Accelerated Mirror Descent Method through Variable and Operator Splitting

Anonymous authors
Paper under double-blind review

Abstract

Accelerated Mirror Descent (Acc-MD) is derived from a discretization of an accelerated mirror ODE system using a variable–operator splitting framework. A new Cauchy–Schwarz type inequality enables the first proof of linear accelerated convergence for mirror descent on a broad class of problems. Unlike prior methods based on the triangle scaling exponent (TSE), Acc-MD achieves acceleration in some cases where TSE fails. Experiments on smooth and composite optimization tasks show that Acc-MD consistently outperforms existing accelerated variants, both theoretically and empirically.

1 Introduction

Consider the unconstrained convex optimization problem

$$\min_{x\in\mathbb{R}^n} f(x),$$

where f is convex and differentiable. Mirror descent (Nemirovskij and Yudin, 1983) updates as

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \left\{ \langle \nabla f(x_k), x\rangle + \frac{1}{\alpha_k} D_\phi(x, x_k) \right\}, \tag{1}$$

where $\alpha_k > 0$, ϕ is a smooth, strictly convex *mirror function* defining the geometry, and $D_\phi(\cdot,\cdot)$ is the associated Bregman divergence. Nesterov (2005) proposed an accelerated variant.

Recent work views optimization algorithms as discretizations of continuous-time dynamics whose trajectories converge to minimizers. Wibisono et al. (2016) derived a Bregman Lagrangian that captures accelerated flows and showed that its discretization gives accelerated mirror descent. Krichene et al. (2015) extended the ODE framework for Nesterov acceleration (Su et al., 2016) to accelerated mirror descent, with discretizations yielding a class of first-order methods with $\mathcal{O}(L_f/k^2)$ convergence, where L_f is the Lipschitz constant of ∇f. Yuan and Zhang (2024) developed high-resolution ODEs and recovered the optimal $\mathcal{O}(L_f/k^2)$ rate for accelerated mirror descent methods in (Nesterov, 2005).

However, the analyses in (Nesterov, 2005; Wibisono et al., 2016; Krichene et al., 2015; Yuan and Zhang, 2024) do not incorporate the notion of relative smoothness, i.e., $D_f(x,y) \le L D_\phi(x,y)$, where D_f and D_ϕ denote the Bregman divergence of f and ϕ, respectively. As a result, their convergence guarantees may be loose outside the Euclidean setting, especially when L is bounded but L_f is not. This reflects a limitation of Nesterov's original theory (Nesterov, 2005), which does not fully use the geometry induced by the mirror map.

Under the relative smoothness condition, mirror descent achieves an $\mathcal{O}(L/k)$ convergence rate, as first shown in (Birnbaum et al., 2011). The notion of relative strong convexity, $\mu D_\phi(x,y) \le D_f(x,y)$, was later introduced in (Lu et al., 2018), leading to a linear rate $1-\mu/L$. For composite problems of the form $F(x) = f(x) + g(x)$, where f is smooth and g is convex but possibly non-smooth, the Bregman Proximal Gradient (BPG) method (Teboulle, 2018) achieves an $\mathcal{O}(L/k)$ rate under relative smoothness; see also (Bauschke et al., 2017; Lu et al., 2018; Zhou et al., 2019). The Accelerated Bregman Proximal Gradient (ABPG) method (Hanzely et al., 2021) obtains an accelerated rate $\mathcal{O}(L/k^\gamma)$ using the triangle scaling exponent (TSE) $\gamma \le 2$. Full acceleration $\mathcal{O}(L/k^2)$ requires $\gamma = 2$, and no accelerated linear rate is known under relative strong convexity. We summarize the convergence results of several mirror descent methods in Table 1.

Table 1: Convergence rates of accelerated mirror descent methods. The objective f and mirror map ϕ are assumed strongly convex and differentiable, where $C_{f,\phi}$ is a Cauchy-Schwarz constant in **(A2)**.

Algorithm / Theory	Assumptions	Convergence rate
Nesterov (2005) Yuan and Zhang (2024)	f: L_f-smooth	$\mathcal{O}(L_f/k^2)$
Krichene et al. (2015)	f: L_f-smooth	$\mathcal{O}(L_f/k^2)$
Birnbaum et al. (2011)	f: L-relative smooth	$\mathcal{O}(L/k)$
Hanzely et al. (2021)	triangle scaling exponent $\gamma \leq 2$ f: L-relative smooth	$\mathcal{O}(L/k^{\gamma})$
Lu et al. (2018)	f: μ-relative convex, L-relative smooth	$\mathcal{O}\big((1-\mu/L)^k\big)$
(**New**) Algorithm 1 Theorem 3.4	f: μ-relative convex, L-relative smooth Assumption **(A2)**	$\mathcal{O}\big((1-\sqrt{\mu/C_{f,\phi}})^k\big)$
(**New**) Algorithm 2 Theorem B.3	f: L-relative smooth, Assumption **(A2)**	$\mathcal{O}(C_{f,\phi}/k^2)$

Contributions. The main contributions of this work are as follows:

- Building on the variable and operator splitting framework of (Chen et al., 2025), we propose a new accelerated mirror descent (Acc-MD) flow with initial conditions $x(0) = x_0$, $y(0) = y_0$:

$$\begin{cases} x' = y - x, \\ (\nabla\phi(y))' = -\mu^{-1}\nabla f(x) + \nabla\phi(x) - \nabla\phi(y), \end{cases} \tag{2}$$

  We split the variable into x and y. At equilibrium, $y^* = x^*$ and $\nabla f(x^*) = 0$, recovering stationarity. The auxiliary variable y decouples the primal update and inertial motion.
- From an implicit–explicit discretization of (2), we design an Acc-MD method; see Algorithm 1. The algorithm is much simpler than existing counterparts in the literature, e.g., algorithms in (Nesterov, 2005; Krichene et al., 2015; Hanzely et al., 2021).
- We introduce a new Assumption **(A2)** based on a Cauchy-Schwarz-type inequality and, under this assumption, prove the first accelerated linear convergence guarantee under relative strong convexity μ, with rate $(1+\sqrt{\mu/C_{f,\phi}})^{-1}$, where $C_{f,\phi}$ is determined by **(A2)**. We present an example showing that our method accelerates whereas ABPG (Hanzely et al., 2021) does not.
- For the convex case $\mu = 0$, a perturbation–homotopy argument yields the optimal $\mathcal{O}(C_{f,\phi}/k^2)$ accelerated rate under Assumption **(A2)**. Without such structural assumptions, mirror descent methods are limited to $\mathcal{O}(1/k)$ complexity (Dragomir et al., 2022).
- We further extend the algorithm to composite optimization $\min_x f(x) + g(x)$ via the split gradient $\nabla f(x) + \nabla g(y)$:

$$(\nabla\phi(y))' = -\mu^{-1}\big(\nabla f(x) + \nabla g(y)\big) + \nabla\phi(x) - \nabla\phi(y), \tag{3}$$

  with an implicit step in $\nabla g(y)$. This yields a proximal variant that handles non-smooth g and covers constrained problems expressible in composite form.
- We benchmark Acc-MD on a range of smooth, non-smooth, and constrained convex problems, showing consistent advantages in both theory and practice over existing mirror descent methods. An adaptive parameter-update rule is proposed to further improve the numerical performance.

Limitation Although we provide one example where **(A2)** holds while TSE fails to yield acceleration, we do not expect **(A2)** to be universally verifiable or uniformly superior to TSE. Relaxing this assumption is therefore an important direction. A second limitation is the convexity requirement: extending the analysis to non-convex objectives remains open, and suitable non-convex mirror maps may be needed to guide convergence, though theoretical guarantees are currently unavailable. Extension to the stochastic setting is also important for applications in neural network training.

Preliminaries Let V be a normed vector space. For $f \in \mathcal{C}^1(V)$, define the *Bregman divergence*

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

For $f \in \mathcal{C}^1(V)$, f is convex if and only if $D_f(y,x) \geq 0$ for all $x, y \in V$. If f is strictly convex, then $D_f(y,x) = 0$ holds if and only if $x = y$. In general, the Bregman divergence is not symmetric, i.e., $D_f(y,x) \neq D_f(x,y)$. Its symmetrization is

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

A key tool in the convergence analysis is the three-point identity of Bregman divergence (Chen and Teboulle, 1993), which follows directly from the definition:

$$\langle \nabla f(y) - \nabla f(x), y - z\rangle = D_f(y,x) + D_f(z,y) - D_f(z,x), \tag{4}$$

We fix a smooth, strictly convex reference function ϕ, known as the *mirror function*. Let ϕ^* be the convex conjugate of ϕ. The mappings between primal and dual variables are

$$\chi = \nabla\phi(x), \quad x = \nabla\phi^*(\chi), \quad \eta = \nabla\phi(y), \quad y = \nabla\phi^*(\eta).$$

Here (x, y) are the primal variables and (χ, η) are their dual counterparts. The maps $\nabla\phi : V \to V^*$ and $\nabla\phi^* : V^* \to V$ are assumed to be one-to-one and efficiently computable. To clarify, we refer to ϕ as the mirror function and $\nabla\phi$ as the mirror map.

An important symmetry relation connects the Bregman divergences of ϕ and its conjugate:

$$D_\phi(x,y) = D_{\phi^*}(\eta,\chi), \tag{5}$$

with reversed argument order. Moreover, the gradient of the Bregman divergence with respect to its first argument satisfies

$$\nabla D_f(\cdot, x) = \nabla f(\cdot) - \nabla f(x), \quad \nabla D_{\phi^*}(\cdot, \chi) = \nabla\phi^*(\cdot) - \nabla\phi^*(\chi). \tag{6}$$

Let A be a self-adjoint, positive definite operator on a Hilbert space V with inner product $(\cdot,\cdot)$. Then

$$(x,y)_A := (Ax, y)$$

defines a new inner product, and the associated norm is denoted by $\|\cdot\|_A$. The corresponding dual norm is $\|\cdot\|_{A^{-1}}$. The convexity and Lipschitz constants of a differentiable function f relative to $\|\cdot\|_A$ are defined by

$$\mu_f(A)\,\|x-y\|_A^2 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle \leq L_f(A)\,\|x-y\|_A^2 \quad \forall\, x, y \in V.$$

When A is identity, L_f and μ_f are standard smoothness and convexity constants of f in the Euclidean norm. Choosing an appropriate inner product may reduce the condition number $\kappa_A(f) := L_f(A)/\mu_f(A)$, in which case A is called a *preconditioner*.

Analogously, the mirror function ϕ may be viewed as a *nonlinear preconditioner*. This is especially useful when L_f is unbounded but the relative smoothness L is finite.

2 Mirror Descent Methods

Given an initial condition $x(0) = x_0$, the mirror descent flow is defined by

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla\phi(x(t)) = -\nabla f(x(t)). \tag{7}$$

Discretizing the flow (7) using the explicit Euler method yields the iteration

$$\nabla\phi(x_{k+1}) - \nabla\phi(x_k) = -\alpha_k \nabla f(x_k), \tag{8}$$

where $\alpha_k > 0$ is the step size. This is equivalent to the classical mirror descent method (1) (Beck and Teboulle, 2003; Bubeck, 2015; Lu et al., 2018).

As noted by Krichene et al. (2015), the flow (7) can also be written in dual form:

$$\chi' = -\nabla f(x), \quad x = \nabla\phi^*(\chi), \tag{9}$$

which is closely related to the Bregman Inverse Scale Space dynamics (Osher et al., 2016).

We present the following three-point identity connecting the Bregman divergences of ϕ and f, which simplifies the convergence analysis compared to prior work, e.g., (Lu et al., 2018).

Lemma 2.1. *Let $\{x_k\}$ be the sequence generated by the mirror descent method* (8)*. Then*

$$\begin{aligned} &D_\phi(x^*, x_{k+1}) - D_\phi(x^*, x_k) + D_\phi(x_{k+1}, x_k) \\ =&\alpha_k \Big[- D_f(x_{k+1}, x^*) - D_f(x^*, x_k) + D_f(x_{k+1}, x_k) \Big]. \end{aligned} \tag{10}$$

Proof. Using the Bregman identity (4) and the update rule (8), we compute

$$\begin{aligned} D_\phi(x^*, x_{k+1}) - D_\phi(x^*, x_k) + D_\phi(x_{k+1}, x_k) &= \langle \nabla\phi(x_{k+1}) - \nabla\phi(x_k), x_{k+1} - x^* \rangle \\ = - \alpha_k \langle \nabla f(x_k) - \nabla f(x^*), x_{k+1} - x^* \rangle &= \alpha_k \left[-D_f(x_{k+1}, x^*) - D_f(x^*, x_k) + D_f(x_{k+1}, x_k) \right]. \end{aligned}$$

□

To establish linear convergence, we adopt the notion of relative smoothness and relative strong convexity introduced in (Lu et al., 2018). The significance of relative smoothness is that f may not be smooth, i.e., $L_f = \infty$, while L remains bounded.

Assumption (A1) The convex function f is said to be *relatively smooth* and *relatively convex* with respect to a mirror function ϕ if there exist constants $\mu \geq 0$ and $L \geq \mu$ such that

$$\mu D_\phi(x, y) \leq D_f(x, y) \leq L D_\phi(x, y) \quad \forall\, x, y \in V. \tag{11}$$

Using the linearity of the Bregman divergence, this is equivalent to $L\phi - f$ and $f - \mu\phi$ being convex.

Theorem 2.2. *Suppose f satisfies Assumption **(A1)**, and let $\{x_k\}$ be the sequence generated by the mirror descent iteration* (8)*. Then for any step size $\alpha_k \leq 1/L$, we have the decay property*

$$D_\phi(x^*, x_k) - D_\phi(x^*, x_{k+1}) \geq \alpha_k \big[D_f(x_{k+1}, x^*) + D_f(x^*, x_k) \big]. \tag{12}$$

In particular, if $\alpha_k = 1/L$ for all $k \geq 0$, then linear convergence holds:

$$D_\phi(x^*, x_k) \leq \left(1 - \frac{\mu}{L}\right)^k D_\phi(x^*, x_0). \tag{13}$$

Proof. From the identity (10) and Assumption **(A1)**, we have

$$\begin{aligned} D_\phi(x^*, x_{k+1}) - D_\phi(x^*, x_k) &\leq -\alpha_k \big[D_f(x_{k+1}, x^*) + D_f(x^*, x_k) \big] + (\alpha_k L - 1) D_\phi(x_{k+1}, x_k) \\ &\leq -\alpha_k \big[D_f(x_{k+1}, x^*) + D_f(x^*, x_k) \big] \leq -\alpha_k \mu D_\phi(x^*, x_k). \end{aligned}$$

Choosing $\alpha_k \equiv 1/L$ and rearranging the inequality to get the desired linear convergence. □

Similar results hold for proximal mirror descent methods (e.g., BPG) for composite optimization. We omit the details here, as an accelerated version will be presented in the next section.

3 Accelerated Mirror Descent Methods

We follow the recent variable and operator splitting (VOS) framework of (Chen et al., 2025) to develop accelerated mirror descent methods. We assume that f satisfies Assumption **(A1)** with constants L and $\mu > 0$. Notice that VOS is not inherently designed for mirror geometry; to extend it beyond the Euclidean setting, we introduce a geometric Assumption **(A2)**. Roughly speaking, **(A1)** controls distances while **(A2)** imposes an angle-type condition.

Assumption (A2) There exists $C_{f,\phi} > 0$ such that the relative Cauchy Schwarz inequality holds:

$$|\,\langle \nabla f_{-\mu}(x) - \nabla f_{-\mu}(\hat{x}), y - \hat{y} \rangle\,| \leq 2\sqrt{C_{f,\phi}}\, D_{f_{-\mu}}^{1/2}(x, \hat{x}) D_\phi^{1/2}(\hat{y}, y), \quad \forall x, \hat{x}, y, \hat{y} \in V.$$

Flow Introducing dual variables $\chi = \nabla\phi(x)$ and $\eta = \nabla\phi(y)$, we can write the accelerated mirror descent flow (2) as

$$x' = y - x, \quad \eta' = -\mu^{-1}\nabla f_{-\mu}(x) - \eta, \tag{14}$$

where $f_{-\mu} := f - \mu\phi$, or equivalently $f = f_{-\mu} + \mu\phi$. Under Assumption **(A1)**, $f_{-\mu}$ is convex.

Stability We define a Lyapunov function:

$$\mathcal{E}(x,\eta) := D_{f_{-\mu}}(x,x^*) + \mu D_{\phi^*}(\eta,\chi^*), \tag{15}$$

where x^* is a minimizer of f and $\chi^* = \nabla\phi(x^*)$. This energy couples primal and dual Bregman divergences, capturing the geometry induced by the mirror map. We establish exponential stability by verifying a strong Lyapunov property (Chen and Luo, 2021).

Lemma 3.1. *Let $\mathcal{E}(x,\eta)$ be defined by* (15), *and define the vector field $\mathcal{G}(x,\eta) = (y - x,\, -\mu^{-1}\nabla f_{-\mu}(x) - \eta)$, where the dual variables satisfy $\chi = \nabla\phi(x)$ and $\eta = \nabla\phi(y)$. Then*

$$-\nabla\mathcal{E}(x,\eta)\cdot\mathcal{G}(x,\eta) = \mathcal{E}(x,\eta) + D_{f_{-\mu}}(x^*,x) + \mu D_\phi(y,x^*). \tag{16}$$

As a consequence, any solution $(x(t),y(t))$ of the flow (2) *satisfies the exponential decay bound*

$$\mathcal{E}(x(t),\eta(t)) \le e^{-t}\mathcal{E}(x(0),\eta(0)). \tag{17}$$

Proof. Using the identity (6), we compute the gradients of the Lyapunov function:

$$\partial_x\mathcal{E} = \nabla f_{-\mu}(x) - \nabla f_{-\mu}(x^*), \quad \partial_\eta\mathcal{E} = \mu\left(\nabla\phi^*(\eta) - \nabla\phi^*(\chi^*)\right) = \mu(y - x^*).$$

Thus, by the direct calculation and symmetry relation (5), we have

$$\begin{aligned}
-\nabla\mathcal{E}(x,\eta)\cdot\mathcal{G}(x,\eta) &= \langle\nabla f_{-\mu}(x) - \nabla f_{-\mu}(x^*), x - y\rangle \\
&\quad + \mu\left\langle y - x^*,\, \mu^{-1}(\nabla f_{-\mu}(x) - \nabla f_{-\mu}(x^*)) + \eta - \chi^*\right\rangle \\
&= \langle\nabla f_{-\mu}(x) - \nabla f_{-\mu}(x^*), x - x^*\rangle + \mu\langle\nabla\phi^*(\eta) - \nabla\phi^*(\chi^*), \eta - \chi^*\rangle \\
&= D_{f_{-\mu}}(x,x^*) + D_{f_{-\mu}}(x^*,x) + \mu D_{\phi^*}(\chi^*,\eta) + \mu D_{\phi^*}(\eta,\chi^*) \\
&= \mathcal{E}(x,\eta) + D_{f_{-\mu}}(x^*,x) + \mu D_\phi(y,x^*).
\end{aligned}$$

Since $D_{f_{-\mu}}(x^*,x) \ge 0$ and $D_\phi(y,x^*) \ge 0$ by convexity of $f_{-\mu}$ and ϕ, we obtain the inequality

$$\nabla\mathcal{E}(x,\eta)\cdot\mathcal{G}(x,\eta) \le -\mathcal{E}(x,\eta),$$

from which the exponential decay (17) follows by Grönwall's inequality. □

As a corollary, we have $x(t), y(t) \in B(x^*,R)$ for all $t \ge 0$, where $B(x^*,R)$ denotes the ball of radius R centered at x^*.

Accelerated Mirror Descent Methods We propose an accelerated mirror descent (Acc-MD) method by an implicit–explicit scheme of (2):

$$\frac{x_{k+1} - x_k}{\alpha} = 2y_{k+1} - y_k - x_{k+1}, \tag{18a}$$

$$\frac{\nabla\phi(y_{k+1}) - \nabla\phi(y_k)}{\alpha} = -\frac{1}{\mu}\nabla f(x_k) + \nabla\phi(x_k) - \nabla\phi(y_{k+1}). \tag{18b}$$

In (18b), $\nabla f_{-\mu}(x_k)$ is explicit and $-\nabla\phi(y_{k+1})$ is implicit. The implicit term improves stability, similar to the proximal point algorithm. The discretization $y \approx (2y_{k+1} - y_k)$ in (18a) acts as an accelerated over-relaxation (AOR) (Wei and Chen, 2025), symmetrizing the error equation and enabling our proof of the accelerated rate.

Algorithm 1 Accelerated mirror descent (Acc-MD) method

```
Parameters: $x_0, y_0 \in \mathbb{R}^n, \mu, C_{f,\phi}$.
Set $\alpha = \sqrt{\mu/C_{f,\phi}}$.
for $k = 0, 1, 2, \dots$ do
    $y_{k+1} = \arg\min_{y\in\mathbb{R}^n} (1+\alpha)\phi(y) - \left\langle -\frac{\alpha}{\mu}\nabla f(x_k) + \alpha\nabla\phi(x_k) + \nabla\phi(y_k), y\right\rangle.$
    $x_{k+1} = \frac{1}{1+\alpha}[x_k + \alpha(2y_{k+1} - y_k)].$
end for
```

An equivalent computation favorable form using $C_{f,\phi}$ in **(A2)** is given as Algorithm 1. When $C_{f,\phi}$ is difficult to estimate, a line-search procedure can be used to obtain an adaptive estimate; see (29).

Convergence analysis For four points $(x, \hat{x}, y, \hat{y})$ and $\alpha \in \mathbb{R}$, introduce the cross term

$$\mathcal{B}^{\alpha}(x, \hat{x}, \hat{y}, y) = D_{f_{-\mu}}(x, \hat{x}) + \mu D_{\phi}(\hat{y}, y) - \alpha \langle \nabla f_{-\mu}(x) - \nabla f_{-\mu}(\hat{x}),\, y - \hat{y} \rangle. \tag{19}$$

Assumption **(A2)** is essential both for defining the algorithm and for establishing its convergence.

Lemma 3.2. *Let Assumption* **(A2)** *hold. Then for* $|\alpha| \leq \sqrt{\mu / C_{f,\phi}}$*, we have* $\mathcal{B}^{\alpha}(x, \hat{x}, \hat{y}, y) \geq 0$.

Proof. By **(A2)** and the inequality $2ab \leq a^2 + b^2$, we have

$$\begin{aligned} \alpha |\, \langle \nabla f_{-\mu}(x) - \nabla f_{-\mu}(\hat{x}), y - \hat{y} \rangle \,| &\leq 2\alpha \sqrt{C_{f,\phi}}\, D_{f_{-\mu}}^{1/2}(x, \hat{x}) D_{\phi}^{1/2}(\hat{y}, y) \\ &\leq \alpha \sqrt{C_{f,\phi}/\mu}\, \left(D_{f_{-\mu}}(x, \hat{x}) + \mu D_{\phi}(\hat{y}, y) \right). \end{aligned}$$

Therefore $\mathcal{B}^{\alpha}(x, \hat{x}, \hat{y}, y) \geq 0$ if $\alpha \leq \sqrt{\mu / C_{f,\phi}}$. □

Define the modified Lyapunov function

$$\mathcal{E}^{\alpha}(x, y) := \mathcal{B}^{\alpha}(x, x^*, x^*, y) := \mathcal{E}(x, \eta) - \alpha \langle \nabla f_{-\mu}(x) - \nabla f_{-\mu}(x^*), y - x^* \rangle. \tag{20}$$

As shown in Lemma 3.2, when $\alpha \leq \sqrt{\mu / C_{f,\phi}}$, $\mathcal{E}^{\alpha}(x, y) \geq 0$ is indeed a Lyapunov function.

Lemma 3.3. *Let* (x_k, y_k) *be the sequence generated by Acc-MD iterations (18), then it holds*

$$\begin{aligned} \mathcal{E}^{\alpha}(x_{k+1}, y_{k+1}) - \mathcal{E}^{\alpha}(x_k, y_k) = &- \alpha \mathcal{E}^{\alpha}(x_{k+1}, y_{k+1}) \\ &- \alpha\, \mathcal{B}^{-\alpha}(x^*, x_{k+1}, y_{k+1}, x^*) - \mathcal{B}^{\alpha}(x_k, x_{k+1}, y_{k+1}, y_k). \end{aligned} \tag{21}$$

Proof. Denote $\boldsymbol{z} = (x, \eta)$. Expand the difference of the Lyapunov function $\mathcal{E}(\boldsymbol{z})$ at $\boldsymbol{z}_{k+1}$:

$$\begin{aligned} \mathcal{E}(\boldsymbol{z}_{k+1}) - \mathcal{E}(\boldsymbol{z}_k) &= \langle \nabla \mathcal{E}(\boldsymbol{z}_{k+1}), \boldsymbol{z}_{k+1} - \boldsymbol{z}_k \rangle - D_{\mathcal{E}}(\boldsymbol{z}_k, \boldsymbol{z}_{k+1}) \\ &= \alpha \langle \nabla \mathcal{E}(\boldsymbol{z}_{k+1}), \mathcal{G}(\boldsymbol{z}_{k+1}) \rangle - D_{\mathcal{E}}(\boldsymbol{z}_k, \boldsymbol{z}_{k+1}) \\ +\alpha \langle \nabla f_{-\mu}(x_{k+1}) - \nabla f_{-\mu}(x^*), y_{k+1} - y_k \rangle &+ \alpha \langle y_{k+1} - x^*, \nabla f_{-\mu}(x_{k+1}) - \nabla f_{-\mu}(x_k) \rangle. \end{aligned}$$

The last line is the difference with the implicit Euler discretization and can be symmetrized as

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

Then we use identity (16) to expand $\langle \nabla \mathcal{E}(\boldsymbol{z}_{k+1}), \mathcal{G}(\boldsymbol{z}_{k+1}) \rangle$ and rearrange the terms to get

$$\begin{aligned} \mathcal{E}^{\alpha}(x_{k+1}, y_{k+1}) - \mathcal{E}^{\alpha}(x_k, y_k) = &- \alpha \mathcal{E}(x_{k+1}, \eta_{k+1}) - \alpha D_{f_{-\mu}}(x^*, x_{k+1}) - \mu \alpha D_{\phi}(y_{k+1}, x^*) \\ &- D_{\mathcal{E}}(\boldsymbol{z}_k, \boldsymbol{z}_{k+1}) + \alpha \langle \nabla f_{-\mu}(x_{k+1}) - \nabla f_{-\mu}(x_k), y_{k+1} - y_k \rangle \\ = &- \alpha \mathcal{E}^{\alpha}(x_{k+1}, y_{k+1}) - \alpha D_{f_{-\mu}}(x^*, x_{k+1}) - \mu \alpha D_{\phi}(y_{k+1}, x^*) \\ &- \alpha \langle \nabla f_{-\mu}(x_{k+1}) - \nabla f_{-\mu}(x^*), y_{k+1} - x^* \rangle \\ &- \mathcal{B}^{\alpha}(x_k, x_{k+1}, y_{k+1}, y_k). \end{aligned}$$

□

Theorem 3.4 (Convergence of Acc-MD method)**.** *Suppose* f *is* μ*-relatively convex with* $\mu > 0$ *and* **(A2)** *holds with constant* $C_{f,\phi}$*. Let* (x_k, y_k) *be generated by scheme (18) with initial value* (x_0, y_0)*,* $\eta_k = \nabla \phi(y_k)$*, and step size* $\alpha = \sqrt{\mu / C_{f,\phi}}$*. Then there exists a constant* $C_0 = C_0(x_0, y_0, \mu, C_{f,\phi})$ *so that we have the accelerated linear convergence*

$$D_{f_{-\mu}}(x_{k+1}, x^*) + \mu D_{\phi^*}(\eta_{k+1}, \chi^*) \leq C_0 \left(\frac{1}{1 + \sqrt{\mu / C_{f,\phi}}} \right)^k, \quad k \geq 1. \tag{22}$$

Proof. By Lemma 3.2, for $\alpha = \sqrt{\mu / C_{f,\phi}}$, we can drop negative terms from the identity (21) to get the linear convergence

$$\mathcal{E}^{\alpha}(x_{k+1}, y_{k+1}) \leq \frac{1}{1 + \alpha} \mathcal{E}^{\alpha}(x_k, y_k) \leq \left(\frac{1}{1 + \sqrt{\mu / C_{f,\phi}}} \right)^{k+1} \mathcal{E}^{\alpha}(x_0, y_0). \tag{23}$$

From the proof of Lemma 3.3, we have

$$\alpha \mathcal{E}(x_{k+1}, \eta_{k+1}) \le \mathcal{E}^\alpha(x_k, y_k) - \mathcal{E}^\alpha(x_{k+1}, y_{k+1}) \le \left(\frac{1}{1 + \sqrt{\mu / C_{f,\phi}}} \right)^k \mathcal{E}^\alpha(x_0, y_0),$$

which leads to (22). □

Discussion on Assumption (A2) Without additional structure, Dragomir et al. (2022) show that mirror descent type methods cannot surpass $\mathcal{O}(1/k)$ complexity. Thus, Assumption **(A2)** is essential for achieving acceleration. We next verify **(A2)** for a broad class of functions. Due to space limitations, the proofs are deferred to Appendix A.

Theorem 3.5. *Let A be a self-adjoint, positive definite (SPD) operator. Assume $f_{-\mu} = f - \mu\phi$ is $L_{f_{-\mu}}$-smooth and ϕ is μ_ϕ-strongly convex. Then*

$$C_{f,\phi} \le \inf_{SPD\ A} \frac{L_{f_{-\mu}}(A)}{\mu_\phi(A)} \le \frac{L_{f_{-\mu}}}{\mu_\phi}.$$

Example 3.6 (Log-linear model)**.** Consider the log-linear dual model in supervised machine learning (Collins et al., 2008):

$$f(x) = \sum_{i=1}^d x_i \log x_i + \frac{1}{2} x^\top A x,$$

where $x = (x_i) \in \mathbb{R}^d$ denotes the dual variables, and $A = \mathbf{g}\mathbf{g}^\top$ with a given $\mathbf{g} \in \mathbb{R}^d$. Taking Shannon's entropy function

$$\phi(x) = \sum_i x_i \log x_i$$

as the mirror function and setting $\mu = 1$, we obtain $f_{-\mu} = \frac{1}{2} x^\top A x$, which is quadratic with $L_{f_{-\mu}}(A) = 1$.

It is well known that the Bregman divergence $D_\phi(x, z)$ can be reformulated as the KL-divergence between two discrete probability measures: $D_\phi(x, z) = \mathrm{KL}(x, z) := \sum_i x_i \log\left(\frac{x_i}{z_i}\right)$. (Collins et al., 2008, Lemma 7) proved that f is 1-relatively strongly convex and $(1 + |A|_\infty)$-relatively smooth with respect to ϕ, where $|A|_\infty$ refers to the largest entry in magnitude across all rows and columns of A. Thus, Assumption **(A1)** holds with $\mu = 1$ and $L = |A|_\infty$.

By Pinsker's inequality, $\|y - \hat{y}\|_1 \le \sqrt{2\,\mathrm{KL}(\hat{y}, y)}$, we obtain

$$\|y - \hat{y}\|_A^2 \le \|\mathbf{g}\|_2^2 \|y - \hat{y}\|_2^2 \le \|\mathbf{g}\|_2^2 \|y - \hat{y}\|_1^2 \le 2\|\mathbf{g}\|_2^2 \mathrm{KL}(\hat{y}, y) = 2\|\mathbf{g}\|_2^2 D_\phi(\hat{y}, y). \tag{24}$$

Therefore, by Theorem 3.5, Assumption **(A2)** holds with constant

$$C_{f,\phi} = \frac{L_{f_{-\mu}}(A)}{\mu_\phi(A)} = \|\mathbf{g}\|_2^2.$$

Another important class is when the mirror function ϕ has continuous Hessian in the ball $B(x^*, R)$.

Theorem 3.7. *Assume f satisfies Assumption **(A1)**, and $\phi \in C^{2,1}(V)$ with ϕ being 1-strongly convex. When $(x, \hat{x}, y, \hat{y}) \in B(x^*, R)$, Assumption **(A2)** holds with*

$$C_{f,\phi} \le (L - \mu)\big(1 + 2L_{\nabla^2\phi} R\big),$$

where $L_{\nabla^2\phi}$ is the Lipschitz constant of $\nabla^2\phi$ restricted to $B(x^*, R)$.

In Section 4, we present the quartic objective where $C_{f,\phi}$ could be estimated locally. In particular, when ϕ is quadratic, $\nabla^2\phi$ is constant and $L_{\nabla^2\phi} = 0$, recovering a sharper estimate $C_{f,\phi} \le L - \mu$.

Extension to convex optimization For convex objectives, i.e., $\mu = 0$, we consider the following perturbed flow:

$$\begin{cases} \qquad\quad x' = y - x, \\ (\nabla\phi(y))' = \epsilon^{-1}\left[\epsilon\big(\nabla\phi(x) - \nabla\phi(y)\big) - \nabla f(x)\right], \end{cases} \tag{25}$$

for a fixed perturbation level $\epsilon > 0$. In dual variables, this becomes

$$x' = y - x, \quad \eta' = -\epsilon^{-1}\nabla f(x) + \chi - \eta,$$

where $\chi = \nabla\phi(x)$ and $\eta = \nabla\phi(y)$. Owing to the variable splitting structure, the perturbation does not change the stationary points of the original system.

To accelerate convergence, we employ a homotopy strategy that gradually decreases the perturbation level. Starting from ϵ_0, we run accelerated mirror descent for m_k iterations at level ϵ_k, and update

$$\epsilon_{k+1} = \frac{\epsilon_k}{2}, \qquad m_{k+1} = \sqrt{2}\, m_k.$$

This geometric schedule ensures an accelerated sublinear convergence rate $\mathcal{O}(C_{f,\phi}/k^2)$. The convergence analysis follows (Chen et al., 2025) and is detailed in Appendix B.

Our algorithm and analysis also extend to constrained optimization, as long as all iterates remain inside the constraint set. By selecting a suitable mirror function ϕ, the constraints can be naturally incorporated. This is another advantage of mirror descent methods.

Example 3.8 (Max-margin model)**.** Consider the max-margin dual model in Collins et al. (2008):

$$\min_{x \in \Delta^d} f(x) = b^\top x + \frac{1}{2} x^\top A x,$$

where Δ^d denotes the d-dimensional probability simplex, $b \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$ is SPD. We take Shannon's entropy function $\phi(x) = \sum_{i=1}^d x_i \log x_i$ as the mirror function which will preserve the iterate x_k remains in the constraint set Δ^d. f is relative convex with $\mu = 0$. It is straight forward to verity that $L_{f_{-\mu}}(A) = L_f(A) \le 1$. Similar to (24), we can get $\mu_\phi(A) = 1/\|A\|_2$ and thus $C_{f,\phi} = \|A\|_2$.

Extension to composite optimization Consider the composite optimization problem

$$\min_{x \in \mathbb{R}^n} \quad F(x) := f(x) + g(x), \tag{26}$$

where f satisfies Assumption **(A1)**, and g is convex but may be non-smooth, with a well-defined generalized proximal operator with respect to the mirror function ϕ; see (27). Assumptions **(A1)** and **(A2)** are imposed solely on the smooth component f. When $g(\cdot)$ denotes the indicator function on a convex set, we can cover a large class of optimization problems over convex domain.

To extend the accelerated mirror descent method to the composite setting, the only change lies in Line 4 of Algorithm 1, where y_{k+1} is computed via a generalized proximal operator:

$$y_{k+1} = \arg\min_{y \in \mathbb{R}^n} (1+\alpha)\,\phi(y) + \frac{\alpha}{\mu} g(y) - \left\langle -\frac{\alpha}{\mu}\nabla f(x_k) + \alpha\nabla\phi(x_k) + \nabla\phi(y_k),\, y \right\rangle. \tag{27}$$

Since y is treated implicitly, the convergence analysis remains the same as in the smooth case, showing the flexibility of the VOS framework. For details, see Appendix C. A perturbation and homotopy argument can be further applied if $\mu = 0$.

Example 3.9 (LASSO problem)**.** Consider the over-parameterized LASSO problem:

$$\min_{x \in \mathbb{R}^d} F(x) := \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1, \tag{28}$$

where $A \in \mathbb{R}^{n\times d}$ with $n < d$. We adopt the mirror function $\phi(x) = \frac{1}{2}x^\top D x$, where $D = \operatorname{diag}(A^\top A)$. As D is diagonal and positive definite, the Acc-MD subproblem (27) can be solved by a generalized soft-thresholding operator in closed form. The relative smoothness constant is $L = \rho(D^{-1/2}A^\top A D^{-1/2})$ where $\rho(\cdot)$ denotes the spectral radius. Then one can simply take $C_{f,\phi} = \rho(D^{-1/2}A^\top A D^{-1/2})$ according to Theorem 3.7.

4 Numerical Examples

We evaluate the performance of Acc-MD on a variety of convex optimization problems. All experiments were conducted in MATLAB R2023a on a desktop with an Intel Core i5-7200U CPU (2.50 GHz) and 8 GB RAM. Random seeds were fixed for reproducibility.

We compare Acc-MD against several state-of-the-art first-order methods from the literature. In all tested scenarios, Acc-MD consistently outperforms competing algorithms by a large margin.

Quartic Objective We evaluate Acc-MD on smooth convex minimization problems and compare it with the following first-order methods: Nesterov's accelerated mirror descent (NAMD) (Nesterov, 2005); Nesterov's accelerated gradient (NAG) (Nesterov, 1983) with step size $1/(k+3)$; Accelerated over-relaxation heavy ball (AOR-HB) (Wei and Chen, 2025) and Mirror descent (MD) (Lu et al., 2018). All methods use the stopping criterion $\|\nabla f(x_k)\| \leq \mathrm{tol} \cdot \|\nabla f(x_0)\|$ with tolerance $\mathrm{tol} = 10^{-6}$.

We consider the quartic objective (Lu et al., 2018, Section 2.1):

$$f(x) = \frac{1}{4}\|Ex\|_2^4 + \frac{1}{4}\|Ax - b\|_4^4 + \frac{1}{2}\|Cx - d\|_2^2,$$

where C and E are $n \times n$ positive definite matrices. The mirror function is chosen as

$$\phi(x) = \frac{1}{4}\|x\|_2^4 + \frac{1}{2}\|x\|_2^2.$$

The update step takes the form $y_{k+1} = \nabla\phi^*(c) = \theta c$, where $\theta > 0$ solves the cubic equation $\|c\|_2^2\theta^3 + \theta - 1 = 0$, and can be calculated by a root-finding routine in MATLAB, and $c = \frac{1}{1+\alpha}\left[\alpha\nabla\phi(x_k) + \nabla\phi(y_k) - \frac{\alpha}{\mu}\nabla f(x_k)\right]$. We refer to (Lu et al., 2018) for the full derivation.

Let λ_C and λ_E denote the smallest eigenvalues of C and E, respectively. Since $\|\nabla^2 f\|$ grows quadratically with $\|x\|_2$, the gradient ∇f is not globally Lipschitz. To apply NAG and AOR-HB, we assume $\|x\|_2 \leq R$ and estimate the global parameters as

$$L_f = (3\|E\|^4 + 3\|A\|^4)R^2 + 6\|A\|^3\|b\|_2 R + 3\|A\|^2\|b\|_2^2 + \|C\|^2, \qquad \mu = \lambda_C^2.$$

According to Lu et al. (2018), **(A1)** holds with the relative smoothness constant $L = L_f$ when $R = 1$, and the relative strong convexity constant $\mu = \min\left\{\lambda_E^4/3,\ \lambda_C^2\right\}$. Since ϕ is smooth, we apply Theorem 3.7 to estimate $C_{f,\phi}$ locally. Furthermore, we propose an adaptive strategy to compute $C_{f,\phi}$.

Adaptive update for parameter $C_{f,\phi}$ When **(A2)** may be hard to verify globally, a practical choice is to estimate $C_{f,\phi}$ from two consecutive iterations

$$\frac{|\langle \nabla f_{-\mu}(x_k) - \nabla f_{-\mu}(x_{k-1}), y_k - y_{k-1}\rangle|^2}{4D_{f_{-\mu}}(x_k, x_{k-1})D_\phi(y_{k-1}, y_k)} = C_{f,\phi}^k, \tag{29}$$

and then use $\alpha = \sqrt{\mu/C_{f,\phi}}^{k}$ to compute (x_{k+1}, y_{k+1}) and verify whether condition **(A2)** holds with $(x_k, x_{k+1}, y_{k+1}, y_k)$. If **(A2)** fails, we increase $C_{f,\phi}^k$ by a factor $r > 1$ and recompute (x_{k+1}, y_{k+1}) until **(A2)** is satisfied. This leads to a natural line-search variant. We follow (Cavalcanti et al., 2024) and employ an adaptive backtracking approach to determine the factor r.

Set $n = 256$, $A \sim \frac{1}{\sqrt{n}}\mathcal{N}(\mathbf{0}, I_{n\times n})$, $C = I_n + C_0C_0^\top/n$ with $C_0 \sim \mathcal{N}(\mathbf{0}, I_{n\times n})$, $E = 2I_n + E_0E_0^\top/n$ with $E_0 \sim \mathcal{N}(\mathbf{0}, I_{n\times n})$, $b = 0$, and $d \sim \mathrm{Unif}(0,1)^n$. This gives $\mu = \mu_\phi = 1$, L remains fixed, while L_f grows with $R = \|d\|_2$, highlighting the benefit of using relative smoothness.

We compare Acc-MD with the adaptive rule (29) (Acc-MD-ad), the original version in Algorithm 1 (Acc-MD-origin), and several other first-order methods. The iteration counts and running times are reported in Figures 1 and 2. Both Acc-MD variants and MD converge faster and more stably than all non-mirror-type methods, and Acc-MD further improves over MD, confirming the acceleration effect. Acc-MD-ad achieves substantial gains over Acc-MD-origin, reducing iterations by about 90% and total running time by about 80%. Although each Acc-MD-ad iteration is more expensive due to line search, the reduction in iteration count dominates, yielding the best overall performance.

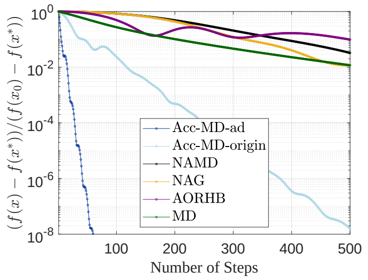

Figure 1: Log relative error vs. iteration curves on the quartic problem.

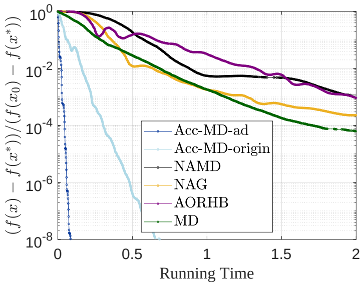

Figure 2: Log relative error vs. running time curves on the quartic problem.

Entropic Mirror Descent Consider the log-linear models in Example 3.6, where we have shown that $\mu = 1$ and $C_{f,\phi} = \|A\|_2 = \|\mathbf{g}\|_2^2$. Therefore, Acc-MD achieves the accelerated linear rate $(1 + 1/\|\mathbf{g}\|_2)^{-1}$ when applied to this example, whereas existing entropic mirror descent methods attain at best the sublinear rate $\mathcal{O}(1/k^2)$.

We compare Acc-MD with Accelerated Bregman Proximal Gradient (ABPG) and Bregman Proximal Gradient (BPG) from Hanzely et al. (2021) and report the results in Figure 3. For this example, the TSE component is $\gamma = 1$, and ABPG achieves only a non-accelerated rate of $\mathcal{O}(1/k)$, converging slightly slower than BPG, which agrees with the results in (Hanzely et al., 2021, Section 6.1.1).

Acc-MD significantly outperforms the other methods and shows a steep and stable error curve. The adaptive variant Acc-MD-ad can also be used, but our experiments show that its performance is not robust and may stagnate. This instability likely arises from numerical errors that violate condition (A2). Designing a more reliable adaptive scheme remains an interesting direction for future work.

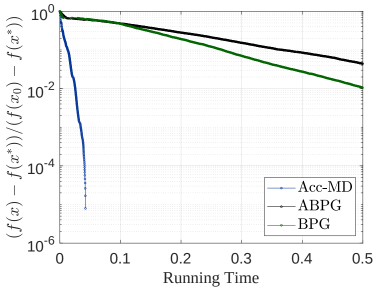

Figure 3: Log relative error vs. iteration curves on the entropic MD problem.

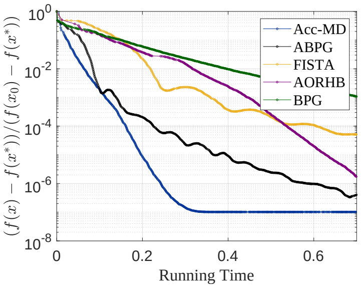

Figure 4: Relative error vs. running time curves for the LASSO problem.

LASSO problem on real-world datasets We evaluate the over-parameterized LASSO problem (Example 3.9) on the Leukemia dataset, a standard benchmark in the Lasso literature (Šehić et al., 2022). The dataset contains 7,129 gene-expression features across 72 samples used to predict Leukemia type. We fix $\lambda = 0.05$, noting that larger regularization induces stronger sparsity and accelerates convergence for all methods; we intentionally choose a small value to accentuate performance differences.

As the model is over-parameterized ($n < d$), $\mu = 0$, and we therefore apply the perturbed variant of Acc-MD (Algorithm 2). As shown in Figure 4, Acc-MD converges markedly faster than competing approaches and naturally exhibits an early-stopping effect.

We provide more examples, including applications to composite convex optimization (LASSO on synthetic data) and constrained convex optimization (quadratic optimization on the simplex), in Appendix D. In all tested scenarios, Acc-MD consistently outperforms competing algorithms by a large margin.

References

Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003. ISSN 0167-6377. doi: https://doi.org/10.1016/S0167-6377(02)00231-6.

Benjamin Birnbaum, Nikhil R. Devanur, and Lin Xiao. Distributed algorithms via gradient descent for Fisher markets. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, EC '11, pages 127–136, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302616. doi: https://doi.org/10.1145/1993574.1993594.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3–4):231–357, November 2015. ISSN 1935-8237. doi: https://doi.org/10.1561/2200000050.

Joao V Cavalcanti, Laurent Lessard, and Ashia C Wilson. Adaptive backtracking line search. *arXiv preprint arXiv:2408.13150*, 2024.

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. doi: https://doi.org/10.1137/0803026.

Long Chen and Hao Luo. A unified convergence analysis of first order convex optimization methods via strong Lyapunov functions, 2021. URL `https://arxiv.org/abs/2108.00132`.

Long Chen, Luo Hao, and Jingrong Wei. Accelerated gradient methods through variable and operator splitting, 2025. URL `https://arxiv.org/abs/2505.04065`.

Michael Collins, Amir Globerson, Terry Koo, Xavier Carreras, and Peter L. Bartlett. Exponentiated gradient algorithms for conditional random fields and max-margin Markov networks. *Journal of Machine Learning Research*, 9(58):1775–1822, 2008. URL `http://jmlr.org/papers/v9/collins08a.html`.

Radu-Alexandru Dragomir, Adrien B. Taylor, Alexandre d'Aspremont, and Jérôme Bolte. Optimal complexity and certification of Bregman first-order methods. *Mathematical Programming*, 194(1): 41–83, 2022. doi: https://doi.org/10.1007/s10107-021-01618-1.

Filip Hanzely, Peter Richtárik, and Lin Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *Computational Optimization and Applications*, 79(2): 405–440, June 2021. doi: https://doi.org/10.1007/s10589-021-00273-8.

Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL `https://proceedings.neurips.cc/paper_files/paper/2015/file/f60bb6bb4c96d4df93c51bd69dcc15a0-Paper.pdf`.

Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018. doi: https://doi.org/10.1137/16M1099546.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983. ISBN 9780471103455.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $o\left(\frac{1}{k^2}\right)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983.

Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1): 127–152, 2005. doi: https://doi.org/10.1007/s10107-004-0552-5.

Stanley Osher, Feng Ruan, Jiechao Xiong, Yuan Yao, and Wotao Yin. Sparse recovery via differential inclusions. *Applied and Computational Harmonic Analysis*, 41(2):436–469, 2016. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2016.01.002. Sparse Representations with Applications in Imaging Science, Data Analysis, and Beyond, Part II.

Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17 (153):1–43, 2016. URL `http://jmlr.org/papers/v17/15-084.html`.

Marc Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, 170(1):67–96, 2018. doi: https://doi.org/10.1007/s10107-018-1284-2.

Kenan Šehić, Alexandre Gramfort, Joseph Salmon, and Luigi Nardi. Lassobench: A high-dimensional hyperparameter optimization benchmark suite for lasso. In Isabelle Guyon, Marius Lindauer, Mihaela van der Schaar, Frank Hutter, and Roman Garnett, editors, *Proceedings of the First International Conference on Automated Machine Learning*, volume 188 of *Proceedings of Machine Learning Research*, pages 2/1–24. PMLR, 25–27 Jul 2022. URL `https://proceedings.mlr.press/v188/sehic22a.html`.

Jingrong Wei and Long Chen. Accelerated over-relaxation Heavy-Ball method: Achieving global accelerated convergence with broad generalization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=SWEqzy7IQB`.

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. doi: https://doi.org/10.1073/pnas.1614734113.

Ya-Xiang Yuan and Yi Zhang. Analyze accelerated mirror descent via high-resolution ODEs. *Journal of the Operations Research Society of China*, 2024. doi: https://doi.org/10.1007/s40305-024-00542-3.

Yi Zhou, Yingbin Liang, and Lixin Shen. A simple convergence analysis of Bregman proximal gradient algorithm. *Computational Optimization and Applications*, 73(3):903–912, 2019. doi: https://doi.org/10.1007/s10589-019-00092-y.

A Missing proofs

Theorem 3.5. *Let A be a self-adjoint, positive definite operator. Assume $f_{-\mu} = f - \mu\phi$ is smooth and ϕ is strongly convex. Then*

$$C_{f,\phi} \le \inf_{SPD\ A} \frac{L_{f_{-\mu}}(A)}{\mu_\phi(A)} \le \frac{L_{f_{-\mu}}}{\mu_\phi}.$$

Proof. For any SPD A, we have

$$\begin{aligned}\langle \nabla f_{-\mu}(x) - \nabla f_{-\mu}(\hat{x}),\, y - \hat{y}\rangle &= \langle A^{-1/2}(\nabla f_{-\mu}(x) - \nabla f_{-\mu}(\hat{x})),\, A^{1/2}(y-\hat{y})\rangle \\ &\le \|\nabla f_{-\mu}(x) - \nabla f_{-\mu}(\hat{x})\|_{A^{-1}} \|y - \hat{y}\|_A.\end{aligned}$$

Applying strong convexity of ϕ in $\|\cdot\|_A$ norm and co-coercivity of $f_{-\mu}$ in $\|\cdot\|_{A^{-1}}$ norm, we obtain

$$2D_\phi(\hat{y}, y) \ge \mu_\phi(A)\|y - \hat{y}\|_A^2, \quad 2D_{f_{-\mu}}(x, \hat{x}) \ge \frac{1}{L_{f_{-\mu}}(A)} \|\nabla f_{-\mu}(x) - \nabla f_{-\mu}(\hat{x})\|_{A^{-1}}^2.$$

Combining the above yields the first estimate. Taking $A = I$ to get the second. □

Theorem 3.7. *Assume f satisfies (A1), and $\phi \in C^{2,1}(V)$ with ϕ being 1-strongly convex. When $(x, \hat{x}, y, \hat{y}) \in B(x^*, R)$, (A2) holds with*

$$C_{f,\phi} \le (L - \mu)\big(1 + 2L_{\nabla^2\phi}R\big),$$

where $L_{\nabla^2\phi}$ is the Lipschitz constant of $\nabla^2\phi$ restricted to $B(x^*, R)$.

Proof. Take $A = \nabla^2\phi(\xi_y)$ for some point ξ_y between y and $\hat{y}$ such that $D_\phi(y, \hat{y}) = \frac{1}{2}\|y - \hat{y}\|_A^2$. As ϕ is 1-strongly convex, $\|e\| \le \|e\|_A$ for any $e \in V$. Let ξ_x be a point between x and $\hat{x}$ satisfying $\langle \nabla\phi(x) - \nabla\phi(\hat{x}),\, x - \hat{x}\rangle = \|x - \hat{x}\|_{\nabla^2\phi(\xi_x)}^2$. By assumption $\xi_x, \xi_y \in B(x^*, R)$ and

$$\big|\|e\|_{\nabla^2\phi(\xi_x)}^2 - \|e\|_{\nabla^2\phi(\xi_y)}^2\big| \le L_{\nabla^2\phi}\|\xi_x - \xi_y\|\|e\|^2 \le 2L_{\nabla^2\phi}R\|e\|^2 \quad \forall e \in B(x^*, R).$$

We use the relative smoothness to get

$$\begin{aligned}|\langle \nabla f_{-\mu}(x) - \nabla f_{-\mu}(\hat{x}),\, x - \hat{x}\rangle| &\le (L-\mu)\|x - \hat{x}\|_{\nabla^2\phi(\xi_x)}^2 \\ &\le (L-\mu)\left(\|x - \hat{x}\|_{\nabla^2\phi(\xi_y)}^2 + L_{\nabla^2\phi}\|\xi_x - \xi_y\|\|x - \hat{x}\|^2\right) \\ &\le (L-\mu)(1 + 2L_{\nabla^2\phi}R)\|x - \hat{x}\|_A^2,\end{aligned}$$

which implies $L_{f_{-\mu}}(A) \le (L-\mu)(1 + 2L_{\nabla^2\phi}R)$. Applying Theorem 3.5 to get the desired estimate. □

B Extension to convex optimization

For convex functions, i.e, $\mu = 0$, we consider the perturbed system

$$\begin{cases} x' = y - x, \\ (\nabla\phi(y))' = \epsilon^{-1}\big[\epsilon(\nabla\phi(x) - \nabla\phi(y)) - \nabla f(x)\big], \end{cases} \tag{30}$$

for some fixed $\epsilon > 0$. With the dual variables, the perturbed flow is equivalent to

$$x' = y - x, \quad \eta' = -\epsilon^{-1}\nabla f(x) + \chi - \eta.$$

Consider the Lyapunov function

$$\mathcal{E}(x, \eta; \epsilon) := D_f(x, x^*) + \epsilon D_{\phi^*}(\eta, \chi^*). \tag{31}$$

We treat ϵ as a fixed parameter and take derivative with respect to x and η.

Lemma B.1. *Let $\mathcal{E}$ be the Lyapunov function* (31)*, and let $\mathcal{G}$ be the vector field of the perturbed acclerated mirror descent flow* (30)*. The following perturbed strong Lyapunov property holds:*

$$-\nabla\mathcal{E}(x,\eta;\epsilon)\cdot\mathcal{G}(x,\eta)=\mathcal{E}(x,\eta;\epsilon)+D_f(x^*,x)+\epsilon D_{\phi^*}(\chi,\eta)-\epsilon D_{\phi^*}(\chi,\chi^*). \tag{32}$$

Proof. By direct calculations,

$$-\nabla\mathcal{E}(x,\eta;\epsilon)\cdot\mathcal{G}(x,\eta)=\langle\nabla f(x)-\nabla f(x^*),x-x^*\rangle+\epsilon(\nabla\phi^*(\eta)-\nabla\phi^*(\chi^*),\eta-\chi).$$

Using the three-point identity of the Bregaman divergence, the additional cross term expands as:

$$\epsilon\langle\nabla\phi^*(\eta)-\nabla\phi^*(\chi^*),\eta-\chi\rangle=\epsilon\left(D_{\phi^*}(\eta,\chi^*)+D_{\phi^*}(\chi,\eta)-D_{\phi^*}(\chi,\chi^*)\right).$$

Therefore,

$$-\nabla\mathcal{E}(x,\eta;\epsilon)\cdot\mathcal{G}(x,\eta)=\mathcal{E}(x,y;\epsilon)+D_f(x^*,x)+\epsilon D_{\phi^*}(\chi,\eta)-\epsilon D_{\phi^*}(\chi,\chi^*).$$

□

Consider the perturbed AOR accelerated mirror descent method:

$$\frac{x_{k+1}-x_k}{\alpha}=y_k-x_{k+1}, \tag{33a}$$

$$\frac{\nabla\phi(y_{k+1})-\nabla\phi(y_k)}{\alpha}=-\epsilon^{-1}(2\nabla f(x_{k+1})-\nabla f(x_k))+\nabla\phi(x_{k+1})-\nabla\phi(y_{k+1}). \tag{33b}$$

An equivalent but computation favorable form is given as

$$\begin{aligned}x_{k+1}&=\frac{1}{1+\alpha}(x_k+\alpha y_k),\\ y_{k+1}&=\arg\min_{y\in\mathbb{R}^n}(1+\alpha)\,\phi(y)-\left\langle\alpha\nabla\phi(x_{k+1})+\nabla\phi(y_k)-\frac{\alpha}{\epsilon}(2\nabla f(x_{k+1})-\nabla f(x_k)),\,y\right\rangle.\end{aligned} \tag{34}$$

Introduce the dual variables $\eta_k=\nabla\phi(y_k),\chi_k=\nabla\phi(x_k)$, $k=0,1,2,\cdots$ to rewrite (33b) as

$$\frac{\eta_{k+1}-\eta_k}{\alpha}=-\frac{1}{\epsilon}(2\nabla f(x_{k+1})-\nabla f(x_k))+\chi_{k+1}-\eta_{k+1}. \tag{35}$$

For four points $x,\hat{x},y,\hat{y}$, introduce the term

$$\mathcal{B}^\alpha(x,\hat{x},\hat{y},y;\epsilon)=D_f(x,\hat{x})+\epsilon D_\phi(\hat{y},y)+\alpha\left\langle\nabla f(x)-\nabla f(\hat{x}),y-\hat{y}\right\rangle. \tag{36}$$

Define the modified Lyapunov function

$$\mathcal{E}^\alpha(x,y;\epsilon):=\mathcal{B}^\alpha(x,x^*,x^*,y;\epsilon):=\mathcal{E}(x,\eta;\epsilon)+\alpha\left\langle\nabla f(x)-\nabla f(x^*),y-x^*\right\rangle. \tag{37}$$

For simplicity, we shall use $\mathcal{B}^\alpha(x,\hat{x},\hat{y},y)$ and $\mathcal{E}^\alpha(x,y)$ when there is no confusion to skip ϵ.

Lemma B.2. *Let (x_k,y_k) be the sequence generated by perturbed Acc-MD (33), then it holds*

$$\begin{aligned}\mathcal{E}^\alpha(x_{k+1},y_{k+1})-\mathcal{E}^\alpha(x_k,y_k)\le\;&-\alpha\mathcal{E}^\alpha(x_{k+1},y_{k+1})-\alpha D_f(x^*,x_{k+1})\\&+\alpha\left\langle\nabla f(x_{k+1})-\nabla f(x^*),y_{k+1}-x^*\right\rangle\\&-\mathcal{B}^\alpha(x_k,x_{k+1},y_{k+1},y_k)+\alpha\epsilon D_\phi(x^*,x_{k+1})\end{aligned} \tag{38}$$

Proof. Denote $z=(x,\eta)$. Expand the difference $\mathcal{E}(x,\eta)$ at (x_{k+1},η_{k+1}),

$$\begin{aligned}\mathcal{E}(x_{k+1},\eta_{k+1})-\mathcal{E}(x_k,\eta_k)=&\;\langle\nabla\mathcal{E}(\boldsymbol{z}_{k+1}),\boldsymbol{z}_{k+1}-\boldsymbol{z}_k\rangle-D_{\mathcal{E}}(\boldsymbol{z}_k,\boldsymbol{z}_{k+1})\\=&\;\alpha\left\langle\nabla\mathcal{E}(\boldsymbol{z}_{k+1};\epsilon),\mathcal{G}(z_{k+1})\right\rangle-D_{\mathcal{E}}(z_k,z_{k+1})\\&-\alpha\left\langle\nabla f(x_{k+1})-\nabla f(x^*),y_{k+1}-y_k\right\rangle\\&-\alpha\left\langle y_{k+1}-x^*,\nabla f(x_{k+1})-\nabla f(x_k)\right\rangle\end{aligned}$$

We write the cross term as

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

Then use identity (32) to expand $\langle \nabla \mathcal{E}(\boldsymbol{z}_{k+1}), \mathcal{G}(\boldsymbol{z}_{k+1})\rangle$ we get

$$\begin{aligned} \mathcal{E}^\alpha(x_{k+1},y_{k+1}) - \mathcal{E}^\alpha(x_k,y_k) = {}& -\alpha\mathcal{E}(x_{k+1},\eta_{k+1}) - \alpha D_f(x^*,x_{k+1}) \\ & -\alpha\epsilon D_\phi(y_{k+1},x_{k+1}) + \alpha\epsilon D_\phi(x^*,x_{k+1})\\ & - D_{\mathcal{E}}(\boldsymbol{z}_k,\boldsymbol{z}_{k+1}) - \alpha\left\langle\nabla f(x_{k+1})-\nabla f(x_k), y_{k+1}-y_k\right\rangle \\ ={}& -\alpha\mathcal{E}^\alpha(x_{k+1},y_{k+1}) - \alpha D_f(x^*,x_{k+1})\\ & -\alpha\epsilon D_\phi(y_{k+1},x_{k+1}) + \alpha\epsilon D_\phi(x^*,x_{k+1}) \\ & +\alpha\left\langle \nabla f(x_{k+1})-\nabla f(x^*), y_{k+1}-x^*\right\rangle \\ & -\mathcal{B}^\alpha(x_k,x_{k+1},y_k,y_{k+1}),\end{aligned}$$

where noted that $D_{\phi^*}(\chi_{k+1},\chi^*) = D_\phi(x^*,x_{k+1})$. Drop the negative terms and rearrangement we get the desired result. □

Theorem B.3 (Convergence of perturbed Acc-MD method)**.** *Suppose f is convex and the relative Cauchy-Schwarz inequality (A2) holds with constant $C_{f,\phi}$. For any $\epsilon > 0$, let (x_k, y_k) be generated by scheme (33) with initial value (x_0, y_0), $\eta_k = \nabla\phi(y_k)$, and step size $0 < \alpha = \sqrt{\epsilon/C_{f,\phi}}$. Assume there exists $R > 0$ such that*

$$D_\phi(x^*,x_k) \le R/2, \quad D_\phi(x^*,y_k) \le R/2, \quad \forall k \ge 0. \tag{39}$$

Then we have the linear convergence with perturbation

$$\mathcal{E}^\alpha(x_k,y_k,\epsilon) \le \left(\frac{1}{1+\sqrt{\epsilon/C_{f,\phi}}}\right)^k \mathcal{E}^\alpha(x_0,y_0,\epsilon) + \epsilon R, \quad k \ge 0. \tag{40}$$

Proof. As before,

$$\begin{aligned} \mathcal{E}^\alpha(x_{k+1},y_{k+1}) - \mathcal{E}^\alpha(x_k,y_k) \le {}& -\alpha\mathcal{E}^\alpha(x_{k+1},y_{k+1}) - \alpha D_f(x^*,x_{k+1}) - \alpha\epsilon D_\phi(x^*,y_{k+1}) \\ & +\alpha\left\langle\nabla f(x_{k+1})-\nabla f(x^*), y_{k+1}-x^*\right\rangle \\ & -\mathcal{B}^\alpha(x_k,x_{k+1},y_{k+1},y_k) \\ & +\alpha\epsilon(D_\phi(x^*,x_{k+1}) + \epsilon D_\phi(x^*,y_{k+1})). \end{aligned} \tag{41}$$

For $\alpha = \sqrt{\epsilon/C_{f,\phi}}$, we can drop negative terms from the inequality (41) to get the accelerated linear convergence

$$\begin{aligned} \mathcal{E}^\alpha(x_{k+1},y_{k+1}) &\le \frac{1}{1+\alpha}\mathcal{E}^\alpha(x_k,y_k) + \frac{\alpha\epsilon}{1+\alpha}(D_\phi(x^*,x_{k+1}) + D_\phi(x^*,y_{k+1})) \\ &\le \frac{1}{1+\alpha}\mathcal{E}^\alpha(x_k,y_k,\epsilon) + \frac{\alpha\epsilon}{1+\alpha}R \\ &\le \left(\frac{1}{1+\alpha}\right)^{k+1}\mathcal{E}^\alpha(x_0,y_0,\epsilon) + \frac{\alpha\epsilon R}{1+\alpha}\sum_{i=0}^{k}\frac{1}{(1+\alpha)^i}. \end{aligned}$$

Summing up the geometric series we get the desired result. □

To achieve an accuracy $\mathcal{E}^\alpha(x_k,y_k) = O(\epsilon)$, the number of iterations is bounded by

$$\left(1+\sqrt{\epsilon/C_{f,\phi}}\right)^{-k} = O(\epsilon) \quad \Longrightarrow \quad k = O\left(\sqrt{\frac{C_{f,\phi}}{\epsilon}}|\ln\epsilon|\right).$$

Compared to the dominant complexity $O(\epsilon^{-1/2})$, the logarithmic factor $O(|\ln\epsilon|)$ is negligible. This establishes the nearly optimal complexity of accelerated gradient methods.

Since the perturbation will not change the equilibrium point x^*, we can choose strictly decreasing ϵ_{k+1} and use the homotopy argument to remove the $|\ln\epsilon|$ dependence (Chen et al., 2025, Theorem 8.4). The modified algorithm is summarized in Algorithm 2. Consequently, the method achieves an effective sublinear convergence rate of $\mathcal{O}(1/k^2)$ in terms of gradient evaluations.

Algorithm 2 AccMD with homotopy perturbation.

```
**Parameters: Initital value and tolerance** $(x_0, y_0, \epsilon_0)$ **and termination tolerance** $\epsilon$**.**
Set $k = 0$ and $m_0 = (\sqrt{C_{f,\phi}} + \sqrt{\epsilon_0}) \ln(2(R+1))\epsilon_0^{-1/2}$
**while** $\epsilon_k > \epsilon$ **do**
    $\epsilon_{k+1} = \epsilon_k/2, \quad m_{k+1} = \sqrt{2}\, m_k$
    Apply perturbed AccMD scheme (34) with the initial value $(x_k, y_k)$, the parameter $\epsilon_{k+1}$ and the step size $\alpha = \sqrt{\epsilon_{k+1}/C_{f,\phi}}$ for $m_{k+1}$ iterations to get $(x_{k+1}, y_{k+1})$
    $k = k + 1$
**end while**
**return** $(x_k, y_k)$
```

C Extension to composite optimization

onsider the composite optimization problem

$$\min_{x \in \mathbb{R}^n} \quad F(x) := f(x) + g(x), \tag{42}$$

where f satisfies the relative smoothness and convexity condition, and g is convex but possibly non-smooth, with a well-defined generalized proximal operator. We propose the accelerated mirror descent flow for composite case

$$\begin{cases} x' = y - x, \\ -\mu\left(\nabla\phi(y)\right)' + \mu(\nabla\phi(x) - \nabla\phi(y)) - \nabla f(x) \in \partial g(y), \end{cases} \tag{43}$$

with initial conditions $x(0) = x_0, y(0) = y_0$. Introducing dual variables $\chi = \nabla\phi(x)$ and $\eta = \nabla\phi(y)$, the flow (43) becomes

$$x' = y - x, \quad \eta' = -\mu_\phi^{-1}[\nabla f_{-\mu}(x) + q(y)] - \eta,$$

where $f_{-\mu} := f - \mu_\phi \phi$ and $q(y) \in \partial g(y)$ is one sub-gradient of g at y.

We make an analysis analogous to the smooth convex case. Define a Lyapunov function:

$$\mathcal{E}(x, \eta) := D_{f_{-\mu}}(x, x^*) + \mu D_{\phi^*}(\eta, \chi^*), \tag{44}$$

where x^* is a minimizer of $f + g$ and $\chi^* = \nabla\phi(x^*)$.

Lemma C.1. *Let* $\mathcal{E}(x, \eta)$ *be defined by* (44)*, and define the vector field* $\mathcal{G}(x, \eta) = (y - x,\ -\mu_\phi^{-1}[\nabla f_{-\mu}(x) + q(y)] - \eta)$*, where* $q(y) \in \partial g(y)$*, and the dual variables satisfy* $\chi = \nabla\phi(x)$ *and* $\eta = \nabla\phi(y)$*. Then*

$$-\nabla\mathcal{E}(x, \eta) \cdot \mathcal{G}(x, \eta) \geq \mathcal{E}(x, \eta) + D_{f_{-\mu}}(x^*, x) + \mu_\phi D_\phi(y, x^*). \tag{45}$$

As a consequence, any solution $(x(t), y(t))$ *of the flow* (43) *satisfies the exponential decay bound*

$$\mathcal{E}(x(t), \eta(t)) \leq e^{-t}\mathcal{E}(x(0), \eta(0)). \tag{46}$$

Proof. First, observe that at minimum x^*, $0 \in \nabla f(x^*) + \partial g(x^*)$, so there is $q(x^*) \in \partial g(x^*)$ that $\nabla f(x^*) + q(x^*) = 0$ holds. From the gradients of the Lyapunov function:

$$\partial_x \mathcal{E} = \nabla f_{-\mu}(x) - \nabla f_{-\mu}(x^*), \quad \partial_\eta \mathcal{E} = \mu_\phi\left(\nabla\phi^*(\eta) - \nabla\phi^*(\chi^*)\right) = \mu_\phi(y - x^*),$$

we compute

$$\begin{aligned}
& -\nabla\mathcal{E}(x, \eta) \cdot \mathcal{G}(x, \eta) \\
&= \langle \nabla f_{-\mu}(x) - \nabla f_{-\mu}(x^*), x - y\rangle + \mu_\phi \left\langle y - x^*,\ \mu_\phi^{-1}(\nabla f_{-\mu}(x) + q(y)) + \eta \right\rangle \\
&= \langle \nabla f_{-\mu}(x) - \nabla f_{-\mu}(x^*), x - x^*\rangle + \mu_\phi \langle \nabla\phi^*(\eta) - \nabla\phi^*(\chi^*), \eta - \chi^*\rangle \\
&= \langle \nabla f_{-\mu}(x) - \nabla f_{-\mu}(x^*), x - x^*\rangle + \langle y - x^*, \nabla f(x^*) + q(y)\rangle + \mu_\phi \langle \nabla\phi^*(\eta) - \nabla\phi^*(\chi^*), \eta - \chi^*\rangle \\
&= D_{f_{-\mu}}(x, x^*) + D_{f_{-\mu}}(x^*, x) + \langle y - x^*, q(y) - q(x^*)\rangle + \mu_\phi D_{\phi^*}(\chi^*, \eta) + \mu_\phi D_{\phi^*}(\eta, \chi^*) \\
&\geq \mathcal{E}(x, \eta) + D_{f_{-\mu}}(x^*, x) + \mu_\phi D_\phi(y, x^*),
\end{aligned}$$

where the last inequality follows from convexity of g as $\langle y - x^*, q(y) - q(x^*)\rangle \geq 0$.

Since $D_{f_{-\mu}}(x^*, x) \geq 0$ and $D_\phi(y, x^*) \geq 0$ by convexity of $f_{-\mu}$ and ϕ, we obtain the inequality

$$\nabla\mathcal{E}(x, \eta) \cdot \mathcal{G}(x, \eta) \leq -\mathcal{E}(x, \eta),$$

from which the exponential decay (46) follows by Grönwall's inequality. □

We generalize the Acc-MD scheme for composite case

$$\frac{x_{k+1} - x_k}{\alpha} = 2y_{k+1} - y_k - x_{k+1}, \tag{47a}$$

$$\frac{\nabla\phi(y_{k+1}) - \nabla\phi(y_k)}{\alpha} \in -\frac{1}{\mu}[\nabla f(x_k) + \partial g(y_{k+1})] + \nabla\phi(x_k) - \nabla\phi(y_{k+1}). \tag{47b}$$

An equivalent computation favorable form is given as Algorithm 3.

Algorithm 3 Accelerated mirror descent (Acc-MD) method for composite optimization

```
Parameters: $x_0, y_0 \in \mathbb{R}^n, \mu, C_{f,\phi}$.
Set $\alpha = \sqrt{\mu/C_{f,\phi}}$.
for $k = 0, 1, 2, \dots$ do
    $y_{k+1} = \arg\min_{y\in\mathbb{R}^n} (1+\alpha)\,\phi(y) + \frac{\alpha}{\mu} g(y) - \left\langle \alpha\nabla\phi(x_k) + \nabla\phi(y_k) - \frac{\alpha}{\mu_\phi}\nabla f(x_k),\, y \right\rangle.$
    $x_{k+1} = \frac{1}{1+\alpha}[x_k + \alpha(2y_{k+1} - y_k)].$
end for
```

As the subgradient of g is evaluated implicitly in scheme (47), the convergence analysis becomes identical to the smooth convex case. Combining with the convergence result of ODE flow Lemma C.1, we derive the following theorem for Algorithm 3:

Theorem C.2 (Convergence of Acc-MD method for composite optimization)**.** *Suppose f is μ-relatively convex and relative Cauchy Schwarz inequality holds with constant $C_{f,\phi}$ and g is convex. Let (x_k, y_k) be generated by scheme (47) with initial value (x_0, y_0), $\eta_k = \nabla\phi(y_k)$, and step size $\alpha = \sqrt{\mu/C_{f,\phi}}$. Then there exists a constant $C_0 = C_0(x_0, y_0, \mu, C_{f,\phi})$ so that we have the accelerated linear convergence*

$$D_{f_\mu}(x_{k+1}, x^*) + \mu D_{\phi^*}(\eta_{k+1}, \chi^*) \leq C_0 \left(\frac{1}{1 + \sqrt{\mu/C_{f,\phi}}}\right)^k, \quad k \geq 1.$$

D Additional Numerical Tests

LASSO problem on synthetic data We consider the over-parameterized LASSO problem (28), on large-size synthetic data. The ground truth x^* is assumed to be sparse, supported on 50 randomly selected indices. Each nonzero entry is drawn as $x_i = \eta_i + \mathrm{sgn}(\eta_i)$, with $\eta_i \sim \mathcal{N}(0, 1)$. The design matrix A is Gaussian with column-wise variance scaling: $a_{ij} \sim \mathcal{N}(0, j^2)$. The response vector is generated as $b = Ax^* + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2 I_n)$ with $\sigma = 1$. We set $\lambda = 1$ throughout the experiments. The mirror function and parameter discussion remain the same as in Section 4. We consider two settings: (1) sample size $n = 150$, dimension $d = 200$; (2) sample size $n = 1000$, dimension $d = 2000$.

As shown in Figure 5 and Figure 6, under both settings, the proposed Acc-MD method converges significantly faster than the competing approaches.

Quadratic on the simplex We consider the constrained optimization problem over the $(n-1)$-dimensional simplex $\Delta_n = \{x \in \mathbb{R}^n \mid \sum_{i=1}^n x_i = 1,\ x_i \geq 0\}$:

$$\min_{x\in\Delta_n} f(x) := \frac{1}{2}\|Ax - b\|^2.$$

This can be viewed as a composite optimization problem, where g is the indicator function of Δ_n.

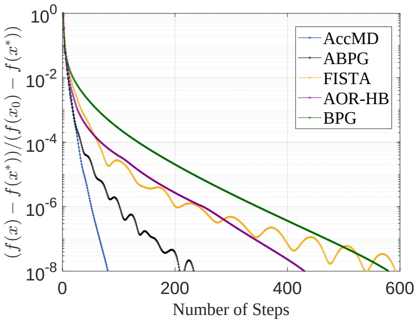

Figure 5: Function error curve of LASSO problem, $n = 150$, $d = 200$.

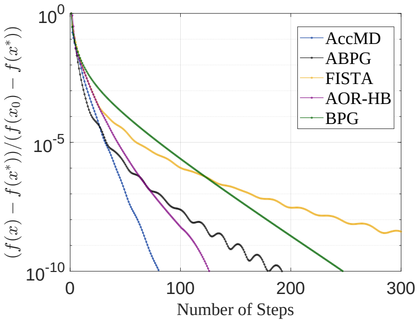

Figure 6: Function error curve of LASSO problem, $n = 1000$, $d = 2000$.

Let $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ be a Gaussian matrix with $a_{ij} \sim \mathcal{N}(0, j^2)$, and set the ground truth to $x^* = (1, \dots, 1)^\top/n$ with $b = Ax^*$. We use the mirror function $\phi(x) = \frac{1}{2}x^\top Dx$, where $D = \operatorname{diag}(A^\top A)$ is strictly positive definite. The mirror maps $\nabla\phi$ and $\nabla\phi^*$ are then simple to evaluate, and the relative smoothness and convexity constants L, μ can be estimated analogously. As shown in Table 2, L remains nearly constant across problem sizes, and the relative condition number L/μ is significantly better than the Euclidean counterpart L_f/μ_f.

The Acc-MD subproblem (27) becomes

$$x_{k+1} \in \arg\min_{x\in\Delta_n} \|x - z_{k+1}\|_D^2, \qquad \text{with } z_{k+1} = D^{-1}\tilde{z}_{k+1}, \tag{48}$$

where $\tilde{z}_{k+1} = \frac{1}{1+\alpha}\left[\nabla\phi(y_k) + \alpha\nabla\phi(x_k) - \frac{\alpha}{\mu}\nabla f(x_k)\right]$. Problem (48) is a projection onto the simplex under a weighted norm. We solve it by introducing the Lagrangian:

$$\mathcal{L}(x, \lambda, \beta) = \frac{1}{2}\|x - z\|_D^2 - \lambda\left(\sum_{i=1}^{n} x_i - 1\right) - \beta^\top x,$$

with KKT conditions:

$$d_i(x_i - z_i) - \lambda - \beta_i = 0, \quad x_i \geq 0, \quad \beta_i \geq 0, \quad x_i\beta_i = 0, \quad 1 \leq i \leq n, \quad \sum_{i=1}^{n} x_i = 1.$$

The solution satisfies $x_i = \max(0, z_i + \lambda/d_i)$, and λ is the unique solution to

$$\sum_{i=1}^{n} \max\left(0, z_i + \frac{\lambda}{d_i}\right) = 1,$$

which can be efficiently found by the bisection method.

As shown in Table 2, Acc-MD outperforms competing methods in both iteration count and runtime. The non-accelerated BPG method is omitted as it fails to converge within a reasonable time.

Table 2: Quadratic optimization on the simplex. Stopping criterion: $f(x) < 10^{-12}f(x_0)$.

Problem Size					Acc-MD		ABPG		FISTA		AOR-HB	
n	L_f	μ_f	L	μ	#Iter	Time	#Iter	Time	#Iter	Time	#Iter	Time
125	4.04×10^6	0.6228	3.78	9.86×10^{-5}	1159	0.16	5225	0.47	39145	1.14	22930	0.53
250	3.17×10^7	0.0146	3.96	1.99×10^{-7}	1669	0.38	6602	1.20	81633	5.25	24133	1.68
500	2.46×10^8	0.1176	3.97	3.79×10^{-7}	4439	1.96	6183	3.18	102666	24.57	67621	14.49
1000	2.00×10^9	0.0592	3.95	1.34×10^{-7}	4960	6.07	6221	10.26	118364	177.83	100283	129.71

LLM USAGE

In preparing this manuscript, large language models (LLMs) were employed exclusively to assist with language-related tasks, such as improving readability, grammar, and style. The models were not used for research ideation, development of methods, data analysis, or interpretation of results. All scientific content, including problem formulation, theoretical analysis, and experimental validation, was conceived, executed, and verified entirely by the authors. The authors bear full responsibility for the accuracy and integrity of the manuscript.

Ethics statement

This work is purely theoretical and algorithmic, focusing on convex optimization methods. It does not involve human subjects, sensitive data, or applications that raise ethical concerns related to privacy, security, fairness, or potential harm. All experiments are based on publicly available datasets or synthetic data generated by standard procedures. The authors believe that this work fully adheres to the ICLR Code of Ethics.

Reproducibility statement

We have taken several measures to ensure the reproducibility of our results. All theoretical assumptions are explicitly stated, and complete proofs are provided in the appendix. For the experimental evaluation, we describe the setup, parameter choices, and baselines in detail in the main text. The source code for our algorithms and experiments are available as supplementary materials. Together, these resources should allow others to reproduce and verify our theoretical and empirical findings.