Primal-Dual Hybrid Gradient (PDHG) Method for LP

Yinyu Ye

Stanford University, MS&E and ICME

http://www.stanford.edu/~yyye

(Chapters 8, 14)

$\boxed{2008}$: IPM

10 $\boxed{2018}$ = 140

10^{-2} $\boxed{2022}$ = 3 hours

10^{-6} PDHG GPU $\boxed{2023}$ = 17 minutes

Features of FOM

FOM requires only matrix-vector multiplication:

1. Matrix-vector multiplication is very suitable for utilizing modern hardware and distributed computation
2. Matrix factorization free. The memory usage is relatively low.

FOM is particularly good at solving large instances

With help of GPU, recent research shows PDHG is fast when the problem becomes large, see

- *"cuPDLP. jl: A GPU implementation of restarted primal-dual hybrid gradient for linear programming in Julia", Lu/Yang, arXiv:2311.12180, 2023*.
- *"cuPDLP-C: A Strengthened Implementation of cuPDLP for Linear Programming by C language," Lu et al. arXiv:2312.14832, 2023*.

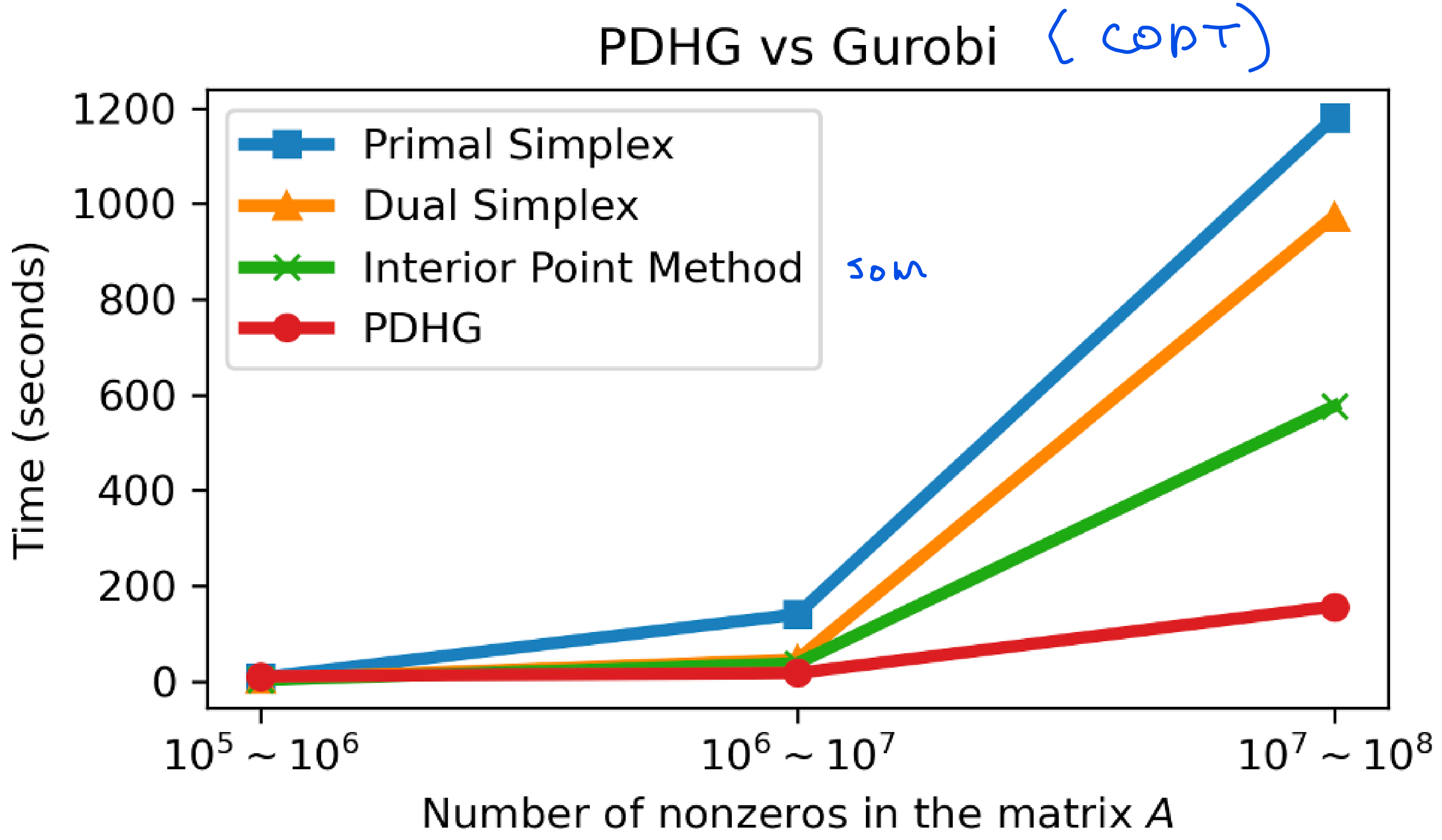


Figure 1: Geometric average runtime on problems from LP relaxations of MIPLIB 2017 (termination tolerance 10^{-4})

Primal-Dual Hybrid Gradient (PDHG) Method

For the standard-form LP:

$$\text{(Primal)} \quad \min\ c^\top x \quad \text{s.t. } Ax = b\ ,\ x \geq 0$$

Instead of directly dealing with the original problem, we deal with the primal-dual form to avoid doing expensive projections onto the feasible set $\{x : Ax = b, x \geq 0\}$.

$$\text{(Dual)} \quad \max\ b^\top y \quad \text{s.t. } A^\top y \leq c$$

$c^T x - y^T (Ax - b)$

$\Downarrow$

Via Lagrangian:

$$\text{(Primal-dual)} \quad \min_{x \geq 0} \max_{y}\ L(x, y) := c^\top x + b^\top y - x^\top A^\top y$$

Saddle point: The $(x^\star, y^\star)$ with $x^\star \geq 0$ is a saddle point of $L(x, y)$ if for any $x \geq 0$ and y:
$L(x^\star, y) \leq L(x^\star, y^\star) \leq L(x, y^\star)$

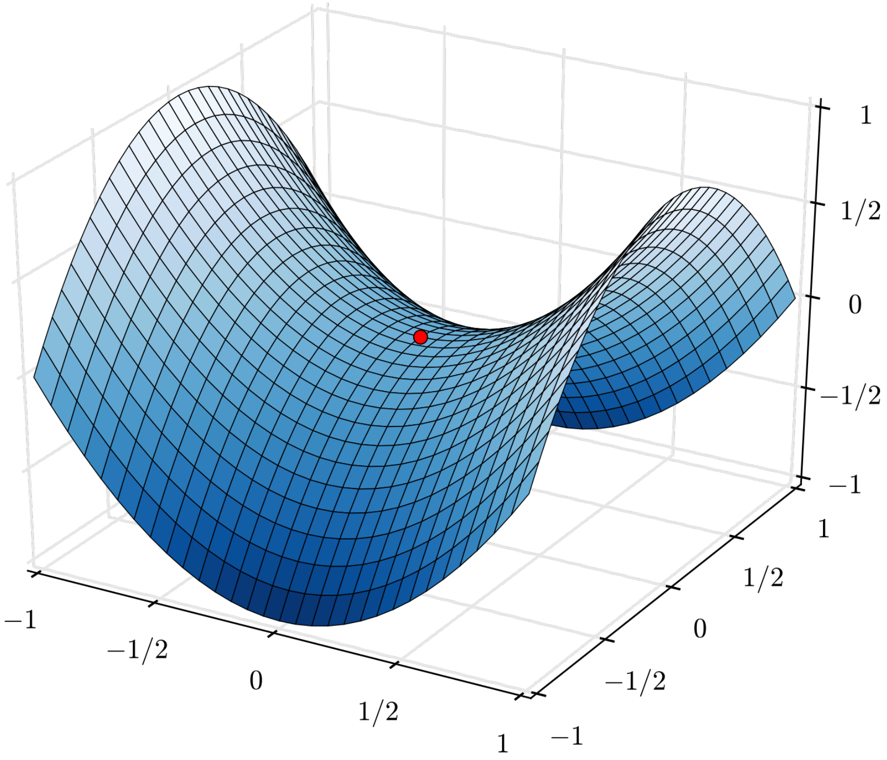


Figure 2: Saddle point

Classical Gradient Descent Ascent:

$$x^{k+1} = \text{Proj}_{x>0}\underline{\left(x^k - \eta\overset{\nabla L_x}{\left(c - A^\top y\right)}\right)} = \underline{\left(x^k - \eta\left(c - A^\top y\right)\right)^+}$$

$$y^{k+1} = y^k + \eta\left(b - A\underset{\boxed{x^{k+1}}}{x^k}\right)$$

$$\begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix} = \underline{\begin{pmatrix} 1 & \eta \\ -\eta & 1 \end{pmatrix} \begin{pmatrix} x^k \\ y^k \end{pmatrix}}$$

Gradient Descent Ascent May Not Converge

For example, in the saddle point problem $\max_x \min_y \; xy$, we have closed form of each iteration:

$$\begin{pmatrix} x^k \\ y^k \end{pmatrix} = \begin{pmatrix} 1 & \eta \\ -\eta & 1 \end{pmatrix}^k \begin{pmatrix} x^0 \\ y^0 \end{pmatrix}$$

L $1+\eta^2$

The determinant of $\begin{pmatrix} 1 & \eta \\ -\eta & 1 \end{pmatrix}$ is always strictly larger than 1 no matter how small η is, which means this approach never converges.

Primal-Dual Hybrid Gradient (PDHG) Method

$$x^{k+1} = \mathrm{Proj}_{x \geq 0}\left(x^k - \eta\left(c - A^\top \overset{k}{y}\right)\right)$$

$$y^{k+1} = y^k + \eta\left(b - A(2x^{k+1} - x^k)\right)$$

$$b - Ax^k - A(x^{k+1} - x^k)$$

The update of y^k uses the idea of momentum.

For the small saddle point problem $\max_x \min_y \; xy$, we also have closed form of each iteration:

$$\begin{pmatrix} x^k \\ y^k \end{pmatrix} = \begin{pmatrix} 1 & \eta \\ -\eta & 1 - 2\eta^2 \end{pmatrix}^k \begin{pmatrix} x^0 \\ y^0 \end{pmatrix}$$

$$1 - 2\eta^2 + \eta^2$$
$$= 1 - \eta^2$$

The modulus of eigenvalues of $\begin{pmatrix} 1 & \eta \\ -\eta & 1 - 2\eta^2 \end{pmatrix}$ are smaller than 1 if $\eta < 1$, so PDHG converges.

"A first-order primal-dual algorithm for convex problems with applications to imaging," Chambolle/Pock, Journal of mathematical imaging and vision, 2011.

Metrics of LP

For the original LP problem, usually we use the following metrics for termination:

$x \geq 0$

1. Primal feasibility $\|Ax - b\|_2 \leq \varepsilon$
2. Dual feasibility $\|(c - A^\top y)^-\|_2 \leq \varepsilon$
3. Primal-dual gap $|c^\top x - b^\top y| \leq \varepsilon$

However, for a general LP $\min_{x \in X} \max_{y \in Y} L(x, y)$, the duality gap measurement at $z = (x, y)$ is

$$\max_{\hat{y} \in Y} L(x, \hat{y}) - \min_{\hat{x} \in X} L(\hat{x}, y) = \max_{\hat{z} := (\hat{x}, \hat{y}) \in Z := X \times Y} \{L(x, \hat{y}) - L(\hat{x}, y)\}$$

$\left\| \begin{matrix} x \\ y \end{matrix} \right\|$

Obviously, when $X = R^n_+,\ Y = R^m$, the duality gap is $+\infty$ almost always.

Normalized duality gap at $z = (x, y)$ with radius r:

$$\rho(r; z) := \frac{1}{r} \cdot \max_{\hat{z} \in \{\hat{z} \in Z : \|z - \hat{z}\| \leq r\}} \{L(x, \hat{y}) - L(\hat{x}, y)\}$$

Key Properties Normalized Duality Gap

1. No longer equal to ∞.

2. Computing or approximating $\rho(r; z)$ is very cheap.

3. For any r, $\rho(r; z) = 0$ if $z \in Z^\star$ and $\rho(r; z) > 0$ if $z \notin Z^\star$.

4. For any r, z, $\rho(r; z) \geq \|Ax - b\|$ and $\rho(r; z) \geq \|(c - A^\top y)^-\|$. The normalized duality gap is always an upper bound of the primal and dual infeasibility.

5. $\rho_r(z) \geq \min\left\{\frac{1}{\|z\|}, \frac{1}{r}\right\}\left(c^\top x - b^\top y\right)^+$. The normalized duality gap is a good upper bound of the duality gap if r and $\|z\|$ are not too large.

“Faster first-order primal-dual methods for linear programming using restarts and sharpness,” Applegate et al.,, Mathematical Programming, 2023.

Ergodic Convergence of PDHG

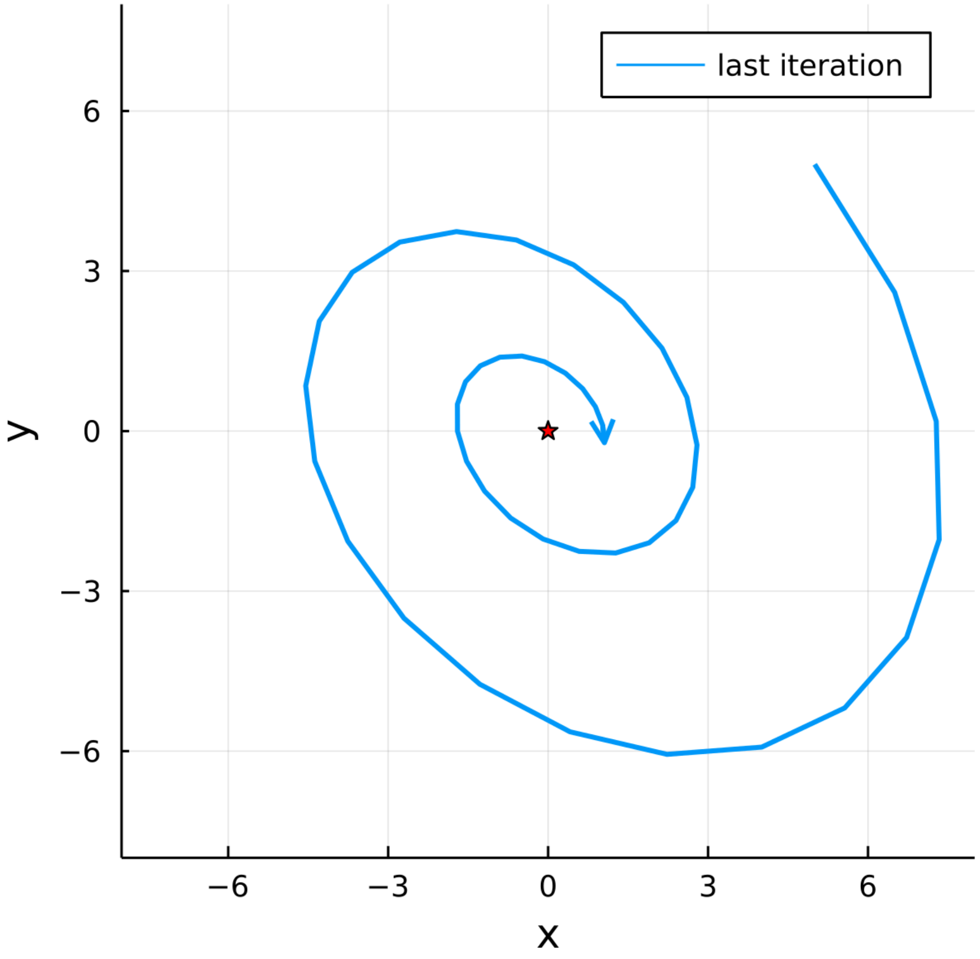
Theorem 1 (Applegate et al. 2023) *Consider the PDHG iterates $\left(x^k, y^k\right)$ with initial solution $\left(x^0, y^0\right)$, it holds for any $x \geq 0, y$ and $\eta \leq \frac{1}{\|A\|_2}$ that*

1. $\rho_{\|\bar{z}^k - z^0\|}\left(\bar{z}^k\right) \leq \frac{4\|\bar{z}^k - z^0\|}{k}$ $\leq \varepsilon$
2. $\left\|\bar{z}^k - z^0\right\| \leq 2 \cdot \text{dist}\left(z^0, Z^\star\right)$

Here $\bar{z}^k := (\bar{x}^k, \bar{y}^k) := \frac{1}{k}\sum_{i=1}^k (x^i, y^i)$, *and* $z^0 := (x^0, y^0)$.

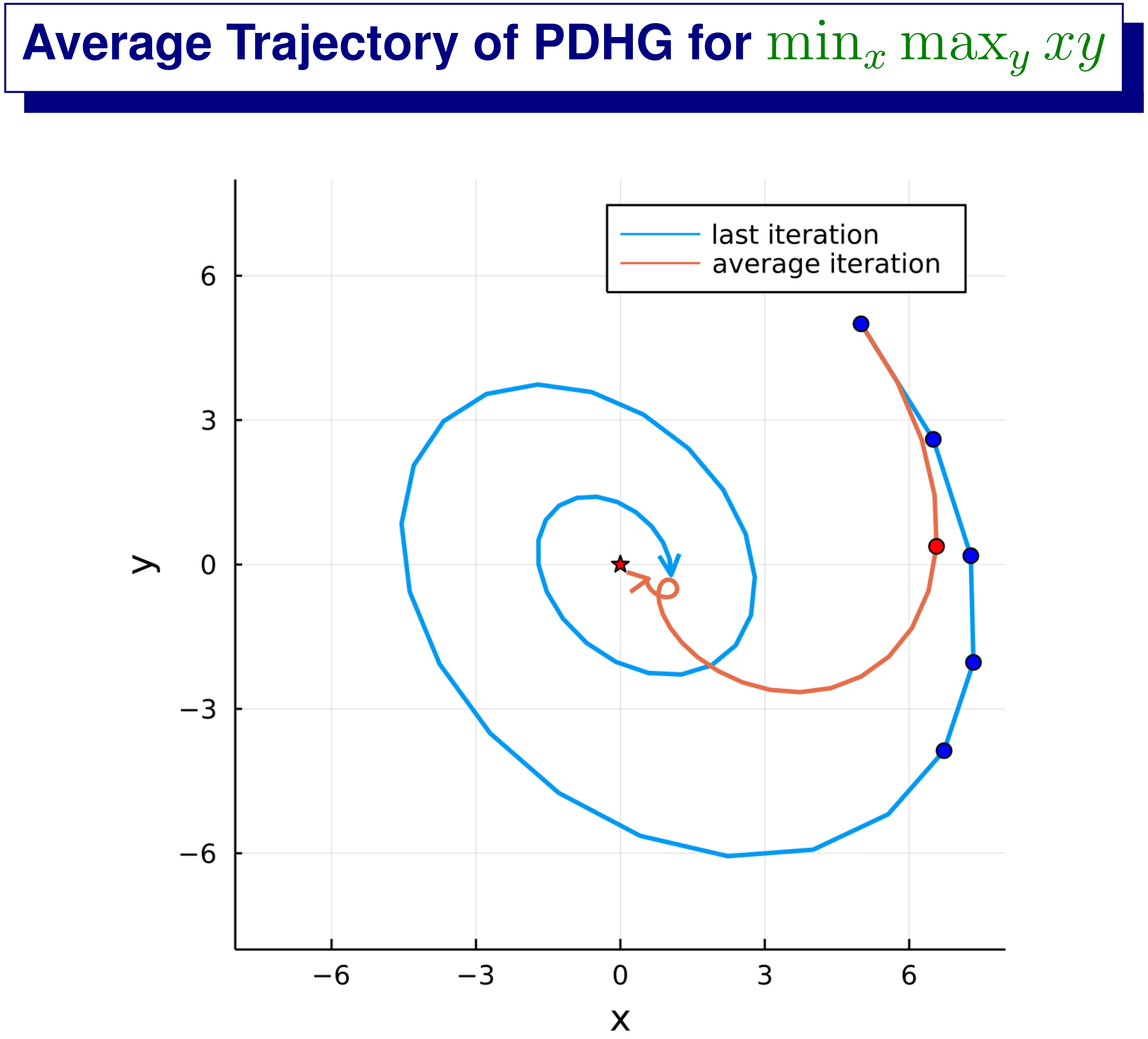
1. Here the norm $\|\cdot\|$ is the norm induced by $\begin{pmatrix} \mathbf{I} & -\eta A^\top \\ -\eta A & \mathbf{I} \end{pmatrix}$
2. η needs to be small to guarantee the induced "norm" is a norm.
3. Item 2 means $\|\bar{z}^k - z^0\|$, $\|\bar{z}^k\|$ are never too large. Therefore, item 1 actually presents an "$O(1/\epsilon)$" convergence rate of the standard metrics (primal and dual infeasibility, duality gap), which rate is typical for FOM but LP applications we usually expect faster rate.

Trajectory of PDHG for $\min_x \max_y xy$



$\begin{pmatrix} x^k \\ y^k \end{pmatrix}$

Figure 3: Final iterates of PDHG . Figure courtesy Lu.



Average Trajectory of PDHG for $\min_x \max_y xy$

Figure 4: Average iterates of PDHG . Figure courtesy Lu.

Restart Trajectory of PDHG for $\min_x \max_y xy$

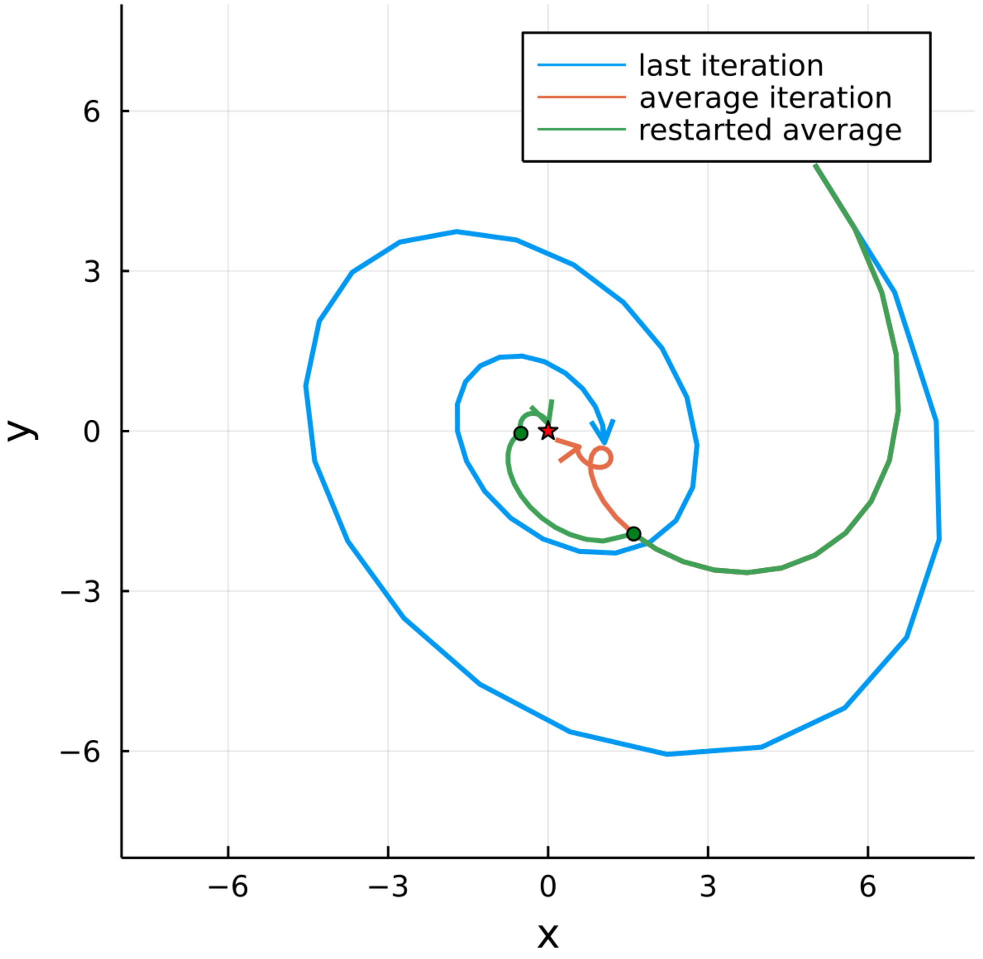


Figure 5: Restarted iterates of PDHG . Figure courtesy Lu.

PHDG with Restarts ← PDLP

```
Algorithm 1: Restarted-PDHG
Input: Initial iterate $z^{0,0} := (x^{0,0}, y^{0,0})$, $n \leftarrow 0$ ;
repeat
    initialize the inner loop: inner loop counter $k \leftarrow 0$ ;
    repeat
        conduct one step of PDHG: $z^{n,k+1} \leftarrow \text{PDHG}(z^{n,k})$ ;
        compute the average iterate in the inner loop.
          $\bar{z}^{n,k+1} \leftarrow \frac{1}{k+1}\sum_{i=1}^{k+1} z^{n,i}$ ;
        $k \leftarrow k+1$ ;
    until $\bar{z}^{n,k}$ satisfies some (verifiable) restart condition ;
    restart the outer loop: $z^{n+1,0} \leftarrow \bar{z}^{n,k}$, $n \leftarrow n+1$ ;
until $z^{n,0}$ satisfies some convergence condition ;
Output: $z^{n,0}$ $( = (x^{n,0}, y^{n,0}))$
```

Restart Scheme: Fixed frequency restart, adaptive restart ...

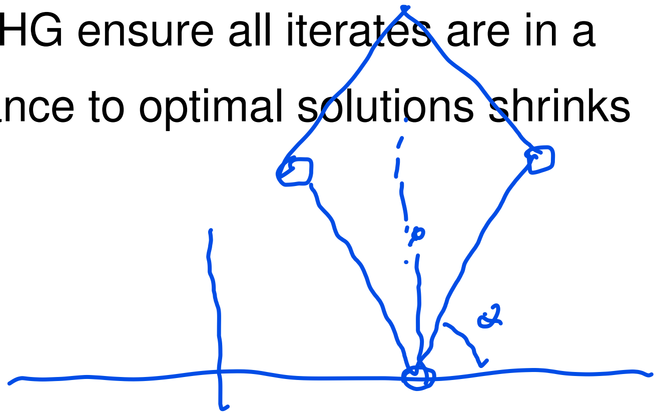
Analyses of PDHG with Restarts I

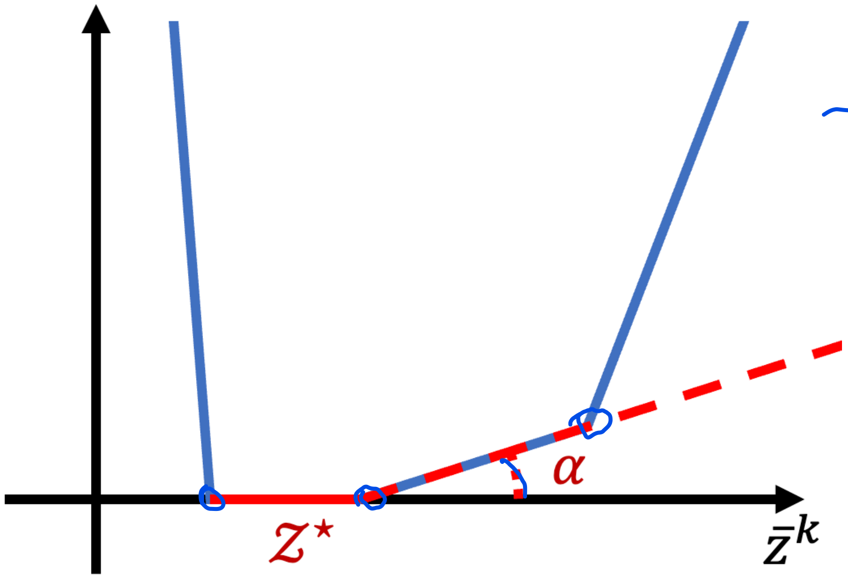
Sharpness of $\rho(r; z)$ The primal-dual LP problem is α-sharp on $S \subseteq Z$ if it holds for all $r \in (0, \text{diam}(S)]$ and $z \in S$ that

Linear

$$\rho(r; z) \geq \alpha \cdot \text{dist}(z, Z^\star)$$

In any given bounded region S, the $\rho(r; z)$ is sharp. The updates of PDHG ensure all iterates are in a bounded region (Theorem 1), with which the sharpness ensures the distance to optimal solutions shrinks one half every $\left\lceil \frac{8\|A\|}{\alpha} \right\rceil$ iterations.





Theorem 2 (Applegate et al. 2023) *All iterates of PDHG are in* $B_{4\cdot \textit{dist}(z^0, Z^\star)}(z^0)$*. And LP is* α*-sharp with*

$$\alpha \geq \frac{1}{H(K)\,(1 + 4 \cdot \textit{dist}\,(z^0, Z^*))}$$

where $H(K)$ *is the Hoffman constant of the KKT system of LP.*

$H(K)$ *is the Hoffman constant of the inequality set* $\{Kz \geq h\}$*, where*

$$K := \begin{pmatrix} I & 0 \\ -A & 0 \\ A & 0 \\ 0 & -A^T \\ -c^T & b^T \end{pmatrix}, \quad h := \begin{pmatrix} 0 \\ -b \\ b \\ -c \\ 0 \end{pmatrix}$$

$H(K)$ is hard to compute or analyze, but *"Computational Guarantees for Restarted PDHG for LP based on "Limiting Error Ratios" and LP Sharpness," Xiong/Freund, arXiv preprint arXiv:2312.14774* shows that sharpness α is connected with the geometry and stability of LP.

Restart Strategies I

Fixed Frequency Restart:

Suppose α and $\|A\|$ are known to the user, restart the algorithm every

$$\left\lceil \frac{8\|A\|}{\alpha} \right\rceil$$

iterations

Adaptive Restart:

Restart the algorithm whenever the normalized duality gap has sufficient decay

$$\rho\left(\left\|\bar{z}^{n,t} - z^{n,0}\right\|; \bar{z}^{n,t}\right) \leq 0.5 \cdot \rho\left(\left\|z^{n,0} - z^{n-1,0}\right\|; z^{n,0}\right).$$

Restart Strategies II

With either of the above restart schemes, we have the following linear convergence guarantee:

Theorem 3 (Applegate et al. 2023) *The restarted PDHG finds a solution* z *such that* dist$(z, Z^\star) \leq \varepsilon$ *within*

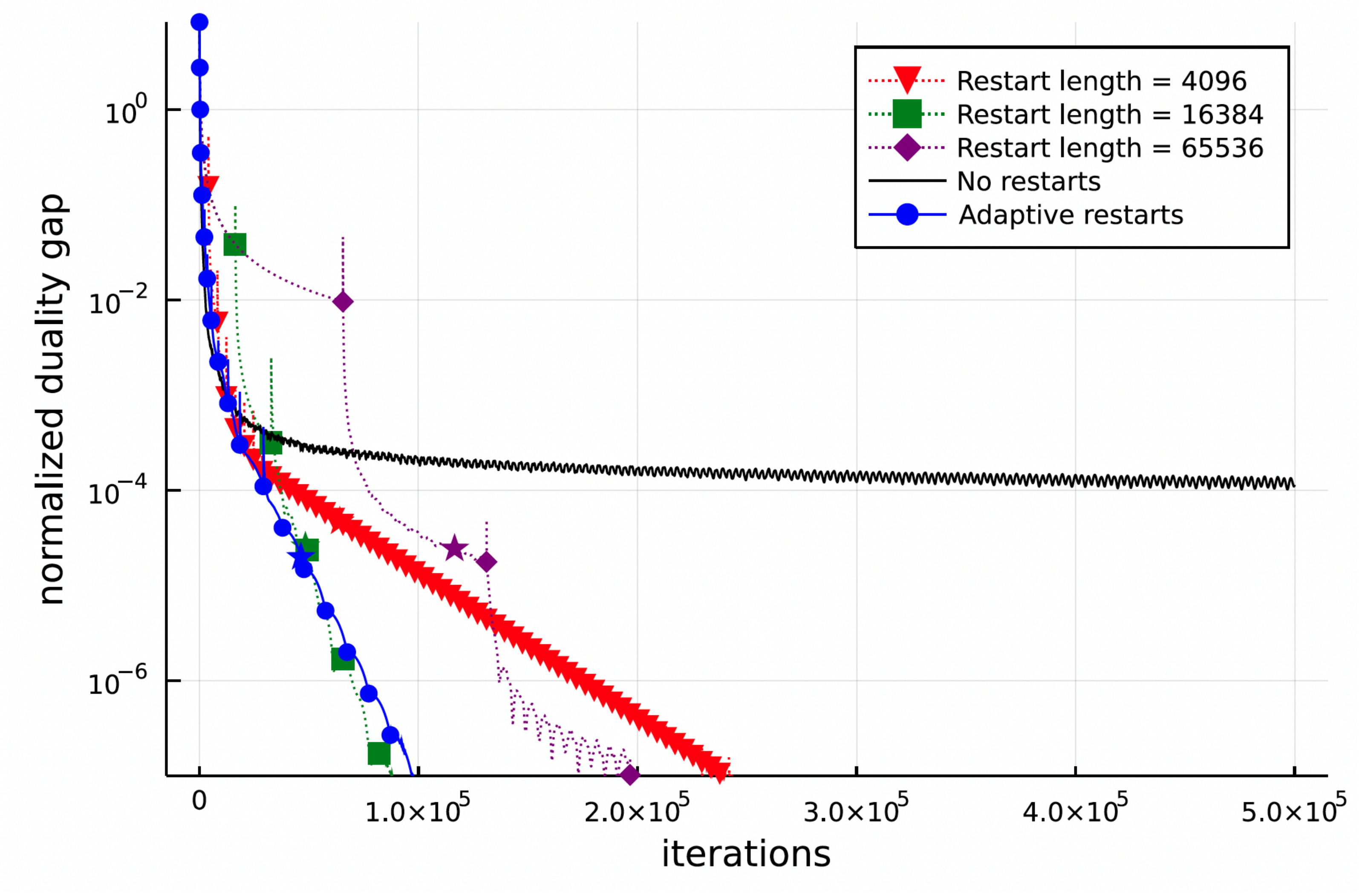
$$O\left(\frac{\|A\|}{\alpha} \cdot \log\left(\frac{1}{\varepsilon}\right)\right)$$

iterations.

An important research question: How to explain and understand α? And how to improve α? [see *Xiong/Freund arXiv:2312.14774, Lu/Yang arXiv:2307.03664, Hinders arXiv:2309.03988*]

Numerical Experiments

LP relaxation of a quadratic assignment problem:



Other Enhancements PDHG I

Infeasibility detection

Use PDHG to find the certificates of infeasibility:

1. The primal problem is infeasible if, and only if, there exists $y \in R^m$, s.t. $b^\top y > 0$ and $A^\top y \leq 0$.
2. The dual problem is infeasible if, and only if, there exists $x \in R^m$, s.t. $c^\top x < 0$, $Ax = 0$ and $x \geq 0$.

Adaptive step sizes

Different step-size for primal and for dual, aiming to balance the primal infeasibility and dual infeasibility.

Presolve: Apply transformations such as detecting inconsistent bounds, removing empty rows and columns, and removing fixed variables ...

Other Enhancements PDHG II

Diagonal preconditioning

Solve the preconditioner problem

$$\min_{x \geq 0} (D_2 c)^\top x \quad \text{s.t.} (D_1 A D_2) x = D_1 b$$

instead. And then $D_2 x^\star$ recovers an optimal solution for the original problem. The preconditioner is supposed to improve the condition number of the matrix.

Addressing general form LP

Directly address the general form LP instead of the standard form, to work on the better sharpness.

$$\begin{array}{ll} \min_{x \in R^n} & c^\top x \\ \text{s.t.} & Gx \geq h, \ Ax = b, \ l \leq x \leq u \end{array}$$