ACCELERATED OVER-RELAXATION HEAVY-BALL METHOD: ACHIEVING GLOBAL ACCELERATED CONVERGENCE WITH BROAD GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The heavy-ball momentum method accelerates gradient descent with a momentum term but lacks accelerated convergence for general smooth strongly convex problems. This work introduces the Accelerated Over-Relaxation Heavy-Ball (AOR-HB) method, the first variant with provable global and accelerated convergence for such problems. AOR-HB closes a long-standing theoretical gap, extends to composite convex optimization and min-max problems, and achieves optimal complexity bounds. It offers three key advantages: (1) broad generalization ability, (2) potential to reshape acceleration techniques, and (3) conceptual clarity and elegance compared to existing methods.

1 INTRODUCTION

We first consider the convex optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x), \tag{1}$$

where the objective function f is μ-strongly convex and L-smooth. Later on we consider extension to composite convex optimization $\min_x f(x) + g(x)$ with a non-smooth function g, and a class of min-max problems $\min_{u\in\mathbb{R}^m} \max_{p\in\mathbb{R}^n} f(u) - g(p) + \langle Bu, p\rangle$ with bilinear coupling.

Notation. $\mathbb{R}^d$ is d-dimensional Euclidean space with standard ℓ_2-inner product $\langle\cdot,\cdot\rangle$ and the induced norm $\|\cdot\|$. $f:\mathbb{R}^d\to\mathbb{R}$ is a differentiable function. We say f is μ-strongly convex function when there exists $\mu>0$ such that

$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle \ge \frac{\mu}{2}\|x-y\|^2, \quad \forall\, x, y \in \mathbb{R}^d.$$

We say f is L-smooth with $L>0$ if its gradient is Lipschitz continuous:

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|, \quad \forall\, x, y \in \mathbb{R}^d.$$

The condition number κ is defined as $\kappa = L/\mu$.

When f is μ-strongly convex and L-smooth, the optimization problem (1) has a unique global minimizer x^*. We focus on iterative methods to find x^*. A useful measure of convergence is the Bregman divergence of f, defined as $D_f(y,x) := f(y) - f(x) - \langle \nabla f(x), y-x\rangle$. In particular, $D_f(x,x^*) = f(x) - f(x^*)$, since $\nabla f(x^*) = 0$. Various bounds and identities on the Bregman divergence can be found in Appendix A.

Heavy-ball methods and flow. Over the past two decades, first-order methods, which rely solely on gradient information rather than the Hessian as required by Newton's method, have gained significant interest due to their efficiency and adaptability to large-scale data-driven applications and machine learning tasks (Bottou et al., 2018). Among these methods, the gradient descent method is the most straightforward and well-established algorithms. However, for ill-conditioned problems, where the condition number $\kappa_f \gg 1$, the gradient descent method suffers from slow convergence.

In order to accelerate the gradient descent methods, a momentum term was introduced, encouraging the method to move along search directions that utilize not only current but also previously seen information. The heavy-ball (HB) method (also known as the momentum method (Polyak, 1964) was in the form of:

$$x_{k+1} = x_k - \gamma \nabla f(x_k) + \beta(x_k - x_{k-1}), \tag{2}$$

where (β, γ) are constant parameters. Polyak motivated the method by an analogy to a "heavy ball" moving in a potential well defined by the objective function f. The corresponding ordinary differential equation (ODE) model is commonly referred to as the heavy-ball flow (Polyak, 1964):

$$x'' + \theta x' + \eta \nabla f(x) = 0, \tag{3}$$

where $x = x(t)$, x' is taking the derivative of t, and (θ, η) are positive constant parameters.

Non-convergence and non-acceleration of the HB method. Polyak (1964) showed that for (2) with $\beta = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2$, $\gamma = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$, and when x_k is sufficiently close to the optimal solution x^*, $\|x_k - x^*\|$ converges at rate $\frac{1-\sqrt{\rho}}{1+\sqrt{\rho}}$, where $\rho = 1/\kappa$. Polyak's choice relies on the spectral analysis of the linear system and thus the accelerated rate is only limited to the convex and quadratic objectives and is local for iterates near x^*. Indeed Lessard et al. (2016) designed a non-convergent example to show that the Polyak's choice of parameters does not guarantee the global convergence for general strongly convex optimization. By changing the parameters (β, γ), in Ghadimi et al. (2015); Sun et al. (2019); Saab Jr et al. (2022); Shi et al. (2022), the global linear convergence of the HB method has been established. However, the best rate is $1 - \mathcal{O}(\rho)$ given by Shi et al. (2022), which coincides with that of the gradient descent and not the accelerated rate $1 - \mathcal{O}(\sqrt{\rho})$.

The absence of acceleration is not due to a technical difficulty in the convergence analysis. Recently, Goujaud et al. (2023) have demonstrated that the HB method provably fails to achieve an accelerated convergence rate for smooth and strongly convex problems. Specifically, for any positive parameters (β, γ) in (2), either there exists an L-smooth, μ-strongly convex function, and an initialization such that HB fails to converge; or even in the class of smooth and strongly convex quadratic function f, the convergence rate is not accelerated: $1 - \mathcal{O}(\rho)$. For more related works on HB methods, see Appendix B.1.

Accelerated first-order methods. To accelerate the HB method, one can introduce either an additional gradient step or a line search step with adequate decay into the algorithm; see Wilson et al. (2021); Siegel (2019); Chen & Luo (2021).

Nesterov accelerated gradient (NAG) method (Nesterov, 1983, page 81) can be viewed as an alternative enhancement of the HB method. Nesterov's approach calculates the gradient at points that are extrapolated based on the inertial force:

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \gamma \nabla f(x_k + \beta(x_k - x_{k-1})), \text{ with } \beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}, \gamma = \frac{1}{L}. \tag{4}$$

Nesterov devised the method of estimate sequences (Nesterov, 2013) to prove that (4) achieves the accelerated linear convergence rate $1 - \sqrt{\rho}$.

Later on, numerous accelerated gradient methods have been developed for smooth strongly convex optimization problems (Lin et al., 2015; Drusvyatskiy et al., 2018; Bubeck et al., 2015; Aujol et al., 2022; Van Scoy et al., 2017; Cyrus et al., 2018); to name just a few. However, little is known beyond convex optimization. One reason is that the techniques developed are often specialized on the convexity of the objective function, making them difficult to extend to non-convex cases. In machine learning terms, these approaches lack generalization ability.

Main contributions for smooth strongly convex optimization. We propose a variant of the HB method in the form of

$$\begin{gathered} x_{k+1} = x_k - \gamma(2\nabla f(x_k) - \nabla f(x_{k-1})) + \beta(x_k - x_{k-1}), \\ \gamma = \frac{1}{(\sqrt{L}+\sqrt{\mu})^2}, \qquad \beta = \frac{L}{(\sqrt{L}+\sqrt{\mu})^2}. \end{gathered} \tag{5}$$

The most notable yet simple change is using $2\nabla f(x_k) - \nabla f(x_{k-1})$ not $\nabla f(x_k)$ to approximate $\nabla f(x)$. Namely an over-relaxation technique (Hadjidimos, 1978) is applied to the gradient term. Therefore, we name (5) accelerated over-relaxation heavy-ball (AOR-HB) method.

Rather than second-order ODEs, we consider a first-order ODE system proposed in Chen & Luo (2021), referred to as the rotated gradient flow:

$$x' = y - x, \quad y' = x - y - \frac{1}{\mu}\nabla f(x), \tag{6}$$

which is a special case of the HB flow (3) with $\theta = 2$ and $\eta = \frac{1}{\mu}$ when y is eliminated. Although (6) is mathematically equivalent to an HB flow, the structure of this 2×2 first-order ODE system is crucial for convergence analysis and algorithmic design. Interestingly, the same second-order ODE can correspond to different first-order systems. Another one equivalent to (3) is

$$x' = v, \quad v' = -\theta v - \eta \nabla f(x), \tag{7}$$

where v represents velocity, offering a physical interpretation. However, the convergence analysis is less transparent in the form (7) or the original second-order ODE (3).

We obtain AOR-HB by discretization of (6) with two iterates (x_k, y_k) cf. (14). The AOR is used to symmetrize the error equation, cf. (19). The choice of parameters (β, γ) in (5) is derived from the step-size $\alpha = \sqrt{\mu/L}$. We rigorously prove that AOR-HB enjoys the global linear convergence with accelerated rate, which closes a long-standing theoretical gap in optimization theory.

Theorem 1.1 (Convergence of AOR-HB method)**.** *Suppose f is μ-strongly convex and L-smooth. Let (x_k, y_k) be generated by scheme (14) with initial value (x_0, y_0) and step size $\alpha = \sqrt{\mu/L}$. Then there exists a constant $C_0 = C_0(x_0, y_0, \mu, L)$ so that we have the accelerated linear convergence*

$$f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|y_{k+1} - x^*\|^2 \le C_0 \left(\frac{1}{1 + \frac{1}{2}\sqrt{\mu/L}}\right)^k, \quad k \ge 1. \tag{8}$$

Remark on Non-Strongly Convex Optimization. For convex optimization problems where $\mu = 0$, we propose a variant of the AOR-HB method that incorporates the dynamic time rescaling introduced in (Luo & Chen, 2022):

$$x_{k+1} = x_k - \frac{k}{k+3}\frac{1}{L}\big(2\nabla f(x_k) - \nabla f(x_{k-1})\big) + \frac{k}{k+3}(x_k - x_{k-1}), \tag{9}$$

which achieves an accelerated rate of $\mathcal{O}(1/k^2)$, comparable to NAG. Convergence analysis can be performed using a perturbation argument similar to that in (Lessard et al., 2016, Section 5.4).

Relation to other acceleration methods. The AOR term $2\nabla f(x_k) - \nabla f(x_{k-1})$ can be treated as adding a gradient correction in the high resolution ODE model (Shi et al., 2022). AOR-HB (5) can be also rendered as a special case of the 'SIE' iteration described in Zhang et al. (2021). However, the parameter choice ($\sqrt{s} = \frac{1}{L}$ and $m = 1$) recovering AOR-HB, does not satisfy the condition in their convergence analysis (Zhang et al., 2021, Theorem 3).

With a proper change of variables, NAG (4) can be seen as a parameter perturbation of the AOR-HB in the form of (5). The performance of AOR-HB are thus comparable to NAG but less precise than triple momentum (TM) methods (Van Scoy et al., 2017) for strongly convex optimization.

However, extending these methods (NAG, TM, and high-resolution ODEs) beyond the convex optimization is rare and challenging. In contrast, AOR-HB has a superior generalization capability as the true driving force of acceleration is better captured by the 2×2 first-order ODE model (6).

Extension to Composite Convex Minimization. Consider the optimization problem:

$$\min_x f(x) + g(x), \tag{10}$$

where f is μ-strongly convex and L-smooth, and g is a non-smooth convex function. To handle this, we only need to split the gradient term in (6) into $\nabla f(x) + \partial g(y)$ and use an implicit discretization scheme for y. This is equivalent to applying the proximal operator

$$\mathrm{prox}_{\lambda g}(x) := \min_y g(y) + \frac{1}{2\lambda}\|y - x\|^2,$$

which we assume is available. The convergence analysis of the AOR-HB-composite Algorithm 2 directly follows from Theorem 1.1. Numerically, Algorithm 2 also performs well for certain non-smooth and non-convex functions g when its proximal operator is available, further demonstrating the generalization capability of the AOR-HB method.

Extension to a class of saddle point problems. We extend the AOR-HB method to a strongly-convex-strongly-concave saddle point system with bilinear coupling, defined as follows:

$$\min_{u\in\mathbb{R}^m}\max_{p\in\mathbb{R}^n} \mathcal{L}(u,p) := f(u) - g(p) + \langle Bu, p\rangle \tag{11}$$

where $B \in \mathbb{R}^{n\times m}$ is a matrix and $B^\top$ denotes its transpose, $f : \mathbb{R}^m \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are strongly convex functions with convexity constant μ_f and μ_g, respectively. Problem (11) has a large number of applications, some of which we briefly introduce in Appendix B.3.

Only few optimal first-order algorithms for saddle point problems have been developed recently (Metelev et al., 2024; Thekumparampil et al., 2022; Kovalev et al., 2022; Jin et al., 2022). We design AOR-HB-saddle method (Algorithm 3), which achieves more efficient and significantly improved performance compared to these methods, while preserving the simplicity of the HB method and offering a solid convergence theory.

THEOREM 1.2 (Convergence of AOR-HB-saddle method)**.** *Suppose f is μ_f-strongly convex and L_f-smooth, g is μ_g-strongly convex and L_g-smooth and let* $\sqrt{\rho} = (\sqrt{2}-1)\min\left\{\sqrt{\frac{\mu_f}{L_f}}, \sqrt{\frac{\mu_g}{L_g}}, \frac{\sqrt{\mu_f\mu_g}}{\|B\|}\right\}$. *Let (u_k, v_k, p_k, q_k) be generated by Algorithm 3 with initial value (u_0, v_0, p_0, q_0) and step size $\alpha = \sqrt{\rho}$. Then there exists a non-negative constant $C_0 = C_0(u_0, v_0, p_0, q_0, \mu_f, L_f, \mu_g, L_g)$ so that we have the linear convergence*

$$D_f(u_{k+1}, u^*) + D_g(p_{k+1}, p^*) + \frac{\mu_f}{2}\|v_{k+1} - u^*\|^2 + \frac{\mu_g}{2}\|q_{k+1} - p^*\|^2 \le C_0\left(\frac{1}{1+\frac{1}{2}\sqrt{\rho}}\right)^k.$$

Extension to a class of monotone operator equations. At the continuous level, AOR-HB extends the rotated gradient flow Chen & Luo (2021). At the discrete level, the AOR discretization technique produces accelerated algorithms. Our approach provides a unified framework for convex optimization, saddle-point problems, and monotone operator equations. Specifically, these cases can be represented by the following form:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} -I & I \\ I - \mu^{-1}A & -I - \mu^{-1}N \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

- Strongly convex optimization: $A(x) = \nabla f(x)$, $N = 0$.
- Composite convex optimization: $A(x) = \nabla f(x)$, $N(y) = \partial g(y)$.
- Saddle-point problem with bilinear coupling: $A(x) = \begin{pmatrix} \nabla f(u) & 0 \\ 0 & \nabla g(p) \end{pmatrix}$, $N(x) = \begin{pmatrix} 0 & B^\top \\ B & 0 \end{pmatrix}$.
- A class of monotone operators: $A(x) = \nabla F(x)$, N is linear and skew-symmetric.

Accelerated Gradient and Skew-Symmetric Splitting (AGSS) methods can be developed, significantly expanding the scope of acceleration beyond its traditional boundaries. Due to space constraints, we will address the most general form of AGSS in future work and focus on AOR-HB method for $\mu > 0$ to present the core ideas more clearly.

We summarize the merits of AOR-HB approach as follows:
1. Generality: It applies to a broader class of optimization problems.
2. Impact: It introduces new techniques that may greatly influence the understanding of acceleration.
3. Elegance: It is more conceptually clear and intuitive.

2 AOR-HB METHOD FOR CONVEX OPTIMIZATION

Strong Lyapunov property. To simplify notation, introduce $\mathbf{z} = (x, y)^\top$ and $\mathbf{z}^* = (x^*, x^*)^\top$. Let $\mathcal{G}(\mathbf{z})$ be a vector field with $\mathcal{G}(\mathbf{z}^*) = 0$. For a generic ODE $\mathbf{z}' = \mathcal{G}(\mathbf{z})$, a Lyapunov function is a non-negative function $\mathcal{E}(\mathbf{z})$ satisfying $-\langle \nabla\mathcal{E}(\mathbf{z}), \mathcal{G}(\mathbf{z})\rangle \ge 0$ for all $\mathbf{z}$ near $\mathbf{z}^*$ and $\mathcal{E}(\mathbf{z}) = 0$ if and only if $\mathbf{z} = \mathbf{z}^*$. In order to get the exponential stability, Chen & Luo (2021) introduce the so-called strong Lyapunov property: there exists $c > 0$ such that

$$-\langle \nabla\mathcal{E}(\mathbf{z}), \mathcal{G}(\mathbf{z})\rangle \ge c\,\mathcal{E}(\mathbf{z}) \quad \forall \mathbf{z} \in \mathbb{R}^d. \tag{12}$$

LEMMA 2.1. *Let $\mathbf{z}' = \mathcal{G}(\mathbf{z})$ with $\mathbf{z}(0) = \mathbf{z}_0$. Assume there exists a Lyapunov function $\mathcal{E}(\mathbf{z})$ satisfying the strong Lyapunov property (12). Then we have the exponential stability*

$$\mathcal{E}(\mathbf{z}) \le e^{-ct}\mathcal{E}(\mathbf{z}_0).$$

Proof. By the chain rule and the strong Lyapunov property:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}\left(\mathbf{z}(t)\right) = \langle \nabla\mathcal{E}(\mathbf{z}), \mathbf{z}'\rangle = \langle \nabla \mathcal{E}(\mathbf{z}), \mathcal{G}(\mathbf{z})\rangle \le -c\mathcal{E}(\mathbf{z}(t)), \tag{13}$$

and the exponential stability follows. □

For the system (6), let $\mathcal{E}(x,y) = f(x) - f(x^*) + \frac{\mu}{2}\|y - x^*\|^2$. By direct calculation and the μ-convexity of f, we obtain the strong Lyapunov property

$$\begin{aligned} -\left\langle \nabla\mathcal{E}(x,y), \mathcal{G}(x,y)\right\rangle &= \langle \nabla f(x), x - x^*\rangle - \mu\langle x - y, y - x^*\rangle \\ &= \langle \nabla f(x), x - x^*\rangle - \frac{\mu}{2}\left(\|x - x^*\|^2 - \|x - y\|^2 - \|y - x^*\|^2\right) \\ &\ge f(x) - f(x^*) + \frac{\mu}{2}\|y - x^*\|^2 + \frac{\mu}{2}\|x - y\|^2 = \mathcal{E}(x,y) + \frac{\mu}{2}\|x - y\|^2. \end{aligned}$$

As shown above, the strong Lyapunov property and exponential stability can be established more straightforwardly using the first-order ODE system (6) rather than the original second-order ODE (3) or its equivalent form (7).

AOR-HB method for smooth strongly convex minimization problems. Based on discretization of (6), we propose an accelerated over-relaxation heavy-ball (AOR-HB) method:

$$\frac{x_{k+1} - x_k}{\alpha} = y_k - x_{k+1}, \tag{14a}$$

$$\frac{y_{k+1} - y_k}{\alpha} = x_{k+1} - y_{k+1} - \frac{1}{\mu}(2\nabla f(x_{k+1}) - \nabla f(x_k)). \tag{14b}$$

Setting $\alpha = \sqrt{\mu/L}$ and eliminating y_k leads to the formulation (5) and summarized in Algorithm 1. The gradient $\nabla f(x_{k+1})$ can be re-used in the next iteration and thus essentially only one gradient evaluation is required for one iteration in Algorithm 1.

Algorithm 1 Accelerated Over-Relaxation Heavy-Ball Method (AOR-HB)

```
Parameters: $x_0, x_1 \in \mathbb{R}^d, L, \mu$. Set $\gamma = \frac{1}{(\sqrt{L}+\sqrt{\mu})^2}, \beta = \frac{L}{(\sqrt{L}+\sqrt{\mu})^2}$.
for $k = 1, 2, \ldots$ do
    $x_{k+1} = x_k - \gamma(2\nabla f(x_k) - \nabla f(x_{k-1})) + \beta(x_k - x_{k-1})$
end for
return $x_{k+1}$
```

The convergence rate in Theorem 1.1 is global and accelerated, in the sense that to obtain $f(x_k) - f(x^*) \le \epsilon$ and $\|x_k - x^*\| \le \epsilon$, we need at most $\mathcal{O}\left(\sqrt{\kappa}|\log \epsilon|\right)$ iterations. The iteration complexity is optimal for first-order methods (Nesterov, 2013). We give a proof for Theorem 1.1 in Appendix C and will outline the key steps using quadratic objectives here.

Preliminaries for convergence analysis. For a symmetric matrix A, introduce $\langle x, y\rangle_A = \langle Ax, y\rangle = \langle x, Ay\rangle$ and $\|\cdot\|_A^2 := \langle\cdot,\cdot\rangle_A$. Consider the quadratic and convex function $f(x) = \frac{1}{2}\|x\|_A^2 - \langle b, x\rangle + c$ for $b \in \mathbb{R}^d, c \in \mathbb{R}$ and SPD matrix A with bound $0 < \mu \le \lambda(A) \le L$, where $\lambda(A)$ denotes a generic eigenvalue of matrix A, and $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are minimal and maximal eigenvalues of A, respectively.

For two symmetric matrices D, A, we denote $A \prec D$ if $D - A$ is symmetric and positive definite (SPD) and $A \preceq D$ if $D - A$ is positive semi-definite. One can easily verify that $A \preceq D$ is equivalent to $\lambda_{\max}(D^{-1}A) \le 1$ when D is non-singular or $\lambda_{\min}(A^{-1}D) \ge 1$ when A is non-singular.

Denote by $\mathcal{D} = \begin{pmatrix} A & 0 \\ 0 & \mu I \end{pmatrix}$ and $\mathcal{A}^{\mathrm{sym}} = \begin{pmatrix} 0 & A \\ A & 0 \end{pmatrix}$. Define two Lyapunov functions

$$\mathcal{E}(\mathbf{z}) := \frac{1}{2}\|\mathbf{z} - \mathbf{z}^*\|_{\mathcal{D}}^2, \quad \text{and} \quad \mathcal{E}^{\alpha}(\mathbf{z}) := \frac{1}{2}\|\mathbf{z} - \mathbf{z}^*\|_{\mathcal{D}+\alpha\mathcal{A}^{\mathrm{sym}}}^2.$$

By direct calculation, we have $\lambda(\mathcal{D}^{-1}\mathcal{A}^{\mathrm{sym}}) = \pm\sqrt{\lambda(A)/\mu}$ which implies

$$(1 - \alpha/\sqrt{\rho})\,\mathcal{D} \preceq \mathcal{D} + \alpha\mathcal{A}^{\mathrm{sym}} \preceq (1 + \alpha/\sqrt{\rho})\,\mathcal{D}, \qquad \rho = \mu/L. \tag{15}$$

Therefore for $0 \le \alpha \le \sqrt{\rho}$,

$$0 \le (1 - \alpha/\sqrt{\rho})\,\mathcal{E}(\mathbf{z}) \le \mathcal{E}^{\alpha}(\mathbf{z}) \le 2\mathcal{E}(\mathbf{z}), \quad \forall \mathbf{z} = (x,y)^\top \in \mathbb{R}^{2d}. \tag{16}$$

Convergence analysis for quadratic functions. The AOR-HB method (14) can be written as a correction of the implicit Euler discretization of the flow (6)

$$\mathbf{z}_{k+1} - \mathbf{z}_k = \alpha \mathcal{G}(\mathbf{z}_{k+1}) - \alpha \begin{pmatrix} 0 & I \\ \frac{1}{\mu} A & 0 \end{pmatrix} (\mathbf{z}_{k+1} - \mathbf{z}_k). \tag{17}$$

As $\mathcal{E}$ is quadratic, $D_{\mathcal{E}}(\mathbf{z}, \mathbf{y}) = \frac{1}{2}\|\mathbf{z} - \mathbf{y}\|_{\mathcal{D}}^2$. So use the definition of Bregman divergence, we have the identity for the difference of the Lyapunov function $\mathcal{E}$ at consecutive steps:

$$\mathcal{E}(\mathbf{z}_{k+1}) - \mathcal{E}(\mathbf{z}_k) = \langle \nabla \mathcal{E}(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{z}_k \rangle - \frac{1}{2}\|\mathbf{z}_k - \mathbf{z}_{k+1}\|_{\mathcal{D}}^2. \tag{18}$$

Substitute (17) into (18) and expand the cross term as:

$$\langle \nabla \mathcal{E}(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{z}_k \rangle = \alpha \langle \nabla \mathcal{E}(\mathbf{z}_{k+1}), \mathcal{G}(\mathbf{z}_{k+1}) \rangle - \alpha \langle \mathbf{z}_{k+1} - \mathbf{z}^*, \mathbf{z}_{k+1} - \mathbf{z}_k \rangle_{\mathcal{A}^{\text{sym}}}. \tag{19}$$

The AOR $Ax \approx 2Ax_{k+1} - Ax_k = Ax_{k+1} + A(x_{k+1} - x_k)$ is used to make the last term in (19) associated to a symmetric matrix $\mathcal{A}^{\text{sym}}$. With such symmetrization, we can use identity of squares

$$\langle \mathbf{z}_{k+1} - \mathbf{z}^*, \mathbf{z}_{k+1} - \mathbf{z}_k \rangle_{\mathcal{A}^{\text{sym}}} = \frac{1}{2}\left(\|\mathbf{z}_{k+1} - \mathbf{z}^*\|_{\mathcal{A}^{\text{sym}}}^2 + \|\mathbf{z}_{k+1} - \mathbf{z}_k\|_{\mathcal{A}^{\text{sym}}}^2 - \|\mathbf{z}_k - \mathbf{z}^*\|_{\mathcal{A}^{\text{sym}}}^2 \right).$$

Substitute back to (18) and rearrange terms, we obtain the following identity:

$$\mathcal{E}^{\alpha}(\mathbf{z}_{k+1}) - \mathcal{E}^{\alpha}(\mathbf{z}_k) = \alpha \langle \nabla \mathcal{E}(\mathbf{z}_{k+1}), \mathcal{G}(\mathbf{z}_{k+1}) \rangle - \frac{1}{2}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|_{\mathcal{D}+\alpha\mathcal{A}^{\text{sym}}}^2, \tag{20}$$

which holds for arbitrary step size α.

Now take $\alpha = \sqrt{\rho}$. Due to (15), the last term in (20) is non-positive. Use the strong Lyapunov property $-\langle \nabla \mathcal{E}(\mathbf{z}), \mathcal{G}(\mathbf{z}) \rangle \geq \mathcal{E}(\mathbf{z})$, and the bound $\mathcal{E}^{\alpha}(\mathbf{z}) \leq 2\mathcal{E}(\mathbf{z})$ in (16) to get

$$\mathcal{E}^{\alpha}(\mathbf{z}_{k+1}) - \mathcal{E}^{\alpha}(\mathbf{z}_k) \leq -\alpha \mathcal{E}(\mathbf{z}_{k+1}) \leq -\frac{\alpha}{2}\mathcal{E}^{\alpha}(\mathbf{z}_{k+1}), \tag{21}$$

which implies the global linear convergence of $\mathcal{E}^{\alpha}$:

$$\mathcal{E}^{\alpha}(\mathbf{z}_{k+1}) \leq \frac{1}{1+\alpha/2}\mathcal{E}^{\alpha}(\mathbf{z}_k) \leq \left(\frac{1}{1+\alpha/2}\right)^{k+1} \mathcal{E}^{\alpha}(\mathbf{z}_0), \quad k \geq 0. \tag{22}$$

Moreover, (21) implies

$$\mathcal{E}(\mathbf{z}_{k+1}) \leq \frac{1}{\alpha}(\mathcal{E}^{\alpha}(\mathbf{z}_k) - \mathcal{E}^{\alpha}(\mathbf{z}_{k+1})) \leq \frac{1}{\alpha}\mathcal{E}^{\alpha}(\mathbf{z}_k) \leq \frac{1}{\alpha}\left(\frac{1}{1+\alpha/2}\right)^{k} \mathcal{E}^{\alpha}(\mathbf{z}_0),$$

which implies (8) with $C_0 = \frac{1}{\alpha}\mathcal{E}^{\alpha}(\mathbf{z}_0) \geq 0$. Note that $C_0 = \mathcal{O}(\sqrt{\kappa})$ is large. However, when considering the convergence of $\mathcal{E}^{\alpha}$ as in (22), this dependency is absent.

AOR-HB method for composite convex optimization. For the composite convex optimization problem (10), AOR-HB can be seamlessly extended by using:

$$x' = y - x, \quad \mu y' - [\mu(x-y) - \nabla f(x)] \in \partial g(y), \tag{23}$$

where $\partial g(\cdot)$ denotes the set of subgradient. The strong Lyapunov property and exponential stability still hold (see Appendix D) as the modification only bring an additional monotone term.

We use implicit discretization in g and apply AOR to ∇f:

$$\frac{x_{k+1} - x_k}{\alpha} = y_k - x_{k+1}, \quad \mu\frac{y_{k+1} - y_k}{\alpha} - \mu(x_{k+1} - y_{k+1}) + (2\nabla f(x_{k+1}) - \nabla f(x_k)) \in \partial g(y_{k+1}).$$

Given the proximal operator of g, an equivalent and computation-favorable formulation is proposed in Algorithm 2. In the convergence analysis, since the non-smooth part is discretized implicitly, no difference arises in the error equation (17). Therefore, the convergence rate and proof are identical to those in Theorem 1.1; see Theorem D.1. In contrast, it requires considerable effort to generalize Nesterov's accelerated gradient to the composite case Beck & Teboulle (2009).

Algorithm 2 Accelerated Over-Relaxation Heavy-Ball Method for Convex Composite Minimization (AOR-HB-composite)

```
1: Parameters: $x_0, y_0 \in \mathbb{R}^d, L, \mu$. Set $\alpha = \sqrt{\frac{\mu}{L}}$ and $\lambda = \frac{\alpha}{(1+\alpha)\mu}$.
2: for $k = 1, 2, \dots$ do
3:     $x_{k+1} = \dfrac{1}{1+\alpha}(x_k + \alpha y_k)$
4:     $y_{k+1} = \mathrm{prox}_{\lambda g}(z_k)$ where $z_k = \frac{1}{1+\alpha}(y_k + \alpha x_{k+1}) - \lambda(2\nabla f(x_{k+1}) - \nabla f(x_k))$
5: end for
6: return $y_{k+1}$
```

3 AOR-HB METHOD FOR A CLASS OF SADDLE POINT PROBLEMS

In this section, we extend the flow (3) to min-max problems and propose an accelerated first-order method for strongly-convex-strongly-concave saddle point problems.

Strongly-convex-strongly-concave saddle point problem with bilinear coupling. The solution (u^*, p^*) to the min-max problem (11) is called a saddle point of $\mathcal{L}(u, p) = f(u) - g(p) + \langle Bu, p\rangle$:

$$\mathcal{L}(u^*, p) \le \mathcal{L}(u^*, p^*) \le \mathcal{L}(u, p^*), \quad \forall\, u \in \mathbb{R}^m, p \in \mathbb{R}^n.$$

As $\mathcal{L}(\cdot, p)$ is strongly convex for any given p and $\mathcal{L}(u, \cdot)$ is strongly concave for any given u, we called it a strongly-convex-strongly-concave saddle point problem. Under this setting, the saddle point (u^*, p^*) exists uniquely and is necessarily a critical point of $\mathcal{L}(u, p)$ satisfying:

$$\nabla \mathcal{L}(u^*, p^*) := \begin{pmatrix} \nabla f & B^\top \\ B & -\nabla g \end{pmatrix} \begin{pmatrix} u^* \\ p^* \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{24}$$

where $\langle \nabla f, u\rangle := \nabla f(u)$ and $\langle \nabla g, p\rangle := \nabla g(p)$.

HB-saddle flow. We propose the following HB-saddle flow for solving the min-max problem (11):

$$\begin{aligned} u' &= v - u, \quad v' = u - v - \frac{1}{\mu_f}(\nabla f(u) + B^\top q), \\ p' &= q - p, \quad q' = p - q - \frac{1}{\mu_g}(\nabla g(p) - Bv). \end{aligned} \tag{25}$$

The HB-saddle flow (25) preserves the strong Lyapunov property (see Appendix F.2) and thus the exponential stability of (u^*, p^*) for (25) follows.

For the gradient term, we use $J\nabla\mathcal{L}$ instead of $\nabla\mathcal{L}$ with $J = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}$. When ∇f and ∇g are linear, $\nabla\mathcal{L}$ is symmetric but not monotone. $J\nabla\mathcal{L}$ is non-symmetric but strongly monotone. Namely

$$\langle J\nabla\mathcal{L}(x) - J\nabla\mathcal{L}(y), x - y\rangle \ge \min\{\mu_f, \mu_g\}\|x - y\|^2, \quad \forall\, x = (u, p)^\top, y = (v, q)^\top.$$

The strong monotonicity has essential similarity to strongly convexity. Comparing to the convex optimization, one extra difficulty is the bilinear coupling $\langle Bu, p\rangle$ which can be again handled by the AOR technique in discretization.

AOR-HB-saddle method. We propose the accelerated over-relaxation heavy-ball method (AOR-HB-saddle) for solving the min-max problem (11) in Algorithm 3. Each iteration requires 2 matrix-vector products and 2 gradient evaluations if we store $\nabla f(u_k)$ and $\nabla g(p_k)$.

The convergence rate in Theorem 1.2 is global and accelerated, meaning that to obtain $\|(u_k, p_k) - (u^*, p^*)\| \le \epsilon$ and $\|(v_k, q_k) - (u^*, p^*)\| \le \epsilon$, we need at most $\mathcal{O}\left(\sqrt{L_f/\mu_f + L_g/\mu_g + \|B\|^2/(\mu_f\mu_g)}\,|\log\epsilon|\right)$ iterations. The iteration complexity is optimal for first-order methods for saddle point problems (Zhang et al., 2022). We refer to the proof of Theorem 1.2 in Appendix F.3.

Algorithm 3 Accelerated over-relaxation Heavy-ball method for strongly-convex-strongly-concave saddle point problems with bilinear coupling (AOR-HB-saddle)

```
1: Parameters: $u_0, v_0 \in \mathbb{R}^m, p_0, q_0 \in \mathbb{R}^n, L_f, L_g, \mu_f, \mu_g, \|B\|$.
2: Set $\alpha = (\sqrt{2}-1)\min\left\{\sqrt{\frac{\mu_f}{L_f}}, \sqrt{\frac{\mu_g}{L_g}}, \frac{\sqrt{\mu_f \mu_g}}{\|B\|}\right\}$.
3: for $k = 0, 1, 2, \dots$ do
4:     $u_{k+1} = \frac{1}{1+\alpha}(u_k + \alpha v_k); \quad p_{k+1} = \frac{1}{1+\alpha}(p_k + \alpha q_k)$
5:     $v_{k+1} = \frac{1}{1+\alpha}\left[v_k + \alpha u_{k+1} - \frac{\alpha}{\mu_f}(2\nabla f(u_{k+1}) - \nabla f(u_k) + B^\top q_k)\right]$
6:     $q_{k+1} = \frac{1}{1+\alpha}\left[q_k + \alpha p_{k+1} - \frac{\alpha}{\mu_g}(2\nabla g(p_{k+1}) - \nabla g(p_k) - B(2v_{k+1} - v_k))\right]$
7: end for
8: return $u_{k+1}, v_{k+1}, p_{k+1}, q_{k+1}$
```

***Remark* 3.1.** We can treat the coupling implicitly: Line 5 and 6 in Algorithm 3 are replaced by

$$\begin{aligned} v_{k+1} &= \frac{1}{1+\alpha}\left[v_k + \alpha u_{k+1} - \frac{\alpha}{\mu_f}(2\nabla f(u_{k+1}) - \nabla f(u_k) + B^\top q_{k+1})\right], \\ q_{k+1} &= \frac{1}{1+\alpha}\left[q_k + \alpha p_{k+1} - \frac{\alpha}{\mu_g}(2\nabla g(p_{k+1}) - \nabla g(p_k) - Bv_{k+1})\right]. \end{aligned} \tag{26}$$

Now (v_{k+1}, q_{k+1}) are coupled together and can be computed by inverting $\begin{pmatrix} (1+\alpha)I & \frac{\alpha}{\mu_f}B^\top \\ -\frac{\alpha}{\mu_g}B & (1+\alpha)I \end{pmatrix}$. It is sufficient to compute $\left((1+\alpha)^2 I + \frac{\alpha^2}{\mu_f \mu_g}BB^\top\right)^{-1}$ or $\left((1+\alpha)^2 I + \frac{\alpha^2}{\mu_f \mu_g}B^\top B\right)^{-1}$, whichever is a relative small size matrix and can be further replaced by an inexact inner solver. We name the method as AOR-HB-saddle-I(implicit). For AOR-HB-saddle-I, the convergence rate can be improved to $1 - \mathcal{O}(\sqrt{\rho})$ with $\rho = \min\{\mu_f/L_f, \mu_g/L_g\}$; see Appendix G for the formal results. It is preferable when the size $BB^\top$ or $B^\top B$ is small.

4 NUMERICAL RESULTS

In this section, we evaluate the performance of our AOR-HB methods on a suite of optimization problems. All numerical experiments were conducted using MATLAB R2022a on a desktop computer equipped with an Intel Core i7-6800K CPU operating at 3.4 GHz and 32GB of RAM. We compare the results against several state-of-the-art optimization algorithms from the literature.

4.1 SMOOTH CONVEX MINIMIZATION

We test our AOR-HB method (Algorithm 1) and compare it with other first-order algorithms including: (i) GD: gradient descent; (ii) NAG: Nesterov acceleration (Nesterov, 2013); (iii) HB: Polyak's momentum method (Polyak, 1964); (iv) TM: triple momentum method (Van Scoy et al., 2017) and (v) ADR (Aujol et al., 2022).

First, we test the algorithms using smooth multidimensional piecewise objective functions borrowed from Van Scoy et al. (2017). Let

$$f(x) = \sum_{i=1}^{p} h\left(a_i^\top x - b_i\right) + \frac{\mu}{2}\|x\|^2, \quad h(x) = \begin{cases} \frac{1}{2}x^2 e^{-r/x}, & x > 0 \\ 0, & x \le 0 \end{cases}, \tag{27}$$

where $A = [a_1, \dots, a_p] \in \mathbb{R}^{d \times p}$ and $b \in \mathbb{R}^p$ with $\|A\| = \sqrt{L-\mu}$. Then f is μ- strongly convex and L-smooth. We randomly generate the components of A and b from the normal distribution and then scale A so that $\|A\| = \sqrt{L-\mu}$. We set $\mu = 1, L = 10^4, d = 100, p = 5$ and $r = 10^{-6}$.

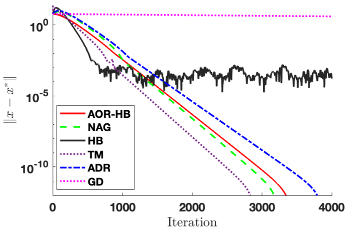

Figure 1: Simulation results for the objective function (27).

Figure 1 illustrates the logarithm of $\|x_k - x^*\|$ for each algorithm's iterates. The HB method reaches a plateau and fails to converge, while the other algorithms demonstrate global and linear convergence. As expected, GD, being non-accelerated, converges slowly due to the large condition number. The AOR-HB method performs as efficiently as NAG and ADR. The TM method achieves a convergence rate of $(1 - \sqrt{\mu/L})^2$, slightly outpacing other accelerated algorithms. However, TM requires more parameter tuning and has not yet been extended beyond strongly smooth convex optimization.

Next, we report the numerical simulations on a logistic regression problem with ℓ_2 regularizer:

$$\min_{x \in \mathbb{R}^d} \left\{ \sum_{i=1}^{m} \log\left(1 + \exp\left(-b_i a_i^\top x\right)\right) + \frac{\lambda}{2}\|x\|^2 \right\}, \tag{28}$$

where $(a_i, b_i) \in \mathbb{R}^d \times \{-1, 1\}, i = 1, 2, \ldots, m$. For the logistic regression problem, $\mu = \lambda$ and $L = \lambda_{\max}\left(\sum_{i=1}^m a_i a_i^\top\right) + \lambda$. The data a_i and b_i are generated by the normal distribution and Bernoulli distribution, respectively. We set $\lambda = 0.1, d = 1000$, and $n = 50$. As illustrated by Figure 2, this is an example where HB converges and it converges fastest. However, such fast convergence lacks theoretical guarantees and may fail in cases like the one depicted in Figure 1. This highlights the importance of robust theoretical convergence analysis rather than relying on empirical success alone.

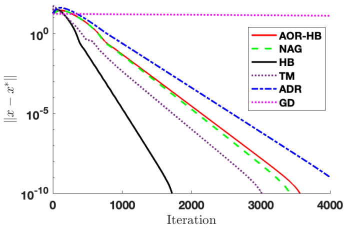

Figure 2: Simulation results for the logistic regression problem (28).

4.2 COMPOSITE CONVEX AND NON-CONVEX MINIMIZATION

We test the performance of AOR-HB-composite method (Algorithm 2) on non-smooth optimization problems and compare with two well-known methods: the fast iterative shrinkage-thresholding algorithm (FISTA) Beck & Teboulle (2009) and the accelerated proximal gradient (APG) algorithm proposed in Li & Lin (2015).

We first consider the Lasso problem:

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1,$$

which has wide applications in compressed sensing, signal processing, and statistical learning, among other fields (Tibshirani, 1996). We generate the matrix A with size 1024×256 from Gaussian random matrices and a ground-truth sparse vector x of sparsity 5. The data vector obtained by matrix-vector multiplication $b = Ax$. We set $\lambda = 0.8$ and use the step size $\alpha = 1/L$ in FISTA and APG. We plot the decay of the relative ℓ_2 error in Figure 3. All the methods cannot reach the accuracy of 10^{-3}. This is because the ground-truth solution may not be the minimizer of the corresponding minimization problem.

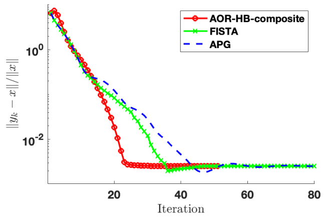

Figure 3: Comparison of ℓ_1 minimization methods.

Then we test the non-convex $\ell_1 - \ell_2$ case:

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \lambda(\|x\|_1 - \|x\|_2).$$

The proximal operator for $g(x) = \lambda(\|x\|_1 - \|x\|_2)$ is given in Yin et al. (2015). Using such non-smooth and non-convex function g will produce solution with better sparsity. The other settings remain the same as the Lasso problem. Figure 3 and 4 show the convergence of the relative ℓ_2 error, where all methods converges but AOR-HB outperforms others.

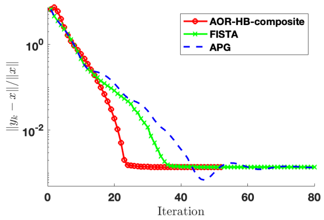

Figure 4: Comparison of $\ell_1 - \ell_2$ minimization methods.

4.3 SADDLE POINT PROBLEMS

We compare AOR-HB-saddle (Algorithm 3) with the following algorithms: (i) AG-OG: accelerated gradient-optimistic gradient method with restarting regime (Li et al., 2023); (ii) APDG: accelerated primal-dual gradient method (Kovalev et al., 2022); (iii) EG: a variant of extra-gradient method (Mokhtari et al., 2020).

We consider policy evaluation problems in reinforcement learning (Du et al., 2017) when finding minimum of the mean squared projected Bellman error (MSPBE):

$$\arg\min_{u} \frac{1}{2}\|Bu - b\|_{C^{-1}}^2 + \frac{1}{2}\|u\|^2.$$

The corresponding saddle point problem (see Appendix B.3) is

$$\min_{u\in\mathbb{R}^m} \max_{p\in\mathbb{R}^n} \frac{1}{2}\|u\|^2 - \frac{1}{2}\|p\|_C^2 - \langle b, p\rangle + \langle Bu, p\rangle. \tag{29}$$

The saddle point formulation saves the computation cost of inverting C. In this example, $\mu_f = L_f = 1$, $\mu_g = \lambda_{\min}(C)$ and $L_g = \lambda_{\max}(C)$. We generate random matrices B and C such that $\mu_g = 1$ and the condition number $\kappa_g = \|B\|^2$. We set $m = 2500$ and $n = 50$.

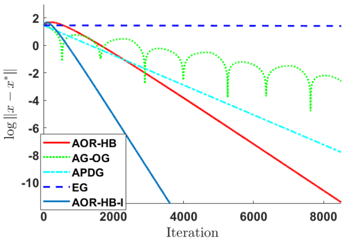

Figure 5: Simulation results for MSPBE (29): $\kappa_g = 10^4, \|B\| = 10^2$.

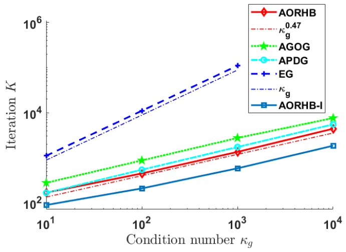

Figure 6: Iteration steps versus condition numbers.

We plot the convergence of error versus number of iterations in Figure 5 for the case where $\kappa_g = 10^4$. Different choices of condition numbers yield similar results, where our AOR-HB-saddle methods achieve much faster linear convergence than other algorithms. In particular, AOR-HB-saddle-I is the semi-implicit scheme (26) with step size $\alpha = \sqrt{1/\kappa(g)}$. We use the direct solver \ to compute $((1+\alpha)I + \frac{\alpha^2}{(1+\alpha)\mu_f\mu_g}BB^\top)^{-1}$ since it is a small 50×50 SPD matrix. AOR-HB-saddle-I converges fastest and is overall time-efficient especially when the coupling term $\|B\|^2/(\mu_f\mu_g)$ is larger than $\sqrt{\kappa_f}$ and $\sqrt{\kappa_g}$.

In Figure 6, we plot the iteration number K versus κ_g such that $\frac{\|x_K - x^*\|}{\|x_0 - x^*\|} \leq 10^{-6}$. In this log-log scale plot, we can observe that the growth of iteration complexity is $\mathcal{O}(\sqrt{\kappa_g})$ for accelerated algorithms and $\mathcal{O}(\kappa_g)$ for EG, which matches the convergence analysis.

Among all accelerated algorithms we have tested, our AOR-HB-saddle method requires fewer iteration steps to achieve the desired accuracy. In addition, AOR-HB-saddle has a single-loop structure and require fewer parameters to tune, which is favorable for implementation.

5 CONCLUDING REMARKS

In conclusion, we have introduced the Accelerated Over-Relaxation Heavy-Ball (AOR-HB) methods, a significant advancement of the accelerated first order optimization algorithms. Our method has been proven to exhibit global and accelerated convergence for not only smooth and strongly convex optimization problems but also a class of non-smooth optimization and min-max problems. This fills a theoretical gap in heavy-ball momentum methods and opens the door to developing accelerated methods with potential forays into non-convex optimization scenarios.

Although AOR-HB methods are accelerated first-order algorithms for both convex and non-convex optimization, they come with certain limitations. First, AOR-HB methods require the parameters μ and L. While this is typical for accelerated methods, exploring adaptive strategies to reduce parameter dependence is an interesting direction for future research. Second, the convergence rates established in some theorems are not as tight as those achieved by the TM method (Van Scoy et al., 2017; Taylor & Drori, 2023). Finally, investigating the algorithm's performance under stochastic conditions is essential for assessing its robustness in real-world machine learning applications.

REFERENCES

Hedy Attouch, Xavier Goudou, and Patrick Redont. The heavy ball with friction method, i. the continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. *Communications in Contemporary Mathematics*, 2(01):1–34, 2000.

J-F Aujol, Ch Dossal, and Aude Rondepierre. Convergence rates of the heavy ball method for quasi-strongly convex optimization. *SIAM Journal on Optimization*, 32(3):1817–1842, 2022.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Raghu Bollapragada, Tyler Chen, and Rachel Ward. On the fast convergence of minibatch heavy ball momentum. *arXiv preprint arXiv:2206.07553*, 2022.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Felix E Browder and Wolodymyr V Petryshyn. Construction of fixed points of nonlinear mappings in Hilbert space. *Journal of Mathematical Analysis and Applications*, 20(2):197–228, 1967.

Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

Long Chen and Hao Luo. A unified convergence analysis of first order convex optimization methods via strong Lyapunov functions. *arXiv preprint arXiv:2108.00132*, 4(9):10, 2021.

Saman Cyrus, Bin Hu, Bryan Van Scoy, and Laurent Lessard. A robust accelerated optimization algorithm for strongly convex functions. In *2018 Annual American Control Conference (ACC)*, pp. 1376–1381. IEEE, 2018.

Jelena Diakonikolas and Michael I Jordan. Generalized momentum-based methods: A Hamiltonian perspective. *SIAM Journal on Optimization*, 31(1):915–944, 2021.

Jelena Diakonikolas and Lorenzo Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization*, 29(1):660–689, 2019.

Dmitriy Drusvyatskiy, Maryam Fazel, and Scott Roy. An optimal first order method based on optimal quadratic averaging. *SIAM Journal on Optimization*, 28(1):251–271, 2018.

Simon S Du and Wei Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 196–205. PMLR, 2019.

Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pp. 1049–1058. PMLR, 2017.

Euhanna Ghadimi, Iman Shames, and Mikael Johansson. Multi-step gradient methods for networked optimization. *IEEE Transactions on Signal Processing*, 61(21):5417–5429, 2013.

Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pp. 310–315. IEEE, 2015.

Baptiste Goujaud, Adrien Taylor, and Aymeric Dieuleveut. Provable non-accelerations of the heavy-ball method. *arXiv preprint arXiv:2307.11291*, 2023.

Apostolos Hadjidimos. Accelerated overrelaxation method. *Mathematics of Computation*, 32(141): 149–157, 1978.

Yujia Jin, Aaron Sidford, and Kevin Tian. Sharper rates for separable minimax and finite sum optimization via primal-dual extragradient methods. In *Conference on Learning Theory*, pp. 4362–4415. PMLR, 2022.

Dmitry Kovalev, Alexander Gasnikov, and Peter Richtárik. Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling. *Advances in Neural Information Processing Systems*, 35:21725–21737, 2022.

Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. *Advances in neural information processing systems*, 28, 2015.

Qi Lei, Ian En-Hsu Yen, Chao-yuan Wu, Inderjit S Dhillon, and Pradeep Ravikumar. Doubly greedy primal-dual coordinate descent for sparse empirical risk minimization. In *International Conference on Machine Learning*, pp. 2034–2042. PMLR, 2017.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Chris Junchi Li, Huizhuo Yuan, Gauthier Gidel, Quanquan Gu, and Michael Jordan. Nesterov meets optimism: rate-optimal separable minimax optimization. In *International Conference on Machine Learning*, pp. 20351–20383. PMLR, 2023.

Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. *Advances in neural information processing systems*, 28, 2015.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. *Advances in neural information processing systems*, 28, 2015.

Fengshan Liu and M Zuhair Nashed. Regularization of nonlinear ill-posed variational inequalities and convergence rates. *Set-Valued Analysis*, 6(4):313–344, 1998.

Hao Luo and Long Chen. From differential equation solvers to accelerated first-order methods for convex optimization. *Mathematical Programming*, 195(1):735–781, 2022.

Dmitry Metelev, Alexander Rogozin, Alexander Gasnikov, and Dmitry Kovalev. Decentralized saddle-point problems with different constants of strong convexity and strong concavity. *Computational Management Science*, 21(1):5, 2024.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pp. 1497–1507. PMLR, 2020.

Michael Muehlebach and Michael Jordan. A dynamical systems perspective on nesterov acceleration. In *International Conference on Machine Learning*, pp. 4656–4662. PMLR, 2019.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543, 1983.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Peter Ochs, Thomas Brox, and Thomas Pock. iPiasco: inertial proximal algorithm for strongly convex optimization. *Journal of Mathematical Imaging and Vision*, 53:171–181, 2015.

Rui Pan, Yuxing Liu, Xiaoyu Wang, and Tong Zhang. Accelerated convergence of stochastic heavy ball method under anisotropic gradient noise. *arXiv preprint arXiv:2312.14567*, 2023.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

Samer Saab Jr, Shashi Phoha, Minghui Zhu, and Asok Ray. An adaptive polyak heavy-ball method. *Machine Learning*, 111(9):3245–3277, 2022.

Jesús María Sanz Serna and Konstantinos C Zygalakis. The connections between Lyapunov functions for some optimization algorithms and differential equations. *SIAM Journal on Numerical Analysis*, 59(3):1542–1565, 2021.

Damien Scieur, Vincent Roulet, Francis Bach, and Alexandre d'Aspremont. Integration methods and optimization algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.

Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, pp. 1–70, 2022.

Jonathan W Siegel. Accelerated first-order methods: Differential equations and Lyapunov functions. *arXiv preprint arXiv:1903.05671*, 2019.

Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17 (153):1–43, 2016.

Boya Sun, Jemin George, and Solmaz Kia. High-resolution modeling of the fastest first-order optimization method for strongly convex functions. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 4237–4242. IEEE, 2020.

Tao Sun, Penghang Yin, Dongsheng Li, Chun Huang, Lei Guan, and Hao Jiang. Non-ergodic convergence analysis of heavy-ball algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5033–5040, 2019.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. PMLR, 2013.

Adrien Taylor and Yoel Drori. An optimal gradient method for smooth strongly convex minimization. *Mathematical Programming*, 199(1):557–594, 2023.

Adrien Taylor, Bryan Van Scoy, and Laurent Lessard. Lyapunov functions for first-order methods: Tight automated convergence guarantees. In *International Conference on Machine Learning*, pp. 4897–4906. PMLR, 2018.

Kiran K Thekumparampil, Niao He, and Sewoong Oh. Lifted primal-dual method for bilinearly coupled smooth minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 4281–4308. PMLR, 2022.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Kansei Ushiyama, Shun Sato, and Takayasu Matsuo. A unified discretization framework for differential equation approach with Lyapunov arguments for convex optimization. *Advances in Neural Information Processing Systems*, 36, 2024.

Bryan Van Scoy, Randy A Freeman, and Kevin M Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1): 49–54, 2017.

Hongbin Wang and Paul C Miller. Scaled heavy-ball acceleration of the Richardson-Lucy algorithm for 3D microscopy image restoration. *IEEE Transactions on Image Processing*, 23(2):848–854, 2013.

Jun-Kun Wang, Chi-Heng Lin, and Jacob D Abernethy. A modular analysis of provable acceleration via Polyak's momentum: Training a wide ReLU network and a deep linear network. In *International Conference on Machine Learning*, pp. 10816–10827. PMLR, 2021.

Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

Ashia C Wilson, Ben Recht, and Michael I Jordan. A Lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021.

Hedi Xia, Vai Suliafu, Hangjie Ji, Tan Nguyen, Andrea Bertozzi, Stanley Osher, and Bao Wang. Heavy ball neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 34:18646–18659, 2021.

Penghang Yin, Yifei Lou, Qi He, and Jack Xin. Minimization of 1-2 for compressed sensing. *SIAM Journal on Scientific Computing*, 37(1):A536–A563, 2015.

Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, 194(1):901–935, 2022.

Peiyuan Zhang, Antonio Orvieto, Hadi Daneshmand, Thomas Hofmann, and Roy S Smith. Revisiting the role of euler numerical integration on acceleration and stability in convex optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3979–3987. PMLR, 2021.

Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *Journal of Machine Learning Research*, 18(84):1–42, 2017.

Dao Li Zhu and Patrice Marcotte. Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. *SIAM Journal on Optimization*, 6(3):714–726, 1996.

A DEFINITIONS AND STANDARD RESULTS

We review some foundational results from the convex analysis.

Recall that the Bregman divergence of f is defined as

$$D_f(y, x) := f(y) - f(x) - \langle \nabla f(x), y - x \rangle,$$

which is in general non-symmetric, i.e., $D_f(y, x) \neq D_f(x, y)$. A symmetrized Bregman divergence is defined as

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle = D_f(y, x) + D_f(x, y).$$

We have the following bounds on the Bregman divergence and the symmetrized Bregman divergence.

LEMMA A.1 (Section 2.1 in Nesterov (2018)). *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth. For any $x, y \in \mathbb{R}^d$,*

$$\frac{\mu}{2}\|x - y\|^2 \le D_f(y, x) \le \frac{L}{2}\|x - y\|^2, \tag{30a}$$

$$\frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \le D_f(y, x) \le \frac{1}{2\mu}\|\nabla f(x) - \nabla f(y)\|^2, \tag{30b}$$

$$\mu\|x - y\|^2 \le \langle \nabla f(x) - \nabla f(y), x - y \rangle \le L\|x - y\|^2, \tag{30c}$$

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \le \langle \nabla f(x) - \nabla f(y), x - y \rangle \le \frac{1}{\mu}\|\nabla f(x) - \nabla f(y)\|^2. \tag{30d}$$

The following three-point identity on the Bregman divergence will be used to replace the identity of squares.

LEMMA A.2 (Bregman divergence identity (Chen & Teboulle, 1993)). *If function $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, then for any $x, y, z \in \mathbb{R}^d$, it holds that*

$$\langle \nabla f(y) - \nabla f(x), y - z \rangle = D_f(z, y) + D_f(y, x) - D_f(z, x). \tag{31}$$

Proof. By definition,

$$\begin{aligned} D_f(z, y) &= f(z) - f(y) - \langle \nabla f(y), z - y \rangle, \\ D_f(y, x) &= f(y) - f(x) - \langle \nabla f(x), y - x \rangle, \\ D_f(z, x) &= f(z) - f(x) - \langle \nabla f(x), z - x \rangle. \end{aligned}$$

Direct calculation gives the identity. □

B RELATED WORKS

B.1 EXTENSIONS/APPLICATIONS OF HEAVY-BALL METHODS

While the acceleration guarantee has not yet proved rigorously, applications of heavy-ball methods including extension to constrained and distributed optimization problems have confirmed its performance benefits over the standard gradient-based methods (Wang & Miller, 2013; Ochs et al., 2015; Ghadimi et al., 2013; Diakonikolas & Jordan, 2021).

Sutskever et al. (2013) showed that stochastic gradient descent with momentum improves the training of deep and recurrent neural networks. Also, using heavy-ball flow improves neural ODEs training and inference (Xia et al., 2021). Recently, accelerated convergence of stochastic heavy-ball methods has been established in Pan et al. (2023); Bollapragada et al. (2022); Wang et al. (2021) but only for quadratic objectives. To get rid of the hyperparameters used in the heavy-ball methods, Saab Jr et al. (2022) proposed an adaptive heavy-ball that estimates the Polyak's optimal hyper-parameters at each iteration.

B.2 FIRST-ORDER METHODS AND DYNAMIC SYSTEM

One approach to better understand the mechanism of the iterative method is the continuous-time analysis: derive an ordinary differential equation (ODE) model which coincides with the iterative method taking step size close to zero. Starting from the accelerated first-order methods for unconstrained optimization, an important milestone in this direction is to understand the acceleration from the variational perspective (Su et al., 2016; Wibisono et al., 2016). While the iterative methods are first-order, the continuous-time dynamics proposed for accelerated methods are high-order or high-resolution ODEs (Shi et al., 2022; Sun et al., 2020; Muehlebach & Jordan, 2019; Attouch et al., 2000). With the continuous-time dynamic, novel accelerated methods are proposed by time discretization of the ODE model and usually the behaviour of the dynamic facilitates the convergence analysis of the iterative methods (Krichene et al., 2015; Luo & Chen, 2022; Aujol et al., 2022; Wilson et al., 2021; Siegel, 2019).

Due to the appealing results, systematic framework to draw connection between the dynamic system and the accelerated iterative method gains lots of interest (Scieur et al., 2017; Ushiyama et al., 2024; Taylor et al., 2018; Sanz Serna & Zygalakis, 2021). In fact, the theory and methods developed for other problem classes often build upon the work done in unconstrained optimization (Diakonikolas & Orecchia, 2019).

B.3 APPLICATIONS OF STRONGLY-CONVEX-STRONGLY CONCAVE SADDLE POINT PROBLEMS WITH BILINEAR COUPLING

A classical application is the regularized empirical risk minimization (ERM) with linear predictors, which is a classical supervised learning problem. Given a data matrix $B = [b_1, b_2, \cdots, b_n]^\top \in \mathbb{R}^{n\times m}$ where $b_i \in \mathbb{R}^m$ is the feature vector of the i-th data entry, the ERM problem aims to solve

$$\min_{u\in\mathbb{R}^m} g(Bu) + f(u), \tag{32}$$

where $g : \mathbb{R}^n \to \mathbb{R}$ is some strongly convex loss function, $f : \mathbb{R}^m \to \mathbb{R}$ is a strongly convex regularizer and $u \in \mathbb{R}^m$ is the linear predictor. Equivalently, we can solve the saddle point problem

$$\min_{u\in\mathbb{R}^m} \max_{p\in\mathbb{R}^n} \left\{ p^\top Bu - g^*(p) + f(u) \right\}. \tag{33}$$

The saddle point formulation is favorable in many scenarios, for instance (Zhang & Xiao, 2017; Du & Hu, 2019; Lei et al., 2017).

Another application is policy evaluation problems in reinforcement learning when finding minimum the mean squared projected Bellman error (Du et al., 2017):

$$\arg\min_u \frac{1}{2}\|Bu - b\|^2_{C^{-1}} + \frac{1}{2}\|u\|^2,$$

where B, C are given matrices. The corresponding saddle point problem is

$$\min_{u\in\mathbb{R}^m} \max_{p\in\mathbb{R}^n} \frac{1}{2}\|u\|^2 - \frac{1}{2}\|p\|^2_C - \langle b, p\rangle + \langle Bu, p\rangle.$$

The saddle point formulation saves the computation cost of inverting C.

C PROOF OF THEOREM 1.1

Consider the Lyapunov function

$$\mathcal{E}(\mathbf{z}) = \mathcal{E}(x, y) : = f(x) - f(x^*) + \frac{\mu}{2}\|y - x^*\|^2 = D_f(x, x^*) + \frac{\mu}{2}\|y - x^*\|^2, \tag{34}$$

and the modified Lyapunov function

$$\begin{aligned}\mathcal{E}^\alpha(\mathbf{z}) = \mathcal{E}^\alpha(x, y) &: = \mathcal{E}(x, y) + \alpha\langle \nabla f(x) - \nabla f(x^*), y - x^*\rangle \\ &= D_f(x, x^*) + \frac{\mu}{2}\|y - x^*\|^2 + \alpha\langle \nabla f(x) - \nabla f(x^*), y - x^*\rangle.\end{aligned} \tag{35}$$

The novelty of $\mathcal{E}^\alpha(x, y)$ is the inclusion of the cross term $\alpha\langle \nabla f(x) - \nabla f(x^*), y - x^*\rangle$.

We denote the right-hand side of the accelerated gradient descent flow (6) as

$$\mathcal{G}(\mathbf{z}) = \mathcal{G}(x, y) := \begin{pmatrix} y - x, \\ x - y - \frac{1}{\mu}\nabla f(x) \end{pmatrix}.$$

LEMMA C.1 (Strong Lyapunov property (Luo & Chen, 2022))**.**

$$-\langle \nabla \mathcal{E}(x, y), \mathcal{G}(x, y)\rangle \geq \mathcal{E}(x, y) + \frac{\mu}{2}\|x - y\|^2, \quad \forall\, x, y \in \mathbb{R}^d. \tag{36}$$

Proof. By direct calculation and the μ-convexity of f,

$$\begin{aligned}&-\langle \nabla \mathcal{E}(x, y), \mathcal{G}(x, y)\rangle = \langle \nabla f(x), x - x^*\rangle - \mu\langle x - y, y - x^*\rangle \\ &= \langle \nabla f(x), x - x^*\rangle - \frac{\mu}{2}\left(\|x - x^*\|^2 - \|x - y\|^2 - \|y - x^*\|^2\right) \\ &\geq f(x) - f(x^*) + \frac{\mu}{2}\|y - x^*\|^2 + \frac{\mu}{2}\|x - y\|^2 = \mathcal{E}(x, y) + \frac{\mu}{2}\|x - y\|^2.\end{aligned}$$

□

LEMMA C.2. *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth. Denote by $\rho = \mu/L$. For any two vectors $\mathbf{z}_k, \mathbf{z}_{k+1} \in \mathbb{R}^{2d}$ and $\alpha > 0$, we have the following inequality on the Bregman divergence of $\mathcal{E}$ defined by (34)*

$$\begin{aligned}\left(1 - \frac{\alpha}{\sqrt{\rho}}\right) D_{\mathcal{E}}(\mathbf{z}_k, \mathbf{z}_{k+1}) &\leq D_{\mathcal{E}}(\mathbf{z}_k, \mathbf{z}_{k+1}) \pm \alpha\langle y_{k+1} - y_k, \nabla f(x_{k+1}) - \nabla f(x_k)\rangle \\ &\leq \left(1 + \frac{\alpha}{\sqrt{\rho}}\right) D_{\mathcal{E}}(\mathbf{z}_k, \mathbf{z}_{k+1}).\end{aligned}$$

Proof. Using the identity of squares (for y) and Bregman divergence identity (31) (for x), we have the component form of

$$D_{\mathcal{E}}(\mathbf{z}_k, \mathbf{z}_{k+1}) = D_f(x_k, x_{k+1}) + \frac{\mu}{2}\|y_k - y_{k+1}\|^2.$$

By Cauchy-Schwarz inequality and bound (30b) in Lemma A.1, we have

$$\begin{aligned}\alpha\left|\langle y_{k+1} - y_k, \nabla f(x_{k+1}) - \nabla f(x_k)\rangle\right| &\leq \frac{\alpha}{2\sqrt{\mu L}}\|\nabla f(x_k) - \nabla f(x_{k+1})\|^2 + \frac{\alpha\sqrt{L\mu}}{2}\|y_k - y_{k+1}\|^2 \\ &\leq \alpha\sqrt{\frac{L}{\mu}} D_f(x_k, x_{k+1}) + \frac{\alpha\sqrt{L\mu}}{2}\|y_k - y_{k+1}\|^2 \\ &= \alpha\sqrt{\frac{L}{\mu}}\left(D_f(x_k, x_{k+1}) + \frac{\mu}{2}\|y_k - y_{k+1}\|^2\right).\end{aligned}$$

□

By Lemma C.2 and $\mathcal{E}(\mathbf{z}) = D_{\mathcal{E}}(\mathbf{z}, \mathbf{z}^*)$ where $\mathbf{z}^* := (x^*, x^*)^\top$, it is straightforward to verify the following bounds for $\mathcal{E},\ \mathcal{E}^\alpha$ as Lyapunov functions. Assume $0 < \alpha \le \sqrt{\rho}$ with $\rho = \frac{\mu}{L}$. Then

$$0 \le \mathcal{E}^\alpha(\mathbf{z}) \le 2\mathcal{E}(\mathbf{z}), \quad \forall\, \mathbf{z} = (x, y)^\top \in \mathbb{R}^{2d}. \tag{37}$$

Now we are in the position to prove our first main result.

Proof of Theorem 1.1. We write (14) as a correction of Implicit Euler discretization of (6):

$$\mathbf{z}_{k+1} - \mathbf{z}_k = \alpha \mathcal{G}(\mathbf{z}_{k+1}) - \alpha \begin{pmatrix} y_{k+1} - y_k \\ \frac{1}{\mu}(\nabla f(x_{k+1}) - \nabla f(x_k)) \end{pmatrix}. \tag{38}$$

and substitude into the identity of $\mathcal{E}$

$$\begin{aligned} &\mathcal{E}(\mathbf{z}_{k+1}) - \mathcal{E}(\mathbf{z}_k) = (\nabla \mathcal{E}(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{z}_k) - D_{\mathcal{E}}(\mathbf{z}_k, \mathbf{z}_{k+1}) \\ &= \alpha \langle \nabla \mathcal{E}(\mathbf{z}_{k+1}), \mathcal{G}(\mathbf{z}_{k+1}) \rangle - \alpha \langle \nabla \mathcal{E}(\mathbf{z}_{k+1}), \begin{pmatrix} y_{k+1} - y_k \\ \frac{1}{\mu}(\nabla f(x_{k+1}) - \nabla f(x_k)) \end{pmatrix} \rangle - D_{\mathcal{E}}(\mathbf{z}_k, \mathbf{z}_{k+1}). \end{aligned}$$

We write out the component form of the middle term and split it as

$$\begin{aligned} &\alpha \langle \nabla f(x_{k+1}) - \nabla f(x^*), y_{k+1} - y_k \rangle + \alpha \langle y_{k+1} - x^*, \nabla f(x_{k+1}) - \nabla f(x_k) \rangle \\ =\ &\alpha \langle \nabla f(x_{k+1}) - \nabla f(x^*), y_{k+1} - x^* \rangle + \alpha \langle y_{k+1} - y_k, \nabla f(x_{k+1}) - \nabla f(x_k) \rangle \\ &- \alpha \langle y_k - x^*, \nabla f(x_k) - \nabla f(x^*) \rangle. \end{aligned}$$

Substitute back to (18) and rearrange terms, we obtain the following identity:

$$\begin{aligned} \mathcal{E}^\alpha(\mathbf{z}_{k+1}) - \mathcal{E}^\alpha(\mathbf{z}_k) = &\ \alpha \langle \nabla \mathcal{E}(\mathbf{z}_{k+1}), \mathcal{G}(\mathbf{z}_{k+1}) \rangle \\ &- D_{\mathcal{E}}(\mathbf{z}_k, \mathbf{z}_{k+1}) - \alpha \langle y_{k+1} - y_k, \nabla f(x_{k+1}) - \nabla f(x_k) \rangle, \end{aligned} \tag{39}$$

By Lemma C.2, we can drop the terms in the second line for $0 \le \alpha \le \sqrt{\rho}$. Use the strong Lyapunov property (36 and bound (37, we have

$$\mathcal{E}^\alpha(\mathbf{z}_{k+1}) - \mathcal{E}^\alpha(\mathbf{z}_k) \le -\alpha \mathcal{E}(\mathbf{z}_{k+1}) \le -\frac{\alpha}{2} \mathcal{E}^\alpha(\mathbf{z}_{k+1}), \tag{40}$$

which implies the global linear convergence of $\mathcal{E}^\alpha$:

$$\mathcal{E}^\alpha(\mathbf{z}_{k+1}) \le \frac{1}{1 + \alpha/2} \mathcal{E}^\alpha(\mathbf{z}_k) \le \left(\frac{1}{1 + \alpha/2} \right)^{k+1} \mathcal{E}^\alpha(\mathbf{z}_0), \quad k \ge 0.$$

Moreover, (40 implies

$$\mathcal{E}(\mathbf{z}_{k+1}) \le \frac{1}{\alpha}(\mathcal{E}^\alpha(\mathbf{z}_k) - \mathcal{E}^\alpha(\mathbf{z}_{k+1})) \le \frac{1}{\alpha} \mathcal{E}^\alpha(\mathbf{z}_k) \le \frac{1}{\alpha} \left(\frac{1}{1 + \alpha/2} \right)^k \mathcal{E}^\alpha(\mathbf{z}_0),$$

which coincides with (8) with $C_0 = \frac{1}{\alpha} \mathcal{E}^\alpha(\mathbf{z}_0) \ge 0$. □

D STRONGLY LYAPUNOV PROPERTY FOR AOR-HB-COMPOSITE FLOW

Consider the Lyapunov function

$$\mathcal{E}(x, y) := D_f(x, x^*) + \frac{\mu}{2} \|y - x^*\|^2,$$

where the Bregman divergence $D_f(x, x^*)$ is generalization of $f(x) - f(x^*)$ in the single convex function case.

We rewrite the flow (23 as

$$\begin{aligned} x' &= y - x, \\ y' &= x - y - \frac{1}{\mu}(\nabla f(x) + \xi), \quad \xi \in \partial g(y). \end{aligned}$$

With this formulation, we denote the right-hand side as

$$\mathcal{G}(\mathbf{z}) = \mathcal{G}(x, y) := \begin{pmatrix} y - x, \\ x - y - \frac{1}{\mu}\nabla(f(x) + \xi) \end{pmatrix}.$$

Using the first order optimality condition, $\mathcal{G}(\mathbf{z}^*) = 0$ for some $\xi^* \in \partial g(x^*)$.

We restate and prove the strong Lyapunov property in the following Lemma. The difference compared with the proof of Lemma C.1 is highlighted in blue.

LEMMA D.1 (Strong Lyapunov property for AOR-HB-composite flow (23)**.**

$$-\langle \nabla \mathcal{E}(x, y), \mathcal{G}(x, y)\rangle \geq \mathcal{E}(x, y) + \frac{\mu}{2}\|x - y\|^2, \quad \forall\, x, y \in \mathbb{R}^d.$$

Proof. By direct calculation and the μ-convexity of f,

$$\begin{aligned}
&- \langle \nabla \mathcal{E}(x, y), \mathcal{G}(x, y)\rangle = -\langle \nabla \mathcal{E}(x, y), \mathcal{G}(x, y) - \mathcal{G}(x^*, x^*)\rangle \\
&= \langle \nabla f(x) - \nabla f(x^*), x - x^*\rangle - \mu\langle x - y, y - x^*\rangle + \langle \xi - \xi^*, y - x^*\rangle \\
&= \langle \nabla f(x) - \nabla f(x^*), x - x^*\rangle - \frac{\mu}{2}\left(\|x - x^*\|^2 - \|x - y\|^2 - \|y - x^*\|^2\right) + \langle \xi - \xi^*, y - x^*\rangle \\
&\geq D_f(x, x^*) + \frac{\mu}{2}\|y - x^*\|^2 + \frac{\mu}{2}\|x - y\|^2 = \mathcal{E}(x, y) + \frac{\mu}{2}\|x - y\|^2,
\end{aligned}$$

where the inequality used the convexity of g:

$$\langle \xi - \eta, y - x\rangle \geq 0, \quad \forall\, x, y \in \mathbb{R}^d \text{ and } \xi \in \partial g(y), \eta \in \partial g(x).$$

□

THEOREM D.1 (Convergence rate of AOR-HB method for composite convex minimization)**.** *Suppose f is μ-strongly convex and L-smooth and g is non-smooth and convex. Let (x_k, y_k) be generated by Algorithm 2 with initial value (x_0, y_0) and step size $\alpha = \sqrt{\mu/L}$. Then there exists a non-negative constant $C_0 = C_0(x_0, y_0, \mu, L)$ so that we have the accelerated linear convergence*

$$D_f(x, x^*) + \frac{\mu}{2}\|y_{k+1} - x^*\|^2 \leq C_0\left(\frac{1}{1 + \frac{1}{2}\sqrt{\mu/L}}\right)^k, \quad k \geq 1. \tag{41}$$

Proof. See Appendix C. □

E EQUIVALENT FORMULATION OF AOR-HB-SADDLE

Algorithm 2 is equivalent to the following discretization of the HB-saddle flow (48:

$$\frac{u_{k+1} - u_k}{\alpha} = v_k - u_{k+1}, \tag{42a}$$

$$\frac{p_{k+1} - p_k}{\alpha} = q_k - p_{k+1}, \tag{42b}$$

$$\frac{v_{k+1} - v_k}{\alpha} = u_{k+1} - v_{k+1} - \frac{1}{\mu_f}\left(2\nabla f(u_{k+1}) - \nabla f(u_k) + B^\top q_k\right), \tag{42c}$$

$$\frac{q_{k+1} - q_k}{\alpha} = p_{k+1} - q_{k+1} - \frac{1}{\mu_g}\left(2\nabla g(p_{k+1}) - \nabla g(p_k) - B(2v_{k+1} - v_k)\right). \tag{42d}$$

The discretization is a mixture of implicit Euler and explicit Euler with time step size α. For both the gradient terms and Bv term, we use the AOR technique, i.e., $\nabla f(u) \approx 2\nabla f(u_{k+1}) - \nabla f(u_k)$, $\nabla g(p) \approx 2\nabla g(p_{k+1}) - \nabla g(p_k)$ and $Bv \approx B(2v_{k+1} - v_k)$. Algorithm 3 is implementation-friendly while (42 is convenient for deriving convergence analysis.

F PROOFS OF SECTION 3

F.1 A CLASS OF MONOTONE OPERATOR EQUATIONS

In fact, we can extend HB flow to a broad class of monotone operator equation $\mathcal{A}(x) = 0$ with

$$\mathcal{A}(x) = \nabla F(x) + \mathcal{N}x, \tag{43}$$

where F is a strongly convex function and L_F smooth, and $\mathcal{N}$ is a linear and skew-symmetric operator, i.e., $\mathcal{N}^\top = -\mathcal{N}$. Then $\mathcal{A}$ is Lipschitz continuous with constant $L_{\mathcal{A}} \leq L_F + \|\mathcal{N}\|$. Therefore $\mathcal{A}$ is monotone and Lipschitz continuous which is also known as inverse-strongly monotonicity (Browder & Petryshyn, 1967; Liu & Nashed, 1998) or co-coercitivity (Zhu & Marcotte, 1996). Consequently equation $\mathcal{A}(x) = 0$ has a unique solution x^* (Rockafellar, 1976).

As a special case, we recover strongly-convex-strongly-concave saddle point problems with bilinear coupling when $x := (u, p)^\top$, $F(x) := f(u) + g(p)$ and $\mathcal{N} := \begin{pmatrix} 0 & B^\top \\ -B & 0 \end{pmatrix}$.

We introduce an accelerated gradient flow

$$\begin{cases} x' = y - x, \\ y' = x - y - \mu^{-1}(\nabla F(x) + \mathcal{N}y). \end{cases} \tag{44}$$

Comparing with the accelerated gradient flow (6) for convex optimization, the difference is the gradient and skew-symmetric splitting $\mathcal{A}(x) \to \nabla F(x) + \mathcal{N}y$.

Denote the vector field on the right hand side of (44 by $\mathcal{G}(x, y)$. Then $\mathcal{G}(x^*, x^*) = 0$ and thus (x^*, x^*) is an equilibrium point of (44.

We first show (x^*, x^*) is exponentially stable. Consider the Lyapunov function:

$$\mathcal{E}(x, y) = D_F(x, x^*) + \frac{\mu}{2}\|y - x^*\|^2. \tag{45}$$

For μ-strongly convex F, function $D_F(\cdot, x^*) \in \mathcal{S}_\mu$. Then $\mathcal{E}(x, y) \geq 0$ and $\mathcal{E}(x, y) = 0$ iff $x = y = x^*$.

We then verify the strong Lyapunov property. The proof is similar to that of Lemma C.1.

THEOREM F.1 (Strong Lyapunov Property). *Assume function F is μ-strongly convex. Then for the Lyapunov function (45 and the accelerated gradient flow vector field $\mathcal{G}$ asoociated to (44, the following strong Lyapunov property holds*

$$-\nabla\mathcal{E}(x, y) \cdot \mathcal{G}(x, y) \geq \mathcal{E}(x, y) + \frac{\mu}{2}\|y - x\|^2, \quad \forall\, x, y \in \mathcal{V}. \tag{46}$$

Proof. First of all, as $\mathcal{G}(x^*, x^*) = 0$,

$$-\langle \nabla\mathcal{E}(x, y), \mathcal{G}(x, y)\rangle = -\langle \nabla\mathcal{E}(x, y), \mathcal{G}(x, y) - \mathcal{G}(x^*, x^*)\rangle.$$

Direct computation gives

$$\begin{aligned} -\langle \nabla\mathcal{E}(x, y), \mathcal{G}(x, y)\rangle &= \langle \nabla D_F(x, x^*), x - x^* - (y - x^*)\rangle - \mu\langle y - x^*, x - x^*\rangle \\ &\quad + \mu\|y - x^*\|^2 + \langle \nabla F(x) - \nabla F(x^*), y - x^*\rangle + \langle y - x^*, \mathcal{N}(y - x^*)\rangle \\ &= \langle \nabla F(x) - \nabla F(x^*), x - x^*\rangle + \mu\|y - x^*\|^2 - \mu\langle y - x^*, x - x^*\rangle, \end{aligned}$$

where we have used $\nabla D_F(x, x^*) = \nabla F(x) - \nabla F(x^*)$ and $\langle y - x^*, \mathcal{N}(y - x^*)\rangle = 0$ since $\mathcal{N}$ is skew-symmetric. We expand the last two term using the identity of squares:

$$\frac{1}{2}\|y - x^*\|^2 - \langle y - x^*, x - x^*\rangle = \frac{1}{2}\|y - x\|^2 - \frac{1}{2}\|x - x^*\|^2.$$

Using the bound (30a, we get

$$\langle \nabla F(x) - \nabla F(x^*), x - x^*\rangle = D_F(x, x^*) + D_F(x^*, x) \geq D_F(x, x^*) + \frac{\mu}{2}\|x - x^*\|^2$$

and obtain (46. □

The calculation is more clear when $\nabla F(x) = Ax$ is linear with $A = \nabla^2 F \geq \mu I$. We denote by $\mathbf{z} = (x, y)^\top$ and $\mathcal{E}(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}-\mathbf{z}^*\|^2_{\mathcal{D}}$ with $\mathcal{D} = \begin{pmatrix} A & 0 \\ 0 & \mu I \end{pmatrix}$. Then $-\langle \nabla \mathcal{E}(x, y), \mathcal{G}(x, y) - \mathcal{G}(x^*, x^*)\rangle$ is a quadratic form $(\mathbf{z} - \mathbf{z}^*)^\top \mathcal{M}(\mathbf{z} - \mathbf{z}^*)$. We calculate the matrix $\mathcal{M}$ as

$$\begin{pmatrix} A & 0 \\ 0 & \mu I \end{pmatrix} \begin{pmatrix} I & -I \\ -I + \mu^{-1} A & I + \mu^{-1}\mathcal{N} \end{pmatrix} = \begin{pmatrix} A & -A \\ -\mu I + A & \mu I + \mathcal{N} \end{pmatrix}.$$

As $v^\top M v = v^\top \operatorname{sym}(M) v$ where $\operatorname{sym}(\cdot)$ is the symmetric part of a matrix, direct computation gives

$$\begin{aligned} \operatorname{sym}\begin{pmatrix} A & -A \\ -\mu I + A & \mu I + \mathcal{N} \end{pmatrix} &= \begin{pmatrix} A & -\mu I/2 \\ -\mu I/2 & \mu I \end{pmatrix} \\ &\geq \begin{pmatrix} A/2 & 0 \\ 0 & \mu I/2 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \mu I & -\mu I \\ -\mu I & \mu I \end{pmatrix}, \end{aligned}$$

where in the last step we use the convexity $A \geq \mu I$. Then (46 follows.

F.2 EXPONENTIAL STABILITY OF HB-SADDLE FLOW

Return to the saddle point problems, we let $\mathbf{z} = (x, y)^\top$ with $x = (u, p)^\top, y = (v, q)^\top \in \mathbb{R}^m \times \mathbb{R}^n$ and denote by

$$F(x) := f(u) + g(p),\ \mathcal{N} := \begin{pmatrix} 0 & B^\top \\ -B & 0 \end{pmatrix},\ \mathcal{D}_\mu := \begin{pmatrix} \mu_f & 0 \\ 0 & \mu_g \end{pmatrix}.$$

Consider the Lyapunov function

$$\begin{aligned} \mathcal{E}(\mathbf{z}) = \mathcal{E}(u, v, p, q) &:= D_F(x, x^*) + \frac{1}{2}\|y - x^*\|^2_{\mathcal{D}_\mu} \\ &= D_f(u, u^*) + D_g(p, p^*) + \frac{\mu_f}{2}\|v - u^*\|^2 + \frac{\mu_g}{2}\|q - p^*\|^2, \end{aligned} \tag{47}$$

As f, g are strongly convex, $\mathcal{E}(\mathbf{z}) \geq 0$ and $\mathcal{E}(\mathbf{z}) = 0$ iff $u = v = u^*$ and $p = q = p^*$. Recall that the HB-saddle flow is

$$\begin{aligned} u' &= v - u, \\ p' &= q - p, \\ v' &= u - v - \frac{1}{\mu_f}(\nabla f(u) + B^\top q), \\ q' &= p - q - \frac{1}{\mu_g}(\nabla g(p) - Bv). \end{aligned} \tag{48}$$

Denoted the vector field on the right hand side of (48 by $\mathcal{G}(u, v, p, q)$, as a special case of Theorem F.1, we obtain the following strong Lyapunov property.

THEOREM F.2 (Strong Lyapunov property for HB-saddle flow)**.** *Suppose f is μ_f-strongly convex g is μ_g-strongly convex. The following strong Lyapunov property holds for all $u, v \in \mathbb{R}^m, p, q \in \mathbb{R}^n$.*

$$-\langle \nabla \mathcal{E}(u, v, p, q), \mathcal{G}(u, v, p, q)\rangle \geq \mathcal{E}(u, v, p, q) + \frac{\mu_f}{2}\|v - u\|^2 + \frac{\mu_g}{2}\|q - p\|^2. \tag{49}$$

Consequently, for solutions $(u(t), v(t), p(t), q(t))$ of the HB-saddle flow (25), we have the exponential stability:

$$D_f(u, u^*) + D_g(p, p^*) + \frac{\mu_f}{2}\|v - u^*\|^2 + \frac{\mu_g}{2}\|q - p^*\|^2 \leq e^{-t}\mathcal{E}(u(0), v(0), p(0), q(0)), \quad t > 0.$$

F.3 PROOF OF THEOREM 1.2

Consider the modified Lyapunov function

$$\mathcal{E}^\alpha(\mathbf{z}) := \mathcal{E}(\mathbf{z}) + \alpha\langle \nabla F(x) - \nabla F(x^*), y - x^*\rangle - \alpha\|y - y^*\|^2_{\mathcal{B}^{\text{sym}}}, \tag{50}$$

where $\mathcal{E}(\mathbf{z})$ is defined as (47 and $\mathcal{B}^{\rm sym} = \begin{pmatrix} 0 & B^\top \\ B & 0 \end{pmatrix}$. Due to the bilinear coupling, additional cross term $\alpha\|y - y^*\|^2_{\mathcal{B}^{\rm sym}} = 2\alpha(B(v - v^*), q - q^*)$ is included in $\mathcal{E}^\alpha$.

We shall split $\frac{1}{2}\|y - x^*\|^2_{\mathcal{D}_\mu}$ to $\frac{\beta}{2}\|y - x^*\|^2_{\mathcal{D}_\mu} + \frac{1-\beta}{2}\|y - x^*\|^2_{\mathcal{D}_\mu}$ to bound the two cross terms. The cross term $\alpha\langle \nabla F(x) - \nabla F(x^*), y - x^*\rangle$ can be bounded using the vector form of Lemma C.2. Next we focus on the $\alpha\|y - y^*\|^2_{\mathcal{B}^{\rm sym}}$.

LEMMA F.1. *Denoted by* $\mathcal{D}_\mu = \begin{pmatrix} \mu_f I & 0 \\ 0 & \mu_g I \end{pmatrix}$ *with* $\mu_f, \mu_g > 0$ *and* $\mathcal{B}^{\rm sym} = \begin{pmatrix} 0 & B^\top \\ B & 0 \end{pmatrix}$*. For any* $\alpha, c > 0$*, we have*

$$\left(1 - \frac{\alpha\|B\|}{c\sqrt{\mu_f\mu_g}}\right) c\mathcal{D}_\mu \le c\mathcal{D}_\mu \pm \alpha\mathcal{B}^{\rm sym} \le \left(1 + \frac{\alpha\|B\|}{c\sqrt{\mu_f\mu_g}}\right) c\mathcal{D}_\mu. \tag{51}$$

Proof. We first calculate the eigenvalues of $\mathcal{D}_\mu^{-1}\mathcal{B}^{\rm sym}$. By choosing the SVD basis of $B = U\Sigma V$, its eigenvalue is given by the 2×2 matrix $\begin{pmatrix} 0 & \mu_f^{-1}\sigma(B) \\ \mu_g^{-1}\sigma(B) & 0 \end{pmatrix}$, where $\sigma(B)$ is a singular value of B. So $\lambda(\mathcal{D}_\mu^{-1}\mathcal{B}^{\rm sym}) = \pm\frac{\sigma(B)}{\sqrt{\mu_f\mu_g}}$ and consequently

$$-\frac{\|B\|}{\sqrt{\mu_f\mu_g}} \le \lambda_{\min}(\mathcal{D}_\mu^{-1}\mathcal{B}^{\rm sym}) \le \lambda_{\max}(\mathcal{D}_\mu^{-1}\mathcal{B}^{\rm sym}) \le \frac{\|B\|}{\sqrt{\mu_f\mu_g}}. \tag{52}$$

Then write $c\mathcal{D}_\mu \pm \alpha\mathcal{B}^{\rm sym} = c\mathcal{D}_\mu(I \pm \frac{\alpha}{c}\mathcal{D}_\mu^{-1}\mathcal{B}^{\rm sym})$ and apply the bound (52 to get the desired result. ☐

LEMMA F.2. *Suppose* $f : \mathbb{R}^m \to \mathbb{R}$ *is* μ_f*-strongly convex and* L_f*-smooth,* $g : \mathbb{R}^n \to \mathbb{R}$ *is* μ_g*-strongly convex and* L_g*-smooth. Let* $F(\mathbf{z}) = f(u) + g(p)$*. Denoted by*

$$\sqrt{\rho} = (\sqrt{2} - 1)\min\left\{\sqrt{\frac{\mu_f}{L_f}}, \sqrt{\frac{\mu_g}{L_g}}, \frac{\sqrt{\mu_f\mu_g}}{\|B\|}\right\}.$$

Then for any $\alpha > 0$ *and* $\mathcal{E}$ *defined by (47 and any two vectors* $\mathbf{z}_k, \mathbf{z}_{k+1} \in \mathbb{R}^{2(m+n)}$

$$\begin{aligned}
&\left(1 - \frac{\alpha}{\sqrt{\rho}}\right) D_{\mathcal{E}}(\mathbf{z}_k, \mathbf{z}_{k+1}) \\
\le\ & D_{\mathcal{E}}(\mathbf{z}_k, \mathbf{z}_{k+1}) \pm \alpha\langle y_{k+1} - y_k, \nabla F(x_{k+1}) - \nabla F(x_k)\rangle \pm \alpha\|y_k - y_{k+1}\|^2_{\mathcal{B}^{\rm sym}}, \\
\le\ & \left(1 + \frac{\alpha}{\sqrt{\rho}}\right) D_{\mathcal{E}}(\mathbf{z}_k, \mathbf{z}_{k+1}).
\end{aligned} \tag{53}$$

In particular, for $\alpha \le \sqrt{\rho}$*, we have the bound*

$$0 \le \mathcal{E}^\alpha(\mathbf{z}) \le 2\mathcal{E}(\mathbf{z}) \quad \mathbf{z} \in \mathbb{R}^{2d}. \tag{54}$$

Proof. For any $\beta \in (0, 1)$, using the vector form of Lemma C.2, we have bound

$$\begin{aligned}
&\left(1 - \frac{\alpha}{\sqrt{\rho_F}}\right)\left(D_F(x_k, x_{k+1}) + \frac{\beta}{2}\|y_k - y_{k+1}\|^2_{\mathcal{D}_\mu}\right) \\
\le\ & D_F(x_k, x_{k+1}) + \frac{\beta}{2}\|y_k - y_{k+1}\|^2_{\mathcal{D}_\mu} \pm \alpha\langle y_{k+1} - y_k, \nabla F(x_{k+1}) - \nabla F(x_k)\rangle \\
\le\ & \left(1 + \frac{\alpha}{\sqrt{\rho_F}}\right)\left(D_F(x_k, x_{k+1}) + \frac{\beta}{2}\|y_k - y_{k+1}\|^2_{\mathcal{D}_\mu}\right)
\end{aligned} \tag{55}$$

with $\rho_F = \beta\min\left\{\frac{\mu_f}{L_f}, \frac{\mu_g}{L_g}\right\}$.

Use Lemma F.1, we have the bound

$$\begin{aligned}
\left(1 - \frac{2\alpha\|B\|}{(1-\beta)\sqrt{\mu_f\mu_g}}\right)\frac{1-\beta}{2}\|y_k - y_{k+1}\|^2_{\mathcal{D}_\mu} &\le \frac{1-\beta}{2}\|y_k - y_{k+1}\|^2_{\mathcal{D}_\mu} \pm \alpha\|y_k - y_{k+1}\|^2_{\mathcal{B}^{\rm sym}} \\
&\le \left(1 + \frac{2\alpha\|B\|}{(1-\beta)\sqrt{\mu_f\mu_g}}\right)\frac{1-\beta}{2}\|y_k - y_{k+1}\|^2_{\mathcal{D}_\mu}.
\end{aligned} \tag{56}$$

Choose $\beta \in (0,1)$ s.t. $\sqrt{\beta} = \frac{1-\beta}{2}$, i.e., $\beta = 3 - 2\sqrt{2}$ and $\sqrt{\beta} = \sqrt{2} - 1$. Adding (55 and (56 implies (53.

Apply (53 with $\mathbf{z}_k = \mathbf{z}$ and $\mathbf{z}_{k+1} = \mathbf{z}^*$ to get (54. $\square$

Now we are ready to prove the theorem.

Proof of Theorem 1.2. We write (42 as a correction of the implicit Euler method

$$\mathbf{z}_{k+1} - \mathbf{z}_k = \alpha \mathcal{G}(\mathbf{z}_{k+1}) - \alpha \begin{pmatrix} 0 & I \\ \mathcal{D}_\mu^{-1} \nabla F & 0 \end{pmatrix} (\mathbf{z}_{k+1} - \mathbf{z}_k) - \alpha \begin{pmatrix} 0 \\ \mathcal{D}_\mu^{-1} \mathcal{B}^{\mathrm{sym}} (y_{k+1} - y_k) \end{pmatrix}. \tag{57}$$

Use the definition of Bregman divergence, we have the identity for the difference of the Lyapunov function $\mathcal{E}$ at consecutive steps:

$$\mathcal{E}(\mathbf{z}_{k+1}) - \mathcal{E}(\mathbf{z}_k) = (\nabla \mathcal{E}(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{z}_k) - D_{\mathcal{E}}(\mathbf{z}_k, \mathbf{z}_{k+1})$$

Substitute (57 and expand the cross term:

$$\begin{aligned} \langle \nabla \mathcal{E}(\mathbf{z}_{k+1}), &\mathbf{z}_{k+1} - \mathbf{z}_k \rangle = \alpha \langle \nabla \mathcal{E}(\mathbf{z}_{k+1}), \mathcal{G}(\mathbf{z}_{k+1}) \rangle \\ &- \alpha \langle \nabla F(x_{k+1}) - \nabla F(x^*), y_{k+1} - y_k \rangle - \alpha \langle y_{k+1} - x^*, \nabla F(x_{k+1}) - \nabla F(x_k) \rangle \\ &- \alpha \langle y_{k+1} - y^*, y_{k+1} - y_k \rangle_{\mathcal{B}^{\mathrm{sym}}} \end{aligned} \tag{58}$$

We split the gradient term as

$$\begin{aligned} &\alpha \langle \nabla F(x_{k+1}) - \nabla F(x^*), y_{k+1} - y_k \rangle + \alpha \langle y_{k+1} - x^*, \nabla F(x_{k+1}) - \nabla F(x_k) \rangle \\ = \; &\alpha \langle \nabla F(x_{k+1}) - \nabla F(x^*), y_{k+1} - x^* \rangle + \alpha \langle y_{k+1} - y_k, \nabla F(x_{k+1}) - \nabla F(x_k) \rangle \\ &- \alpha \langle y_k - x^*, \nabla F(x_k) - \nabla F(x^*) \rangle. \end{aligned}$$

We can use identity of squares to expand

$$\langle y_{k+1} - y^*, y_{k+1} - y_k \rangle_{\mathcal{B}^{\mathrm{sym}}} = \frac{1}{2} \left(\|y_{k+1} - y^*\|^2_{\mathcal{B}^{\mathrm{sym}}} + \|y_{k+1} - y_k\|^2_{\mathcal{B}^{\mathrm{sym}}} - \|y_k - y^*\|^2_{\mathcal{B}^{\mathrm{sym}}} \right).$$

Substitute back to (18) and rearrange terms, we obtain the following identity:

$$\begin{aligned} \mathcal{E}^\alpha(\mathbf{z}_{k+1}) - \mathcal{E}^\alpha(\mathbf{z}_k) = \; &\alpha \langle \nabla \mathcal{E}(\mathbf{z}_{k+1}), \mathcal{G}(\mathbf{z}_{k+1}) \rangle \\ - D_{\mathcal{E}}(\mathbf{z}_k, \mathbf{z}_{k+1}) &+ \alpha \langle y_{k+1} - y_k, \nabla F(x_{k+1}) - \nabla F(x_k) \rangle - \alpha \|y_k - y_{k+1}\|^2_{\mathcal{B}^{\mathrm{sym}}}, \end{aligned} \tag{59}$$

which holds for arbitrary step size α.

Now take $\alpha \le \sqrt{\rho}$. The last term in (59 is non-positive. Use the strong Lyapunov property, and the bound $\mathcal{E}^\alpha(\mathbf{z}) \le 2\mathcal{E}(\mathbf{z})$ to get

$$\mathcal{E}^\alpha(\mathbf{z}_{k+1}) - \mathcal{E}^\alpha(\mathbf{z}_k) \le -\alpha \mathcal{E}(\mathbf{z}_{k+1}) \le -\frac{\alpha}{2} \mathcal{E}^\alpha(\mathbf{z}_{k+1}), \tag{60}$$

which implies the global linear convergence:

$$\mathcal{E}^\alpha(\mathbf{z}_{k+1}) \le \frac{1}{1 + \alpha/2} \mathcal{E}^\alpha(\mathbf{z}_k) \le \left(\frac{1}{1 + \alpha/2} \right)^{k+1} \mathcal{E}^\alpha(\mathbf{z}_0), \quad k \ge 0.$$

Moreover, (60 implies

$$\mathcal{E}(\mathbf{z}_{k+1}) \le \frac{1}{\alpha} (\mathcal{E}^\alpha(\mathbf{z}_k) - \mathcal{E}^\alpha(\mathbf{z}_{k+1})) \le \frac{1}{\alpha} \mathcal{E}^\alpha(\mathbf{z}_k) \le \frac{1}{\alpha} \left(\frac{1}{1 + \alpha/2} \right)^k \mathcal{E}^\alpha(\mathbf{z}_0),$$

which coincides with (8) with $C_0 = \frac{1}{\alpha} \mathcal{E}^\alpha(\mathbf{z}_0) \ge 0$. $\square$

Algorithm 4 AOR-HB-saddle-I Algorithm

```
**Parameters:** $u_0, v_0 \in \mathbb{R}^m, p_0, q_0 \in \mathbb{R}^n, L_f, L_g, \mu_f, \mu_g, \|B\|$.
**Set** $\alpha = \min\left\{\sqrt{\frac{\mu_f}{L_f}}, \sqrt{\frac{\mu_g}{L_g}}\right\}$.
**for** $k = 0, 1, 2, \dots$ **do**
    $u_{k+1} = \frac{1}{1+\alpha}(u_k + \alpha v_k); \quad p_{k+1} = \frac{1}{1+\alpha}(p_k + \alpha q_k)$
    Find $(v_{k+1}, q_{k+1})$ such that
        $v_{k+1} = \frac{1}{1+\alpha}\left(v_k + \alpha u_{k+1} - \frac{\alpha}{\mu_f}(2\nabla f(u_{k+1}) - \nabla f(u_k) + B^\top q_{k+1})\right),$
        $q_{k+1} = \frac{1}{1+\alpha}\left(q_k + \alpha p_{k+1} - \frac{\alpha}{\mu_g}(2\nabla g(p_{k+1}) - \nabla g(p_k) - B v_{k+1})\right).$
**end for**
**return** $u_{k+1}, v_{k+1}, p_{k+1}, q_{k+1}$
```

G AOR-HB-SADDLE-I ALGORITHM AND CONVERGENCE ANALYSIS

We propose the AOR-HB-saddle-I algorithm in the following Algorithm 4.

Consider the modified Lyapunov function

$$\mathcal{E}^\alpha(\mathbf{z}) := \mathcal{E}(\mathbf{z}) + \alpha\langle \nabla F(x) - \nabla F(x^*), y - x^* \rangle, \tag{61}$$

The following theorem show the convergence rate of AOR-HB-saddle-I method.

THEOREM G.1 (Convergence of AOR-HB-saddle-I method)**.** *Suppose f is μ_f-strongly convex and L_f-smooth, g is μ_g-strongly convex and L_g-smooth and let $\sqrt{\rho} = \min\left\{\sqrt{\frac{\mu_f}{L_f}}, \sqrt{\frac{\mu_g}{L_g}}\right\}$. Let (u_k, v_k, p_k, q_k) be generated by Algorithm 4 with initial value (u_0, v_0, p_0, q_0) and step size $\alpha = \sqrt{\rho}$. Then there exists a non-negative constant $C_0 = C_0(u_0, v_0, p_0, q_0, \mu_f, L_f, \mu_g, L_g)$ so that we have the linear convergence*

$$D_f(u_{k+1}, u^*) + D_g(p_{k+1}, p^*) + \frac{\mu_f}{2}\|v_{k+1} - u^*\|^2 + \frac{\mu_g}{2}\|q_{k+1} - p^*\|^2 \le C_0\left(\frac{1}{1 + \frac{1}{2}\sqrt{\rho}}\right)^k. \tag{62}$$

Proof. AOR-HB-saddle-I can be written as a correction of the implicit Euler method

$$\mathbf{z}_{k+1} - \mathbf{z}_k = \alpha\mathcal{G}(\mathbf{z}_{k+1}) - \alpha\begin{pmatrix} 0 & I \\ \mathcal{D}_\mu^{-1}\nabla F & 0 \end{pmatrix}(\mathbf{z}_{k+1} - \mathbf{z}_k). \tag{63}$$

The proof follows the proof of Theorem 1.1 in Appendix C with $f = F$.

□