Convergence of Adaptive Stochastic Mirror Descent

Ting Hu, Xiaotong Liu, Kai Ji, and Yunwen Lei

***Abstract*—In this paper, we present a family of adaptive stochastic optimization methods, which are associated with mirror maps that are widely used to capture the geometry properties of optimization problems during iteration processes. The well-known adaptive moment estimation (Adam) type algorithm falls into the family when the mirror maps take the form of temporal adaptation. In the context of convex objective functions, we show that with proper step sizes and hyperparameters, the average regret can achieve the convergence rate $O(T^{-\frac{1}{2}})$ after T iterations under some standard assumptions. We further improve it to $O(T^{-1}\log T)$ when the objective functions are strongly convex. In the context of smooth objective functions (not necessarily convex), based on properties of the strongly convex differentiable mirror map, our algorithms achieve convergence rates of order $O(T^{-\frac{1}{2}})$ up to a logarithmic term, requiring large or increasing hyperparameters that are coincident with practical usage of Adam-type algorithms. Thus, our work gives explanations for the selection of the hyperparameters in Adam-type algorithms' implementation.**

***Index Terms*—Adam, mirror descent, non-convex stochastic optimization, convergence analysis.**

I. Introduction

WIth the development of deep neural networks and reinforcement learning, adaptive gradient stochastic methods have been constantly drawing attention in the fields of science and engineering. Adaptive methods such as AdaGrad [21], RMSProp [47], AMSGrad [46], and AdamW [42] are widely used to train modern deep neural networks including GANs [1], GPTs [2] and ViTs [3]. It is often known that adaptive gradient methods converge considerably faster than vanilla Stochastic Gradient Descent (SGD) for training certain models like transformers, as demonstrated in [4], [5]. Similar phenomena are also reported in large-scale convex and non-convex optimization problems including natural language processing (NLP), generative adversarial networks (GANs) and computer vision [13], [18], [28]. Unlike vanilla SGD, which moves forward/backward at uniform learning rates in all directions, adaptive methods adjust the step sizes in each coordinate direction by using the information of past gradients to scale down gradients. Thus, adaptive methods achieve a rapid convergence rate and a robust training process compared to standard stochastic gradient methods when the gradients are sparse or small. Due to the empirical successes of adaptive gradient stochastic methods, a significant amount of literature has recently emerged to investigate their theoretical foundations and convergence guarantees. See [6], [7], [33], [40], [41], [50], [54].

Mirror descent is a classical method for computation and optimization, which was originally introduced in the pioneering work [10] for solving convex optimization problems. By an appropriate choice of the mirror map, the geometry of optimization models can be captured when their estimators of interest lie in a non-Hilbert space, e.g., a Banach space with the ℓ_1-norm is preferred for problems with sparse solutions.

Given $(\Omega, \|\cdot\|)$, we define the dual space $(\Omega^*, \|\cdot\|_*)$, which is a linear bounded operator acting on Ω, endowed with the norm $\|\mathbf{w}^*\|_* = \sup_{\|\mathbf{w}\|\le 1} |\langle \mathbf{w}^*, \mathbf{w}\rangle|$ for any $\mathbf{w}^* \in \Omega^*$. A mirror map Ψ is said to be *Fréchet differentiable*, meaning that at each $\mathbf{x} \in \Omega$, the gradient $\nabla\Psi(\mathbf{x})$ lies in Ω^*. A map Ψ is *σ-strongly convex* with respect to the norm $\|\cdot\|$ for some $\sigma > 0$, satisfying

$$\begin{aligned} D_\Psi(\mathbf{x},\mathbf{y}) : &= \Psi(\mathbf{x}) - \Psi(\mathbf{y}) - \langle \nabla\Psi(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \\ &\ge \sigma/2\|\mathbf{x}-\mathbf{y}\|^2, \ \forall \mathbf{x},\mathbf{y} \in \Omega, \end{aligned} \tag{1}$$

where $\langle \nabla\Psi(\mathbf{y}), \cdot\rangle$ denotes the linear functional acting on $(\Omega, \|\cdot\|)$. The quantity $D_\Psi(\mathbf{x},\mathbf{y})$ induced by Ψ is defined as the *Bregman distance* between $\mathbf{x}$ and $\mathbf{y}$.

Given a differentiable objective function $f : \Omega \to \mathbb{R}$ and the stepsize sequence $\{\eta_t\}_t$, mirror descent approximates the minimizer of f by starting with $\mathbf{w}_1 \in dom(\nabla\Psi)$ and generating iterates $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ via

$$\nabla\Psi(\mathbf{w}_{t+1}) - \nabla\Psi(\mathbf{w}_t) = -\eta_t \nabla f(\mathbf{w}_t), \tag{2}$$

where the gradient descent is performed in the dual space $(\Omega^*, \|\cdot\|_*)$. When $\Omega = \mathbb{R}^d$, $\|\cdot\| = \|\cdot\|_{\ell^2} = \|\cdot\|_*$ and the map $\Psi(\cdot) = \Psi^*(\cdot) = \frac{1}{2}\|\cdot\|^2_{\ell^2}$, the mirror descent algorithm recovers the standard gradient descent.

In many machine learning algorithms, mirror descent is a preferred learning paradigm since the mirror map can be chosen to adapt to different types of algorithms by considering the geometry of the optimization problems. A classical example is reconstructing 3D medical images from Positron Emission Tomography (PET), where the optimization problem is defined on the unit simplex $\Omega = \{\mathbf{w} \in \mathbb{R}^d_+ : \|\mathbf{w}\|_{\ell^1} = 1\}$, rather than the structure of Hilbert distance $\|\cdot\|_{\ell^2}$. In the work by [15], the authors used the Bregman distance $D_{\Psi^{(KL)}}$ to replace the distance $\|\cdot\|_{\ell^2}$, which is induced by $\Psi^{(KL)}(\mathbf{w}) = \sum_{j=1}^{d} w_j \ln w_j$, with $\Omega = \{\mathbf{w} \in \mathbb{R}^d_+ : \|\mathbf{w}\|_{\ell^1} = 1\}$, $\|\cdot\| = \|\cdot\|_{\ell^1}$. Another example is policy optimization in reinforcement learning (RL). For a

The work is supported partially by the National Natural Science Foundation of China under Grants 12071356 and 12471095, partially by the Research Grants Council of Hong Kong under Grant 22303723. (Corresponding author: Yunwen Lei).

Ting Hu is with School of Management, Xi'an Jiaotong University, Xi'an, China (e-mail: tinghu@xjtu.edu.cn).

Xiaotong Liu is with School of Management, Xi'an Jiaotong University, Xi'an, China (e-mail: ariesoomoon@gmail.com).

Kai Ji is with Huawei Technologies Co., Ltd., Shenzhen, China (e-mail: jikai9@huawei.com).

Yunwen Lei is with Department of Mathematics, The University of Hong Kong, Hongkong, China (e-mail: leiyw@hku.hk).

discounted Markov decision process, the goal of RL is to find a policy π by maximizing

$$V_p(\pi) := \mathbf{E}\Big[\sum_{h=0}^{\infty}\gamma^h r_h\Big], \pi \in \Omega \tag{3}$$

where $\gamma \in [0,1]$ is the discount factor, r_h is the reward and Ω is the probability simplex supported on the state-action pair (s,a) satisfying $\Omega := \{\sum_s \pi_{s,a} = 1, \pi_{s,a} \geq 0 \text{ for all } a\}$. Thus, mirror descent with $\Psi^{(KL)}(\pi) = \sum_{s,a} \pi_{s,a} \ln \pi_{s,a}$ is employed and a sequence of promising policies π_t is generated by mirror descent. For details, we refer the readers to the recent work in [29]. Compressed sensing is also a successful example of mirror descent, where a solution of a linear equation $A\mathbf{w} = f$ has minimal ℓ_1-norm $\|\mathbf{w}\|_{\ell^1}$ for sparsity. By using the mirror map $\Psi^{(\epsilon)}$ (to be stated in Example 1), a linearized Bregman iteration was proposed in [16], which is efficient and robust to noise. Besides, mirror descent is also successfully applied to distributed learning, manifold learning, differential privacy and so on. For more applications of mirror descent, please see [15], [16], [23], [21], [37], [49]. Some commonly used mirror maps will be listed in Example 1 in Subsection II-B.

Due to the fruitful applications of mirror descent, it has received considerable study in the literature on more learning tasks [12], [22], [24], [30], [38]. Now most studies of mirror descent focus on the convex setting. However, the existing results for mirror descent in nonconvex situations are far from mature, as they are just applied to some special mirror maps or objective functions [26], [29].

In this paper, considering the optimization problems' structures and fast training speed in the implementation, we integrate adaptive gradient methods into mirror descent and present a family of adaptive mirror descent (named AdMD), which will be described in the Section II. The main implications of our results are summarized as follows.

(1) AdMD is a new family of adaptive gradient methods with general mirror maps, which can be used to capture various geometric properties of learning problems. In our work, vanilla Adam and other adaptive gradient methods are essentially mirror descent with temporal adaptation of mirror maps. In the framework of online learning that processes samples sequentially [53], we prove that the convergence of AdMD is theoretically guaranteed in the convex setting and demonstrate an explicit convergence rate $O(T^{-\frac{1}{2}})$ with a decreasing $\beta_{1,t}$ and increasing or large $\beta_{2,t}$. Furthermore, we improve it to $O(T^{-1}\log T)$ in the strongly convex situation.

(2) In the L-smooth setting (not necessarily convex), our work refines the selection strategy of hyperparameters to guarantee convergence, where hyperparameters are chosen to be large or increasing, making them more consistent with the implementation in practice. In a lot of previous theoretical work on Adam, the hyperparameters to ensure the convergence were usually set to small value or a decreasing sequence, not guaranteed to be feasible in the practical design of Adam or its variants. Thus, our work also contributes to a better understanding of the convergence of Adam and its variants.

Outline. In Section II, after introducing some notations, we present the AdMD algorithm and then provide the convergence analysis of AdMD in the convex setting. The convergence analysis of AdMD in the smooth setting is presented in Section III. In Section IV we discuss the most relevant literature and make comparisons with our results. Simulations are performed in Section V to illustrate our theory. We make some further discussions and conclusions in Section VI. In the appendix, some useful lemmas are given and the proofs of the main convergence results for AdMD are provided.

II. Adaptive Stochastic Mirror Descent

Notations. To introduce our adaptive gradient stochastic methods, we establish the necessary notations used in the paper. Let $d \in \mathbb{N}$ be the dimension of the optimization problem and we use $[d]$ to denote the set $\{1, \cdots, d\}$. Scalars are denoted by lowercase italic letters such as x, y. Vectors are denoted by lowercase boldface letters, such as $\mathbf{x} \in \mathbb{R}^d$. A sequence of vectors is denoted by subscripts, i.e., $\mathbf{x}_1, \cdots, \mathbf{x}_t, \mathbf{x}_{t+1}, \cdots \in \mathbb{R}^d$. An additional subscript is used to denote the entries of a vector $\mathbf{x}_t$, meaning $\mathbf{x}_t = (x_{t,j})_{j=1}^d \in \mathbb{R}^d$. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\mathbf{x} \otimes \mathbf{y}$ denotes Hadamard Product, the element-wise multiplication of two vectors. Let ℓ^q denote $(\mathbb{R}^d, \|\cdot\|_{\ell^q})$ space with the norm $\|\cdot\|_{\ell^q}$ defined by $\|\mathbf{x}\|_{\ell^q} = (\sum_{j=1}^d |x_j|^q)^{\frac{1}{q}} < \infty$ for any $\mathbf{x} \in \mathbb{R}^d, q > 0$. $a = O(T)$ denotes $a \leq CT$ for some universal constant C (independent of T).

A. Adaptive stochastic mirror descent

In the paper, we study stochastic optimization problems of the form

$$\min_{\mathbf{w}} \Big\{ f(\mathbf{w}) := \mathbb{E}_{\xi \sim \mathbb{P}}\left[f(\mathbf{w};\xi)\right] \Big\}, \tag{4}$$

where the random variable ξ is drawn from some unknown probability distribution $\mathbb{P}$, the function f is differentiable with respect to $\mathbf{w}$ (or the left derivative of $f(\mathbf{w};\xi)$ at $\mathbf{w}$ exists). Concrete examples include the hinge loss $f(\mathbf{w};\xi) = \max\{1 - \mathbf{w}^T\xi, 0\}$, the logistic loss $f(\mathbf{w};\xi) = \log\left(1 + \exp(-\mathbf{w}^T\xi)\right)$, the least squares loss $f(\mathbf{w};\xi) = (1 - \mathbf{w}^T\xi)^2$, and the ϵ-insensitive loss $f(\mathbf{w};\xi) = \max\{|1 - \mathbf{w}^T\xi| - \epsilon, 0\}$. Let $\nabla f(\mathbf{w};\xi)$ denote the stochastic gradient of $f(\mathbf{w};\xi)$ depending on the random variable ξ and assume $\nabla f(\mathbf{w};\xi)$ is an unbiased estimate of $\nabla f(\mathbf{w})$, i.e., $\mathbb{E}_{\xi\sim\mathbb{P}}\nabla f(\mathbf{w};\xi) = \nabla f(\mathbf{w})$ for any $\mathbf{w} \in \Omega$. The general form of adaptive stochastic algorithm for solving the optimization problem (4) is given by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t}}, \quad t \in \mathbb{N}, \tag{5}$$

where $\{\eta_t\}$ is the base stepsizes, $\mathbf{g}_t := \nabla f(\mathbf{w}_t;\xi_t)$ stands for the stochastic gradient arriving at the t-th iteration, $\mathbf{m}_t := \mathbf{m}_t(\mathbf{g}_1, \cdots, \mathbf{g}_t)$ is a function for the historical gradient estimation, $\mathbf{v}_t := \mathbf{v}_t(\mathbf{g}_1, \cdots, \mathbf{g}_t)$ is the scaling function to produce the adaptive stepsizes $\frac{\eta_t}{\sqrt{\mathbf{v}_t}}$. Different choices of $\mathbf{m}_t, \mathbf{v}_t$ lead to different adaptive gradient stochastic algorithms. For example, when $\mathbf{m}_t = \mathbf{g}_t, \mathbf{v}_t = \sum_{j=1}^t \mathbf{g}_j^2 := \sum_{j=1}^t \mathbf{g}_j \otimes \mathbf{g}_j$, it is referred to as AdaGrad [21]. Combining the characteristic of adaptive gradients (5) with mirror descent (2), we present a family of adaptive gradient stochastic algorithms with general mirror maps beyond $\Psi(\mathbf{w})=\frac{1}{2}\|\mathbf{w}\|_{\ell^2}^2$, termed AdMD for short, where the updates satisfy

$$\nabla\Psi(\mathbf{w}_{t+1})=\nabla\Psi(\mathbf{w}_t)-\eta_t\frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t}}.$$

In this work, we are interested in the Adam-style algorithms, where $\mathbf{m}_t$ and $\mathbf{v}_t$ are defined as the exponential moving averages of $\mathbf{g}_t$ and $\mathbf{g}_t^2$. Given the initial $\mathbf{m}_0,\mathbf{v}_0\in\mathbb{R}^d$, for $t\in\mathbb{N}$,

$$\mathbf{m}_t=\beta_{1,t}\mathbf{m}_{t-1}+(1-\beta_{1,t})\mathbf{g}_t,\ \ \mathbf{v}_t=\beta_{2,t}\mathbf{v}_{t-1}+(1-\beta_{2,t})\mathbf{g}_t^2$$

where $\beta_{1,t},\beta_{2,t}\in(0,1)$ are the hyperparameters, adjusting the exponential decay rates for $\mathbf{m}_t$ and $\mathbf{v}_t$, respectively. Then, AdMD in the Adam-style is executed in Algorithm 1.

```
Algorithm 1: AdMD
Input: the iteration number $T$, stepsizes $\{\eta_t>0\}_{t=1}^T$,
       the hyperparameters $\beta_{1,t},\beta_{2,t}\in(0,1)$, initial
       value $\mathbf{w}_1\in dom(\nabla\Phi)$
Output: $\mathbf{w}_1,\cdots,\mathbf{w}_{T+1}$
for $t=1,\cdots,T$ do
    $\mathbf{g}_t=\nabla f(\mathbf{w}_t;\xi_t)$;
    $\mathbf{m}_t=\beta_{1,t}\mathbf{m}_{t-1}+(1-\beta_{1,t})\mathbf{g}_t$;
    $\mathbf{v}_t=\beta_{2,t}\mathbf{v}_{t-1}+(1-\beta_{2,t})\mathbf{g}_t^2$;
    $\nabla\Psi(\mathbf{w}_{t+1})=\nabla\Psi(\mathbf{w}_t)-\eta_t\frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t}}$;
end
```

The update rule of AdMD shows that the convergence behavior of AdMD involves three ingredients: mirror maps, the hyperparameters and stepsizes, which are obviously more complicated in comparison with the classical mirror descent or Adam. Vanilla Adam algorithm is obviously a special case of AdMD with $\Psi^{(\ell^2)}(\mathbf{w}):=\frac{1}{2}\|\mathbf{w}\|_{\ell^2}^2$. Thus, our work can promote the research on adaptive stochastic methods from a more widespread point of view. We aim to design provably convergent algorithms and then present theoretical results for AdMD in both convex and nonconvex cases. Some necessary assumptions are given below.

B. Convergence of AdMD in the convex setting

Our first assumption is regarding the properties of mirror maps.

Definition 1: We say that the mirror map $\Psi(\mathbf{w})$ is **coordinatewise** if

$$\Psi(\mathbf{w}):=\sum_{j=1}^d J(w_j),\ \ \mathbf{w}=(w_1,\cdots,w_d)\in\Omega,\tag{6}$$

where $J(\cdot):\mathbb{R}\to\mathbb{R}$ is convex.
We list some of the most popular choices for Ψ of the form (6) which are well documented in the literature, see e.g., [13], [16], [17], [19], [27], [44], [45].

Example 1:

1) (ℓ^2-norm) $\Psi^{(\ell^2)}(\mathbf{w})=\frac{1}{2}\|\mathbf{w}\|_{\ell^2}^2=\frac{1}{2}\sum_{i=1}^d w_i^2$ with $\Omega=\mathbb{R}^d,\|\cdot\|=\|\cdot\|_{\ell^2}$, where $J(u)=\frac{1}{2}u^2$.
2) (KL-divergence) $\Psi^{(KL)}(\mathbf{w})=\sum_{j=1}^d w_j\ln w_j$ with $\Omega=\{\mathbf{w}\in\mathbb{R}_+^d:\|\mathbf{w}\|_{\ell^1}=1\},\|\cdot\|=\|\cdot\|_{\ell^1}$, where $J(u)=u\ln u$.
3) (Itakura-Saito divergence) $\Psi^{(IS)}(\mathbf{w})=-\sum_{j=1}^d\ln w_j$ with $\Omega=\{\mathbf{w}\in\mathbb{R}_+^d:\|\mathbf{w}\|_{\ell^1}=1\},\|\cdot\|=\|\cdot\|_{\ell^1}$, where $J(u)=-\ln u$.
4) (Fermi-Dirac entropy) $\Psi^{(FD)}(\mathbf{w})=\sum_{i=1}^d w_i\ln w_i+(1-w_i)\ln(1-w_i)$ with $\Omega=\{\mathbf{w}\in\mathbb{R}_+^d:w_i\in(0,1),i\in[d]\}$ and $\|\cdot\|=\|\cdot\|_{\ell^2}$, where $J(u)=u\log u+(1-u)\log(1-u)$.
5) (Hellinger) $\Psi^{(He)}(\mathbf{w})=-\sum_{i=1}^d\sqrt{1-w_i^2}$ with $\Omega=\{\mathbf{w}\in\mathbb{R}_+^d:\|\mathbf{w}\|_{\ell^2}\le 1\}$ and $\|\cdot\|=\|\cdot\|_{\ell^2}$, where $J(u)=-\sqrt{1-u^2}$.
6) $\Psi^{(\epsilon)}(\mathbf{w})=F_\epsilon(\mathbf{w})+\frac{1}{2}\|\mathbf{w}\|_{\ell^2}^2$ with $F_\epsilon(\mathbf{u})=\sum_{i=1}^d\tilde{F}_\epsilon(u_i)$,
$$\tilde{F}_\epsilon(u_i)=\begin{cases}\frac{|u_i|^2}{2\epsilon}, & \text{if } |u_i|\le\epsilon,\\ |u_i|-\frac{\epsilon}{2}, & \text{if } |u_i|>\epsilon.\end{cases}$$
Here $\Omega=\mathbb{R}^d,\|\cdot\|=\|\cdot\|_{\ell^2}$ and $J(u)=\tilde{F}_\epsilon(u)+\frac{1}{2}u^2$.

Assumption 1: Given the coefficient vector $\mathbf{c}=(c_1,\cdots,c_d)\in\mathbb{R}_+^d$ and the coordinatewise map $\Psi(\cdot)$ of the form (6), we define a coefficient-based map $\Psi_{(\mathbf{c})}(\cdot):\mathbb{R}^d\to\mathbb{R}_+$ as

$$\Psi_{(\mathbf{c})}(\mathbf{w}):=\sum_{j=1}^d c_jJ(w_j),\ \ \mathbf{w}=(w_1,\cdots,w_d)\in\Omega.$$

We assume that for any $\mathbf{u},\mathbf{v}\in\Omega$, there exists some $\sigma>0$ such that

$$\langle\nabla\Psi_{(\mathbf{c})}(\mathbf{u})-\nabla\Psi_{(\mathbf{c})}(\mathbf{v}),\mathbf{u}-\mathbf{v}\rangle\ge\sigma\|\mathbf{c}^{\frac{1}{2}}(\mathbf{u}-\mathbf{v})\|^2.\tag{7}$$

The mirror maps stated in Example 1 all satisfy Assumption (7) and the proofs will be given in the appendix. It is straightforward to derive

$$D_{\Psi_{(\mathbf{c})}}(\mathbf{u},\mathbf{v})\ge 2^{-1}\sigma\|\mathbf{c}^{\frac{1}{2}}(\mathbf{u}-\mathbf{v})\|^2,\tag{8}$$

where $D_{\Psi_{(\mathbf{c})}}$ is the Bregman distance induced by $\Psi_{(\mathbf{c})}$. The conjugate function of Ψ is defined by $\Psi^*(\mathbf{y})=\max_{\mathbf{x}\in\Omega}\{\langle\mathbf{x},\mathbf{y}\rangle-\Psi(\mathbf{x})\}$. (Note: all subscripts in this section are uniformly changed to j). We also notice that in AdMD, the iterates $\mathbf{w}_t$ are implicitly in $\nabla\Psi(\mathbf{w}_t)$. When Ψ is strongly convex on $\mathcal{X}$, Proposition 3 of [15] tells us that if Ψ^* is finite everywhere and differentiable, one has $\nabla\Psi^*=(\partial\Psi)^{-1}$. Thus, in practice, we usually employ Ψ for which the corresponding Ψ^* is easily computable. For $\Psi^{(\ell^2)}(\mathbf{w})=\dfrac{1}{2}\|\mathbf{w}\|_{\ell^2}^2$, the conjugate function coincides with itself. For $\Psi^{(KL)}(w)=w\ln w$, the conjugate function is $\Psi^*(v)=\exp\{v\}-1$. For $\Psi^{(IS)}(w)=-\ln w$, the conjugate function is $\Psi^*(v)=-\log(-v)-1$. For $\Psi^{(FD)}(w)=w\ln w+(1-w)\ln(1-w)$, the conjugate function is $\Psi^*(v)=\log(1+\exp\{v\})-1$. For $\Psi^{(He)}(w)=-\sqrt{1-w^2}$, the conjugate function is $\Psi^*(v)=\sqrt{1+v^2}$. For more details, see [11], [15].

Assumption 2: Assume that there exists some t_0 such that $\mathbf{v}_{t_*} = (v_{t_*,i})_{i=1}^d$ satisfies

$$\min_{i\in[d]} v_{t_*,i}^{\frac{1}{2}} > c_0, \text{ for some } c_0 > 0. \tag{9}$$

The assumption is not difficult to satisfy during the iteration and if some t_* is given, the value of c_0 is determined by the accumulated influence of previous gradients $\mathbf{g}_1, \cdots, \mathbf{g}_{t_*}$. It ensures that the scaling term $\mathbf{v}_t$ is lower bounded starting from t_* if $\beta_{2,t}$ is a constant close to 1 and non-decreasing. For example, when $\beta_{2,t} \equiv 1 - \frac{1}{T}$ or $\beta_{2,t} = 1 - \frac{1}{t^2}$, by the fact

$$v_{t+1,i} \geq \beta_{2,t} v_{t,i} \geq \beta_{2,t}\beta_{2,t-1} v_{t-1,i} \geq \cdots \geq \prod_{k=t_0}^{t} \beta_{2,k} v_{t_0,i}$$

$$\geq \prod_{k=t_0}^{t} \beta_{2,k} c_0^2, \ \forall t \geq t_0, i \in [d]$$

and with basic calculations, we can check that $\prod_{k=t_0}^t \beta_{2,k} \geq \prod_{k=2}^t \beta_{2,k} > C$ where C is a constant independent of t far from 0. It shows that

$$v_{t+1,i}^{\frac{1}{2}} \geq C^{\frac{1}{2}} c_0, \text{ for any } t \geq t_0, i \in [d], \tag{10}$$

which can be attributed to the fact that the two strategies of $\beta_{2,t}$ assign more weight to the past gradients [46] and this will be discussed after Theorem 1.

So, this assumption essentially guarantees that along each coordinate direction i the scaling factor $v_{t_*,i}^{\frac{1}{2}}$ is bounded from above uniformly, and the value of c_0 increases as t_* grows. In our convergence rates, the value of c_0 only appears in the constant before the convergence order $O(\cdot)$, which does not affect the order of convergence. Thus, we can select an integer t_* such that the related c_0 is not too small. Meanwhile, in executing adaptive algorithms, the scaling factor $v_{t_*,i}^{\frac{1}{2}}$ is usually replaced with $(v_{t_*,i} + \varepsilon)^{\frac{1}{2}}$ where ε is a constant to enhance the stability of AdMD. Thus, the scaling factor $(v_{t_*,i} + \varepsilon)^{\frac{1}{2}}$ is at least bounded from above by ε. In the recent work [33] the selection of ε appears to be critical for the performance of the algorithm. This is beyond the scope of this paper and hence will be left for our future research.

In the paper, we consider the inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^d x_i y_i$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. We present the convergence result for AdMD in the convex setting, whose proof will be given in Appendix.

Theorem 1: Let $\{\mathbf{w}_t\}$ be generated by AdMD. Suppose $\mathbf{w}^* \in \Omega$, $\sup_t \|\mathbf{w}_t - \mathbf{w}^*\|_\infty < D$ for some D, and assumption (8) holds. Let the base stepsizes η_t and $\beta_{1,t}$ be non-increasing with the iteration time t, then for any $t_* \leq T$,

$$\begin{aligned}
&\frac{1}{T - t_*} \sum_{t=t_*+1}^{T} \mathbb{E}\left[f(\mathbf{w}_t) - f(\mathbf{w}^*)\right] \\
&\leq \frac{1}{T - t_*} \mathbb{E}\Bigg[\sum_{t=t_*+1}^{T} \Bigg(\sum_{j=1}^{d} \left| \eta_{t+1}^{-1} v_{t+1,j}^{\frac{1}{2}} - \eta_t^{-1} v_{t,j}^{\frac{1}{2}} \right| D' \\
&\quad + \beta_{1,t} \|\mathbf{m}_{t-1}\|_{\ell^1} D + 2^{-1}\sigma^{-1}\eta_t \left\| \mathbf{v}_t^{-\frac{1}{4}} \mathbf{m}_t \right\|_*^2 \Bigg) \\
&\quad + \eta_{t_0+1}^{-1} D_{\Psi_{(\mathbf{v}_{t_*+1})}}(\mathbf{w}^*, \mathbf{w}_{t_*+1}) \Bigg] (1 - \beta_{1,1})^{-1},
\end{aligned} \tag{11}$$

where D' is a constant independent of t or d, and will be given in the proof.

Let $\mathbf{w}^*$ be a desired solution for the convex optimization problem (4). The theorem establishes a general regret bound for AdMD in terms of $\mathbf{v}_t, \mathbf{m}_t, \eta_t$. We observe that the quantity $\left|\eta_{t+1}^{-1} v_{t+1,j}^{\frac{1}{2}} - \eta_t^{-1} v_{t,j}^{\frac{1}{2}}\right|$ plays an essential role in the convergence analysis since it quantifies the change of adaptive learning rate $\frac{\eta_t}{\sqrt{\mathbf{v}_t}}$ during the iteration. In previous works on vanilla Adam-style algorithms [46], $\hat{\mathbf{v}}_t = \max\{\mathbf{v}_t, \mathbf{v}_{t-1}\}$ is employed in the updating rule instead of $\mathbf{v}_t$ to keep non-increasing adaptive learning rates. In fact, keeping quantity $\left|\eta_{t+1}^{-1} v_{t+1,j}^{\frac{1}{2}} - \eta_t^{-1} v_{t,j}^{\frac{1}{2}}\right|$ positive is not the only way to guarantee convergence, in the current works, more and more attention is paid to the selection of hyperparameters in order to solve the convergence issue of Adam-style algorithms. Keeping this in mind, we address the issue by focusing on the selection of $\beta_{2,t}$ as follows.

Selection of $\beta_{2,t}$. There are usually two strategies for the selection of $\beta_{2,t}$ in Adam-style algorithms. The first one is that $\beta_{2,t}$ should be a constant close to 1, matching the typical setting $\beta_2 = 0.999$ in practice [20], [33], [39]. The second one is that one can use an increasing schedule of $\beta_{2,t}$ as the iteration progresses [35], [46], [54]. The common idea behind the two strategies is that the sign-based direction of iteration plays a decisive role in the convergence behaviors of Adam-style algorithms. Thus, when the current gradient does not have the right iteration direction, undesirable convergence behaviors will happen. Typically, if a small gradient is accompanied by a relatively large step size, a large update will occur in the wrong direction and lead to divergence. In [14], [35], [48], a set of examples has been provided to depict the phenomenon. With this in mind, the two strategies give more weights on the past gradients, which takes into consideration the update speeds and direction of iteration at the same time. In the following theorem, we will show that AdMD can achieve good convergence rates when adopting the two strategies mentioned above for the selection of $\beta_{2,t}$. For the theoretical analysis, we let $\beta_{2,t} \equiv 1 - \frac{1}{T}$ for the large enough case and $\beta_{2,t} = 1 - \frac{1}{t^2}$ for the increasing case.

We say a norm $\|\cdot\|$ on $\mathbb{R}^d$ is **monotonic** if $\|\mathbf{x}\| \leq \|\mathbf{y}\|$ when $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ satisfy $|x_i| \leq |y_i|$ for $i \in [d]$.

Theorem 2: Let $\{\mathbf{w}_t\}$ be generated by AdMD. Suppose $\|\cdot\|_*$ is monotonic and (9) holds. Under the same conditions of Theorem 1, let $\beta_{1,t} = ct^{-\frac{1}{2}}$ with some $0 < c < 1$, and $\eta_t = t^{-\frac{1}{2}}$. If $\beta_{2,t} \equiv 1 - T^{-1}$ or $\beta_{2,t} = 1 - t^{-2}$, then

$$\begin{aligned}
&\frac{1}{T - t_0} \sum_{t=t_*+1}^{T} \mathbb{E}\left[f(\mathbf{w}_t) - f(\mathbf{w}^*)\right] \\
&\leq \frac{1}{T - t_*} \mathbb{E}\Bigg[D' \sum_{j=1}^{d} (T+1)^{\frac{1}{2}} v_{T+1,j}^{\frac{1}{2}} \\
&\quad + D \sum_{t=t_*+1}^{T} \beta_{1,t} \|\mathbf{m}_{t-1}\|_{\ell^1} + \frac{1}{2\sigma C^{\frac{1}{2}} c_0} \sum_{t=t_*+1}^{T} \eta_t \|\mathbf{m}_t\|_*^2 \\
&\quad + \frac{D_{\Psi_{(\mathbf{v}_{t_*+1})}}(\mathbf{w}^*, \mathbf{w}_{t_*+1})}{\eta_{t_*+1}} \Bigg] (1 - \beta_{1,1})^{-1},
\end{aligned} \tag{12}$$

where C is given in (10).

The proof of Theorem 2 is deferred to the appendix. By ignoring the constant, it gives the convergence order

$$O\left(T^{-\frac{1}{2}}\left(\sum_{j=1}^{d} v_{T+1,j}^{\frac{1}{2}} + \sup_t \mathbb{E}\Big[\|\mathbf{m}_t\|_{\ell^1} + \|\mathbf{m}_t\|_*^2\Big]\right)\right). \quad (13)$$

The explicit convergence rate is determined by the bound of stochastic gradient $\mathbf{g}_t$, and mirror maps, as shown in the following remark.

Remark 1: If $\sup_t \|\mathbf{g}_t\|_\infty < G$ for some $G > 0$, by Lemma 2 and Proposition 1 in the appendix, it is easy to check that $\sum_{j=1}^{d} v_{T+1,j}^{\frac{1}{2}} = O(d)$, $\|\mathbf{m}_{t-1}\|_{\ell^1} = O(d)$. When we use the map $\Psi^{(\epsilon)}(\mathbf{x})$ or $\Psi^{(\ell^2)}(\mathbf{x})$ (Adam), $\|\mathbf{m}_t\|_*^2 = \|\mathbf{m}_t\|_{\ell^2}^2 = O(d)$, the average regret achieves $O(T^{-\frac{1}{2}}d)$. When we use the map $\Psi^{(KL)}(\mathbf{x})$, $\|\mathbf{m}_t\|_*^2 = \|\mathbf{m}_t\|_\infty^2 = O(1)$, the average regret achieves $O(T^{-\frac{1}{2}}d)$.

If $\sup_t \|\mathbf{g}_t\|_{\ell^2} \le G$ for some $G > 0$, it is easy to check that $\sum_{j=1}^{d} \mathbf{v}_{T+1,j}^{\frac{1}{2}} = O(\sqrt{d})$, $\|\mathbf{m}_{t-1}\|_{\ell^1} = O(\sqrt{d})$. When we use the map $\Psi^{(\epsilon)}(\mathbf{x})$ or $\Psi^{(\ell^2)}(\mathbf{x})$ (Adam), $\|\mathbf{m}_t\|_*^2 = \|\mathbf{m}_t\|_{\ell^2}^2 = O(\sqrt{d})$, the average regret achieves $O(T^{-\frac{1}{2}}\sqrt{d})$. When we use the map $\Psi^{(KL)}(\mathbf{x})$, $\|\mathbf{m}_t\|_*^2 = \|\mathbf{m}_t\|_\infty^2 = O(1)$, the average regret achieves $O(T^{-\frac{1}{2}}\sqrt{d})$.

In summary, as shown above, the degree of gradient sparseness will affect the convergence rates, which matches the previous result in the standard Adam [36]. At the same time, we find that AdMD executed by the map $\Psi^{(KL)}$ has the same convergence order as $\Psi^{(\epsilon)}$ or $\Psi^{(\ell^2)}$. From a theoretical view, that is because we use the second momentum $\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2)\mathbf{g}_t \otimes \mathbf{g}_t$. It suggests that in theory we should employ $\mathbf{v}_t = \max\{\beta_2 \mathbf{v}_{t-1}, |\mathbf{g}_t|\}$ to scale down the gradient alternatively and expect to get faster convergence rates. Actually, it also has been pointed out in [18] that the role of $\mathbf{v}_t^{\frac{1}{2}}$ is just to scale down the gradients and can be replaced by other scaling functions defined on the historical gradients, such as $\mathbf{v}_t^{\frac{p}{2}}$ for any $p \in (0, 1)$.

In the following, we show the convergence rate of AdMD for strongly convex problems. We say that a function $f(\mathbf{w}) : \Omega \to \mathbb{R}$ is **λ-strongly convex** with respect to the Bregman distance D_Ψ if

$$f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \ge \lambda D_\Psi(\mathbf{x}, \mathbf{y}), \ \forall \mathbf{x}, \mathbf{y} \in \Omega. \quad (14)$$

Theorem 3: Under the same conditions of Theorem 1, assume that (9) holds and $f(\mathbf{w}; \xi) : \Omega \to \mathbb{R}$ is λ-strongly convex for any given random variable ξ.

(a) When $\beta_{2,t} = 1 - t^{-2}$, let $\beta_{1,t} = ct^{-1}$ with some $0 < c < 1$, and $\eta_t = \eta t^{-1}$ with $\eta = \frac{2G}{(1-c)\lambda}$. Then

$$\frac{1}{T - t_*}\sum_{t=t_*+1}^{T} \mathbb{E}[f(\mathbf{w}_t) - f(\mathbf{w}^*)]$$
$$= O\left(T^{-1}\log T \sup_t \mathbb{E}\Big[\|\mathbf{m}_t\|_{\ell^1} + \|\mathbf{m}_t\|_*^2\Big]\right). \quad (15)$$

(b) When $\beta_{2,t} \equiv 1 - T^{-1}$, let $\beta_{1,t} = ct^{-1}$ with some $0 < c < 1$, and $\eta_t = \eta t^{-1}$ with $\eta = \frac{G + C^{-\frac{1}{2}} c_0^{-1} G^2}{(1-c)\lambda}$, the statement (15) holds.

In comparison with (13), the convergence rate in (15) improves to $O(T^{-1}\log T)$ when $\|\mathbf{m}_t\|_{\ell^1}, \|\mathbf{m}_t\|_*$ are bounded. This improvement shows that AdMD can achieve fast convergence rates as classical SGD, and improve the best rates obtained in the previous study on vanilla Adam-type algorithms in the strongly convex setting. However, we note that in the above theorems, a decreasing $\beta_{1,t}$ is employed and bounded far away from 1. It does not give a convincing explanation for the selection of Adam-type parameters in experiments, which suggest $\beta_{1,t} \equiv 0.9$. At the same time, due to the successes of adaptive gradient methods for training deep neural networks, there have been many recent attempts to prove their nonconvex optimization guarantees [20], [28], [33], [39], [51]. So we shall provide theoretical analysis for AdMD in smooth scenarios (not necessarily convex) in the Section III, where more natural selections of $\beta_{1,t}$ are employed.

III. Main results with Lipschitz continuous gradient

In this section, we study the convergence issue of AdMD algorithms based on the L-smooth objective functions.

Definition 2: Let $h(\cdot)$ be a differentiable function on $(\Omega, \|\cdot\|)$. We say that h is L-smooth if the gradient $\nabla h(\cdot)$ is L-Lipschitz continuous, that is

$$\|\nabla h(\mathbf{u}) - \nabla h(\mathbf{v})\|_* \le L\|\mathbf{u} - \mathbf{v}\|, \quad \mathbf{u}, \mathbf{v} \in \Omega. \quad (16)$$

Moreover, we say that a function $h(\cdot)$ on $\mathbb{R}^d$ is coordinatewise smooth if for $i \in [d]$, there exists a constant $L_i > 0$ such that

$$|\nabla_i h(\mathbf{u} + s\mathbf{e}_i) - \nabla_i h(\mathbf{u})| \le L_i s, \ \forall s > 0, \quad (17)$$

where $\mathbf{e}_i$ denotes the i-th unit vector and ∇_i the derivative with respect to the i-th coordinate.

It is not difficult to prove that when h satisfies (16), the following inequality holds

$$h(\mathbf{u}) \le h(\mathbf{v}) + \langle \nabla h(\mathbf{v}), \mathbf{u} - \mathbf{v}\rangle + \frac{L\|\mathbf{u} - \mathbf{v}\|^2}{2}, \quad \mathbf{u}, \mathbf{v} \in \Omega, \quad (18)$$

which will be used repeatedly in the rest of the paper.

Assumption 3: The random gradient $\nabla f(\mathbf{w}; \xi)$ on ξ is an unbiased stochastic estimation of $\nabla f(\mathbf{w})$, that is $\mathbb{E}_\xi[\nabla f(\mathbf{w}; \xi)] = \nabla f(\mathbf{w})$ for all $\mathbf{w} \in \Omega$. In addition, $\nabla f(\mathbf{w}; \xi)$ has a bounded variance such that for some constants $\delta^2, c_0 > 0$,

$$\mathbb{E}_\xi[\|\nabla f(\mathbf{w}; \xi) - \nabla f(\mathbf{w})\|_*^2] \le \delta^2(1 + c_0\|\nabla f(\mathbf{w}_t)\|_*^2). \quad (19)$$

The variance bounded condition is weaker than the boundedness assumption $\mathbb{E}_\xi[\|\nabla f(\mathbf{w}; \xi) - \nabla f(\mathbf{w})\|_*^2] \le \delta^2$, which is commonly used in analyzing the convergence behaviors of stochastic gradient methods (see [28], [33]).

To prove our main results, we first introduce auxiliary variables $\mathbf{z}_t = \beta_{1,t}\mathbf{z}_{t-1} + (1 - \beta_{1,t})\mathbf{g}_t$. Then the lemma below holds, whose proof will be provided in the appendix.

Lemma 1: Let $c_3 > 0$. Suppose f is L-smooth satisfying (16), $\min_{i\in[d]} v_{t-1,i} \geq c_3$ and $\|\cdot\|_*$ is monotonic. Then we get

$$\begin{aligned}
&\mathbb{E}_{\xi_t}\left[\|\mathbf{z}_t - \nabla f(\mathbf{w}_t)\|_*^2\right] \\
&\leq \beta_{1,t}\|\mathbf{z}_{t-1} - \nabla f(\mathbf{w}_{t-1})\|_*^2 + \frac{4c_3^{-1}L^2\eta_{t-1}^2}{(1-\beta_{1,t})\sigma^2}\|\mathbf{m}_{t-1}\|_*^2 \\
&\quad + \delta^2(1-\beta_{1,t})^2(1 + c_0\|\nabla f(\mathbf{w}_t)\|_*^2). \qquad (20)
\end{aligned}$$

When $\mathbf{z}_t = \mathbf{m}_t$, the lemma tracks the objective function gradient $\nabla f(\cdot)$ at iterate $\mathbf{w}_t$ by using the past stochastic gradients $\mathbf{g}_i, 1 \leq i \leq t$. It should be noted that the first two terms in the above bound increase as $\beta_{1,t}$ increases, while the last term decreases. So, based on Lemma 1, we show how $\beta_{1,t}$ balances in the iteration process, and then fast convergence rates for AdMD are derived when $\|\cdot\| = \|\cdot\|_* = \|\cdot\|_{\ell^2}$.

Theorem 4: Let $\{\mathbf{w}_t\}$ be generated by AdMD. Suppose Ψ satisfies (8) and is coordinatewise smooth with (17). Let f be L-smooth satisfying (16). Assume (19) holds, and for any $i \in [d]$ and $t_* \leq t \leq T$, $c_3 < v_{t,i} \leq c_4$ with some $c_3, c_4 > 0$.

(a) If $\beta_{1,t} \equiv 1 - T^{-\frac{1}{2}}$, and $\eta_t \equiv \eta_1 T^{-\frac{1}{2}}$ satisfying

$$\eta_1 \leq \min\{\sqrt{(8c_uL^2)^{-1}c_3c_l}\sigma, (2c_u^2L)^{-1}c_l\}, \qquad (21)$$

we get for $T \geq \max\{(2c_0c_u\delta^2c_l^{-1})^2, 2t_*\}$,

$$\frac{1}{T-t_*}\sum_{t=t_*+1}^{T} \mathbb{E}[\|\nabla f(\mathbf{w}_t)\|_{\ell^2}^2] \leq C_1T^{-\frac{1}{2}}. \qquad (22)$$

(b) If $\beta_{1,t} = 1 - t^{-\frac{1}{2}}$, and $\eta_t = \eta_1 t^{-\frac{1}{2}}$ satisfying

$$\eta_1 \leq 4^{-1}\sqrt{c_l(c_u\eta_1L^2)^{-1}c_3}\sigma, \qquad (23)$$

we get for $T \geq \max\{2t_*, (4c_uc_0\delta^2c_l^{-1})^2\}$,

$$\frac{1}{T-t_0}\sum_{t=t_*+1}^{T} \mathbb{E}[\|\nabla f(\mathbf{w}_t)\|_{\ell^2}^2] \leq C_2T^{-\frac{1}{2}}\log T, \qquad (24)$$

where $c_l = \min_{i\in[d]} L_i^{-1}c_4^{-\frac{1}{2}}$, $c_u = 2c_3^{-\frac{1}{2}}c_0/\sigma$, and C_1, C_2 are constants independent of T (given explicitly in the proof). The theorem highlights the selection of the first hyperparameter $\beta_{1,t}$ for AdMD. Here, an increasing or large $\beta_{1,t}$ is used, which is more general and closer to the practical setting of Adam-type algorithms. As a concrete example, vanilla Adam satisfies (17) with $\Psi(\mathbf{w}) = \|\mathbf{w}\|_{\ell^2}^2$. In the following, we further illustrate the above theorem by considering vanilla Adam with suitable selection parameters.

Corollary 1: Let $\{\mathbf{w}_t\}$ be generated with $\Psi(\mathbf{w}) = \|\mathbf{w}\|_{\ell^2}^2$. Suppose $\sup_t \|\mathbf{g}_t\|_\infty < G$ for some $G > 0$ and (9) holds. Then we have the following statements.

(a) If $\beta_{1,t} \equiv 1-T^{-\frac{1}{2}}$, $\beta_{2,t} \equiv 1-T^{-1}$ and $\eta_t \equiv \eta_1 T^{-\frac{1}{2}}$ satisfying (21), the convergence rate (22) holds.

(b) If $\beta_{1,t} = 1-t^{-\frac{1}{2}}$, $\beta_{2,t} = 1-t^{-2}$ and $\eta_t = \eta_1 t^{-\frac{1}{2}}$ satisfying (23), the convergence rate (24) holds.

Remark 2: The corollary suggests two selection strategies of $\beta_{1,t}, \beta_{2,t}$ in vanilla Adam-type algorithms. The first one requires that the two hyperparameters are both large enough, the second one requires that they are both increasing. Meanwhile, in the two selection strategies above, we observe that $\beta_{1,t} \leq \sqrt{\beta_{2,t}}$ holds, which is in good agreement with the practical setting.

IV. Discussions and Comparisons

The motivation of the combination of mirror descent and adaptive gradient methods can be traced back to [21], [49], which aims to exploit optimization problems' regularization structure. For coordinatewise mirror maps (6) with $J(\cdot) \geq 0$, update of AdMD in Algorithm 1 can be rewritten equivalently as

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\Omega}\left\{\eta_t\langle\mathbf{m}_t, \mathbf{w}\rangle_{\ell^2} + D_{\Psi^{(t)}}(\mathbf{w}, \mathbf{w}_t)\right\}, \qquad (25)$$

where the time-varying Bregman distances $D_{\Psi^{(t)}}$ are induced by the squared Mahalanobis-like norm $\Psi^{(t)}(\mathbf{w}) = \langle\mathcal{J}^{\frac{1}{2}}(\mathbf{w}), \mathbf{M}_t\mathcal{J}^{\frac{1}{2}}(\mathbf{w})\rangle_{\ell^2}$ for a symmetric positive matrix $\mathbf{M}_t = \mathbf{diag}(\mathbf{v}_t\mathbf{v}_t^T)^{\frac{1}{4}}$ and $\mathcal{J}^{\frac{1}{2}}(\mathbf{w}) = (J^{\frac{1}{2}}(w_1), \cdots, J^{\frac{1}{2}}(w_d)) \in \mathbb{R}^d$. Following the same idea, update of Adam also takes the form of (25) by setting $\mathcal{J}^{\frac{1}{2}}(\mathbf{w}) = \frac{1}{\sqrt{2}}\mathbf{w}$, which is reformulated as

$$\arg\min_{\mathbf{w}\in\Omega}\left\{\eta_t\langle\mathbf{m}_t, \mathbf{w}\rangle_{\ell^2} + D_{\Psi^{(t)}}(\mathbf{w}, \mathbf{w}_t)\right\}$$

where $D_{\Psi^{(t)}}(\mathbf{w}, \mathbf{w}_t) = \langle\mathbf{w} - \mathbf{w}_t, \mathbf{M}_t(\mathbf{w} - \mathbf{w}_t)\rangle_{\ell^2}$. In this sense, vanilla Adam and other adaptive gradient methods are essentially mirror descent with temporal adaptation of squared Mahalanobis metrics. Thus, AdMD is a natural extension of adaptive gradient methods with more different choices of mirror maps, where optimization problem structures are dynamically incorporated into learning processes so that the learning algorithm can be performed more effectively. To our knowledge, this is the first time to investigate the convergence issue of Adam-like algorithms in the context of general mirror maps. In this work, we prove that AdMD can indeed achieve the optimal convergence rates or regret bounds that are standard in the literature on stochastic and online optimization.

We review some related works on the convergence properties of vanilla Adam and its variants. First, we compare our results in a generic framework of solving convex optimization problems. Adam [36] has witnessed a great success in training many tasks in machine learning area and promising alternative algorithms to Adam have been proposed, such as AMSGrad, AdamNC [46], AdaShift [52], NosAdam [35], AdaBound [43]. In the convex optimization, their theoretical analyses can achieve the rate of order $O(T^{-\frac{1}{2}})$ if the first hyperparameter $\beta_{1,t}$ is chosen to be non-increasing. A comparison with these existing results is stated in Theorem 2, and the same convergence order is obtained as shown in Remark 1. In addition, we improve it to $O(T^{-1}\log T)$ in Theorem 3 for strongly convex objective functions, which has not been addressed in the aforementioned works.

Next, we turn to recent works on solving L-smooth nonconvex problems. In practice, large hyperparameters $\beta_{1,t}, \beta_{2,t}$ promote the convergence of Adam-style algorithms and the recommended choice is $\beta_{1,t} = 0.9, \beta_{2,t} = 0.999$. In theory, a series of works have emerged to understand the selection of hyperparameters. For instance, in the smooth non-convex setting, Zou et al. [54] provided a convergence bound for Adam under increasing $\beta_{2,t} = 1 - t^{-1}$. D*é*fossez et al.

[20] established a probabilistic convergence bound by letting $\beta_{2,t} \equiv \beta_2$ be a fixed constant. Hong et al. [33] recently derived a high probability convergence rate of $O(T^{-\frac{1}{2}} \log T)$ under $\beta_{2,t} \equiv 1 - T^{-1}$. Chen et al. [18] showed that in the Adam update $\mathbf{v}^{\frac{1}{2}}$ can be replaced with $\mathbf{v}^p$ for any $p \in (0, \frac{1}{2})$, which enjoys the fast convergence rate under the hyperparameters restriction $\beta_{1,t} \leq \beta_{2,t}^p$. Their theoretical works focused on how to interpret or modify the scaling term $\mathbf{v}_t$ by properly selecting $\beta_{2,t}$ and treating $\beta_{1,t}$ as a constant β_1. The corresponding convergence rate exhibits a dependence on $(1 - \beta_1)^{-1}$. It means their obtained convergence rate becomes considerably large as β_1 is close to 1, which cannot explain why large $\beta_{1,t}$ is advisable in Adam-type algorithms. In the recent work by [28], a convergence rate was established by allowing for an increasing or large $\beta_{1,t}$, which applies to all variants of Adam under standard conditions. Similarly, in Theorem 4 of this paper, AdMD has been shown to converge in L-smooth optimization problems with the choice $\beta_{1,t} = 1 - t^{-\frac{1}{2}}$ or $\beta_{1,t} \equiv 1 - T^{-\frac{1}{2}}$, removing the dependence on $(1 - \beta_1)^{-1}$. It has been discussed in Remark 2. Our results cover both the selection of $\beta_{1,t}$, $\beta_{2,t}$, which gives reasonable explanations for the success of Adam-type algorithms in smooth (not necessarily convex) problems.

At the end of this section, we remark on the selection of parameters when and how different Adam-like algorithms can perform well according to our results. In the convex setting, as shown in Theorem 2, we show that with $\beta_{2,t} \equiv 1 - T^{-1}$ or $\beta_{2,t} = 1 - t^{-2}$, the proposed AdMD achieves a convergence rate of $O(T^{-\frac{1}{2}})$, which matches the best known convergence rate in adaptive gradient algorithms. In Theorem 3, with the same choice as Theorem 2, the convergence rate improves to $O(T^{-1} \log T)$ in the strongly-convex setting, which has not been derived in Adam-like's literature. Meanwhile, in the non-convex L-smooth setting of Theorem 4 and Corollary 1, increasing $\beta_{1,t} = 1 - t^{-\frac{1}{2}}, \beta_{2,t} = 1 - t^{-2}$ and large $\beta_{1,t} \equiv 1 - T^{-\frac{1}{2}}, \beta_{2,t} \equiv 1 - T^{-1}$ will lead to convergence rate of $O(T^{-\frac{1}{2}})$ and $O(T^{-\frac{1}{2}} \log T)$, respectively. These results suggest that we should not be rely heavily on the gradient information we access in the present, and the past gradients are useful since the direction of the current gradient may be opposite to the correct optimization direction. We illustrate the choice of the results with the help of the typical example stated in [55]:

$$f_t(x) = \begin{cases} Cx, & \text{if } t \textbf{ mod } C = 1, \\ -x, & \text{otherwise.} \end{cases}$$

where C is larger than 2 and $x \in [-1, 1]$. The right optimization direction should go towards -1, while the iterates would progress in the opposite direction. That is because the frequency of the large gradient C is much less than the gradient -1, which moves the algorithm in the wrong direction. For the small value of $\beta_{2,t}$, the large gradient C cannot offset the effect since the base learning rate η_t is scaled down by a relatively large $\mathbf{v}_t$, and hence the algorithm converges to 1 rather than -1. In adaptive methods, $\mathbf{v}_t$ in fact serves as a coordinate-wise scaling of the base learning rates. If $\mathbf{v}_t$ is determined by its most current gradients, when the current $\mathbf{g}_t$ keeps being small, the adaptive learning rate could be large, which may lead to oscillation of iterates and the possibility of getting trapped in local minimum. Since the direction of iteration plays a crucial role in the convergence behaviors of Adam-style algorithms, either increasing or choosing large parameters in our choices guarantee a more stable calculation of past gradients, which do not heavily depend on the few recent gradients.

V. Simulation Validation

In this section, we perform both toy simulations and a CT-reconstruction experiment to demonstrate the performance of AdMD in Algorithm 1, and to verify the theoretical statements. All experiments were executed using Python 3.7 on a PC with an Intel Core i5 2GHz processor. Scripts reproducing these experiments are available at https://github.com/Ariesoomoon/AdMD_experiments.

A. Toy simulations

We conducted four simulations. The first two simulations demonstrate the effectiveness and superior performance of AdMD in the convex and strongly convex cases, respectively, validating Theorems 2 and 3. The third simulation explores the convergence behavior of AdMD in the nonconvex setting using the two proposed parameter settings, thereby verifying Corollary 1. Table I summarizes the experimental settings for the first three simulations. To ensure a fair comparison, consistent parameter settings are applied across all algorithms. Based on these validations, the final simulation extends the analysis by comparing AdMD with other adaptive optimization algorithms.

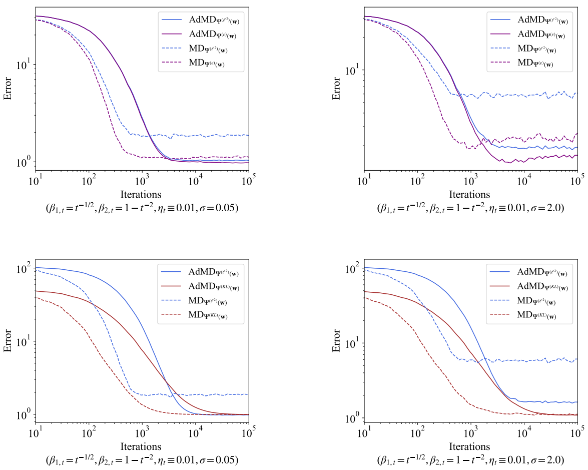
We generate a dataset of $|D| = 1000$ samples for training, where $|D|$ denotes the cardinality of the dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|D|}$, and $\mathbf{x}_i = (x_i^j)_{j=1}^d$, where d represents the dimension of $\mathbf{x}_i$. To showcase the algorithm's ability to capture geometric information, we generate each dimension of the inputs $\{x_i\}_{i=1}^{|D|}$ independently from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$ with $\sigma = 1$. The outputs are defined as $y_i = y_i' + \varepsilon_i$, where y_i' is independently drawn from the same Gaussian distribution, and ε_i is sampled from $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.05$ or 2.0. We use linear regression to fit the data, aiming to gradually approach the optimal coefficient $\mathbf{w} \in \mathbb{R}^d, \mathbf{w} = (w_j)_{j=1}^d$ through gradient descent. In our experiments, we run each algorithm over the noisy training samples $D = \{(\mathbf{x_i}, y_i)\}_{i=1}^{|D|}$ and evaluate their performance on the noise-free dataset $D' = \{(\mathbf{x}_i, y_i')\}_{i=1}^{|D|}$. We use the error $\frac{1}{|D|} \sum_{i=1}^{|D|} (y_i' - \mathbf{x}_i^\top \mathbf{w}_t)^2$ to evaluate the algorithm's performance at each iteration t. Note that in experiments, we did not provide the optimal $\mathbf{w}$, but instead observed the convergence of such errors during the iteration process.

Except for the $\Psi^{(\ell^2)}(\mathbf{w})$, we choose $\Psi^{(KL)}(\mathbf{w})$ and $\Psi^{(\epsilon)}(\mathbf{w})$ as representative mirror maps because of their focus on sparsity and smoothness, respectively. $\Psi^{(KL)}(\mathbf{w})$ emphasizes small weights, making it ideal for capturing sparsity in optimization tasks (similarly, $\Psi^{(IS)}(\mathbf{w})$ performs well in tasks where sparsity is crucial). In contrast, $\Psi^{(\epsilon)}(\mathbf{w})$ ensures smooth and stable optimization through regularization, similar to $\Psi^{(FD)}(\mathbf{w})$ and $\Psi^{(He)}(\mathbf{w})$. We compare AdMD with three mirror maps $\Psi^{(\ell^2)}(\mathbf{w})$, $\Psi^{(\epsilon)}(\mathbf{w})$, and $\Psi^{(KL)}(\mathbf{w})$. For simplicity, we refer to these as $\text{AdMD}_{\Psi^{(\ell^2)}(\mathbf{w})}$, $\text{AdMD}_{\Psi^{(\epsilon)}(\mathbf{w})}$, and $\text{AdMD}_{\Psi^{(KL)}(\mathbf{w})}$,

TABLE I: Detailed settings in simulations.

Simulations	Verification	Loss function	$\beta_{1,t}$ in $\mathbf{m}_t$	$\beta_{2,t}$ in $\mathbf{v}_t$	Stepsize η_t
Simulation 1	Theorem 2	Convex: *Least square loss* $f(\mathbf{w}_t) = \frac{1}{\lvert D\rvert}\sum_{i=1}^{\lvert D\rvert}(y_i - \mathbf{x}_i^\top \mathbf{w}_t)^2$	$\beta_{1,t} = ct^{-1/2}, c = 0.3$	- $\beta_{2,t} \equiv 1 - T^{-1}$ - $\beta_{2,t} = 1 - t^{-2}$	0.01
Simulation 2	Theorem 3	Strongly convex ($\lambda = 0.8$): *L_2-regularized least square loss* $f(\mathbf{w}_t) = \frac{1}{\lvert D\rvert}\sum_{i=1}^{\lvert D\rvert}(y_i - \mathbf{x}_i^\top \mathbf{w}_t)^2 + \lambda\lVert \mathbf{w}_t\rVert_2^2$	$\beta_{1,t} = ct^{-1}, c = 0.3$	$\beta_{2,t} = 1 - t^{-2}$	0.01
Simulation 3	Corollary 1	Nonconvex: Nonconvex logarithmic loss	$\beta_{1,t} \equiv 1 - T^{-1/2}$	$\beta_{2,t} \equiv 1 - T^{-1}$	$T^{-1/2}$
			$\beta_{1,t} = 1 - t^{-1/2}$	$\beta_{2,t} = 1 - t^{-2}$	$t^{-1/2}$

Note: A small $\epsilon_0 = 10^{-8}$ is introduced into the stepsize $\frac{\eta_t}{\sqrt{\mathbf{v}_t}}$ and modified as $\frac{\eta_t}{\sqrt{\mathbf{v}_t}+\epsilon_0}$ to ensure numerical stability.



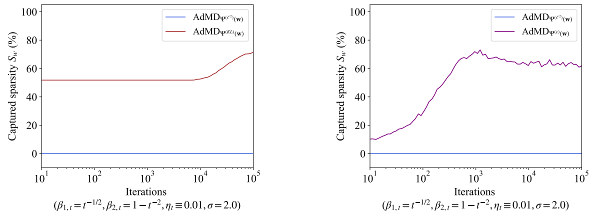
$(\beta_{1,t} = t^{-1/2}, \beta_{2,t} = 1 - t^{-2}, \eta_t \equiv 0.01, \sigma = 0.05)$

$(\beta_{1,t} = t^{-1/2}, \beta_{2,t} = 1 - t^{-2}, \eta_t \equiv 0.01, \sigma = 2.0)$

$(\beta_{1,t} = t^{-1/2}, \beta_{2,t} = 1 - t^{-2}, \eta_t \equiv 0.01, \sigma = 0.05)$

$(\beta_{1,t} = t^{-1/2}, \beta_{2,t} = 1 - t^{-2}, \eta_t \equiv 0.01, \sigma = 2.0)$

Fig. 1: Results in the convex case (for Theorem 2).



$(\beta_{1,t} = t^{-1/2}, \beta_{2,t} = 1 - t^{-2}, \eta_t \equiv 0.01, \sigma = 2.0)$

$(\beta_{1,t} = t^{-1/2}, \beta_{2,t} = 1 - t^{-2}, \eta_t \equiv 0.01, \sigma = 2.0)$

Fig. 2: Captured sparsity during the iteration process.

respectively. Additionally, we compare AdMD with general mirror descent (MD) defined in (2) with the same mirror map. Note that for AdMD$_{\Psi^{(\ell^2)}(\mathbf{w})}$, the update process of $\mathbf{w}_t$ follows (5). For AdMD$_{\Psi^{(KL)}(\mathbf{w})}$, the update process of $\mathbf{w}_t$ is:

$$\mathbf{w}_{t+1} = \mathbf{w}_t e^{-\eta_t \frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t}}}.$$

For AdMD$_{\Psi^{(\epsilon)}(\mathbf{w})}$, the update process of $\mathbf{w}_t$ is:

$$\begin{cases} \nabla\Psi^{(\epsilon)}(\mathbf{w}_{t+1}) = \nabla\Psi^{(\epsilon)}(\mathbf{w}_t) - \eta_t \frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t}}, \\ \mathbf{w}_{t+1} = M_\epsilon\left(\nabla\Psi^{(\epsilon)}(\mathbf{w}_{t+1})\right), \end{cases}$$

where $M_\epsilon : \mathbb{R}^d \to \mathbb{R}^d$ is defined component-wise in terms of the function $M_\epsilon : \mathbb{R} \to \mathbb{R}$ given by

$$M_\epsilon(u) = \begin{cases} \frac{u\epsilon}{1+\epsilon}, & \text{if } |u| \le 1+\epsilon, \\ \operatorname{sgn}(u)(|u|-1), & \text{otherwise.} \end{cases}$$

We set $\mathbf{m}_1 = \mathbf{v}_1 = 0 \in \mathbb{R}^d$ and $\epsilon = 10^{-8}$ (as suggested in [16]) here. To assess AdMD$_{\Psi^{(KL)}(\mathbf{w})}$, we initialized $\mathbf{w}_0$ from a Gaussian distribution $\mathcal{N}(0,1)$ in the experiments involving the mirror map $\Psi^{(KL)}(\mathbf{w})$. In other experiments, $\mathbf{w}_0$ was initialized from a uniform distribution $\mathcal{U}[0,1)$.

In all simulations, we set the maximum number of iterations to $T = 10^5$. To present the results more clearly, we did not display the results at every iteration. Instead, we selected 100 iterations using a logarithmic scale between 1 and T. The presented iteration indices were calculated by taking the integer part of the logarithmic scale values: $\left\{\left\lfloor 10^{\frac{k}{n-1}\log_{10}(T)}\right\rfloor \mid k = 0,1,2,\ldots,n-1\right\}$ where $n = 100$ and $\lfloor\cdot\rfloor$ denotes the floor function. Each simulation was conducted over 10 trials for averaging. For reproducibility, we used a fixed seed for random sample selection at each iteration of every trial.

1) Simulation 1: Effectiveness and superior performance of AdMD in the convex case: In this simulation, our goals are twofold: (1) to verify that AdMD benefits from using mirror maps that capture geometric information of data, and (2) to demonstrate that AdMD exhibits better convergence performance than MD due to its Adam-style updates. Note that MD$_{\Psi^{(\ell^2)}(\mathbf{w})}$ corresponds to classical SGD.

Figure 1 presents four main observations. (1) From the two subplots in the first row, AdMD$_{\Psi^{(\epsilon)}(\mathbf{w})}$ outperforms AdMD$_{\Psi^{(\ell^2)}(\mathbf{w})}$ in convergence, demonstrating the effectiveness of AdMD and its superiority in leveraging the geometric properties of the data. This is because the mirror map $\Psi^{(\epsilon)}(\mathbf{w})$ provides better smoothness by applying larger updates to smaller w_js and smaller updates to larger ones. This adaptive behavior accelerates convergence by focusing updates on more important coefficients. (2) From the two subplots in the second row, AdMD$_{\Psi^{(KL)}(\mathbf{w})}$ also achieves better convergence than AdMD$_{\Psi^{(\ell^2)}(\mathbf{w})}$, further validating the effectiveness of AdMD. Due to the logarithmic nature of $\Psi^{(KL)}(\mathbf{w})$, AdMD$_{\Psi^{(KL)}(\mathbf{w})}$ applies larger updates to small w_js and prevents them from being amplified by noise. (3) The superior performance of AdMD$_{\Psi^{(\epsilon)}(\mathbf{w})}$ and AdMD$_{\Psi^{(KL)}(\mathbf{w})}$ is more evident at higher noise levels, as these mirror maps correct noise-affected w_js more efficiently than $\Psi^{(\ell^2)}(\mathbf{w})$, which struggles to suppress the impact of noise due to uniform linear adjustments. (4) Lastly, AdMD consistently outperforms MD with the same mirror map, due to its Adam-style gradient updates.

We also present the changes in the sparsity of the coefficient vector $\mathbf{w}_t$ during iterations under different mirror

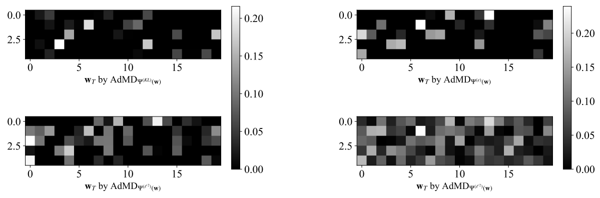


Fig. 3: Visualization of $\mathbf{w}_T$ with $T = 10^5$ (results of a single trial).

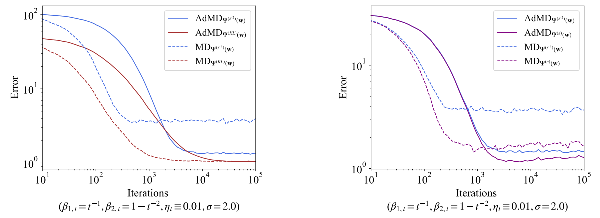


Fig. 4: Results in the strongly convex case (for Theorem 3).

maps: $\Psi^{(\ell^2)}(\mathbf{w})$, $\Psi^{(KL)}(\mathbf{w})$, and $\Psi^{(\epsilon)}(\mathbf{w})$. The sparsity, S_w, is defined as $S_w = \frac{\sum_{j=1}^{d} \mathbf{I}\{|w_j| \le 10^{-8}\}}{d} \times 100\%$. As shown in Figure 2, the sparsity remains constant throughout the iterations with $\Psi^{(\ell^2)}(\mathbf{w})$, while it increases with $\Psi^{(KL)}(\mathbf{w})$ and $\Psi^{(\epsilon)}(\mathbf{w})$. To clearly visualize the sparsity after T iterations, we transformed the 100-dimensional vector $\mathbf{w}_T$ into a 5×20 pixel image, as shown in Figure 3. Both AdMD$_{\Psi^{(\epsilon)}(\mathbf{w})}$ and AdMD$_{\Psi^{(KL)}(\mathbf{w})}$ exhibit better sparsity compared with AdMD$_{\Psi^{(\ell^2)}(\mathbf{w})}$, as small values of w_j are driven to zero.

2) Simulation 2: Effectiveness and superior performance of AdMD in the strongly convex case: In this simulation, we use L_2-regularized least squares as a strongly convex loss function. The regularization parameter λ is fixed at 0.8. The results, shown in Figure 4, are for a noise level of 2.0, with the rest of the experimental setup identical to Simulation 1. We make two key observations: (1) Both AdMD$_{\Psi^{(KL)}(\mathbf{w})}$ and AdMD$_{\Psi^{(\epsilon)}(\mathbf{w})}$ demonstrate better convergence performance than AdMD$_{\Psi^{(\ell^2)}(\mathbf{w})}$. This trend is also observed with MD, further illustrating that $\Psi^{(KL)}(\mathbf{w})$ and $\Psi^{(\epsilon)}(\mathbf{w})$ are effective at capturing the geometric properties of the data. (2) Within the current T iterations, AdMD consistently converges to a lower error than MD, which can be attributed to the acceleration effect of the Adam-style updates in AdMD. These observations support Theorem 3, confirming that under a strongly convex loss function, AdMD achieves the same convergence rate as SGD (i.e., MD$_{\Psi^{(\ell^2)}(\mathbf{w})}$).

3) Simulation 3: Power of AdMD in the nonconvex case: The aim of this simulation is to verify Corollary 1 by showing the convergence of AdMD under the assumption that the loss function f is L-smooth that satisfies the inequality (16), and the mirror map fulfills the coordinate-wise smoothness requirement in inequality (17). To meet these assumptions, we use the nonconvex logarithmic loss as follows:

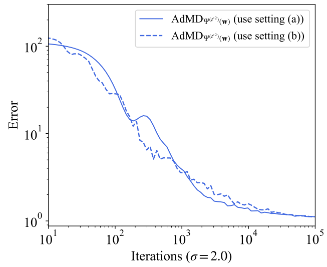


Fig. 5: Results in the nonconvex case (setting (a) and setting (b) are given by Corollary 1).

$$f(\mathbf{w}_t) = \frac{1}{|D|} \sum_{i=1}^{|D|} \log\left(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w}_t)\right) + \alpha \log\left(1 + \|\mathbf{w}_t\|^2\right),$$

which combines both smooth and nonconvex elements.

We then use the mirror map $\Psi^{(\ell^2)}(\mathbf{w})$ that satisfies the coordinate-wise smoothness requirement. The selected loss function and mirror map ensure that the theoretical assumptions of Corollary 1 are satisfied, allowing us to validate the convergence behavior of the AdMD algorithm. As shown in Figure 5, with the two parameter settings from Corollary 1, the error exhibits a converging trend, validating Corollary 1.

4) Simulation 4: Comparison of AdMD with other adaptive optimization algorithms: We also conducted a detailed comparison between AdMD and other adaptive optimization algorithms, including AdaGrad [21], Adam [57], and RMSProp [47]. For simplicity, we only performed the experiments in the convex case, with x_i sampled separately from both the uniform distribution $\mathcal{U}(0,1)$ and the Gaussian distribution $\mathcal{N}(0,1)$ to generate data with distinct geometric structures. Additionally, we tested on both noisy and noise-free data. To ensure fairness in comparison, we kept the parameters η_t, $\beta_{1,t}$, and $\beta_{2,t}$ the same for all algorithms.

Figure 6 highlights three key observations: (1) Except for AdaGrad, the convergence performance of these adaptive algorithms is quite similar. Notably, the results for Adam and AdMD$_{\Psi^{(\ell^2)}(\mathbf{w})}$ are almost identical, demonstrating the effectiveness of AdMD. (2) AdMD$_{\Psi^{(\epsilon)}(\mathbf{w})}$ (or #AdMD$_{\Psi^{(\epsilon)}(\mathbf{w})}$) outperforms AdMD$_{\Psi^{(\ell^2)}(\mathbf{w})}$ (or #AdMD$_{\Psi^{(\ell^2)}(\mathbf{w})}$), where # indicates AdMD using $\beta_{1,t} = 0.3t^{-1/2}$ and $\beta_{2,t} \equiv 1 - T^{-1}$. This is because the mirror map $\Psi^{(\epsilon)}(\mathbf{w})$ adapts to gradient magnitudes more effectively by leveraging data sparsity. This differs from the mirror map $\Psi^{(\ell^2)}(\mathbf{w})$, which applies uniform updates across all coefficients. We also observe that this advantage is pronounced when x_i is sampled from a Gaussian distribution, suggesting that AdMD with mirror maps capable of capturing geometric information performs better when the data itself exhibits specific geometric structures. (3) Compared to noise-free data, the advantage of AdMD$_{\Psi^{(\epsilon)}(\mathbf{w})}$ and #AdMD$_{\Psi^{(\epsilon)}(\mathbf{w})}$ becomes more evident with noisy data.

B. CT reconstruction

In this section, we conduct a CT reconstruction experiment on the Shepp-Logan (SL) phantom image (with distinct geo-

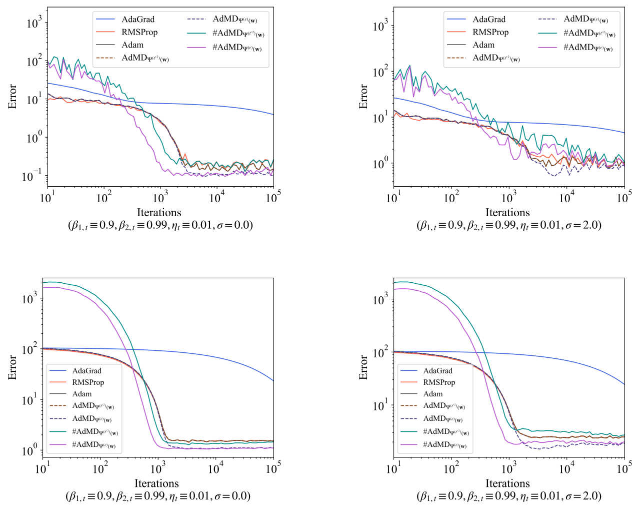


Fig. 6: Comparison of different algorithms in the convex case, where # represents AdMD with parameters recommended by Theorem 2, specifically $\beta_{1,t} = 0.3t^{-1/2}$ and $\beta_{2,t} \equiv 1 - T^{-1}$.

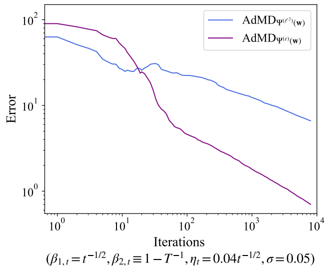


Fig. 7: Results of $\text{AdMD}_{\Psi^{(\ell^2)}(\mathbf{w})}$ and $\text{AdMD}_{\Psi^{(\epsilon)}(\mathbf{w})}$.

metric shapes) to verify our theoretical assessments. Specifically, we utilize the AIR toolbox [56] to construct a projection matrix $A \in \mathbb{R}^{n\times d}$ and an $N \times N$ sparse image, represented by a vector $\mathbf{w} \in \mathbb{R}^d$ with $d = N^2$. We set $N = 32$ and $n = 11520$. Our goal is to reconstruct the image of coefficient $\mathbf{w}$ based on the noisy measurement dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i$ is the i-th row of the matrix A and $y_i = y_i' + \epsilon_i'$ with $y_i' = \mathbf{x}_i^\top \mathbf{w}$ and ϵ_i' independently drawn from a Gaussian distribution $\mathcal{N}(0, 0.05^2)$. For simplicity, we only conduct the experiment using $\text{AdMD}_{\Psi^{(\ell^2)}(\mathbf{w})}$ and $\text{AdMD}_{\Psi^{(\epsilon)}(\mathbf{w})}$ to demonstrate the superior performance of AdMD when employing a mirror map that captures the geometric structure of $\mathbf{w}$.

Figure 7 illustrates the performance of $\text{AdMD}_{\Psi^{(\ell^2)}(\mathbf{w})}$ and $\text{AdMD}_{\Psi^{(\epsilon)}(\mathbf{w})}$ in the convex case. The error is calculated as in the previous simulations. As observed, $\text{AdMD}_{\Psi^{(\epsilon)}(\mathbf{w})}$ demonstrates superior convergence as compared to $\text{AdMD}_{\Psi^{(\ell^2)}(\mathbf{w})}$, thereby verifying Theorem 2. Figure 8 shows the ground truth and the reconstruction of $\mathbf{w}$ using $\text{AdMD}_{\Psi^{(\ell^2)}(\mathbf{w})}$ and $\text{AdMD}_{\Psi^{(\epsilon)}(\mathbf{w})}$ in the convex case, respectively. It is clear that $\text{AdMD}_{\Psi^{(\epsilon)}(\mathbf{w})}$ better captures the sparse structure of $\mathbf{w}$.

VI. Concluding Remark and Future Researches

In this paper, we present a family of adaptive gradient methods associated with mirror maps (AdMD) exploiting the structure of learning problems such as sparsity. We establish convergence rates of AdMD in both convex and nonconvex settings under some standard assumptions in the analysis of classical SGD and adaptive gradient methods. Our convergence results provide suggestions for tuning parameters $\beta_{1,t}, \beta_{2,t}$ of AdMD in practice and boost a better understanding on the convergence behaviors of generic Adam that are extensively studied in existing literature. Below, we discuss several related questions that are worthwhile for future research.

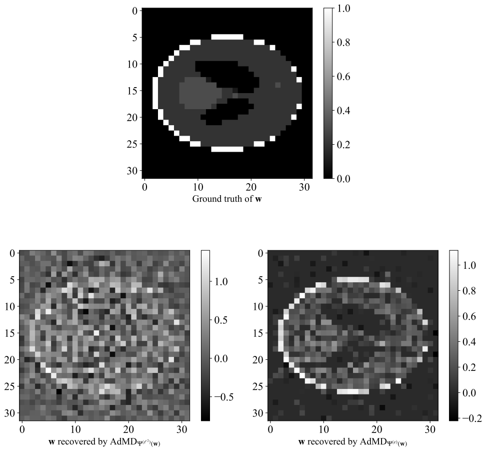


Fig. 8: Ground truth of $\mathbf{w}$ and reconstructions by $\text{AdMD}_{\Psi^{(\ell^2)}(\mathbf{w})}$ and $\text{AdMD}_{\Psi^{(\epsilon)}(\mathbf{w})}$ in the convex case.

First of all, our theoretical analysis provides very useful insights into the practical application of AdMD. However, there is still space for the improvement of theoretical works. Our current convergence results for non-convex AdMD rely on the L-smooth condition. Consequently, they can not directly apply to many deep learning tasks. To our knowledge, in the non-convex setting, even in the case of the vanilla Adam the existing theoretical researches are far from maturity and the L-smooth condition is commonly used to illustrate convergence properties [28], [33]. It remains an open question on how to establish the convergence of AdMD under a relaxed version of the L-smooth condition such as the self-contained condition used in many practical learning problems. See [9], [25].

Secondly, in the general stochastic schemes of mirror maps, there are two typical updates. The first one is related to the Nesterov's primal-dual subgradient method [31], and Xiao's extension [49] and the follow-the-regularized-leader (FTRL) family of algorithms [32], which can be expressed as

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\Omega} \left\{\eta_t \langle \mathbf{m}_t, \mathbf{w}\rangle_{\ell^2} + \eta_t r(\mathbf{w}) + \Psi^{(t)}(\mathbf{w})\right\}, \quad (26)$$

where $r(\mathbf{w})$ is a regularized term and $\Psi^{(t)}$ is a temporal adaptation of maps (changing with t). The second update uses proximal gradient and composite mirror descent [9], [22]

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\Omega} \left\{\eta_t \langle \mathbf{m}_t, \mathbf{w}\rangle_{\ell^2} + \eta_t r(\mathbf{w}) + D_{\Psi^{(t)}}(\mathbf{w}, \mathbf{w}_t)\right\}, \quad (27)$$

making $\mathbf{w}_{t+1}$ close to $\mathbf{w}_t$ by the Bregman divergence associated with $\Psi^{(t)}$. The work in this paper mainly considers (27) without the regularization term $r(\cdot)$. It remains an interesting work to investigate (26), (27) with adaptive gradient methods.

Meanwhile, our AdMD is built on the strongly-convex mirror maps and then it is natural to generalize AdMD to convex and non-differentiable mirror maps [11].

Lastly, it is worth mentioning that, under some Polyak-Łojasiewicz (PL), linear convergence rates have been derived for standard SGD and other gradient methods in non-convex settings. For example, the PL condition can be checked by neural networks with one-hidden layers, ResNets with linear activation and objective functions in matrix factorization [26], [8]. It is yet unknown to us whether the conditions like PL can be incorporated in our convergence analysis. It is also interesting to study whether the linear convergence for AdMD can be derived under such conditions.

References

[1] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations, 2018.*

[2] T. B. Brown , B. Mann, N. Ryder, M. Subbiah and D. Amodei. Language models are few-shot learners. In *Advances in neural information processing systems, 2020.*

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly . An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations, 2021.*

[4] J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. J. Reddi, S. Kumar and S. Sra. Why are adaptive methods good for attention models. In *Neural Information Processing Systems, 2020.*

[5] J. Devlin, M.W.Chang, K. Lee, K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT, 2019*, pages 4171–4186, 2019.

[6] Y. Ida, Y. Fujiwara. Improving generalization performance of adaptive learning rate by switching from block diagonal matrix preconditioning to SGD. In *International Joint Conference on Neural Networks (IJCNN), 2020.*

[7] N. Keskar, R. Socher. Improving generalization performance by switching from Adam to SGD. *CoRR, abs/1712.07628, 2017.*

[8] D. J. Foster, A. Sekhari, and K. Sridharan. Uniform convergence of gradients for non-convex learning and optimization. roc. Adv. Neural Inf. Process. Syst., pp. 8759¨C8770, 2018.

[9] Y. Lei and K. Tang. Stochastic composite mirror descent: Optimal bounds with high probabilities. In *Neural Information Processing Systems, 2018.*

[10] A. Nemirovskij and D. Yudin. Problem Complexity and Method Efficiency in Optimization. *Wiley , New York, 1983.*

[11] H. Bauschke, B. Jerome, M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research.* 42, 2017.

[12] G. Raskutti, S. Mukherjee. The information geometry of mirror descent. *Springer International Publishing.* 2015.

[13] K. Antonakopoulos and P. Mertikopoulos. Adaptive first-order methods revisited: Convex optimization without Lipschitz requirements. *arXiv.2107.08011*, 2021.

[14] L. Balles and P. Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pages 413–422, 2018.

[15] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[16] J. F. Cai, S. J. Osher, and Z. Shen. Linearized Bregman iterations for compressed sensing. *Mathematics of Computation*, 267(78):1515–1536, 2009.

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

[18] J. Chen, D. Zhou, Y. Tang, Z. Yang, Y. Cao, and Q. Gu. Closing the generalization gap of adaptive gradient methods in training deep neural networks. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3267–3275, 2020.

[19] J. S. Chen and S. Pan. An entropy-like proximal algorithm and the exponential multiplier method for convex symmetric cone programming. *Computational Optimization and Applications*, 47(3):477–499, 2010.

[20] A. Défossez, L. Bottou, B. Francis, and U. Nicolas. A simple convergence proof of Adam and Adagrad. In *Transactions on Machine Learning Research*, 2022.

[21] J. C. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[22] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Twenty Third Conference on Computational Learning Theory*, 2010.

[23] A. Borovykh, N. Kantas, P. Parpas, and G. Pavliotis. Stochastic mirror descent for convex optimization with consensus constraints. *SIAM Journal on Applied Dynamical Systems*, 23(3):2208-2241, 2024.

[24] T. D. Nguyen, T. H. Nguyen, A. Ene, and H. Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In Advances in Neural Information Processing Systems, 2023.

[25] B. Wang, Y. Zhang, H. Zhang, Q. Meng, R. Sun, Z. Ma, T. Liu, Z. Luo, W. Chen. Provable Adaptivity of Adam under Non-uniform Smoothness. arXiv:2208.09900 , 2024.

[26] D. Wang and J. Xu. Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 1182¨C1189, 2019.

[27] J. Eckstein. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Mathematics of Operations Research*, 18(1):202–226, 1993.

[28] Z. Guo, Y. Xu, W. Yin, R. Jin, and T. Yang. A novel convergence analysis for algorithms of the Adam family. In *Advances in Neural Information Processing Systems 34(NeurIPS)*, 2021.

[29] I. Fatkhullin, and N. He. Taming nonconvex stochastic mirror descent with general Bregman divergence. In *International Conference on Artificial Intelligence and Statistics*, pages 3493-3501. PMLR, 2024.

[30] N. M. Vural, L. Yu, K. Balasubramanian, S. Volgushev, and M. A. Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Conference on Learning Theory*, pages 65-102. PMLR, 2022.

[31] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221-259, 2009.

[32] E. Hazan, A. Kalai, S. Kale, and A. Agarwal. Logarithmic regret algorithms for online convex optimization. *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2006.

[33] Y. Hong and J. Lin. High probability convergence of Adam under unbounded gradients and affine variance noise. *arXiv:2311.02000*, 2023.

[34] T. Hu, Q. Wu, and D. X. Zhou. Distributed kernel gradient descent algorithm for minimum error entropy principle. *Applied and Computational Harmonic Analysis*, 49:229–256, 2020.

[35] H. Huang, C. Wang, and B. Dong. Nostalgic Adam: Weighting more of the past gradients when designing the adaptive learning rate. In *International Joint Conference on Artificial Intelligence*, 2019.

[36] D. P. Kingma and J. Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[37] Y. Lei and D. X. Zhou. Analysis of online composite mirror descent algorithm. *Neural Computing*, 29:1–36, 2017.

[38] Y. Lei and D. X. Zhou. Convergence of online mirror descent. *Applied and Computational Harmonic Analysis*, 48:342–373, 2020.

[39] H. Li, A. Jadbabaie, and A. Rakhlin. Convergence of Adam under relaxed assumptions. *arXiv:2304.13972*, 2023.

[40] X. Li and F. Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 983–992, 2019.

[41] Y. Liu, Y. Gao, and W. Yin. An improved analysis of stochastic gradient descent with momentum. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 18261–18271, 2020.

[42] I. Loshchilov and F. Hutter. Fixing weight decay regularization in Adam. In *International Conference on Learning Representations*, 2019.

[43] L. Luo, Y. Xiong, Y. Liu, and X. Sun. Adaptive gradient methods with dynamicbound of learning rate. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

[44] P. Ochs, J. Fadili, and T. Brox. Non-smooth non-convex Bregman minimization: Unification and new algorithms. *Journal of Optimization Theory and Applications*, 181(1):244–278, 2019.

[45] A. R. D. Pierro and A. N. Iusem. A relaxed version of Bregman's method for convex programming. *Journal of Optimization Theory and Applications*, 51(3):421–440, 1986.

[46] S. J. Reddi, S. Kale, and S. Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

[47] T. Tieleman and G. Hinton. RMSProp: Divide the gradient by a running average of its recent magnitude. *Coursera*, 2012.

[48] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems 30*, pages 4148–4158, 2017.

[49] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

[50] M. Zaheer, S. J. Reddi, D. S. Sachan, S. Kale, and S. Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems 31(NeurIPS)*, 2018.

[51] D. Zhou, Y. Tang, Z. Yang, Y. Cao, and Q. Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv.1808.05671.*, 2018.

[52] Z. Zhou, Q. Zhang, G. Lu, H. Wang, W. Zhang, and Y. Yu. Adashift: Decorrelation and convergence of adaptive learning rate methods. In *International Conference on Learning Representations*, 2019.

[53] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, 2003.

[54] F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu. A sufficient condition for convergences of Adam and RMSProp. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[55] N. Shi, D. Li, M. Hong, R. Sun. RMSprop converges with proper hyper-parameter. In *International Conference on Learning Representations*, 2021.

[56] P. C. Hansen and M. Saxild-Hansen. AIR tools—a MATLAB package of algebraic iterative reconstruction methods. *Journal of Computational and Applied Mathematics*, 236(8):2167–2178, 2012.

[57] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv.1412.6980.*, 2014.

Appendix

A. Useful Lemmas

We need the following elementary inequalities in the proof, which can be found in [34].

Lemma 2: For $0 < \gamma < 1$, then

$$\frac{1}{1-\gamma}\left[(T+1)^{1-\gamma}-1\right] \leq \sum_{t=1}^{T} t^{-\gamma} \leq \frac{T^{1-\gamma}}{1-\gamma}.$$

For $\gamma > 1$, then $\sum_{t=1}^{T} t^{-\gamma} \leq \frac{1}{\gamma-1}$. For $\gamma = 1$, then $\sum_{t=1}^{T} t^{-1} \leq \log eT$.

We can get the Pythagorean Theorem for coefficient-based Bregman distance $D_{\Psi_{(\mathbf{c})}}(\mathbf{z},\mathbf{x})$.

Lemma 3: For any $\mathbf{c} \in \mathbb{R}_+^d, \mathbf{x},\mathbf{y},\mathbf{z} \in \mathbb{R}^d$, we have

$$\begin{aligned} &D_{\Psi_{(\mathbf{c})}}(\mathbf{z},\mathbf{x}) - D_{\Psi_{(\mathbf{c})}}(\mathbf{z},\mathbf{y}) + D_{\Psi_{(\mathbf{c})}}(\mathbf{x},\mathbf{y}) \\ &= \langle \nabla\Psi_{(\mathbf{c})}(\mathbf{y}) - \nabla\Psi_{(\mathbf{c})}(\mathbf{x}), \mathbf{z}-\mathbf{x}\rangle. \end{aligned} \tag{28}$$

When $\mathbf{c} = \mathbf{1} \in \mathbb{R}^d$, it is the Pythagorean Theorem for the standard Bregman distance $D_\Psi(\mathbf{z},\mathbf{x})$, which can be found in [15], [37]. Its proof is straightforward by writing out the definition of $D_{\Psi_{(\mathbf{c})}}(\mathbf{z},\mathbf{x})$. Here we omit it for simplicity.

Proposition 1: Define $\mathbf{m}_t, \mathbf{v}_t$ by AdMD. Then for $t \in \mathbb{N}$,

$$\mathbf{m}_t = \sum_{s=1}^{t}(1-\beta_{1,s})\left(\prod_{k=s+1}^{t}\beta_{1,k}\right)\mathbf{g}_s \tag{29}$$

and

$$\mathbf{v}_t = \sum_{s=1}^{t}(1-\beta_{2,s})\left(\prod_{k=s+1}^{t}\beta_{2,k}\right)\mathbf{g}_s^2. \tag{30}$$

Assuming $\sup\limits_t \|\mathbf{g}_t\|_* < G$, then

$$\|\mathbf{m}_t\|_* < \left(1-\prod_{k=1}^{t}\beta_{1,k}\right)G,\ t = 1,\cdots,T. \tag{31}$$

In addition, assuming $\sup\limits_t \|\mathbf{g}_t\|_\infty < G$, then

$$\|\mathbf{v}_t\|_\infty < \left(1-\prod_{k=1}^{t}\beta_{2,k}\right)G^2,\ t = 1,\cdots,T. \tag{32}$$

The proof is direct from the definition of $\mathbf{m}_t, \mathbf{v}_t$. We omit it for simplicity.

B. Proofs of main results

***Proof of Theorem 1**:* Note that update in AdMD satisfies $\sqrt{\mathbf{v}_t}\,(\nabla\Psi(\mathbf{w}_{t+1}) - \nabla\Psi(\mathbf{w}_t)) = -\eta_t\mathbf{m}_t$. With the definition of $\Psi_{(\mathbf{c})}$ and $\mathbf{c} = \mathbf{v}_t^{\frac{1}{2}}$, it is rewritten as

$$\begin{aligned} &\nabla\Psi_{(\mathbf{c})}(\mathbf{w}_{t+1}) - \nabla\Psi_{(\mathbf{c})}(\mathbf{w}_t) = -\eta_t\mathbf{m}_t \\ &= -\eta_t\left((1-\beta_{1,t})\mathbf{g}_t + \beta_{1,t}\mathbf{m}_{t-1}\right). \end{aligned} \tag{33}$$

To alleviate the notation burden, we write $\Psi^{(t)}$ as $\Psi_{(\mathbf{c})}$ with $\mathbf{c} = \mathbf{v}_t^{\frac{1}{2}}$ for $t = 1,\cdots$. The associated Bregman distance is denoted by $D_{\Psi^{(t)}}$. Applying Lemma 3 with $\mathbf{c} = \mathbf{v}_t^{\frac{1}{2}}$, $\mathbf{z} = \mathbf{w}^*$, $\mathbf{x} = \mathbf{w}_{t+1}$ and $\mathbf{y} = \mathbf{w}_t$, we get

$$\begin{aligned} &D_{\Psi^{(t)}}(\mathbf{w}^*,\mathbf{w}_{t+1}) - D_{\Psi^{(t)}}(\mathbf{w}^*,\mathbf{w}_t) + D_{\Psi^{(t)}}(\mathbf{w}_{t+1},\mathbf{w}_t) \\ &= \left\langle \nabla\Psi^{(t)}(\mathbf{w}_t) - \nabla\Psi^{(t)}(\mathbf{w}_{t+1}), \mathbf{w}^* - \mathbf{w}_{t+1}\right\rangle. \end{aligned} \tag{34}$$

Putting (33) into (34), we get

$$\begin{aligned} &D_{\Psi^{(t)}}(\mathbf{w}^*,\mathbf{w}_{t+1}) - D_{\Psi^{(t)}}(\mathbf{w}^*,\mathbf{w}_t) + D_{\Psi^{(t)}}(\mathbf{w}_{t+1},\mathbf{w}_t) \\ &\leq \left\langle \nabla\Psi^{(t)}(\mathbf{w}_t) - \nabla\Psi^{(t)}(\mathbf{w}_{t+1}), \mathbf{w}^* - \mathbf{w}_t\right\rangle \\ &\quad + \left\langle \nabla\Psi^{(t)}(\mathbf{w}_t) - \nabla\Psi^{(t)}(\mathbf{w}_{t+1}), \mathbf{w}_t - \mathbf{w}_{t+1}\right\rangle \\ &\leq \eta_t(1-\beta_{1,t})\left(f(\mathbf{w}^*;\xi_t) - f(\mathbf{w}_t;\xi_t)\right) + \\ &\eta_t\beta_{1,t}\langle\mathbf{m}_{t-1},\mathbf{w}^*-\mathbf{w}_t\rangle + \left\langle \mathbf{v}_t^{\frac{1}{4}}\left(\nabla\Psi(\mathbf{w}_t) - \nabla\Psi(\mathbf{w}_{t+1})\right), \mathbf{v}_t^{\frac{1}{4}}(\mathbf{w}_t - \mathbf{w}_{t+1})\right\rangle \\ &\leq \eta_t(1-\beta_{1,t})\left(f(\mathbf{w}^*;\xi_t) - f(\mathbf{w}_t;\xi_t)\right) \\ &\quad + \eta_t\beta_{1,t}\langle\mathbf{m}_{t-1},\mathbf{w}^*-\mathbf{w}_t\rangle + 2^{-1}\sigma\|\mathbf{v}_t^{\frac{1}{4}}(\mathbf{w}_t - \mathbf{w}_{t+1})\|^2 \\ &\quad + 2^{-1}\sigma^{-1}\left\|\mathbf{v}_t^{\frac{1}{4}}\left(\nabla\Psi(\mathbf{w}_t) - \nabla\Psi(\mathbf{w}_{t+1})\right)\right\|_*^2 \\ &= \eta_t(1-\beta_{1,t})\left(f(\mathbf{w}^*;\xi_t) - f(\mathbf{w}_t;\xi_t)\right) + \eta_t\beta_{1,t}\langle\mathbf{m}_{t-1},\mathbf{w}^*-\mathbf{w}_t\rangle \\ &\quad + 2^{-1}\sigma^{-1}\eta_t^2\left\|\mathbf{v}_t^{-\frac{1}{4}}\mathbf{m}_t\right\|_*^2 + 2^{-1}\sigma\left\|\mathbf{v}_t^{\frac{1}{4}}(\mathbf{w}_t - \mathbf{w}_{t+1})\right\|^2, \end{aligned} \tag{35}$$

where the third inequality is obtained by Cauchy inequality. Then, using (8) and dividing η_t on both sides of (35), we get

$$\begin{aligned} &(1-\beta_{1,t})\left(f(\mathbf{w}_t;\xi_t) - f(\mathbf{w}^*;\xi_t)\right) \\ &\leq \left[\eta_t^{-1}D_{\Psi^{(t)}}(\mathbf{w}^*,\mathbf{w}_t) - \eta_{t+1}^{-1}D_{\Psi^{(t+1)}}(\mathbf{w}^*,\mathbf{w}_{t+1})\right] \\ &\quad + \beta_{1,t}\langle\mathbf{m}_{t-1},\mathbf{w}^*-\mathbf{w}_t\rangle \\ &\quad + \left[\eta_{t+1}^{-1}D_{\Psi^{(t+1)}}(\mathbf{w}^*,\mathbf{w}_{t+1}) - \eta_t^{-1}D_{\Psi^{(t)}}(\mathbf{w}^*,\mathbf{w}_{t+1})\right] \\ &\quad + 2^{-1}\sigma^{-1}\eta_t\left\|\mathbf{v}_t^{-\frac{1}{4}}\mathbf{m}_t\right\|_*^2. \end{aligned} \tag{36}$$

For the second term $\beta_{1,t}\mathbf{m}_{t-1}(\mathbf{w}^*-\mathbf{w}_t)$, it is bounded by

$$\begin{aligned} \beta_{1,t}\langle\mathbf{m}_{t-1},\mathbf{w}^*-\mathbf{w}_t\rangle &\leq \beta_{1,t}\|\mathbf{m}_{t-1}\|_{\ell^1}\|\mathbf{w}^*-\mathbf{w}_t\|_\infty \\ &\leq \beta_{1,t}\|\mathbf{m}_{t-1}\|_{\ell^1}D. \end{aligned} \tag{37}$$

For the third term

$$\left[\eta_{t+1}^{-1}D_{\Psi^{(t+1)}}(\mathbf{w}^*,\mathbf{w}_{t+1}) - \eta_t^{-1}D_{\Psi^{(t)}}(\mathbf{w}^*,\mathbf{w}_{t+1})\right],$$

by $\sup\limits_t \|\mathbf{w}_t - \mathbf{w}^*\|_\infty < D$ and the continuity of J, there exists a constant D' such that

$$J(w_j^*) - J(w_{t+1,j}) - J'(w_{t+1,j})(w_j^* - w_{t+1,j}) \leq D',\ i \in [d].$$

Therefore,

$$\begin{aligned} &\eta_{t+1}^{-1}D_{\Psi^{(t+1)}}(\mathbf{w}^*,\mathbf{w}_{t+1}) - \eta_t^{-1}D_{\Psi^{(t)}}(\mathbf{w}^*,\mathbf{w}_{t+1}) \\ &\leq \sum_{j=1}^{d}\left(\eta_{t+1}^{-1}v_{t+1,j}^{\frac{1}{2}} - \eta_t^{-1}v_{t,j}^{\frac{1}{2}}\right)\left(J(w_j^*) - J(w_{t+1,j})\right. \\ &\quad \left. - J'(w_{t+1,j})(w_j^* - w_{t+1,j})\right) \leq \sum_{j=1}^{d}\left|\eta_{t+1}^{-1}v_{t+1,j}^{\frac{1}{2}} - \eta_t^{-1}v_{t,j}^{\frac{1}{2}}\right|D'. \end{aligned}$$

Note that $\mathbf{w}_t$ depends only on $\xi_1,\cdots,\xi_{t-1}$, is independent of ξ_t, then for given ξ, $\mathbb{E}_{\xi_t|\xi_1,\cdots,\xi_{t-1}}f(\mathbf{w}_t;\xi_t) = f(\mathbf{w}_t)$. Putting the estimates above into (36) and taking expectation with respect to $\xi_1, \cdots, \xi_t$ on both sides, we have

$$\begin{aligned}
&(1-\beta_{1,1})\mathbb{E}\left[f(\mathbf{w}_t)-f(\mathbf{w}^*)\right] \\
\le\ & (1-\beta_{1,t})\mathbb{E}\left[f(\mathbf{w}_t)-f(\mathbf{w}^*)\right] \\
\le\ & \mathbb{E}\Big[\left[\eta_t^{-1}D_{\Psi(t)}(\mathbf{w}^*,\mathbf{w}_t)-\eta_{t+1}^{-1}D_{\Psi(t+1)}(\mathbf{w}^*,\mathbf{w}_{t+1})\right] \\
& + \beta_{1,t}\|\mathbf{m}_{t-1}\|_{\ell^1}D + \sum_{j=1}^{d}\left|\eta_{t+1}^{-1}v_{t+1,j}^{\frac{1}{2}}-\eta_t^{-1}v_{t,j}^{\frac{1}{2}}\right|D' \\
& + 2^{-1}\sigma^{-1}\eta_t\left\|\mathbf{v}_t^{-\frac{1}{4}}\mathbf{m}_t\right\|_*^2\Big].
\end{aligned}$$

Dividing $1-\beta_{1,1}$ and $T-t_0$ on both sides and summing up the aforementioned inequality from $t=t_*+1$ to T, the desired conclusion (11) is obtained. The proof is completed. ■

Proof of Theorem 2: We shall prove (12) by Theorem 1. When $\beta_{2,t}=1-\frac{1}{t^2}$, by (30), we have for $i\in[d]$

$$\begin{aligned}
&\eta_{t+1}^{-2}v_{t+1,j}-\eta_t^{-2}v_{t,j} \\
=\ & (t+1)\sum_{s=1}^{t+1}(1-\beta_{2,s})\left(\prod_{k=s+1}^{t+1}\beta_{2,k}\right)g_{s,j}^2 \\
& - t\sum_{s=1}^{t}(1-\beta_{2,s})\left(\prod_{k=s+1}^{t}\beta_{2,k}\right)g_{s,j}^2 \\
\ge\ & (t+1)\sum_{s=1}^{t}(1-\beta_{2,s})\left(\prod_{k=s+1}^{t+1}\beta_{2,k}\right)g_{s,j}^2 \\
& - t\sum_{s=1}^{t}(1-\beta_{2,s})\left(\prod_{k=s+1}^{t}\beta_{2,k}\right)g_{s,j}^2 \\
=\ & (t+1)\beta_{2,t+1}\sum_{s=1}^{t}(1-\beta_{2,s})\left(\prod_{k=s+1}^{t}\beta_{2,k}\right)g_{s,j}^2 \\
& - t\sum_{s=1}^{t}(1-\beta_{2,s})\left(\prod_{k=s+1}^{t}\beta_{2,k}\right)g_{s,j}^2 > 0,
\end{aligned}$$

where the last inequality is obtained by $(t+1)\beta_{2,t+1}=(t+1)(1-\frac{1}{(t+1)^2})\ge t$. So $\eta_{t+1}^{-1}v_{t+1,j}^{\frac{1}{2}}>\eta_t^{-1}v_{t,j}^{\frac{1}{2}}$ and

$$\sum_{t=1}^{T}\sum_{j=1}^{d}\left|\eta_{t+1}^{-1}v_{t+1,j}^{\frac{1}{2}}-\eta_t^{-1}v_{t,j}^{\frac{1}{2}}\right|D' = \eta_{T+1}^{-1}\sum_{j=1}^{d}v_{T+1,j}^{\frac{1}{2}}D'.$$

By the monotonicity of $\|\cdot\|_*$ and Assumption 2, we know that $\|\mathbf{v}_t^{-\frac{1}{4}}\mathbf{m}_t\|_*^2 \le C^{-\frac{1}{2}}c_0^{-1}\|\mathbf{m}_t\|_*^2$ using (10). Collecting the above analysis, we get (12) by Theorem 1.

When $\beta_{2,t}\equiv 1-\frac{1}{T}$, we get (12) similarly. The proof is completed. ■

Proof of Theorem 3: Applying the strong convexity (14) to the second inequality of (35), we have

$$\begin{aligned}
& D_{\Psi(t)}(\mathbf{w}^*,\mathbf{w}_{t+1})-D_{\Psi(t)}(\mathbf{w}^*,\mathbf{w}_t)+D_{\Psi(t)}(\mathbf{w}_{t+1},\mathbf{w}_t) \\
\le\ & \eta_t(1-\beta_{1,t})\left(f(\mathbf{w}^*;\xi_t)-f(\mathbf{w}_t;\xi_t)\right) \\
& - \eta_t(1-\beta_{1,t})\lambda D_\Psi(\mathbf{w}^*,\mathbf{w}_t)+\eta_t\beta_{1,t}\langle\mathbf{m}_{t-1},\mathbf{w}^*-\mathbf{w}_t\rangle \\
& + \mathbf{v}_t^{\frac{1}{4}}\Big\langle\nabla\Psi(\mathbf{w}_t)-\nabla\Psi(\mathbf{w}_{t+1}),\mathbf{v}_t^{\frac{1}{4}}(\mathbf{w}_t-\mathbf{w}_{t+1})\Big\rangle.
\end{aligned}$$

Following the similar proof of (36) in Theorem 1, we get

$$\begin{aligned}
&(1-\beta_{1,t})\left(f(\mathbf{w}_t;\xi_t)-f(\mathbf{w}^*;\xi_t)\right) \\
\le\ & \left[\eta_{t-1}^{-1}D_{\Psi(t-1)}(\mathbf{w}^*,\mathbf{w}_t)-\eta_t^{-1}D_{\Psi(t)}(\mathbf{w}^*,\mathbf{w}_{t+1})\right] \\
& + \beta_{1,t}\|\mathbf{m}_{t-1}\|_{\ell^1}D\ + 2\sigma^{-1}C^{-1}c_0^{-1}\eta_t\|\mathbf{m}_t\|_*^2 + \Big[\eta_t^{-1}D_{\Psi(t)}(\mathbf{w}^*,\mathbf{w}_t) \\
& - \eta_{t-1}^{-1}D_{\Psi(t-1)}(\mathbf{w}^*,\mathbf{w}_t)-(1-\beta_{1,t})\lambda D_\Psi(\mathbf{w}^*,\mathbf{w}_t)\Big]. \qquad (38)
\end{aligned}$$

When $\beta_{2,t}=1-\frac{1}{t^2}$, by $\sup_t\|g_t\|_\infty\le G$, we have for any $t\le T$

$$t v_{t,j}^{\frac{1}{2}}-(t-1)v_{t-1,j}^{\frac{1}{2}} = \sqrt{t^2(1-t^{-2})v_{t-1,j}+g_{t,i}^2}-(t-1)v_{t-1,j}^{\frac{1}{2}} \le 2G.$$

It shows the last term of the above inequality is bounded by

$$\begin{aligned}
& \eta_t^{-1}D_{\Psi(t)}(\mathbf{w}^*,\mathbf{w}_t)-\eta_{t-1}^{-1}D_{\Psi(t-1)}(\mathbf{w}^*,\mathbf{w}_t)-(1-\beta_{1,t})\lambda D_\Psi(\mathbf{w}^*,\mathbf{w}_t) \\
\le\ & \eta_t^{-1}D_{\Psi(t)}(\mathbf{w}^*,\mathbf{w}_t)-\eta_{t-1}^{-1}D_{\Psi(t-1)}(\mathbf{w}^*,\mathbf{w}_t)-(1-c)\lambda D_\Psi(\mathbf{w}^*,\mathbf{w}_t) \\
\le\ & \sup_{i\in[d]}\{\eta^{-1}t v_{t,j}^{\frac{1}{2}}-\eta^{-1}(t-1)v_{t-1,j}^{\frac{1}{2}}-(1-c)\lambda\}D_\Psi(\mathbf{w}^*,\mathbf{w}_t) \\
\le\ & \sup_{i\in[d]}\left\{2\eta^{-1}G-(1-c)\lambda\right\}D_\Psi(\mathbf{w}^*,\mathbf{w}_t)=0.
\end{aligned}$$

It implies that

$$\begin{aligned}
&(1-\beta_{1,t})\left(f(\mathbf{w}_t;\xi_t)-f(\mathbf{w}^*;\xi_t)\right) \\
\le\ & \left[\eta_{t-1}^{-1}D_{\Psi(t-1)}(\mathbf{w}^*,\mathbf{w}_t)-\eta_t^{-1}D_{\Psi(t)}(\mathbf{w}^*,\mathbf{w}_{t+1})\right] \\
& + \beta_{1,t}\|\mathbf{m}_{t-1}\|_{\ell^1}D + 2\sigma^{-1}C^{-1}c_0^{-1}\eta_t\|\mathbf{m}_t\|_*^2 .
\end{aligned}$$

Following the proof procedure of Theorem 1, we take expectation on the aforementioned inequality with respect to $\xi_1,\cdots,\xi_t$, and get

$$\begin{aligned}
&(1-\beta_{1,1})\mathbb{E}\left[f(\mathbf{w}_t)-f(\mathbf{w}^*)\right] \le (1-\beta_{1,t})\mathbb{E}\left[f(\mathbf{w}_t)-f(\mathbf{w}^*)\right] \\
\le\ & \mathbb{E}\Big[\left[\eta_t^{-1}D_{\Psi(t)}(\mathbf{w}^*,\mathbf{w}_t)-\eta_{t+1}^{-1}D_{\Psi(t+1)}(\mathbf{w}^*,\mathbf{w}_{t+1})\right] \\
& + \beta_{1,t}\|\mathbf{m}_{t-1}\|_{\ell^1}D + 2\sigma^{-1}C^{-1}c_0^{-1}\eta_t\|\mathbf{m}_t\|_*^2\Big] \\
\le\ & \mathbb{E}\Big[\left[\eta_t^{-1}D_{\Psi(t)}(\mathbf{w}^*,\mathbf{w}_t)-\eta_{t+1}^{-1}D_{\Psi(t+1)}(\mathbf{w}^*,\mathbf{w}_{t+1})\right] \\
& + ct^{-1}\|\mathbf{m}_{t-1}\|_{\ell^1}D + 2\sigma^{-1}\eta C^{-\frac{1}{2}}c_0^{-1}t^{-1}\|\mathbf{m}_t\|_*^2\Big].
\end{aligned}$$

We divide $1-\beta_{1,1}$ and $T-t_0$ on both sides and sum up the aforementioned inequality from $t=t_*+1$ to T. The desired conclusion (15) holds.

When $\beta_{2,t}\equiv 1-\frac{1}{T}$, we have for any $t\le T$

$$\begin{aligned}
& t v_{t,j}^{\frac{1}{2}}-(t-1)v_{t-1,j}^{\frac{1}{2}} = v_{t,j}^{\frac{1}{2}}+(t-1)\left(v_{t,j}^{\frac{1}{2}}-v_{t-1,j}^{\frac{1}{2}}\right) \\
=\ & v_{t,j}^{\frac{1}{2}}+\frac{(t-1)(v_{t,j}-v_{t-1,j})}{v_{t,j}^{\frac{1}{2}}+v_{t-1,j}^{\frac{1}{2}}} = v_{t,j}^{\frac{1}{2}}+\frac{(t-1)T^{-1}(g_{t,j}^2-v_{t-1,j})}{v_{t,j}^{\frac{1}{2}}+v_{t-1,j}^{\frac{1}{2}}} \\
\le\ & G+C^{-\frac{1}{2}}c_0^{-1}G^2.
\end{aligned}$$

Similarly, by $\eta=\frac{G+C^{-\frac{1}{2}}c_0^{-1}G^2}{(1-c)\lambda}$, we have

$$\eta_t^{-1}D_{\Psi(t)}(\mathbf{w}^*,\mathbf{w}_t)-\eta_{t-1}^{-1}D_{\Psi(t-1)}(\mathbf{w}^*,\mathbf{w}_t)\le(1-\beta_{1,t})\lambda D_\Psi(\mathbf{w}^*,\mathbf{w}_t).$$

Following the same proof procedures in the above case of $\beta_{2,t}=1-\frac{1}{t^2}$, we derive the same convergence order in (15). The proof is finished. ■

Proof of Lemma 1: Note that

$$
\begin{aligned}
&\mathbf{z}_t - \nabla f(\mathbf{w}_t) = \beta_{1,t}\left[\mathbf{z}_{t-1} - \nabla f(\mathbf{w}_{t-1})\right] \\
&\quad + \beta_{1,t}\left[\nabla f(\mathbf{w}_{t-1}) - \nabla f(\mathbf{w}_t)\right] + (1-\beta_{1,t})\left[g_t - \nabla f(\mathbf{w}_t)\right]
\end{aligned}
$$

and $\mathbf{z}_{t-1}, \mathbf{w}_t$ depend on $\xi_1, \cdots, \xi_{t-1}$ (independent of ξ_t).

Hence, we get $\mathbb{E}_{\xi_t}\left[g_t - \nabla f(\mathbf{w}_t)\right] = 0$, and by the Lipschitz continuity of ∇f,

$$
\begin{aligned}
&\mathbb{E}_{\xi_t}\left[\|\mathbf{z}_t - \nabla f(\mathbf{w}_t)\|_*^2\right] \\
&= \mathbb{E}_{\xi_t}\left[\beta_{1,t}^2\|\left(\mathbf{z}_{t-1} - \nabla f(\mathbf{w}_{t-1})\right) + \left(\nabla f(\mathbf{w}_{t-1}) - \nabla f(\mathbf{w}_t)\right)\|_*^2\right] \\
&\quad + (1-\beta_{1,t})^2\mathbb{E}_{\xi_t}\left[\|g_t - \nabla f(\mathbf{w}_t)\|_*^2\right] \\
&\le \beta_{1,t}\|\mathbf{z}_{t-1} - \nabla f(\mathbf{w}_{t-1})\|_*^2 + \frac{\beta_{1,t}^2}{1-\beta_{1,t}}\|\nabla f(\mathbf{w}_{t-1}) - \nabla f(\mathbf{w}_t)\|_*^2 \\
&\quad + \delta^2(1-\beta_{1,t})^2(1 + c_0\|\nabla f(\mathbf{w}_t)\|_*^2) \\
&\le \beta_{1,t}\|\mathbf{z}_{t-1} - \nabla f(\mathbf{w}_{t-1})\|_*^2 + \frac{L^2}{1-\beta_{1,t}}\|\mathbf{w}_{t-1} - \mathbf{w}_t\|^2 \\
&\quad + \delta^2(1-\beta_{1,t})^2(1 + c_0\|\nabla f(\mathbf{w}_t)\|_*^2) \\
&\le \beta_{1,t}\|\mathbf{z}_{t-1} - \nabla f(\mathbf{w}_{t-1})\|_*^2 + \frac{4L^2}{(1-\beta_{1,t})\sigma^2} \\
&\quad + \delta^2(1-\beta_{1,t})^2(1 + c_0\|\nabla f(\mathbf{w}_t)\|_*^2) \\
&\le \beta_{1,t}\|\mathbf{z}_{t-1} - \nabla f(\mathbf{w}_{t-1})\|_*^2 + \frac{4c_3^{-1}L^2\|\mathbf{v}_{t-1}^{\frac{1}{2}}(\nabla\Phi(\mathbf{w}_t) - \nabla\Phi(\mathbf{w}_{t-1}))\|_*^2}{(1-\beta_{1,t})\sigma^2} \\
&\quad + \delta^2(1-\beta_{1,t})^2(1 + c_0\|\nabla f(\mathbf{w}_t)\|_*^2) \\
&= \beta_{1,t}\|\mathbf{z}_{t-1} - \nabla f(\mathbf{w}_{t-1})\|_*^2 + \frac{4c_3^{-1}L^2\eta_{t-1}^2}{(1-\beta_{1,t})\sigma^2}\|\mathbf{m}_{t-1}\|_*^2 \\
&\quad + \delta^2(1-\beta_{1,t})^2(1 + c_0\|\nabla f(\mathbf{w}_t)\|_*^2),
\end{aligned}
$$

where the penultimate inequality is obtained by the monotonicity of $\|\cdot\|_*$. The proof is completed. ■

Proof of Theorem 4: According to the form (6), we know that $\nabla\Psi(\mathbf{w}_{t+1}) - \nabla\Psi(\mathbf{w}_t) = (J'(w_{t+1,i}) - J'(w_{t,i}))_{i=1}^d \in \mathbb{R}^d$. Let $\mathbf{a}_t = (a_{t,i})_{i=1}^d := \frac{\mathbf{w}_{t+1} - \mathbf{w}_t}{\nabla\Psi(\mathbf{w}_{t+1}) - \nabla\Psi(\mathbf{w}_t)} = \left(\frac{w_{t+1,i} - w_{t,i}}{J'(w_{t+1,i}) - J'(w_{t,i})}\right)_{i=1}^d$. The σ-strong convexity of Ψ tells us that $0 \le a_{t,i} < 2/\sigma$ for any $t \in \mathbb{N}, i \in [d]$. Meanwhile, by (17), we have $a_{t,i} \ge L_i^{-1}$ for any $t \in \mathbb{N}, i \in [d]$. This together with (30), $c_3 \le v_{t,i} \le c_4$ yields that for $t \ge t_*$,

$$
c_l < a_{t,i}v_{t,i}^{-\frac{1}{2}} \le c_u, \ i \in [d]. \tag{39}
$$

In the rest of proof, let $\|\cdot\|$ denote $\|\cdot\|_{\ell^2}$ for simplicity. Applying (18) with $g = f, \mathbf{u} = \mathbf{w}_{t+1}, \mathbf{v} = \mathbf{w}_t$, using $\mathbf{v}_t^{\frac{1}{2}}(\nabla\Psi(\mathbf{w}_{t+1}) - \nabla\Psi(\mathbf{w}_t)) = -\eta_t\mathbf{m}_t$ again, we have

$$
\begin{aligned}
&f(\mathbf{w}_{t+1}) \le f(\mathbf{w}_t) + \langle\nabla f(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \frac{L\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2}{2} \\
&\le f(\mathbf{w}_t) + \langle\nabla f(\mathbf{w}_t), \mathbf{a}_t(\nabla\Psi(\mathbf{w}_{t+1}) - \nabla\Psi(\mathbf{w}_t))\rangle + \frac{L\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2}{2} \\
&= f(\mathbf{w}_t) + \langle\mathbf{a}_t^{\frac{1}{2}}\mathbf{v}_t^{-\frac{1}{4}}\nabla f(\mathbf{w}_t), \mathbf{a}_t^{\frac{1}{2}}\mathbf{v}_t^{\frac{1}{4}}(\nabla\Psi(\mathbf{w}_{t+1}) - \nabla\Psi(\mathbf{w}_t))\rangle \\
&\quad + \frac{L\|\mathbf{a}_t(\nabla\Psi(\mathbf{w}_{t+1}) - \nabla\Psi(\mathbf{w}_t))\|^2}{2} \\
&= f(\mathbf{w}_t) - \eta_t\langle\mathbf{a}_t^{\frac{1}{2}}\mathbf{v}_t^{-\frac{1}{4}}\nabla f(\mathbf{w}_t), \mathbf{a}_t^{\frac{1}{2}}\mathbf{v}_t^{-\frac{1}{4}}\mathbf{m}_t\rangle \\
&\quad + \frac{L\|\mathbf{a}_t(\nabla\Psi(\mathbf{w}_{t+1}) - \nabla\Psi(\mathbf{w}_t))\|^2}{2} \\
&= f(\mathbf{w}_t) + \frac{\eta_t}{2}\|\mathbf{a}_t^{\frac{1}{2}}\mathbf{v}_t^{-\frac{1}{4}}(\nabla f(\mathbf{w}_t) - \mathbf{m}_t)\|^2 \\
&\quad - \frac{\eta_t}{2}\|\mathbf{a}_t^{\frac{1}{2}}\mathbf{v}_t^{-\frac{1}{4}}\nabla f(\mathbf{w}_t)\|^2 - \frac{\eta_t}{2}\|\mathbf{a}_t^{\frac{1}{2}}\mathbf{v}_t^{-\frac{1}{4}}\mathbf{m}_t\|^2 \\
&\quad + \frac{L\|\mathbf{a}_t\mathbf{v}_t^{-\frac{1}{2}}\mathbf{v}_t^{\frac{1}{2}}(\nabla\Psi(\mathbf{w}_{t+1}) - \nabla\Psi(\mathbf{w}_t))\|^2}{2} \\
&\le f(\mathbf{w}_t) + \frac{\eta_t c_u}{2}\|\nabla f(\mathbf{w}_t) - \mathbf{m}_t\|^2 \\
&\quad - \frac{\eta_t c_l}{2}\|\nabla f(\mathbf{w}_t)\|^2 - \frac{\eta_t c_l}{2}\|\mathbf{m}_t\|^2 + \frac{c_u^2 L\eta_t^2}{2}\|\mathbf{m}_t\|^2 \\
&\le f(\mathbf{w}_t) + \frac{\eta_t c_u}{2}\|\nabla f(\mathbf{w}_t) - \mathbf{m}_t\|^2 - \frac{\eta_t c_l}{2}\|\nabla f(\mathbf{w}_t)\|^2 - \frac{\eta_t c_l}{4}\|\mathbf{m}_t\|^2,
\end{aligned}
$$

where the last inequality is obtained by the restriction (21).

Rearranging the above inequality yields

$$
\frac{\eta_t c_l}{2}\|\nabla f(\mathbf{w}_t)\|^2 \le f(\mathbf{w}_t) - f(\mathbf{w}_{t+1}) + \frac{\eta_t c_u}{2}\Delta_t - \frac{\eta_t c_l}{4}\|\mathbf{m}_t\|^2,
$$

where $\Delta_t = \|\nabla f(\mathbf{w}_t) - \mathbf{m}_t\|^2$. Taking expectation on the aforementioned inequality with respect to $\xi_T, \cdots, \xi_1$ and summing it up from $t = t_* + 1$ to T, we get

$$
\begin{aligned}
&\frac{c_l}{2}\sum_{t=t_*+1}^{T}\eta_t\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2] \le \mathbb{E}[f(\mathbf{w}_{t_*+1})] \\
&\quad + \frac{c_u}{2}\sum_{t=t_*+1}^{T}\eta_t\mathbb{E}\left[\Delta_t\right] - \frac{c_l}{4}\sum_{t=t_*+1}^{T}\eta_t\mathbb{E}[\|\mathbf{m}_t\|^2].
\end{aligned} \tag{40}
$$

Case (a): When $\beta_{1,t} \equiv \beta_1 = 1 - T^{-\frac{1}{2}}$, and $\eta_t \equiv \eta := \eta_1 T^{-\frac{1}{2}}$.

Applying (20) with $\mathbf{z}_t = \mathbf{m}_t$ and summing up from $t = t_* + 1$ to T, we get by taking expectation with respect to $\xi_T, \cdots, \xi_1$ in sequence,

$$
\begin{aligned}
\sum_{t=t_*+1}^{T}\mathbb{E}[\Delta_t] \le \sum_{t=t_*+1}^{T}\Bigg(&\beta_1\mathbb{E}[\Delta_{t-1}] + \frac{4c_3^{-1}L^2\eta^2}{(1-\beta_1)\sigma^2}\mathbb{E}\left[\|\mathbf{m}_{t-1}\|^2\right] \\
&+ \delta^2(1-\beta_1)^2(1 + c_0\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2])\Bigg).
\end{aligned}
$$

Note that $\sum_{t=t_0+1}^{T}\beta_1\mathbb{E}[\Delta_{t-1}] \le \beta_1\mathbb{E}[\Delta_{t_*}] + \beta_1\sum_{t=t_*+1}^{T}\mathbb{E}[\Delta_t]$. Then putting it into the inequality above yields

$$
\begin{aligned}
(1-\beta_1)\sum_{t=t_*+1}^{T}\mathbb{E}[\Delta_t] \le \beta_1\mathbb{E}[\Delta_{t_*}] + \sum_{t=t_*+1}^{T}\Bigg(&\frac{4c_3^{-1}L^2\eta^2}{(1-\beta_1)\sigma^2}\mathbb{E}\left[\|\mathbf{m}_{t-1}\|^2\right] \\
&+ \delta^2(1-\beta_1)^2(1 + c_0\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2])\Bigg).
\end{aligned}
$$

Dividing $1-\beta_1$ on both sides of the inequality above and putting it into (40) yield

$$\begin{aligned}&\frac{c_l\eta}{2}\sum_{t=t_*+1}^{T}\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2]\le\mathbb{E}[f(\mathbf{w}_{t_*+1})]-\frac{c_l\eta}{4}\sum_{t=t_*+1}^{T}\mathbb{E}[\|\mathbf{m}_t\|^2]\\&+\frac{c_u\eta}{2(1-\beta_1)}\bigg\{\beta_1\mathbb{E}[\Delta_{t_*}]+\sum_{t=t_*+1}^{T}\Big(\frac{4c_3^{-1}L^2\eta^2}{(1-\beta_1)\sigma^2}\mathbb{E}\left[\|\mathbf{m}_{t-1}\|^2\right]\\&+\delta^2(1-\beta_1)^2(1+c_0\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2])\Big)\bigg\}.\end{aligned}$$

Subtracting $\frac{c_0c_u\delta^2(1-\beta_1)\eta}{2}\sum_{t=t_*+1}^{T}\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2]$ on both sides, we get

$$\begin{aligned}&\Big(\frac{c_l-c_0c_u\delta^2(1-\beta_1)}{2}\Big)\eta\sum_{t=t_*+1}^{T}\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2]\\&\le\mathbb{E}[f(\mathbf{w}_{t_*+1})]+\frac{c_u\eta}{2(1-\beta_1)}\Big\{\beta_1\mathbb{E}[\Delta_{t_*}]\\&+\delta^2(1-\beta_1)^2(T-t_*)+\frac{4c_3^{-1}L^2\eta_1^2}{\sigma^2}\mathbb{E}[\|\mathbf{m}_{t_*}\|^2]\Big\}\\&+\left\{\frac{2c_uc_3^{-1}L^2\eta^3}{(1-\beta_1)^2\sigma^2}-\frac{c_l\eta}{4}\right\}\sum_{t=t_*+1}^{T}\mathbb{E}[\|\mathbf{m}_t\|^2].\end{aligned}\tag{41}$$

When $T\ge(2c_0c_u\delta^2c_l^{-1})^2$, we can check by (21) that

$$\frac{c_l-c_0c_u\delta^2(1-\beta_1)}{2}\ge\frac{c_l}{4},\ \frac{2c_uc_3^{-1}L^2\eta^3}{(1-\beta_1)^2\sigma^2}-\frac{c_l\eta}{4}<0$$

and

$$\begin{aligned}&\frac{c_u\eta}{2(1-\beta_1)}\Big\{\beta_1\mathbb{E}[\Delta_{t_*}]+\delta^2(1-\beta_1)^2(T-t_*)+\frac{4c_3^{-1}L^2\eta_1^2}{\sigma^2}\mathbb{E}[\|\mathbf{m}_{t_*}\|^2]\Big\}\\&\le\frac{c_u\eta_1}{2}\Big(\mathbb{E}[\Delta_{t_*}]+\delta^2+\frac{4c_3^{-1}L^2\eta_1^2}{\sigma^2}\mathbb{E}[\|\mathbf{m}_{t_*}\|^2]\Big).\end{aligned}$$

Putting the estimate above into (41) gives

$$\begin{aligned}&\frac{c_l\eta}{4}\sum_{t=t_*+1}^{T}\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2]\le\mathbb{E}[f(\mathbf{w}_{t_*+1})]\\&+\frac{c_u\eta_1}{2}\Big(\mathbb{E}[\Delta_{t_*}]+\delta^2+\frac{4c_3^{-1}L^2\eta_1^2}{\sigma^2}\mathbb{E}[\|\mathbf{m}_{t_*}\|^2]\Big).\end{aligned}$$

Note that when $T\ge2t_*$, $T^{-\frac{1}{2}}(T-t_*)\ge\frac{1}{2}T^{\frac{1}{2}}$. It implies that $(\frac{c_l\eta}{4})^{-1}(T-t_*)^{-1}\le8c_l^{-1}\eta_1^{-1}T^{-\frac{1}{2}}$ for $T\ge2t_*$. Dividing $\frac{c_l\eta}{4}(T-t_*)$ on both sides of the above inequality shows the desired conclusion (22) with

$$\begin{aligned}C_1:=&8c_l^{-1}\eta_1^{-1}\Big\{\mathbb{E}[f(\mathbf{w}_{t_*+1})]\\&+\frac{c_u\eta_1}{2}\Big(\mathbb{E}[\Delta_{t_*}]+\delta^2+\frac{4c_3^{-1}L^2\eta_1^2}{\sigma^2}\mathbb{E}[\|\mathbf{m}_{t_*}\|^2]\Big)\Big\}.\end{aligned}$$

Case (b): $\beta_{1,t}=1-t^{-\frac{1}{2}}$, and $\eta_t=\eta_1t^{-\frac{1}{2}}$. Applying (20) with $\mathbf{z}_t=\mathbf{m}_t$ and subtracting $\beta_{1,t}\mathbb{E}_{\xi_t}[\Delta_t]$ on both sides, we get

$$\begin{aligned}&(1-\beta_{1,t})\mathbb{E}_{\xi_t}[\Delta_t]\le\beta_{1,t}\left(\Delta_{t-1}-\mathbb{E}_{\xi_t}[\Delta_t]\right)\\&+\frac{4c_3^{-1}L^2\eta_{t-1}^2}{(1-\beta_{1,t})\sigma^2}\|\mathbf{m}_{t-1}\|^2+\delta^2(1-\beta_{1,t})^2(1+c_0[\|\nabla f(\mathbf{w}_t)\|^2]).\end{aligned}$$

Note that $\frac{1}{2}\le\beta_{1,t}\le1$ for $t\ge4$. Then, $(1-\beta_{1,t})/\beta_{1,t}\ge1-\beta_{1,t}=\eta_1^{-1}\eta_t$. So, dividing $\eta_1^{-1}\beta_{1,t}$ on both sides of the inequality above gives

$$\begin{aligned}\eta_t\mathbb{E}_{\xi_t}[\Delta_t]\le{}&\eta_1\left(\Delta_{t-1}-\mathbb{E}_{\xi_t}[\Delta_t]\right)+\frac{8\eta_1c_3^{-1}L^2\eta_{t-1}^2}{(1-\beta_{1,t})\sigma^2}\|\mathbf{m}_{t-1}\|^2\\&+2\eta_1\delta^2(1-\beta_{1,t})^2(1+c_0[\|\nabla f(\mathbf{w}_t)\|^2]).\end{aligned}$$

Summing up from $t=t_*+1$ to T and taking expectation with respect to $\xi_T,\cdots,\xi_1$, we get

$$\begin{aligned}&\sum_{t=t_*+1}^{T}\eta_t\mathbb{E}[\Delta_t]\\&\le\eta_1\sum_{t=t_*+1}^{T}(\mathbb{E}[\Delta_{t-1}]-\mathbb{E}[\Delta_t])+\frac{8\eta_1c_3^{-1}L^2\eta_{t-1}^2}{(1-\beta_{1,t})\sigma^2}\mathbb{E}\left[\|\mathbf{m}_{t-1}\|^2\right]\\&+2\eta_1\delta^2(1-\beta_{1,t})^2(1+c_0\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2])\\&\le\eta_1\mathbb{E}[\Delta_{t_*}]+\sum_{t=t_*+1}^{T}\Big(\frac{8\eta_1c_3^{-1}L^2\eta_{t-1}^2}{(1-\beta_{1,t})\sigma^2}\mathbb{E}\left[\|\mathbf{m}_{t-1}\|^2\right]\\&+2\eta_1\delta^2(1-\beta_{1,t})^2(1+c_0\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2])\Big).\end{aligned}$$

Putting the inequality above into (40) yields

$$\begin{aligned}&\frac{c_l}{2}\sum_{t=t_*+1}^{T}\eta_t\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2]\le\mathbb{E}[f(\mathbf{w}_{t_*+1})]+\frac{c_u\eta_1}{2}\mathbb{E}[\Delta_{t_*}]\\&-\sum_{t=t_*+1}^{T}\Big(\frac{c_l\eta_t}{4}-\frac{4c_u\eta_1c_3^{-1}L^2\eta_t^2}{(1-\beta_{1,t})\sigma^2}\Big)\mathbb{E}[\|\mathbf{m}_t\|^2]\\&+\sum_{t=t_*+1}^{T}c_u\eta_1\delta^2(1-\beta_{1,t})^2(1+c_0\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2])\\&+\frac{4c_u\eta_1c_3^{-1}L^2\eta_{t_*}^2}{(1-\beta_{1,t_*})\sigma^2}\mathbb{E}[\|\mathbf{m}_{t_*}\|^2].\end{aligned}\tag{42}$$

The restriction (23) shows that $\frac{c_l\eta_t}{4}-\frac{4c_u\eta_1c_3^{-1}L^2\eta_t^2}{(1-\beta_{1,t})\sigma^2}>0$. The condition $t\ge t_1:=\max\{t_*,(4c_uc_0\delta^2c_l^{-1})^2\}$ implies that $c_u\eta_1\delta^2(1-\beta_{1,t})^2c_0\le\frac{c_l\eta_t}{4}$. Then, subtracting $\frac{c_l}{4}\sum_{t=t_*+1}^{T}\eta_t\mathbf{E}[\|\nabla f(\mathbf{w}_t)\|^2]$ on both sides of (42) and rearranging it, we get

$$\begin{aligned}&\frac{c_l\eta_T}{4}\sum_{t=t_*+1}^{T}\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2]\le\frac{c_l}{4}\sum_{t=t_*+1}^{T}\eta_t\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2]\\&\le c_2+\sum_{t=t_*+1}^{T}c_u\eta_1\delta^2(1-\beta_{1,t})^2\le c_2+c_u\eta_1\delta^2\log(eT),\end{aligned}$$

where the last inequality is derived by Lemma 2 in the appendix and

$$\begin{aligned}c_2:={}&\mathbb{E}[f(\mathbf{w}_{t_*+1})]+\frac{c_u\eta_1}{2}\mathbb{E}[\Delta_{t_*}]+\frac{4c_u\eta_1c_3^{-1}L^2\eta_{t_*}^2}{(1-\beta_{1,t_*})\sigma^2}\mathbb{E}\|\mathbf{m}_{t_*}\|^2\\&+\sum_{t=t_*+1}^{t_1}\Big(c_u\eta_1\delta^2(1-\beta_{1,t})^2c_0-\frac{c_l\eta_t}{4}\Big)\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2].\end{aligned}$$

Note that $(\frac{c_l\eta_T}{4})^{-1}(T-t_*)^{-1} \leq 8c_l^{-1}\eta_1^{-1}T^{-\frac{1}{2}}$ for $T \geq 2t_*$. Then dividing $\frac{c_l\eta_T}{2}(T-t_*)$ on both sides of the inequality above shows the desired conclusion (24) with

$$C_2 := 8c_l^{-1}\eta_1^{-1}(c_2 + c_u\eta_1\delta^2\log(e)).$$

The proof is finished. ■

Proof of Corollary 1: Note that $\sup\limits_t \|\mathbf{g}_t\|_\infty < G$ guarantees that there exists some $c_4 > 0$ such that $v_{t,i} \leq c_4$ for $i \in [d]$. Meanwhile, by (10), we know that there exists some $c_3 > 0$ such that $v_{t,i} > c_3, i \in [d], t \geq t_*$ when $\beta_{2,t} \equiv 1 - \frac{1}{T}$ or $\beta_{2,t} = 1 - \frac{1}{t^2}$. So, Corollary 1 is a direct application of Theorem 4. The proof is finished. ■

Proof of Example 1: To check assumption (6), we only need to prove for some $\sigma > 0$,

$$\langle \nabla\Psi(\mathbf{u}) - \nabla\Psi(\mathbf{v}), \mathbf{u} - \mathbf{v}\rangle \geq \sigma\|\mathbf{u} - \mathbf{v}\|^2, \ \ \forall \mathbf{u}, \mathbf{v} \in \Omega.$$

For (1), it is easy to derive the strong convexity. For (2), the proofs were given in [15]. For (6), the proofs were given in [37]. For (3), we have

$$\begin{aligned}
&\langle \nabla\Psi^{(IS)}(\mathbf{u}) - \nabla\Psi^{(IS)}(\mathbf{v}), \mathbf{u} - \mathbf{v}\rangle \\
&= \sum_{i=1}^{d}\left(-\frac{1}{u_i} + \frac{1}{v_i}\right)(u_i - v_i) = \sum_{i=1}^{d}\frac{(u_i - v_i)^2}{u_i v_i} \\
&\geq \sum_{i=1}^{d}\frac{2(u_i - v_i)^2}{u_i + v_i} = \sum_{i=1}^{d}\frac{u_i + v_i}{2}\left(\frac{u_i - v_i}{\frac{u_i+v_i}{2}}\right)^2 \\
&\geq \sum_{i=1}^{d}\left(\frac{u_i + v_i}{2}\frac{u_i - v_i}{\frac{u_i+v_i}{2}}\right)^2 = \|\mathbf{u} - \mathbf{v}\|_{\ell^1}^2,
\end{aligned}$$

where the third inequality is obtained by $\frac{1}{u_i v_i} \geq \frac{2}{u_i+v_i}$ for $0 < u_i, v_1 \leq 1$ and the fifth inequality is obtained by the convexity of the quadratic function and $\sum_{i=1}^d u_i = \sum_{i=1}^d v_i = 1$.

For (4), the quadratic gradient of $\Psi^{(FD)}(\mathbf{w})$ on the i-th component is $\frac{1}{w_i} + \frac{1}{1-w_i}$ for $i \in [d]$. Thus, for $0 < w_i < 1$ we get $\frac{1}{w_i} + \frac{1}{1-w_i} \geq 4$. It then follows that

$$\langle \nabla\Psi^{(FD)}(\mathbf{u}) - \nabla\Psi^{(FD)}(\mathbf{v}), \mathbf{u} - \mathbf{v}\rangle \geq 4\|\mathbf{u} - \mathbf{v}\|_{\ell^2}^2.$$

For (5), following the similar method in (4), we can derive the strong convexity. The proof is finished. ■