OPT2022: 14th Annual Workshop on Optimization for Machine Learning

TiAda: A Time-scale Adaptive Algorithm for Nonconvex Minimax Optimization

Xiang Li XIANG.LI@INF.ETHZ.CH
Junchi Yang JUNCHI.YANG@INF.ETHZ.CH
Niao He NIAO.HE@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Switzerland

Abstract

Adaptive gradient methods have shown their ability to adjust the stepsizes on the fly in a parameter-agnostic manner, and empirically achieve faster convergence for solving minimization problems. When it comes to nonconvex minimax optimization, however, current convergence analyses of gradient descent ascent (GDA) combined with adaptive stepsizes require careful tuning of hyper-parameters and the knowledge of problem-dependent parameters. Such a discrepancy arises from the primal-dual nature of minimax problems and the necessity of delicate *time-scale separation* between the primal and dual updates in attaining convergence. In this work, we propose a *single-loop* adaptive GDA algorithm called TiAda for nonconvex minimax optimization that automatically adapts to the time-scale separation. Our algorithm is *fully parameter-agnostic* and can achieve *near-optimal complexities* simultaneously in deterministic and stochastic settings of nonconvex-strongly-concave minimax problems. The effectiveness of the proposed method is further justified numerically for a number of machine learning applications.

1. Introduction

Adaptive gradient methods, such as AdaGrad [10], Adam [24] and AMSGrad [43], have become the default choice of optimization algorithms in many machine learning applications owing to their robustness to hyper-parameter selection and fast empirical convergence. Classic analyses of gradient descent for smooth functions require the stepsize to be less than $2/l$, where l is the smoothness parameter and often unknown, whereas many adaptive schemes can automatically adapt to them [48, 49]. Such tuning-free algorithms are called *parameter-agnostic*, as they do not require any prior knowledge of problem-specific parameters, e.g., the smoothness or strong-convexity parameter.

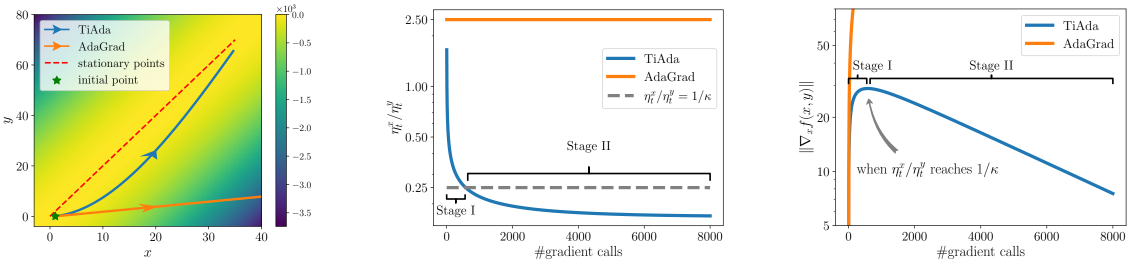
In this work, we aim to bring the benefits of adaptive stepsizes to solving the following problem:

$$\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathcal{Y}} f(x,y) = \mathbb{E}_{\xi\in P}\left[F(x,y;\xi)\right], \tag{1}$$

where P is an unknown distribution from which we can drawn i.i.d. samples, $\mathcal{Y} \subset \mathbb{R}^{d_2}$ is closed and convex, and $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ is nonconvex in x. We call x the primal variable and y the dual variable. This minimax formulation has found vast applications in modern machine learning [2, 7, 11–14, 37, 39]. Albeit theoretically underexplored, adaptive methods are widely deployed in combination with popular minimax optimization algorithms such as (stochastic) gradient descent ascent (GDA), extragradient (EG) [25], and optimistic GDA [41, 42]; see, e.g., [8, 16, 38, 44].

While it seems natural to directly extend adaptive stepsizes to minimax optimization algorithms, a recent work by Yang et al. [50] pointed out that such schemes may not always converge without knowing problem-dependent parameters. Unlike the case of minimization, convergent analyses of

© X. Li, J. Yang & N. He.



(a) trajectory (b) effective stepsize ratio (c) convergence

Figure 1: Comparison between TiAda and vanilla GDA with AdaGrad stepsizes (labeled as AdaGrad) on the quadratic function (2) with $L = 2$ under a poor initial stepsize ratio, i.e., $\eta^x/\eta^y = 5$. Here, η_t^x and η_t^y are the effective stepsizes respectively for x and y, and κ is the condition number[1]. The background color in (a) demonstrates the function value $f(x, y)$.

GDA and EG for nonconvex minimax optimization are subject to *time-scale separation* [4, 35, 46, 51] — the stepsize ratio of primal and dual variables needs to be smaller than a problem-dependent threshold — which is recently shown to be necessary even when the objective is strongly concave in y with true gradients [32]. Moreover, Yang et al. [50] showed that GDA with standard adaptive stepsizes, fails to adapt to the time-scale separation requirement. Take the following nonconvex-strongly-concave function as a concrete example:

$$f(x, y) = -\frac{1}{2}y^2 + Lxy - \frac{L^2}{2}x^2, \tag{2}$$

where $L > 0$ is a constant. Yang et al. [50] proved that directly using adaptive stepsizes like AdaGrad, Adam and AMSGrad will fail to converge if the ratio of initial stepsizes of x and y (denoted as η^x and η^y) is large. We illustrate this phenomenon in Figures 1(a) and 1(c), where AdaGrad diverges. To sum up, adaptive stepsizes designed for minimization, are not *time-scale adaptive* for minimax optimization and thus not *parameter-agnostic*.

To circumvent this time-scale separation bottleneck, Yang et al. [50] introduced NeAda for problem (1) with nonconvex-strongly-concave objectives. Although the algorithm is agnostic to the smoothness and strong-concavity parameters, there are several limitations: (a) In the stochastic setting, it gradually increases the number of inner loop steps (k steps for the k-th outer loop) to improve the inner maximization problem accuracy, resulting in a possible waste of inner loop updates if the maximization problem is already well solved; (b) NeAda needs a large batchsize of order $\Omega\left(\epsilon^{-2}\right)$ to achieve the near-optimal convergence rate in theory; (c) It is not fully adaptive to the gradient noise, since it deploys different strategies for deterministic and stochastic settings.

In this work, we address all of the issues above by proposing TiAda (**Ti**me-scale **Ada**ptive Algorithm), a single-loop algorithm with time-scale adaptivity for minimax optimization. Specifically, one of our major modifications is setting the effective stepsize, i.e., the scale of (stochastic) gradient used in the updates, of the primal variable to the reciprocal of the *maximum* between the primal and dual variables' second moments, i.e., the sums of their past gradient norms. This ensures the effective stepsize ratio of x and y being upper bounded by a decreasing sequence, which eventually reaches the desired time-scale separation. Taking the test function (2) as an example, Figure 1 illus-

[1] Please refer to Section 2 for formal definitions of initial stepsize and effective stepsize. Note that the initial stepsize ratio, η^x/η^y, does not necessarily equal to the first effective stepsize ratio, η_0^x/η_0^y.

trates the time-scale adaptivity of TiAda: In Stage I, the stepsize ratio quickly decreases below the threshold; in Stage II, the ratio is stabilized and the gradient norm starts to converge fast.

In summary, our contributions are as follows:

- We introduce the first *single-loop* and *fully parameter-agnostic* adaptive algorithm, TiAda, for nonconvex-strongly-concave (NC-SC) minimax optimization. It adapts to the necessary time-scale separation without large batchsize or any knowledge of problem-dependant parameters or target accuracy. TiAda finds an ϵ-stationary point with an optimal complexity of $\mathcal{O}\left(\epsilon^{-2}\right)$ in the deterministic case, and a near-optimal sample complexity of $\mathcal{O}\left(\epsilon^{-(4+\delta)}\right)$ for any small $\delta > 0$ in the stochastic case. It shaves off the extra logarithmic terms in the complexity of NeAda with AdaGrad stepsize for both primal and dual variables [50]. TiAda is proven to be noise-adaptive, which is the first of its kind among nonconvex minimax optimization algorithms.
- While TiAda is based on AdaGrad stepsize, in Appendix A, we generalize TiAda with other existing adaptive schemes, and conduct experiments on several tasks. We show that TiAda converges faster and is more robust compared with NeAda or GDA with other existing adaptive stepsizes.

1.1. Related Work

Adaptive gradient methods. The original AdaGrad was introduced for online convex optimization and maintains coordinate-wise stepsizes. In nonconvex stochastic optimization, AdaGrad-Norm with one learning rate for all directions is shown to achieve the same complexity as SGD [33, 48], even with the high probability bound [23, 34]. In comparison, RMSProp [19] and Adam [24] use the decaying moving average of past gradients, but may suffer from divergence [43]. Many variants of Adam are proposed, and a wide family of them, including AMSGrad, are provided with convergence guarantees [6, 9, 53, 54]. One of the distinguishing traits of adaptive algorithms is that they can achieve order-optimal rates without knowledge about the problem parameters [22, 29, 48].

Adaptive minimax optimization algorithms. In the convex-concave regime, several adaptive algorithms are designed based on EG and AdaGrad stepsize, and they inherit the parameter-agnostic characteristic [1, 3]. In sharp contrast, when the objective function is nonconvex about one variable, most existing adaptive algorithms require knowledge of the problem parameters [17, 20, 21]. Very recently, it was proved that a parameter-dependent ratio between two stepsizes is necessary for GDA in NC-SC minimax problems with non-adaptive stepsize [32] and most existing adaptive stepsizes [50]. Heusel et al. [18] shows the two-time-scaled GDA with non-adaptive stepsize or Adam will converge, but assuming the existence of an asymptotically stable attractor.

1.2. Notations

We denote l as the smoothness parameter, μ as the strong-concavity parameter, and $\kappa := l/\mu$ as the condition number. For the minimax problem (1), we denote $y^*(x) := \arg\max_{y\in\mathcal{Y}} f(x, y)$ as the solution of the inner maximization problem, $\Phi(x) := f(x, y^*(x))$ as the primal function, and $\mathcal{P}_{\mathcal{Y}}(\cdot)$ as projection operator onto set $\mathcal{Y}$. We assume access to stochastic gradient oracle returning $\nabla_x F(x, y; \xi)$ for x and $\nabla_y F(x, y; \xi)$ for y.

2. Method

Algorithm 1 TiAda (Time-scale Adaptive Algorithm)

```
1: Input: $(x_0, y_0)$, $v_0^x > 0$, $v_0^y > 0$, $\eta^x > 0$, $\eta^y > 0$, $\alpha > 0$, $\beta > 0$ and $\alpha > \beta$.
2: for $t = 0, 1, 2, ...$ do
3:     sample i.i.d. $\xi_t^x$ and $\xi_t^y$, and let $g_t^x = \nabla_x F(x_t, y_t; \xi_t^x)$ and $g_t^y = \nabla_y F(x_t, y_t; \xi_t^y)$
4:     $v_{t+1}^x = v_t^x + \|g_t^x\|^2$ and $v_{t+1}^y = v_t^y + \|g_t^y\|^2$
5:     $x_{t+1} = x_t - \frac{\eta^x}{\max\{v_{t+1}^x, v_{t+1}^y\}^\alpha} g_t^x$ and $y_{t+1} = \mathcal{P}_\mathcal{Y}\left(y_t + \frac{\eta^y}{\left(v_{t+1}^y\right)^\beta} g_t^y\right)$
6: end for
```

We formally introduce the TiAda method in Algorithm 1, and the major difference with AdaGrad lies in line 5. Like AdaGrad, TiAda stores the accumulated squared (stochastic) gradient norm of the primal and dual variables in v_t^x and v_t^y, respectively. We refer to hyper-parameters η^x and η^y as the *initial stepsizes*, and the actual stepsizes for updating in line 5 as *effective stepsizes* which are denoted by η_t^x and η_t^y. TiAda adopts effective stepsizes $\eta_t^x = \eta^x / \max\left\{v_{t+1}^x, v_{t+1}^y\right\}^\alpha$ and $\eta_t^y = \eta^y / \left(v_{t+1}^y\right)^\beta$, while AdaGrad uses $\eta^x / \left(v_{t+1}^x\right)^{1/2}$ and $\eta^y / \left(v_{t+1}^y\right)^{1/2}$. In Section 3, our theoretical analysis suggests to choose $\alpha > 1/2 > \beta$.

2.1. The Time-Scale Adaptivity of TiAda

Current analyses of GDA with non-adaptive stepsizes require the time-scale, η_t^x / η_t^y, to be smaller than a threshold depending on problem constants such as the smoothness and the strong-concavity parameter [35, 51]. It is tempting to expect adaptive stepsizes to automatically find a suitable time-scale separation. However, the quadratic example (2) given by Yang et al. [50] shattered the illusion. In this example, the effective stepsize ratio stays the same along the run of existing adaptive algorithms, including AdaGrad (see Figure 1(*b*)), Adam and AMSGrad, and they fail to converge if the initial stepsizes are not carefully chosen (see Yang et al. [50] for details). As v_t^x and v_t^y only separately contain the gradients of x and y, the effective stepsizes of two variables in these methods depend on their own history, which prevents them from cooperating to adjust the ratio.

Now we explain how TiAda adapts to both the required time-scale separation and small enough stepsizes. First, the ratio of our modified effective stepsizes is upper bounded by a decreasing sequence when $\alpha > \beta$:

$$\frac{\eta_t^x}{\eta_t^y} = \frac{\eta^x / \max\left\{v_{t+1}^x, v_{t+1}^y\right\}^\alpha}{\eta^y / \left(v_{t+1}^y\right)^\beta} \leq \frac{\eta^x / \left(v_{t+1}^y\right)^\alpha}{\eta^y / \left(v_{t+1}^y\right)^\beta} = \frac{\eta^x}{\eta^y \left(v_{t+1}^y\right)^{\alpha-\beta}}, \tag{3}$$

as v_t^y is the sum of previous gradient norms and is increasing. Regardless of the initial stepsize ratio η^x / η^y, we expect the effective stepsize ratio to eventually drop below the desirable threshold for convergence. On the other hand, the effective stepsizes for the primal and dual variables are also upper bounded by decreasing sequences, $\eta^x / \left(v_{t+1}^x\right)^\alpha$ and $\eta^y / \left(v_{t+1}^y\right)^\beta$, respectively. Similar to AdaGrad, such adaptive stepsizes will reduce to small enough, e.g., $\mathcal{O}(1/l)$, to ensure convergence.

To demonstrate the time-scale adaptivity of TiAda, we conducted experiments on the quadratic minimax example (2) with $L = 2$. As shown in Figure 1(*b*), while the effective stepsize ratio of AdaGrad stays unchanged for this particular function, TiAda progressively decreases the ratio. According to Lemma 2.1 of Yang et al. [50], $1/\kappa$ is the threshold where GDA starts to converge. We label the time period before reaching this threshold as Stage I, during which as shown in Figure 1(c), the gradient norm for TiAda increases. However, as soon as it enters Stage II, i.e., when the ratio drops below $1/\kappa$, TiAda converges fast to the stationary point. In contrast, since the stepsize ratio of AdaGrad never reaches this threshold, the gradient norm keeps growing.

3. Theoretical Analysis of TiAda

In this section, we study the convergence of TiAda under NC-SC setting with both deterministic and stochastic gradient oracles. We make the following assumptions to develop our convergence results.

Assumption 3.1 (smoothness) *Function $f(x, y)$ is l-smooth ($l > 0$) in both x and y, that is, for any $x_1, x_2 \in \mathbb{R}^{d_1}$ and $y_1, y_2 \in \mathcal{Y}$, we have*

$$\max\{\|\nabla_x f(x_1, y_1) - \nabla_x f(x_2, y_2)\|, \|\nabla_y f(x_1, y_1) - \nabla_y f(x_2, y_2)\|\} \leq l\left(\|x_1 - x_2\| + \|y_1 - y_2\|\right).$$

Assumption 3.2 (strong-concavity in y) *Function $f(x, y)$ is μ-strongly-concave ($\mu > 0$) in y, that is, for any $x \in \mathbb{R}^{d_1}$ and $y_1, y_2 \in \mathcal{Y}$, we have*

$$f(x, y_1) \geq f(x, y_2) + \langle \nabla_y f(x, y_2), y_1 - y_2 \rangle + \frac{\mu}{2}\|y_1 - y_2\|^2.$$

Assumption 3.3 (interior optimal point) *For any $x \in \mathbb{R}^{d_1}$, $y^*(x)$ is in the interior of $\mathcal{Y}$.*

Remark 1 *The last assumption ensures $\nabla_y f(x, y^*(x)) = 0$, which is important for AdaGrad-like stepsizes. If the gradient about y is not 0 at $y^*(x)$, the stepsize will keep decreasing even near the optimal point, leading to slow convergence. This assumption could be potentially alleviated by using generalized AdaGrad stepsizes [3].*

We aim to find a near stationary point for the minimax problem (1). Here, (x, y) is defined to be an ϵ stationary point if $\|\nabla_x f(x, y)\| \leq \epsilon$ and $\|\nabla_y f(x, y)\| \leq \epsilon$ in the deterministic setting, or $\mathbb{E}\|\nabla_x f(x, y)\|^2 \leq \epsilon^2$ and $\mathbb{E}\|\nabla_y f(x, y)\|^2 \leq \epsilon^2$ in the stochastic setting, where the expectation is taken over all the randomness in the algorithm. This stationarity notion can be easily translated to the near-stationarity of the primal function $\Phi(x) = \max_{y \in \mathcal{Y}}(x, y)$ [51]. Under our analyses, TiAda is able to achieve the optimal $\mathcal{O}\left(\epsilon^{-2}\right)$ complexity in the deterministic setting and a near-optimal $\mathcal{O}\left(\epsilon^{-(4+\delta)}\right)$ sample complexity for any small $\delta > 0$ in the stochastic setting.

Firstly we assume to have access to the exact gradients of $f(\cdot, \cdot)$, and therefore we can replace $\nabla_x F(x_t, y_t; \xi_t^x)$ and $\nabla_y F(x_t, y_t; \xi_t^y)$ by $\nabla_x f(x_t, y_t)$ and $\nabla_y f(x_t, y_t)$ in Algorithm 1.

Theorem 2 (deterministic setting) *Under Assumptions 3.1 to 3.3, Algorithm 1 with deterministic gradient oracles satisfies that for any $0 < \beta < \alpha < 1$, after T iterations,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_x f(x_t, y_t)\|^2 + \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_y f(x_t, y_t)\|^2 \leq \mathcal{O}\left(\frac{1}{T}\right).$$

This theorem implies that for any initial stepsizes, TiAda finds an ϵ-stationary point within $\mathcal{O}(\epsilon^{-2})$ iterations. Such complexity is comparable to that of nonadaptive methods, such as vanilla GDA [35], and is optimal in the dependency of ϵ [52]. Like NeAda [50], TiAda does not need any prior knowledge about μ and l, but it improves over NeAda by removing the logarithmic term in the complexity. Notably, we provide a unified analysis for a wide range of α and β. while most existing literature on only validates a specific choice, e.g., $\alpha = 1/2$, for AdaGrad-like stepsizes [22, 48].

Now, we assume the access to a stochastic gradient oracle, that returns unbiased noisy gradients, $\nabla_x F(x, y; \xi)$ and $\nabla_y F(x, y; \xi)$. Also, we make the following additional assumptions.

Assumption 3.4 (stochastic gradients) *For $z \in \{x, y\}$, we have $\mathbb{E}_\xi [\nabla_z F(x, y, \xi)] = \nabla_z f(x, y)$. In addition, there exists a constant G such that $\|\nabla_z F(x, y, \xi)\| \leq G$ for any $x \in \mathbb{R}^{d_1}$ and $y \in \mathcal{Y}$.*

Assumption 3.5 (bounded primal function value) *There exists a constant $\Phi_{\max} \in \mathbb{R}$ such that for any $x \in \mathbb{R}^{d_1}$, $\Phi(x)$ is upper bounded by $\Phi_{\max}$.*

Remark 3 *Many works on adaptive algorithms have the two assumptions above [23, 29]. This implies the domain of y is bounded, which is also assumed in the analyses of AdaGrad [28, 30].*

Assumption 3.6 (second order Lipschitz continuity for y) *For any $x_1, x_2 \in \mathbb{R}^{d_1}$ and $y_1, y_2 \in \mathcal{Y}$, there exists constant L such that $\left\|\nabla^2_{xy} f(x_1, y_1) - \nabla^2_{xy} f(x_2, y_2)\right\| \leq L\left(\|x_1 - x_2\| + \|y_1 - y_2\|\right)$ and $\left\|\nabla^2_{yy} f(x_1, y_1) - \nabla^2_{yy} f(x_2, y_2)\right\| \leq L\left(\|x_1 - x_2\| + \|y_1 - y_2\|\right)$.*

Remark 4 *Chen et al. [5] also impose this assumption to achieve the optimal $\mathcal{O}\left(\epsilon^{-4}\right)$ complexity for GDA with non-adaptive stepsizes for solving NC-SC minimax problems. Together with Assumption 3.3, we can show that $y^*(\cdot)$ is smooth. Nevertheless, without this assumption, Lin et al. [35] only show a worse complexity of $\mathcal{O}\left(\epsilon^{-5}\right)$ for GDA without large batchsize.*

Theorem 5 (stochastic setting) *Under Assumptions 3.1 to 3.6, Algorithm 1 with stochastic gradient oracles satisfies that for any $0 < \beta < \alpha < 1$, after T iterations,*

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_x f(x_t, y_t)\|^2 + \sum_{t=0}^{T-1}\|\nabla_y f(x_t, y_t)\|^2\right] \leq \mathcal{O}\left(T^{\alpha-1} + T^{-\alpha} + T^{\beta-1} + T^{-\beta}\right).$$

TiAda can achieve the complexity arbitrarily close to the optimal sample complexity, $\mathcal{O}\left(\epsilon^{-4}\right)$ [31], by choosing α and β arbitrarily close to 0.5. Specifically, TiAda achieves a complexity of $\mathcal{O}\left(\epsilon^{-(4+\delta)}\right)$ for any small $\delta > 0$ if we set $\alpha = 0.5 + \delta/(8+2\delta)$ and $\beta = 0.5 - \delta/(8+2\delta)$. Notably, this matches the complexity of NeAda with AdaGrad stepsizes for both variables [50].

Theorem 5 implies that TiAda is fully agnostic to problem parameters, e.g., μ, l and σ. Compared with the only parameter-agnostic algorithm, NeAda, our algorithm has several advantages. First, TiAda is a single-loop algorithm, while NeAda [50] needs increasing inner-loop steps and a huge batchsize of order $\Omega\left(\epsilon^{-2}\right)$ to achieve its best complexity. Second, our stationary guarantee is for $\mathbb{E}\|\nabla_x f(x, y)\|^2 \leq \epsilon^2$, which is stronger than $\mathbb{E}\|\nabla_x f(x, y)\| \leq \epsilon$ guarantee in NeAda. Last but not least, NeAda needs to know whether $\sigma = 0$ as it uses different stopping criteria for the inner loop in deterministic and stochastic settings. In comparison, TiAda achieves the (near) optimal complexity in both settings with the same strategy.

4. Conclusion

In this work, we bring in adaptive stepsizes to nonconvex minimax problems in a parameter-agnostic manner. We designed the first time-scale adaptive algorithm, TiAda, which progressively adjusts the effective stepsize ratio and reaches the desired time-scale separation. TiAda is also noise adaptive and does not require large batchsizes compared with the existing parameter-agnostic algorithm for nonconvex minimax optimization. Furthermore, TiAda is able to achieve optimal and near-optimal complexities respectively wtih deterministic and stochastic gradient oracles. We also empirically showcased the advantages of TiAda over NeAda and GDA with adaptive stepsizes.

References

[1] Kimon Antonakopoulos, Veronica Belmega, and Panayotis Mertikopoulos. An adaptive mirror-prox method for variational inequalities with singular operators. *NeurIPS*, 32, 2019.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223. PMLR, 2017.

[3] Francis Bach and Kfir Y Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Conference on Learning Theory*, pages 164–194. PMLR, 2019.

[4] Radu Ioan Boţ and Axel Böhm. Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *arXiv preprint arXiv:2007.13605*, 2020.

[5] Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. 2021.

[6] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*, 2018.

[7] Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. Learning from conditional distributions via dual embeddings. In *AISTATS*, pages 1458–1467. PMLR, 2017.

[8] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *ICLR*, 2018.

[9] Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

[10] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[11] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2021.

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014.

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[15] Green9. Pytorch code for gan models. https://github.com/Zeleni9/pytorch-wgan, 2018.

[16] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *NeurIPS*, 30, 2017.

[17] Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. A novel convergence analysis for algorithms of the adam family. *arXiv preprint arXiv:2112.03459*, 2021.

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[19] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.

[20] Feihu Huang and Heng Huang. Adagda: Faster adaptive gradient descent ascent methods for minimax optimization. *arXiv preprint arXiv:2106.16101*, 2021.

[21] Feihu Huang, Xidong Wu, and Heng Huang. Efficient mirror descent ascent methods for nonsmooth minimax problems. *NeurIPS*, 34:10431–10443, 2021.

[22] Ali Kavis, Kfir Y Levy, Francis Bach, and Volkan Cevher. Unixgrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. *NeurIPS*, 32, 2019.

[23] Ali Kavis, Kfir Levy, and Volkan Cevher. High probability bounds for a class of nonconvex algorithms with adagrad stepsize. In *ICLR*, 2022.

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[25] Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[27] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[28] Kfir Levy. Online to offline conversions, universality and adaptive minibatch sizes. *NeurIPS*, 30, 2017.

[29] Kfir Levy, Ali Kavis, and Volkan Cevher. Storm+: Fully adaptive sgd with recursive momentum for nonconvex optimization. *NeurIPS*, 34:20571–20582, 2021.

[30] Kfir Y Levy, Alp Yurtsever, and Volkan Cevher. Online adaptive methods, universality and acceleration. *NeurIPS*, 31, 2018.

[31] Haochuan Li, Yi Tian, Jingzhao Zhang, and Ali Jadbabaie. Complexity lower bounds for nonconvex-strongly-concave min-max optimization. *NeurIPS*, 34:1792–1804, 2021.

[32] Haochuan Li, Farzan Farnia, Subhro Das, and Ali Jadbabaie. On convergence of gradient descent ascent: A tight local analysis. In *ICML*, pages 12717–12740. PMLR, 2022.

[33] Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd international conference on artificial intelligence and statistics*, pages 983–992. PMLR, 2019.

[34] Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive sgd with momentum. *arXiv preprint arXiv:2007.14294*, 2020.

[35] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *ICML*, pages 6083–6093. PMLR, 2020.

[36] Louis Lv. Reproducing "certifying some distributional robustness with principled adversarial training". https://github.com/Louis-udm/Reproducing-certifiable-distributional-robustness, 2019.

[37] David J Miller, Zhen Xiang, and George Kesidis. Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks. *Proceedings of the IEEE*, 108(3):402–433, 2020.

[38] Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pages 4573–4582. PMLR, 2020.

[39] Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *arXiv preprint arXiv:2102.07035*, 2021.

[40] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[41] Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

[42] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019. PMLR, 2013.

[43] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *ICLR*, 2018.

[44] Amirhossein Reisizadeh, Farzan Farnia, Ramtin Pedarsani, and Ali Jadbabaie. Robust federated learning: The case of affine distribution shifts. *Advances in Neural Information Processing Systems*, 33:21554–21565, 2020.

[45] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 29, 2016.

[46] Othmane Sebbouh, Marco Cuturi, and Gabriel Peyré. Randomized stochastic gradient descent ascent. In *AISTATS*, pages 2941–2969. PMLR, 2022.

[47] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *ICLR*, 2018.

[48] Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *The Journal of Machine Learning Research*, 21(1):9047–9076, 2020.

[49] Yuege Xie, Xiaoxia Wu, and Rachel Ward. Linear convergence of adaptive stochastic gradient descent. In *AISTATS*, pages 1475–1485. PMLR, 2020.

[50] Junchi Yang, Xiang Li, and Niao He. Nest your adaptive algorithm for parameter-agnostic nonconvex minimax optimization. *arXiv preprint arXiv:2206.00743*, 2022.

[51] Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *AISTATS*, pages 5485–5517. PMLR, 2022.

[52] Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pages 482–492. PMLR, 2021.

[53] Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. *arXiv preprint arXiv:2208.09632*, 2022.

[54] Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

Appendix A. Experiments

In this section, we first present extensions of TiAda that accommodate other adaptive schemes besides AdaGrad and are more practical in deep models. Then we present empirical results of TiAda and compare it with (i) simple combinations of GDA and adaptive stepsizes, which are commonly used in practice, and (ii) NeAda with different adaptive mechanism [50]. For fair comparisons, we use the same hyper-parameters when comparing our TiAda with other algorithms. Our experiments include test functions proposed by Yang et al. [50], the NC-SC distributional robustness optimization [47], and training the nonconvex-nonconcave Wasserstein GAN with gradient penalty [16]. We believe that this not only validates our theoretical results but also shows the potential of our algorithm in real-world scenarios. To show the strength of being parameter-agnostic of TiAda, in all the experiments, we merely select $\alpha = 0.6$ and $\beta = 0.4$ without further tuning those two hyper-parameters. For notational simplicity, we will use the name of an existing adaptive algorithm to refer to the simple combination of GDA and it, i.e., setting the stepsize of GDA to that adaptive scheme separately for both x and y. For instance "AdaGrad" for minimax problems stands for the algorithm that uses AdaGrad stepsizes separately for x and y in GDA.

A.1. Extensions to Other Adaptive Stepsizes and High-dimensional Models

Although we design TiAda upon AdaGrad-Norm, it is easy and intuitive to apply other adaptive schemes like Adam and AMSGrad. To do so, for $z \in \{x, y\}$, we replace the definition of g_t^z and v_{t+1}^z in line 3 and 4 of Algorithm 1 to

$$g_t^z = \beta_t^z g_{t-1}^z + (1 - \beta_t^z)\nabla_z F(x_t, y_t; \xi_t^z), \quad v_{t+1}^z = \psi\left(v_0, \left\{\|\nabla_z F(x_i, y_i; \xi_i^z)\|^2\right\}_{i=0}^t\right),$$

Table 1: Stepsize schemes fit in generalized TiAda. See also [50].

Algorithms	first moment parameter β_t	second moment function $\psi\left(v_0, \{u_i^2\}_{i=0}^t\right)$
AdaGrad (TiAda)	$\beta_t = 0$	$v_0 + \sum_{i=0}^t u_i^2$
GDA	$\beta_t = 0$	1
Adam	$0 < \beta_t < 1$	$\gamma^{t+1} v_0 + (1-\gamma) \sum_{i=0}^t \gamma^{t-i} u_i^2$
AMSGrad	$0 < \beta_t < 1$	$\max_{m=0,\ldots,t} \gamma^{m+1} v_0 + (1-\gamma) \sum_{i=0}^m \gamma^{m-i} u_i^2$

where $\{\beta_t^z\}$ is the momentum parameters and ψ is the second moment function. Some common stepsizes that fit in this generalized framework can be seen in Table 1. Since Adam is widely used in many deep learning tasks, we also implement generalized TiAda with Adam stepsizes in our experiments for real-world applications, and we label it "TiAda-Adam".

Besides generalizing TiAda to accommodate different stepsize schemes, for high-dimensional models, we also provide a coordinate-wise version of TiAda. Note that we cannot simply change everything in Algorithm 1 to be coordinate-wise, because we use the gradients of y in the stepsize of x and there are no corresponding relationships between the coordinates of x and y. Therefore we use the global accumulated gradient norms to dynamically adjust the stepsize of x. Denote the second moment (analogous to v_{t+1}^x in Algorithm 1) for the i-th coordinate of x at the t-th step as $v_{t+1,i}^x$ and globally $v_{t+1}^x := \sum_{i=1}^{d_1} v_{t+1,i}^x$. We also use similar notations for y. Then, the update for the i-th parameter, i.e., x^i and y^i, can be written as

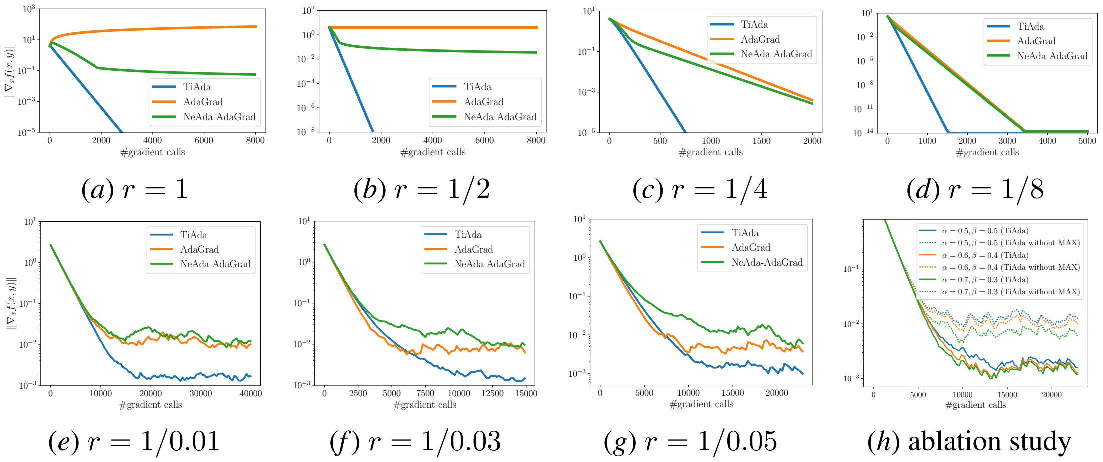
$$\begin{cases} x_{t+1}^i = x_t^i - \frac{\left(v_{t+1}^x\right)^\alpha}{\max\left\{v_{t+1}^x, v_{t+1}^y\right\}^\alpha} \cdot \frac{\eta^x}{\left(v_{t+1,i}^x\right)^\alpha} \nabla_{x^i} f(x_t, y_t) \\ y_{t+1}^i = y_t^i + \frac{\eta^y}{\left(v_{t+1,i}^y\right)^\beta} \nabla_{y^i} f(x_t, y_t). \end{cases}$$

If we look at the effective stepsize of x in the above update scheme, when the overall gradients of y are small (i.e., $v_{t+1}^y < v_{t+1}^x$), meaning the inner maximization problem is well solved, then the first factor becomes 1 and the effective stepsize of x is just the second factor, similar to the AdaGrad updates. If the term v_{t+1}^y dominates over v_{t+1}^x, the first factor would be smaller than 1, allowing to slow down the update of x and waiting for a better approximation of $y^*(x)$.

Our results in the following subsections provide strong empirical evidence for the effectiveness of these TiAda variants, and developing convergence guarantees for them would be an interesting future work. We believe our proof techniques for TiAda, together with existing convergence results for coordinate-wise AdaGrad and AMSGrad [6, 9, 54], can shed light on the theoretical analyses of these variants.

A.2. Test Functions

Firstly, we examine TiAda on the quadratic function (2) that shows the non-convergence of simple combinations of GDA and adaptive stepsizes [50]. Since our TiAda is based on AdaGrad, we compare it with GDA with AdaGrad stepsize and NeAda-AdaGrad [50]. For Figure 1 and the first row of Figure 2, we conduct experiments on problem (2) with $L = 2$. We use initial stepsize $\eta^y = 0.2$ and initial point $(1, 0.01)$ for all runs. As shown in Figure 2, when the initial ratio is poor,



$(a)\, r = 1$ $(b)\, r = 1/2$ $(c)\, r = 1/4$ $(d)\, r = 1/8$

$(e)\, r = 1/0.01$ $(f)\, r = 1/0.03$ $(g)\, r = 1/0.05$ (h) ablation study

Figure 2: Comparison of the algorithms on test functions. $r = \eta^x/\eta^y$ is the initial stepsize ratio. In the first row, we use the quadratic function (2) with $L = 2$ under deterministic gradient oracles. For the second row, we test the methods on the McCormick function with noisy gradients.

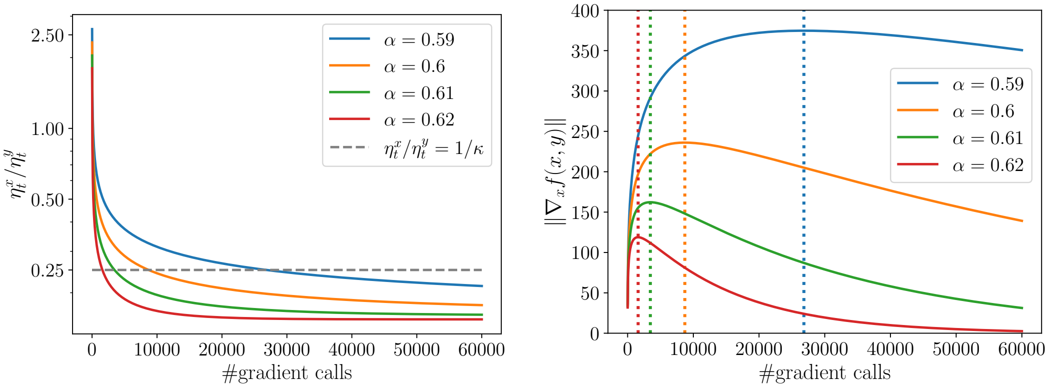
TiAda and NeAda-AdaGrad always converge while AdaGrad diverges. NeAda also suffers from slow convergence when the initial ratio is poor, e.g., 1 and 1/2 after 2000 iterations. In contrast, TiAda automatically balances the stepsizes and converges fast under all ratios.

Ablation Study on Convergence Behavior with Different α and β We conduct more experiments on the quadratic function to study the effect of hyper-parameters α and β on the convergence behavior of TiAda. As discussed in Sections 1 and 2, we refer to the period before the stepsize ratio reduce to the convergence threshold as Stage I, and the period after that as Stage II. In order to accentuate the difference between these two stages, we pick a large initial stepsize ratio $\eta^x/\eta^y = 20$. We compare 4 different pairs of α and β: $\alpha \in \{0.59, 0.6, 0.61, 0.62\}$ and $\beta = 1-\alpha$. From Figure 3, we observed that as soon as TiAda enters Stage II, the norm of gradients start to drop. Moreover, the closer α and β are to 0.5, the more time TiAda remains in Stage I, which confirms the intuitions behind our analysis in Appendix C.3.

For the stochastic case, we follow Yang et al. [50] and conduct experiments on the McCormick function which is more complicated and 2-dimensional: $f(x,y) = \sin(x_1+x_2)+(x_1-x_2)^2-\frac{3}{2}x_1+\frac{5}{2}x_2+1+x_1y_1+x_2y_2-\frac{1}{2}(y_1^2+y_2^2)$. We chose $\eta^y = 0.01$, and the noises added to the gradients are from zero-mean Gaussian distribution with variance 0.01. TiAda consistently outperforms AdaGrad and NeAda-AdaGrad as demonstrated in the second row of Figure 2 regardless of the initial ratio. In this function, we also run an ablation study on the effect of our design that uses max-operator in the update of x. We compare TiAda with and its variant without the max-operator, TiAda without MAX (Algorithm 2 in Appendix C.4) whose effective stepsizes of x are $\eta^x/\left(v_{t+1}^x\right)^\alpha$. According to Figure 2(h), TiAda converges to smaller gradient norms under all configurations of α and β.

A.3. Distributional robustness optimization

In this subsection, we consider the distributional robustness optimization [47]. We target training the model weights, the primal variable x, to be robust to the perturbations in the image inputs, the



(*a*) effective stepsize ratio

(*b*) convergence

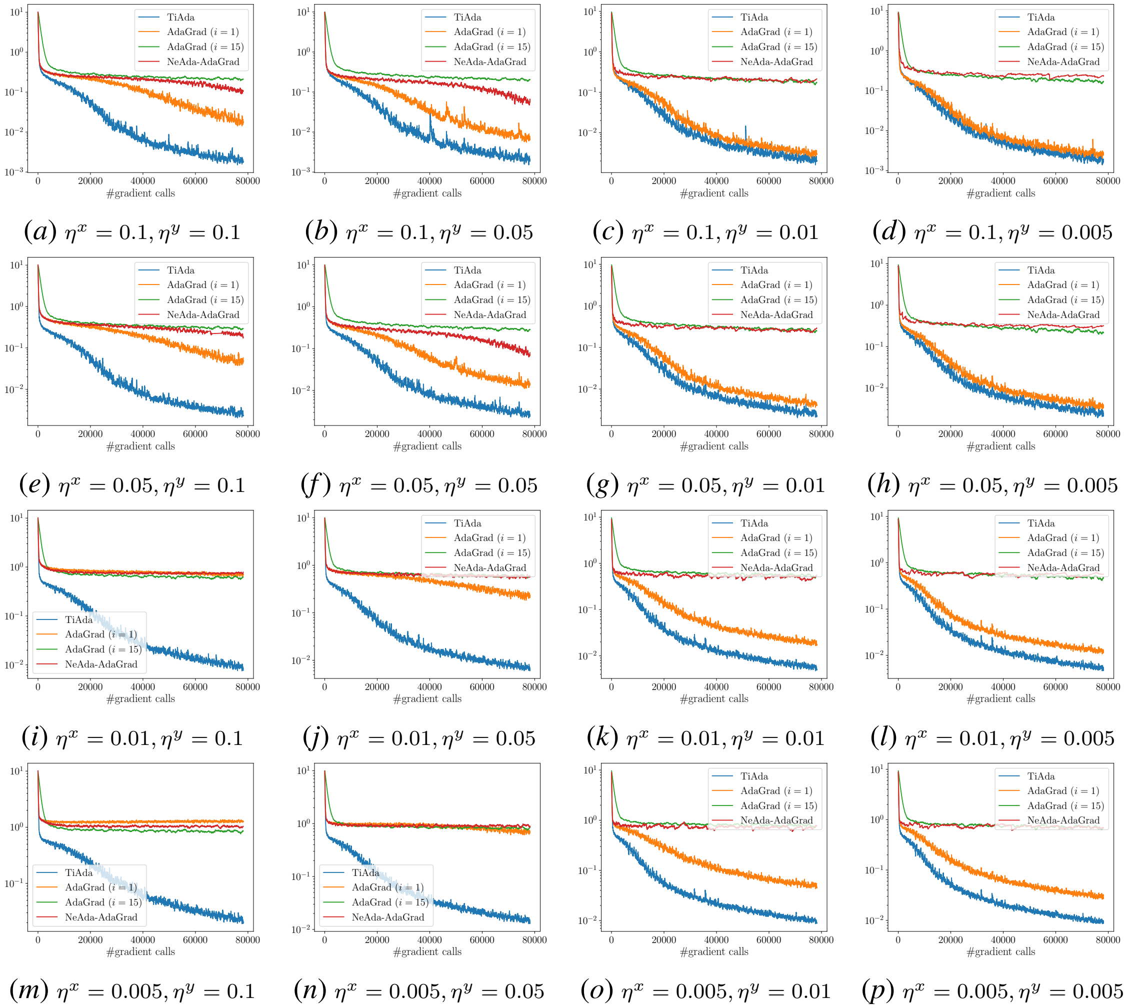
Figure 3: Illustration of the effect of α and β on the two stages in TiAda's time-scale adaptation process. We set $\beta = 1 - \alpha$. The dashed line on the right plot represents the first iteration when the effective stepsize ratio is below $1/\kappa$.

dual variable y. The problem can be formulated as:

$$\min_{x} \max_{y=[y_1,\ldots,y_n]} \frac{1}{n} \sum_{i=1}^{n} f_i(x, y_i) - \gamma \|y_i - v_i\|^2, \tag{4}$$

where f_i is the loss function of the i-th sample, v_i is the i-th input image, and y_i is the corresponding perturbation. There are a total of n samples and γ is a trade-off hyper-parameter between the original loss and the penalty of the perturbations. If γ is large enough, the problem is NC-SC. We adapt code from Lv [36], and used the same hyper-parameter setting as Sebbouh et al. [46], Sinha et al. [47], i.e., $\gamma = 1.3$. The model we used is a three layer convolutional neural network (CNN) with a final fully-connected layer. For each layer, batch normalization and ELU activation are used. The width of each layer is $(32, 64, 128, 512)$. The setting is the same as Sinha et al. [47], Yang et al. [50]. We set the batchsize as 128, and for the Adam-like optimizers, including Adam, NeAda-Adam and TiAda-Adam, we use $\beta_1 = 0.9, \beta_2 = 0.999$ for the first moment and second moment parameters.

We conduct the experiments on the MNIST dataset [27]. Since it is common in practice to update y 15 times after each x update [47] for better generalization error, we implement AdaGrad using both single and 15 iterations of inner loop (update of y). We use a grid of stepsize combinations to evaluate TiAda and compare it with NeAda and GDA with corresponding adaptive stepsizes. For AdaGrad-like algorithms, we use $\{0.1, 0.05, 0.01, 0.0005\}$ for both η^x and η^y, and the results are reported in Figure 4. In all cases, TiAda outperforms NeAda and AdaGrad, especially when $\eta^y = 0.1$ or 0.05, the performance gap is large. For Adam-like algorithms, we use $\{0.001, 0.0005, 0.0001\}$ for η^x and $\{0.1, 0.05, 0.005, 0.001\}$ for η^y, and the results are shown in Figure 5. In this case, we find that TiAda is not only faster, but also more stable comparing to Adam with single inner loop iteration. We note that since Adam uses the reciprocal of the moving average of gradient norms, it is extremely unstable when the gradients are small. Therefore, Adam-like algorithms often experience instability when they are near stationary points.



(a) $\eta^x = 0.1, \eta^y = 0.1$ (b) $\eta^x = 0.1, \eta^y = 0.05$ (c) $\eta^x = 0.1, \eta^y = 0.01$ (d) $\eta^x = 0.1, \eta^y = 0.005$

(e) $\eta^x = 0.05, \eta^y = 0.1$ (f) $\eta^x = 0.05, \eta^y = 0.05$ (g) $\eta^x = 0.05, \eta^y = 0.01$ (h) $\eta^x = 0.05, \eta^y = 0.005$

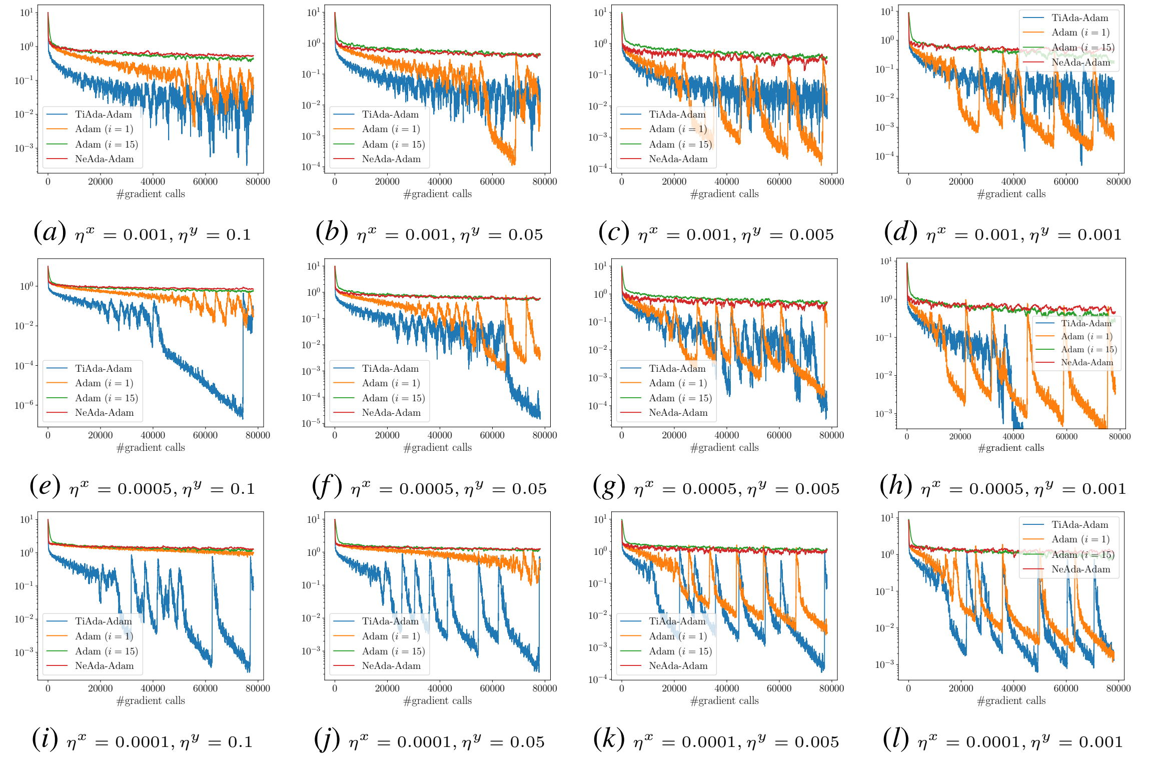
(i) $\eta^x = 0.01, \eta^y = 0.1$ (j) $\eta^x = 0.01, \eta^y = 0.05$ (k) $\eta^x = 0.01, \eta^y = 0.01$ (l) $\eta^x = 0.01, \eta^y = 0.005$

(m) $\eta^x = 0.005, \eta^y = 0.1$ (n) $\eta^x = 0.005, \eta^y = 0.05$ (o) $\eta^x = 0.005, \eta^y = 0.01$ (p) $\eta^x = 0.005, \eta^y = 0.005$

Figure 4: Gradient norms in x of AdaGrad-like algorithms on distributional robustness optimization (4). We use i in the legend to indicate the number of inner loops.

A.4. Generative Adversarial Networks

Another successful and popular application of minimax optimization is generative adversarial networks. In this task, a discriminator (or critic) is trained to distinguish whether an image is from the dataset. At the same time, a generator is mutually trained to synthesize samples with the same distribution as the training dataset so as to fool the discriminator. We use WGAN-GP loss [16], which imposes the discriminator to be a 1-Lipschitz function, with CIFAR-10 dataset [26] in our experiments. We use the code adapted from Green9 [15]. A four layer CNN and a four layer CNN with transpose convolution layers are used respectively for the discriminator and generator. Following a similar setting as Daskalakis et al. [8], we set batchsize as 512, the dimension of latent variable as 50 and the weight of gradient penalty term as 10^{-4}. For the Adam-like optimizers, we set $\beta_1 = 0.5, \beta_2 = 0.9$. To get the inception score [45], we feed the pre-trained inception network



(a) $\eta^x = 0.001, \eta^y = 0.1$ (b) $\eta^x = 0.001, \eta^y = 0.05$ (c) $\eta^x = 0.001, \eta^y = 0.005$ (d) $\eta^x = 0.001, \eta^y = 0.001$

(e) $\eta^x = 0.0005, \eta^y = 0.1$ (f) $\eta^x = 0.0005, \eta^y = 0.05$ (g) $\eta^x = 0.0005, \eta^y = 0.005$ (h) $\eta^x = 0.0005, \eta^y = 0.001$

(i) $\eta^x = 0.0001, \eta^y = 0.1$ (j) $\eta^x = 0.0001, \eta^y = 0.05$ (k) $\eta^x = 0.0001, \eta^y = 0.005$ (l) $\eta^x = 0.0001, \eta^y = 0.001$

Figure 5: Gradient norms in x of Adam-like algorithms on distributional robustness optimization (4). We use i in the legend to indicate the number of inner loops.

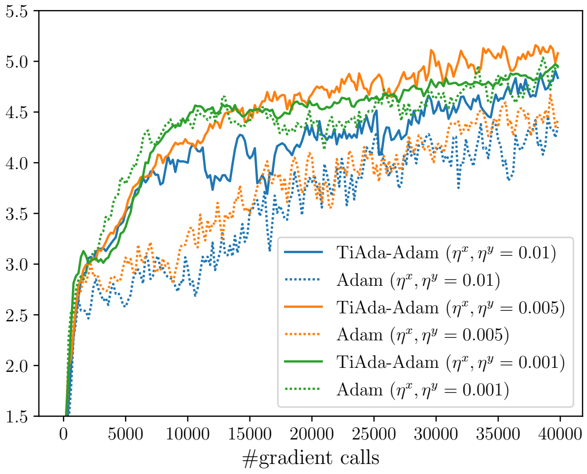


Figure 6: Inception score on WGAN-GP.

with 8000 synthesized samples. Since TiAda is a single-loop algorithm, for fair comparisons, we also update the discriminator only once for each generator update in Adam.

In Figure 6, we plot the inception scores of TiAda-Adam and Adam under different initial stepsizes. We use the same color for the same initial stepsizes, and different line styles to distinguish the two methods, i.e., solid lines for TiAda-Adam and dashed lines for Adam. For all the three initial stepsizes we consider, TiAda-Adam achieves higher inception scores. Also, TiAda-Adam is more robust to initial stepsize selection, as the gap between different solid lines at the end of training is smaller than the dashed lines.

Appendix B. Helper Lemmas

Lemma 6 *Let $x_1, ..., x_T$ be a sequence of non-negative real numbers, $x_1 > 0$ and $0 < \alpha < 1$. Then we have*

$$\left(\sum_{t=1}^{T} x_t\right)^{1-\alpha} \leq \sum_{t=1}^{T} \frac{x_t}{\left(\sum_{k=1}^{t} x_k\right)^{\alpha}} \leq \frac{1}{1-\alpha}\left(\sum_{t=1}^{T} x_t\right)^{1-\alpha}.$$

When $\alpha = 1$, we have

$$\sum_{t=1}^{T} \frac{x_t}{\left(\sum_{k=1}^{t} x_k\right)^{\alpha}} \leq 1 + \log\left(\frac{\sum_{t=1}^{t} x_t}{x_1}\right).$$

Proof For the first inequality in the case $0 < \alpha < 1$, we have

$$\sum_{t=1}^{T} \frac{x_t}{\left(\sum_{k=1}^{t} x_k\right)^{\alpha}} \geq \sum_{t=1}^{T} \frac{x_t}{\left(\sum_{k=1}^{T} x_k\right)^{\alpha}} = \frac{\sum_{t=1}^{T} x_t}{\left(\sum_{t=1}^{T} x_t\right)^{\alpha}} = \left(\sum_{t=1}^{T} x_t\right)^{1-\alpha}.$$

The second inequality in the case $0 < \alpha < 1$ is proved in Lemma 3 of [29]. For $\alpha = 1$, it is proved in Lemma 3.2 of Ward et al. [48]. ■

Lemma 7 (smoothness of $\Phi(\cdot)$ and Lipschitzness of $y^*(\cdot)$. Lemma 4.3 in Lin et al. [35]) *Under Assumptions 3.1 and 3.2, we have $\Phi(\cdot)$ is $(l + \kappa l)$-smooth with $\nabla \Phi(x) = \nabla_x f(x, y^*(x))$, and $y^*(\cdot)$ is κ-Lipschitz.*

Lemma 8 (smoothness of $y^*(\cdot)$. Lemma 2 in Chen et al. [5]) *Under Assumptions 3.1, 3.2 and 3.6, we have for some constant $\widehat{L}$,*

$$\|\nabla y^*(x_1) - \nabla y^*(x_2)\| \leq \widehat{L}\|x_1 - x_2\|.$$

Appendix C. Proofs

In this section, we first show the proofs for theorems in Section 3 and then present an analysis for a TiAda variant without the max-operator.

For notational convenience in the proofs, we denote the stochastic gradient as $\nabla_x \widetilde{f}(x_t, y_t)$ and $\nabla_y \widetilde{f}(x_t, y_t)$. Also denote $y_t^* = y^*(x_t)$, $\eta_t = \frac{\eta^x}{\max\left\{v_{t+1}^x, v_{t+1}^y\right\}^{\alpha}}$, $\gamma_t = \frac{\eta^y}{\left(v_{t+1}^y\right)^{\beta}}$, $\Phi^* = \min_{x \in \mathbb{R}^{d_1}} \Phi(x)$, and $\Delta\Phi = \Phi_{\max} - \Phi^*$. We use $\mathbf{1}$ as the indicator function.

C.1. Proof of Theorem 2

We present a formal version of Theorem 2.

Theorem 9 (deterministic setting) *Under Assumptions 3.1 to 3.3, Algorithm 1 with deterministic gradient oracles satisfies that for any $0 < \beta < \alpha < 1$, after T iterations,*

$$\sum_{t=0}^{T-1} \|\nabla_x f(x_t, y_t)\|^2 \le \max\left\{5C_1, 2C_2\right\},$$

where

$$\begin{aligned}
C_1 &= v_0^x + \left(\frac{2\Delta\Phi}{\eta^x}\right)^{\frac{1}{1-\alpha}} + \left(\frac{4\kappa l e^{(1-\alpha)(1-\log v_0^x)/2}}{e(1-\alpha)\left(v_0^x\right)^{2\alpha-1}}\right)^{\frac{2}{1-\alpha}} \mathbf{1}_{2\alpha\ge1} + \left(\frac{2\kappa l}{1-2\alpha}\right)^{\frac{1}{\alpha}} \mathbf{1}_{2\alpha<1} \\
&\quad + \left(\frac{c_1 c_5}{\eta^x}\right)^{\frac{1}{1-\alpha}} + \left(\frac{2c_1 c_4 \eta^x e^{(1-\alpha)(1-\log v_0^x)/2}}{e(1-\alpha)\left(v_0^x\right)^{2\alpha-\beta-1}}\right)^{\frac{2}{1-\alpha}} \mathbf{1}_{2\alpha-\beta\ge1} + \left(\frac{c_1 c_4 \eta^x}{1-2\alpha+\beta}\right)^{\frac{1}{\alpha-\beta}} \mathbf{1}_{2\alpha-\beta<1} \\
C_2 &= v_0^x + \left[\left(\frac{2\Delta\Phi + c_1 c_5}{\eta^x \left(v_0^x\right)^{1-2\alpha+\beta}} + \frac{c_1 c_4 \eta^x}{1-2\alpha+\beta} + \frac{2\kappa l e^{(1-2\alpha+\beta)(1-\log v_0^x)}}{e(1-2\alpha+\beta)\left(v_0^x\right)^{2\alpha-1}}\mathbf{1}_{2\alpha\ge1}\right.\right. \\
&\quad \left.\left. + \frac{2\kappa l}{(1-2\alpha)\left(v_0^x\right)^{\beta}}\mathbf{1}_{2\alpha<1}\right)\left(\frac{c_5}{\left(v_0^x\right)^{1-2\alpha+\beta}} + \frac{c_4\left(\eta^x\right)^2}{1-2\alpha+\beta}\right)^{\frac{\alpha}{1-\beta}}\right]^{\frac{1}{1-(1-2\alpha+\beta)\left(1+\frac{\alpha}{1-\beta}\right)}} \mathbf{1}_{2\alpha-\beta<1} \\
&\quad + \left[\left(\frac{2\Delta\Phi + c_1 c_5}{\eta^x \left(v_0^x\right)^{1/4}} + \frac{8\kappa l e^{(1-\log v_0^x)/4}}{e\left(v_0^x\right)^{2\alpha-1}} + \frac{4c_1 c_4 \eta^x e^{(1-\log v_0^x)/4}}{e\left(v_0^x\right)^{2\alpha-\beta-1}}\right)\right. \\
&\quad \left.\left(\frac{c_5}{\left(v_0^x\right)^{\frac{(1-\beta)}{4\alpha}}} + \frac{4c_4\alpha\left(\eta^x\right)^2 e^{(1-\beta)(1-\log v_0^x)/(4\alpha)}}{e(1-\beta)\left(v_0^x\right)^{2\alpha-\beta-1}}\right)^{\frac{\alpha}{1-\beta}}\right]^2 \mathbf{1}_{2\alpha\ge1},
\end{aligned}$$

with $\Delta\Phi = \Phi(x_0) - \Phi^*, \quad c_1 = \dfrac{\eta^x\kappa^2}{\eta^y\left(v_{t_0}^y\right)^{\alpha-\beta}}, \quad c_2 = \max\left\{\dfrac{4\eta^y\mu l}{\mu+l}, \eta^y(\mu+l)\right\},$

$$c_3 = 4(\mu+l)\left(\frac{1}{\mu^2} + \frac{\eta^y}{\left(v_{t_0}^y\right)^{\beta}}\right)c_2^{1/\beta}, \quad c_4 = (\mu+l)\left(\frac{2\kappa^2}{\left(v_0^y\right)^{\alpha}} + \frac{(\mu+l)\kappa^2}{\eta^y\mu l}\right),$$

$$c_5 = c_3 + \frac{\eta^y v_0^y}{\left(v_0^y\right)^{\beta}} + \frac{\eta^y c_2^{\frac{1-\beta}{\beta}}}{1-\beta}.$$

In addition, denoting the above upper bound for $\sum_{t=0}^{T-1}\|\nabla_x f(x_t, y_t)\|^2$ as C_3, we have

$$\sum_{t=0}^{T-1}\|\nabla_y f(x_t, y_t)\|^2 \le \left(c_5 + c_4\left(\eta^x\right)^2\left(\frac{1+\log C_3 - \log v_0^x}{\left(v_0^x\right)^{2\alpha-\beta-1}}\mathbf{1}_{2\alpha-\beta\ge1} + \frac{C_3^{1-2\alpha+\beta}}{1-2\alpha+\beta}\mathbf{1}_{2\alpha-\beta<1}\right)\right)^{\frac{1}{1-\beta}}.$$

Proof Let us start from the smoothness of the primal function $\Phi(\cdot)$. By Theorem 7,

$$\Phi(x_{t+1})$$

$$
\begin{aligned}
&\leq \Phi(x_t) - \eta_t \langle \Phi(x_{t+1}), \nabla_x f(x_t, y_t) \rangle + kl\eta_t^2 \|\nabla_x f(x_t, y_t)\|^2 \\
&= \Phi(x_t) - \eta_t \|\nabla_x f(x_t, y_t)\|^2 + \eta_t \langle \nabla_x f(x_t, y_t) - \nabla \Phi(x_t), \nabla_x f(x_t, y_t) \rangle + kl\eta_t^2 \|\nabla_x f(x_t, y_t)\|^2 \\
&\leq \Phi(x_t) - \eta_t \|\nabla_x f(x_t, y_t)\|^2 + \frac{\eta_t}{2} \|\nabla_x f(x_t, y_t)\|^2 + \frac{\eta_t}{2} \|\nabla_x f(x_t, y_t) - \nabla \Phi(x_t)\|^2 + kl\eta_t^2 \|\nabla_x f(x_t, y_t)\|^2 \\
&= \Phi(x_t) - \frac{\eta_t}{2} \|\nabla_x f(x_t, y_t)\|^2 + kl\eta_t^2 \|\nabla_x f(x_t, y_t)\|^2 + \frac{\eta_t}{2} \|\nabla_x f(x_t, y_t) - \nabla \Phi(x_t)\|^2 \\
&= \Phi(x_t) - \frac{\eta_t}{2} \|\nabla_x f(x_t, y_t)\|^2 + kl\eta_t^2 \|\nabla_x f(x_t, y_t)\|^2 + \frac{\eta^x}{2 \max\left\{v_{t+1}^x, v_{t+1}^y\right\}^\alpha} \|\nabla_x f(x_t, y_t) - \nabla \Phi(x_t)\|^2 \\
&\leq \Phi(x_t) - \frac{\eta_t}{2} \|\nabla_x f(x_t, y_t)\|^2 + kl\eta_t^2 \|\nabla_x f(x_t, y_t)\|^2 + \frac{\eta^x}{2\left(v_{t_0}^y\right)^{\alpha-\beta} \left(v_{t+1}^y\right)^\beta} \|\nabla_x f(x_t, y_t) - \nabla \Phi(x_t)\|^2 \\
&\leq \Phi(x_t) - \frac{\eta_t}{2} \|\nabla_x f(x_t, y_t)\|^2 + kl\eta_t^2 \|\nabla_x f(x_t, y_t)\|^2 + \frac{\eta^x \kappa^2}{2\left(v_{t_0}^y\right)^{\alpha-\beta} \left(v_{t+1}^y\right)^\beta} \|\nabla_y f(x_t, y_t)\|^2 \\
&\leq \Phi(x_t) - \frac{\eta_t}{2} \|\nabla_x f(x_t, y_t)\|^2 + kl\eta_t^2 \|\nabla_x f(x_t, y_t)\|^2 + \frac{\eta^x \kappa^2}{2\eta^y \left(v_{t_0}^y\right)^{\alpha-\beta}} \cdot \gamma_t \|\nabla_y f(x_t, y_t)\|^2,
\end{aligned}
$$

where in the second to last inequality, we used the strong-concavity of $f(x, \cdot)$:

$$
\|\nabla_x f(x_t, y_t) - \nabla \Phi(x_t)\| \leq l \|y_t - y_t^*\| \leq \kappa \|\nabla_y f(x_t, y_t)\|.
$$

Telescoping and rearranging the terms, we have

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \eta_t \|\nabla_x f(x_t, y_t)\|^2 \\
&\leq 2 \underbrace{(\Phi(x_0) - \Phi^*)}_{\Delta\Phi} + 2\kappa l \sum_{t=0}^{T-1} \eta_t^2 \|\nabla_x f(x_t, y_t)\|^2 + \underbrace{\frac{\eta^x \kappa^2}{\eta^y \left(v_{t_0}^y\right)^{\alpha-\beta}}}_{c_1} \sum_{t=0}^{T-1} \gamma_t \|\nabla_y f(x_t, y_t)\|^2 \\
&= 2\Delta\Phi + \sum_{t=0}^{T-1} \frac{2\kappa l \eta^x}{\max\left\{v_{t+1}^x, v_{t+1}^y\right\}^{2\alpha}} \|\nabla_x f(x_t, y_t)\|^2 + c_1 \sum_{t=0}^{T-1} \gamma_t \|\nabla_y f(x_t, y_t)\|^2 \\
&\leq 2\Delta\Phi + \sum_{t=0}^{T-1} \frac{2\kappa l \eta^x}{\left(v_{t+1}^x\right)^{2\alpha}} \|\nabla_x f(x_t, y_t)\|^2 + c_1 \sum_{t=0}^{T-1} \gamma_t \|\nabla_y f(x_t, y_t)\|^2 \\
&\leq 2\Delta\Phi + 2\kappa l \eta^x \left(\frac{1 + \log v_T^x - \log v_0^x}{\left(v_0^x\right)^{2\alpha-1}} \cdot \mathbf{1}_{2\alpha \geq 1} + \frac{\left(v_T^x\right)^{1-2\alpha}}{1 - 2\alpha} \cdot \mathbf{1}_{2\alpha < 1} \right) + c_1 \sum_{t=0}^{T-1} \gamma_t \|\nabla_y f(x_t, y_t)\|^2.
\end{aligned} \tag{5}
$$

We proceed to bound $\sum_{t=0}^{T-1} \gamma_t \|\nabla_y f(x_t, y_t)\|^2$. Let t_0 be the first iteration such that $\left(v_{t_0+1}^y\right)^\beta > c_2 := \max\left\{\frac{4\eta^y \mu l}{\mu + l}, \eta^y(\mu + l)\right\}$. We have $v_{t_0}^y \leq c_2^{1/\beta}$, and for $t \geq t_0$,

$$
\begin{aligned}
&\left\|y_{t+1} - y_{t+1}^*\right\|^2 \\
&\leq (1 + \lambda_t) \|y_{t+1} - y_t^*\|^2 + \left(1 + \frac{1}{\lambda_t}\right) \left\|y_{t+1}^* - y_t^*\right\|^2
\end{aligned}
$$

$$\leq \underbrace{(1+\lambda_t)\left(\|y_t - y_t^*\|^2 + \frac{(\eta^y)^2}{\left(v_{t+1}^y\right)^{2\beta}}\|\nabla_y f(x_t,y_t)\|^2 + \frac{2\eta^y}{\left(v_{t+1}^y\right)^{\beta}}\langle y_t - y_t^*, \nabla_y f(x_t,y_t)\rangle\right)}_{\text{(A)}}$$
$$+\left(1+\frac{1}{\lambda_t}\right)\left\|y_{t+1}^* - y_t^*\right\|^2,$$

where $\lambda_t > 0$ will be determined later. For l-smooth and μ-strongly convex function $g(x)$, according to Theorem 2.1.12 in Nesterov [40], we have

$$\langle \nabla g(x) - \nabla g(y), x-y\rangle \geq \frac{\mu l}{\mu+l}\|x-y\|^2 + \frac{1}{\mu+l}\|\nabla g(x) - \nabla g(y)\|^2.$$

Therefore,

$$\begin{aligned}
&\text{Term (A)}\\
&\leq (1+\lambda_t)\left(\left(1 - \frac{2\eta^y \mu l}{(\mu+l)\left(v_{t+1}^y\right)^{\beta}}\right)\|y_t - y_t^*\|^2 + \left(\frac{(\eta^y)^2}{\left(v_{t+1}^y\right)^{2\beta}} - \frac{2\eta^y}{(\mu+l)\left(v_{t+1}^y\right)^{\beta}}\right)\|\nabla_y f(x_t,y_t)\|^2\right).
\end{aligned}$$

Let $\lambda_t = \frac{\eta^y \mu l}{(\mu+l)\left(v_{t+1}^y\right)^{\beta} - 2\eta^y \mu l}$. Note that $\lambda_t > 0$ after t_0. Then

$$\begin{aligned}
&\text{Term (A)}\\
&\leq \left(1 - \frac{\eta^y \mu l}{(\mu+l)\left(v_{t+1}^y\right)^{\beta}}\right)\|y_t - y_t^*\|^2 + (1+\lambda_t)\left(\frac{(\eta^y)^2}{\left(v_{t+1}^y\right)^{2\beta}} - \frac{2\eta^y}{(\mu+l)\left(v_{t+1}^y\right)^{\beta}}\right)\|\nabla_y f(x_t,y_t)\|^2\\
&\leq \|y_t - y_t^*\|^2 + \underbrace{(1+\lambda_t)\left(\frac{(\eta^y)^2}{\left(v_{t+1}^y\right)^{2\beta}} - \frac{2\eta^y}{(\mu+l)\left(v_{t+1}^y\right)^{\beta}}\right)}_{\text{(B)}}\|\nabla_y f(x_t,y_t)\|^2.
\end{aligned}$$

As $1+\lambda_t \geq 1$ and $\left(v_{t+1}^y\right)^{\beta} \geq \eta^y(\mu+l)$, we have term (B) $\leq -\frac{\eta^y}{(\mu+l)\left(v_{t+1}^y\right)^{\beta}}$. Putting them back, we can get

$$\begin{aligned}
&\left\|y_{t+1} - y_{t+1}^*\right\|^2\\
&\leq \|y_t - y_t^*\|^2 - \frac{\eta^y}{(\mu+l)\left(v_{t+1}^y\right)^{\beta}}\|\nabla_y f(x_t,y_t)\|^2 + \left(1+\frac{1}{\lambda_t}\right)\left\|y_{t+1}^* - y_t^*\right\|^2\\
&\leq \|y_t - y_t^*\|^2 - \frac{\eta^y}{(\mu+l)\left(v_{t+1}^y\right)^{\beta}}\|\nabla_y f(x_t,y_t)\|^2 + \frac{(\mu+l)\left(v_{t+1}^y\right)^{\beta}}{\eta^y \mu l}\left\|y_{t+1}^* - y_t^*\right\|^2\\
&\leq \|y_t - y_t^*\|^2 - \frac{\eta^y}{(\mu+l)\left(v_{t+1}^y\right)^{\beta}}\|\nabla_y f(x_t,y_t)\|^2 + \frac{(\mu+l)\kappa^2\left(v_{t+1}^y\right)^{\beta}}{\eta^y \mu l}\|x_{x+1} - x_t\|^2\\
&= \|y_t - y_t^*\|^2 - \frac{\eta^y}{(\mu+l)\left(v_{t+1}^y\right)^{\beta}}\|\nabla_y f(x_t,y_t)\|^2 + \frac{(\mu+l)\kappa^2\left(v_{t+1}^y\right)^{\beta}\eta_t^2}{\eta^y \mu l}\|\nabla_x f(x_t,y_t)\|^2.
\end{aligned}$$

Then, by telescoping, we have

$$\sum_{t=t_0}^{T-1} \frac{\eta^y}{(\mu+l)\left(v_{t+1}^y\right)^\beta} \|\nabla_y f(x_t, y_t)\|^2 \leq \left\|y_{t_0} - y_{t_0}^*\right\|^2 + \sum_{t=t_0}^{T-1} \frac{(\mu+l)\kappa^2 \left(v_{t+1}^y\right)^\beta \eta_t^2}{\eta^y \mu l} \|\nabla_x f(x_t, y_t)\|^2. \tag{6}$$

For the first term in the RHS, using Young's inequality with τ to be determined later, we have

$$\begin{aligned}
\left\|y_{t_0} - y_{t_0}^*\right\|^2 &\leq 2\left\|y_{t_0} - y_{t_0-1}^*\right\|^2 + 2\left\|y_{t_0}^* - y_{t_0-1}^*\right\|^2 \\
&= 2\left\|\mathcal{P}_{\mathcal{Y}}\left(y_{t_0-1} + \gamma_{t_0-1}\nabla_y f(x_{t_0-1}, y_{t_0-1})\right) - y_{t_0-1}^*\right\|^2 + 2\left\|y_{t_0}^* - y_{t_0-1}^*\right\|^2 \\
&\leq 2\left\|y_{t_0-1} + \gamma_{t_0-1}\nabla_y f(x_{t_0-1}, y_{t_0-1}) - y_{t_0-1}^*\right\|^2 + 2\left\|y_{t_0}^* - y_{t_0-1}^*\right\|^2 \\
&\leq 4\left(\left\|y_{t_0-1} - y_{t_0-1}^*\right\|^2 + \gamma_{t_0-1}^2\|\nabla_y f(x_{t_0-1}, y_{t_0-1})\|^2\right) + 2\left\|y_{t_0}^* - y_{t_0-1}^*\right\|^2 \\
&\leq 4\left(\frac{1}{\mu^2}\|\nabla_y f(x_{t_0-1}, y_{t_0-1})\|^2 + \gamma_{t_0-1}^2\|\nabla_y f(x_{t_0-1}, y_{t_0-1})\|^2\right) + 2\left\|y_{t_0}^* - y_{t_0-1}^*\right\|^2 \\
&= 4\left(\frac{1}{\mu^2} + \gamma_{t_0-1}^2\right)\|\nabla_y f(x_{t_0-1}, y_{t_0-1})\|^2 + 2\left\|y_{t_0}^* - y_{t_0-1}^*\right\|^2 \\
&\leq 4\left(\frac{1}{\mu^2} + \gamma_0^2\right) v_{t_0}^y + 2\left\|y_{t_0}^* - y_{t_0-1}^*\right\|^2 \\
&\leq 4\left(\frac{1}{\mu^2} + \frac{\eta^y}{\left(v_{t_0}^y\right)^\beta}\right) c_2^{1/\beta} + 2\left\|y_{t_0}^* - y_{t_0-1}^*\right\|^2 \\
&\leq 4\left(\frac{1}{\mu^2} + \frac{\eta^y}{\left(v_{t_0}^y\right)^\beta}\right) c_2^{1/\beta} + 2\kappa^2\|x_{t_0} - x_{t_0-1}\|^2 \\
&\leq 4\left(\frac{1}{\mu^2} + \frac{\eta^y}{\left(v_{t_0}^y\right)^\beta}\right) c_2^{1/\beta} + 2\kappa^2\eta_{t_0-1}^2\|\nabla_x f(x_{t_0-1}, y_{t_0-1})\|^2 \\
&\leq 4\left(\frac{1}{\mu^2} + \frac{\eta^y}{\left(v_{t_0}^y\right)^\beta}\right) c_2^{1/\beta} + \frac{2\kappa^2\left(v_{t+1}^y\right)^\beta}{\left(v_0^y\right)^\beta}\eta_{t_0-1}^2\|\nabla_x f(x_{t_0-1}, y_{t_0-1})\|^2.
\end{aligned}$$

Combined with Equation (6), we have

$$\begin{aligned}
&\sum_{t=t_0}^{T-1} \frac{\eta^y}{\left(v_{t+1}^y\right)^\beta}\|\nabla_y f(x_t, y_t)\|^2 \\
&\leq \underbrace{4(\mu+l)\left(\frac{1}{\mu^2} + \frac{\eta^y}{\left(v_{t_0}^y\right)^\beta}\right) c_2^{1/\beta}}_{c_3} + \underbrace{(\mu+l)\left(\frac{2\kappa^2}{\left(v_0^y\right)^\alpha} + \frac{(\mu+l)\kappa^2}{\eta^y\mu l}\right)}_{c_4} \sum_{t=t_0-1}^{T-1}\left(v_{t+1}^y\right)^\beta \eta_t^2\|\nabla_x f(x_t, y_t)\|^2.
\end{aligned}$$

By adding terms from 0 to $t_0 - 1$ and $\frac{\eta^y v_0^y}{\left(v_0^y\right)^\beta}$ from both sides, we have

$$\frac{\eta^y v_0^y}{\left(v_0^y\right)^\beta} + \sum_{t=0}^{T-1} \frac{\eta^y}{\left(v_{t+1}^y\right)^\beta}\|\nabla_y f(x_t, y_t)\|^2$$

$$
\begin{aligned}
&\leq c_3 + \frac{\eta^y v_0^y}{\left(v_0^y\right)^\beta} + c_4 \sum_{t=0}^{T-1} \left(v_{t+1}^y\right)^\beta \eta_t^2 \|\nabla_x f(x_t, y_t)\|^2 + \sum_{t=t=0}^{t_0-1} \frac{\eta^y}{\left(v_{t+1}^y\right)^\beta} \|\nabla_y f(x_t, y_t)\|^2 \\
&\leq c_3 + \frac{\eta^y v_0^y}{\left(v_0^y\right)^\beta} + c_4 \sum_{t=0}^{T-1} \left(v_{t+1}^y\right)^\beta \eta_t^2 \|\nabla_x f(x_t, y_t)\|^2 + \frac{\eta^y v_0^y}{\left(v_0^y\right)^\beta} + \sum_{t=t=0}^{t_0-1} \frac{\eta^y}{\left(v_{t+1}^y\right)^\beta} \|\nabla_y f(x_t, y_t)\|^2 \\
&\leq c_3 + \frac{\eta^y v_0^y}{\left(v_0^y\right)^\beta} + c_4 \sum_{t=0}^{T-1} \left(v_{t+1}^y\right)^\beta \eta_t^2 \|\nabla_x f(x_t, y_t)\|^2 + \frac{\eta^y}{1-\beta} v_{t_0}^{1-\beta} \\
&\leq c_3 + \frac{\eta^y v_0^y}{\left(v_0^y\right)^\beta} + c_4 \sum_{t=0}^{T-1} \left(v_{t+1}^y\right)^\beta \eta_t^2 \|\nabla_x f(x_t, y_t)\|^2 + \frac{\eta^y c_2^{\frac{1-\beta}{\beta}}}{1-\beta} \\
&= c_3 + \frac{\eta^y v_0^y}{\left(v_0^y\right)^\beta} + \frac{\eta^y c_2^{\frac{1-\beta}{\beta}}}{1-\beta} + c_4 \left(\eta^x\right)^2 \sum_{t=0}^{T-1} \frac{\left(v_{t+1}^y\right)^\beta}{\max\left\{v_{t+1}^x, v_{t+1}^y\right\}^{2\alpha}} \|\nabla_x f(x_t, y_t)\|^2 \\
&= \underbrace{c_3 + \frac{\eta^y v_0^y}{\left(v_0^y\right)^\beta} + \frac{\eta^y c_2^{\frac{1-\beta}{\beta}}}{1-\beta}}_{c_5} + c_4 \left(\eta^x\right)^2 \sum_{t=0}^{T-1} \frac{1}{\left(v_{t+1}^x\right)^{2\alpha-\beta}} \|\nabla_x f(x_t, y_t)\|^2 \\
&\leq c_5 + c_4 \left(\eta^x\right)^2 \left(\frac{1 + \log v_T^x - \log v_0^x}{\left(v_0^x\right)^{2\alpha-\beta-1}} \cdot \mathbf{1}_{2\alpha-\beta\geq 1} + \frac{\left(v_T^x\right)^{1-2\alpha+\beta}}{1-2\alpha+\beta} \cdot \mathbf{1}_{2\alpha-\beta<1} \right).
\end{aligned}
$$

The LHS can be bounded by $\left(v_T^y\right)^{1-\beta}$ by Theorem 6. Then we get two useful inequalities from above:

$$
\begin{cases}
\sum_{t=0}^{T-1} \gamma_t \|\nabla_y f(x_t, y_t)\|^2 \leq c_5 + c_4 \left(\eta^x\right)^2 \left(\frac{1+\log v_T^x - \log v_0^x}{\left(v_0^x\right)^{2\alpha-\beta-1}} \cdot \mathbf{1}_{2\alpha-\beta\geq 1} + \frac{\left(v_T^x\right)^{1-2\alpha+\beta}}{1-2\alpha+\beta} \cdot \mathbf{1}_{2\alpha-\beta<1} \right) \\
v_T^y \leq \left(c_5 + c_4 \left(\eta^x\right)^2 \left(\frac{1+\log v_T^x - \log v_0^x}{\left(v_0^x\right)^{2\alpha-\beta-1}} \cdot \mathbf{1}_{2\alpha-\beta\geq 1} + \frac{\left(v_T^x\right)^{1-2\alpha+\beta}}{1-2\alpha+\beta} \cdot \mathbf{1}_{2\alpha-\beta<1} \right) \right)^{\frac{1}{1-\beta}}.
\end{cases}
\tag{7}
$$

Now bring it back to Equation (5), we get

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \eta_t \|\nabla_x f(x_t, y_t)\|^2 \\
&\leq 2\Delta\Phi + 2\kappa l \eta^x \left(\frac{1 + \log v_T^x - \log v_0^x}{\left(v_0^x\right)^{2\alpha-1}} \cdot \mathbf{1}_{2\alpha\geq 1} + \frac{\left(v_T^x\right)^{1-2\alpha}}{1-2\alpha} \cdot \mathbf{1}_{2\alpha<1} \right) \\
&\quad + c_1 c_5 + c_1 c_4 \left(\eta^x\right)^2 \left(\frac{1 + \log v_T^x - \log v_0^x}{\left(v_0^x\right)^{2\alpha-\beta-1}} \cdot \mathbf{1}_{2\alpha-\beta\geq 1} + \frac{\left(v_T^x\right)^{1-2\alpha+\beta}}{1-2\alpha+\beta} \cdot \mathbf{1}_{2\alpha-\beta<1} \right).
\end{aligned}
$$

For the LHS, we have

$$
\sum_{t=0}^{T-1} \eta_t \|\nabla_x f(x_t, y_t)\|^2 = \sum_{t=0}^{T-1} \frac{\eta^x}{\max\left\{v_{t+1}^x, v_{t+1}^y\right\}^\alpha} \|\nabla_x f(x_t, y_t)\|^2
$$

$$\geq \frac{\eta^x}{\max\left\{v_T^x, v_T^y\right\}^\alpha} \sum_{t=0}^{T-1} \|\nabla_x f(x_t, y_t)\|^2$$

From here, by combining two inequalites above and noting that $\sum_{t=0}^{T-1} \|\nabla_x f(x_t, y_t)\|^2 \leq v_T^x$, we can already conclude that $\sum_{t=0}^{T-1} \|\nabla_x f(x_t, y_t)\|^2 = \mathcal{O}(1)$. Now we will provide an explicit bound. We consider two cases:

(1) If $v_T^y \leq v_T^x$, then

$$\begin{aligned}
&\sum_{t=0}^{T-1} \|\nabla_x f(x_t, y_t)\|^2 \\
&\leq \frac{2\Delta\Phi \left(v_T^x\right)^\alpha}{\eta^x} + 2\kappa l \left(\frac{\left(v_T^x\right)^\alpha \left(1 + \log v_T^x - \log v_0^x\right)}{\left(v_0^x\right)^{2\alpha - 1}} \cdot \mathbf{1}_{2\alpha \geq 1} + \frac{\left(v_T^x\right)^{1-\alpha}}{1 - 2\alpha} \cdot \mathbf{1}_{2\alpha < 1} \right) \\
&\quad + \frac{c_1 c_5 \left(v_T^x\right)^\alpha}{\eta^x} + c_1 c_4 \eta^x \left(\frac{\left(v_T^x\right)^\alpha \left(1 + \log v_T^x - \log v_0^x\right)}{\left(v_0^x\right)^{2\alpha - \beta - 1}} \cdot \mathbf{1}_{2\alpha - \beta \geq 1} + \frac{\left(v_T^x\right)^{1-\alpha+\beta}}{1 - 2\alpha + \beta} \cdot \mathbf{1}_{2\alpha - \beta < 1} \right) \\
&= \frac{2\Delta\Phi \left(v_T^x\right)^\alpha}{\eta^x} + 2\kappa l \left(\frac{\left(v_T^x\right)^\alpha \left(v_T^x\right)^{\frac{1-\alpha}{2}} \left(v_T^x\right)^{\frac{\alpha-1}{2}} \left(1 + \log v_T^x - \log v_0^x\right)}{\left(v_0^x\right)^{2\alpha - 1}} \cdot \mathbf{1}_{2\alpha \geq 1} + \frac{\left(v_T^x\right)^{1-\alpha}}{1 - 2\alpha} \cdot \mathbf{1}_{2\alpha < 1} \right) \\
&\quad + \frac{c_1 c_5 \left(v_T^x\right)^\alpha}{\eta^x} + c_1 c_4 \eta^x \left(\frac{\left(v_T^x\right)^\alpha \left(v_T^x\right)^{\frac{1-\alpha}{2}} \left(v_T^x\right)^{\frac{\alpha-1}{2}} \left(1 + \log v_T^x - \log v_0^x\right)}{\left(v_0^x\right)^{2\alpha - \beta - 1}} \cdot \mathbf{1}_{2\alpha - \beta \geq 1} + \frac{\left(v_T^x\right)^{1-\alpha+\beta}}{1 - 2\alpha + \beta} \cdot \mathbf{1}_{2\alpha - \beta < 1} \right) \\
&\leq \frac{2\Delta\Phi \left(v_T^x\right)^\alpha}{\eta^x} + 2\kappa l \left(\frac{2 e^{(1-\alpha)(1 - \log v_0^x)/2} \left(v_T^x\right)^{\frac{1+\alpha}{2}}}{e(1-\alpha) \left(v_0^x\right)^{2\alpha - 1}} \cdot \mathbf{1}_{2\alpha \geq 1} + \frac{\left(v_T^x\right)^{1-\alpha}}{1 - 2\alpha} \cdot \mathbf{1}_{2\alpha < 1} \right) \\
&\quad + \frac{c_1 c_5 \left(v_T^x\right)^\alpha}{\eta^x} + c_1 c_4 \eta^x \left(\frac{2 e^{(1-\alpha)(1 - \log v_0^x)/2} \left(v_T^x\right)^{\frac{1+\alpha}{2}}}{e(1-\alpha) \left(v_0^x\right)^{2\alpha - \beta - 1}} \cdot \mathbf{1}_{2\alpha - \beta \geq 1} + \frac{\left(v_T^x\right)^{1-\alpha+\beta}}{1 - 2\alpha + \beta} \cdot \mathbf{1}_{2\alpha - \beta < 1} \right),
\end{aligned} \tag{8}$$

where we used $x^{-m}(c + \log x) \leq \frac{e^{cm}}{em}$ for $x > 0$, $m > 0$ and $c \in \mathbb{R}$ in the last inequality. Also, if $0 < \alpha_i < 1$ and b_i are positive constants, and $x \leq \sum_{i=1}^n b_i x^{\alpha_i}$, then we get $x \leq n \sum_{i=1}^n b_i^{1/(1-\alpha_i)}$. Now consider v_T^x as the x in the previous statement, and note that the LHS of Equation (8) equals to $v_T^x - v_0^x$. Then we can get

$$\begin{aligned}
v_T^x \leq 5 v_0^x &+ 5 \left(\frac{2\Delta\Phi}{\eta^x} \right)^{\frac{1}{1-\alpha}} + 5 \left(\frac{4\kappa l e^{(1-\alpha)(1 - \log v_0^x)/2}}{e(1-\alpha) \left(v_0^x\right)^{2\alpha - 1}} \right)^{\frac{2}{1-\alpha}} \cdot \mathbf{1}_{2\alpha \geq 1} + 5 \left(\frac{2\kappa l}{1 - 2\alpha} \right)^{\frac{1}{\alpha}} \cdot \mathbf{1}_{2\alpha < 1} \\
&+ 5 \left(\frac{c_1 c_5}{\eta^x} \right)^{\frac{1}{1-\alpha}} + 5 \left(\frac{2 c_1 c_4 \eta^x e^{(1-\alpha)(1 - \log v_0^x)/2}}{e(1-\alpha) \left(v_0^x\right)^{2\alpha - \beta - 1}} \right)^{\frac{2}{1-\alpha}} \cdot \mathbf{1}_{2\alpha - \beta \geq 1} + 5 \left(\frac{c_1 c_4 \eta^x}{1 - 2\alpha + \beta} \right)^{\frac{1}{\alpha - \beta}} \cdot \mathbf{1}_{2\alpha - \beta < 1}.
\end{aligned} \tag{9}$$

Note that the RHS is a constant and also an upper bound for $\sum_{t=0}^{T-1} \|\nabla_x f(x_t, y_t)\|^2$.

(2) If $v_T^y \leq v_T^x$, then we can use the upper bound for v_T^y from Equation (7). We now discuss two cases:

1. $2\alpha < 1 + \beta$. Then we have

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \|\nabla_x f(x_t, y_t)\|^2 \\
\leq &\left(\frac{2\Delta\Phi + c_1 c_5}{\eta^x} + 2\kappa l \left(\frac{1 + \log v_T^x - \log v_0^x}{\left(v_0^x\right)^{2\alpha-1}} \cdot \mathbf{1}_{2\alpha \geq 1} + \frac{\left(v_T^x\right)^{1-2\alpha}}{1-2\alpha} \cdot \mathbf{1}_{2\alpha<1} \right) + \frac{c_1 c_4 \eta^x \left(v_T^x\right)^{1-2\alpha+\beta}}{1-2\alpha+\beta} \right) \\
&\left(c_5 + \frac{c_4 \left(\eta^x\right)^2 \left(v_T^x\right)^{1-2\alpha+\beta}}{1-2\alpha+\beta} \right)^{\frac{\alpha}{1-\beta}} \\
\leq &\left(\frac{2\Delta\Phi + c_1 c_5}{\eta^x \left(v_0^x\right)^{1-2\alpha+\beta}} + 2\kappa l \left(\frac{1 + \log v_T^x - \log v_0^x}{\left(v_0^x\right)^{2\alpha-1} \left(v_T^x\right)^{1-2\alpha+\beta}} \cdot \mathbf{1}_{2\alpha \geq 1} + \frac{1}{(1-2\alpha)\left(v_0^x\right)^{\beta}} \cdot \mathbf{1}_{2\alpha<1} \right) + \frac{c_1 c_4 \eta^x}{1-2\alpha+\beta} \right) \\
&\left(\frac{c_5}{\left(v_0^x\right)^{1-2\alpha+\beta}} + \frac{c_4 \left(\eta^x\right)^2}{1-2\alpha+\beta} \right)^{\frac{\alpha}{1-\beta}} \cdot \left(v_T^x\right)^{1-2\alpha+\beta+\frac{(1-2\alpha+\beta)\alpha}{1-\beta}} \\
\leq &\left(\frac{2\Delta\Phi + c_1 c_5}{\eta^x \left(v_0^x\right)^{1-2\alpha+\beta}} + 2\kappa l \left(\frac{e^{(1-2\alpha+\beta)(1-\log v_0^x)}}{e(1-2\alpha+\beta)\left(v_0^x\right)^{2\alpha-1}} \cdot \mathbf{1}_{2\alpha \geq 1} + \frac{1}{(1-2\alpha)\left(v_0^x\right)^{\beta}} \cdot \mathbf{1}_{2\alpha<1} \right) + \frac{c_1 c_4 \eta^x}{1-2\alpha+\beta} \right) \\
&\left(\frac{c_5}{\left(v_0^x\right)^{1-2\alpha+\beta}} + \frac{c_4 \left(\eta^x\right)^2}{1-2\alpha+\beta} \right)^{\frac{\alpha}{1-\beta}} \cdot \left(v_T^x\right)^{1-2\alpha+\beta+\frac{(1-2\alpha+\beta)\alpha}{1-\beta}},
\end{aligned}
$$

Note that since $\alpha > \beta$, we have

$$
1 - 2\alpha + \beta + \frac{(1-2\alpha+\beta)\alpha}{1-\beta} \leq \frac{(1-\alpha)\alpha}{1-\beta} + 1 - \alpha = 1 + \frac{\alpha(\beta-\alpha)}{1-\beta} < 1.
$$

Therefore, with the same reasoning as Equation (9),

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \|\nabla_x f(x_t, y_t)\|^2 \leq v_T^x \\
\leq 2 &\Bigg[\left(\frac{2\Delta\Phi + c_1 c_5}{\eta^x \left(v_0^x\right)^{1-2\alpha+\beta}} + \frac{c_1 c_4 \eta^x}{1-2\alpha+\beta} + \frac{2\kappa l e^{(1-2\alpha+\beta)(1-\log v_0^x)}}{e(1-2\alpha+\beta)\left(v_0^x\right)^{2\alpha-1}} \cdot \mathbf{1}_{2\alpha \geq 1} + \frac{2\kappa l}{(1-2\alpha)\left(v_0^x\right)^{\beta}} \cdot \mathbf{1}_{2\alpha<1} \right) \\
&\left(\frac{c_5}{\left(v_0^x\right)^{1-2\alpha+\beta}} + \frac{c_4 \left(\eta^x\right)^2}{1-2\alpha+\beta} \right)^{\frac{\alpha}{1-\beta}} \Bigg]^{\frac{1}{1-(1-2\alpha+\beta)\left(1+\frac{\alpha}{1-\beta}\right)}} + 2v_0^x,
\end{aligned}
$$

which gives us constant RHS.

2. $2\alpha \geq 1 + \beta$. Then we have

$$
\sum_{t=0}^{T-1} \|\nabla_x f(x_t, y_t)\|^2
$$

$$\begin{aligned}
&\le \left(\frac{2\Delta\Phi + c_1c_5}{\eta^x} + \frac{2\kappa l\left(1+\log v_T^x - \log v_0^x\right)}{\left(v_0^x\right)^{2\alpha-1}} + \frac{c_1c_4\eta^x\left(1+\log v_T^x - \log v_0^x\right)}{\left(v_0^x\right)^{2\alpha-\beta-1}}\right) \\
&\quad \left(c_5 + \frac{c_4\left(\eta^x\right)^2\left(1+\log v_T^x - \log v_0^x\right)}{\left(v_0^x\right)^{2\alpha-\beta-1}}\right)^{\frac{\alpha}{1-\beta}} \\
&\le \left(\frac{2\Delta\Phi + c_1c_5}{\eta^x\left(v_0^x\right)^{1/4}} + \frac{2\kappa l\left(1+\log v_T^x - \log v_0^x\right)}{\left(v_0^x\right)^{2\alpha-1}\left(v_T^x\right)^{1/4}} + \frac{c_1c_4\eta^x\left(1+\log v_T^x - \log v_0^x\right)}{\left(v_0^x\right)^{2\alpha-\beta-1}\left(v_T^x\right)^{1/4}}\right) \\
&\quad \left(\frac{c_5}{\left(v_0^x\right)^{\frac{(1-\beta)}{4\alpha}}} + \frac{c_4\left(\eta^x\right)^2\left(1+\log v_T^x - \log v_0^x\right)}{\left(v_0^x\right)^{2\alpha-\beta-1}\left(v_T^x\right)^{\frac{(1-\beta)}{4\alpha}}}\right)^{\frac{\alpha}{1-\beta}} \cdot \left(v_T^x\right)^{1/2} \\
&\le \left(\frac{2\Delta\Phi + c_1c_5}{\eta^x\left(v_0^x\right)^{1/4}} + \frac{8\kappa l e^{(1-\log v_0^x)/4}}{e\left(v_0^x\right)^{2\alpha-1}} + \frac{4c_1c_4\eta^x e^{(1-\log v_0^x)/4}}{e\left(v_0^x\right)^{2\alpha-\beta-1}}\right) \\
&\quad \left(\frac{c_5}{\left(v_0^x\right)^{\frac{(1-\beta)}{4\alpha}}} + \frac{4c_4\alpha\left(\eta^x\right)^2 e^{(1-\beta)(1-\log v_0^x)/(4\alpha)}}{e(1-\beta)\left(v_0^x\right)^{2\alpha-\beta-1}}\right)^{\frac{\alpha}{1-\beta}} \cdot \left(v_T^x\right)^{1/2},
\end{aligned}$$

which implies

$$\begin{aligned}
&\sum_{t=0}^{T-1}\left\|\nabla_x f(x_t, y_t)\right\|^2 \le v_T^x \\
&\le 2\left[\left(\frac{2\Delta\Phi + c_1c_5}{\eta^x\left(v_0^x\right)^{1/4}} + \frac{8\kappa l e^{(1-\log v_0^x)/4}}{e\left(v_0^x\right)^{2\alpha-1}} + \frac{4c_1c_4\eta^x e^{(1-\log v_0^x)/4}}{e\left(v_0^x\right)^{2\alpha-\beta-1}}\right)\right. \\
&\quad \left.\left(\frac{c_5}{\left(v_0^x\right)^{\frac{(1-\beta)}{4\alpha}}} + \frac{4c_4\alpha\left(\eta^x\right)^2 e^{(1-\beta)(1-\log v_0^x)/(4\alpha)}}{e(1-\beta)\left(v_0^x\right)^{2\alpha-\beta-1}}\right)^{\frac{\alpha}{1-\beta}}\right]^2 + 2v_0^x.
\end{aligned}$$

Now we also get only a constant on the RHS.

Summarizing all the cases, we finish the proof. ■

C.2. Intermediate Lemmas for Theorem 5

Lemma 10 *Under the same setting as Theorem 5, if for $t = t_0$ to $t_1 - 1$ and any $\lambda_t > 0$, S_t,*

$$\left\|y_{t+1} - y_{t+1}^*\right\|^2 \le (1+\lambda_t)\left\|y_{t+1} - y_t^*\right\|^2 + S_t,$$

then we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=t_0}^{t_1-1}\left(f(x_t, y_t^*) - f(x_t, y_t)\right)\right] &\le \mathbb{E}\left[\sum_{t=t_0+1}^{t_1-1}\left(\frac{1-\gamma_t\mu}{2\gamma_t}\left\|y_t - y_t^*\right\|^2 - \frac{1}{2\gamma_t(1+\lambda_t)}\left\|y_{t+1} - y_{t+1}^*\right\|^2\right)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=t_0}^{t_1-1}\frac{\gamma_t}{2}\left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2\right] + \mathbb{E}\left[\sum_{t=t_0}^{t_1-1}\frac{S_t}{2\gamma_t(1+\lambda_t)}\right].
\end{aligned}$$

Proof Letting $\lambda_t := \frac{\mu\eta^y}{2\left(v^y_{t+1}\right)^\beta}$, we have

$$\begin{aligned}
&\left\|y_{t+1} - y^*_{t+1}\right\|^2 \\
&\leq (1+\lambda_t)\|y_{t+1} - y^*_t\|^2 + S_t \\
&= (1+\lambda_t)\left\|\mathcal{P}_{\mathcal{Y}}\left(y_t + \gamma_t \nabla_y \widetilde{f}(x_t, y_t)\right) - y^*_t\right\|^2 + S_t \\
&\leq (1+\lambda_t)\left\|y_t + \gamma_t \nabla_y \widetilde{f}(x_t, y_t) - y^*_t\right\|^2 + S_t \\
&= (1+\lambda_t)\left(\|y_t - y^*_t\|^2 + \gamma_t^2\left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2 + 2\gamma_t\left\langle \nabla_y \widetilde{f}(x_t, y_t), y_t - y^*_t\right\rangle\right) + S_t \\
&= (1+\lambda_t)\Bigg(\|y_t - y^*_t\|^2 + \gamma_t^2\left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2 + 2\gamma_t\left\langle \nabla_y \widetilde{f}(x_t, y_t), y_t - y^*_t\right\rangle \\
&\qquad + \gamma_t\mu\|y_t - y^*_t\|^2 - \gamma_t\mu\|y_t - y^*_t\|^2\Bigg) + S_t
\end{aligned}$$

By multiplying $\frac{1}{\gamma_t(1+\lambda_t)}$ and rearranging the terms, we can get

$$\begin{aligned}
&2\left\langle \nabla_y \widetilde{f}(x_t, y_t), y^*_t - y_t\right\rangle - \mu\|y_t - y^*_t\|^2 \\
&\leq \frac{1-\gamma_t\mu}{\gamma_t}\|y_t - y^*_t\|^2 - \frac{1}{\gamma_t(1+\lambda_t)}\left\|y_{t+1} - y^*_{t+1}\right\|^2 + \gamma_t\left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2 + \frac{S_t}{\gamma_t(1+\lambda_t)}.
\end{aligned}$$

By telescoping from $t = t_0$ to $t_1 - 1$, we have

$$\begin{aligned}
&\sum_{t=t_0}^{t_1-1}\left(\left\langle \nabla_y \widetilde{f}(x_t, y_t), y^*_t - y_t\right\rangle - \frac{\mu}{2}\|y_t - y^*_t\|^2\right) \\
&\leq \sum_{t=t_0+1}^{t_1-1}\left(\frac{1-\gamma_t\mu}{2\gamma_t}\|y_t - y^*_t\|^2 - \frac{1}{2\gamma_t(1+\lambda_t)}\left\|y_{t+1} - y^*_{t+1}\right\|^2\right) + \sum_{t=t_0}^{t_1-1}\frac{\gamma_t}{2}\left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2 \\
&\quad + \sum_{t=t_0}^{t_1-1}\frac{S_t}{2\gamma_t(1+\lambda_t)}.
\end{aligned}$$

Now we take the expectation and get

$$\begin{aligned}
\mathbb{E}\left[\text{LHS}\right] &\geq \mathbb{E}\left[\sum_{t=t_0}^{t_1-1}\mathbb{E}_{\xi^y_t}\left[\left(\left\langle \nabla_y \widetilde{f}(x_t, y_t), y^*_t - y_t\right\rangle - \frac{\mu}{2}\|y_t - y^*_t\|^2\right)\right]\right] \\
&= \mathbb{E}\left[\sum_{t=t_0}^{t_1-1}\left(\langle \nabla_y f(x_t, y_t), y^*_t - y_t\rangle - \frac{\mu}{2}\|y_t - y^*_t\|^2\right)\right] \\
&\geq \mathbb{E}\left[\sum_{t=t_0}^{t_1-1}\left(f(x_t, y^*_t) - f(x_t, y_t)\right)\right],
\end{aligned}$$

where we used strong-concavity in the last inequality.

■

Lemma 11 *Under the same setting as Theorem 5, if $v_{t+1}^y \le C$ for $t = 0, ..., t_0 - 1$, then we have*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{t_0-1} \left(f(x_t, y_t^*) - f(x_t, y_t)\right)\right] \\
\le\ &\mathbb{E}\left[\sum_{t=0}^{t_0-1} \left(\frac{1-\gamma_t\mu}{2\gamma_t}\|y_t - y_t^*\|^2 - \frac{1}{\gamma_t(2+\mu\gamma_t)}\left\|y_{t+1} - y_{t+1}^*\right\|^2\right)\right] + \mathbb{E}\left[\sum_{t=0}^{t_0-1} \frac{\gamma_t}{2}\left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2\right] \\
&+ \frac{\kappa^2\left(\mu\eta^y C^\beta + 2C^{2\beta}\right)(\eta^x)^2}{2\mu\left(\eta^y\right)^2}\mathbb{E}\left[\frac{1 + \log v_{t_0}^x - \log v_0^x}{\left(v_0^x\right)^{2\alpha-1}} \cdot \mathbf{1}_{\alpha \ge 0.5} + \frac{\left(v_{t_0}^x\right)^{1-2\alpha}}{1-2\alpha} \cdot \mathbf{1}_{\alpha < 0.5}\right].
\end{aligned}$$

Proof By Young's inequality, we have

$$\left\|y_{t+1} - y_{t+1}^*\right\|^2 \le (1+\lambda_t)\|y_{t+1} - y_t^*\|^2 + \left(1 + \frac{1}{\lambda_t}\right)\left\|y_{t+1}^* - y_t^*\right\|^2.$$

Then letting $\lambda_t = \frac{\mu\gamma_t}{2}$ and by Theorem 10, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{t_0-1} \left(f(x_t, y_t^*) - f(x_t, y_t)\right)\right] \\
\le\ &\mathbb{E}\left[\sum_{t=0}^{t_0-1} \left(\frac{1-\gamma_t\mu}{2\gamma_t}\|y_t - y_t^*\|^2 - \frac{1}{\gamma_t(2+\mu\gamma_t)}\left\|y_{t+1} - y_{t+1}^*\right\|^2\right)\right] \\
&+ \mathbb{E}\left[\sum_{t=0}^{t_0-1} \frac{\gamma_t}{2}\left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2\right] + \mathbb{E}\left[\sum_{t=0}^{t_0-1} \frac{\left(1 + \frac{2}{\mu\gamma_t}\right)}{\gamma_t(2+\mu\gamma_t)}\left\|y_{t+1}^* - y_t^*\right\|^2\right].
\end{aligned}$$

We now remain to bound the last term:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{t_0-1} \frac{\left(1 + \frac{2}{\mu\gamma_t}\right)}{\gamma_t(2+\mu\gamma_t)}\left\|y_{t+1}^* - y_t^*\right\|^2\right] \\
\le\ &\mathbb{E}\left[\sum_{t=0}^{t_0-1} \frac{\left(1 + \frac{2}{\mu\gamma_t}\right)}{2\gamma_t}\left\|y_{t+1}^* - y_t^*\right\|^2\right] \\
=\ &\mathbb{E}\left[\sum_{t=0}^{t_0-1} \frac{\mu\eta^y\left(v_{t+1}^y\right)^\beta + 2\left(v_{t+1}^y\right)^{2\beta}}{2\mu\left(\eta^y\right)^2}\left\|y_{t+1}^* - y_t^*\right\|^2\right] \\
\le\ &\frac{\mu\eta^y C^\beta + 2C^{2\beta}}{2\mu\left(\eta^y\right)^2}\mathbb{E}\left[\sum_{t=0}^{t_0-1}\left\|y_{t+1}^* - y_t^*\right\|^2\right].
\end{aligned}$$

By Theorem 7 we have

$$\sum_{t=0}^{t_0-1}\left\|y_{t+1}^* - y_t^*\right\|^2 \le \kappa^2\sum_{t=0}^{t_0-1}\|x_{t+1} - x_t\|^2$$

$$
\begin{aligned}
&= \kappa^2 \sum_{t=0}^{t_0-1} \eta_t^2 \left\| \nabla_x \widetilde{f}(x_t, y_t) \right\|^2 \\
&= \kappa^2 \left(\eta^x\right)^2 \sum_{t=0}^{t_0-1} \frac{1}{\max\left\{v_{t+1}^x, v_{t+1}^y\right\}^{2\alpha}} \left\| \nabla_x \widetilde{f}(x_t, y_t) \right\|^2 \\
&\le \kappa^2 \left(\eta^x\right)^2 \sum_{t=0}^{t_0-1} \frac{1}{\left(v_{t+1}^x\right)^{2\alpha}} \left\| \nabla_x \widetilde{f}(x_t, y_t) \right\|^2 \\
&\le \kappa^2 \left(\eta^x\right)^2 \left(\frac{v_0^x}{\left(v_0^x\right)^{2\alpha}} + \sum_{t=0}^{t_0-1} \frac{1}{\left(v_{t+1}^x\right)^{2\alpha}} \left\| \nabla_x \widetilde{f}(x_t, y_t) \right\|^2 \right) \\
&\le \kappa^2 \left(\eta^x\right)^2 \left(\frac{1 + \log v_{t_0}^x - \log v_0^x}{\left(v_0^x\right)^{2\alpha-1}} \cdot \mathbf{1}_{\alpha \ge 0.5} + \frac{\left(v_{t_0}^x\right)^{1-2\alpha}}{1-2\alpha} \cdot \mathbf{1}_{\alpha < 0.5} \right)
\end{aligned}
$$

where we applied Theorem 6 in the last inequality. Bringing back this result, we finish the proof. ■

Lemma 12 *Under the same setting as Theorem 5, if t_0 is the first iteration such that $v_{t_0+1}^y > C$, then we have*

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_0}^{T-1} \left(f(x_t, y_t^*) - f(x_t, y_t)\right)\right] \\
&\le \mathbb{E}\left[\sum_{t=t_0}^{T-1} \left(\frac{1-\gamma_t \mu}{2\gamma_t} \|y_t - y_t^*\|^2 - \frac{1}{\gamma_t(2+\mu\gamma_t)} \left\|y_{t+1} - y_{t+1}^*\right\|^2 \right)\right] + \mathbb{E}\left[\sum_{t=t_0}^{T-1} \frac{\gamma_t}{2} \left\| \nabla_y \widetilde{f}(x_t, y_t) \right\|^2\right] \\
&\quad + \left(\kappa^2 + \frac{\widehat{L}^2 G^2 \left(\eta^x\right)^2}{\mu \eta^y \left(v_0^y\right)^{2\alpha-\beta}} \right) \frac{\left(\eta^x\right)^2}{2(1-\alpha)\eta^y \left(v_0^y\right)^{\alpha-\beta}} \mathbb{E}\left[\left(v_T^x\right)^{1-\alpha}\right] \\
&\quad + \frac{2\kappa^2 \left(\eta^x\right)^2}{\mu \left(\eta^y\right)^2 C^{2\alpha-2\beta}} \mathbb{E}\left[\sum_{t=t_0}^{T-1} \|\nabla_x f(x_t, y_t)\|^2\right] + \left(\frac{1}{\mu} + \frac{\eta^y}{\left(v_0^y\right)^\beta} \right) \frac{4\kappa \eta^x G^2}{\eta^y \left(v_0^y\right)^\alpha} \mathbb{E}\left[\left(v_T^y\right)^\beta\right].
\end{aligned}
$$

Proof By the Lipschitzness of $y^*(\cdot)$ as in Theorem 7, we have

$$
\begin{aligned}
\left\|y_{t+1} - y_{t+1}^*\right\|^2 &= \|y_{t+1} - y_t^*\|^2 + \left\|y_t^* - y_{t+1}^*\right\|^2 + 2\left\langle y_{t+1} - y_t^*, y_t^* - y_{t+1}^* \right\rangle \\
&\le \|y_{t+1} - y_t^*\|^2 + \kappa^2 \eta_t^2 \left\| \nabla_x \widetilde{f}(x_t, y_t) \right\|^2 + 2\left\langle y_{t+1} - y_t^*, y_t^* - y_{t+1}^* \right\rangle \\
&\le \|y_{t+1} - y_t^*\|^2 + \kappa^2 \eta_t^2 \left\| \nabla_x \widetilde{f}(x_t, y_t) \right\|^2 \underbrace{-2\left(y_{t+1} - y_t^*\right)^\intercal \nabla y^*(x_t) \left(x_{t+1} - x_t\right)}_{\text{(C)}} \\
&\quad + \underbrace{2\left(y_{t+1} - y_t^*\right)^\intercal \left(y_t^* - y_{t+1}^* + \nabla y^*(x_t)\left(x_{t+1} - x_t\right)\right)}_{\text{(D)}}.
\end{aligned}
$$

For Term (C), by the Cauchy-Schwarz and Lipschitzness of $y^*(\cdot)$,

$$
-2\left(y_{t+1} - y_t^*\right)^\intercal \nabla y^*(x_t) \left(x_{t+1} - x_t\right)
$$

$$
\begin{aligned}
&= 2\eta_t \left(y_{t+1} - y_t^*\right)^\intercal \nabla y^*(x_t) \nabla_x f(x_t, y_t) + 2\eta_t \left(y_{t+1} - y_t^*\right)^\intercal \nabla y^*(x_t) \left(\nabla_x \widetilde{f}(x_t, y_t) - \nabla_x f(x_t, y_t)\right) \\
&\le 2\eta_t \|y_{t+1} - y_t^*\| \|\nabla y^*(x_t)\| \|\nabla_x f(x_t, y_t)\| + 2\eta_t \left(y_{t+1} - y_t^*\right)^\intercal \nabla y^*(x_t) \left(\nabla_x \widetilde{f}(x_t, y_t) - \nabla_x f(x_t, y_t)\right) \\
&\le 2\|y_{t+1} - y_t^*\| \kappa \eta_t \|\nabla_x f(x_t, y_t)\| + 2\eta_t \left(y_{t+1} - y_t^*\right)^\intercal \nabla y^*(x_t) \left(\nabla_x \widetilde{f}(x_t, y_t) - \nabla_x f(x_t, y_t)\right) \\
&\le \lambda_t \|y_{t+1} - y_t^*\|^2 + \frac{\kappa^2 \eta_t^2}{\lambda_t} \|\nabla_x f(x_t, y_t)\|^2 + 2\eta_t \left(y_{t+1} - y_t^*\right)^\intercal \nabla y^*(x_t) \left(\nabla_x \widetilde{f}(x_t, y_t) - \nabla_x f(x_t, y_t)\right),
\end{aligned}
$$

where we used Young's inequality in the last step and $\lambda_t > 0$ will be determined later.

For Term (D), according to Cauchy-Schwarz and the smoothness of $y^*(\cdot)$ as shown in Theorem 8,

$$
\begin{aligned}
&2\left(y_{t+1} - y_t^*\right)^\intercal \left(y_t^* - y_{t+1}^* + \nabla y^*(x_t)(x_{t+1} - x_t)\right) \\
&\le 2\|y_{t+1} - y_t^*\| \left\|y_t^* - y_{t+1}^* + \nabla y^*(x_t)(x_{t+1} - x_t)\right\| \\
&\le 2\|y_{t+1} - y_t^*\| \cdot \frac{\widehat{L}}{2} \|x_{t+1} - x_t\|^2 \\
&= \widehat{L} \eta_t^2 \|y_{t+1} - y_t^*\| \left\|\nabla_x \widetilde{f}(x_t, y_t)\right\|^2 \\
&\le \widehat{L} \eta_t^2 \|y_{t+1} - y_t^*\| G \cdot \left\|\nabla_x \widetilde{f}(x_t, y_t)\right\| \\
&\le \frac{\tau \widehat{L} G^2 \eta_t^2}{2} \|y_{t+1} - y_t^*\|^2 + \frac{\widehat{L} \eta_t^2}{2\tau} \left\|\nabla_x \widetilde{f}(x_t, y_t)\right\|^2,
\end{aligned}
$$

where in the last step we used Young's inequality and $\tau > 0$.

Therefore, in total, we have

$$
\begin{aligned}
\left\|y_{t+1} - y_{t+1}^*\right\|^2 \le &\left(1 + \lambda_t + \frac{\tau \widehat{L} G^2 \eta_t^2}{2}\right) \|y_{t+1} - y_t^*\|^2 + \left(\kappa^2 + \frac{\widehat{L}}{2\tau}\right) \eta_t^2 \left\|\nabla_x \widetilde{f}(x_t, y_t)\right\|^2 \\
&+ \frac{\kappa^2 \eta_t^2}{\lambda_t} \|\nabla_x f(x_t, y_t)\|^2 + 2\eta_t \left(y_{t+1} - y_t^*\right)^\intercal \nabla y^*(x_t) \left(\nabla_x \widetilde{f}(x_t, y_t) - \nabla_x f(x_t, y_t)\right).
\end{aligned}
$$

Note that we can upper bound η_t by

$$
\eta_t = \frac{\eta^x}{\max\left\{v_{t+1}^x, v_{t+1}^y\right\}^\alpha} \le \frac{\eta^x}{\left(v_{t+1}^y\right)^\alpha} \le \frac{\eta^x}{\left(v_0^y\right)^\alpha},
$$

and

$$
\eta_t \le \frac{\eta^x}{\left(v_{t+1}^y\right)^\alpha} = \frac{\eta^x}{\left(v_{t+1}^y\right)^{\alpha-\beta} \left(v_{t+1}^y\right)^\beta} \le \frac{\eta^x}{\left(v_0^y\right)^{\alpha-\beta} \left(v_{t+1}^y\right)^\beta},
$$

which, plugged into the previous result, implies

$$
\begin{aligned}
\left\|y_{t+1} - y_{t+1}^*\right\|^2 \le &\left(1 + \lambda_t + \frac{\tau \widehat{L} G^2 \left(\eta^x\right)^2}{2\left(v_0^y\right)^{2\alpha-\beta} \left(v_{t+1}^y\right)^\beta}\right) \|y_{t+1} - y_t^*\|^2 + \left(\kappa^2 + \frac{\widehat{L}}{2\tau}\right) \eta_t^2 \left\|\nabla_x \widetilde{f}(x_t, y_t)\right\|^2 \\
&+ \frac{\kappa^2 \eta_t^2}{\lambda_t} \|\nabla_x f(x_t, y_t)\|^2 + 2\eta_t \left(y_{t+1} - y_t^*\right)^\intercal \nabla y^*(x_t) \left(\nabla_x \widetilde{f}(x_t, y_t) - \nabla_x f(x_t, y_t)\right).
\end{aligned}
$$

Now we choose $\lambda_t = \frac{\mu\eta^y}{4\left(v_{t+1}^y\right)^\beta}$ and $\tau = \frac{\mu\eta^y\left(v_0^y\right)^{2\alpha-\beta}}{2\widehat{L}G^2(\eta^x)^2}$, and get

$$\begin{aligned}
&\left\|y_{t+1}-y_{t+1}^*\right\|^2 \\
&\le \left(1+\frac{\mu\eta^y}{2\left(v_{t+1}^y\right)^\beta}\right)\|y_{t+1}-y_t^*\|^2 + \left(\kappa^2+\frac{\widehat{L}^2G^2\left(\eta^x\right)^2}{\mu\eta^y\left(v_0^y\right)^{2\alpha-\beta}}\right)\eta_t^2\left\|\nabla_x\widetilde{f}(x_t,y_t)\right\|^2 \\
&\quad + \frac{4\kappa^2\left(v_{t+1}^y\right)^\beta\eta_t^2}{\mu\eta^y}\|\nabla_x f(x_t,y_t)\|^2 + 2\eta_t\left(y_{t+1}-y_t^*\right)^\intercal\nabla y^*(x_t)\left(\nabla_x\widetilde{f}(x_t,y_t)-\nabla_x f(x_t,y_t)\right).
\end{aligned}$$

Then Theorem 10 gives us

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_0}^{T-1}\left(f(x_t,y_t^*)-f(x_t,y_t)\right)\right] \\
&\le \mathbb{E}\left[\sum_{t=t_0}^{T-1}\left(\frac{1-\gamma_t\mu}{2\gamma_t}\|y_t-y_t^*\|^2-\frac{1}{\gamma_t(2+\mu\gamma_t)}\left\|y_{t+1}-y_{t+1}^*\right\|^2\right)\right]+\mathbb{E}\left[\sum_{t=t_0}^{T-1}\frac{\gamma_t}{2}\left\|\nabla_y\widetilde{f}(x_t,y_t)\right\|^2\right] \\
&\quad+\underbrace{\mathbb{E}\left[\sum_{t=t_0}^{T-1}\frac{1}{\gamma_t(2+\mu\gamma_t)}\left(\kappa^2+\frac{\widehat{L}^2G^2\left(\eta^x\right)^2}{\mu\eta^y\left(v_0^y\right)^{2\alpha-\beta}}\right)\eta_t^2\left\|\nabla_x\widetilde{f}(x_t,y_t)\right\|^2\right]}_{\text{(E)}} \\
&\quad+\underbrace{\mathbb{E}\left[\sum_{t=t_0}^{T-1}\frac{4\kappa^2\left(v_{t+1}^y\right)^\beta\eta_t^2}{\gamma_t(2+\mu\gamma_t)\mu\eta^y}\|\nabla_x f(x_t,y_t)\|^2\right]}_{\text{(F)}} \\
&\quad+\underbrace{\mathbb{E}\left[\sum_{t=t_0}^{T-1}\frac{2\eta_t}{\gamma_t(2+\mu\gamma_t)}\left(y_{t+1}-y_t^*\right)^\intercal\nabla y^*(x_t)\left(\nabla_x\widetilde{f}(x_t,y_t)-\nabla_x f(x_t,y_t)\right)\right]}_{\text{(G)}}
\end{aligned}$$

Now we proceed to bound each term.

Term (E)

$$\begin{aligned}
\text{Term (E)} &\le \left(\kappa^2+\frac{\widehat{L}^2G^2\left(\eta^x\right)^2}{\mu\eta^y\left(v_0^y\right)^{2\alpha-\beta}}\right)\mathbb{E}\left[\sum_{t=t_0}^{T-1}\frac{\eta_t^2}{2\gamma_t}\left\|\nabla_x\widetilde{f}(x_t,y_t)\right\|^2\right] \\
&= \left(\kappa^2+\frac{\widehat{L}^2G^2\left(\eta^x\right)^2}{\mu\eta^y\left(v_0^y\right)^{2\alpha-\beta}}\right)\mathbb{E}\left[\sum_{t=t_0}^{T-1}\frac{\left(\eta^x\right)^2\left(v_{t+1}^y\right)^\beta}{2\eta^y\max\left\{v_{t+1}^x,v_{t+1}^y\right\}^{2\alpha}}\left\|\nabla_x\widetilde{f}(x_t,y_t)\right\|^2\right] \\
&\le \left(\kappa^2+\frac{\widehat{L}^2G^2\left(\eta^x\right)^2}{\mu\eta^y\left(v_0^y\right)^{2\alpha-\beta}}\right)\mathbb{E}\left[\sum_{t=t_0}^{T-1}\frac{\left(\eta^x\right)^2\left(v_{t+1}^y\right)^\beta}{2\eta^y\left(v_{t+1}^y\right)^\beta\left(v_{t+1}^y\right)^{\alpha-\beta}\left(v_{t+1}^x\right)^\alpha}\left\|\nabla_x\widetilde{f}(x_t,y_t)\right\|^2\right] \\
&\le \left(\kappa^2+\frac{\widehat{L}^2G^2\left(\eta^x\right)^2}{\mu\eta^y\left(v_0^y\right)^{2\alpha-\beta}}\right)\mathbb{E}\left[\sum_{t=t_0}^{T-1}\frac{\left(\eta^x\right)^2}{2\eta^y\left(v_0^y\right)^{\alpha-\beta}\left(v_{t+1}^x\right)^\alpha}\left\|\nabla_x\widetilde{f}(x_t,y_t)\right\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \left(\kappa^2 + \frac{\widehat{L}^2 G^2 \left(\eta^x\right)^2}{\mu \eta^y \left(v_0^y\right)^{2\alpha-\beta}}\right) \mathbb{E}\left[\frac{\left(\eta^x\right)^2}{2\eta^y \left(v_0^y\right)^{\alpha-\beta}} \left(\frac{v_0^x}{\left(v_0^x\right)^\alpha} + \sum_{t=0}^{T-1} \frac{1}{\left(v_{t+1}^x\right)^\alpha} \left\|\nabla_x \widetilde{f}(x_t, y_t)\right\|^2\right)\right] \\
&\leq \left(\kappa^2 + \frac{\widehat{L}^2 G^2 \left(\eta^x\right)^2}{\mu \eta^y \left(v_0^y\right)^{2\alpha-\beta}}\right) \frac{\left(\eta^x\right)^2}{2(1-\alpha)\eta^y \left(v_0^y\right)^{\alpha-\beta}} \mathbb{E}\left[\left(v_T^x\right)^{1-\alpha}\right],
\end{aligned}
$$

where we used Theorem 6 in the last step.

Term (F)

$$
\begin{aligned}
\text{Term (F)} &\leq \mathbb{E}\left[\sum_{t=t_0}^{T-1} \frac{2\kappa^2 \left(v_{t+1}^y\right)^\beta \eta_t^2}{\gamma_t \mu \eta^y} \left\|\nabla_x f(x_t, y_t)\right\|^2\right] \\
&= \frac{2\kappa^2 \left(\eta^x\right)^2}{\mu \left(\eta^y\right)^2} \mathbb{E}\left[\sum_{t=t_0}^{T-1} \frac{\left(v_{t+1}^y\right)^{2\beta}}{\max\left\{v_{t+1}^x, v_{t+1}^y\right\}^{2\alpha}} \left\|\nabla_x f(x_t, y_t)\right\|^2\right] \\
&\leq \frac{2\kappa^2 \left(\eta^x\right)^2}{\mu \left(\eta^y\right)^2} \mathbb{E}\left[\sum_{t=t_0}^{T-1} \frac{\left(v_{t+1}^y\right)^{2\beta}}{\left(v_{t+1}^y\right)^{2\alpha}} \left\|\nabla_x f(x_t, y_t)\right\|^2\right] \\
&\leq \frac{2\kappa^2 \left(\eta^x\right)^2}{\mu \left(\eta^y\right)^2} \mathbb{E}\left[\frac{1}{\left(v_{t_0+1}^y\right)^{2\alpha-2\beta}} \sum_{t=t_0}^{T-1} \left\|\nabla_x f(x_t, y_t)\right\|^2\right] \\
&\leq \frac{2\kappa^2 \left(\eta^x\right)^2}{\mu \left(\eta^y\right)^2 C^{2\alpha-2\beta}} \mathbb{E}\left[\sum_{t=t_0}^{T-1} \left\|\nabla_x f(x_t, y_t)\right\|^2\right]
\end{aligned}
$$

Term (G) For simplicity, denote $m_t := \frac{2}{\gamma_t(2+\mu\gamma_t)} \left(y_{t+1} - y_t^*\right)^\intercal \nabla y^*(x_t) \left(\nabla_x \widetilde{f}(x_t, y_t) - \nabla_x f(x_t, y_t)\right)$ Since $y^*(\cdot)$ is κ-Lipschitz as in Theorem 7, $|m_t|$ can be upper bounded as

$$
\begin{aligned}
|m_t| &\leq \frac{1}{\gamma_t} \left\|y_{t+1} - y_t^*\right\| \left\|\nabla y^*(x_t)\right\| \left(\left\|\nabla_x \widetilde{f}(x_t, y_t)\right\| + \left\|\nabla_x f(x_t, y_t)\right\|\right) \\
&\leq \frac{\kappa}{\gamma_t} \left\|y_{t+1} - y_t^*\right\| \left(\left\|\nabla_x \widetilde{f}(x_t, y_t)\right\| + \left\|\nabla_x f(x_t, y_t)\right\|\right) \\
&\leq \frac{\kappa}{\gamma_t} \left\|\mathcal{P}_\mathcal{Y}\left(y_t + \gamma_t \nabla_y \widetilde{f}(x_t, y_t)\right) - y_t^*\right\| \left(\left\|\nabla_x \widetilde{f}(x_t, y_t)\right\| + \left\|\nabla_x f(x_t, y_t)\right\|\right) \\
&\leq \frac{\kappa}{\gamma_t} \left\|y_t + \gamma_t \nabla_y \widetilde{f}(x_t, y_t) - y_t^*\right\| \left(\left\|\nabla_x \widetilde{f}(x_t, y_t)\right\| + \left\|\nabla_x f(x_t, y_t)\right\|\right) \\
&\leq \frac{\kappa}{\gamma_t} \left(\left\|y_t - y_t^*\right\| + \left\|\gamma_t \nabla_y \widetilde{f}(x_t, y_t)\right\|\right) \left(\left\|\nabla_x \widetilde{f}(x_t, y_t)\right\| + \left\|\nabla_x f(x_t, y_t)\right\|\right) \\
&\leq \frac{\kappa}{\gamma_t} \left(\frac{1}{\mu} \left\|\nabla_y f(x_t, y_t)\right\| + \left\|\gamma_t \nabla_y \widetilde{f}(x_t, y_t)\right\|\right) \left(\left\|\nabla_x \widetilde{f}(x_t, y_t)\right\| + \left\|\nabla_x f(x_t, y_t)\right\|\right) \\
&\leq \underbrace{\frac{2G\kappa}{\gamma_{T-1}} \left(\frac{G}{\mu} + \frac{\eta^y G}{\left(v_0^y\right)^\beta}\right)}_{M}.
\end{aligned}
$$

Also note that γ_t and y_{t+1} does not depend on ξ_t^x, so $\mathbb{E}_{\xi_t^x}\left[m_t\right] = 0$. Next, we look at Term (G).

$$
\begin{aligned}
\text{Term (G)} &= \mathbb{E}\left[\sum_{t=t_0}^{T-1} \eta_t m_t\right] \\
&= \mathbb{E}\left[\eta_{t_0} m_{t_0} + \sum_{t=t_0+1}^{T-1} \eta_{t-1} m_t + \sum_{t=t_0+1}^{T-1} \left(\eta_t - \eta_{t-1}\right) m_t\right] \\
&\le \mathbb{E}\left[\frac{\eta^x}{\left(v_0^y\right)^\alpha} M + \sum_{t=t_0+1}^{T-1} \eta_{t-1} \mathbb{E}_{\xi_t^x}\left[m_t\right] + \sum_{t=t_0+1}^{T-1} \left(\eta_{t-1} - \eta_t\right)\left(-m_t\right)\right] \\
&\le \mathbb{E}\left[\frac{\eta^x}{\left(v_0^y\right)^\alpha} M + \sum_{t=t_0+1}^{T-1} \left(\eta_{t-1} - \eta_t\right) M\right] \\
&\le \mathbb{E}\left[\frac{2\eta^x}{\left(v_0^y\right)^\alpha} M\right] \\
&= \left(\frac{1}{\mu} + \frac{\eta^y}{\left(v_0^y\right)^\beta}\right) \frac{4\kappa \eta^x G^2}{\eta^y \left(v_0^y\right)^\alpha} \mathbb{E}\left[\left(v_T^y\right)^\beta\right].
\end{aligned}
$$

Summarizing all the results, we finish the proof. ■

Lemma 13 *Under the same setting as Theorem 5, we have*

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T-1}\left(\frac{1-\gamma_t\mu}{2\gamma_t}\|y_t - y_t^*\|^2 - \frac{1}{\gamma_t(2+\mu\gamma_t)}\left\|y_{t+1} - y_{t+1}^*\right\|^2\right)\right] \\
\le\ & \frac{\left(v_0^y\right)^\beta G^2}{2\mu^2\eta^y} + \frac{(2\beta G)^{\frac{1}{1-\beta}+2} G^2}{4\mu^{\frac{1}{1-\beta}+3}\left(\eta^y\right)^{\frac{1}{1-\beta}+2}\left(v_0^y\right)^{2-2\beta}}.
\end{aligned}
$$

Proof

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T-1}\left(\frac{1-\gamma_t\mu}{2\gamma_t}\|y_t - y_t^*\|^2 - \frac{1}{\gamma_t(2+\mu\gamma_t)}\left\|y_{t+1} - y_{t+1}^*\right\|^2\right)\right] \\
\le & \left(\frac{\left(v_0^y\right)^\beta}{2\eta^y} - \frac{\mu}{2}\right)\|y_0 - y_0^*\|^2 + \frac{1}{2\eta^y}\sum_{t=1}^{T-1}\left(\left(v_{t+1}^y\right)^\beta - \frac{\mu\eta^y}{2} - \left(v_t^y\right)^\beta - \frac{\mu^2\left(\eta^y\right)^2}{4\left(v_t^y\right)^\beta + 2\mu\eta^y}\right)\|y_t - y_t^*\|^2 \\
\le & \frac{\left(v_0^y\right)^\beta G^2}{2\mu^2\eta^y} + \frac{1}{2\eta^y}\underbrace{\sum_{t=1}^{T-1}\left(\left(v_{t+1}^y\right)^\beta - \frac{\mu\eta^y}{2} - \left(v_t^y\right)^\beta\right)\|y_t - y_t^*\|^2}_{\text{(H)}}.
\end{aligned}
$$

For Term (H),we will bound it using the same strategy as in [50]. The general idea is to show that $\left(v_{t+1}^y\right)^\beta - \frac{\mu\eta^y}{2} - \left(v_t^y\right)^\beta$ is positive for only a constant number of times. If the term is positive at iteration t, then we have

$$
0 < \left(v_{t+1}^y\right)^\beta - \left(v_t^y\right)^\beta - \frac{\mu\eta^y}{2}
$$

$$
\begin{aligned}
&= \left(v_t^y + \left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2\right)^\beta - (v_t^y)^\beta - \frac{\mu\eta^y}{2} \\
&= (v_t^y)^\beta \left(1 + \frac{\left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2}{v_t^y}\right)^\beta - (v_t^y)^\beta - \frac{\mu\eta^y}{2} \\
&\leq (v_t^y)^\beta \left(1 + \frac{\beta\left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2}{v_t^y}\right) - (v_t^y)^\beta - \frac{\mu\eta^y}{2} \\
&= \frac{\beta\left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2}{(v_t^y)^{1-\beta}} - \frac{\mu\eta^y}{2}, \qquad (10)
\end{aligned}
$$

where in the last inequality we used Bernoulli's inequality. By rearranging the terms, we have the two following conditions

$$
\begin{cases}
\left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2 > \frac{\mu\eta^y}{2\beta} (v_t^y)^{1-\beta} \geq \frac{\mu\eta^y}{2\beta} (v_0^y)^{1-\beta} \\
(v_t^y)^{1-\beta} < \frac{2\beta}{\mu\eta^y} \left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2 \leq \frac{2\beta G}{\mu\eta^y},
\end{cases}
$$

This indicates that at each time the term is positive, the gradient norm must be large enough and the accumulated gradient norm, i.e., v_{t+1}^y, must be small enough. Therefore, we can have at most

$$
\frac{\left(\frac{2\beta G}{\mu\eta^y}\right)^{\frac{1}{1-\beta}}}{\frac{\mu\eta^y}{2\beta} (v_0^y)^{1-\beta}}
$$

constant number of iterations when the term is positive. When the term is positive, it is also upper bounded by using the result from Equation (10):

$$
\begin{aligned}
\left((v_{t+1}^y)^\beta - \frac{\mu\eta^y}{2} - (v_t^y)^\beta\right) \|y_t - y_t^*\|^2 &\leq \frac{\beta\left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2}{(v_t^y)^{1-\beta}} \|y_t - y_t^*\|^2 \\
&\leq \frac{\beta G^2}{(v_0^y)^{1-\beta}} \|y_t - y_t^*\|^2 \\
&\leq \frac{\beta G^2}{\mu^2 (v_0^y)^{1-\beta}} \|\nabla_y f(x_t, y_t)\|^2 \\
&\leq \frac{\beta G^4}{\mu^2 (v_0^y)^{1-\beta}}
\end{aligned}
$$

which is a constant. In total, Term (H) is bounded by

$$
\frac{(2\beta G)^{\frac{1}{1-\beta}+2} G^2}{2\mu^{\frac{1}{1-\beta}+3} (\eta^y)^{\frac{1}{1-\beta}+1} (v_0^y)^{2-2\beta}}.
$$

Bringing it back, we get the desired result. ■

Lemma 14 *Under the same setting as Theorem 5, for any constant C, we have*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T-1}\left(f(x_t,y_t^*)-f(x_t,y_t)\right)\right]\\
\leq\ &\frac{2\kappa^2\left(\eta^x\right)^2}{\mu\left(\eta^y\right)^2C^{2\alpha-2\beta}}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_x f(x_t,y_t)\|^2\right]+\frac{\eta^y}{2(1-\beta)}\mathbb{E}\left[\left(v_T^y\right)^{1-\beta}\right]\\
&+\left(\frac{1}{\mu}+\frac{\eta^y}{\left(v_0^y\right)^\beta}\right)\frac{4\kappa\eta^xG^2}{\eta^y\left(v_0^y\right)^\alpha}\mathbb{E}\left[\left(v_T^y\right)^\beta\right]\\
&+\frac{\kappa^2\left(\mu\eta^yC^\beta+2C^{2\beta}\right)\left(\eta^x\right)^2}{2\mu\left(\eta^y\right)^2}\mathbb{E}\left[\frac{1+\log v_T^x-\log v_0^x}{\left(v_0^x\right)^{2\alpha-1}}\cdot\mathbf{1}_{\alpha\geq0.5}+\frac{\left(v_T^x\right)^{1-2\alpha}}{1-2\alpha}\cdot\mathbf{1}_{\alpha<0.5}\right]\\
&+\left(\kappa^2+\frac{\widehat{L}^2G^2\left(\eta^x\right)^2}{\mu\eta^y\left(v_0^y\right)^{2\alpha-\beta}}\right)\frac{\left(\eta^x\right)^2}{2(1-\alpha)\eta^y\left(v_0^y\right)^{\alpha-\beta}}\mathbb{E}\left[\left(v_T^x\right)^{1-\alpha}\right]\\
&+\frac{\left(v_0^y\right)^\beta G^2}{2\mu^2\eta^y}+\frac{\left(2\beta G\right)^{\frac{1}{1-\beta}+2}G^2}{4\mu^{\frac{1}{1-\beta}+3}\left(\eta^y\right)^{\frac{1}{1-\beta}+2}\left(v_0^y\right)^{2-2\beta}}.
\end{aligned}$$

Proof By Theorem 11 and Theorem 12, we have for any constant C,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T-1}\left(f(x_t,y_t^*)-f(x_t,y_t)\right)\right]\\
\leq\ &\mathbb{E}\left[\sum_{t=0}^{T-1}\left(\frac{1-\gamma_t\mu}{2\gamma_t}\|y_t-y_t^*\|^2-\frac{1}{\gamma_t(2+\mu\gamma_t)}\left\|y_{t+1}-y_{t+1}^*\right\|^2\right)\right]\\
&+\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\gamma_t}{2}\left\|\nabla_y\widetilde{f}(x_t,y_t)\right\|^2\right]+\frac{2\kappa^2\left(\eta^x\right)^2}{\mu\left(\eta^y\right)^2C^{2\alpha-2\beta}}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_x f(x_t,y_t)\|^2\right]\\
&+\frac{\kappa^2\left(\mu\eta^yC^\beta+2C^{2\beta}\right)\left(\eta^x\right)^2}{2\mu\left(\eta^y\right)^2}\mathbb{E}\left[\frac{1+\log v_T^x-\log v_0^x}{\left(v_0^x\right)^{2\alpha-1}}\cdot\mathbf{1}_{\alpha\geq0.5}+\frac{\left(v_T^x\right)^{1-2\alpha}}{1-2\alpha}\cdot\mathbf{1}_{\alpha<0.5}\right]\\
&+\left(\kappa^2+\frac{\widehat{L}^2G^2\left(\eta^x\right)^2}{\mu\eta^y\left(v_0^y\right)^{2\alpha-\beta}}\right)\frac{\left(\eta^x\right)^2}{2(1-\alpha)\eta^y\left(v_0^y\right)^{\alpha-\beta}}\mathbb{E}\left[\left(v_T^x\right)^{1-\alpha}\right]\\
&+\left(\frac{1}{\mu}+\frac{\eta^y}{\left(v_0^y\right)^\beta}\right)\frac{4\kappa\eta^xG^2}{\eta^y\left(v_0^y\right)^\alpha}\mathbb{E}\left[\left(v_T^y\right)^\beta\right].
\end{aligned}$$

The first term can be bounded by Theorem 13. For the second term, we have

$$\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\gamma_t}{2}\left\|\nabla_y\widetilde{f}(x_t,y_t)\right\|^2\right]=\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\eta^y}{2\left(v_{t+1}^y\right)^\beta}\left\|\nabla_y\widetilde{f}(x_t,y_t)\right\|^2\right]$$

$$\begin{aligned}
&\leq \frac{\eta^y}{2}\mathbb{E}\left[\frac{v_0^y}{\left(v_0^y\right)^\beta} + \sum_{t=0}^{T-1}\frac{1}{\left(v_{t+1}^y\right)^\beta}\left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2\right] \\
&\leq \frac{\eta^y}{2(1-\beta)}\mathbb{E}\left[\left(v_T^y\right)^{1-\beta}\right],
\end{aligned}$$

where the last inequality follows from Theorem 6. Then the proof is completed. ■

C.3. Proof of Theorem 5

We present a formal version of Theorem 5.

Theorem 15 (stochastic setting) *Under Assumptions 3.1 to 3.6, Algorithm 1 with stochastic gradient oracles satisfies that for any $0 < \beta < \alpha < 1$, after T iterations,*

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_x f(x_t, y_t)\|^2\right] \\
&\leq \frac{4\Delta\Phi G^{2\alpha}}{\eta^x T^{1-\alpha}} + \left(\frac{4l\kappa\eta^x}{1-\alpha} + \left(\kappa^2 + \frac{\widehat{L}^2 G^2 \left(\eta^x\right)^2}{\mu\eta^y \left(v_0^y\right)^{2\alpha-\beta}}\right)\frac{2l\kappa\left(\eta^x\right)^2}{(1-\alpha)\eta^y\left(v_0^y\right)^{\alpha-\beta}}\right)\frac{G^{2(1-\alpha)}}{T^\alpha} \\
&\quad + \frac{2l\kappa\eta^y G^{2(1-\beta)}}{(1-\beta)T^\beta} + \left(\frac{1}{\mu} + \frac{\eta^y}{\left(v_0^y\right)^\beta}\right)\frac{16 l\kappa^2\eta^x G^{2(1+\beta)}}{\eta^y\left(v_0^y\right)^\alpha T^{1-\beta}} \\
&\quad + \frac{2\kappa^4\left(\mu\eta^y C^\beta + 2C^{2\beta}\right)\left(\eta^x\right)^2}{\left(\eta^y\right)^2}\left(\frac{1 + \log(G^2 T) - \log v_0^x}{\left(v_0^x\right)^{2\alpha-1}T}\cdot\mathbf{1}_{\alpha\geq 0.5} + \frac{G^{2(1-2\alpha)}}{(1-2\alpha)T^{2\alpha}}\cdot\mathbf{1}_{\alpha<0.5}\right) \\
&\quad + \frac{2\kappa^2\left(v_0^y\right)^\beta G^2}{\mu\eta^y T} + \frac{l\kappa\left(2\beta G\right)^{\frac{1}{1-\beta}+2}G^2}{\mu^{\frac{1}{1-\beta}+3}\left(\eta^y\right)^{\frac{1}{1-\beta}+2}\left(v_0^y\right)^{2-2\beta}T},
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_y f(x_t, y_t)\|^2\right] \\
&\leq \frac{4\kappa^3\left(\eta^x\right)^2}{\left(\eta^y\right)^2 C^{2\alpha-2\beta}}\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_x f(x_t, y_t)\|^2\right] + \frac{l\eta^y G^{2-2\beta}}{(1-\beta)T^\beta} + \left(\frac{1}{\mu} + \frac{\eta^y}{\left(v_0^y\right)^\beta}\right)\frac{8l\kappa\eta^x G^{2+2\beta}}{\eta^y\left(v_0^y\right)^\alpha T^{1-\beta}} \\
&\quad + \frac{\kappa^3\left(\mu\eta^y C^\beta + 2C^{2\beta}\right)\left(\eta^x\right)^2}{\left(\eta^y\right)^2}\left(\frac{1 + \log TG^2 - \log v_0^x}{\left(v_0^x\right)^{2\alpha-1}T}\cdot\mathbf{1}_{\alpha\geq 0.5} + \frac{G^{2-4\alpha}}{(1-2\alpha)T^{2\alpha}}\cdot\mathbf{1}_{\alpha<0.5}\right) \\
&\quad + \left(\kappa^2 + \frac{\widehat{L}^2 G^2\left(\eta^x\right)^2}{\mu\eta^y\left(v_0^y\right)^{2\alpha-\beta}}\right)\frac{l\left(\eta^x\right)^2 G^{2-2\alpha}}{(1-\alpha)\eta^y\left(v_0^y\right)^{\alpha-\beta}T^\alpha} + \frac{\kappa\left(v_0^y\right)^\beta G^2}{\mu\eta^y T} + \frac{2l\left(2\beta G\right)^{\frac{1}{1-\beta}+2}G^2}{4\mu^{\frac{1}{1-\beta}+3}\left(\eta^y\right)^{\frac{1}{1-\beta}+2}\left(v_0^y\right)^{2-2\beta}T}.
\end{aligned}$$

Proof By smoothness of the primal function, we have

$$\Phi(x_{t+1}) - \Phi(x_t) \leq -\eta_t\left\langle\nabla\Phi(x_t), \nabla_x\widetilde{f}(x_t, y_t)\right\rangle + l\kappa\eta_t^2\left\|\nabla_x\widetilde{f}(x_t, y_t)\right\|^2.$$

By multiplying $\frac{1}{\eta_t}$ on both sides and taking the expectation w.r.t. the noise of current iteration, we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\Phi(x_{t+1}) - \Phi(x_t)}{\eta_t}\right] \\
&\leq -\langle \nabla \Phi(x_t), \nabla_x f(x_t, y_t)\rangle + l\kappa \mathbb{E}\left[\eta_t \left\| \nabla_x \widetilde{f}(x_t, y_t)\right\|^2\right] \\
&= -\left\| \nabla_x f(x_t, y_t)\right\|^2 + \langle \nabla_x f(x_t, y_t) - \nabla \Phi(x_t), \nabla_x f(x_t, y_t)\rangle + l\kappa \mathbb{E}\left[\eta_t \left\| \nabla_x \widetilde{f}(x_t, y_t)\right\|^2\right] \\
&\leq -\left\| \nabla_x f(x_t, y_t)\right\|^2 + \frac{1}{2}\left\| \nabla_x f(x_t, y_t) - \nabla \Phi(x_t)\right\|^2 + \frac{1}{2}\left\| \nabla_x f(x_t, y_t)\right\|^2 + l\kappa \mathbb{E}\left[\eta_t \left\| \nabla_x \widetilde{f}(x_t, y_t)\right\|^2\right] \\
&= -\frac{1}{2}\left\| \nabla_x f(x_t, y_t)\right\|^2 + \frac{1}{2}\left\| \nabla_x f(x_t, y_t) - \nabla \Phi(x_t)\right\|^2 + l\kappa \mathbb{E}\left[\eta_t \left\| \nabla_x \widetilde{f}(x_t, y_t)\right\|^2\right]
\end{aligned}
$$

Summing over $t = 0$ to $T - 1$, rearranging and taking total expectation, we get

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T-1} \left\| \nabla_x f(x_t, y_t)\right\|^2\right] \\
&\leq \underbrace{2\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\Phi(x_t) - \Phi(x_{t+1})}{\eta_t}\right]}_{\text{(I)}} + \underbrace{2l\kappa \mathbb{E}\left[\sum_{t=0}^{T-1} \eta_t \left\| \nabla_x \widetilde{f}(x_t, y_t)\right\|^2\right]}_{\text{(J)}} + \underbrace{\mathbb{E}\left[\sum_{t=0}^{T-1} \left\| \nabla_x f(x_t, y_t) - \nabla \Phi(x_t)\right\|^2\right]}_{\text{(K)}}.
\end{aligned}
\tag{11}
$$

Term (I)

$$
\begin{aligned}
2\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\Phi(x_t) - \Phi(x_{t+1})}{\eta_t}\right] &\leq 2\mathbb{E}\left[\frac{\Phi(x_0)}{\eta_0} - \frac{\Phi(x_T)}{\eta_{T-1}} + \sum_{t=1}^{T-1} \Phi(x_t)\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\right] \\
&\leq 2\mathbb{E}\left[\frac{\Phi_{\max}}{\eta_0} - \frac{\Phi^*}{\eta_{T-1}} + \sum_{t=1}^{T-1} \Phi_{\max}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\right] \\
&= 2\mathbb{E}\left[\frac{\Delta\Phi}{\eta_{T-1}}\right] = 2\mathbb{E}\left[\frac{\Delta\Phi}{\eta^x} \max\left\{v_T^x, v_T^y\right\}^\alpha\right].
\end{aligned}
$$

Term (J)

$$
\begin{aligned}
2l\kappa \sum_{t=0}^{T-1} \mathbb{E}\left[\eta_t \left\| \nabla_x \widetilde{f}(x_t, y_t)\right\|^2\right] &= 2l\kappa \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\eta^x}{\max\left\{v_{t+1}^x, v_{t+1}^y\right\}^\alpha} \left\| \nabla_x \widetilde{f}(x_t, y_t)\right\|^2\right] \\
&\leq 2l\kappa\eta^x \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{1}{\left(v_{t+1}^x\right)^\alpha} \left\| \nabla_x \widetilde{f}(x_t, y_t)\right\|^2\right] \\
&\leq 2l\kappa\eta^x \mathbb{E}\left[\left(\frac{v_0^x}{\left(v_0^x\right)^\alpha} + \sum_{t=0}^{T-1} \frac{1}{\left(v_{t+1}^x\right)^\alpha} \left\| \nabla_x \widetilde{f}(x_t, y_t)\right\|^2\right)\right] \\
&\leq \frac{2l\kappa\eta^x}{1-\alpha} \mathbb{E}\left[\left(v_T^x\right)^{1-\alpha}\right].
\end{aligned}
$$

Term (K) According to the smoothness of $f(x_t, \cdot)$, we have

$$\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_x f(x_t, y_t) - \nabla\Phi(x_t)\|^2\right] \le l^2\mathbb{E}\left[\sum_{t=0}^{T-1}\|y_t - y_t^*\|^2\right] \le 2l\kappa\mathbb{E}\left[\sum_{t=0}^{T-1}\left(f(x_t, y_t^*) - f(x_t, y_t)\right)\right],$$

where the last inequality follows the strong-concavity of y. Now we let

$$C = \left(\frac{8l\kappa^3\left(\eta^x\right)^2}{\mu\left(\eta^y\right)^2}\right)^{\frac{1}{2\alpha-2\beta}},$$

and apply Theorem 14, in total, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_x f(x_t, y_t)\|^2\right] \\
&\le \frac{1}{2}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_x f(x_t, y_t)\|^2\right] + 2\mathbb{E}\left[\frac{\Delta\Phi}{\eta^x}\max\left\{v_T^x, v_T^y\right\}^\alpha\right] + \frac{2l\kappa\eta^x}{1-\alpha}\mathbb{E}\left[\left(v_T^x\right)^{1-\alpha}\right] \\
&\quad + \frac{l\kappa\eta^y}{1-\beta}\mathbb{E}\left[\left(v_T^y\right)^{1-\beta}\right] + \left(\frac{1}{\mu} + \frac{\eta^y}{\left(v_0^y\right)^\beta}\right)\frac{8l\kappa^2\eta^x G^2}{\eta^y\left(v_0^y\right)^\alpha}\mathbb{E}\left[\left(v_T^y\right)^\beta\right] \\
&\quad + \frac{\kappa^4\left(\mu\eta^y C^\beta + 2C^{2\beta}\right)\left(\eta^x\right)^2}{\left(\eta^y\right)^2}\mathbb{E}\left[\frac{1 + \log v_T^x - \log v_0^x}{\left(v_0^x\right)^{2\alpha-1}}\cdot\mathbf{1}_{\alpha\ge 0.5} + \frac{\left(v_T^x\right)^{1-2\alpha}}{1-2\alpha}\cdot\mathbf{1}_{\alpha<0.5}\right] \\
&\quad + \left(\kappa^2 + \frac{\widehat{L}^2 G^2\left(\eta^x\right)^2}{\mu\eta^y\left(v_0^y\right)^{2\alpha-\beta}}\right)\frac{l\kappa\left(\eta^x\right)^2}{(1-\alpha)\eta^y\left(v_0^y\right)^{\alpha-\beta}}\mathbb{E}\left[\left(v_T^x\right)^{1-\alpha}\right] + \frac{\kappa^2\left(v_0^y\right)^\beta G^2}{\mu\eta^y} \\
&\quad + \frac{l\kappa\left(2\beta G\right)^{\frac{1}{1-\beta}+2}G^2}{2\mu^{\frac{1}{1-\beta}+3}\left(\eta^y\right)^{\frac{1}{1-\beta}+2}\left(v_0^y\right)^{2-2\beta}}.
\end{aligned}$$

It remains to bound $\left(v_T^z\right)^m$ for $z \in \{x, y\}$ and $m \ge 0$:

$$\left(v_T^z\right)^m \le \left(TG^2\right)^m.$$

Bringing it back, we conclude our proof.

■

C.4. TiAda without Accessing Opponent's Gradients

The effective stepsize of x requires the knowledge of gradients of y, i.e., v_{t+1}^y. At the end of Section 3, we discussed the situation when such information is not available. Now we formally introduce the algorithm and present the convergence result.

Theorem 16 (stochastic) *Under Assumptions 3.1, 3.2, 3.4 and 3.5, Algorithm 2 with stochastic gradient oracles satisfies that for any* $0 < \beta < \alpha < 1$, *after* T *iterations,*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_x f(x_t, y_t)\|^2\right]$$

Algorithm 2 TiAda without MAX

```
**Input:** $(x_0, y_0)$, $v_0^x > 0$, $v_0^y > 0$, $\eta^x > 0$, $\eta^y > 0$, $\alpha > 0$, $\beta > 0$ and $\alpha > \beta$.
**for** $t = 0, 1, 2, \ldots$ **do**
    sample i.i.d. $\xi_t^x$ and $\xi_t^y$, and let $g_t^x = \nabla_x F(x_t, y_t; \xi_t^x)$ and $g_t^y = \nabla_y F(x_t, y_t; \xi_t^y)$
    $v_{t+1}^x = v_t^x + \|g_t^x\|^2$ and $v_{t+1}^y = v_t^y + \|g_t^y\|^2$
    $x_{t+1} = x_t - \frac{\eta^x}{\left(v_{t+1}^x\right)^\alpha} g_t^x$ and $y_{t+1} = \mathcal{P}_{\mathcal{Y}}\left(y_t + \frac{\eta^y}{\left(v_{t+1}^y\right)^\beta} g_t^y\right)$
**end for**
```

$$
\begin{aligned}
&\leq \frac{2\Delta\Phi G^{2\alpha}}{\eta^x T^{1-\alpha}} + \frac{2l\kappa\eta^x G^{2-2\alpha}}{(1-\alpha)T^\alpha} + \left(\frac{\left(v_0^y\right)^\beta G^2}{2\mu^2\eta^y} + \frac{(2\beta G)^{\frac{1}{1-\beta}+2} G^2}{4\mu^{\frac{1}{1-\beta}+3} (\eta^y)^{\frac{1}{1-\beta}+2} \left(v_0^y\right)^{2-2\beta}}\right)\frac{1}{T} + \frac{\eta^y G^{2-2\beta}}{2(1-\beta)T^\beta} \\
&\quad + \left(\frac{(\eta^x)^2 \kappa^2}{2\left(v_0^y\right)^\beta \eta^y} + \frac{(\eta^x)^2 \kappa^2}{\mu(\eta^y)^2}\right)\left(\frac{\left(1 + \log G^2 T - \log v_0^x\right) G^{4\beta}}{\left(v_0^x\right)^{2\alpha-1} T^{1-2\beta}} \cdot \mathbf{1}_{\alpha \geq 0.5} + \frac{G^{2-4\alpha+4\beta}}{(1-2\alpha)T^{2\alpha-2\beta}} \cdot \mathbf{1}_{\alpha < 0.5}\right),
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla_y f(x_t, y_t)\|^2\right] \\
&\leq \left(\frac{\kappa\left(v_0^y\right)^\beta G^2}{\mu\eta^y} + \frac{2l\,(2\beta G)^{\frac{1}{1-\beta}+2} G^2}{4\mu^{\frac{1}{1-\beta}+3} (\eta^y)^{\frac{1}{1-\beta}+2} \left(v_0^y\right)^{2-2\beta}}\right)\frac{1}{T} + \frac{l\eta^y G^{2-2\beta}}{(1-\beta)T^\beta} \\
&\quad + \left(\frac{l\,(\eta^x)^2 \kappa^2}{\left(v_0^y\right)^\beta \eta^y} + \frac{2\,(\eta^x)^2 \kappa^3}{(\eta^y)^2}\right)\left(\frac{\left(1 + \log G^2 T - \log v_0^x\right) G^{4\beta}}{\left(v_0^x\right)^{2\alpha-1} T^{1-2\beta}} \cdot \mathbf{1}_{\alpha \geq 0.5} + \frac{G^{2-4\alpha+4\beta}}{(1-2\alpha)T^{2\alpha-2\beta}} \cdot \mathbf{1}_{\alpha < 0.5}\right).
\end{aligned}
$$

Remark 17 *The best rate achievable is $\widetilde{\mathcal{O}}\left(\epsilon^{-6}\right)$ by choosing $\alpha = 1/2$ and $\beta = 1/3$.*

Proof Theorems 10 and 13 can be directly used here because they do not have or expand the effective stepsize of x, i.e., η_t. This is also the case for the beginning part of Appendix C.3, the proof of Theorem 5, up to Equation (11). However, we need to bound Terms (I), (J) and (K) in Equation (11) differently. According to our assumption on bounded stochastic gradients, we know that v_T^x and v_T^y are both upper bounded by TG^2, which we will use throughout the proof.

Term (I)

$$
\begin{aligned}
2\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\Phi(x_t) - \Phi(x_{t+1})}{\eta_t}\right] &\leq 2\mathbb{E}\left[\frac{\Phi(x_0)}{\eta_0} - \frac{\Phi(x_T)}{\eta_{T-1}} + \sum_{t=1}^{T-1} \Phi(x_t)\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\right] \\
&\leq 2\mathbb{E}\left[\frac{\Phi_{\max}}{\eta_0} - \frac{\Phi^*}{\eta_{T-1}} + \sum_{t=1}^{T-1} \Phi_{\max}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\right] \\
&= 2\mathbb{E}\left[\frac{\Delta\Phi}{\eta_{T-1}}\right] = 2\mathbb{E}\left[\frac{\Delta\Phi}{\eta^x}\left(v_T^x\right)^\alpha\right] \leq \frac{2\Delta\Phi G^{2\alpha} T^\alpha}{\eta^x}.
\end{aligned}
$$

Term (J)

$$
\begin{aligned}
2l\kappa \sum_{t=0}^{T-1} \mathbb{E}\left[\eta_t \left\|\nabla_x \widetilde{f}(x_t, y_t)\right\|^2\right] &= 2l\kappa\eta^x \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{1}{\left(v_{t+1}^x\right)^\alpha} \left\|\nabla_x \widetilde{f}(x_t, y_t)\right\|^2\right] \\
&\le 2l\kappa\eta^x \mathbb{E}\left[\left(\frac{v_0^x}{\left(v_0^x\right)^\alpha} + \sum_{t=0}^{T-1} \frac{1}{\left(v_{t+1}^x\right)^\alpha} \left\|\nabla_x \widetilde{f}(x_t, y_t)\right\|^2\right)\right] \\
&\le \frac{2l\kappa\eta^x}{1-\alpha} \mathbb{E}\left[\left(v_T^x\right)^{1-\alpha}\right] \le \frac{2l\kappa\eta^x G^{2-2\alpha} T^{1-\alpha}}{1-\alpha}.
\end{aligned}
$$

Term (K) According to the smoothness and strong-concavity of $f(x_t, \cdot)$, we have

$$
\mathbb{E}\left[\sum_{t=0}^{T-1} \|\nabla_x f(x_t, y_t) - \nabla \Phi(x_t)\|^2\right] \le l^2 \mathbb{E}\left[\sum_{t=0}^{T-1} \|y_t - y_t^*\|^2\right] \le 2l\kappa \mathbb{E}\left[\sum_{t=0}^{T-1} \left(f(x_t, y_t^*) - f(x_t, y_t)\right)\right].
$$

To bound the RHS, we use Young's inequality and have

$$
\left\|y_{t+1} - y_{t+1}^*\right\|^2 \le (1 + \lambda_t)\|y_{t+1} - y_t^*\|^2 + \left(1 + \frac{1}{\lambda_t}\right)\|y_{t+1}^* - y_t^*\|^2.
$$

Then applying Theorem 10 with $\lambda_t = \frac{\mu\gamma_t}{2}$ gives us

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T-1} \left(f(x_t, y_t^*) - f(x_t, y_t)\right)\right] \\
&\le \mathbb{E}\left[\sum_{t=0}^{T-1} \left(\frac{1 - \gamma_t\mu}{2\gamma_t}\|y_t - y_t^*\|^2 - \frac{1}{\gamma_t(2 + \mu\gamma_t)}\|y_{t+1} - y_{t+1}^*\|^2\right)\right] \\
&\quad + \underbrace{\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\gamma_t}{2}\left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2\right]}_{\text{(L)}} + \underbrace{\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\left(1 + \frac{2}{\mu\gamma_t}\right)}{\gamma_t(2 + \mu\gamma_t)}\|y_{t+1}^* - y_t^*\|^2\right]}_{\text{(M)}},
\end{aligned}
$$

where the first term is $\mathcal{O}(1)$ according to Theorem 13. The other two terms can be bounded as follow.

Term (L)

$$
\le \mathbb{E}\left[\frac{\eta^y}{2}\left(\frac{v_0^y}{\left(v_0^y\right)^\beta} + \sum_{t=0}^{T-1} \frac{1}{\left(v_{t+1}^y\right)^\beta}\left\|\nabla_y \widetilde{f}(x_t, y_t)\right\|^2\right)\right] \le \mathbb{E}\left[\frac{\eta^y}{2(1-\beta)}\left(v_T^y\right)^{1-\beta}\right] \le \frac{\eta^y G^{2-2\beta} T^{1-\beta}}{2(1-\beta)}.
$$

Term (M)

$$
= \mathbb{E}\left[\sum_{t=0}^{T-1} \left(\frac{1}{\left(v_{t+1}^y\right)^\beta} + \frac{2}{\mu\eta^y}\right) \frac{\left(v_{t+1}^y\right)^{2\beta}}{2\eta^y(1 + \lambda_t)}\|y_{t+1}^* - y_t^*\|^2\right]
$$

$$
\begin{aligned}
&\leq \left(\frac{1}{2\left(v_0^y\right)^{\beta}\eta^y} + \frac{1}{\mu(\eta^y)^2}\right)\mathbb{E}\left[\sum_{t=0}^{T-1}\left(v_{t+1}^y\right)^{2\beta}\left\|y_{t+1}^* - y_t^*\right\|^2\right] \\
&\leq \left(\frac{1}{2\left(v_0^y\right)^{\beta}\eta^y} + \frac{1}{\mu(\eta^y)^2}\right)\mathbb{E}\left[\left(v_T^y\right)^{2\beta}\sum_{t=0}^{T-1}\left\|y_{t+1}^* - y_t^*\right\|^2\right] \\
&\leq \left(\frac{\kappa^2}{2\left(v_0^y\right)^{\beta}\eta^y} + \frac{\kappa^2}{\mu(\eta^y)^2}\right)\mathbb{E}\left[\left(v_T^y\right)^{2\beta}\sum_{t=0}^{T-1}\left\|x_{t+1} - x_t\right\|^2\right] \\
&= \left(\frac{\left(\eta^x\right)^2\kappa^2}{2\left(v_0^y\right)^{\beta}\eta^y} + \frac{\left(\eta^x\right)^2\kappa^2}{\mu(\eta^y)^2}\right)\mathbb{E}\left[\left(v_T^y\right)^{2\beta}\sum_{t=0}^{T-1}\frac{1}{\left(v_{t+1}^x\right)^{2\alpha}}\left\|\nabla_x\widetilde{f}(x_t,y_t)\right\|^2\right] \\
&\leq \left(\frac{\left(\eta^x\right)^2\kappa^2}{2\left(v_0^y\right)^{\beta}\eta^y} + \frac{\left(\eta^x\right)^2\kappa^2}{\mu(\eta^y)^2}\right)\mathbb{E}\left[\left(v_T^y\right)^{2\beta}\left(\frac{1+\log v_T^x - \log v_0^x}{\left(v_0^x\right)^{2\alpha-1}}\cdot\mathbf{1}_{\alpha\geq 0.5} + \frac{\left(v_T^x\right)^{1-2\alpha}}{1-2\alpha}\cdot\mathbf{1}_{\alpha<0.5}\right)\right] \\
&\leq \left(\frac{\left(\eta^x\right)^2\kappa^2}{2\left(v_0^y\right)^{\beta}\eta^y} + \frac{\left(\eta^x\right)^2\kappa^2}{\mu(\eta^y)^2}\right)\left(\frac{\left(1+\log G^2T - \log v_0^x\right)G^{4\beta}T^{2\beta}}{\left(v_0^x\right)^{2\alpha-1}}\cdot\mathbf{1}_{\alpha\geq 0.5} + \frac{G^{2-4\alpha+4\beta}T^{1-2\alpha+2\beta}}{1-2\alpha}\cdot\mathbf{1}_{\alpha<0.5}\right),
\end{aligned}
$$

where we used the the Lipschitzness of $y^*(\cdot)$ in the third inequality.

Summarizing all the terms, we finish the proof.

■