ZO-AdaBound: An Adaptive Zeroth-Order Optimization Algorithm with a Learning Rate Clip

Albert Chua   Liping Yin

I. Introduction

Gradient-free optimization methods are important for problems where one does not have access to gradients or expressions for the gradient are expensive to compute. To create more reliable solutions to optimization problems in the situations described above, one could use Zeroth-Order (ZO) methods, which mimic first order methods by getting approximations of the full gradient. In modern deep learning, the two most popular optimization algorithms are stochastic gradient descent (SGD) [7] and Adam [3].

However, the first order versions of these algorithms have their flaws. In particular, the main disadvantage of SGD is that the gradient scales uniformly in each direction, which slows down the training speed. Thus, adaptive methods, such as Adam, were proposed to speed up the training process by scaling the gradient using the value of past gradients. However, while methods like Adam tend to work well in the beginning stages of training, they plateau on unseen data [9] because of unstable learning rates towards the end of training.

The authors of [4] propose a ZO version of SGD, ZO-signSGD, and [2] propose a ZO version of Adam, ZO-AdaMM, which have shown good performance on adversarial tasks. However, while both these algorithms are effective for adversarial tasks, the drawbacks of the first order versions of these algorithms are still present in the ZO versions.

In [5], the authors propose AdaBound, which is a modification of Adam. To handle unstable learning rates at the end of training, the authors propose a simple clip of the learning rate in an elementwise manner. That is to say, assume there is an lower bound function $\eta_l(t)$ and a upper bound function $\eta_u(t)$. At any iteration t, the learning rate should be in $[\eta_l(t), \eta_u(t)]$. One can think of Adam as a specific case of AdaBound with $\eta_l(t) = 0$ and $\eta_u(t) = \infty$.

In the experiments in [5], the authors show that AdaBound presents modest improvement over Adam with a variety of different networks, such as shallow feedfoward neural networks, convolutional neural networks, residual architectures, and recurrent nets. This suggests the following question: could one also improve [2] via a learning rate clip?

II. Our Contributions

We propose, ZO-AdaBound, the first zeroth order optimization algorithm using an adaptive and bounded learning rate. For our proposed algorithm, we provide $O(d^2/T)$ convergence greatness, where d is the number of optimization variables and T is the number of iterations, which make this algorithm a good candidate for adversarial attacks on systems with a small number of outputs.

III. Preliminaries

A. ZO Oracle Gradients

We define the ZO-gradient as an estimate of ∇f via a finite difference in a random unit direction $\mathbf{u}$ uniformly drawn from unit sphere and a small step parameter $\mu > 0$:

$$\hat{\nabla} f(\mathbf{x}) = \frac{d}{\mu}(f(\mathbf{x} + \mu\mathbf{u}) - f(\mathbf{x}))\mathbf{u}. \tag{1}$$

B. Projection Under Mahalanobis Distance

The square root, maximum, and division operators are elementwise for the rest of this paper. The main tool for our analysis, and the analysis of [6], [2], [5] is the projection operation under Mahalanobis distance with respect to a matrix $\mathbf{H}$. We define

$$\Pi_{\mathcal{X},\mathbf{H}}(\mathbf{a}) = \text{argmin}_{x\in\mathcal{X}}\|\sqrt{\mathbf{H}}(\mathbf{x} - \mathbf{a})\|_2^2.$$

C. ZO-AdaMM

We consider the following optimization problem:

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) = \mathbb{E}_\xi[f(\mathbf{X}, \xi)], \tag{2}$$

where f is a differentiable function and ξ is a random variable. Let

$$f_\mu(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim U_B}[f(\mathbf{x} + \mu\mathbf{u})], \tag{3}$$

where B is the unit ball in $\mathbb{R}^d$ and U_B is the uniform distribution over B. Using the projection from the previous section, the authors of [2] define the algorithm for ZO-AdaMM to solve (2) as follows:

Algorithm 1 ZO-AdaMM

```
Input: $\mathbf{x}_1 \in \mathcal{X}$, step sizes $\{\alpha_t\}_{t=1}^T$, $\{\beta_{1t}\}_{t=1}^T$, $\beta_2 \in (0, 1]$
Set $\mathbf{m}_0 = 0$, $\mathbf{v}_0 = 0$, and $\hat{\mathbf{v}}_0$
for $t = 1$ to $T$ do
    $\mathbf{g}_t = \hat{\nabla}\mathbf{f}_t(x_t)$
    $\mathbf{m}_t = \beta_{1t}\mathbf{m}_{t-1} + (1 - \beta_{1t})\mathbf{g}_t$
    $\mathbf{v}_t = \beta_2\mathbf{v}_{t-1} + (1 - \beta_2)\mathbf{g}_t^2$
    $\hat{\mathbf{v}}_t = \max(\mathbf{v}_{t-1}, \mathbf{v}_t)$
    $\hat{\mathbf{V}}_t = \text{diag}(\hat{\mathbf{v}}_t)$
    $\mathbf{x}_{t+1} = \Pi_{\mathcal{X},\mathbf{V}_t^{1/2}}(\mathbf{x}_t - \mathbf{V}_t^{-1/2}\mathbf{m}_t)$
end for
```

IV. ZO-AdaBound

We define

$$\mathrm{Clip}(x,\eta_l,\eta_u) := \begin{cases} x & x \in [\eta_l, \eta_u] \\ \eta_l & x < \eta_l \\ \eta_u & x > \eta_u. \end{cases} \tag{4}$$

The general idea is that adding (4) into an adaptive optimizer prevents very low learning rates at the beginning of the training process. Additionally, a learning rate clip also prevents very high learning rates at the end of the training process, which will lead to a more stable training process in later iterations. The following is a modification of Algorithm 1 with a learning rate clip:

Algorithm 2 ZO-AdaBound

Input: $\mathbf{x}_1 \in \mathcal{X}$, step sizes $\{\alpha_t\}_{t=1}^T$, $\{\beta_{1t}\}_{t=1}^T$, β_2, lower bound function η_l, upper bound function η_u

```
Set $\mathbf{m}_0 = 0$, $\mathbf{v}_0 = 0$
for $t = 1$ to $T$ do
    $\mathbf{g}_t = \hat{\nabla}\mathbf{f}_t(x_t)$
    $\mathbf{m}_t = \beta_{1t}\mathbf{m}_{t-1} + (1-\beta_{1t})\mathbf{g}_t$
    $\mathbf{v}_t = \beta_2\mathbf{v}_{t-1} + (1-\beta_2)\mathbf{g}_t^2$
    $\hat{\eta}_t = \mathrm{Clip}(\alpha_t/\sqrt{\mathbf{v}_t}, \eta_l(t), \eta_u(t))$
    $\eta_t = \hat{\eta}_t/\sqrt{t}$ and $\mathbf{N}_t = \mathrm{diag}(\eta_t)$
    $\mathbf{x}_{t+1} = \Pi_{\mathcal{X},\mathbf{N}_t^{-1}}(\mathbf{x}_t - \mathbf{N}_t\mathbf{m}_t)$
end for
```

V. Convergence Analysis

We now analyze the convergence of Algorithm 2 in the case of unconstrained nonconvex optimization and constrained convex optimization. We make the following two assumptions throughout our proofs:

- **A1**: $f_t(\cdot) := f(\cdot;\eta_t)$ has an L_g-Lipschitz gradient.
- **A2**: f_t has κ-bounded stochastic gradient: $\|\nabla f_t(\mathbf{x})\|_\infty \leq \kappa$.

A. Preliminary Lemmas

Lemma 1. *If $\mathcal{X} = \mathbb{R}^d$, we have*

$$\Pi_{\mathcal{X},\mathbf{N}_t^{-1}}(\mathbf{x}_t - \mathbf{N}_t\mathbf{m}_t) = \mathbf{x}_t - \mathbf{N}_t\mathbf{m}_t. \tag{5}$$

Lemma 2 ([2])**.** *The following implications hold:*

- *If f is convex, then f_μ is convex.*
- *If f is L_c-Lipschitz, then f_μ is L_c-Lipschitz. Additionally, we have the pointwise bound*

$$|f_\mu(\mathbf{x}) - f(\mathbf{x})| \leq L_c\mu \qquad \forall \mathbf{x} \in \mathbb{R}^d. \tag{6}$$

- *If f has L_g-Lipschitz gradient, then f_μ is L_g-Lipschitz gradient. Moreover, have the following two bounds for any $x \in \mathbb{R}^d$:*

$$|f_\mu(\mathbf{x}) - f(\mathbf{x})| \leq L_g\mu^2 \tag{7}$$

$$\|f_\mu(\mathbf{x}) - f(\mathbf{x})\|^2 \leq \frac{\mu^2 d^2 L_g^2}{4}. \tag{8}$$

Lemma 3 ([2])**.** *For all $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathbb{E}_{\mathbf{u}}\left[\hat{\nabla} f(\mathbf{x})\right] = \nabla f_\mu(\mathbf{x}). \tag{9}$$

Additionally, if f has L_g-Lipschitz gradient, then

$$\mathbb{E}_{\mathbf{u}}\left[\|\hat{\nabla} f(\mathbf{x})\|_2^2\right] \leq 2d\|\nabla f(\mathbf{x})\|_2^2 + \frac{\mu^2 L_g^2 d^2}{2}. \tag{10}$$

B. Unconstrained Nonconvex Optimization

Since we use a Mahalanobis distance projection instead of a Euclidean projection in Algorithm 2, we define a Mahalanobis based convergence measure as well.

Let $\mathbf{x}^+ = \mathbf{x}_{t+1}$, $\mathbf{x}^- = \mathbf{x}_t$, $\mathbf{g} = \mathbf{m}_t$. Then the projection step of ZO-AdaBound can be rewritten as (see the proof in Lemma 4)

$$\mathbf{x}^+ = \mathrm{argmin}_{\mathbf{x}\in\mathcal{X}}\left\{\langle \mathbf{g}, \mathbf{x}\rangle + \frac{1}{2}\left\|\mathbf{N}_t^{-1/2}\left(\mathbf{x} - \mathbf{x}^-\right)\right\|_2^2\right\}. \tag{11}$$

Consider the gradient mapping given by

$$P_{\mathcal{X},\mathbf{H}}(\mathbf{x}^-, \mathbf{g}) := \mathbf{x}^- - \mathbf{x}^+. \tag{12}$$

A natural interpretation is that one is finding a projected version of $\mathbf{g}$ at the point $\mathbf{x}^-$:

$$\mathbf{x}^+ = \mathbf{x}^- - P_{\mathcal{X},\mathbf{H}}(\mathbf{x}^-, \mathbf{g}). \tag{13}$$

Based on (11) and [2], we consider the following Mahalanobis distance based convergence measure for ZO-AdaBound:

$$\|\mathcal{G}(\mathbf{x}_t)\|_2^2 := \|\mathbf{N}_t^{-1/2} P_{\mathcal{X},\mathbf{N}_t^{-1}}(\mathbf{x}_t, \nabla f(\mathbf{x}_t))\|^2. \tag{14}$$

Lemma 4. *In the particular case of $\mathcal{X} = \mathbb{R}^d$, we have*

$$\|\mathcal{G}(\mathbf{x}_t)\|_2^2 = \|\mathbf{N}_t^{1/2}\nabla f(\mathbf{x}_t)\|^2. \tag{15}$$

Note that this Mahalanobis distance based convergence measure in (15) is the squared norm of the gradient in a linearly transformed coordinate system $\mathbf{y}_t = \mathbf{N}_t^{-1/2}\mathbf{x}_t$.

We now present some lemmas that will be necessary for our convergence analysis.

Lemma 5. *Given a sequence $\{\mathbf{x}_t\}$ from the ZO-AdaBound Algorithm, consider the sequence*

$$\mathbf{z}_t = \mathbf{x}_t + \frac{\beta_1}{1-\beta_1}(\mathbf{x}_t - \mathbf{x}_{t-1}), \tag{16}$$

where we let $\mathbf{x}_0 = \mathbf{x}_1$. For $\beta_{1,t} = \beta_1$ and $\mathcal{X} = \mathbb{R}^d$, the following holds for all $t > 1$:

$$\mathbf{z}_{t+1} - z_t = -\frac{\beta_1}{1-\beta_1}(\mathbf{N}_t - \mathbf{N}_{t-1})\mathbf{m}_{t-1} - \mathbf{N}_t\hat{\mathbf{g}}_t.$$

Additionally,

$$\mathbf{z}_2 - \mathbf{z}_1 = -\mathbf{N}_1\mathbf{m}_1.$$

Lemma 6 ([2])**.** *The following bound holds:*

$$\begin{aligned}\mathbb{E}[f_\mu(\mathbf{z}_{t+1}) - f_\mu(\mathbf{z}_1)] \leq & \sum_{t=1}^T \mathbb{E}\left[\langle \nabla f_\mu(\mathbf{x}_t), \mathbf{z}_{t+1} - \mathbf{z}_t\rangle\right] \\ & + \frac{4L_g + 5L_g\beta_1^2}{2(1-\beta_1)^2}\sum_{t=1}^T \mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right].\end{aligned} \tag{17}$$

Lemma 7. *Assume*

$$\|\hat{g}_t\|_\infty \leq G_{zo}, \forall t \in \{1, \ldots, T\}$$

for some constant G_{zo} and $\mathbf{m}_0 = 0$. Additionally, assume that $\eta_u(t) \leq R_\infty$ for some constant R_∞ and $\sum_{i=1}^{\infty} \frac{1}{i^{1/4}} |\eta_u(i) - \eta_l(i-1)| \leq M$ for some constant M. By the ZO-AdaBound update rule, we have

$$\sum_{t=1}^{T} \mathbb{E}[\langle \nabla f_\mu(\mathbf{x}_t), \mathbf{z}_{t+1} - \mathbf{z}_t \rangle] \leq dM \cdot \mathbb{E}\left[\frac{\kappa G_{zo}}{1-\beta_1} + \kappa^2\right] - \sum_{t=1}^{T} \mathbb{E}\left[\langle \nabla f_\mu(\mathbf{x}_t), \mathbf{N}_t \nabla f_\mu(\mathbf{x}_t) \rangle\right]. \tag{18}$$

Lemma 8. *A modified version of the learning rate bound functions from [5], given by*

$$\eta_u(t) = 1 + \frac{1}{\gamma t}$$

and

$$\eta_l(t) = 1 - \frac{1}{1 + \gamma t}$$

with $\gamma > 0$, satisfy the bound

$$\sum_{i=1}^{\infty} \frac{1}{i^{1/2}} |\eta_u(i) - \eta_l(i-1)| \leq M.$$

Lemma 9. *Assume that $\eta_u(t) \leq R_\infty$ for some constant R_∞. Then ZO-AdaBound yields*

$$\mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2] \leq \frac{d^3 \kappa^2 R_\infty^2}{t}. \tag{19}$$

Proposition 10. *Suppose that **A1** and **A2** hold. Let $\mathcal{X} = \mathbb{R}^d$, $f_\mu(\mathbf{x}) - \min_{\mathbf{x}} f_\mu(\mathbf{x}) \leq D_f$, $\beta_{1,t} = \beta_1$, $\mu = \frac{1}{\sqrt{Td}}$, $\eta_u(t) \leq R_\infty$, and$\sum_{i=1}^{\infty} \frac{1}{i^{1/2}} |\eta_u(i) - \eta_l(i-1)| \leq M$ for some constants R_∞ and M. Then*

$$\begin{aligned} &\mathbb{E}[\|\mathbf{N}_R^{1/2} \nabla f(\mathbf{x}_R)\|^2] \\ &\leq \frac{R_\infty d L_g^2}{2T} + \frac{6dM}{T} \mathbb{E}\left[\frac{\kappa G_{zo}}{1-\beta_1} + \kappa^2\right] \\ &+ \frac{2D_f}{T} + \frac{4L_g + 5L_g\beta_1^2}{(1-\beta_1)^2} \frac{d^2\kappa^2 R_\infty^2}{T}. \end{aligned} \tag{20}$$

where $\mathbf{x}_R$ is picked uniformly from $\{\mathbf{x}_t\}_{t=1}^T$.

Remark 1. Note that ZO-AdaMM has a convergence rate of $O\left(\frac{d}{\sqrt{T}}\right)$ and our formulation of ZO-AdaBound has a convergence rate of $O\left(\frac{d^2}{T}\right)$. In other words, theoretically speaking, our convergence is worse for optimzation problems with a large number of variables, but the number of iterations for convergence to a local minimum should be less.

C. Constrained Convex Optimization

Unlike from the nonconvex case, we measure the convergence of ZO-AdaBound for convex optimization by using the the average regret

$$R_T = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{x}^*)\right] \tag{21}$$

where, like mentioned previously, $f_t(\mathbf{x}_t) = f(\mathbf{x}_t; \boldsymbol{\xi}_t)$, and $\mathbf{x}^*$ is the optimal solution. We provide a regret bound in Proposition 11:

Proposition 11. *Assume that **A1** and **A2** hold. Furthermore, suppose that $\mathcal{X}$ has bounded diameter D_∞, $\beta_{1t} \leq \beta_1$ for all $t = 1, \ldots, T$, $\eta_u(t) \leq R_\infty$ and $\eta_l(t) \geq L_\infty$ for some constants R_∞ and L_∞, and $\frac{t}{\eta_\ell(t)} - \frac{t-1}{\eta_u(t-1)} \leq M$ holds for some constant $M > 0$ all $t = 1, \ldots, T$. Then we have the following regret bound:*

$$\begin{aligned} R_{T,\mu} &:= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f_{t,\mu}(\mathbf{x}_t) - \frac{1}{T}\sum_{t=1}^{T} f_{t,\mu}(\mathbf{x}^*)\right] \\ &\leq \frac{2\sqrt{T}-1}{T} \frac{R_\infty G_{zo}^2}{1-\beta_1} \\ &+ \frac{1}{T} \frac{d \cdot D_\infty^2}{2(1-\beta_{11})} \left(2M(\sqrt{T}-1) + \frac{1}{L_\infty}\right) \\ &+ \frac{1}{T} \frac{D_\infty^2}{2(1-\beta_1)} \sum_{i=1}^{d} \mathbb{E}\left[\sum_{t=1}^{T} \beta_{1,t} \eta_{t,i}^{-1}\right]. \end{aligned} \tag{22}$$

Remark 2. Picking $\beta_{1,t} = O(1/t)$ will yield a $O(T^{-1/2})$ regret bound, which is good enough. This proof could probably be improved, but we do not have enough time to do that, so this will be left for future work.

VI. Conclusions and Future Work

We propose, ZO-AdaBound, the first zeroth order optimization algorithm using an adaptive and bounded learning rate. For this project, we prove convergence for unconstrained nonconvex optimization and constrained convex optimization. There is still theoretical work that we could do for constrained nonconvex optimization. Additionally, it would be enlightening to test ZO-AdaBound's performance on Black-Box Adversarial benchmarks in [2].

References

[1] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*. International Conference on Learning Representations, ICLR, 2019.

[2] Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. *Advances in neural information processing systems*, 32, 2019.

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[4] Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. signsgd via zeroth-order oracle. In *International Conference on Learning Representations*.

[5] Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. *arXiv preprint arXiv:1902.09843*, 2019.

[6] Sashank Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[7] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[8] Pedro Savarese. On the convergence of adabound and its connection to sgd. *arXiv preprint arXiv:1908.04457*, 2019.

[9] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Advances in neural information processing systems*, 30, 2017.

Appendix

A. Proof of Lemma 1

Proof. Notice that

$$\begin{aligned}&\|\mathbf{N}_t^{-1/2}(x - (\mathbf{x}_t - \mathbf{N}_t\mathbf{m}_t))\|^2\\&= \|\mathbf{N}_t^{-1/2}\mathbf{x} - \mathbf{N}_t^{-1/2}\mathbf{x}_t + \mathbf{N}_t^{1/2}\mathbf{m}_t\|^2\\&= \mathbf{x}^T\mathbf{N}_t^{-1}\mathbf{x} - 2\mathbf{x}^T\mathbf{N}_t^{-1}\mathbf{x}_t\\&+ 2\mathbf{x}^T\mathbf{m}_t + \|\mathbf{N}_t^{-1/2}\mathbf{x}_t - \mathbf{N}_t^{1/2}\mathbf{m}_t\|^2.\end{aligned}$$

By convexity, it is clear that the minimizer occurs at the stationary point. By the KKT conditions,

$$\mathbf{N}_t^{-1}\mathbf{x} - 2\mathbf{N}_t^{-1}\mathbf{x}_t + 2\mathbf{m}_t = 0. \tag{23}$$

It follows that

$$\Pi_{\mathcal{X},\mathbf{N}_t^{-1}}(\mathbf{x}_t - \mathbf{N}_t\mathbf{m}_t) = \mathbf{x}_t - \mathbf{N}_t\mathbf{m}_t.$$

□

B. Proof of Lemma 4

Proof. Looking at

$$\|\mathbf{N}_t^{1/2}P_{\mathcal{X},\mathbf{N}_t^{-1}}(\mathbf{x}_t, \nabla f(\mathbf{x}_t))\|^2,$$

we consider the system

$$\begin{cases}P_{\mathcal{X},\mathbf{N}_t^{-1}}(\mathbf{x}_t, \mathbf{g}) = \mathbf{x}_t - \mathbf{x}^+\\ \mathbf{x}^+ = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}}\left\{\langle\nabla f(\mathbf{x}_t), \mathbf{x}\rangle + \dfrac{1}{2}\left\|\mathbf{N}_t^{-1/2}\left(\mathbf{x} - \mathbf{x}_t\right)\right\|_2^2\right\}.\end{cases}$$

First, let's start by solving for $\mathbf{x}^+$. Expanding yields

$$\begin{aligned}&\operatorname{argmin}_{\mathbf{x}\in\mathcal{X}}\left\{\langle\nabla f(\mathbf{x}_t), \mathbf{x}\rangle + \frac{1}{2}\left\|\mathbf{N}_t^{-1/2}\left(\mathbf{x} - \mathbf{x}_t\right)\right\|_2^2\right\}\\&= \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}}\left\{\langle\nabla f(\mathbf{x}_t), \mathbf{x}\rangle + \frac{1}{2}\mathbf{x}^T\mathbf{N}_t^{-1}\mathbf{x} - (\mathbf{x}_t)^T\mathbf{N}_t^{-T}\mathbf{x}\right\}.\end{aligned}$$

Taking a derivative, we see that the stationary point $\mathbf{x}^+$ satisfies

$$\nabla f(\mathbf{x}_t) + \mathbf{N}_t^{-1}\mathbf{x}^+ - \mathbf{N}_t^{-1}\mathbf{x}_t = 0. \tag{24}$$

It follows that

$$\mathbf{x}^+ = \mathbf{x}_t - \mathbf{N}_t\nabla f(\mathbf{x}_t).$$

Now substitute this into the system to get

$$\begin{aligned}P_{\mathcal{X},\mathbf{N}_t^{-1}}(\mathbf{x}_t, \mathbf{g}) &= \mathbf{x}_t - \mathbf{x}^+\\&= \mathbf{x}_t - (\mathbf{x}_t - \mathbf{N}_t\nabla f(\mathbf{x}_t)\\&= \mathbf{N}_t\nabla f(\mathbf{x}_t).\end{aligned}$$

Thus, the result follows. □

C. Proof of Lemma 5

Proof. We follow a similar proof given in [1]. When $\mathcal{X} = \mathbb{R}^n$, we have

$$\begin{aligned}\mathbf{x}_{t+1} &= \Pi_{\mathcal{F},\mathbf{N}_t^{-1}}(\mathbf{x}_t - \mathbf{N}_t\mathbf{m}_t)\\&= \mathbf{x}_t - \mathbf{N}_t\mathbf{m}_t.\end{aligned} \tag{25}$$

By the update rule for ZO-AdaBound, when $t > 1$, we have

$$\begin{aligned}&\mathbf{x}_{t+1} - \mathbf{x}_t\\&= \mathbf{N}_t\mathbf{m}_t\\&= -\mathbf{N}_t(\beta_1\mathbf{m}_{t-1} - (1-\beta_1)\hat{\mathbf{g}}_t)\\&= -\beta_1\mathbf{N}_t\mathbf{N}_{t-1}^{-1}(\mathbf{x}_t - \mathbf{x}_{t-1}) - (1-\beta_1)\mathbf{N}_t\hat{\mathbf{g}}_t\\&= \beta_1(\mathbf{x}_t - \mathbf{x}_{t-1}) + \beta_1(\mathbf{N}_t\mathbf{N}_{t-1}^{-1} - \mathbf{I})(\mathbf{x}_t - \mathbf{x}_{t-1}) - (1-\beta_1)\mathbf{N}_t\hat{\mathbf{g}}_t\\&= \beta_1(\mathbf{x}_t - \mathbf{x}_{t-1}) - \beta_1(\mathbf{N}_t\mathbf{N}_{t-1}^{-1} - \mathbf{I})\mathbf{N}_{t-1}\mathbf{m}_{t-1} - (1-\beta_1)\mathbf{N}_t\hat{\mathbf{g}}_t\\&= \beta_1(\mathbf{x}_t - \mathbf{x}_{t-1}) - \beta_1(\mathbf{N}_t - \mathbf{N}_{t-1})\mathbf{m}_{t-1} - (1-\beta_1)\mathbf{N}_t\hat{\mathbf{g}}_t.\end{aligned}$$

Also notice that we can write

$$\mathbf{x}_{t+1} - \mathbf{x}_t = (1-\beta_1)\mathbf{x}_{t+1} + \beta_1(\mathbf{x}_{t+1} - \mathbf{x}_t) - (1-\beta_1)\mathbf{x}_t.$$

By rearranging and substituting, we get

$$\begin{aligned}&(1-\beta_1)\mathbf{x}_{t+1} + \beta_1(\mathbf{x}_{t+1} - \mathbf{x}_t)\\&= (1-\beta_1)\mathbf{x}_t + \beta_1(\mathbf{x}_t - \mathbf{x}_{t-1})\\&- \beta_1(\mathbf{N}_t - \mathbf{N}_{t-1})\mathbf{m}_{t-1} - (1-\beta_1)\mathbf{N}_t\hat{\mathbf{g}}_t.\end{aligned}$$

Now divide both sides by $1-\beta_1$:

$$\begin{aligned}&\mathbf{x}_{t+1} + \frac{\beta_1}{1-\beta_1}(\mathbf{x}_{t+1} - \mathbf{x}_t)\\&= \mathbf{x}_t + \frac{\beta_1}{1-\beta_1}(\mathbf{x}_t - \mathbf{x}_{t-1})\\&- \frac{\beta_1}{1-\beta_1}(\mathbf{N}_t - \mathbf{N}_{t-1})\mathbf{m}_{t-1} - \mathbf{N}_t\hat{\mathbf{g}}_t.\end{aligned} \tag{26}$$

Define

$$\mathbf{z}_t = \mathbf{x}_t + \frac{\beta_1}{1-\beta_1}(\mathbf{x}_t - \mathbf{x}_{t-1}).$$

Substituting into (26) yields

$$\mathbf{z}_{t+1} = z_t - \frac{\beta_1}{1-\beta_1}(\mathbf{N}_t - \mathbf{N}_{t-1})\mathbf{m}_{t-1} - \mathbf{N}_t\hat{\mathbf{g}}_t.$$

In the case where $t = 1$, we see that $\mathbf{z}_1 = \mathbf{x_1}$. It follows by direct calculation that

$$\mathbf{z}_2 - \mathbf{z}_1 = -\mathbf{N}_1\mathbf{m}_1.$$

□

D. Proof of Lemma 7

Proof. We follow the argument from Lemma 2.3 in [2]. By Lemma 17, we have

$$\begin{aligned}\langle\nabla f_\mu(\mathbf{x}_t), \mathbf{z}_{t+1} - \mathbf{z}_t\rangle &= \left\langle\nabla f_\mu(\mathbf{x}_t), -\frac{\beta_1}{1-\beta_1}\left(\mathbf{N}_t - \mathbf{N}_{t-1}\right)\mathbf{m}_t\right\rangle\\&- \langle\nabla f_\mu(\mathbf{x}_t), \mathbf{N}_t\hat{\mathbf{g}}_t\rangle,\end{aligned} \tag{27}$$

and

$$\begin{aligned}\langle \nabla f_\mu(\mathbf{x}_t), \mathbf{N}_t \hat{\mathbf{g}}_t \rangle = & \langle \nabla f_\mu(\mathbf{x}_t), \mathbf{N}_t \nabla f_\mu(\mathbf{x}_t) \rangle \\ & + \langle \nabla f_\mu(\mathbf{x}_t), \mathbf{N}_{t-1}(\hat{\mathbf{g}}_t - \nabla f_\mu(\mathbf{x}_t)) \rangle \\ & + \langle \nabla f_\mu(\mathbf{x}_t), (\mathbf{N}_t - \mathbf{N}_{t-1}) (\hat{\mathbf{g}}_t - \nabla f_\mu(\mathbf{x}_t)) \rangle. \end{aligned} \tag{28}$$

Substitute (28) into (27) to get

$$\begin{aligned} & \langle \nabla f_\mu(\mathbf{x}_t), \mathbf{z}_{t+1} - \mathbf{z}_t \rangle \\ \leq & \frac{1}{1-\beta_1} \left\langle \nabla f_\mu(\mathbf{x}_t), - \left(\mathbf{N}_t^{1/2} - \mathbf{N}_{t-1}^{1/2} \right) \mathbf{m}_{t-1} \right\rangle \\ & - \langle \nabla f_\mu(\mathbf{x}_t), - (\mathbf{N}_t - \mathbf{N}_{t-1}) \nabla f_\mu(\mathbf{x}_t) \rangle \\ & - \langle \nabla f_\mu(\mathbf{x}_t), \mathbf{N}_t^{1/2} \nabla f_\mu(\mathbf{x}_t) \rangle - \langle \nabla f_\mu(\mathbf{x}_t), \mathbf{N}_{t-1}(\hat{\mathbf{g}}_t - \nabla f_\mu(\mathbf{x}_t)) \rangle. \end{aligned} \tag{29}$$

This is where our proof takes advantage of the learning rate bound and deviates from [2]. Notice that

$$\begin{aligned} & \left\langle \nabla f_\mu(\mathbf{x}_t), - (\mathbf{N}_t - \mathbf{N}_{t-1}) \frac{\mathbf{m}_{t-1}}{1-\beta_1} \right\rangle \\ & - \langle \nabla f_\mu(\mathbf{x}_t), - (\mathbf{N}_t - \mathbf{N}_{t-1}) \nabla f_\mu(\mathbf{x}_t) \rangle \\ & = \left\langle \nabla f_\mu(\mathbf{x}_t), (\mathbf{N}_t - \mathbf{N}_{t-1}) \left(\nabla f_\mu(\mathbf{x}_t) - \frac{\mathbf{m}_{t-1}}{1-\beta_1} \right) \right\rangle \\ & = (\nabla f_\mu(\mathbf{x}_t))^T \left[\text{diag} \left(\mathbf{N}_t - \mathbf{N}_{t-1} \right) \odot \left(\nabla f_\mu(\mathbf{x}_t) - \frac{\mathbf{m}_{t-1}}{1-\beta_1} \right) \right] \\ & \leq \| \nabla f_\mu(\mathbf{x}_t) \|_\infty \left\| \text{diag} \left(\mathbf{N}_t - \mathbf{N}_{t-1} \right) \odot (\nabla f_\mu \left(\mathbf{x}_t) - \frac{\mathbf{m}_{t-1}}{1-\beta_1} \right) \right\|_1 \\ & \leq \| \nabla f_\mu(\mathbf{x}_t) \|_\infty \left(\| \nabla f_\mu(\mathbf{x}_t) \|_\infty + \frac{\| \mathbf{m}_{t-1} \|_\infty}{1-\beta_1} \right) \| \text{diag} \left(\mathbf{N}_t - \mathbf{N}_{t-1} \right) \|_1 . \end{aligned} \tag{30}$$

Since the matrix $\mathbf{N}_t$ is diagonal, denote the entries for $t = 1, \ldots, T$ as $n_{t,i}$ with $i = 1 \ldots, d$ and $t = 1, \ldots, T$. It follows that we can bound (30) by

$$\begin{aligned} & \| \nabla f_\mu(\mathbf{x}_t) \|_\infty \left(\| \nabla f_\mu(\mathbf{x}_t) \|_\infty + \frac{\| \mathbf{m}_{t-1} \|_\infty}{1-\beta_1} \right) \\ & \cdot \| \text{diag} \left(\mathbf{N}_t - \mathbf{N}_{t-1} \right) \|_1 \\ & \leq \| \nabla f_\mu(\mathbf{x}_t) \|_\infty \left(\| \nabla f_\mu(\mathbf{x}_t) \|_\infty + \frac{\| \mathbf{m}_{t-1} \|_\infty}{1-\beta_1} \right) \\ & \times \sum_{i=1}^d |n_{t,i} - n_{t-1,i}| \\ & \leq \left(\frac{\kappa G_{zo}}{1-\beta_1} + \kappa^2 \right) \sum_{i=1}^d |n_{t,i} - n_{t-1,i}| . \end{aligned} \tag{31}$$

Sum t from 1 to T and take expectations to get

$$\begin{aligned} & \sum_{t=1}^T \mathbb{E}[\langle \nabla f_\mu(\mathbf{x}_t), \mathbf{z}_{t+1} - \mathbf{z}_t \rangle] \\ \leq & \mathbb{E} \left[\sum_{t=1}^T \left(\frac{\eta G_{zo}}{1-\beta_1} + \eta^2 \right) \sum_{i=1}^d |n_{t,i} - n_{t-1,i}| \right] \\ & - \sum_{t=1}^T \mathbb{E} \left[\langle \nabla f_\mu(\mathbf{x}_t), \mathbf{N}_t \nabla f_\mu(\mathbf{x}_t) \rangle \right] \\ & - \sum_{t=1}^T \mathbb{E} \left[\langle \nabla f_\mu(\mathbf{x}_t), \mathbf{N}_{t-1}(\hat{\mathbf{g}}_t - \nabla f_\mu(\mathbf{x}_t)) \rangle \right] . \end{aligned} \tag{32}$$

Like in [2], we see that $\mathbb{E}[\hat{\mathbf{g}}_t | \hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_{t-1}] = \nabla f_\mu(\mathbf{x}_t)$ by the assumption that $\mathbb{E}[\hat{\mathbf{g}}_t] = \nabla f_\mu(\mathbf{x}_t)$ and the noise on $\hat{\mathbf{g}}_t$ is independent of $\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_{t-1}$. Thus,

$$\mathbb{E} \left[\langle \nabla f_\mu(\mathbf{x}_t), \mathbf{N}_{t-1}(\hat{\mathbf{g}}_t - \nabla f_\mu(\mathbf{x}_t)) \rangle \right] = 0. \tag{33}$$

It follows that the inequality in (32) is actually

$$\begin{aligned} & \sum_{t=1}^T \mathbb{E}[\langle \nabla f_\mu(\mathbf{x}_t), \mathbf{z}_{t+1} - \mathbf{z}_t \rangle] \\ & \leq \mathbb{E} \left[\sum_{t=1}^T \left(\frac{\eta G_{zo}}{1-\beta_1} + \kappa^2 \right) \sum_{i=1}^d |n_{t,i} - n_{t-1,i}| \right] \\ & - \sum_{t=1}^T \mathbb{E} \left[\langle \nabla f_\mu(\mathbf{x}_t), \mathbf{N}_t \nabla f_\mu(\mathbf{x}_t) \rangle \right] . \end{aligned} \tag{34}$$

Now we simplify (34) by considering the first term. We have

$$\begin{aligned} & \mathbb{E} \left[\sum_{t=1}^T \left(\frac{\kappa G_{zo}}{1-\beta_1} + \kappa^2 \right) \sum_{i=1}^d |n_{t,i} - n_{t-1,i}| \right] \\ & \leq \left(\frac{\eta G_{zo}}{1-\beta_1} + \kappa^2 \right) \mathbb{E} \left[\sum_{i=1}^d \sum_{t=1}^T |n_{t,i} - n_{t-1,i}| \right] \\ & = \left(\frac{\eta G_{zo}}{1-\beta_1} + \kappa^2 \right) \mathbb{E} \left[\sum_{i=1}^d \sum_{t=1}^T t^{-1/2} |\hat{\eta}_{t,i} - \hat{\eta}_{t-1,i}| \right] \\ & \leq \left(\frac{\eta G_{zo}}{1-\beta_1} + \kappa^2 \right) \mathbb{E} \left[\sum_{i=1}^d \sum_{t=1}^T t^{-1/2} |\hat{\eta}_{t,i} - \hat{\eta}_{t-1,i}| \right] . \end{aligned} \tag{35}$$

Now notice that

$$\begin{aligned} \sum_{t=1}^T t^{-1/2} |\hat{\eta}_{t,i} - \hat{\eta}_{t-1,i}| \leq & \sum_{t=1}^T t^{-1/2} |\hat{\eta}_{t,i} - \eta_l(t)| \\ & + \sum_{t=1}^T t^{-1/2} |\eta_u(t) - \hat{\eta}_{t-1,i}| \\ & + \sum_{t=1}^T t^{-1/2} |\eta_u(t) - \eta_l(t)| \\ & \leq 3 \sum_{t=1}^T t^{-1/2} |\eta_u(t) - \eta_l(t)| \\ & \leq 3M. \end{aligned} \tag{36}$$

Hence, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{\kappa G_{zo}}{1-\beta_1}+\kappa^2\right)\sum_{i=1}^{d}|n_{t,i}-n_{t-1,i}|\right] \\
&\quad\le 3dM\cdot\mathbb{E}\left[\frac{\kappa G_{zo}}{1-\beta_1}+\kappa^2\right] \qquad (37)
\end{aligned}$$

and

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}[\langle\nabla f_\mu(\mathbf{x}_t),\mathbf{z}_{t+1}-\mathbf{z}_t\rangle] &\le dM\cdot\mathbb{E}\left[\frac{\kappa G_{zo}}{1-\beta_1}+\kappa^2\right] \\
&\quad -\sum_{t=1}^{T}\mathbb{E}\left[\langle\nabla f_\mu(\mathbf{x}_t),\mathbf{N}_t\nabla f_\mu(\mathbf{x}_t)\rangle\right].
\end{aligned}$$

□

E. Proof of Lemma 8

Proof. We have

$$\begin{aligned}
\sum_{i=1}^{\infty}\frac{1}{i^{1/2}}|\eta_u(i)-\eta_l(i-1)| &= \sum_{i=1}^{\infty}\frac{1}{i^{1/2}}\left|\frac{1}{\gamma i}+\frac{1}{1+\gamma(i-1)}\right| \\
&= \sum_{i=1}^{\infty}\frac{1}{i^{1/2}}\frac{2\gamma i-\gamma+1}{\gamma^2 i^2+\gamma i(1-\gamma)}.
\end{aligned}$$

It is very clear that each term of the series is $O(i^{-3/2})$. Thus, the result follows. □

F. Proof of Lemma 9

Proof. Assume that $\beta_{1,t}=\beta_1$. By the update rule,

$$\begin{aligned}
&\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 \\
&\le \|\mathbf{N}_t\|^2\|\mathbf{m}_t\|^2 \\
&\le \sum_{i=1}^{d}\frac{\hat{\eta}_{t,i}^2}{t}\left((1-\beta_1)\sum_{j=0}^{t-1}\beta_1^{t-j}\hat{\mathbf{g}}_{j,i}\right)^2 \\
&\le dR_\infty^2(1-\beta_1)^2\sum_{i=1}^{d}\left(\sum_{j=0}^{t-1}\frac{\beta_1^{t-j}}{t^{1/2}}\hat{\mathbf{g}}_{j,i}\right)^2 \\
&\le dR_\infty^2(1-\beta_1)^2\sum_{i=1}^{d}\left(\sum_{j=0}^{t-1}\beta_1^{t-j}\right)\left(\sum_{j=0}^{t-1}\frac{\beta_1^{t-j}}{t}\hat{\mathbf{g}}_{j,i}^2\right) \\
&\le dR_\infty^2(1-\beta_1)\sum_{j=0}^{t-1}\sum_{i=1}^{d}\frac{\beta_1^{t-j}}{t}\hat{\mathbf{g}}_{j,i}^2 \\
&\le dR_\infty^2(1-\beta_1)\sum_{j=0}^{t-1}\frac{\beta_1^{t-j}}{t}d\|\hat{\mathbf{g}}_j\|_\infty^2.
\end{aligned}$$

Take the expectation to get the desired result. Unfortunately, unlike [2], this bound depends on the size of the stochastic gradient, which is not ideal. □

G. Proof of Proposition 10

Proof. We have already proven that

$$\begin{aligned}
\mathbb{E}[f_\mu(\mathbf{z}_{t+1})-f_\mu(\mathbf{z}_1)] &\le \sum_{t=1}^{T}\mathbb{E}\left[\langle\nabla f_\mu(\mathbf{x}_t),\mathbf{z}_{t+1}-\mathbf{z}_t\rangle\right] \\
&\quad+\frac{4L_g+5L_g\beta_1^2}{2(1-\beta_1)^2}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right] \\
&\le 3dM\cdot\mathbb{E}\left[\frac{\kappa G_{zo}}{1-\beta_1}+\kappa^2\right] \\
&\quad-\sum_{t=1}^{T}\mathbb{E}\left[\langle\nabla f_\mu(\mathbf{x}_t),\mathbf{N}_t\nabla f_\mu(\mathbf{x}_t)\rangle\right] \\
&\quad+\frac{4L_g+5L_g\beta_1^2}{2(1-\beta_1)^2}\sum_{t=1}^{T}\frac{d^3\kappa^2R_\infty^2}{T}. \qquad (38)
\end{aligned}$$

First, we rewrite

$$\begin{aligned}
&\mathbb{E}\left[\langle\nabla f_\mu(\mathbf{x}_t),\mathbf{N}_t\nabla f_\mu(\mathbf{x}_t)\rangle\right] \qquad (39) \\
&= \mathbb{E}\left[\langle\mathbf{N}_t^{1/2}\nabla f_\mu(\mathbf{x}_t),\mathbf{N}_t^{1/2}\nabla f_\mu(\mathbf{x}_t)\rangle\right] \\
&= \mathbb{E}[\|\mathbf{N}_t^{1/2}\nabla f_\mu(\mathbf{x}_t)\|^2]. \qquad (40)
\end{aligned}$$

We now rearrange, substitute, assume $f_\mu(\mathbf{z}_1)-\min_{\mathbf{z}} f_\mu(\mathbf{z})\le D_f$, and divide by T to get

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{N}_t^{1/2}\nabla f_\mu(\mathbf{x}_t)\|^2] &\le \frac{3dM}{T}\mathbb{E}\left[\frac{\kappa G_{zo}}{1-\beta_1}+\kappa^2\right] \\
&\quad+\frac{D_f}{T}+\frac{4L_g+5L_g\beta_1^2}{2(1-\beta_1)^2}\frac{d^3\kappa^2R_\infty^2}{T}. \qquad (41)
\end{aligned}$$

Now we can choose R uniformly from 1 to T, which yields

$$\begin{aligned}
\mathbb{E}[\|\mathbf{N}_R^{1/2}\nabla f_\mu(\mathbf{x}_R)\|^2] &\le \frac{3dM}{T}\mathbb{E}\left[\frac{\kappa G_{zo}}{1-\beta_1}+\kappa^2\right] \\
&\quad+\frac{D_f}{T}+\frac{4L_g+5L_g\beta_1^2}{2(1-\beta_1)^2}\frac{d^3\kappa^2R_\infty^2}{T}. \qquad (42)
\end{aligned}$$

Finally, since we know that we have the bounding functions from before, it follows by Lemma 2 that

$$\begin{aligned}
\|\mathbf{N}_t^{1/2}(\nabla f_\mu(\mathbf{x})-\nabla f(\mathbf{x}))\|^2 &\le \|\mathbf{N}_t^{1/2}\|^2\frac{\mu^2d^2L_g^2}{4} \\
&\le \frac{R_\infty\mu^2d^2L_g^2}{4}. \qquad (43)
\end{aligned}$$

Now it follows that

$$\begin{aligned}
\mathbb{E}[\|\mathbf{N}_R^{1/2}\nabla f(\mathbf{x}_R)\|^2] &\le 2\mathbb{E}[\|\mathbf{N}_R^{1/2}\nabla f(\mathbf{x}_R)\|^2] \\
&\quad+2\mathbb{E}[\|\mathbf{N}_R^{1/2}(\nabla f_\mu(\mathbf{x}_R)-\nabla f(\mathbf{x}_R))\|^2] \\
&\le \frac{R_\infty\mu^2d^2L_g^2}{2}+\frac{6dM}{T}\mathbb{E}\left[\frac{\kappa G_{zo}}{1-\beta_1}+\kappa^2\right] \\
&\quad+\frac{2D_f}{T}+\frac{4L_g+5L_g\beta_1^2}{(1-\beta_1)^2}\frac{d^3\kappa^2R_\infty^2}{T}.
\end{aligned}$$

By our choice of μ, we get

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{N}_R^{1/2}\nabla f_\mu(\mathbf{x}_R)\|^2] \\
&\le \frac{R_\infty d L_g^2}{2T} + \frac{6dM}{T}\mathbb{E}\left[\frac{\kappa G_{zo}}{1-\beta_1} + \kappa^2\right] \\
&+ \frac{2D_f}{T} + \frac{4L_g + 5L_g\beta_1^2}{(1-\beta_1)^2}\frac{d^3\kappa^2 R_\infty^2}{T}.
\end{aligned}$$

□

H. Proof of Proposition 11

Proof. We follow the proof of Proposition 4 from [2] and Theorem 3 in [8]. From the proof in [2], it follows that

$$\mathbb{E}[f_{t,\mu}(\mathbf{x}_t) - f_{t,\mu}(\mathbf{x}^*)] \le \mathbb{E}\langle \hat{\mathbf{g}}_t, \mathbf{x}_t - \mathbf{x}^*\rangle. \tag{44}$$

One can see that $\Pi_{\mathcal{X},\mathbf{N}_t^1}(x^*) = x^*$ by direct calculation. Using Lemma 4 from [6], we see that

$$\begin{aligned}
\left\|\mathbf{N}_t^{-1/2}(\mathbf{x}_{t+1} - \mathbf{x}^*)\right\|^2 &\le \left\|\mathbf{N}_t^{-1/2}(\mathbf{x}_t - \mathbf{N}_t\mathbf{m}_t - \mathbf{x}^*)\right\|^2 \\
&\le \left\|\mathbf{N}_t^{-1/2}(\mathbf{x}_t - \mathbf{x}^*)\right\|^2 \\
&+ \|\mathbf{N}_t^{1/2}\mathbf{m}_t\|^2 \\
&- 2\beta_{1,t}\langle \mathbf{m}_{t-1}, \mathbf{x}_t - \mathbf{x}^*\rangle \\
&- 2(1-\beta_{1,t})\langle \hat{\mathbf{g}}_t, \mathbf{x}_t - \mathbf{x}^*\rangle. \qquad (45)
\end{aligned}$$

We now rearrange to get

$$\begin{aligned}
\langle \hat{\mathbf{g}}_t, \mathbf{x}_t - \mathbf{x}^*\rangle &\le \left\|\mathbf{N}_t^{-1/2}(\mathbf{x}_t - \mathbf{N}_t\mathbf{m}_t - \mathbf{x}^*)\right\|^2 \\
&= \frac{1}{2(1-\beta_{1,t})}\left\|\mathbf{N}_t^{-1/2}(\mathbf{x}_t - \mathbf{x}^*)\right\|^2 \\
&+ \frac{1}{2(1-\beta_{1,t})}\|\mathbf{N}_t^{-1/2}\mathbf{m}_t\|^2 \\
&- \frac{\beta_{1,t}}{1-\beta_{1,t}}\langle \mathbf{m}_{t-1}, \mathbf{x}_t - \mathbf{x}^*\rangle \\
&- \frac{1}{2(1-\beta_{1,t})}\left\|\mathbf{N}_t^{-1/2}(\mathbf{x}_{t+1} - \mathbf{x}^*)\right\|^2. \qquad (46)
\end{aligned}$$

Now, we see that

$$\begin{aligned}
&\frac{\beta_{1,t}}{1-\beta_{1,t}}\langle \mathbf{m}_{t-1}, \mathbf{x}_t - \mathbf{x}^*\rangle \\
&= \frac{\beta_{1,t}}{1-\beta_{1,t}}\langle \mathbf{N}_t^{1/2}\mathbf{m}_{t-1}, \mathbf{N}_t^{-1/2}(\mathbf{x}_t - \mathbf{x}^*)\rangle \\
&\le \frac{\beta_{1,t}}{2(1-\beta_{1,t})}\left(\left\|\mathbf{N}_t^{1/2}\mathbf{m}_{t-1}\right\|^2 + \|\mathbf{N}_t^{-1/2}(\mathbf{x}_t - \mathbf{x}^*)\|^2\right)
\end{aligned} \tag{47}$$

Thus,

$$\begin{aligned}
&\langle \hat{\mathbf{g}}_t, \mathbf{x}_t - \mathbf{x}^*\rangle \qquad (48)\\
&\le \frac{\left\|\mathbf{N}_t^{-1/2}(\mathbf{x}_t - \mathbf{x}^*)\right\|^2 - \left\|\mathbf{N}_t^{-1/2}(\mathbf{x}_{t+1} - \mathbf{x}^*)\right\|^2}{2(1-\beta_{1,t})} \\
&+ \frac{1}{2(1-\beta_{1,t})}\|\mathbf{N}_t^{1/2}\mathbf{m}_t\|^2 \\
&+ \frac{\beta_{1,t}}{2(1-\beta_{1,t})}\|\mathbf{N}_t^{1/2}\mathbf{m}_{t-1}\|^2 \\
&+ \frac{\beta_{1,t}}{2(1-\beta_{1,t})}\|\mathbf{N}_t^{-1/2}(\mathbf{x}_t - \mathbf{x}^*)\|^2. \qquad (49)
\end{aligned}$$

Similar to [2], we can sum from 1 to T, take an expectation, and bound:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\langle \hat{\mathbf{g}}_t, \mathbf{x}_t - \mathbf{x}^*\rangle\right] \\
&\le \sum_{t=1}^{T}\mathbb{E}\left[\frac{\left\|\mathbf{N}_t^{-1/2}(\mathbf{x}_t - \mathbf{x}^*)\right\|^2 - \left\|\mathbf{N}_t^{-1/2}(\mathbf{x}_{t+1} - \mathbf{x}^*)\right\|^2}{2(1-\beta_{1,t})}\right] \\
&+ \frac{1}{2(1-\beta_1)}\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{N}_t^{1/2}\mathbf{m}_t\|^2\right] \\
&+ \frac{\beta_1}{2(1-\beta_1)}\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{N}_t^{1/2}\mathbf{m}_{t-1}\|^2\right] \\
&+ \frac{\beta_1}{2(1-\beta_1)}\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{N}_t^{-1/2}(\mathbf{x}_t - \mathbf{x}^*)\|^2\right] \qquad (50)\\
&:= I_1 + I_2 + I_3 + I_4.
\end{aligned}$$

For $I_2 + I_3$, we use Lemma 3 in [5] to get

$$I_2 + I_3 \le (2\sqrt{T} - 1)\frac{R_\infty G_{zo}^2}{1-\beta_1}. \tag{51}$$

For I_1, we see that

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\frac{\left\|\mathbf{N}_t^{-1/2}(\mathbf{x}_t - \mathbf{x}^*)\right\|^2 - \left\|\mathbf{N}_t^{-1/2}(\mathbf{x}_{t+1} - \mathbf{x}^*)\right\|^2}{2(1-\beta_{1,t})}\right] \\
&\le \sum_{t=1}^{T}\mathbb{E}\left[\frac{\left\|\mathbf{N}_t^{-1/2}(\mathbf{x}_t - \mathbf{x}^*)\right\|^2 - \left\|\mathbf{N}_t^{-1/2}(\mathbf{x}_{t+1} - \mathbf{x}^*)\right\|^2}{2(1-\beta_{1,t})}\right] \\
&\le \mathbb{E}\left[\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{n_{t,i}^{-1}(\mathbf{x}_{t,i} - \mathbf{x}_i^*)^2 - n_{t,i}^{-1}(\mathbf{x}_{t+1,i} - \mathbf{x}_i^*)^2}{2(1-\beta_{1,t})}\right] \\
&\le \sum_{i=1}^{d}\mathbb{E}\left[\sum_{t=1}^{T}\frac{n_{t,i}^{-1}(\mathbf{x}_{t,i} - \mathbf{x}_i^*)^2 - n_{t,i}^{-1}(\mathbf{x}_{t+1,i} - \mathbf{x}_i^*)^2}{2(1-\beta_{1,t})}\right] \qquad (52)
\end{aligned}$$

Now consider the term inside the expectation. By reindexing, we get

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\frac{n_{t,i}^{-1}(\mathbf{x}_{t,i}-\mathbf{x}_i^*)^2 - n_{t,i}^{-1}(\mathbf{x}_{t+1,i}-\mathbf{x}_i^*)^2}{2(1-\beta_{1,t})}\right)\\
&=\sum_{t=1}^{T}\frac{n_{t,i}^{-1}(\mathbf{x}_{t,i}-\mathbf{x}_i^*)^2}{2(1-\beta_{1,t})}-\sum_{t=1}^{T}\frac{n_{t,i}^{-1}(\mathbf{x}_{t+1,i}-\mathbf{x}_i^*)^2}{2(1-\beta_{1,t})}\\
&=\sum_{t=1}^{T}\frac{n_{t,i}^{-1}(\mathbf{x}_{t,i}-\mathbf{x}_i^*)^2}{2(1-\beta_{1,t})}-\sum_{t=2}^{T}\frac{n_{t-1,i}^{-1}(\mathbf{x}_{t,i}-\mathbf{x}_i^*)^2}{2(1-\beta_{1,t-1})}\\
&\leq\sum_{t=1}^{T}\frac{n_{t,i}^{-1}(\mathbf{x}_{t,i}-\mathbf{x}_i^*)^2}{2(1-\beta_{1,t})}-\sum_{t=2}^{T}\frac{n_{t-1,i}^{-1}(\mathbf{x}_{t,i}-\mathbf{x}_i^*)^2}{2(1-\beta_{1,t-1})}\\
&=\frac{n_{1,i}^{-1}}{2(1-\beta_1)}(\mathbf{x}_{1,i}-\mathbf{x}_i^*)^2+\sum_{t=2}^{T}(n_{t,i}^{-1}-n_{t-1,i}^{-1})\frac{(\mathbf{x}_{t,i}-\mathbf{x}_i^*)^2}{2(1-\beta_{1,t-1})}\\
&=\frac{n_{1,i}^{-1}}{2(1-\beta_1)}(\mathbf{x}_{1,i}-\mathbf{x}_i^*)^2+\sum_{t=2}^{T}\left[\frac{\sqrt{t}}{\hat{\eta}_{t,i}}-\frac{\sqrt{t-1}}{\hat{\eta}_{t-1,i}}\right]\frac{(\mathbf{x}_{t,i}-\mathbf{x}_i^*)^2}{2(1-\beta_{1,t-1})}.
\end{aligned} \tag{53}$$

Now consider the second term in last line of (53):

$$\begin{aligned}
&\sum_{t=2}^{T}\left[\frac{\sqrt{t}}{\hat{\eta}_{t,i}}-\frac{\sqrt{t-1}}{\hat{\eta}_{t-1,i}}\right]\frac{(\mathbf{x}_{t,i}-\mathbf{x}_i^*)^2}{2(1-\beta_{1,t-1})}\\
&=\sum_{t=2}^{T}\frac{1}{\sqrt{t}}\left[\frac{t}{\hat{\eta}_{t,i}}-\frac{\sqrt{t}\sqrt{t-1}}{\hat{\eta}_{t-1,i}}\right]\frac{(\mathbf{x}_{t,i}-\mathbf{x}_i^*)^2}{2(1-\beta_{1,t-1})}\\
&\leq\sum_{t=2}^{T}\frac{1}{\sqrt{t}}\left[\frac{t}{\hat{\eta}_{t,i}}-\frac{t-1}{\hat{\eta}_{t-1,i}}\right]\frac{(\mathbf{x}_{t,i}-\mathbf{x}_i^*)^2}{2(1-\beta_{1,t-1})}\\
&\leq\sum_{t=2}^{T}\frac{1}{\sqrt{t}}\left[\frac{t}{\eta_l(t)}-\frac{t-1}{\eta_u(t-1)}\right]\frac{(\mathbf{x}_{t,i}-\mathbf{x}_i^*)^2}{2(1-\beta_{1,t-1})}\\
&\leq\sum_{t=2}^{T}\frac{M}{\sqrt{t}}\frac{(\mathbf{x}_{t,i}-\mathbf{x}_i^*)^2}{2(1-\beta_{1,t-1})}\\
&\leq\frac{D_\infty^2}{2(1-\beta_{11})}\left(\eta_{t-1,i}^{-1}+2M(\sqrt{T}-1)\right).
\end{aligned} \tag{54}$$

It now follows that we can sum over the components to get

$$\begin{aligned}
I_1 &\leq \frac{D_\infty^2}{2(1-\beta_{11})}\left(2dM(\sqrt{T}-1)+\sum_{i=1}^{d}\mathbb{E}[\eta_{1,i}^{-1}]\right)\\
&\leq \frac{d\cdot D_\infty^2}{2(1-\beta_{11})}\left(2M(\sqrt{T}-1)+\frac{1}{L_\infty}\right).
\end{aligned} \tag{55}$$

For I_4, we start by using an argument like [8] to get

$$I_2 \leq \frac{D_\infty^2}{2(1-\beta_1)}\sum_{i=1}^{d}\mathbb{E}\left[\sum_{t=1}^{T}\beta_{1,t}\eta_{t,i}^{-1}\right]. \tag{56}$$

□