal{A} : \mathcal{S}^n \to \mathbb{R}^n$, $\mathcal{B} : \mathcal{S}^n \to \mathbb{R}^l$ and D were generated as before.
- b was chosen so as to make I/n feasible

Results for composite linearly constrained problems ($l = 50,\ n = 20,\ d = 1,\ \bar{\rho} = 10^{-3},\ \bar{\eta} = 10^{-6}$)					
L_f	$\bar{F}$	Iteration Count		Runtime	
		AG	AIPP	AG	AIPP
1000000	-1.49E+03	110415	**17673**	169.22	**30.11**
100000	-1.49E+02	110414	**17673**	169.67	**30.26**
10000	-1.49E+01	110386	**17673**	170.17	**30.02**
1000	-1.49E+00	110135	**17673**	169.15	**30.00**
100	-1.49E-01	107942	**17393**	183.78	**31.56**
10	-1.49E-02	96776	**16499**	170.62	**30.44**

Results for composite linearly constrained problems ($l = 50$, $n = 100$, $d = 0.0015$, $\bar{\rho} = 10^{-3}$, $\bar{\eta} = 10^{-6}$)					
L_f	$\bar{f}$	Iteration Count		Runtime	
		AG	AIPP	AG	AIPP
1000000	-5.22E+04	33330	**6426**	159.30	**27.96**
100000	-5.22E+03	33290	**5405**	173.25	**24.16**
10000	-5.22E+02	32897	**3897**	157.55	**18.58**
1000	-5.22E+01	29611	**8321**	144.01	**36.31**
100	-5.22E+00	17289	**7042**	83.07	**31.80**
10	-5.22E-01	5917	**4644**	29.93	**21.36**

Implementation Remarks

- Even though Phase II is theoretically needed, it was never needed for solving the instances in our test.
- λ_k has been chosen aggressively in all instances, i.e., $\lambda_k > 1/m$.

Additional results

$$p_* := \min_x \{f(x) + h(x) : Ax = b\}$$

where now

$$f(x) = \max_{y \in Y} \Phi(x, y)$$

Assume that Y is a closed convex set whose diameter

$$D_y := \sup_{y, y' \in Y} \|y - y'\|$$

is finite

It is also assumed that

- $\Phi(x,\cdot)$ is concave on Y for every $x \in X$;
- $\Phi(\cdot,y)$ is continuously differentiable on $\mathrm{dom}\, h$ for every $y \in Y$;
- there exist scalars $(L_x, L_y) \in \mathbb{R}^2_{++}$, and $m \in (0, L_x]$ such that

$$\Phi(x',y) - \left[\Phi(x,y) + \left\langle \nabla_x \Phi(x,y), x'-x \right\rangle_{\mathcal{X}}\right] \geq -\frac{m}{2}\|x-x'\|^2_{\mathcal{X}}$$
$$\left\|\nabla_x \Phi(x,y) - \nabla_x \Phi(x',y')\right\|_{\mathcal{X}} \leq L_x\|x-x'\|_{\mathcal{X}} + L_y\|y-y'\|_{\mathcal{Y}}$$

for every $x, x' \in \mathrm{dom}\, h$ and $y, y' \in Y$.

f can now be nonsmooth and nonconvex but it can easily be approximated by a smooth nonconvex function, namely,

$$f_{\xi}(x) := \max_{y \in \mathcal{Y}} \left\{ \Phi_{\xi}(x,y) := \Phi(x,y) - \frac{1}{2\xi}\|y - y_0\|_{\mathcal{Y}}^2 : y \in Y \right\}$$

where $y_0 \in Y$ and $\xi > 0$

Similar to the one used by Nesterov in his smooth approximation acceleration scheme!

Applying the penalty AIPP method to

$$\min_x \{f_{\xi}(x) + h(x) : Ax = b\}$$

for some well-chosen ξ, yields a quintuple $(\bar{u}, \bar{v}, \bar{x}, \bar{y}, \bar{w})$ satisfying

$$\begin{pmatrix} \bar{u} \\ \bar{v} \end{pmatrix} \in \begin{pmatrix} \nabla_x \Phi(\bar{x}, \bar{y}) + \mathcal{A}^* \bar{w} \\ 0 \end{pmatrix} + \begin{pmatrix} \partial h(\bar{x}) \\ [-\Phi(\bar{x}, \cdot)]\,(\bar{y}) \end{pmatrix}$$

$$\|\bar{u}\|_{\mathcal{X}}^* \le \rho_x, \quad \|\bar{v}\|_{\mathcal{Y}}^* \le \rho_y, \quad \|\mathcal{A}\bar{x} - b\|_{\mathcal{U}} \le \eta.$$

in a total number of ACG iterations bounded by

$$\mathcal{O}\left(m^{3/2} D_h^2 \left[\frac{L_x^{1/2}}{\rho_x^2} + \frac{L_y D_y^{1/2}}{\rho_y^{1/2} \rho_x^2} + \frac{m^{1/2} \|\mathcal{A}\| D_h}{\eta \rho_x^2}\right]\right)$$

The complexity is still $\mathcal{O}(1/\eta^3)$ under the assumption that $\rho_x = \rho_y = \eta$.

Concluding Remarks

- We have presented the quadratic penalty AIPP method for "solving" a linearly constrained composite smooth nonconvex program and have shown that its associated bound is

$$\mathcal{O}\left(\frac{1}{\bar{\rho}^2\bar{\eta}}\right)$$

  If instead either the PG or AG method were used to solve subproblems (P_c), the bound would be $\mathcal{O}\left(1/[\bar{\rho}^2\bar{\eta}^2]\right)$
- We have also argued that the above complexity 'remains the same' in the context of linearly constrained composite nonsmooth nonconvex min-max programs.

THE END

Thanks!

Example

On first slide.

Example

On second slide.

Example

On first slide.

Example

On second slide.

Theorem

On first slide.

Corollary

On second slide.

Theorem

On first slide.

Corollary

On second slide.

Theorem

In left column.

Corollary

In right column.
New line

Theorem

In left column.

Corollary

In right column.
New line

- You can control text size using special keywords
  Text Text Text Text Text Text Text Text Text Text
- You can also specify the text size directly
  This sentence has 0.5 centimeters of space between lines.
  This sentence is 1x the size of normal sentences
  This sentence is 2x the size of normal sentences

- You can control spacing between bullet points with the vspace* command
- This bullet point will have addition vertical spacing after it
- This bullet point will have less vertical spacing after it
- This is the last item

- The first main message of your talk in one or two lines.
- The second main message of your talk in one or two lines.
- Perhaps a third message, but not more than that.

- Outlook
  - What we have not done yet.
  - Even more stuff.

A. Author.
Handbook of Everything.
Some Press, 1990.

S. Someone.
On this and that.
Journal on This and That. 2(1):50–100, 2000.