Journal of Optimization Theory and Applications (2024) 201:280–312
https://doi.org/10.1007/s10957-024-02394-6

A Universal Accelerated Primal–Dual Method for Convex Optimization Problems

Hao Luo[1,2]

Received: 8 December 2022 / Accepted: 22 January 2024 / Published online: 1 March 2024

Abstract
This work presents a universal accelerated primal–dual method for affinely constrained convex optimization problems. It can handle both Lipschitz and Hölder gradients but does not need to know the smoothness level of the objective function. In line search part, it uses dynamically decreasing parameters and produces approximate Lipschitz constant with moderate magnitude. In addition, based on a suitable discrete Lyapunov function and tight decay estimates of some differential/difference inequalities, a universal optimal mixed-type convergence rate is established. Some numerical tests are provided to confirm the efficiency of the proposed method.

Keywords Convex optimization · Primal–dual method · Mixed-type estimate · Optimal complexity · Bregman divergence · Lyapunov function

Mathematics Subject Classification 65B99 · 68Q25 · 90C25

1 Introduction

Consider the minimization problem

$$\min_{x\in Q\cap\Omega} f(x) := h(x) + g(x), \tag{1}$$

Communicated by Olivier Fercoq.

This work was supported by the Foundation of Chongqing Normal University (Grant No. 202210000161) and the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJZD-K202300505).

✉ Hao Luo
luohao@cqnu.edu.cn

1 National Center for Applied Mathematics in Chongqing, Chongqing Normal University, Chongqing 401331, China

2 Chongqing Research Institute of Big Data, Peking University, Chongqing 401121, China

where $Q \subset \mathbb{R}^n$ is a simple closed convex subset, $\Omega := \{x \in \mathbb{R}^n : Ax = b\}$ is an affine set with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, and $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is properly closed and convex, with smooth part h and nonsmooth (simple) part g. The model problem (1) arises from many practical applications, such as compressed sensing [6], image processing [8] and decentralized distributed optimization [4].

In the literature, existing algorithms for solving (1) mainly include Bregman iteration [5, 29, 67], quadratic penalty method [35], augmented Lagrangian method (ALM) [26–28, 33, 41, 56], and alternating direction method of multipliers [22, 36, 40, 50, 53, 57, 60, 61, 65]. Generally speaking, these methods have sublinear rate $\mathcal{O}(1/k)$ for convex problems and can be further accelerated to $\mathcal{O}(1/k^2)$ for (partially) strongly convex objectives. We also note that primal–dual methods [7, 20, 25, 31, 58, 59, 62] and operator splitting schemes [13, 18, 19] can be applied to (1) with two-block structure.

However, among these works, it is rare to see the optimal *mixed-type* convergence rate, i.e., the lower complexity bound [51]

$$\mathcal{O}\left(\min\left\{\frac{\|A\|}{\epsilon}, \frac{\|A\|}{\sqrt{\mu\epsilon}}\right\} + \min\left\{\sqrt{L/\epsilon}, \sqrt{L/\mu}\cdot|\ln\epsilon|\right\}\right), \tag{2}$$

where $\mu \geq 0$ is the convexity parameter of f and L is the Lipschitz constant of ∇h. Both Nesterov's smoothing technique [46] and the accelerated primal–dual method in [12] achieve the lower bound for convex case $\mu = 0$. The inexact ALM framework in [66] possesses the optimal complexity (2) but involves a subroutine for inexactly solving the subproblem.

We mention that the second part of (2) corresponding to L and μ agrees with the well-known lower complexity bound of first-order methods for unconstrained convex problems with Lipschitz gradients, namely (the affine set Ω is the entire space $\mathbb{R}^n$)

$$\min_{x\in Q} f(x) := h(x) + g(x). \tag{3}$$

The intermediate non-Lipschitz case is also of interest to be considered [43, 45]. Particularly, when ∇h is Hölder continuous (cf.(11)) with exponent $\nu \in [0, 1)$, Nesterov [48] presented a universal fast gradient method (FGM) for solving (3) that did *not* require à priori knowledge of the smoothness parameter ν and the Hölderian constant $M_\nu(h)$. A key ingredient of FGM is that Hölderian gradients can be recast into the standard Lipschitz case but with inexact computations [15, 54, 55], and it achieves the optimal complexity [44]

$$\mathcal{O}\left(\frac{[M_\nu(h)]^{\frac{2}{1+3\nu}}}{\epsilon^{\frac{2}{1+3\nu}}}\right). \tag{4}$$

More extensions of FGM can be found in [23, 24, 32].

The dual problem of (1) reads equivalently as

$$\min_{\lambda\in\mathbb{R}^m}\left\{d(\lambda) := \langle b, \lambda\rangle + \max_{x\in Q}\left\{-f(x) - \left\langle A^\top\lambda, x\right\rangle\right\}\right\}. \tag{5}$$

If f is uniformly convex of degree $p \geq 2$ (see [48, Definition 1]), then ∇d is Hölder continuous with exponent $\nu = 1/(p-1)$ (cf. [48, Lemma 1]). The methods in [16, 37] work for (5) with strongly convex objective f, i.e., the Lipschitzian case ($\nu = 1$). Yurtsever et al. [68] proposed an accelerated universal primal–dual gradient method (AccUniPDGrad) for general Hölderian case ($\nu < 1$) and established the complexity bound (4) for the objective residual and the feasibility violation, with $M_\nu(h)$ being replaced by $M_\nu(d)$. Similarly with the spirit of FGM, the presented method utilizes the "inexactness" property of ∇d and applies FISTA [2] to (5) with line search.

In this work, we propose a universal accelerated primal–dual method (see Algorithm 1) for solving (1). Compared with existing works, the main contributions are highlighted as follows:

- It is first-order black-box type for both Lipschitz and Hölder cases but does not need to know the smoothness level priorly.
- It uses the Bregman divergence and can handle the non-Euclidean setting.
- In line search part, it adopts dynamically decreasing tolerance while FGM [48] and AccUniPDGrad [68] use the desired fixed accuracy.
- By using the tool of Lyapunov function and tight decay estimates of some differential/difference inequalities, we prove the universal mixed-type estimate that achieves the optimal complexity (including (2),(4) as special cases).

We also provide some numerical tests to validate the practical performance. It is confirmed that: (i) a proper choice of Bregman distance is crucial indeed; (ii) our method outperforms FGM and AccUniPDGrad especially for non-Lipschitz problems and smooth problems with large Lipschitz constants, as the automatically decreasing tolerance leads to approximate Lipschitz constants with moderate magnitude.

Our method here is motivated from an implicit-explicit time discretization of a novel accelerated Bregman primal–dual dynamics (see (24)), which is an extension of the accelerated primal–dual flow in [39] to the non-Euclidean case. For unconstrained problems, there are some existing continuous dynamics [34, 63, 64] with Bregman distances. For linearly constrained case, we see an accelerated primal–dual mirror model [69], which is inspired by the accelerated mirror descent [34] and the primal–dual dynamical approach [21] but without time discretizations.

The rest of the paper is organized as follows. In Sect. 2, we provide some preliminaries including the Bregman divergence and the Hölder continuity. Then, the main algorithm together with its universal mixed-type estimate is presented in Sect. 3, and rigorous proofs of two technical lemmas are summarized in Sects. 4 and 5, respectively. Finally, some numerical results are reported in Sect. 6.

2 Preliminary

2.1 Notations

Let $\langle \cdot, \cdot \rangle$ be the usual inner product of vectors and $\|\cdot\|$ be the standard Euclidean norm (of vectors and matrices). Given a proper function $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the effective domain of g is denoted as usual by $\mathbf{dom}\, g$, and the subdifferential of g at any

$x \in \mathbf{dom}\, g$ is the set of all subgradients:

$$\partial g(x) := \left\{\xi \in \mathbb{R}^n : g(y) \geq g(x) + \langle \xi, y - x\rangle \quad \forall\, y \in \mathbb{R}^n\right\}.$$

Recall that $Q \subset \mathbb{R}^n$ is a nonempty closed convex subset. Let $\iota_Q(\cdot)$ be the indicator function of Q and $N_Q(\cdot) := \partial \iota_Q(\cdot)$ be its normal cone.

Introduce the Lagrangian for the model problem (1):

$$\mathcal{L}(x, \lambda) := f(x) + \iota_Q(x) + \langle \lambda, Ax - b\rangle \quad \forall\, (x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m.$$

We say $(x^*, \lambda^*) \in Q \times \mathbb{R}^m$ is a saddle point of $\mathcal{L}$ if

$$\mathcal{L}(x^*, \lambda) \leq \mathcal{L}(x^*, \lambda^*) \leq \mathcal{L}(x, \lambda^*) \quad \forall\, (x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m, \tag{6}$$

which also implies the optimality condition:

$$Ax^* - b = 0, \quad \partial f(x^*) + N_Q(x^*) + A^\top \lambda^* \ni 0. \tag{7}$$

2.2 Bregman Divergence

Let $\phi : Q \to \mathbb{R}$ be a smooth *prox-function* and define the *Bregman divergence*

$$D_\phi(x, y) := \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle \quad \forall\, x,\ y \in Q.$$

Throughout, suppose ϕ is 1-strongly convex:

$$D_\phi(x, y) \geq \frac{1}{2}\|x - y\|^2 \quad \forall\, x,\ y \in Q. \tag{8}$$

Particularly, $\phi(x) = \frac{1}{2}\|x\|^2$ leads to $D_\phi(x, y) = D_\phi(y, x) = 1/2\,\|x - y\|^2$, which boils down to the standard Euclidean setting. In addition, we have the following *three-term identity*; see [9, Lemma 3.2] or [17, Lemma 3.3].

Lemma 2.1 *([9, 17]) For any $x, y, z \in Q$, it holds that*

$$\langle \nabla\phi(x) - \nabla\phi(y), y - z\rangle = D_\phi(z, x) - D_\phi(z, y) - D_\phi(y, x). \tag{9}$$

If $\phi(x) = \frac{1}{2}\|x\|^2$, then

$$2\langle x - y, y - z\rangle = \|x - z\|^2 - \|y - z\|^2 - \|x - y\|^2. \tag{10}$$

2.3 Hölder Continuity

Let h be a differentiable function on Q. For $0 \le \nu \le 1$, define

$$M_\nu(h) := \sup_{\substack{x,\, y \in Q \\ x \neq y}} \frac{\|\nabla h(x) - \nabla h(y)\|}{\|x - y\|^\nu}.$$

If $M_\nu(h) < \infty$, then ∇h is Hölder continuous with exponent ν:

$$\|\nabla h(x) - \nabla h(y)\| \le M_\nu(h)\, \|x - y\|^\nu \quad \forall\, x, y \in Q, \tag{11}$$

and this also implies that

$$h(x) \le h(y) + \langle \nabla h(y), x - y \rangle + \frac{M_\nu(h)}{1+\nu} \|x - y\|^{1+\nu} \quad \forall\, x, y \in Q. \tag{12}$$

When $\nu = 1$, $M_1(h)$ corresponds to the Lipschitz constant of ∇h, and we also use the conventional notation $L_h = M_1(h)$.

According to [48, Lemma 2], the estimate (12) can be transferred into the usual gradient descent inequality, with "inexact computations". Based on this, (accelerated) gradient methods can be used to minimize functions with Hölder continuous gradients [15, 54, 55].

Proposition 2.1 *(*[48]) Assume $M_\nu(h) < \infty$ and for $\delta > 0$, define

$$M(\nu, \delta) := \delta^{\frac{\nu-1}{\nu+1}} [M_\nu(h)]^{\frac{2}{\nu+1}}. \tag{13}$$

Then, for any $M \ge M(\nu, \delta)$, we have

$$h(x) \le h(y) + \langle \nabla h(y), x - y \rangle + \frac{M}{2} \|x - y\|^2 + \frac{\delta}{2} \quad \forall\, x, y \in Q.$$

3 Main Algorithm

Throughout, we make the following assumption on $f = h + g$:

Assumption 3.1 The nonsmooth part g is properly closed convex on Q. The smooth part h satisfies $\inf_{0 \le \nu \le 1} M_\nu(h) < \infty$ and is μ-convex on Q with $\mu \ge 0$, namely,

$$h(x) \ge h(y) + \langle \nabla h(y), x - y \rangle + \mu D_\phi(x, y) \quad \forall\, x, y \in Q.$$

Remark 3.1 For some cases, h might not be smooth (in C^1 globally) enough but at least Lipschitz continuous. By Rademacher's theorem [30, Theorem 3.1], Lipschitzian functions are differentiable almost everywhere. Thus, the assumption $M_\nu(h) < \infty$ holds true with $\nu = 0$. Besides, in practical computations, we expect that the proximal

calculation (cf.(14)) of the nonsmooth part g with respect to proper $D_\phi(\cdot,\cdot)$ is easy to compute (or has closed solution); see the matrix game problem in Sect. 6.1. □

Our main algorithm, called universal accelerated primal–dual (UAPD) method, is summarized in Algorithm 1, where the subpart "sub-UAPD" in lines 3 and 6 has been given by Algorithm 2. Note that we do *not* require priorly the exponent ν and the smoothness constant $M_\nu(h)$ but perform a line search procedure (see lines 4–7).

Algorithm 1 Universal Accelerated Primal–Dual (UAPD) Method

Require: Problem information: $\mu \geq 0$ and $\|A\|$.
Line search parameters: $\rho_u > 1,\ \rho_d \geq 1$.
Initial parameters: $\gamma_0,\ M_0 > 0,\ \beta_0 = 1$.
Initial guesses: $x_0,\ v_0 \in Q$ and $\lambda_0 \in \mathbb{R}^m$.

1: **for** $k = 0, 1, \cdots$ **do**
2: Set $i = 0,\ M_{k,0} = M_k$ and $S_k = \{x_k, v_k, \lambda_k, \beta_k, \gamma_k\}$.
3: Compute $\{y_{k,i}, x_{k,i}, v_{k,i}, \alpha_{k,i}, \delta_{k,i}, \Delta_{k,i}\} = \texttt{sub-UAPD}(k, S_k, M_{k,i})$.
4: **while** $h(x_{k,i}) - \Delta_{k,i} > \delta_{k,i}/2$ **do** {Line search}
5: Update $i = i + 1$ and $M_{k,i} = \rho_u^i \cdot M_{k,0}$.
6: Compute $\{y_{k,i}, x_{k,i}, v_{k,i}, \alpha_{k,i}, \delta_{k,i}, \Delta_{k,i}\} = \texttt{sub-UAPD}(k, S_k, M_{k,i})$.
7: **end while**
8: Set $i_k = i,\ \alpha_k = \alpha_{k,i_k},\ M_{k+1} = M_{k,i_k}/\rho_d$ and $\delta_{k+1} = \delta_{k,i_k}$.
9: Update $\gamma_{k+1} = (\gamma_k + \mu\alpha_k)/(1 + \alpha_k)$ and $\beta_{k+1} = \beta_k/(1 + \alpha_k)$.
10: Update $x_{k+1} = x_{k,i_k}, v_{k+1} = v_{k,i_k}$ and $\lambda_{k+1} = \lambda_k + \alpha_k/\beta_k(Av_{k+1} - b)$.
11: **end for**

Algorithm 2 $\{\widetilde{y}_k, \widetilde{x}_k, \widetilde{v}_k, \widetilde{\alpha}_k, \widetilde{\delta}_k, \widetilde{\Delta}_k\} = \texttt{sub-UAPD}(k, S_k, \widetilde{M}_k)$

Require: $k \in \mathbb{N},\ \widetilde{M}_k > 0$ and $S_k = \{x_k, v_k, \lambda_k, \beta_k, \gamma_k\}$.

1: Choose the step size $\widetilde{\alpha}_k = \sqrt{\beta_k\gamma_k}/\sqrt{\beta_k\widetilde{M}_k + \|A\|^2}$.
2: Set $\widetilde{\beta}_k = \beta_k/(1 + \widetilde{\alpha}_k)$ and $\widetilde{\delta}_k = \widetilde{\beta}_k/(k + 1)$.
3: Set $\widetilde{y}_k = (x_k + \widetilde{\alpha}_k v_k)/(1 + \widetilde{\alpha}_k)$ and $\widetilde{\lambda}_k = \lambda_k + \widetilde{\alpha}_k/\beta_k\,(Av_k - b)$.
4: Update $\widetilde{x}_k = (x_k + \widetilde{\alpha}_k\widetilde{v}_k)/(1 + \widetilde{\alpha}_k)$ with

$$\widetilde{v}_k = \underset{v\in Q}{\operatorname{argmin}} \left\{ g(v) + \left\langle \nabla h(\widetilde{y}_k) + A^\top\widetilde{\lambda}_k, v\right\rangle + \mu D_\phi(v, \widetilde{y}_k) + \frac{\gamma_k}{\widetilde{\alpha}_k} D_\phi(v, v_k) \right\}. \tag{14}$$

5: Compute $\widetilde{\Delta}_k = h(\widetilde{y}_k) + \langle \nabla h(\widetilde{y}_k), \widetilde{x}_k - \widetilde{y}_k\rangle + \frac{\widetilde{M}_k}{2}\|\widetilde{x}_k - \widetilde{y}_k\|^2$.

3.1 Line Search

In Algorithm 1, the parameter $\rho_u > 1$ enlarges the approximate Lipschitz constant $M_{k,i}$ (cf. line 5) to meet the following descent condition (cf. line 4)

$$h(x_{k,i}) \leq h(y_{k,i}) + \left\langle \nabla h(y_{k,i}), x_{k,i} - y_{k,i}\right\rangle + \frac{M_{k,i}}{2}\left\|x_{k,i} - y_{k,i}\right\|^2 + \frac{\delta_{k,i}}{2}. \tag{15}$$

For each $k \in \mathbb{N}$, the line search part will end up with the smallest integer i_k satisfying (15) and call the subpart `sub-UAPD` $i_k + 1$ times in total. The extra parameter $\rho_d \geq 1$ reduces the output constant M_{k,i_k} and updates $M_{k+1} = M_{k,i_k}/\rho_d$ (cf. line 8).

Remark 3.2 We introduce the pair (ρ_u, ρ_d) in our method for generality. Practically, it is not easy to find the optimal choice. If $\rho_d = 1$, then M_k is nondecreasing. Otherwise, M_k can be nonmonotone. Consider two situations.

- For the standard Lipschitz case ($\nu = 1$), as analyzed in [47, Section 3], if $M_0 \leq L_h$, then the choice $\rho_d \geq \rho_u$ promises that $M_k \leq L_h$ for all $k \geq 0$.
- For the Hölder continuous case ($\nu < 1$), we have $M_k \to \infty$ as $k \to \infty$ (since $M(\nu, \delta) \to \infty$ as $\delta \to 0$). Taking $\rho_d > 1$ may reduce M_k (locally for some k) but will increase the burden on the line search procedure. □

Below, let us show that i_k is finite for $k \in \mathbb{N}$. Indeed, we see $M_{k,i} = \rho_u^i M_{k,0}$ increases as i does, and the step size (cf. line 1 of Algorithm 2)

$$\alpha_{k,i} = \sqrt{\frac{\beta_k \gamma_k}{\beta_k M_{k,i} + \|A\|^2}} \tag{16}$$

has to be decreasing. Thus the tolerance (cf. line 2 of Algorithm 2)

$$\delta_{k,i} = \frac{1}{k+1} \cdot \frac{\beta_k}{1 + \alpha_{k,i}} \tag{17}$$

is increasing and by (13), $M(\nu, \delta_{k,i})$ is decreasing. This together with Proposition 2.1 and Assumption 3.1 concludes that

$$\begin{aligned} &i_k = 0 && \text{if } M(\nu, \delta_{k,0}) \leq M_{k,0}, \\ &i_k \leq \lceil s^* \rceil && \text{else,} \end{aligned}$$

where $\lceil s^* \rceil$ denotes the ceiling function (the minimal integer that is no less than s^*), and $s^* \in [0, \infty)$ solves the equation $M_{k,s^*} = M(\nu, \delta_{k,s^*})$; see Fig. 1. In particular, we have

$$M_{k,s^*} = \rho_u^{s^*} M_{k,0} \leq M(\nu, \delta_{k,0}) \quad \Longrightarrow \quad s^* \leq \log_{\rho_u} \frac{M(\nu, \delta_{k,0})}{M_{k,0}}.$$

This eventually leads to

$$i_k \leq \max\left\{0, \left\lceil \log_{\rho_u} \frac{M(\nu, \delta_{k,0})}{M_{k,0}} \right\rceil\right\} < \infty.$$

Remark 3.3 In the line search part, our Algorithm 1 adopts dynamically decreasing tolerance (17). However, the methods in [48] and [68, Algorithm 2] chose $\delta_k = \epsilon/k$, where ϵ is the desired accuracy. Hence, by Proposition 2.1, the approximate smoothness constant M_k of our method is smaller than the other two methods, especially for the Hölderian case. This will be verified by numerical experiments in Sect. 6. □

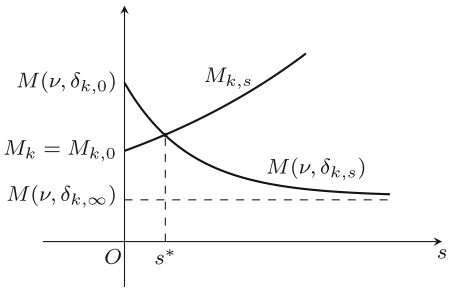

Fig. 1 The illustration of $M_{k,s}$ and $M(\nu, \delta_{k,s})$ as functions of $s \in [0, \infty)$. Here $\delta_{k,\infty} = \lim_{s\to\infty} \delta_{k,s} = \beta_k/(k+1)$ since $\alpha_{k,s} \to 0$ as $s \to \infty$

Below, we give an upper bound of M_k and the total number of line search steps. By Theorem 3.1, β_k corresponds to the convergence rate of Algorithm 1 and admits explicit decay estimate (see Lemma 3.3). If the desired accuracy $\beta_{k+1} = \mathcal{O}(\epsilon)$ is given, then the term $\left|\log_{\rho_u} \beta_{k+1}\right|$ in (19) can be replaced by $\left|\log_{\rho_u} \epsilon\right|$.

Lemma 3.1 *For any $k \in \mathbb{N}$, we have*

$$M_{k+1} \le \frac{1}{\rho_d} \max\left\{ \frac{M_0}{\rho_d^k},\ \sqrt{\rho_u}\rho_u \cdot M(\nu, \delta_{k+1}) \right\}, \tag{18}$$

and consequently, it holds that

$$\sum_{j=0}^{k} i_j \le \frac{3}{2} + k\,\frac{\ln \rho_d}{\ln \rho_u} + \frac{2}{1+\nu}\left|\log_{\rho_u} \frac{M_\nu(h)}{M_0}\right| + \frac{1-\nu}{1+\nu}\left[\log_{\rho_u}(k+1) + \left|\log_{\rho_u} \beta_{k+1}\right|\right]. \tag{19}$$

Proof See Sect. Appendix A. □

3.2 Time Discretization Interpretation

Below, we provide a time discretization interpretation of Algorithm 1. Given the k-th iterations (x_k, v_k, λ_k) and the parameters (γ_k, β_k, M_k), the line search procedure returns $(y_k, x_{k+1}, v_{k+1}, \lambda_{k+1})$ that satisfy

$$\frac{y_k - x_k}{\alpha_k} = v_k - y_k, \tag{20a}$$

$$\gamma_k \frac{\nabla\phi(v_{k+1}) - \nabla\phi(v_k)}{\alpha_k} \in \mu\left[\nabla\phi(y_k) - \nabla\phi(v_{k+1})\right] - \mathcal{G}(y_k, v_{k+1}, \lambda_k), \tag{20b}$$

$$\frac{x_{k+1} - x_k}{\alpha_k} = v_{k+1} - x_{k+1}, \tag{20c}$$

$$\beta_k \frac{\lambda_{k+1} - \lambda_k}{\alpha_k} = Av_{k+1} - b, \tag{20d}$$

where $\mathcal{G}(y_k, v_{k+1}, \lambda_k) := \nabla h(y_k) + \partial g(v_{k+1}) + N_Q(v_{k+1}) + A^\top \widetilde{\lambda}_k$ with $\widetilde{\lambda}_k = \lambda_k + \alpha_k/\beta_k(Av_k - b)$, and the step size α_k solves (cf.(16))

$$\alpha_k^2(\rho_d \beta_k M_{k+1} + \|A\|^2) = \gamma_k \beta_k, \tag{21}$$

where we used the relation $M_{k,i_k} = \rho_d M_{k+1}$. Besides, y_k and x_{k+1} fulfill (cf. (15))

$$h(x_{k+1}) \le h(y_k) + \langle \nabla h(y_k), x_{k+1} - y_k \rangle + \frac{\rho_d M_{k+1}}{2} \|x_{k+1} - y_k\|^2 + \frac{\delta_{k+1}}{2}, \tag{22}$$

and the parameters $(\gamma_{k+1}, \beta_{k+1})$ are governed by (cf. line 9 of Algorithm 1)

$$\frac{\gamma_{k+1} - \gamma_k}{\alpha_k} = \mu - \gamma_{k+1}, \quad \frac{\beta_{k+1} - \beta_k}{\alpha_k} = -\beta_{k+1}, \tag{23}$$

with $\beta_0 = 1$ and $\gamma_0 > 0$. For clarity, in Sect. Appendix B, we give a detailed derivation of the reformulation (20a) from Algorithms 1 and 2.

As one can see, y_k in (20a) is an intermediate which provides a "prediction", and then the "correction" step (20c) is used to update x_{k+1}. From (20a), (20b), and (20c), it is not hard to find that $y_k,\ v_{k+1},\ x_{k+1} \in Q$, as long as $x_k, v_k \in Q$. Therefore, with $x_0, v_0 \in Q$, it holds that $\{x_k, y_k, v_k\}_{k\in\mathbb{N}} \subset Q$.

Furthermore, we mention that the reformulation (20a) itself admits an implicit-explicit time discretization for the following primal–dual dynamics:

$$\begin{aligned} &x' = v - x, \\ &\gamma \frac{\mathrm{d}}{\mathrm{d}t} \nabla\phi(v) \in \mu\big[\nabla\phi(x) - \nabla\phi(v)\big] - \partial f(x) - N_Q(x) - A^\top \lambda, \\ &\beta\lambda' = Av - b, \end{aligned} \tag{24}$$

where γ and β are governed by continuous analogues to (23):

$$\gamma' = \mu - \gamma, \quad \beta' = -\beta. \tag{25}$$

We call (24) the *Accelerated Bregman Primal–Dual* (ABPD) flow. For $\phi(x) = 1/2\,\|x\|^2$, it amounts to the accelerated primal–dual flow in [39].

3.3 A Universal Estimate

Let $\{(x_k, v_k, \lambda_k, \gamma_k, \beta_k)\}_{k\in\mathbb{N}}$ be the sequence generated from Algorithm 1. We introduce the discrete Lyapunov function

$$\mathcal{E}_k := \mathcal{L}\left(x_k, \lambda^*\right) - \mathcal{L}\left(x^*, \lambda_k\right) + \gamma_k D_\phi(x^*, v_k) + \frac{\beta_k}{2}\left\|\lambda_k - \lambda^*\right\|^2. \tag{26}$$

A one-step estimate is presented below.

Lemma 3.2 *Under Assumption 3.1, we have*

$$\mathcal{E}_{k+1} - \mathcal{E}_k \leq -\alpha_k \mathcal{E}_{k+1} + \frac{\delta_{k+1}}{2}(1+\alpha_k) \quad \forall\, k \in \mathbb{N}. \tag{27}$$

Proof See Sect. 4. □

Using this lemma, we obtain the following theorem, which says the final rate is given by the sharp decay estimate of the sequence $\{\beta_k\}_{k\in\mathbb{N}}$; see Lemma 3.3.

Theorem 3.1 *Under Assumption 3.1, we have* $\{x_k, v_k\}_{k\in\mathbb{N}} \subset Q$ *and*

$$\|Ax_k - b\| \leq \beta_k \mathcal{T}_k, \tag{28}$$

$$\left|f(x_k) - f(x^*)\right| \leq \beta_k \mathcal{W}_k, \tag{29}$$

$$\mathcal{L}\left(x_k, \lambda^*\right) - \mathcal{L}\left(x^*, \lambda_k\right) \leq \beta_k \mathcal{R}_k, \tag{30}$$

for all $k \in \mathbb{N}$, *where* $\mathcal{R}_k := \mathcal{E}_0 + \ln(k+1)$, $\mathcal{T}_k := \|Ax_0 - b\| + 2\sqrt{2\mathcal{R}_k}$ *and* $\mathcal{W}_k := \mathcal{R}_k + \|\lambda^*\|\,\mathcal{T}_k$. *Moreover, we have*

$$\gamma_{\min}\left\|v_k - x^*\right\|^2 + \mu\left\|x_k - x^*\right\|^2 \leq 4\beta_k \mathcal{R}_k, \tag{31}$$

where $\gamma_{\min} := \min\{\gamma_0, \mu\} \geq 0$.

Proof From (23) and the contraction estimate (27) follows immediately that

$$\mathcal{E}_{k+1} \leq \frac{1}{1+\alpha_k}\mathcal{E}_k + \frac{\delta_{k+1}}{2} \quad \Longrightarrow \quad \mathcal{E}_k \leq \beta_k \mathcal{E}_0 + \frac{\beta_k}{2}\sum_{i=0}^{k-1}\frac{\delta_{i+1}}{\beta_{i+1}}.$$

By (17,23), we have

$$\delta_{k+1} = \frac{1}{k+1}\cdot\frac{\beta_k}{1+\alpha_k} = \frac{\beta_{k+1}}{k+1},$$

which further implies

$$\mathcal{E}_k \leq \beta_k \mathcal{E}_0 + \frac{\beta_k}{2}\sum_{i=0}^{k-1}\frac{1}{i+1} \leq \beta_k\left[\mathcal{E}_0 + \ln(k+1)\right] = \beta_k \mathcal{R}_k, \tag{32}$$

which proves (30). In view of (23), it holds that $\gamma_{k+1} = (\gamma_k + \mu\alpha_k)/(1+\alpha_k) \geq \gamma_{\min}$. This together with (32) gives

$$\gamma_{\min} D_\phi(x^*, v_k) \leq \gamma_k D_\phi(x^*, v_k) \leq \mathcal{E}_k \leq \beta_k \mathcal{R}_k.$$

It is clear that $\mathcal{L}(\cdot, \lambda)$ is convex and by (7,30,3.1),

$$\mu D_\phi(x_k, x^*) \leq \mathcal{L}\left(x_k, \lambda^*\right) - \mathcal{L}\left(x^*, \lambda^*\right) \leq \beta_k \mathcal{R}_k.$$

Hence, combining the above two estimates with (8) leads to (31).

Following [39, Theorem 3.1], we can establish (28,29). For the sake of completeness, we provide the details as below. Thanks to (20c) and (20d), we have

$$\begin{aligned}\lambda_{k+1}-\lambda_k &= \frac{\alpha_k}{\beta_k}(Av_{k+1}-b) = \frac{\alpha_k}{\beta_k}\left[A\left(x_{k+1}+\frac{x_{k+1}-x_k}{\alpha_k}\right)-b\right]\\ &= \frac{\alpha_k}{\beta_k}\left[\frac{1+\alpha_k}{\alpha_k}(Ax_{k+1}-b)-\frac{1}{\alpha_k}(Ax_k-b)\right]\\ &= \frac{1}{\beta_{k+1}}(Ax_{k+1}-b)-\frac{1}{\beta_k}(Ax_k-b),\end{aligned}$$

which yields that

$$\lambda_k-\lambda_0 = \frac{1}{\beta_k}(Ax_k-b)-(Ax_0-b).$$

By (32), we obtain $\|\lambda_k-\lambda^*\|^2 \le 2\mathcal{R}_k$ and

$$\begin{aligned}\|Ax_k-b\| &= \beta_k\|\lambda_k-\lambda_0+Ax_0-b\|\\ &\le \beta_k\left\|\lambda_k-\lambda^*\right\|+\beta_k\left\|\lambda_0-\lambda^*\right\|+\beta_k\|Ax_0-b\|\\ &\le \beta_k\sqrt{2\mathcal{R}_k}+\beta_k\left\|\lambda_0-\lambda^*\right\|+\beta_k\|Ax_0-b\|.\end{aligned}$$

In view of $\mathcal{R}_k = \mathcal{E}_0+\ln(k+1)$ and $\beta_0=1$, we find that $\|\lambda_0-\lambda^*\|^2 \le 2\mathcal{R}_k$. Plugging this into the above estimate gives (28). Observing (30), it follows that

$$-\left\langle\lambda^*, Ax_k-b\right\rangle \le f(x_k)-f(x^*) \le \beta_k\mathcal{R}_k-\left\langle\lambda^*, Ax_k-b\right\rangle,$$

which promises

$$\left|f(x_k)-f(x^*)\right| \le \beta_k\mathcal{R}_k+\left\|\lambda^*\right\|\|Ax_k-b\| \le \beta_k\mathcal{W}_k.$$

Consequently, this proves (29) and concludes the proof of this theorem. □

Remark 3.4 Note that the choice (17) can be replaced with

$$\delta_{k,i} = \frac{\delta}{k+1}\cdot\frac{\beta_k}{1+\alpha_{k,i}}, \quad \delta>0.$$

Then the quantity $\mathcal{R}_k = \mathcal{E}_0+\ln(k+1)$ in Theorem 3.1 becomes $\mathcal{R}_k = \mathcal{E}_0+\delta\ln(k+1)$. Taking $\delta = 1/\ln(K+1)$ cancels the logarithm factor, where $K\in\mathbb{N}$ is the maximal number of iterations chosen in advance. □

It remains to investigate the decay rate of $\{\beta_k\}_{k\in\mathbb{N}}$. From (21,23), we obtain

$$\beta_{k+1}-\beta_k = -\frac{\sqrt{\gamma_k\beta_k}\beta_{k+1}}{\sqrt{\rho_d\beta_k M_{k+1}+\|A\|^2}}. \tag{33}$$

A careful investigation into this difference equation gives the desired result, which involves the following quantity

$$\Delta := \max\left\{M_0,\ \sqrt{\rho_u}\rho_u\,[M_\nu(h)]^{\frac{2}{1+\nu}}\right\}. \tag{34}$$

Lemma 3.3 *Assume that* $\max\{\gamma_0, \mu\} \le \|A\|^2$, *then*

$$\beta_k \le \frac{4\,\|A\|}{\sqrt{\gamma_0}k} + \frac{(16\sqrt{2})^{1+\nu}\Delta^{\frac{1+\nu}{2}}}{\gamma_0^{\frac{1+\nu}{2}}k^{\frac{1+3\nu}{2}}} \quad \forall k \ge 1, \tag{35}$$

and moreover, we have

$$\beta_k \le \begin{cases} \dfrac{64\,\|A\|^2}{\gamma_{\min}k^2 + \|A\|^2} + \exp\left(-\dfrac{k}{8}\sqrt{\dfrac{\gamma_{\min}}{\Delta}}\right) & \text{if } \nu = 1, \\ \dfrac{64\,\|A\|^2}{\gamma_{\min}k^2 + \|A\|^2} + \left(1 + \dfrac{1-\nu}{32}\sqrt{\dfrac{\gamma_{\min}}{2^{\frac{1-\nu}{1+\nu}}\Delta}}\,k^{\frac{1+3\nu}{2+2\nu}}\right)^{-\frac{2+2\nu}{1-\nu}} & \text{if } \nu < 1, \end{cases} \tag{36}$$

where $\gamma_{\min} = \min\{\gamma_0, \mu\} \ge 0$.

Proof Since $M(\nu, \delta_{k+1}) = \delta_{k+1}^{\frac{\nu-1}{\nu+1}}[M_\nu(h)]^{\frac{2}{\nu+1}}$ and $\delta_{k+1} = \beta_{k+1}/(k+1) \le 1$, by Lemma 3.1, we have

$$\begin{aligned} \rho_d M_{k+1} &\le \max\left\{\frac{M_0}{\rho_d^k},\ \sqrt{\rho_u}\rho_u M(\nu, \delta_{k+1})\right\} \\ &= \max\left\{\frac{M_0}{\rho_d^k},\ \sqrt{\rho_u}\rho_u[M_\nu(h)]^{\frac{2}{\nu+1}}\delta_{k+1}^{\frac{\nu-1}{\nu+1}}\right\} \le \Delta\cdot\delta_{k+1}^{\frac{\nu-1}{\nu+1}}. \end{aligned}$$

Plugging this into (33) gives

$$\beta_{k+1} - \beta_k \le -\frac{\sqrt{\gamma_k\beta_k}\beta_{k+1}}{\sqrt{\Delta\cdot\beta_k\delta_{k+1}^{\frac{\nu-1}{\nu+1}} + \|A\|^2}}. \tag{37}$$

From this we obtain (35,36). Missing proofs are provided in Sect. 5. □

Remark 3.5 Both two estimates (35,36) hold for $\mu \ge 0$. In addition, for the limiting case $\nu \to 1-$, we have

$$\lim_{\nu\to 1-}\left(1 + \frac{1-\nu}{32}\sqrt{\frac{\gamma_{\min}}{2^{\frac{1-\nu}{1+\nu}}\Delta}}\,k^{\frac{1+3\nu}{2+2\nu}}\right)^{-\frac{2+2\nu}{1-\nu}} = \exp\left(-\frac{k}{8}\sqrt{\frac{\gamma_{\min}}{\Delta}}\right),$$

which matches the case $\nu = 1$. □

By the *universal mixed-type estimate* in Lemma 3.3, our Algorithm 1 achieves the lower complexity bound for both the affinely constrained case (i.e., $A \neq 0_{m\times n}$) and the unconstrained case $\Omega = \mathbb{R}^n$ (i.e., $A = 0_{m\times n}$ and $b = 0_m$), with Hölderian smoothness exponent $\nu \in [0, 1]$. Detailed comparisons with existing results are summarized in order.

Remark 3.6 Consider the unconstrained case $\Omega = \mathbb{R}^n$ (i.e.,$A = 0_{m\times n}$ and $b = 0_m$).

- The Lipschitzian case $\nu = 1$: From (29), it follows that (neglecting the logarithm factor $\ln(k + 1)$)

$$f(x_k) - f^* \leq \beta_k \mathcal{E}_0,$$

where by Lemma 3.3, we have (taking $\gamma_0 > \mu$)

$$\beta_k \leq \min\left\{\frac{2\cdot 16^2\Delta}{\gamma_0 k^2},\ \exp\left(-\frac{k}{8}\sqrt{\frac{\mu}{\Delta}}\right)\right\}.$$

Hence, to achieve the accuracy $f(x_k) - f^* \leq \epsilon$, the iteration complexity is no more than (recalling (34) and assuming $\Delta \sim L_h$)

$$\mathcal{O}\left(\min\left\{\sqrt{\frac{L_h}{\epsilon}},\ \sqrt{\frac{L_h}{\mu}}\cdot|\ln\epsilon|\right\}\right).$$

This is the well-known lower bound (cf. [45, 49]) of first-order methods for smooth convex functions with Lipschitzian gradients; see [10, 11, 38, 42, 47].

- The Hölderian case $0 \leq \nu < 1$: Similarly with the above analysis, the iteration complexity for $f(x_k) - f^* \leq \epsilon$ is bounded by

$$\mathcal{O}\left(\min\left\{\left(\frac{M_\nu(h)}{\epsilon}\right)^{\frac{2}{1+3\nu}},\ \left(\frac{M_\nu(h)}{\mu}\right)^{\frac{2}{1+3\nu}}\cdot\left(\frac{\mu}{\epsilon}\right)^{\frac{1-\nu}{1+3\nu}}\right\}\right). \tag{38}$$

This matches the lower bound in [43, 44]. The convex case $\mu = 0$ has been obtained by the methods in [32, 43, 48], and the restarted schemes in [32, 52] attained the complexity bound for $\mu > 0$. Besides, Guminov et al. [24] obtained (38) for nonconvex problems, with an extra one-dimensional line search step. □

Remark 3.7 Then, let us focus on the affine constraint case (i.e., $A \neq 0_{m\times n}$).

- The Lipschitzian case $\nu = 1$: By (28,29), to achieve $|f(x_k) - f^*| \leq \epsilon$ and $\|Ax_k - b\| \leq \epsilon$, the iteration complexity (neglecting the logarithm factor $\ln(k+1)$) is

$$\mathcal{O}\left(\min\left\{\frac{\|A\|}{\epsilon} + \sqrt{\frac{L_h}{\epsilon}},\ \frac{\|A\|}{\sqrt{\mu\epsilon}} + \sqrt{\frac{L_h}{\mu}}\cdot|\ln\epsilon|\right\}\right). \tag{39}$$

This coincides with the lower complexity bound in [51]. The methods in [12, 46, 66] achieved the bound for convex case $\mu = 0$, and the strongly convex case $\mu > 0$ can be found in [66].

- The Hölderian case $0 \leq \nu < 1$:

$$\mathcal{O}\left(\min\left\{\frac{\|A\|}{\epsilon}+\left(\frac{M_\nu(h)}{\epsilon}\right)^{\frac{2}{1+3\nu}},\quad \frac{\|A\|}{\sqrt{\mu\epsilon}}+\left(\frac{M_\nu(h)}{\mu}\right)^{\frac{2}{1+3\nu}}\cdot\left(\frac{\mu}{\epsilon}\right)^{\frac{1-\nu}{1+3\nu}}\right\}\right).$$

Similarly with (39), this universal mixed-type estimate has optimal dependence on $\|A\|$ (corresponding to the affine constraint), and the remainder agrees with (38), which is optimal with respect to μ and $M_\nu(h)$ (related to the objective). □

4 Proof of Lemma 3.2

Start from the difference $\mathcal{E}_{k+1}-\mathcal{E}_k=\mathbb{I}_1+\mathbb{I}_2+\mathbb{I}_3$, where

$$\begin{aligned}
\mathbb{I}_1 &:= \mathcal{L}\left(x_{k+1},\lambda^*\right)-\mathcal{L}\left(x_k,\lambda^*\right),\\
\mathbb{I}_2 &:= \gamma_{k+1}D_\phi(x^*,v_{k+1})-\gamma_k D_\phi(x^*,v_k),\\
\mathbb{I}_3 &:= \frac{\beta_{k+1}}{2}\left\|\lambda_{k+1}-\lambda^*\right\|^2-\frac{\beta_k}{2}\left\|\lambda_k-\lambda^*\right\|^2.
\end{aligned}$$

Notice that $x_k,\ x_{k+1}\in Q$ and the first term is easy to handle:

$$\mathbb{I}_1 = f(x_{k+1})-f(x_k)+\langle\lambda^*,A(x_{k+1}-x_k)\rangle. \tag{40}$$

We give the estimate of $\mathbb{I}_2$ in Sect. 4.1 and finish the proof of (27) in Sect. 4.2.

4.1 Estimate of $\mathbb{I}_2$

Invoking the three-term identity (9) and the difference equation of $\{\gamma_k\}_{k\in\mathbb{N}}$ in (23), we split the second term $\mathbb{I}_2$ as follows

$$\begin{aligned}
\mathbb{I}_2 &= (\gamma_{k+1}-\gamma_k)D_\phi(x^*,v_{k+1})+\gamma_k\left[D_\phi(x^*,v_{k+1})-D_\phi(x^*,v_k)\right]\\
&= \alpha_k(\mu-\gamma_{k+1})D_\phi(x^*,v_{k+1})-\gamma_k D_\phi(v_{k+1},v_k)\\
&\quad+\gamma_k\left\langle\nabla\phi(v_{k+1})-\nabla\phi(v_k),v_{k+1}-x^*\right\rangle.
\end{aligned} \tag{41}$$

Let us prove

$$\begin{aligned}
&\mu\alpha_k D_\phi(x^*,v_{k+1})+\gamma_k\left\langle\nabla\phi(v_{k+1})-\nabla\phi(v_k),v_{k+1}-x^*\right\rangle\\
&\leq h(x_k)-h(y_k)-\alpha_k\left[h(y_k)-h(x^*)+\left\langle\widetilde{\lambda}_k,Av_{k+1}-b\right\rangle\right]\\
&\quad-\alpha_k\left[g(v_{k+1})-g(x^*)+\langle\nabla h(y_k),v_{k+1}-v_k\rangle\right],
\end{aligned} \tag{42}$$

which leads to the desired estimate of $\mathbb{I}_2$:

$$\begin{aligned}\mathbb{I}_2 \leq & -\alpha_k\gamma_{k+1}D_\phi(x^*, v_{k+1}) - \gamma_k D_\phi(v_{k+1}, v_k) - \alpha_k\langle\widetilde{\lambda}_k, Av_{k+1} - b\rangle \\ & - \alpha_k\left[g(v_{k+1}) - g(x^*) + h(y_k) - h(x^*)\right] \\ & + h(x_k) - h(y_k) - \alpha_k\left\langle\nabla h(y_k), v_{k+1} - v_k\right\rangle.\end{aligned} \tag{43}$$

To do this, define ζ_{k+1} by that

$$\begin{aligned}&\gamma_k\big[\nabla\phi(v_{k+1}) - \nabla\phi(v_k)\big] \\ &= \mu\alpha_k\big[\nabla\phi(y_k) - \nabla\phi(v_{k+1})\big] - \alpha_k\big[\nabla h(y_k) + \zeta_{k+1} + A^\top\widetilde{\lambda}_k\big].\end{aligned} \tag{44}$$

Observing (20b), it follows that $\zeta_{k+1} \in \partial g(v_{k+1}) + N_Q(v_{k+1})$ and

$$-\alpha_k\langle\zeta_{k+1}, v_{k+1} - x^*\rangle \leq -\alpha_k\left[g(v_{k+1}) - g(x^*)\right].$$

Thanks to (9), we have the decomposition

$$\begin{aligned}&\mu\alpha_k\left\langle\nabla\phi(y_k) - \nabla\phi(v_{k+1}), v_{k+1} - x^*\right\rangle \\ &= \mu\alpha_k\big[D_\phi(x^*, y_k) - D_\phi(x^*, v_{k+1}) - D_\phi(v_{k+1}, y_k)\big],\end{aligned}$$

and invoking (20a) leads to

$$\begin{aligned}&-\alpha_k\left\langle\nabla h(y_k), v_{k+1} - x^*\right\rangle \\ &= -\alpha_k\left\langle\nabla h(y_k), v_{k+1} - v_k\right\rangle - \left\langle\nabla h(y_k), y_k - x_k\right\rangle - \alpha_k\left\langle\nabla h(y_k), y_k - x^*\right\rangle.\end{aligned}$$

Since $x_k,\ y_k \in Q$, by Assumption 3.1 we obtain

$$\begin{aligned}&-\left\langle\nabla h(y_k), y_k - x_k\right\rangle - \alpha_k\left\langle\nabla h(y_k), y_k - x^*\right\rangle \\ &\leq h(x_k) - h(y_k) - \alpha_k\left[h(y_k) - h(x^*) + \mu D_\phi(x^*, y_k)\right].\end{aligned}$$

Hence, combining the above estimates with (44) proves (42).

4.2 Estimate of $\mathbb{I}_3$

Similarly with (41), by (9), (20d) and (23), the third term $\mathbb{I}_3$ is rearranged by that

$$\begin{aligned}\mathbb{I}_3 = & -\frac{\alpha_k\beta_{k+1}}{2}\left\|\lambda_{k+1} - \lambda^*\right\|^2 - \frac{\beta_k}{2}\left\|\lambda_{k+1} - \lambda_k\right\|^2 \\ & + \alpha_k\langle Av_{k+1} - b, \lambda_{k+1} - \lambda^*\rangle.\end{aligned}$$

To match the cross term $-\alpha_k\langle\widetilde{\lambda}_k, Av_{k+1} - b\rangle$ in the estimate of $\mathbb{I}_2$ (cf.(43)), we rewrite the last term as follows:

$$\begin{aligned}&\alpha_k\langle Av_{k+1} - b, \lambda_{k+1} - \lambda^*\rangle \\ &= \alpha_k\langle Av_{k+1} - b, \lambda_{k+1} - \widetilde{\lambda}_k\rangle + \alpha_k\langle Av_{k+1} - b, \widetilde{\lambda}_k - \lambda^*\rangle.\end{aligned}$$

In view of (10) and (20d), we get

$$\begin{aligned}&\alpha_k\langle Av_{k+1}-b,\lambda_{k+1}-\widetilde{\lambda}_k\rangle=\beta_k\left\langle\lambda_{k+1}-\lambda_k,\lambda_{k+1}-\widetilde{\lambda}_k\right\rangle\\&=\frac{\beta_k}{2}\left\|\lambda_{k+1}-\lambda_k\right\|^2+\frac{\beta_k}{2}\left\|\lambda_{k+1}-\widetilde{\lambda}_k\right\|^2-\frac{\beta_k}{2}\left\|\lambda_k-\widetilde{\lambda}_k\right\|^2,\end{aligned}$$

which gives

$$\mathbb{I}_3\leq-\frac{\alpha_k\beta_{k+1}}{2}\left\|\lambda_{k+1}-\lambda^*\right\|^2+\frac{\beta_k}{2}\left\|\lambda_{k+1}-\widetilde{\lambda}_k\right\|^2+\alpha_k\langle Av_{k+1}-b,\widetilde{\lambda}_k-\lambda^*\rangle.$$

4.3 Proof of (27)

Combining (43) and the estimate of $\mathbb{I}_3$, we obtain

$$\begin{aligned}\mathbb{I}_2+\mathbb{I}_3\leq&-\frac{\alpha_k\beta_{k+1}}{2}\left\|\lambda_{k+1}-\lambda^*\right\|^2+\frac{\beta_k}{2}\left\|\lambda_{k+1}-\widetilde{\lambda}_k\right\|^2-\alpha_k\langle\lambda^*,Av_{k+1}-b\rangle\\&-\alpha_k\gamma_{k+1}D_\phi(x^*,v_{k+1})-\gamma_kD_\phi(v_{k+1},v_k)\\&-\alpha_k\left[g(v_{k+1})-g(x^*)+h(y_k)-h(x^*)\right]\\&+h(x_k)-h(y_k)-\alpha_k\left\langle\nabla h(y_k),v_{k+1}-v_k\right\rangle.\end{aligned}$$

Inserting the identity (40) into the above inequality and observing that

$$\begin{aligned}&-\alpha_k\langle\lambda^*,Av_{k+1}-b\rangle+\langle\lambda^*,A(x_{k+1}-x_k)\rangle\\&=-\alpha_k\left\langle\lambda^*,A\left(x_{k+1}+\frac{x_{k+1}-x_k}{\alpha_k}\right)-b\right\rangle+\langle\lambda^*,A(x_{k+1}-x_k)\rangle\ \text{(by (20c))}\\&=-\alpha_k\langle\lambda^*,Ax_{k+1}-b\rangle\end{aligned}$$

and

$$\begin{aligned}&-\alpha_k\left[g(v_{k+1})-g(x^*)+h(y_k)-h(x^*)\right]\\&=-\alpha_k\left[g(x_{k+1})-g(x^*)+h(x_{k+1})-h(x^*)\right]\\&\quad-\alpha_k\left[g(x_k)-g(x_{k+1})+h(x_k)-h(x_{k+1})\right]\\&\quad-\alpha_k\left[g(v_{k+1})-g(x_k)+h(y_k)-h(x_k)\right]\\&=-\alpha_k\left[f(x_{k+1})-f(x^*)\right]-\alpha_k\left[f(x_k)-f(x_{k+1})\right]\\&\quad-\alpha_k\left[g(v_{k+1})-g(x_k)+h(y_k)-h(x_k)\right],\end{aligned}$$

we arrive at

$$\begin{aligned}\mathcal{E}_{k+1}-\mathcal{E}_k \leq & -\alpha_k\left[f(x_{k+1})-f(x^*)\right]-\alpha_k\langle\lambda^*, Ax_{k+1}-b\rangle\\ & -\alpha_k\gamma_{k+1}D_\phi(x^*,v_{k+1})-\frac{\alpha_k\beta_{k+1}}{2}\left\|\lambda_{k+1}-\lambda^*\right\|^2\\ & +(1+\alpha_k)\left[f(x_{k+1})-f(x_k)\right]+\frac{\beta_k}{2}\left\|\lambda_{k+1}-\widetilde{\lambda}_k\right\|^2\\ & -\alpha_k\left[g(v_{k+1})-g(x_k)+h(y_k)-h(x_k)\right]-\gamma_k D_\phi(v_{k+1},v_k)\\ & +h(x_k)-h(y_k)-\alpha_k\left\langle\nabla h(y_k), v_{k+1}-v_k\right\rangle.\end{aligned}$$

The first two lines equal to $-\alpha_k\mathcal{E}_{k+1}$ and a careful collection of the rest terms gives

$$\begin{aligned}\mathcal{E}_{k+1}-\mathcal{E}_k \leq & -\alpha_k\mathcal{E}_{k+1}+\frac{\beta_k}{2}\left\|\lambda_{k+1}-\widetilde{\lambda}_k\right\|^2-\gamma_k D_\phi(v_{k+1},v_k)\\ & +(1+\alpha_k)\left[h(x_{k+1})-h(y_k)\right]-\alpha_k\left\langle\nabla h(y_k), v_{k+1}-v_k\right\rangle\\ & +(1+\alpha_k)g(x_{k+1})-g(x_k)-\alpha_k g(v_{k+1}).\end{aligned}\tag{45}$$

From (20c), we see x_{k+1} is a convex combination of x_k and v_{k+1}, which implies

$$(1+\alpha_k)g(x_{k+1})\leq g(x_k)+\alpha_k g(v_{k+1}).$$

Thanks to (22) and the relation $\alpha_k(v_{k+1}-v_k)=(1+\alpha_k)(x_{k+1}-y_k)$ (cf.(20a) and (20c)), we obtain the estimate

$$\begin{aligned}&(1+\alpha_k)\left[h(x_{k+1})-h(y_k)\right]-\alpha_k\left\langle\nabla h(y_k), v_{k+1}-v_k\right\rangle\\ &\leq\frac{\alpha_k^2M_{k+1}}{2+2\alpha_k}\left\|v_{k+1}-v_k\right\|^2+\frac{\delta_{k+1}}{2}(1+\alpha_k).\end{aligned}$$

Consequently, plugging these two estimates into (45) and using (8) leads to

$$\begin{aligned}\mathcal{E}_{k+1}-\mathcal{E}_k \leq & -\alpha_k\mathcal{E}_{k+1}+\frac{\delta_{k+1}}{2}(1+\alpha_k)+\frac{\beta_k}{2}\left\|\lambda_{k+1}-\widetilde{\lambda}_k\right\|^2\\ & +\frac{\alpha_k^2M_{k+1}-\gamma_k(1+\alpha_k)}{1+\alpha_k}D_\phi(v_{k+1},v_k).\end{aligned}$$

Recall that $\widetilde{\lambda}_k=\lambda_k+\alpha_k/\beta_k(Av_k-b)$, which together with (20d) gives $\lambda_{k+1}-\widetilde{\lambda}_k=\alpha_k/\beta_k A(v_{k+1}-v_k)$ and

$$\begin{aligned}\mathcal{E}_{k+1}-\mathcal{E}_k \leq & -\alpha_k\mathcal{E}_{k+1}+\frac{\delta_{k+1}}{2}(1+\alpha_k)\\ & +\frac{1}{\beta_k}\left[\alpha_k^2(\beta_{k+1}M_{k+1}+\|A\|^2)-\gamma_k\beta_k\right]D_\phi(v_{k+1},v_k).\end{aligned}$$

Since $\beta_{k+1}\leq\beta_k$ (cf.(23)), the desired estimate (27) follows immediately from (21). This finishes the proof of Lemma 3.2.

5 Proof of Lemma 3.3

The key to complete the proof of Lemma 3.3 is the difference inequality (37). In Sect. 5.1, we shall introduce an auxiliary differential inequality (cf.(46)) that can be viewed as a continuous analogue to (37). Later in Sects. 5.2 and 5.3, we finish the proofs of (35),(36) by using the asymptotic estimate of (46).

5.1 A Differential Inequality

In the sequel, we need some functional spaces in one dimension; see [14, Definitions 1.87 and 2.1]. Denote by $C^1(I)$ the space of (real-valued) continuous functions on the interval $I \subset \mathbb{R}$ with continuous derivatives. Let $L^\infty(I)$ be the space of essentially bounded measurable functions, which means any $\sigma \in L^\infty(I)$ is bounded almost everywhere. The space $L^1(I)$ consists of measurable functions that are absolutely summable (integrable). As usual, we denote by $W^{1,\infty}(I)$ the set of all real-valued functions which, together with their generalized derivatives, belong to $L^\infty(I)$.

Let $\eta,\ R \geq 0$ and $\theta > 1$ be constants such that $\eta \leq \theta - 1$. Suppose $y \in W^{1,\infty}(0,\infty)$ is positive and satisfies the differential inequality

$$y'(t) \leq -\frac{\sigma(t)y^\theta(t)}{\sqrt{\varphi(t)y^{2\eta}(t)+R^2}}, \quad y(0)=1, \tag{46}$$

where $\sigma \in L^1(0,\infty)$ is nonnegative, and $\varphi \in C^1[0,\infty)$ is positive and nondecreasing. The decay rate of $y(t)$ is given below.

Lemma 5.1 *Assume $y \in W^{1,\infty}(0,\infty)$ is positive and satisfies (46). Then, for all $t > 0$, we have*

$$y(t) \leq \begin{cases} \exp\left(-\dfrac{\Sigma(t)}{2\sqrt{\varphi(t)}}\right) + \left(1+\dfrac{\theta-1}{2R}\Sigma(t)\right)^{\frac{1}{1-\theta}} & \text{if } \eta = \theta-1, \\ \left(1+\dfrac{\theta-1}{2R}\Sigma(t)\right)^{\frac{1}{1-\theta}} + \left(1+\dfrac{\theta-\eta-1}{2\sqrt{\varphi(t)}}\Sigma(t)\right)^{\frac{1}{\eta+1-\theta}} & \text{if } \eta < \theta-1, \end{cases}$$

where $\Sigma(t) := \int_0^t \sigma(s)\,\mathrm{d}s$.

Proof Write (46) as follows

$$y'(t) \leq -\frac{\sigma(t)}{\sqrt{\varphi(t)y^{2\eta-2\theta}(t)+R^2y^{-2\theta}(t)}}.$$

Shifting the denominator from the right to the left and using the trivial estimate

$$\sqrt{\varphi(t)y^{2\eta-2\theta}(t)+R^2y^{-2\theta}(t)} \leq \sqrt{\varphi(t)}y^{\eta-\theta}(t)+Ry^{-\theta}(t),$$

we obtain that

$$\left(\frac{\sqrt{\varphi(t)}}{y^{\theta-\eta}(t)}+\frac{R}{y^{\theta}(t)}\right)y'(t)\le-\sigma(t). \tag{47}$$

To the end, we shall discuss in two cases: $\eta=\theta-1$ and $\eta<\theta-1$. Detailed proofs can be found in Sect. Appendix C. □

5.2 Proof of (35)

By (23), we have

$$\gamma_{k+1}-\gamma_0\beta_{k+1}=\frac{\mu\alpha_k+\gamma_k}{1+\alpha_k}-\frac{\gamma_0\beta_k}{1+\alpha_k}\ge\frac{\gamma_k-\gamma_0\beta_k}{1+\alpha_k}.$$

Since $\gamma_0=\gamma_0\beta_0$, it follows that $\gamma_k\ge\gamma_0\beta_k$ and (37) becomes

$$\beta_{k+1}-\beta_k\le-\frac{\sqrt{\gamma_0}\beta_k\beta_{k+1}}{\sqrt{\Delta\cdot\beta_k\delta_{k+1}^{\frac{\nu-1}{\nu+1}}+\|A\|^2}}, \tag{48}$$

where $\delta_{k+1}=\beta_{k+1}/(k+1)$.

Define a piecewise continuous linear interpolation

$$y(t):=\beta_k(k+1-t)+\beta_{k+1}(t-k)\quad\forall\,t\in[k,k+1),\ k\in\mathbb{N}. \tag{49}$$

Clearly, $y\in W^{1,\infty}(0,\infty)$ is positive and $0<y(t)\le y(0)=1$. In particular, we have $\beta_k=y(k)$ for all $k\in\mathbb{N}$, and the decay estimate of β_k is transferred into the asymptotic behavior of $y(t)$, which satisfies (the proof is given below)

$$y'(t)\le-\frac{\sqrt{\gamma_0}y^2(t)/2}{\sqrt{\varphi(t)[y(t)]^{\frac{2\nu}{1+\nu}}+\|A\|^2}}, \tag{50}$$

where $\varphi(t):=4\Delta(t+1)^{\frac{1-\nu}{1+\nu}}$. Thus, utilizing Lemma 5.1 gives

$$\beta_k=y(k)\le\frac{4\|A\|}{\sqrt{\gamma_0}k}+\frac{(16\sqrt{2})^{1+\nu}\Delta^{\frac{1+\nu}{2}}}{\gamma_0^{\frac{1+\nu}{2}}k^{\frac{1+3\nu}{2}}}\quad\forall\,k\ge1,$$

which establishes (35).

Below, let us verify (50). Since $\gamma_k\le\max\{\gamma_0,\mu\}\le\|A\|^2$, from (21) we find that $\alpha_k\le\sqrt{\gamma_k\beta_k}/\|A\|\le1$. For any $t\in(k,k+1)$, it is clear that

$$1\ge\frac{\beta_{k+1}}{y(t)}\ge\frac{\beta_{k+1}}{\beta_k}=\frac{1}{1+\alpha_k}\ge\frac{1}{2},\quad\text{and}\quad1\le\frac{\beta_k}{y(t)}\le\frac{\beta_k}{\beta_{k+1}}\le2,$$

which implies

$$\Delta \cdot \beta_k \delta_{k+1}^{\frac{\nu-1}{\nu+1}} = \frac{\varphi(k)\beta_k \beta_{k+1}^{\frac{\nu-1}{\nu+1}}}{4} \le \varphi(t)[y(t)]^{\frac{2\nu}{1+\nu}}.$$

Since $y'(t) = \beta_{k+1} - \beta_k$, plugging the above estimate into (48) proves (50).

5.3 Proof of (36)

Again, by (23), we have $\gamma_k \ge \gamma_{\min} = \min\{\gamma_0, \mu\}$, and (37) becomes

$$\beta_{k+1} - \beta_k \le -\frac{\sqrt{\gamma_{\min}\beta_k}\beta_{k+1}}{\sqrt{\Delta \cdot \beta_k \delta_{k+1}^{\frac{\nu-1}{\nu+1}} + \|A\|^2}}.$$

Recall the interpolation $y(t)$ defined by (49). Similarly with (50), we claim that

$$y'(t) \le -\frac{\sqrt{\gamma_{\min}} y^{3/2}(t)/2}{\sqrt{\varphi(t)[y(t)]^{\frac{2\nu}{1+\nu}} + \|A\|^2}},$$

and invoking Lemma 5.1 again gives

$$\beta_k \le \begin{cases} \dfrac{64\|A\|^2}{\gamma_{\min}k^2 + \|A\|^2} + \exp\left(-\dfrac{k}{8}\sqrt{\dfrac{\gamma_{\min}}{\Delta}}\right) & \text{if } \nu = 1, \\ \dfrac{64\|A\|^2}{\gamma_{\min}k^2 + \|A\|^2} + \left(1 + \dfrac{1-\nu}{32}\sqrt{\dfrac{\gamma_{\min}}{2^{\frac{1-\nu}{1+\nu}}\Delta}}\, k^{\frac{1+3\nu}{2+2\nu}}\right)^{-\frac{2+2\nu}{1-\nu}} & \text{if } \nu < 1. \end{cases}$$

This proves (36) and completes the proof of Lemma 3.3.

6 Numerical Examples

In this part, we provide several experiments to validate the performance of our Algorithm 1 (denoted by UAPD). It is compared with Nesterov's FGM [48] and the AccUniPDGrad method [68], for solving unconstrained and affinely constrained problems.

Both UAPD and FGM involve the proximal mapping of the nonsmooth part g, and FGM performs one more proximal calculation for updating v_{k+1}. In line search part, FGM and AccUniPDGrad use the tolerance $\delta_k = \epsilon\tau_k$ with $\tau_k = \mathcal{O}(1/k)$, where ϵ is the desired accuracy. Clearly, this is smaller than ours $\delta_k = \beta_k/k$. As discussed in Remark 3.3, this will lead to over-estimate issue, especially for Hölderian case (cf. Sect. 6.1) and smooth problems with large Lipschitz constants (cf. Sect. 6.2).

6.1 Matrix Game

The problem reads as

$$\min_{x\in\Delta_n}\max_{y\in\Delta_m}\langle x, Py\rangle = \min_{x\in\Delta_n}\left\{\varphi(x) := \max_{1\le j\le m}\langle p_j, x\rangle\right\}, \tag{51}$$

where $P = (p_1, p_2, \cdots, p_m) \in \mathbb{R}^{n\times m}$ is the given payoff matrix and $\Delta_\times \subset \mathbb{R}^\times$ is the simplex with $\times = m$ or n. By von Neumann's minimax theorem [1, Corollary 15.30], we can change the min-max order of (51) and obtain

$$\max_{y\in\Delta_m}\min_{x\in\Delta_n}\langle x, Py\rangle = \max_{y\in\Delta_m}\left\{\psi(y) := \min_{1\le i\le n}\langle q_i, y\rangle\right\}, \tag{52}$$

where $P^\top = (q_1, q_2, \cdots, q_n)$ with $q_i \in \mathbb{R}^m$. Moreover, there is no duality gap, i.e., $\varphi^* = \psi^*$. According to [1, Proposition 5.12], we claim that $\varphi(x^\#) = \psi(y^\#)$ for some $(x^\#, y^\#) \in \Delta_n \times \Delta_m$ if and only if $\varphi(x^\#) = \varphi^*$ and $\psi(y^\#) = \psi^*$.

As we do not know the optimal value, consider the unconstrained problem

$$\min_{x\in\Delta_n,\, y\in\Delta_m}\{f(x, y) := \varphi(x) - \psi(y)\}. \tag{53}$$

Clearly, the minimal value is zero:

$$f^* = \min_{x\in\Delta_n}\varphi(x) - \max_{y\in\Delta_m}\psi(y) = \varphi^* - \psi^* = 0.$$

Moreover, $(x^\#, y^\#)$ is an optimal solution to (53) if and only if $x^\#$ and $y^\#$ are optimal solutions to (51,52), respectively.

Since ψ is concave, f is convex but nonsmooth (Lipschitz continuous). In view of Remark 3.1, f satisfies Assumption 3.1 without the simple part g. In addition, to work with the simplex constraint, a proper prox-function is the entropy $\phi(x) = \langle x, \ln x\rangle$, which gives closed solution of the proximal calculation (14).

Numerical results are displayed in Fig. 2. We record (i) the decay behavior of the objective residual $|f(x_k, y_k) - f^*|$ with respect to both iteration number k and running time t (in seconds), (ii) the total number $\#i_k$ of the line search step i_k and its average $\bar{i}_k$, and (iii) the approximate Lipschitz constant M_k. The pay off matrix P is generated from the normal distribution. For FGM, the priori accuracy is $\epsilon = 1e\text{-}5$. For our method, the line search parameters are $\rho_u = 2$ and $\rho_d = 1$.

From Fig. 2, our UAPD outperforms FGM, with faster convergence, less number of line search steps and smaller Lipschitz constants. Within the same iteration number $k = 1e5$, UAPD achieves ten times smaller error ($1e\text{-}2$ *v.s.* $1e\text{-}1$) but takes only about half less time (20 s *v.s.* 40 s). It can be seen that, FGM has much more line search steps and the average is 1, which means it performs at least one step of line search in each iteration. Since $\rho_d = 1$, according to Remark 3.2, the Lipschitz constant M_k of our UAPD is nondecreasing, which agrees with the numerical illustration. As mentioned in Remark 3.3, FGM suffers from over-estimated Lipschitz parameters

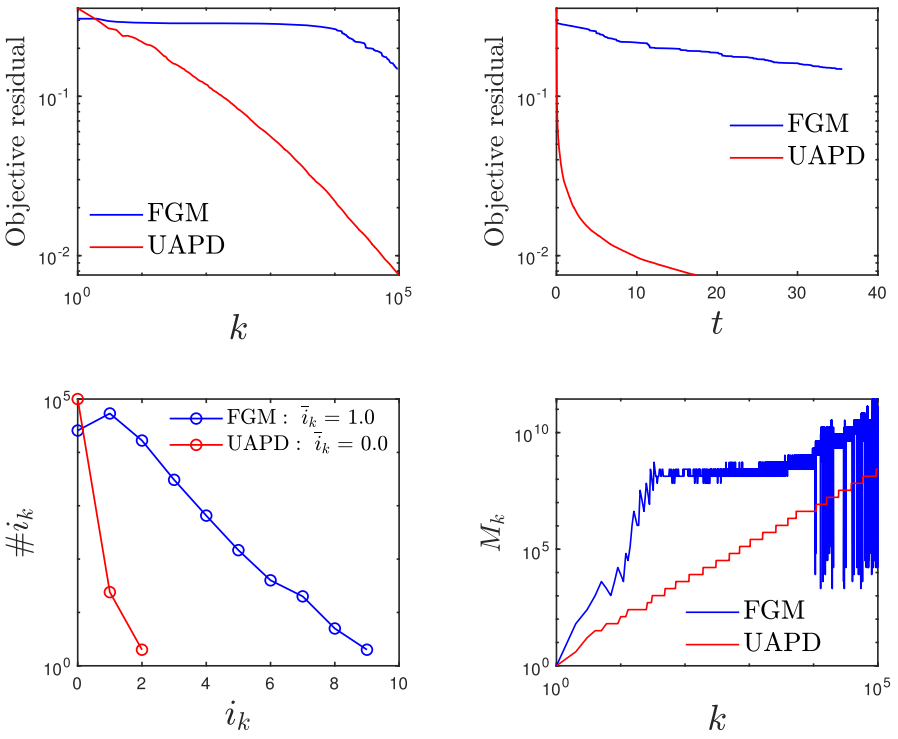

Fig. 2 Numerical results of FGM and UAPD for solving (53) with $m = 100,\ n = 400$. The desired accuracy of FGM is $\epsilon = 1e\text{-}5$

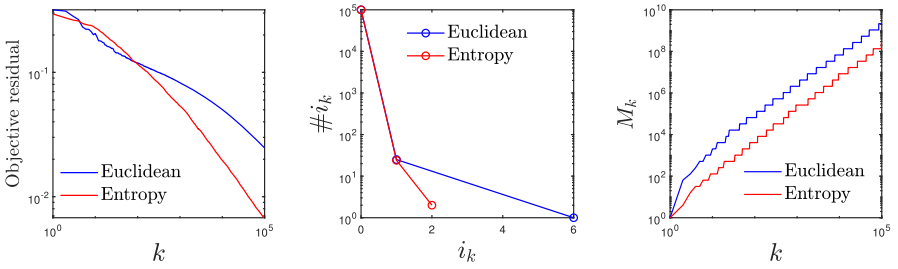

Fig. 3 Numerical performances of UAPD on the problem (53) with different prox-functions

with dramatically growth behavior since it adopts smaller tolerance ϵ/k (comparable to our dynamically decreasing choice β_k/k). Besides, in step 3 of FGM, it updates the Lipschitz constant by $L_{k+1} = L_{k,i_k}/2 = 2^{i_k} L_k/2$ (here L agrees with our notation M), which corresponds to choosing $\rho_d = 2$ in our method. Hence, the Lipschitz constant sequence of FGM is nonmonotone and changes with high oscillation. This verifies the claim in Remark 3.2 that large $\rho_d > 1$ reduces M_k locally but increases the burden on the line search procedure.

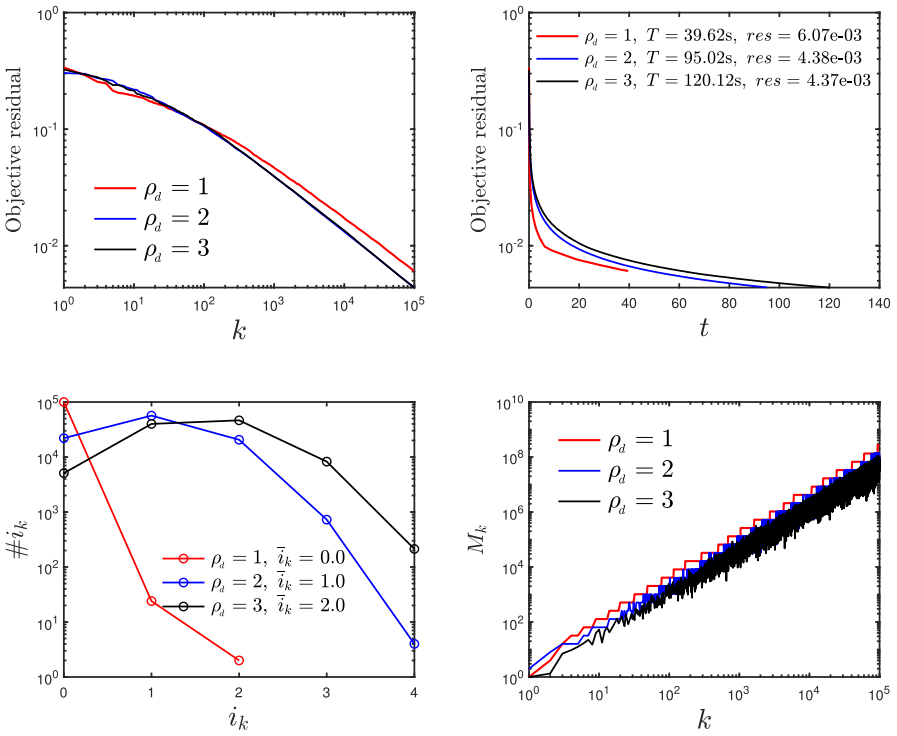

Fig. 4 Numerical performances of UAPD on the problem (53) with $\rho_u = 2$ and different ρ_d

In Fig. 3, we check also the difference between the Euclidean distance $\phi(x) = 1/2\,\|x\|^2$ and the entropy $\phi(x) = \langle x, \ln x\rangle$. It is observed that these two cases are very similar in line search part but the latter leads to better convergence rate.

Besides, we report the performance of our UAPD with different ρ_d. From Fig. 4, we see little improvement on the convergence rate with respect to the iteration k, and large ρ_d does not win smaller choices because it runs with more than triple time (120.12 s *v.s.* 39.62 s) but has not reduced the residual by third (6.07*e*-3 *v.s.* 4.37*e*-3). Moreover, the magnitude of M_k does not differ too much but large ρ_d call more line search steps, which increases the computational cost. In conclusion, for matrix game problem, $\rho_d = 1$ seems a doable choice.

6.2 Regularized Matrix Game Problem

The objective of (51) admits an approximation (cf. [46])

$$\varphi_\sigma(x) := \sigma \ln\left(\sum_{j=1}^{m} e^{\langle p_j, x\rangle/\sigma}\right), \tag{54}$$

where $\sigma > 0$ denotes the smoothing parameter. This regularized objective is smoother than the original one. According to [46, Eq.(4.8)], we choose $\sigma = \epsilon/(2\ln m)$, and the

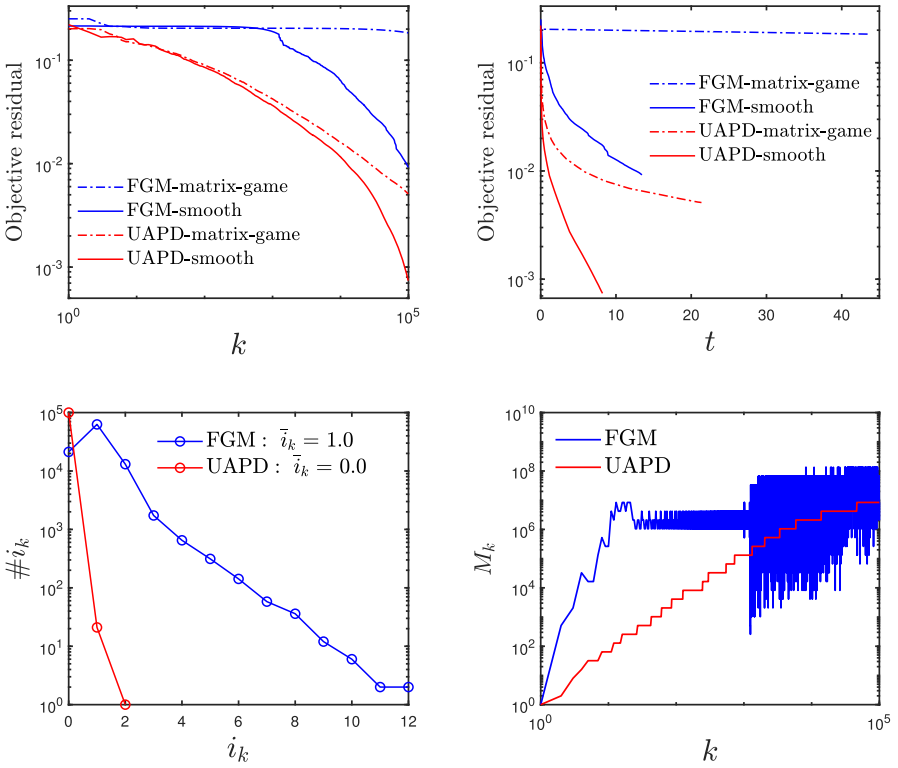

Fig. 5 Numerical performances of FGM and UAPD on minimizing the regularized objective (54) with $m = 100$, $n = 400$. The desired accuracy of FGM is $\epsilon = 1e\text{-}6$

Lipschitz constant of $\nabla\varphi_\sigma$ is $L_\sigma = \max_{i,j} |P_{i,j}|^2/(4\sigma)$. To avoid the overflow issue, we adopt the shifting technique [3].

We then apply UAPD and FGM (with $\epsilon = 1e\text{-}6$) to the smooth objective (54) and report the numerical outputs in Fig. 5. The optimal value φ_σ^* is obtained by running UAPD with enough iterations. Similarly as before, our UAPD is superior to FGM in convergence, line search cost and approximate Lipschitz constant. Also, we plot the results of the original matrix game problem (see the top row in Fig. 5) and find that with smoothing technique both two methods perform better than before.

6.3 Continuous Steiner Problem

Let us consider one more unconstrained problem

$$\min_{x\in\mathbb{R}^n_+} f(x) = \sum_{j=1}^{m} \|x - a_j\|, \tag{55}$$

where $a_j \in \mathbb{R}^n$ denotes the given location. Note that the objective is actually quite smooth far away from each a_j. We generate a_j from the uniform distribution and run

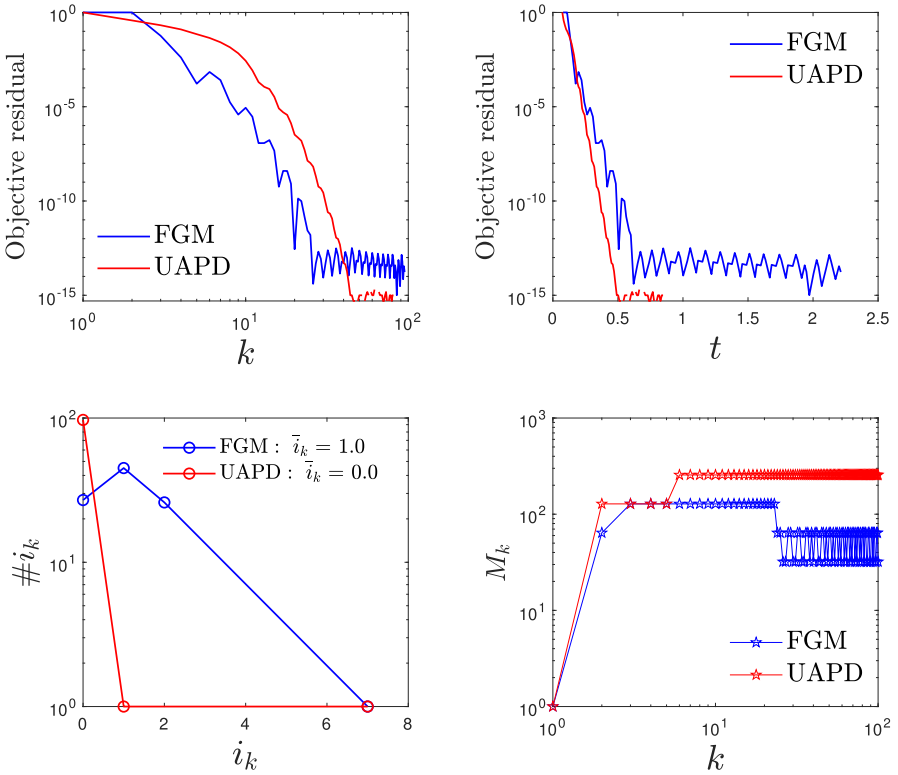

Fig. 6 Numerical results of FGM and UAPD for solving (55) with $m = 800,\ n = 400$. The desired accuracy of FGM is $\epsilon = 1e\text{-}8$

UAPD with enough times to obtain an approximated optimal value f^*. Numerical results in Fig. 6 show that both FGM (with $\epsilon = 1e\text{-}8$) and UAPD work well and possess similar convergence behaviors. Moreover, as ∇f is almost Lipschitz continuous and the magnitude of the Lipschitz constant is not large, the over-estimated issue of FGM is negligible, and the approximated constant M_k is the same as that of our UAPD.

6.4 Basis Pursuit Problem

In the last example, we look at the basis pursuit problem

$$\min_{x \in \mathbb{R}^n} \|x\|_1 \quad \text{s.t. } Ax = b,$$

where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. To be compatible with the problem setting of AccUniPDGrad [68], consider an equivalent formulation

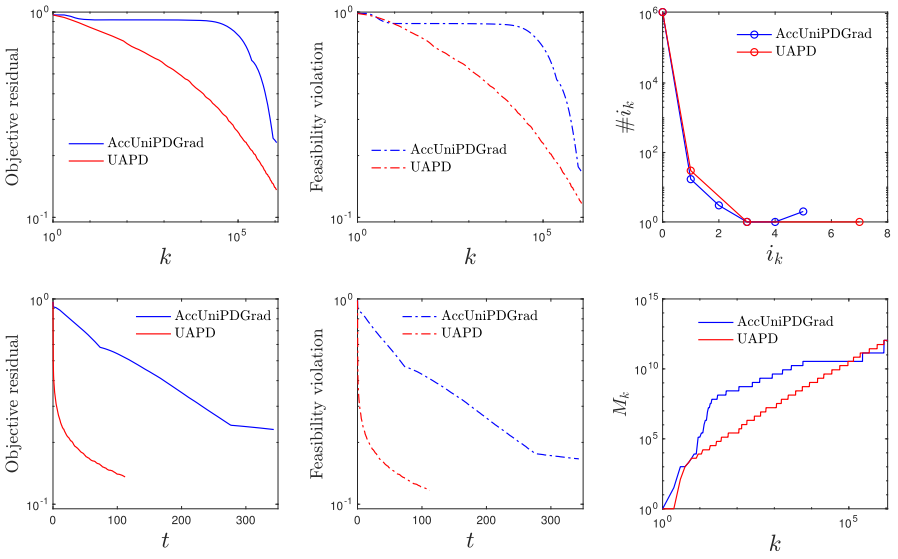

Fig. 7 Numerical performances of AccUniPDGrad and UAPD on the basis pursuit problem (56) with $m = 100$, $n = 400$. The desired accuracy for AccUniPDGrad is $\epsilon = 1e\text{-}3$

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}\|x\|_1^2 \quad \text{s.t. } Ax = b. \tag{56}$$

The dual problem reads as

$$\min_{\lambda\in\mathbb{R}^m} \left\{\varphi(\lambda) := \langle b, \lambda\rangle + \frac{1}{2}\left\|A^\top\lambda\right\|_\infty^2\right\}.$$

Note that existing accelerated methods in [29, 39, 65] can be applied to solving (56) with theoretical rate $\mathcal{O}(1/k)$. But we only focus on the comparison between UAPD and AccUniPDGrad, as black-box type methods with line search procedure. We mention that AccUniPDGrad also uses smaller tolerance ϵ/k as that in FGM, where $\epsilon > 0$ denotes the desired accuracy, and it takes $M_{k+1} = M_{k,i_k} = 2^{i_k} M_k$, which coincides with our choice $\rho_d = 1$.

To obtain a reasonable approximate minimal value f^*, we run the accelerated primal–dual method proposed in [39] with enough iterations. Numerical results are shown in Fig. 7, which indicate that (i) our UAPD has smaller objective residual and feasibility violation, (ii) the line search steps are very close and the Lipschitz constant sequences are nondecreasing (since $\rho_d = 1$), and (iii) AccUniPDGrad generates overestimated M_k because of its small tolerance ϵ/k.

Acknowledgements The author would like to thank the Editor and two anonymous referees, for their careful readings and valuable comments that improve significantly the early version of the paper.

Appendix A Proof of Lemma 3.1

Let us first prove (18). Recall that i_k is the smallest nonnegative integer such that (cf. (15))

$$h(x_{k,i_k}) - \Delta_{k,i_k} \le \frac{\delta_{k,i_k}}{2}.$$

If $i_k = 0$, then $M_{k+1} = M_{k,0}/\rho_d = M_k/\rho_d$. Otherwise (i.e., $i_k \ge 1$), we claim that

$$M_{k,i_k} \le \rho_u \cdot M(\nu, \delta_{k,i_k-1}). \tag{57}$$

If this is violated, then $M_{k,i_k-1} = M_{k,i_k}/\rho_u > M(\nu, \delta_{k,i_k-1})$. According to Proposition 2.1, this implies immediately that

$$h(x_{k,i_k-1}) - \Delta_{k,i_k-1} \le \frac{\delta_{k,i_k-1}}{2},$$

which yields a contradiction and thus verifies (57). Additionally, by (16), we have $\alpha_{k,i_k-1} \le \sqrt{\rho_u}\alpha_{k,i_k}$. Therefore, collecting (13,17,57) leads to

$$M_{k+1} = M_{k,i_k}/\rho_d \le \frac{\sqrt{\rho_u}\rho_u}{\rho_d} \cdot M(\nu, \delta_{k,i_k}) = \frac{\sqrt{\rho_u}\rho_u}{\rho_d} \cdot M(\nu, \delta_{k+1}). \tag{58}$$

This means that for all $k \in \mathbb{N}$, we have

$$M_{k+1} \le \frac{1}{\rho_d} \max\left\{M_k,\ \sqrt{\rho_u}\rho_u \cdot M(\nu, \delta_{k+1})\right\}. \tag{59}$$

Since $\delta_{k+1} = \beta_{k+1}/(k+1)$ with $\{\beta_k\}_{k\in\mathbb{N}}$ being decreasing (cf.(17)), it follows that $\delta_{k+1} \le \delta_{\ell+1}$ and $M(\nu, \delta_{\ell+1}) \le M(\nu, \delta_{k+1})$ for all $0 \le \ell \le k$, which together with (59) indicates the estimate

$$\begin{aligned} M_{k+1} &\le \frac{1}{\rho_d} \max\left\{\frac{1}{\rho_d} \max\left\{M_{k-1},\ \sqrt{\rho_u}\rho_u \cdot M(\nu, \delta_k)\right\},\ \sqrt{\rho_u}\rho_u \cdot M(\nu, \delta_{k+1})\right\} \\ &\le \frac{1}{\rho_d} \max\left\{\frac{M_{k-1}}{\rho_d},\ \sqrt{\rho_u}\rho_u \cdot M(\nu, \delta_{k+1})\right\} \\ &\le \cdots \le \frac{1}{\rho_d} \max\left\{\frac{M_0}{\rho_d^k},\ \sqrt{\rho_u}\rho_u \cdot M(\nu, \delta_{k+1})\right\}. \end{aligned}$$

This proves the desired result (18).

Then, let us verify (19). Observing that $M_{k+1} = M_{k,i_k}/\rho_d = \rho_u^{i_k} M_k/\rho_d$, we have

$$i_k = \log_{\rho_u} \frac{\rho_d M_{k+1}}{M_k} \quad \Longrightarrow \quad \sum_{j=0}^{k} i_j = \log_{\rho_u} \frac{\rho_d^{k+1} M_{k+1}}{M_0}.$$

Invoking the estimate of M_{k+1} gives

$$\begin{aligned}\sum_{j=0}^{k} i_j &\le \max\left\{0, \frac{3}{2} + k\log_{\rho_u}\rho_d + \log_{\rho_u}\frac{M(v,\delta_{k+1})}{M_0}\right\}\\ &\le \frac{3}{2} + k\log_{\rho_u}\rho_d + \left|\log_{\rho_u}\frac{M(v,\delta_{k+1})}{M_0}\right|.\end{aligned}$$

Since $M(v,\delta_{k+1}) = \delta_{k+1}^{\frac{v-1}{v+1}}[M_v(h)]^{\frac{2}{v+1}}$, we obtain (19) and complete the proof of Lemma 3.1.

Appendix B Derivation of the Reformulation (20a)

Observing line 10 of Algorithm 1, we obtain (20d). Given (x_k, v_k, λ_k) and (γ_k, β_k, M_k), (y_k, x_{k+1}, v_{k+1}) are nothing but the output of Algorithm 2 with the input (k, S_k, M_{k,i_k}), where $S_k = \{x_k, v_k, \lambda_k, \beta_k, \gamma_k\}$. Hence, from lines 3 and 4, we obtain

$$\begin{aligned}\frac{y_k - x_k}{\alpha_k} &= v_k - y_k,\\ \frac{x_{k+1} - x_k}{\alpha_k} &= v_{k+1} - x_{k+1},\end{aligned}$$

which gives (20a) and (20c). Besides, we have

$$v_{k+1} = \operatorname*{argmin}_{v\in Q}\left\{g(v) + \left\langle \nabla h(y_k) + A^\top\widetilde{\lambda}_k, v\right\rangle + \mu D_\phi(v, y_k) + \frac{\gamma_k}{\alpha_k}D_\phi(v, v_k)\right\},$$

with $\widetilde{\lambda}_k = \lambda_k + \alpha_k/\beta_k(Av_k - b)$. The optimality condition reads as

$$\begin{aligned}0 \in{}& \nabla h(y_k) + \partial g(v_{k+1}) + N_Q(v_{k+1}) + A^\top\widetilde{\lambda}_k\\ &+ \mu\left[\nabla\phi(v_{k+1}) - \nabla\phi(y_k)\right] + \frac{\gamma_k}{\alpha_k}\left[\nabla\phi(v_{k+1}) - \nabla\phi(v_k)\right].\end{aligned}$$

After rearranging, we get (20b).

Appendix C Proof of Lemma 5.1

Appendix C.1 The Case $\eta = \theta - 1$

The estimate (47) becomes

$$\sqrt{\varphi(t)}\frac{y'(t)}{y(t)} + R\frac{y'(t)}{y^\theta(t)} \le -\sigma(t). \tag{60}$$

Since $y(0) = 1$ and $y'(t) \leq 0$, it holds that $0 < y(t) \leq 1$ for all $t \geq 0$. As $\varphi(t)$ is positive and nondecreasing, we obtain

$$\left(\sqrt{\varphi}\ln y\right)' = \frac{\varphi'}{2\sqrt{\varphi}}\ln y + \sqrt{\varphi}\frac{y'}{y} \leq \sqrt{\varphi}\frac{y'}{y}.$$

Combining this with (60) gives

$$\left(\sqrt{\varphi(t)}\ln y(t) + \frac{R}{1-\theta}y^{1-\theta}(t)\right)' \leq -\sigma(t),$$

and integrating over $(0, t)$ leads to

$$\sqrt{\varphi(t)}\ln\frac{1}{y(t)} + \frac{R}{\theta-1}\left(y^{1-\theta}(t) - 1\right) \geq \int_0^t \sigma(s)\,\mathrm{d}s = \Sigma(t). \tag{61}$$

Define

$$Y_1(t) := \exp\left(-\frac{\Sigma(t)}{2\sqrt{\varphi(t)}}\right) \quad \text{and} \quad Y_2(t) := \left(1 + \frac{\theta-1}{2R}\Sigma(t)\right)^{\frac{1}{1-\theta}}. \tag{62}$$

Then one finds that

$$\begin{aligned}
\sqrt{\varphi(t)}\ln\frac{1}{Y_1(t)} &= \frac{1}{2}\Sigma(t), & Y_1(0) &= 1,\\
\frac{R}{\theta-1}\left(Y_2^{1-\theta}(t) - 1\right) &= \frac{1}{2}\Sigma(t), & Y_2(0) &= 1.
\end{aligned}$$

This also implies

$$\sqrt{\varphi(t)}\ln\frac{1}{Y(t)} + \frac{R}{\theta-1}\left(Y^{1-\theta}(t) - 1\right) \leq \Sigma(t), \tag{63}$$

where $Y(t) := Y_1(t) + Y_2(t)$. For fixed $t > 0$, the function

$$v \longmapsto \sqrt{\varphi(t)}\ln\frac{1}{v} + \frac{R}{\theta-1}\left(v^{1-\theta} - 1\right)$$

is monotonously decreasing in terms of $v \in (0, \infty)$. Collecting (61,63) yields that

$$y(t) \leq Y(t) = \exp\left(-\frac{\Sigma(t)}{2\sqrt{\varphi(t)}}\right) + \left(1 + \frac{\theta-1}{2R}\Sigma(t)\right)^{\frac{1}{1-\theta}}.$$

This completes the proof of Lemma 5.1 with $\eta = \theta - 1$.

Appendix C.2 The Case $\eta < \theta - 1$

The proof is in line with the previous case. With some elementary calculus computations, the estimate (61) now becomes

$$G(\varphi(t), y(t)) \geq \Sigma(t),$$

where $G : (0, \infty) \times (0, \infty) \to \mathbb{R}$ is defined by

$$G(w, v) := \frac{\sqrt{w}}{\theta - \eta - 1}(v^{\eta+1-\theta} - 1) + \frac{R}{\theta - 1}(v^{1-\theta} - 1),$$

for all $w,\ v > 0$. In addition to $Y_2(t)$ defined in (62), we need

$$Y_3(t) := \left(1 + \frac{\theta - \eta - 1}{2\sqrt{\varphi(t)}}\Sigma(t)\right)^{\frac{1}{\eta+1-\theta}}.$$

Since $G(w, \cdot)$ is monotonously decreasing and

$$G(\varphi(t), Y_2(t) + Y_3(t)) \leq \Sigma(t) \leq G(\varphi(t), y(t)),$$

we obtain

$$y(t) \leq \left(1 + \frac{\theta - 1}{2R}\Sigma(t)\right)^{\frac{1}{1-\theta}} + \left(1 + \frac{\theta - \eta - 1}{2\sqrt{\varphi(t)}}\Sigma(t)\right)^{\frac{1}{\eta+1-\theta}}.$$

This concludes the proof of Lemma 5.1 with $\eta < \theta - 1$.

References

1. Bauschke, H., Combettes, P.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS Books in Mathematics. Springer Science+Business Media, New York (2011)
2. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
3. Blanchard, P., Higham, D.J., Higham, N.J.: Accurately computing the log-sum-exp and softmax functions. IMA J. Numer. Anal. **41**(4), 2311–2330 (2021)
4. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2010)
5. Cai, J.-F., Osher, S., Shen, Z.: Linearized Bregman iterations for compressed sensing. Math. Comput. **78**(267), 1515–1536 (2009)
6. Candès, E., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
7. Chambolle, A., Pock, T.: A first-order primal–dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
8. Chambolle, A., Pock, T.: An introduction to continuous optimization for imaging. Acta Numer. **25**, 161–319 (2016)
9. Chen, G., Teboulle, M.: Convergence analysis of a proximal-like minimization algorithm using Bregman functions. SIAM J. Optim. **3**(3), 538–543 (1993)

10. Chen, L., Luo, H.: First order optimization methods based on Hessian-driven Nesterov accelerated gradient flow. arXiv:1912.09276 (2019)
11. Chen, L., Luo, H.: A unified convergence analysis of first order convex optimization methods via strong Lyapunov functions. arXiv: 2108.00132 (2021)
12. Chen, Y., Lan, G., Ouyang, Y.: Optimal primal–dual methods for a class of saddle point problems. SIAM J. Optim. **24**(4), 1779–1814 (2014)
13. Davis, D., Yin, W.: Convergence rate analysis of several splitting schemes. Splitting Methods in Communication, Imaging, Science, and Engineering, pages 115–163 (2016)
14. Demengel, F., Demengel, G., Erné, R.: Functional Spaces for the Theory of Elliptic Partial Differential Equations. Universitext. Springer, London (2012)
15. Devolder, O., Glineur, F., Nesterov, Y.: First-order methods of smooth convex optimization with inexact oracle. Math. Program. **146**(1–2), 37–75 (2014)
16. Dvurechensky, P, Gasnikov, A., Kroshnin, A.: Computational optimal transport: complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In: Proceedings of the 35 th International Conference on Machine Learning, volume 80, Stockholm, Sweden (2018). PMLR
17. Dvurechensky, P., Staudigl, M., Shtern, S.: First-order methods for convex optimization. arXiv:2101.00935 (2021)
18. Eckstein, J.: Splitting Methods for Monotone Operators with Applications to Parallel Optimization. PhD Thesis, Massachusetts Institute of Technology (1989)
19. Eckstein, J., Bertsekas, D.P.: On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**(1), 293–318 (1992)
20. Esser, E., Zhang, X., Chan, T.F.: A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. SIAM J. Imaging Sci. **3**(4), 1015–1046 (2010)
21. Feijer, D., Paganini, F.: Stability of primal–dual gradient dynamics and applications to network optimization. Automatica **46**(12), 1974–1981 (2010)
22. Goldstein, T., O'Donoghue, B., Setzer, S., Baraniuk, R.: Fast alternating direction optimization methods. SIAM J. Imaging Sci. **7**(3), 1588–1623 (2014)
23. Guminov, S., Gasnikov, A., Anikin, A., Gornov, A.: A universal modification of the linear coupling method. Optim. Methods Softw. **34**(3), 560–577 (2019)
24. Guminov, S.V., Nesterov, Y.E., Dvurechensky, P.E., Gasnikov, A.V.: Primal–dual accelerated gradient descent with line search for convex and nonconvex optimization problems. arXiv:1809.05895 (2018)
25. He, B., You, Y., Yuan, X.: On the convergence of primal-dual hybrid gradient algorithm. SIAM J. Imaging Sci. **7**(4), 2526–2537 (2014)
26. He, B., Yuan, X.: On the acceleration of augmented Lagrangian method for linearly constrained optimization. https://optimization-online.org/2010/10/2760/ (2010)
27. He, X., Hu, R., Fang, Y.-P.: Fast primal–dual algorithm via dynamical system for a linearly constrained convex optimization problem. Automatica **146**, 110547 (2022)
28. He, X., Hu, R., Fang, Y.-P.: Inertial accelerated primal–dual methods for linear equality constrained convex optimization problems. Numer. Algorithms **90**(4), 1669–1690 (2022)
29. Huang, B., Ma, S., Goldfarb, D.: Accelerated linearized Bregman method. J. Sci. Comput. **54**, 428–453 (2013)
30. Heinonen, J.: Lectures on Lipschitz analysis. Technical Report vol. 100, Rep. Univ. Jyväskylä Dept. Math. Stat., University of Jyväskylä (2005)
31. Jiang, F., Cai, X., Wu, Z., Han, D.: Approximate first-order primal-dual algorithms for saddle point problems. Math. Comput. **90**(329), 1227–1262 (2021)
32. Kamzolov, D., Dvurechensky, P., Gasnikov, A.: Universal intermediate gradient method for convex problems with inexact oracle. arXiv:1712.06036 (2019)
33. Kang, M., Kang, M., Jung, M.: Inexact accelerated augmented Lagrangian methods. Comput. Optim. Appl. **62**(2), 373–404 (2015)
34. Krichene, W., Bayen, A., Bartlett, P.: Accelerated mirror descent in continuous and discrete time. Adv. Neural Inf. Process. Syst. (NIPS) **28**, 2845–2853 (2015)
35. Li, H., Fang, C., Lin, Z.: Convergence rates analysis of the quadratic penalty method and its applications to decentralized distributed optimization. arXiv:1711.10802 (2017)
36. Li, H., Lin, Z.: Accelerated alternating direction method of multipliers: an optimal $O(1/K)$ nonergodic analysis. J. Sci. Comput. **79**(2), 671–699 (2019)

37. Lin, T., Ho, N., Jordan, M.I.: On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In International Conference on Machine Learning, pp. 3982–3991. PMLR (2019)
38. Luo, H.: Accelerated differential inclusion for convex optimization. Optimization (2021). https://doi.org/10.1080/02331934.2021.2002327
39. Luo, H.: Accelerated primal-dual methods for linearly constrained convex optimization problems. arXiv:2109.12604 (2021)
40. Luo, H., Zhang, Z.-H.: A unified differential equation solver approach for separable convex optimization: splitting, acceleration and nonergodic rate. arXiv:2109.13467 (2023)
41. Luo, H.: A primal–dual flow for affine constrained convex optimization. ESAIM: Control Optim. Calc. Var. **28**, 33 (2022)
42. Luo, H., Chen, L.: From differential equation solvers to accelerated first-order methods for convex optimization. Math. Program. (2021). https://doi.org/10.1007/s10107-021-01713-3
43. Nbmirovskii, A.S., Nrsterov, Y.E.: Optimal methods of smooth convex minimization. USSR Comput. Math. Math. Phys. **25**(2), 21–30 (1985)
44. Nemirovsky, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. John Wiley & Sons, New York (1983)
45. Nesterov, Y.: Introductory Lectures on Convex Optimization. Applied Optimization, vol. 87. Springer, US, Boston, MA (2004)
46. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
47. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. Ser. B **140**(1), 125–161 (2013)
48. Nesterov, Y.: Universal gradient methods for convex optimization problems. Math. Program. **152**, 381–404 (2015)
49. Nesterov, Y.: Lectures on Convex Optimization. Springer Optimization and Its Applications, vol. 137. Springer International Publishing, Cham (2018)
50. Ouyang, Y., Chen, Y., Lan, G., Pasiliao, E.: An accelerated linearized alternating direction method of multipliers. SIAM J. Imaging Sci. **8**(1), 644–681 (2015)
51. Ouyang, Y., Xu, Y.: Lower complexity bounds of first-order methods for convex–concave bilinear Saddle-point problems. Math. Program. **185**(1–2), 1–35 (2021)
52. Roulet, V., d'Aspremont, A.: Sharpness, restart, and acceleration. In: 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA (2017)
53. Sabach, S., Teboulle, M.: Faster Lagrangian-based methods in convex optimization. SIAM J. Optim. **32**(1), 204–227 (2022)
54. Stonyakin, F., Dvinskikh, D., Dvurechensky, P., Kroshnin, A., Kuznetsova, O., Agafonov, A., Gasnikov, A., Tyurin, A., Uribe, C.A., Pasechnyuk, D., Artamonov, S.: Gradient methods for problems with inexact model of the objective. arXiv:1902.09001 (2019)
55. Stonyakin, F., Gasnikov, A., Dvurechensky, P., Alkousa, M., Titov, A.: Generalized mirror prox for monotone variational inequalities: Universality and inexact oracle. arXiv:1806.05140 (2022)
56. Tao, M., Yuan, X.: Accelerated Uzawa methods for convex optimization. Math. Comput. **86**(306), 1821–1845 (2016)
57. Tian, W., Yuan, X.: An alternating direction method of multipliers with a worst-case $O(1/n^2)$ convergence rate. Math. Comput. **88**(318), 1685–1713 (2018)
58. Tran-Dinh, Q.: A unified convergence rate analysis of the accelerated smoothed gap reduction algorithm. Optim. Lett. (2021). https://doi.org/10.1007/s11590-021-01775-4
59. Tran-Dinh, Q., Fercoq, O., Cevher, V.: A smooth primal-dual optimization framework for nonsmooth composite convex minimization. SIAM J. Optim. **28**(1), 96–134 (2018)
60. Tran-Dinh, Q., Zhu, Y.: Augmented Lagrangian-based decomposition methods with non-ergodic optimal rates. arXiv:1806.05280 (2018)
61. Tran-Dinh, Q., Zhu, Y.: Non-stationary first-order primal–dual algorithms with faster convergence rates. SIAM J. Optim. **30**(4), 2866–2896 (2020)
62. Valkonen, T.: Inertial, corrected, primal–dual proximal splitting. SIAM J. Optim. **30**(2), 1391–1420 (2020)
63. Wibisono, A., Wilson, A.C., Jordan, M.I.: A variational perspective on accelerated methods in optimization. Proc. Natl. Acad. Sci. **113**(47), E7351–E7358 (2016)
64. Wilson, A., Recht, B., Jordan, M.: A Lyapunov analysis of momentum methods in optimization. arXiv:1611.02635 (2016)

65. Xu, Y.: Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming. SIAM J. Optim. **27**(3), 1459–1484 (2017)
66. Xu, Y.: Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. Math. Program. **185**(1–2), 199–244 (2021)
67. Yin, W., Osher, S., Goldfarb, D., Darbon, J.: Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. SIAM J. Imaging Sci. **1**(1), 143–168 (2008)
68. Yurtsever, A., Tran-Dinh, Q., Cevher, V.: A universal primal-dual convex optimization framework. arXiv: 1502.03123 (2015)
69. Zhao, Y., Liao, X., He, X., Li, C.: Accelerated primal-dual mirror dynamical approaches for constrained convex optimization. arXiv:2205.15983 (2022)