Optimistic Dual Extrapolation for Coherent Non-monotone Variational Inequalities

Chaobing Song[†][*] Zhengyuan Zhou[+] Yichao Zhou[‡] Yong Jiang[†] Yi Ma[‡]

[†]Tsinghua-Berkeley Shenzhen Institute, Tsinghua University
songcb16@mails.tsinghua.edu.cn, jiangy@sz.tsinghua.edu.cn
[‡]Department of EECS, University of California, Berkeley
zyc@berkeley.edu, yima@eecs.berkeley.edu
[+]Stern School of Business, New York University, zzhou@stern.nyu.edu

Abstract

The optimization problems associated with training generative adversarial neural networks can be largely reduced to certain *non-monotone* variational inequality problems (VIPs), whereas existing convergence results are mostly based on monotone or strongly monotone assumptions. In this paper, we propose *optimistic dual extrapolation (OptDE)*, a method that only performs *one* gradient evaluation per iteration. We show that OptDE is provably convergent to *a strong solution* under different coherent non-monotone assumptions. In particular, when a *weak solution* exists, the convergence rate of our method is $O(1/\epsilon^2)$, which matches the best existing result of the methods with two gradient evaluations. Further, when a σ*-weak solution* exists, the convergence guarantee is improved to the linear rate $O(\log\frac{1}{\epsilon})$. Along the way–as a byproduct of our inquiries into non-monotone variational inequalities–we provide the near-optimal $O\big(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\big)$ convergence guarantee in terms of restricted strong merit function for monotone variational inequalities. We also show how our results can be naturally generalized to the stochastic setting, and obtain corresponding new convergence results. Taken together, our results contribute to the broad landscape of variational inequality–both non-monotone and monotone alike–by providing a novel and more practical algorithm with the state-of-the-art convergence guarantees.

1 Introduction

Variational inequality (VI) provides a principled framework for minimax problems via their first-order optimality conditions. Given a closed convex set $\mathcal{W} \subset \mathbb{R}^d$ and an operator $F : \mathcal{W} \to \mathbb{R}^d$, the variational inequality problem VIP$(F, \mathcal{W})$ aims to find a solution $\boldsymbol{w}^* \in \mathcal{W}$ such that:

$$\forall \boldsymbol{w} \in \mathcal{W},\ \langle F(\boldsymbol{w}^*), \boldsymbol{w} - \boldsymbol{w}^* \rangle \geq 0, \tag{1}$$

where $\boldsymbol{w}^*$ is called a *strong solution* of VIP$(F, \mathcal{W})$. For the minimax problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}), \tag{2}$$

let $\mathcal{W} \equiv \mathcal{X} \times \mathcal{Y}, \boldsymbol{w} \equiv \left[\begin{smallmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{smallmatrix}\right], F(\boldsymbol{w}) \equiv \left[\begin{smallmatrix}\nabla_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{y})\\ -\nabla_{\boldsymbol{y}} f(\boldsymbol{x},\boldsymbol{y})\end{smallmatrix}\right]$. Then solving (1) is equivalent to finding a first-order Nash equilibrium of the minimax problem (2) [32].

[*]This work was conducted during Chaobing Song's visit to Professor Yi Ma's group at UC Berkeley.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Convex-Concave Minimax Problems. The operator $F(\boldsymbol{w})$ will be *monotone* if

$$\forall \boldsymbol{w}, \boldsymbol{v} \in \mathcal{W}, \ \langle F(\boldsymbol{w}) - F(\boldsymbol{v}), \boldsymbol{w} - \boldsymbol{v} \rangle \geq 0. \tag{3}$$

VI with monotone operators has been well studied, which provides a concise and optimal framework for convex-concave minimax problems [29]. For monotone $\text{VIP}(F, \mathcal{W})$, it is well known that the strong solution satisfying (1) is also equivalent to the solution $\boldsymbol{w}^* \in \mathcal{W}$ satisfying:

$$\forall \boldsymbol{w} \in \mathcal{W}, \ \langle F(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^* \rangle \geq 0, \tag{4}$$

where $\boldsymbol{w}^*$ is called a *weak solution* of $\text{VIP}(F, \mathcal{W})$. A classical result [29] under the monotone and Lipschitz continuous assumptions is that the *Mirror-Prox* algorithm [29] can converge to an ϵ-accurate weak solution in terms of ergodic averaging in $O(1/\epsilon)$ iterations, which is optimal for first-order methods in solving monotone VIPs [30, 33]. Nemirovski's Mirror-Prox is a non-Euclidean extension of the extragradient method [22] from the perspective of mirror descent. Another important non-Euclidean extension is Nesterov's *dual extrapolation* [31] from the perspective of dual averaging, which also has the optimal $O(1/\epsilon)$ convergence rate. The main difference between mirror descent and dual averaging is the way of combining the constraint (or the regularization term if exists) into the projection (or the proximal) step [31].

Despite obtaining the optimal convergence rate, both Mirror-Prox and dual extrapolation are *two-call* extragradient methods that need to evaluate gradients *twice per iteration*. In some contexts such as training deep neural networks, evaluating gradients can be expensive. Thus it will have significant practical benefits if we only need one gradient evaluation per iteration and still maintain the same convergence rate. In terms of *single-call* methods for minimax problems, vanilla gradient descent ascent (and its mirror descent generalizations) might be a natural choice. Unfortunately, it is not guaranteed and it can diverge even in simple monotone settings [24]. Consequently, after the (two-call) extragradient method [22], several *single-call* extragradient methods [35, 3, 6, 27] have been analyzed under the monotone setting and share the same convergence rates with Mirror-Prox and dual extrapolation [17]. However, there is an increasing trend in applying these single-call extragradient methods to stabilize the training of generative adversarial networks (GAN) [8, 12, 34], which is *nonconvex-nonconcave* in general and hence has remained underexplored.

Nonconvex-Nonconcave Minimax Problems. Despite the well-developed convergence theory for monotone VIPs and thus for convex-concave minimax problems, many minimax problems arising in modern machine learning are nevertheless *nonconvex-nonconcave*, such as GAN [14], adversarial training [15], gradient reversal for domain adaption [11], and multi-agent reinforcement learning [38]. As a result, the corresponding VI is not monotone and the aforementioned theoretical guarantees for monotone VIPs no longer apply. First, for non-monotone VIPs, it is nontrivial to obtain the rate of convergence to a weak solution, thus one may explore the rate of convergence to a strong solution instead. Second, without the monotone property, the ergodic averaging technique [22] will no longer have theoretical guarantees, thus we might need to choose the *last iterate* or *best iterate*. However, the classical convergence result [29] said little about the rate of convergence to a weak solution or the convergence of last iterate or best iterate.[2]

To obtain theoretical guarantees beyond the monotone setting, a common approach is to relax the lower bound (3) in the monotone assumption. Along this research line, several more general assumptions have been proposed, such as the *pseudo-monotone* assumption [20, 16] and its variants [19], and the *generalized monotone* assumption [7]. In the machine learning community, similar concepts have also been proposed, such as variational coherence [41, 42]. For simplicity, we coin the problem class along this research line as *coherent non-monotone variational inequalities*. Among them, [7] is the first to provide explicit global convergence results such that the best iterate of the N-EG method [7] can converge to an ϵ-accurate strong solution in $O(1/\epsilon^2)$ iterations under the generalized monotone and Lipschitz continuous assumptions. However, N-EG needs to evaluate gradient *twice per iteration*, which is less desirable when gradient evaluation is expensive. For the single-call extragradient method [4], under a second-order condition[3], very recently [17] has provided local linear convergence results in certain non-monotone setting, while the constants in these results remain implicit. The following problem remains open: *Can single-call extragradient methods have explicit global convergence results beyond the monotone setting?*

[2] Recently, [13] shows the first tight last iterate result for general smooth convex-concave minimax problems with Lipschitz derivatives of operators.

[3] As we will see, it is a localized version of our assumption.

Table 1: **Iteration complexity for finding an ϵ-accurate solution in the deterministic setting.** (In both Tables 1 and 2, "—" denotes the corresponding results are not known or can not be obtained.)

Convergence measure	Merit function (Definition 1)		Distance $\|\cdot - \boldsymbol{w}^*\|^2$
Algorithm	N-EG [7]	OptDE **(this Paper)**	OptDE **(this Paper)**
Weak solution exists	$O(1/\epsilon^2)$	$O(1/\epsilon^2)$	—
σ-weak solution exists	—	$O(\log \frac{1}{\epsilon})$	$O(\log \frac{1}{\epsilon})$
No. of gradient calls	2	1	1

Table 2: **Stochastic oracle complexity for finding an expected ϵ-accurate solution in the stochastic setting.**

Convergence measure	Merit function (Definition 1)		Distance $\mathbb{E}[\|\cdot - \boldsymbol{w}^*\|^2]$	
Algorithm	SEG [18]	SOptDE **(this paper)**	ESA [19]	SOptDE **(this paper)**
Weak solution exists	$O(1/\epsilon^4)$	$O(1/\epsilon^4)$	—	—
σ-weak solution exists	—	$O(1/\epsilon^2 \log \frac{1}{\epsilon})$	$O(1/\epsilon)$	$O(1/\epsilon)$
No. of gradient calls	2	1	2	1

Contributions of This Paper. In this paper we develop an *Optimistic Dual Extrapolation (OptDE)* method that provably converges to a strong solution for coherent non-monotone VIPs. The OptDE method can be viewed as a single-call variant of Nesterov's dual extrapolation that maintains its "anticipatory" properties. We characterize convergence rates of the best iterate[4] of OptDE under two coherent non-monotone assumptions, where the merit function is given in Definition 1 and $\|\cdot\|$ is the natural norm used in algorithms. As shown in Table 1, when the problem has a *weak solution* $\boldsymbol{w}^*$, our method matches the best known rate $O(1/\epsilon^2)$ of N-EG [7]. Further strengthening the assumption to that a *σ-weak solution* $\boldsymbol{w}^*$ exists with $\sigma > 0$ – nevertheless a weaker condition than the strongly monotone assumption required in previous work, we are able to obtain a linear convergence rate of $O(\log \frac{1}{\epsilon})$. For this setting, we can also use the distance $\|\cdot - \boldsymbol{w}^*\|^2$ to measure the progress and obtain a linear convergence result; meanwhile, despite not shown in Table 1, we also obtain a linear convergence result of the last iterate. Our result shows that even under the two coherent non-monotone assumptions, the convergence rate of single-call extragradient methods can be comparable to that of the N-EG method with two gradient evaluations per iteration.

Our coherent non-monotone analysis for the setting that a σ-weak solution exists has two meaningful corollaries about best iterate and last iterate *in the monotone setting*, respectively: With *a regularization trick*, both the best iterate and last iterate[5] of OptDE can be an ϵ-accurate solution in $O(\frac{1}{\epsilon} \log \frac{1}{\epsilon})$ number of iterations. To our knowledge, the near-optimal result $O(\frac{1}{\epsilon} \log \frac{1}{\epsilon})$ for attaining an ϵ-accurate strong solution was only appeared in [9] very recently with a two-loop Halpern iteration method, while our result is obtained by the simpler single-loop single-call OptDE method.

Meanwhile, we extend the OptDE algorithm to the stochastic setting as *Stochastic OptDE (SOptDE)* and show that our results in the deterministic setting can be naturally generalized to the *stochastic setting*. This allows us to characterize the stochastic oracle complexity (*i.e.,* the number of stochastic oracles we access) of SOptDE under the coherent non-monotone assumptions. The results under the stochastic setting are summarized in Table 2.[6] As we see, the results match the best-known results of SEG [7] [18] and ESA [19] respectively, while both SEG and ESA need two gradient evaluations per iteration. Meanwhile, under the assumption that a σ-weak solution exists, we obtain the first theoretical guarantee in terms of the merit function in Definition 1.

Last but not least, different from N-EG [7] and ESA [19], the proposed OptDE and SOptDE algorithms only need the norm square $\|\cdot\|^2$ being strongly convex but not necessarily globally Lipschitz continuous, which will be significant if $\|\cdot\|$ is a non-Euclidean norm: $\|\cdot\|^2$ can not be strongly convex and globally Lipschitz continuous simultaneously in general.

[4] For given a number of iterations, the best iterate can be explicitly found and happen before the last iterate.

[5] Here the last iterate is not in the classical sense, which will be explained in Section 3.

[6] The results of the SEG [18], ESA [19] algorithms are given under pseudomonotone and strongly pseudomonotone assumptions respectively, which are slightly stronger than our assumptions.

[7] The original result of SEG is given by "square natural residual", which can be used to derive the strong solution guarantee in Table 2 (see the supplementary material for detail).

2 Technical Assumptions

Notations: For $K \in \mathbb{Z}_+$, let $[K] := \{1, 2, \ldots, K\}$. Let lower case boldface alphabets denote vectors, such as $\boldsymbol{x} \in \mathbb{R}^d$ and lower case alphabets with subscript denote elements, such as $x_1, x_2, \ldots, x_d$. Let $\|\cdot\|$ denote a general norm. Let $\|\cdot\|_*$ denote the dual norm of $\|\cdot\|$ defined by $\|\boldsymbol{y}\|_* := \max_{\|\boldsymbol{x}\|\leq 1}\langle \boldsymbol{x}, \boldsymbol{y}\rangle$. For $\boldsymbol{x} \in \mathbb{R}^d$ and $p \geq 1$, let $\|\boldsymbol{x}\|_p := \left(\sum_{i=1}^d |x_i|^p\right)^{\frac{1}{p}}$.

To measure the accuracy of iterates to a strong solution, we consider the following "restricted strong merit function".

Definition 1 (Restricted strong merit function) *$\tilde{\boldsymbol{w}} \in \mathcal{W}$ is an ϵ-accurate strong solution of the VIP$(F, \mathcal{W})$ with a fixed parameter $D > 0$ if*

$$\sup_{\boldsymbol{w}\in\mathcal{W}, \|\boldsymbol{w}-\tilde{\boldsymbol{w}}\|_2\leq D} \langle F(\tilde{\boldsymbol{w}}), \tilde{\boldsymbol{w}} - \boldsymbol{w}\rangle \leq \epsilon. \tag{5}$$

With $\epsilon \to 0$ and $D \to +\infty$, Definition 1 becomes the definition of the strong solution in (1). In the nonconvex-nonconcave minimax setting, Definition 1 has been proposed as the definition of the ϵ-accurate first-order Nash equilibrium [32]. If $\mathcal{W}$ is a bounded set, then we still have an effective measure even if $D \to +\infty$; if $\mathcal{W}$ is unbounded, then D needs to be a finite positive parameter. To give a unified measure for both bounded and unbounded settings, we set D to be a finite positive parameter.

Throughout this paper, we make the following standard Lipschitz continuous assumption.

Assumption 1 *For the VIP$(F, \mathcal{W})$ in (1), $\forall \boldsymbol{w}, \boldsymbol{v} \in \mathcal{W}$, $\|F(\boldsymbol{w}) - F(\boldsymbol{v})\|_* \leq L\|\boldsymbol{w} - \boldsymbol{v}\|$, where $L > 0$ is the Lipschitz constant.*

Meanwhile, we assume that the (possible non-Euclidean) norm $\|\cdot\|$ satisfies Assumption 2.

Assumption 2 *$\frac{1}{2}\|\boldsymbol{w}\|^2$ is γ-strongly convex ($0 < \gamma \leq 1$) with respect to (w.r.t.) $\|\cdot\|$ and the dual norm of gradient $\nabla\frac{1}{2}\|\boldsymbol{w}\|^2$ is bounded by $\delta\|\boldsymbol{w}\|(\delta > 0)$:*

$$\frac{1}{2}\|\boldsymbol{w}\|^2 \geq \frac{1}{2}\|\boldsymbol{v}\|^2 + \langle\nabla\frac{1}{2}\|\boldsymbol{v}\|^2, \boldsymbol{w} - \boldsymbol{v}\rangle + \frac{\gamma}{2}\|\boldsymbol{w} - \boldsymbol{v}\|^2, \tag{6}$$

$$\left\|\nabla\frac{1}{2}\|\boldsymbol{w}\|^2\right\|_* \leq \delta\|\boldsymbol{w}\|. \tag{7}$$

From [1], $\frac{1}{2}\|\cdot\|_p^2 (1 < p \leq 2)$ is $(p-1)$-strongly convex $w.r.t.$ $\|\cdot\|_p$. Without loss of generality, in Assumption 2, we assume $0 < \gamma \leq 1$. For all the norm setting $\frac{1}{2}\|\cdot\|_p^2 (1 < p \leq 2)$, we have $\delta = 1$.

For the norm $\|\cdot\|$, we define the prox-mapping as

$$P_{\boldsymbol{v}}(\boldsymbol{w}) := \arg\min_{\boldsymbol{z}\in\mathcal{W}} \left\{\langle \boldsymbol{w}, \boldsymbol{z}\rangle + \frac{1}{2\gamma}\|\boldsymbol{z} - \boldsymbol{v}\|^2\right\}, \tag{8}$$

and assume that it can be solved efficiently. Meanwhile, we also define the corresponding Bregman divergence of $\frac{1}{2}\|\cdot\|^2$: $\forall \boldsymbol{w}, \boldsymbol{v} \in \mathcal{W}$,

$$V_{\boldsymbol{v}}(\boldsymbol{w}) := \frac{1}{2}\|\boldsymbol{w}\|^2 - \frac{1}{2}\|\boldsymbol{v}\|^2 - \left\langle\nabla\frac{1}{2}\|\boldsymbol{v}\|^2, \boldsymbol{w} - \boldsymbol{v}\right\rangle. \tag{9}$$

Obviously we have $V_{\boldsymbol{v}}(\boldsymbol{w}) \geq \frac{\gamma}{2}\|\boldsymbol{w} - \boldsymbol{v}\|^2$.

Then we make Assumptions 3 and 4 for the coherent non-monotone VIP$(F, \mathcal{W})$ we study.

Assumption 3 (Existence of a weak solution) *For the VIP$(F, \mathcal{W})$ in (1), there exists a weak solution $\boldsymbol{w}^* \in \mathcal{W}$ such that $\forall \boldsymbol{w} \in \mathcal{W}$, $\langle F(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^*\rangle \geq 0$.*

Assumption 4 (Existence of a σ-weak solution) *For the VIP$(F, \mathcal{W})$ in (1), given $\boldsymbol{w}_0 \in \mathcal{W}$, there exists a σ-weak solution $\boldsymbol{w}^* \in \mathcal{W}$ with parameter $\sigma > 0$ such that $\forall \boldsymbol{w} \in \mathcal{W}$, $\langle F(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^*\rangle \geq \frac{\sigma}{\gamma}(V_{\boldsymbol{w}-\boldsymbol{w}_0}(\boldsymbol{w}^* - \boldsymbol{w}_0) + V_{\boldsymbol{w}^*-\boldsymbol{w}_0}(\boldsymbol{w} - \boldsymbol{w}_0))$.*

Algorithm 1 Optimistic Dual Extrapolation

```
1:  Input: Lipschitz constant $L > 0$ from Assumption 1, $\gamma, \delta > 0$ from Assumption 2. The VIP($F, \mathcal{W}$) satisfying Assumption 3 ($\sigma = 0$) or Assumption 4 ($\sigma > 0$).
2:  $A_0 = 0, 0 < \alpha \le \min\left\{\frac{1}{4\sqrt{2}}, \frac{\sqrt{3}}{4\sqrt{\gamma}}\right\}$.
3:  $\boldsymbol{w}_0 = \boldsymbol{z}_0 \in \mathcal{W}, \boldsymbol{g}_0 = \mathbf{0}$.
4:  for $k = 1, 2, 3, \ldots, K$ do
5:      $a_k = \frac{\alpha\gamma(1+\sigma A_{k-1})}{L}, A_k = A_{k-1} + a_k$.
6:      $\boldsymbol{w}_k = P_{\boldsymbol{z}_{k-1}}\big(\frac{\alpha}{L}F(\boldsymbol{w}_{k-1})\big)$.
7:      $\boldsymbol{g}_k = \boldsymbol{g}_{k-1} + a_k\big(F(\boldsymbol{w}_k) - \frac{\sigma}{\gamma}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{w}_0\|^2\big)$.
8:      $\boldsymbol{z}_k = P_{\boldsymbol{w}_0}\big(\frac{1}{1+\sigma A_k}\boldsymbol{g}_k\big)$.
9:  end for
10: $\tilde{\boldsymbol{w}}_K = \arg\min_{\boldsymbol{w}_k : k \in [K]}(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|)$.
11: return $\tilde{\boldsymbol{w}}_K$.
```

Assumption 3 assumes the existence of weak solutions, which is also adopted in [25]. Assumption 3 is slightly weaker than the variational coherence assumption [41, 42] or the generalized monotone assumption [7]. Some nontrivial examples satisfying the generalized monotone assumption can be found in [7, 44, 28]. The generalized monotone assumption is in turn weaker than the pseudo-monotone assumption [20, 16], which is weaker than the monotone assumption (3).

Remark 1 *In the monotone setting, the weak solution set and strong solution set are equivalent to each other; meanwhile, an approximate strong solution is also an approximate weak solution, while the reverse does not hold in general (which can explain the terms "weak" and "strong"). However, in the non-monotone setting, if the operator F is continuous, a weak solution is a strong solution, while the reverse is not true in general [21, Chapter 3]. For instance, consider the minimax problem $\min_{x\in\mathbb{R}}\max_{y\in\mathbb{R}} x^2y^2$ and let $F(x, y) = (2xy^2, -2yx^2)^T$ with $(x, y) \in \mathbb{R}^2$. Then we can verify that $(0, 0)$ is the only weak solution of VIP($F, \mathbb{R}^2$), while the set of strong solution is the x-axis or the y-axis, and the set of Nash equilibrium is the y-axis.*

Assumption 4 further assumes a stronger variant of Assumption 3, which is also called as strongly variational stability in [41]. For the Euclidean setting where $\|\cdot\| := \|\cdot\|_2$ and thus $\gamma = 1$, the inequality is simplified to $\langle F(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^*\rangle \ge \sigma\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2$. Assumption 4 is weaker than the strongly pseudo-monotone [19] and strongly monotone assumptions, but as we will see, is already sufficient to ensure a linear convergence rate for our method.

Remark 2 *Our main motivation in making Assumptions 3 and 4 is to prove explicit global convergence results for VIP($F, \mathcal{W}$) under conditions* as weak as possible. *However, the non-monotone subsets of Assumptions 3 and 4,* a.k.a., *pseudomonotone and strongly pseudomonotone respectively, also have many real applications in competitive exchange economy [2], fractional programming [10, 37], and product pricing [5]. Meanwhile, the restriction of Assumption 4 in minimization problems such as one-point convexity [23] is also used in analyzing neural networks.*

3 Optimistic Dual Extrapolation

In this section, we present the *optimistic dual extrapolation (OptDE)* algorithm for solving the VIP($F, \mathcal{W}$) in (1). The method is a single-call variant of Nesterov's dual extrapolation [31]. The overall algorithm is summarized as Algorithm 1. The algorithm works under either Assumption 3 by setting $\sigma = 0$ or Assumption 4 with $\sigma > 0$.

For Algorithm 1, we define two constants A_0 and α in Step 2. Then we initialize three vectors $\boldsymbol{w}_0, \boldsymbol{z}_0$ and $\boldsymbol{g}_0$ in Step 3. In the main loop, we update the two positive numbers a_k and A_k in Step 5. Then we perform an "extrapolation" step in Step 6 and then "dual averaging" steps in Steps 7 and 8. As we see, as Algorithm 1 only performs one new gradient evaluation in Step 8, it is "optimistic" [36] hence the name "optimistic dual extrapolation". Once Algorithm 1 runs K iterations, we return the best iterate measured by the sum of residual norms $\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|$[8].

[8] This return value is given according to our convergence analysis.

Compared with Nesterov's dual extrapolation, the main difference is that the extrapolation Step 6 is a prox-mapping on $F(\boldsymbol{w}_{k-1})$, not on $F(\boldsymbol{z}_{k-1})$. Compared with past extra-gradient [17, 36], the main difference is that we perform dual averaging by Steps 7 and 8, instead of a "mirror descent" step. Compared with N-EG which is claimed to be a non-Euclidean extragradient method [7], not only we perform just one gradient evaluation per iteration but also do not require $\frac{1}{2}\|\cdot\|^2$ to have bounded Lipschitz continuous gradients, which is significant in the non-Euclidean setting since the norm square $\frac{1}{2}\|\cdot\|_p^2$ for $p \in (1, 2)$ may not have globally bounded Lipschitz continuous gradients.

In the following, we assume $\boldsymbol{w}^*$ is a solution that satisfies Assumption 3 if $\sigma = 0$ or satisfies Assumption 4 if $\sigma > 0$.

Theorem 1 *Let Assumptions 1 and 2 hold. For both settings $\sigma = 0$ (i.e., Assumption 3 holds) and $\sigma > 0$ (i.e., Assumption 4 holds), after K iterations, Algorithm 1 returns a $\tilde{\boldsymbol{w}}_K$ such that*

$$\sup_{\boldsymbol{w}\in\mathcal{W}, \|\tilde{\boldsymbol{w}}_K - \boldsymbol{w}\| \leq D} \langle F(\tilde{\boldsymbol{w}}_K), \tilde{\boldsymbol{w}}_K - \boldsymbol{w}\rangle \leq C_0 D \|\boldsymbol{w}_0 - \boldsymbol{w}^*\| \sqrt{\frac{L}{A_{K-1} + a_1}}, \tag{10}$$

with $C_0 = \left(1 + \frac{\delta}{\alpha\gamma}\right)\sqrt{\frac{8\alpha}{\gamma}}$, $a_1 = \frac{\alpha\gamma}{L}$, and

$$A_{K-1} = \begin{cases} \frac{\alpha\gamma(K-1)}{L} & \text{if } \sigma = 0, \\ \frac{1}{\sigma}\left(1 + \frac{\alpha\gamma\sigma}{L}\right)^{K-1} - \frac{1}{\sigma} & \text{if } \sigma > 0. \end{cases} \tag{11}$$

Particularly if $\sigma > 0$, we also have

$$\|\tilde{\boldsymbol{w}}_K - \boldsymbol{w}^*\| \leq \frac{C_0}{\sigma}\|\boldsymbol{w}_0 - \boldsymbol{w}^*\| \sqrt{\frac{L}{A_{K-1} + a_1}}. \tag{12}$$

Proof. See Section C.4. ■

Theorem 1 implies our main result in Table 1. As we see, for $\sigma = 0$, except for constants, our result is the same with the two-call extragradient method N-EG [7]. However, to analyze single-call methods, particularly for the setting $\sigma = 0$, the analysis is much more involved and leads to an interesting criterion of return value in Step 10 of Algorithm 1. For the setting $\sigma > 0$, then linear convergence rates can be obtained in terms of both restricted strong merit solution and solution distance. Meanwhile, for the setting $\sigma > 0$, our result in terms of restricted strong merit solution (10) can not be implied by the result of the solution distance (12), while the reverse side is true. Furthermore, when $\sigma > 0$, the result (10) is also used in deriving Corollary 1 for the monotone setting. Finally, to simplify our analysis, we did not yet optimize the constants in (10) and (12), which probably can be further improved.

In Theorem 1, we provide a unified result for the two settings $\sigma = 0$ and $\sigma > 0$ in terms of the best iterate. However, when $\sigma > 0$, we can also prove linear convergence rates in terms of last iterate, which is given in Proposition 1 below.

Proposition 1 *Let Assumptions 1 and 2 hold. For the setting $\sigma > 0$ (*i.e.*, Assumption 4 holds), $\forall K \geq 1$, after K iterations, Algorithm 1 returns a $\boldsymbol{w}_K$ such that*

$$\sup_{\boldsymbol{w}\in\mathcal{W}, \|\boldsymbol{w}_K - \boldsymbol{w}\| \leq D} \langle F(\boldsymbol{w}_K), \boldsymbol{w}_K - \boldsymbol{w}\rangle \leq C_0 D \|\boldsymbol{w}_0 - \boldsymbol{w}^*\| \sqrt{\frac{L}{a_{K-1}}}, \tag{13}$$

with C_0 defined in Theorem 1, $a_0 = a_1$ and $\forall K \geq 1$,

$$a_K = \frac{\alpha\gamma}{L}\left(1 + \frac{\alpha\gamma\sigma}{L}\right)^{K-1}. \tag{14}$$

Meanwhile, we also have

$$\|\boldsymbol{w}_K - \boldsymbol{w}^*\| \leq \frac{C_0}{\sigma}\|\boldsymbol{w}_0 - \boldsymbol{w}^*\| \sqrt{\frac{L}{a_{K-1}}}. \tag{15}$$

Proof. See Section C.5. ■

By Proposition 1, to prove the linear convergence of the last iterate, we do not need the strongly monotone assumption, but only Assumption 4. Despite the last iterate also has a linear convergence rate, it is slower than the rate of best iterate in Theorem 1. As we will see, Proposition 1 will also be used to prove the last iterate convergence for the monotone setting in a non-classical sense.

Remark 3 *The motivation behind OptDE is that by generalizing Nesterov's estimation sequence, we can perform a unified convergence analysis under Assumptions 3 and 4. However, as shown in [40], if a regularizer exists, the (regularized) dual averaging steps (Steps 7 and 8 of Algorithm 1) can help us better explore the structure of regularizers such as sparsity when it exists.*

Remark 4 *[17] has given local convergence analysis in terms of solution distance by assuming that Assumption 4 holds in a neighbourhood of the optimal solution. The analysis in [17] needs extra techniques, while the constants in the rates of [17] are implicit. Our solution distance result in* (12) *can be viewed as a global and explicit version of [17] by assuming Assumption 4 holds globally. Meanwhile, [17] does not give any result under Assumption 3 or in terms of restricted strong solution under Assumption 4 whereas our analysis does.*

Our results are mainly given under the coherent non-monotone Assumptions 3 and 4. As shown in Theorem 1, under Assumption 3 that includes the monotone assumption, we can obtain an ϵ-accurate strong solution in $O(\epsilon^{-2})$ iterations. However, in the following we show that with *a regularization trick*, the rate can be much better in the monotone setting by using our results in Theorem 1 and Proposition 1.

First, to give our results in the monotone setting, we have Lemma 1.

Lemma 1 *If the VIP$(F, \mathcal{W})$ is monotone, then the regularized problem VIP$(F + \epsilon\nabla\frac{1}{2\gamma}\|\cdot - \boldsymbol{w}_0\|^2, \mathcal{W})$ satisfies Assumption 4 with $\sigma = \epsilon$.*

Proof. See Section C.6. ■

Due to Lemma 1, we can apply Theorem 1 and Proposition 1 to the regularized problem VIP$(F + \epsilon\nabla\frac{1}{2\gamma}\|\cdot - \boldsymbol{w}_0\|^2, \mathcal{W})$, and then obtain Corollaries 1 and 2 for the VIP$(F, \mathcal{W})$, respectively.

Corollary 1 (Best iterate convergence in the monotone setting) *Given $\boldsymbol{w}_0 \in \mathcal{W}$, let Assumptions 1 and 2 hold for the regularized problem VIP$(F + \epsilon\nabla\frac{1}{2\gamma}\|\cdot - \boldsymbol{w}_0\|^2, \mathcal{W})$. By optimizing the regularized problem by Algorithm 4, then the best iterate returned by Algorithm 4 satisfies*

$$\begin{aligned}&\sup_{\boldsymbol{w}\in\mathcal{W}, \|\tilde{\boldsymbol{w}}_K - \boldsymbol{w}\| \leq D, \|\boldsymbol{w} - \boldsymbol{w}_0\| \leq D} \langle F(\tilde{\boldsymbol{w}}_K), \tilde{\boldsymbol{w}}_K - \boldsymbol{w}\rangle \\ \leq\ & D\epsilon + DC_0\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|\sqrt{\frac{L\epsilon}{\left(1 + \frac{\alpha\gamma\epsilon}{L}\right)^{K-1} - 1 + \frac{\alpha\gamma}{L}}},\end{aligned}$$

where C_0 is defined in Theorem 1.

Proof. See Section C.7. ■

Compared with Theorem 1 and Proposition 1, we need an extra condition $\|\boldsymbol{w} - \boldsymbol{w}_0\| \leq D$ in Corollary 1, which can be satisfied by choosing a large enough D. By Corollary 1, by choosing $K = O\left(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\right)$, we will obtain an $O(D\epsilon)$-accurate solution. Note that D does not appear in our algorithm and is not relevant to the choice of ϵ.

Corollary 2 (Last iterate convergence in the monotone setting) *Given $\boldsymbol{w}_0 \in \mathcal{W}$, let Assumptions 1 and 2 hold for the regularized problem VIP$(F + \epsilon\nabla\frac{1}{2\gamma}\|\cdot - \boldsymbol{w}_0\|^2, \mathcal{W})$. By optimizing the regularized problem by Algorithm 4, the last iterate of Algorithm 4 satisfies*

$$\sup_{\boldsymbol{w}\in\mathcal{W}, \|\boldsymbol{w}_K - \boldsymbol{w}\| \leq D, \|\boldsymbol{w} - \boldsymbol{w}_0\| \leq D} \langle F(\boldsymbol{w}_K), \boldsymbol{w}_K - \boldsymbol{w}\rangle \leq D\epsilon + DC_0L\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|\sqrt{\frac{1}{\alpha\gamma\left(1 + \frac{\alpha\gamma\epsilon}{L}\right)^{K-1}}},$$

where C_0 is defined in Theorem 1.

Algorithm 2 Stochastic Optimistic Dual Extrapolation

```
Input: Lipschitz constant $L > 0$ from Assumption 1, $\gamma, \delta > 0$ from Assumption 2. The VIP$(F, \mathcal{W})$ satisfying Assumption 3 ($\sigma = 0$) or Assumption 4 ($\sigma > 0$).
$A_0 = 0, \alpha = \min\{\frac{\gamma}{32}, \frac{1}{16}\}$.
$\boldsymbol{w}_0 = \boldsymbol{z}_0 \in \mathcal{W}, \boldsymbol{g}_0 = \mathbf{0}$.
for $k = 1, 2, 3, \ldots, K$ do
    $a_k = \frac{\alpha\gamma\sqrt{1+\sigma A_{k-1}}}{L}, A_k = A_{k-1} + a_k$.
    $\boldsymbol{w}_k = P_{\boldsymbol{z}_{k-1}}\big(\frac{\alpha^2\gamma}{L^2 a_k} F(\boldsymbol{w}_{k-1}; \xi_{k-1})\big)$.
    $\boldsymbol{g}_k = \boldsymbol{g}_{k-1} + a_k\big(F(\boldsymbol{w}_k; \xi_k) - \frac{\sigma}{\gamma}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{w}_0\|^2\big)$.
    $\boldsymbol{z}_k = P_{\boldsymbol{w}_0}\big(\frac{1}{1+\sigma A_k}\boldsymbol{g}_k\big)$.
end for
$\tilde{\boldsymbol{w}}_K = \boldsymbol{w}_k$, where $k$ is chosen at random with probability distribution $\{\frac{a_1}{A_K}, \frac{a_2}{A_K}, \ldots, \frac{a_K}{A_K}\}$.
return $\tilde{\boldsymbol{w}}_K$.
```

Proof. See Section C.8. ■

Similar to Corollary 1 for best iterate, in Corollary 2, by choosing $K = O\big(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\big)$, the last iterate will be an $O(D\epsilon)$-accurate strong solution, which is significantly better than the tight bound $O(1/\epsilon^2)$ for last iterate [13]. Nevertheless, it should be noted that Corollary 2 is in a non-classical sense: we do not guarantee last iterate convergence for all $K \geq 1$, but only after $K = O\big(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\big)$ with a prescribed accuracy parameter ϵ. Thus our result does not contradict with the lower bound of last iterate [13].

Meanwhile, our proof only relies on the regularized problem VIP$(F + \epsilon\nabla\frac{1}{2\gamma}\|\cdot - \boldsymbol{w}_0\|^2, \mathcal{W})$ satisfying Assumption 4 with $\sigma = \epsilon$, which holds if the VIP$(F, \mathcal{W})$ is monotone. However, it is not necessary for the VIP$(F, \mathcal{W})$ to be monotone. For instance, if the VIP$(F, \mathcal{W})$ satisfies Assumption 3 and $\boldsymbol{w}_0 = \boldsymbol{w}^*$, then the VIP$(F + \epsilon\nabla\frac{1}{2\gamma}\|\cdot - \boldsymbol{w}_0\|^2, \mathcal{W})$ also satisfies Assumption 4 with $\sigma = \epsilon$. Of course, letting $\boldsymbol{w}_0 = \boldsymbol{w}^*$ is impractical and we leave the more general setting of $\boldsymbol{w}_0$ under non-monotone settings for further research.

Remark 5 *Recently, [9] has proposed a different Halpern iteration method under the monotone and Lipschitz assumptions. The Halpern iteration method does not need to know the Lipschitz constant and thus is parameter-free, and also attains the $O\big(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\big)$ convergence rate. Nevertheless, there are two major differences: The Halpern iteration method has two-loop, while our OptDE method is a single-loop single-call method; now the Halpern iteration method is limited to the Euclidean setting, while ours can have theoretical guarantees in the non-Euclidean setting.*

4 Stochastic Optimistic Dual Extrapolation

In this section, we present a stochastic version of the above OptDE method, *a.k.a., stochastic optimistic dual extrapolation (SOptDE)*, which is given in Algorithm 2. Compared with the OptDE method in Algorithm 1, the main difference is that Algorithm 2 approximates $\{F(\boldsymbol{w}_k)\}$ by the unbiased stochastic estimations $\{F(\boldsymbol{w}_k; \xi_k)\}$, where the randomness is from the *i.i.d* random variables $\{\xi_k\}$. For simplicity, in this section, we use $\mathbb{E}_\xi[\cdot]$ to denote the expectation *w.r.t.* ξ while fixing the previous randomness; meanwhile, we use $\mathbb{E}[\cdot]$ to denote the expectation *w.r.t.* the randomness of all the history. Formally, we make Assumption 5.

Assumption 5 *$\forall \boldsymbol{w} \in \mathcal{W}$, $F(\boldsymbol{w}; \xi)$ is an unbiased estimation of $F(\boldsymbol{w})$ such that $\mathbb{E}_\xi[F(\boldsymbol{w}; \xi)] = F(\boldsymbol{w})$; meanwhile the variance of $F(\boldsymbol{w}; \xi)$ is bounded by s^2 such that $\mathbb{E}_\xi[\|F(\boldsymbol{w}; \xi) - F(\boldsymbol{w})\|_*^2] \leq s^2$.*

Meanwhile, to cancel the error from randomness, in Algorithm 2, when $\sigma > 0$, we consider a more conservative parameter setting $a_k = \frac{\alpha\gamma\sqrt{1+\sigma A_{k-1}}}{L}$ rather than $a_k = \frac{\alpha\gamma(1+\sigma A_{k-1})}{L}$ of Algorithm 1. Furthermore, because of the randomness, choosing the exact best iterate as in the deterministic case is no longer meaningful as its expectation is impossible to compute. In this case, we choose $\tilde{\boldsymbol{w}}_K$ at random according to the distribution $\{\frac{a_1}{A_K}, \frac{a_2}{A_K}, \ldots, \frac{a_K}{A_K}\}$, which also facilitates theoretical analysis[9].

[9] In practice, nevertheless, one may often consider choosing the last iterate for simplicity.

Theorem 2 *For the setting $\sigma = 0$ (i.e., Assumption 3 holds), after K iterations, Algorithm 5 returns a $\tilde{\boldsymbol{w}}_K$ such that*

$$\mathbb{E}\Big[\sup_{\boldsymbol{w}\in\mathcal{W},\|\tilde{\boldsymbol{w}}_K-\boldsymbol{w}\|\le D}\langle F(\tilde{\boldsymbol{w}}_K),\tilde{\boldsymbol{w}}_K-\boldsymbol{w}\rangle\Big]$$
$$\le \sqrt{2}(1+\delta)LD\sqrt{\frac{\|\boldsymbol{w}^*-\boldsymbol{w}_0\|^2}{8\alpha\gamma K}+\frac{s^2}{L^2}+L^2\Big(\frac{\|\boldsymbol{w}^*-\boldsymbol{w}_0\|^2}{8\alpha\gamma K}+\frac{s^2}{L^2}\Big)}+\frac{s^2}{2L^2}. \tag{16}$$

For $\sigma > 0$, (i.e., Assumption 4 holds), we have

$$\mathbb{E}[\|\tilde{\boldsymbol{w}}_K-\boldsymbol{w}^*\|^2]\le\frac{32L^2}{\sigma^2(\alpha\gamma)^2(K+1)^2}\Big(\frac{8\alpha s^2K}{L^2}+\frac{1}{2\gamma}\|\boldsymbol{w}^*-\boldsymbol{w}_0\|^2\Big). \tag{17}$$

Proof. See Section D.4. ■

As show in (16), for the setting $\sigma = 0$ (*a.k.a.*, Assumption 3) even if the number of iterations $K\to\infty$, the expected restricted strong merit function can only be upper bounded by $O\big(\frac{s}{L}\big)$. Thus to guarantee the convergence of SOptDE, the variance should be $o(1)$, such as $s^2 = O\big(\frac{1}{K}\big)$. In the Euclidean setting that $\|\cdot\| := \|\cdot\|_2$, by the concentration inequality [39], to attain a variance of $O\big(\frac{1}{K}\big)$, we need $O(K)$ samples. Thus combining the setting $s^2 = O(\frac{1}{K})$ and the result in (16), it can be verified that the single-call SOptDE method needs $O(1/\epsilon^4)$ number of samples to obtain an ϵ-accurate solution in terms of the expected restricted strong merit function.

To develop the two-call stochastic extragradient method SEG [18] under the pseudomonotone assumption[10], [18] has also considered variance reduction with a large batch size and used a "quadratic natural residual" (in our notation, it is $\mathbb{E}[\|\boldsymbol{w}_k-\boldsymbol{z}_{k-1}\|^2]$) to measure the accuracy, which in turns can be used to derive the same complexity result $O(1/\epsilon^4)$ as SOptDE in terms of expected restricted strong merit function (see the supplementary material). OSG [26] is a single-call version of SEG, which also uses quadratic natural residual as a convergence measure. However in the general constrained setting, it is not know how to convert the guarantee of quadratic natural residual of the single-call OSG into the guarantee of expected restricted strong merit function. In fact, in our single-call setting, the "(quadratic) natural residual" $\mathbb{E}[\|\boldsymbol{w}_k-\boldsymbol{z}_{k-1}\|^2]$ is no longer useful in deriving the theoretical guarantee by expected restricted strong merit function. As a result, we consider the term $\mathbb{E}[\|\boldsymbol{w}_k-\boldsymbol{z}_{k-1}\|^2+\|\boldsymbol{w}_{k-1}-\boldsymbol{z}_{k-1}\|^2]$, which makes our proof quite different from that in [18].

Under the stronger Assumption 4, our result is given in terms of the expected solution distance. As shown in (17), under Assumption 4, SOptDE can converge provably even when the variance s^2 is constant. In fact, the $O\big(\frac{1}{K}\big)$ is optimal and has been obtained by the two-call extragradient method ESA [19] under the pseudomonotone assumption. Meanwhile, [43] used the plain stochastic gradient descent algorithm and obtained the $O\big(\frac{1}{K}\big)$ result for strongly monotone variational inequalities, which can also be extended to the setting that σ-weak solution exists.

With the aggressive parameter setting $a_k = \frac{\alpha\gamma(1+\sigma A_{k-1})}{L}$ and a large batch size strategy, we also obtain the first convergence guarantee $O(1/\epsilon^2\log\frac{1}{\epsilon})$ in terms of restricted strong merit function as shown in Table 2 (see details in the supplementary material).

5 Concluding Remarks

In this paper, we proposed a single-call extragradient method *optimistic dual extrapolation (OptDE)* beyond the monotone setting and also extended it to the stochastic setting as *stochastic optimistic dual extrapolation (SOptDE)*. We systematically proved the convergence results of OptDE and SOptDE under the Assumption 3 that a weak solution exists and Assumption 4 that a strongly weak solution exists. We also show beneficial implications of our analysis in both non-monotone and monotone settings. In the future, we will further study how the proposed new methods may lead to improved computational efficiency and performance guarantees in a wide range of machine learning problems such as the training of adversarial deep neural networks.

[10] We can verify that the result in [18] can be extended under our Assumption 3.

Broader Impact

In this paper, we discuss a systematic theoretical analysis for single-call extragradient methods, which has been widely used for modern machine learning applications. The theoretical results in this paper can bring in meaningful insight and understanding for practical algorithms.

Acknowledgement

Chaobing and Yi acknowledge support from Tsinghua-Berkeley Shenzhen Institute (TBSI) Research Fund. Yichao and Yi acknowledge funding from Sony Research. Yi acknowledges support from ONR grant N00014-20-1-2002 and the joint Simons Foundation-NSF DMS grant #2031899, as well as support from Berkeley AI Research (BAIR), Berkeley FHL Vive Center for Enhanced Reality, and Berkeley Center for Augmented Cognition.

References

[1] Keith Ball, Eric A Carlen, and Elliott H Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones mathematicae*, 115(1):463–482, 1994. 4

[2] Luigi Brighi and Reinhard John. Characterizations of pseudomonotone maps and economic equilibrium. *Journal of Statistics and Management Systems*, 5(1-3):253–273, 2002. 5

[3] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011. 2

[4] Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Conference on Learning Theory*, pages 6–1, 2012. 2

[5] S Chan Choi, Wayne S DeSarbo, and Patrick T Harker. Product positioning under price competition. *Management Science*, 36(2):175–199, 1990. 5

[6] Shisheng Cui and Uday V Shanbhag. On the analysis of reflected gradient and splitting methods for monotone stochastic variational inequality problems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 4510–4515. IEEE, 2016. 2

[7] Cong D Dang and Guanghui Lan. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and applications*, 60(2):277–310, 2015. 2, 3, 5, 6

[8] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *International Conference on Learning Representations (ICLR 2018)*, 2018. 2

[9] Jelena Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *33rd Annual Conference on Learning Theory*, pages vol 125:1–24, 2020. 3, 8

[10] Alexandre Mikhajlovich Elizarov and AN Kalimullina. Maximization of the lift/drag ratio of airfoils with a turbulent boundary layer: Sharp estimates, approximation, and numerical solutions. *Computational Mathematics and Mathematical Physics*, 49(3):559–572, 2009. 5

[11] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014. 2

[12] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018. 2

[13] Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. *arXiv preprint arXiv:2002.00057*, 2020. 2, 8

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[16] Nicolas Hadjisavvas, Siegfried Schaible, and N-C Wong. Pseudomonotone operators: a survey of the theory and its applications. *Journal of Optimization Theory and Applications*, 152(1):1–20, 2012. 2, 5

[17] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems*, pages 6936–6946, 2019. 2, 6, 7

[18] Alfredo N Iusem, Alejandro Jofré, Roberto Imbuzeiro Oliveira, and Philip Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017. 3, 9, 15

[19] Aswin Kannan and Uday V Shanbhag. Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Computational Optimization and Applications*, 74(3):779–820, 2019. 2, 3, 5, 9

[20] S Karamardian. Complementarity problems over cones with monotone and pseudomonotone maps. *Journal of Optimization Theory and Applications*, 18(4):445–454, 1976. 2, 5

[21] David Kinderlehrer and Guido Stampacchia. *An introduction to variational inequalities and their applications*. SIAM, 2000. 5

[22] GM Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976. 2

[23] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in neural information processing systems*, pages 597–607, 2017. 5

[24] Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. *arXiv preprint arXiv:1802.06132*, 2018. 2

[25] Qihang Lin, Mingrui Liu, Hassan Rafique, and Tianbao Yang. Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality. *arXiv preprint arXiv:1810.10207*, 2018. 5

[26] Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. In *International Conference on Learning Representations*, 2019. 9

[27] Yu Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015. 2

[28] Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1-2):465–507, 2019. 5

[29] Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004. 2

[30] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983. 2

[31] Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007. 2, 5

[32] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, pages 14905–14916, 2019. 1, 4

[33] Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *arXiv preprint arXiv:1808.02901*, 2018. 2

[34] Wei Peng, Yu-Hong Dai, Hui Zhang, and Lizhi Cheng. Training gans with centripetal acceleration. *Optimization Methods and Software*, pages 1–19, 2020. 2

[35] 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. 2

[36] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019, 2013. 5, 6

[37] Aymeric Rousseau, Phil Sharer, Sylvain Pagerit, and Sujit Das. Trade-off between fuel economy and cost for advanced vehicle configurations. In *20th International Electric Vehicle Symposium (EVS20), Monaco*, 2005. 5

[38] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016. 2

[39] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015. 9

[40] Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010. 7

[41] Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen Boyd, and Peter W Glynn. Stochastic mirror descent in variationally coherent optimization problems. In *Advances in Neural Information Processing Systems*, pages 7040–7049, 2017. 2, 5

[42] Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen P Boyd, and Peter W Glynn. On the convergence of mirror descent beyond stochastic convex programming. *SIAM Journal on Optimization*, 30(1):687–716, 2020. 2, 5

[43] Zhengyuan Zhou, Panayotis Mertikopoulos, A Moustakas, Nicholas Bambos, and Peter Glynn. Robust power management via learning and game design. *Operations Research*, 2020. 9

[44] Zhengyuan Zhou, Panayotis Mertikopoulos, Aris L Moustakas, Nicholas Bambos, and Peter Glynn. Mirror descent learning in continuous games. In *2017 IEEE 56th Conference on Decision and Control (CDC)*, pages 5776–5783. IEEE, 2017. 5

A Convergence Analysis of Optimistic Dual Extrapolation

Based on the optimality condition of $\boldsymbol{z}_k$ and Assumption 2, we have Lemma 2.

Lemma 2 *In Algorithm 1, let*

$$E_{1k} := a_k \Big\langle F(\boldsymbol{w}_k) + \frac{L}{\alpha\gamma} \nabla_{\boldsymbol{w}_k} \frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{z}_k \Big\rangle \\ - \frac{L}{\alpha\gamma}\Big(\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{\gamma}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2\Big), \tag{18}$$

then $\forall \boldsymbol{u}, \boldsymbol{w}_0 \in \mathcal{W}$, *we have*

$$\sum_{k=1}^{K} a_k \left(\langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{u} \rangle - \frac{\sigma}{\gamma} V_{\boldsymbol{w}_k - \boldsymbol{w}_0}(\boldsymbol{u} - \boldsymbol{w}_0) \right) \le \sum_{k=1}^{K} E_{1k} + \frac{1}{2\gamma}\|\boldsymbol{u} - \boldsymbol{w}_0\|^2. \tag{19}$$

Proof. See Section C.1. ■

In Lemma 2, the sequence $\{E_{1k}\}$ can be viewed as the errors we need to bound in each iteration. The upper bound of the sum of $\{E_{1k}\}$ is given in Lemma 3 below.

Lemma 3 *In Algorithm 1,* $\forall k \in [K]$, *we have*

$$\sum_{k=1}^{K} E_{1k} \le -\frac{L}{8\alpha} \sum_{k=1}^{K} a_{k-1}\big(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2\big),$$

where we define $a_0 := a_1$ *for convenience.*

Proof. See Section C.2. ■

By Lemma 3, $\forall\ 0 < \alpha \le \min\left\{\frac{1}{4\sqrt{2}}, \frac{\sqrt{3}}{4\sqrt{\gamma}}\right\}$ and $k \in [K]$, $\sum_{k=1}^{K} E_{1k}$ is upper bounded by the sum of strictly negative terms about $\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2$, which makes it possible to give a upper bound about $\min_{k\in[K]}(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|)$. To show the guarantees by restricted strong merit function and the distance $\|\boldsymbol{w}_k - \boldsymbol{w}^*\|$, we give Lemma 4.

Lemma 4 *In Algorithm 1,* $\forall k \in [K]$, *we have,*

$$\sup_{\boldsymbol{w}\in\mathcal{W}, \|\boldsymbol{w}_k - \boldsymbol{w}\| \le D} \langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{w} \rangle \le \big(1 + \frac{\delta}{\alpha\gamma}\big) DL(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{z}_{k-1} - \boldsymbol{w}_{k-1}\|). \tag{20}$$

If $\sigma > 0$, *then we have*

$$\|\boldsymbol{w}_k - \boldsymbol{w}^*\| \le \big(1 + \frac{\delta}{\alpha\gamma}\big)\frac{L}{\sigma}(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{z}_{k-1} - \boldsymbol{w}_{k-1}\|). \tag{21}$$

Proof. See Section C.3. ■

Then combining Lemmas 2, 3 and 4, we obtain Theorem 1 in main body (see Section C.4 for the proof.).

B Convergence Analysis of Stochastic Optimistic Dual Extrapolation

We can extend the proof for the OptDE method in Section A to the stochastic setting for Lemmas 5, 6 and 7 and then obtain Theorems 2. First, we extend Lemma 2 into Lemma 5.

Lemma 5 *In Algorithm 5,* $\forall k \in [K]$, *we have the following inequality:* $\forall \boldsymbol{u}, \boldsymbol{w}_0 \in \mathcal{W}$, *let*

$$E_{2k} := a_k \Big\langle F(\boldsymbol{w}_k; \xi_k) + \frac{L^2 a_k}{(\alpha\gamma)^2} \nabla_{\boldsymbol{w}_k} \frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{z}_k \Big\rangle \\ - \frac{L^2 a_k}{(\alpha\gamma)^2}\Big(\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{\gamma}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2\Big) + a_k\langle F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k; \xi_k), \boldsymbol{w}_k - \boldsymbol{u} \rangle, \tag{22}$$

then we have

$$\mathbb{E}\Big[\sum_{k=1}^{K} a_k\Big(\langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{u}\rangle - \frac{\sigma}{\gamma} V_{\boldsymbol{w}_k - \boldsymbol{w}_0}(\boldsymbol{u} - \boldsymbol{w}_0)\Big)\Big] \le \mathbb{E}\Big[\sum_{k=1}^{K} E_{2k}\Big] + \frac{1}{2\gamma}\|\boldsymbol{u} - \boldsymbol{w}_0\|^2. \quad (23)$$

Proof. See Section D.1. ■

Compared with the E_{1k} of Lemma 2, E_{2k} contains an extra term $a_k\langle F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k;\xi_k), \boldsymbol{w}_k - \boldsymbol{u}\rangle$. Then based on the definition of E_{2k} and Assumption 5, we have Lemma 6.

Lemma 6 *In Algorithm 5, $\forall k \in [K]$ and $\forall \boldsymbol{u} \in \mathcal{W}$, we have*

$$\mathbb{E}\Big[\sum_{k=1}^{K} E_{2k}\Big] \le -\mathbb{E}\Big[4\alpha \sum_{k=1}^{K}(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2)\Big] + \frac{8\alpha s^2 K}{L^2}. \quad (24)$$

Proof. See Section D.2. ■

Lemma 6 extends Lemma 3 into the stochastic setting. Meanwhile, by the optimality condition of $\boldsymbol{w}_k$, and Assumptions 1, 2 and 5, we can extend Lemma 4 to Lemma 7.

Lemma 7 *In Algorithm 5, for the setting $\sigma = 0$ and $\forall k \in [K]$, we have,*

$$\begin{aligned}&\mathbb{E}_{\xi_{k-1}}\Big[\sup_{\boldsymbol{w}\in\mathcal{W}, \|\boldsymbol{w}_k - \boldsymbol{w}\| \le D} \langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{w}\rangle\Big] \\ \le & \big(1 + \frac{\delta}{\alpha\gamma}\big) LD\,\mathbb{E}_{\xi_{k-1}}[(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|)] + \frac{L^2}{2}\mathbb{E}_{\xi_{k-1}}[\|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\|^2] + \frac{s^2}{2L^2}.\end{aligned}$$

Proof. See Section D.3. ■

Then combining Lemmas 5, 6 and 7, we obtain Theorem 2 for the SOptDE method in the main body (see Section D.4 for the proof).

B.1 The $O\left(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon}\right)$ rate in terms of restricted strong merit function under Assumption 4

It turns out that with the conservative setting $a_k = \frac{\alpha\gamma\sqrt{1+\sigma A_{k-1}}}{L}$, we can not obtain strong convergence results in terms of restricted strong merit function. To obtain the rate $O\left(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon}\right)$, we need adopt the more aggressive setting $a_k = \frac{\alpha\gamma(1+\sigma A_{k-1})}{L}$ with a large batch size strategy, which is given in Algorithm 3. With this setting, we have Proposition 2.

Algorithm 3 Stochastic Optimistic Dual Extrapolation (**Version 2**)

```
Input: Lipshitz constant $L > 0$ from Assumption 1, $\gamma, \delta > 0$ from Assumption 2. The
  VIP$(F, \mathcal{W})$ satisfying Assumption 3 ($\sigma = 0$) or Assumption 4 ($\sigma > 0$).
$A_0 = 0, 0 < \alpha \le \min\left\{\frac{1}{8}, \frac{\sqrt{3}}{4\sqrt{2}\gamma}\right\}$.
$\boldsymbol{w}_0 = \boldsymbol{z}_0 \in \mathcal{W}, \boldsymbol{g}_0 = \boldsymbol{0}$.
for $k = 1, 2, 3, \ldots, K$ do
    $a_k = \frac{\alpha\gamma(1+\sigma A_{k-1})}{L}, A_k = A_{k-1} + a_k$.
    $\boldsymbol{w}_k = \arg\min_{\boldsymbol{w}\in\mathcal{W}}\left\{\langle F(\boldsymbol{w}_{k-1};\xi_{k-1}), \boldsymbol{w}\rangle + \frac{L}{2\alpha\gamma}\|\boldsymbol{w} - \boldsymbol{z}_{k-1}\|^2\right\}$.
    $\boldsymbol{g}_k = \boldsymbol{g}_{k-1} + a_k\big(F(\boldsymbol{w}_k;\xi_k) - \frac{\sigma}{\gamma}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{w}_0\|^2\big)$.
    $\boldsymbol{z}_k = \arg\min_{\boldsymbol{z}\in\mathcal{W}}\left\{\langle \boldsymbol{g}_k, \boldsymbol{z}\rangle + \frac{1+\sigma A_k}{2\gamma}\|\boldsymbol{z} - \boldsymbol{w}_0\|^2\right\}$
end for
$\tilde{\boldsymbol{w}}_K = \boldsymbol{w}_k$, where $k$ is chosen at random with probability distribution $\{\frac{a_1}{A_K}, \frac{a_2}{A_K}, \ldots, \frac{a_K}{A_K}\}$.
return $\tilde{\boldsymbol{w}}_K$.
```

Proposition 2 *Let Assumptions 1, 2, 4 and 5 hold. After K iterations, Algorithm 3 returns a $\tilde{\boldsymbol{w}}_K$ such that*

$$\begin{aligned}&\mathbb{E}\Big[\sup_{\boldsymbol{w}\in\mathcal{W},\|\tilde{\boldsymbol{w}}_K-\boldsymbol{w}\|\le D}\langle F(\tilde{\boldsymbol{w}}_K),\tilde{\boldsymbol{w}}_K-\boldsymbol{w}\rangle\Big]\\&\le C_1 LD\sqrt{\frac{\|\boldsymbol{w}^*-\boldsymbol{w}_0\|^2}{L(A_{K-1}+a_1)}+\frac{s^2}{L^2}}+L^2\Big(\frac{\|\boldsymbol{w}^*-\boldsymbol{w}_0\|^2}{L(A_{K-1}+a_1)}+\frac{s^2}{L^2}\Big)+\frac{s^2}{2L^2}.\end{aligned}\tag{25}$$

with $C_1 := 4\big(1+\frac{\delta}{\alpha\gamma}\big)\sqrt{\frac{\alpha}{\gamma}}$, $a_1=\frac{\alpha\gamma}{L}$, and

$$A_{K-1}=\frac{1}{\sigma}\Big(1+\frac{\alpha\gamma\sigma}{L}\Big)^{K-1}-\frac{1}{\sigma}.\tag{26}$$

The proof of Proposition 2 follows the same pipeline of proving Theorem 2, except that we use the setting $a_k=\frac{\alpha\gamma(1+\sigma A_{k-1})}{L}$ that is also used in Algorithm 4. We leave the proof of Proposition 2 as a simple exercise.

In Proposition 2, if we hope the variance of the stochastic estimation $\{F(\boldsymbol{w}_k;\xi_k)\}$ as $s^2 = O(\frac{1}{A_{K-1}+a_1})$, then we need $O(A_{K-1}+a_1)$ stochastic samples per iteration. Meanwhile, to attain an expected ϵ-accurate strong solution, we will need $O(\log\frac{1}{\epsilon})$ number of iterations. Thus the total number of stochastic samples we need is $O(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon})$.

B.2 The “ (quadratic) natural residual function” [18] and restricted strong merit function

In our notation, for any $\boldsymbol{w}\in\mathcal{W}$, the (quadratic) natural residual function in [18] is defined by: given $\eta>0$,

$$r_\eta(\boldsymbol{w})=\|\boldsymbol{w}-P_{\boldsymbol{w}}(\eta F(\boldsymbol{w}))\|^2,\tag{27}$$

which can be used to derive the restricted strong merit function as Proposition 3.

Proposition 3 *Let $\boldsymbol{w}' := P_{\boldsymbol{w}}(\eta F(\boldsymbol{w})) = \arg\min_{\boldsymbol{z}\in\mathcal{W}}\{\langle\eta F(\boldsymbol{w}),\boldsymbol{z}\rangle+\frac{1}{2\gamma}\|\boldsymbol{z}-\boldsymbol{w}\|^2\}$. Then we have*

$$\sup_{\boldsymbol{z}\in\mathcal{W},\|\boldsymbol{w}'-\boldsymbol{z}\|\le D}\langle F(\boldsymbol{w}'),\boldsymbol{w}'-\boldsymbol{z}\rangle\le\Big(L+\frac{\delta}{\eta\gamma}\Big)D\sqrt{r_\eta(\boldsymbol{w})}.\tag{28}$$

Proof. It follows that

$$\begin{aligned}\langle F(\boldsymbol{w}'),\boldsymbol{w}'-\boldsymbol{z}\rangle &= \langle F(\boldsymbol{w}')-\big(F(\boldsymbol{w})+\frac{1}{\eta}\nabla_{\boldsymbol{w}'}\frac{1}{2\gamma}\|\boldsymbol{w}'-\boldsymbol{w}\|^2\big),\boldsymbol{w}'-\boldsymbol{z}\rangle\\&\quad+\Big\langle F(\boldsymbol{w})+\frac{1}{\eta}\nabla_{\boldsymbol{w}'}\frac{1}{2\gamma}\|\boldsymbol{w}'-\boldsymbol{w}\|^2,\boldsymbol{w}'-\boldsymbol{z}\Big\rangle\\&\overset{(a)}{\le} \langle F(\boldsymbol{w}')-\big(F(\boldsymbol{w})+\frac{1}{\eta}\nabla_{\boldsymbol{w}'}\frac{1}{2\gamma}\|\boldsymbol{w}'-\boldsymbol{w}\|^2\big),\boldsymbol{w}'-\boldsymbol{z}\rangle\\&\overset{(b)}{\le} \|F(\boldsymbol{w}')-F(\boldsymbol{w})\|_*\|\boldsymbol{w}'-\boldsymbol{z}\|+\frac{1}{\eta}\Big\|\nabla_{\boldsymbol{w}'}\frac{1}{2\gamma}\|\boldsymbol{w}'-\boldsymbol{w}\|^2\Big\|_*\|\boldsymbol{w}'-\boldsymbol{z}\|\\&\overset{(c)}{\le} L\|\boldsymbol{w}'-\boldsymbol{w}\|\|\boldsymbol{w}'-\boldsymbol{z}\|+\frac{1}{\eta\gamma}\delta\|\boldsymbol{w}'-\boldsymbol{w}\|\|\boldsymbol{w}'-\boldsymbol{z}\|\\&= \Big(L+\frac{\delta}{\eta\gamma}\Big)\|\boldsymbol{w}'-\boldsymbol{w}\|\|\boldsymbol{w}'-\boldsymbol{z}\|.\end{aligned}\tag{29}$$

where (a) is by the optimality condition of $\boldsymbol{w}'$, (b) is by the Cauchy-Schwarz inequality, (c) is by the Lipschitz continuity of $F(\mathbf{w})$ and the bounded assumption (7). So we have

$$\sup_{\boldsymbol{z}\in\mathcal{W},\|\boldsymbol{w}'-\boldsymbol{z}\|\le D}\langle F(\boldsymbol{w}'),\boldsymbol{w}'-\boldsymbol{z}\rangle\le\Big(L+\frac{\delta}{\eta\gamma}\Big)D\|\boldsymbol{w}'-\boldsymbol{w}\|=\Big(L+\frac{\delta}{\eta\gamma}\Big)D\sqrt{r_\eta(\boldsymbol{w})}.\tag{30}$$

■

C Proof of Section A

By the definition of proximal operator (8), we can equivalently reformulate the optimistic dual extrapolation (OptDE) algorithm in the main body as Algorithm 4. Then based on the definition of $\boldsymbol{g}_k$ in Step 7 and the definition of the Bregman divergence $V_{\boldsymbol{w}}(\boldsymbol{u})(\boldsymbol{w}, \boldsymbol{u} \in \mathcal{W})$, we can verify that

$$\boldsymbol{z}_k = \arg\min_{\boldsymbol{z}\in\mathcal{W}} \left\{ \psi_k(\boldsymbol{z}) := \sum_{i=1}^{k} a_i \Big(\langle F(\boldsymbol{w}_i), \boldsymbol{z}-\boldsymbol{u}\rangle + \frac{\sigma}{\gamma} V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{z} - \boldsymbol{w}_0) \Big) + \frac{1}{2\gamma}\|\boldsymbol{z}-\boldsymbol{w}_0\|^2 \right\}, \quad (31)$$

where $\boldsymbol{u}$ is an arbitrary vector in $\mathcal{W}$ and is irrelevant to the minimizer $\boldsymbol{z}_k$. In our context, $\psi_k(\boldsymbol{z})$ plays the role of a "generalized estimation sequence" to help us conduct convergence analysis. By the γ-strong convexity of the Bregman divergence $V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{z} - \boldsymbol{w}_0)$, we know that $\psi_k(\boldsymbol{z})$ is strongly convex with strong convexity parameter $1 + \sigma \sum_{i=1}^{k} a_i = 1 + \sigma A_k$.

Algorithm 4 Optimistic Dual Extrapolation **(Reformulation)**

```
 1: **Input:** Lipschitz constant $L > 0$ from Assumption 1, $\gamma, \delta > 0$ from Assumption 2. The
    VIP$(F, \mathcal{W})$ satisfying Assumption 3 ($\sigma = 0$) or Assumption 4 ($\sigma > 0$).
 2: $A_0 = 0$, $0 < \alpha \le \min\left\{\frac{1}{4\sqrt{2}}, \frac{\sqrt{3}}{4\sqrt{\gamma}}\right\}$.
 3: $\boldsymbol{w}_0 = \boldsymbol{z}_0 \in \mathcal{W}$.
 4: for $k = 1, 2, 3, \ldots, K$ do
 5:     $a_k = \frac{\alpha\gamma(1+\sigma A_{k-1})}{L}$, $A_k = A_{k-1} + a_k$.
 6:     $\boldsymbol{w}_k = \arg\min_{\boldsymbol{w}\in\mathcal{W}} \left\{\langle F(\boldsymbol{w}_{k-1}), \boldsymbol{w}\rangle + \frac{L}{2\alpha\gamma}\|\boldsymbol{w} - \boldsymbol{z}_{k-1}\|^2\right\}$.
 7:     $\boldsymbol{g}_k = \boldsymbol{g}_{k-1} + a_k\big(F(\boldsymbol{w}_k) - \frac{\sigma}{\gamma}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{w}_0\|^2\big)$.
 8:     $\boldsymbol{z}_k = \arg\min_{\boldsymbol{z}\in\mathcal{W}} \left\{\langle \boldsymbol{g}_k, \boldsymbol{z}\rangle + \frac{1+\sigma A_k}{2\gamma}\|\boldsymbol{z} - \boldsymbol{w}_0\|^2\right\}$.
 9: end for
10: $\tilde{\boldsymbol{w}}_K = \arg\min_{\boldsymbol{w}_k : k\in[K]}(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|)$.
11: return $\tilde{\boldsymbol{w}}_K$.
```

C.1 Proof of Lemma 2

Proof. Given the definition of the generalized estimation sequence $\psi_k(\boldsymbol{z})$ in (31) and the minimizer $\boldsymbol{z}_k$ in Algorithm 4, by the optimality condition of $\boldsymbol{z}_k$, we have: $\forall \boldsymbol{u} \in \mathcal{W}$,

$$\left\langle \sum_{i=1}^{k} a_i \Big(F(\boldsymbol{w}_i) + \frac{\sigma}{\gamma}\nabla V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{z}_k - \boldsymbol{w}_0)\Big) + \nabla_{\boldsymbol{z}_k}\frac{1}{2\gamma}\|\boldsymbol{z}_k - \boldsymbol{w}_0\|^2, \boldsymbol{u} - \boldsymbol{z}_k \right\rangle \ge 0. \quad (32)$$

Then we have $\forall k \in [K]$,

$$\begin{aligned}
\psi_k(\boldsymbol{z}_k) &= \sum_{i=1}^{k} a_i \Big(\langle F(\boldsymbol{w}_i), \boldsymbol{z}_k - \boldsymbol{u}\rangle + \frac{\sigma}{\gamma} V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{z}_k - \boldsymbol{w}_0) \Big) + \frac{1}{2\gamma}\|\boldsymbol{z}_k - \boldsymbol{w}_0\|^2 \\
&\overset{(a)}{\le} \frac{\sigma}{\gamma}\sum_{i=1}^{k} a_i \left(\langle \nabla V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{z}_k - \boldsymbol{w}_0), \boldsymbol{u} - \boldsymbol{z}_k\rangle + V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{z}_k - \boldsymbol{w}_0) \right) \\
&\quad + \left\langle \nabla_{\boldsymbol{z}_k}\frac{1}{2\gamma}\|\boldsymbol{z}_k - \boldsymbol{w}_0\|^2, \boldsymbol{u} - \boldsymbol{z}_k \right\rangle + \frac{1}{2\gamma}\|\boldsymbol{z}_k - \boldsymbol{w}_0\|^2 \\
&\overset{(b)}{\le} \frac{\sigma}{\gamma}\sum_{i=1}^{k} a_i V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{u} - \boldsymbol{w}_0) + \frac{1}{2\gamma}\|\boldsymbol{u} - \boldsymbol{w}_0\|^2, \qquad (33)
\end{aligned}$$

where (a) is by the optimality condition (32), and (b) is by the convexity of both $V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{u} - \boldsymbol{w}_0)$ and $\frac{1}{2\gamma}\|\boldsymbol{u} - \boldsymbol{w}_0\|^2$.

Meanwhile for $k \geq 1$, by the definition of $\psi_k(\boldsymbol{z}_k)$, we have

$$\begin{aligned}
\psi_k(\boldsymbol{z}_k) &= \psi_{k-1}(\boldsymbol{z}_k) + a_k\langle F(\boldsymbol{w}_k), \boldsymbol{z}_k - \boldsymbol{u}\rangle \\
&\overset{(a)}{\geq} \psi_{k-1}(\boldsymbol{z}_{k-1}) + \frac{1+\sigma A_{k-1}}{2}\|\boldsymbol{z}_k - \boldsymbol{z}_{k-1}\|^2 + a_k\langle F(\boldsymbol{w}_k), \boldsymbol{z}_k - \boldsymbol{u}\rangle \\
&= \psi_{k-1}(\boldsymbol{z}_{k-1}) + \frac{1+\sigma A_{k-1}}{2}\|\boldsymbol{z}_k - \boldsymbol{z}_{k-1}\|^2 + a_k\langle F(\boldsymbol{w}_k), \boldsymbol{z}_k - \boldsymbol{w}_k\rangle \\
&\quad + a_k\langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{u}\rangle, \qquad (34)
\end{aligned}$$

where (a) is by the $(1+\sigma A_{k-1})$-strong convexity of $\psi_{k-1}(\boldsymbol{z})$. Meanwhile, by the strong convexity of $\frac{1}{2}\|\cdot\|^2$ in Assumption 2, we have

$$\begin{aligned}
& a_k\left\langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{z}_k\right\rangle - \frac{1+\sigma A_{k-1}}{2}\|\boldsymbol{z}_k - \boldsymbol{z}_{k-1}\|^2 \\
&\leq \Big\langle a_k F(\boldsymbol{w}_k) + (1+\sigma A_{k-1})\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{z}_k\Big\rangle \\
&\quad - (1+\sigma A_{k-1})\Big(\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{\gamma}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2\Big). \qquad (35)
\end{aligned}$$

Then combining (34) and (35), and after simple arrangements, we have

$$\begin{aligned}
a_k\langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{u}\rangle &\leq \psi_k(\boldsymbol{z}_k) - \psi_{k-1}(\boldsymbol{z}_{k-1}) \\
&\quad + \Big\langle a_k F(\boldsymbol{w}_k) + (1+\sigma A_{k-1})\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{z}_k\Big\rangle \\
&\quad - (1+\sigma A_{k-1})\Big(\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{\gamma}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2\Big). \qquad (36)
\end{aligned}$$

Summing (36) from $k=1$ to K, we have

$$\begin{aligned}
& \sum_{k=1}^K a_k\langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{u}\rangle \\
\leq\; & \sum_{k=1}^K \Big(\Big\langle a_k F(\boldsymbol{w}_k) + (1+\sigma A_{k-1})\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{z}_k\Big\rangle \\
& - (1+\sigma A_{k-1})\Big(\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{\gamma}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2\Big)\Big) + \psi_K(\boldsymbol{z}_K) - \psi_0(\boldsymbol{z}_0) \\
\overset{(a)}{\leq}\; & \sum_{k=1}^K \Big(\Big\langle a_k F(\boldsymbol{w}_k) + (1+\sigma A_{k-1})\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{z}_k\Big\rangle \\
& \quad - (1+\sigma A_{k-1})\Big(\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{\gamma}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2\Big)\Big) \\
& + \frac{\sigma}{\gamma}\sum_{k=1}^K a_k V_{\boldsymbol{w}_k - \boldsymbol{w}_0}(\boldsymbol{u} - \boldsymbol{w}_0) + \frac{1}{2\gamma}\|\boldsymbol{u} - \boldsymbol{w}_0\|^2, \qquad (37)
\end{aligned}$$

where (a) is by the fact $\psi_0(\boldsymbol{z}_0) = 0$ and the upper bound of $\psi_K(\boldsymbol{z}_K)$ by (33). By the setting $a_k = \frac{\alpha\gamma(1+\sigma A_{k-1})}{L}$ in Algorithm 4 and (37), we have

$$\begin{aligned}
& \sum_{k=1}^K a_k\langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{u}\rangle \\
\leq\; & \sum_{k=1}^K a_k\Bigg(\Big\langle F(\boldsymbol{w}_k) + \frac{L}{\alpha\gamma}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{z}_k\Big\rangle \\
& \qquad - \frac{L}{\alpha\gamma}\Big(\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{\gamma}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2\Big)\Bigg) \\
& + \frac{\sigma}{\gamma}\sum_{k=1}^K a_k V_{\boldsymbol{w}_k - \boldsymbol{w}_0}(\boldsymbol{u} - \boldsymbol{w}_0) + \frac{1}{2\gamma}\|\boldsymbol{u} - \boldsymbol{w}_0\|^2. \qquad (38)
\end{aligned}$$

Then based on the definition of E_{1k} in Lemma 2, after simple arrangements, Lemma 2 is proved.

■

C.2 Proof of Lemma 3

Proof. By the definition of E_{1k} in Lemma 2, we have: $\forall k \in [K]$,

$$\begin{aligned}
E_{1k} &= a_k \Big(\Big\langle F(\boldsymbol{w}_k) + \frac{L}{\alpha\gamma} \nabla_{\boldsymbol{w}_k} \frac{1}{2} \|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{z}_k \Big\rangle \\
&\quad - \frac{L}{\alpha\gamma} \Big(\frac{1}{2} \|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{\gamma}{2} \|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2 \Big) \Big) \\
&\leq a_k \Big(\Big\langle F(\boldsymbol{w}_k) - F(\boldsymbol{w}_{k-1}), \boldsymbol{w}_k - \boldsymbol{z}_k \Big\rangle \\
&\quad + \Big\langle F(\boldsymbol{w}_{k-1}) + \frac{L}{\alpha\gamma} \nabla_{\boldsymbol{w}_k} \frac{1}{2} \|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{z}_k \Big\rangle \\
&\quad - \frac{L}{\alpha\gamma} \Big(\frac{1}{2} \|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{\gamma}{2} \|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2 \Big) \Big).
\end{aligned} \tag{39}$$

Meanwhile, we have for all $\alpha > 0$,

$$\begin{aligned}
&\Big\langle F(\boldsymbol{w}_k) - F(\boldsymbol{w}_{k-1}), \boldsymbol{w}_k - \boldsymbol{z}_k \Big\rangle \\
\overset{(a)}{\leq}\;& \|F(\boldsymbol{w}_k) - F(\boldsymbol{w}_{k-1})\|_* \|\boldsymbol{w}_k - \boldsymbol{z}_k\| \\
\overset{(b)}{\leq}\;& L\|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\| \|\boldsymbol{w}_k - \boldsymbol{z}_k\| \\
\overset{(c)}{\leq}\;& L\alpha\|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\|^2 + \frac{L}{4\alpha}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2 \\
\overset{(d)}{\leq}\;& L\alpha(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{z}_{k-1} - \boldsymbol{w}_{k-1}\|)^2 + \frac{L}{4\alpha}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2 \\
\overset{(e)}{\leq}\;& 2L\alpha(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \|\boldsymbol{z}_{k-1} - \boldsymbol{w}_{k-1}\|^2) + \frac{L}{4\alpha}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2,
\end{aligned} \tag{40}$$

where (a) is by the Cauchy–Schwarz inequality, (b) is the Lipschitz continuous Assumption 1, (c) is by the fact $ab \leq a^2 + \frac{b^2}{4}$, (d) is by the triangle inequality of norm $\|\cdot\|$ and (e) is by the fact $(a+b)^2 \leq 2(a^2+b^2)$.

Then by the optimality condition of $\boldsymbol{w}_k$ in the k-th iteration of Algorithm 4, we have: $\forall \boldsymbol{z} \in \mathcal{W}$,

$$\Big\langle F(\boldsymbol{w}_{k-1}) + \frac{L}{\alpha\gamma} \nabla_{\boldsymbol{w}_k} \frac{1}{2} \|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{z} \Big\rangle \leq 0. \tag{41}$$

By combining (39), (40) and (41), we have

$$\begin{aligned}
E_{1k} &\leq a_k \Big(-\frac{L}{2\alpha\gamma}(1 - 4\alpha^2\gamma)\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + 2L\alpha\|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2 - \frac{L}{4\alpha}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2 \Big) \\
&\overset{(a)}{\leq} a_k \Big(-\frac{L}{8\alpha\gamma}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{L}{8\alpha}\frac{a_{k-1}}{a_k}\|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2 - \frac{L}{4\alpha}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2 \Big) \\
&\leq -\frac{La_k}{8\alpha\gamma}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{L}{8\alpha}(a_{k-1}\|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2 - 2a_k\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2) \\
&\overset{(b)}{\leq} -\frac{La_k}{8\alpha}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{L}{8\alpha}(a_{k-1}\|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2 - 2a_k\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2),
\end{aligned} \tag{42}$$

where (a) is by the fact

$$a_k = \frac{\alpha\gamma}{L}\Big(1 + \frac{\alpha\gamma\sigma}{L}\Big)^{K-1} \tag{43}$$

from Step 5 of Algorithm 4 and the setting

$$\alpha \leq \min\left\{\frac{1}{4\sqrt{2}}, \frac{\sqrt{3}}{4\sqrt{\gamma}}\right\} \leq \frac{1}{4}\sqrt{\frac{L}{L + \alpha\gamma\sigma}} = \frac{1}{4}\sqrt{\frac{a_{k-1}}{a_k}},$$

and (b) is by the setting that $0 < \gamma \le 1$.

With the $\boldsymbol{w}_0 = \boldsymbol{z}_0$, for convenience, we set $a_0 := a_1$. By summing (42) from $k = 1$ to K, we have

$$\begin{aligned}
\sum_{k=1}^{K} E_{1k} \quad &\le \quad -\frac{L}{8\alpha}\Big(\sum_{k=1}^{K} a_k\big(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2\big) - a_0\|\boldsymbol{w}_0 - \boldsymbol{z}_0\|^2 + a_K\|\boldsymbol{w}_K - \boldsymbol{z}_K\|^2\Big) \\
&\overset{(a)}{=} \quad -\frac{L}{8\alpha}\Big(\sum_{k=1}^{K} \big(a_k\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + a_{k-1}\|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2\big) + 2a_K\|\boldsymbol{w}_K - \boldsymbol{z}_K\|^2\Big) \\
&\overset{(b)}{\le} \quad -\frac{L}{8\alpha}\sum_{k=1}^{K} a_{k-1}\big(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2\big), \qquad (44)
\end{aligned}$$

where (a) is by the fact $\boldsymbol{w}_0 = \boldsymbol{z}_0$, and (b) is by the fact that $a_k \ge a_{k-1} > 0$. Lemma 3 is proved. ■

C.3 Proof of Lemma 4

Proof. It follows that

$$\begin{aligned}
& \langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{w}\rangle \\
= \quad & \langle F(\boldsymbol{w}_k) - \big(F(\boldsymbol{w}_{k-1}) + \frac{L}{\alpha\gamma}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2\big), \boldsymbol{w}_k - \boldsymbol{w}\rangle \\
& + \Big\langle F(\boldsymbol{w}_{k-1}) + \frac{L}{\alpha\gamma}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{w}\Big\rangle \\
\overset{(a)}{\le} \quad & \langle F(\boldsymbol{w}_k) - \big(F(\boldsymbol{w}_{k-1}) + \frac{L}{\alpha\gamma}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2\big), \boldsymbol{w}_k - \boldsymbol{w}\rangle \\
\overset{(b)}{\le} \quad & \|F(\boldsymbol{w}_k) - F(\boldsymbol{w}_{k-1})\|_*\|\boldsymbol{w}_k - \boldsymbol{w}\| + \frac{L}{\alpha\gamma}\Big\|\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2\Big\|_*\|\boldsymbol{w}_k - \boldsymbol{w}\| \\
\overset{(c)}{\le} \quad & L\|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\|\|\boldsymbol{w}_k - \boldsymbol{w}\| + \frac{L}{\alpha\gamma}\delta\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|\|\boldsymbol{w}_k - \boldsymbol{w}\| \\
\overset{(d)}{\le} \quad & L(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{z}_{k-1} - \boldsymbol{w}_{k-1}\|)\|\boldsymbol{w}_k - \boldsymbol{w}\| + \frac{L}{\alpha\gamma}\delta\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|\|\boldsymbol{w}_k - \boldsymbol{w}\| \\
\le \quad & \Big(\big(1 + \frac{\delta}{\alpha\gamma}\big)(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{z}_{k-1} - \boldsymbol{w}_{k-1}\|)\Big)L\|\boldsymbol{w}_k - \boldsymbol{w}\|, \qquad (45)
\end{aligned}$$

where (a) is by the optimality condition of $\boldsymbol{w}_k$, (b) is by the Cauchy-Schwarz inequality, (c) is by the Lipschitz continuity of $F(\mathbf{w})$ and the bounded assumption (7), (d) is by the triangle inequality of norm $\|\cdot\|$. So we have

$$\sup_{\boldsymbol{w}\in\mathcal{W}, \|\boldsymbol{w}_k - \boldsymbol{w}\| \le D} \langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{w}\rangle \le \big(1 + \frac{\delta}{\alpha\gamma}\big)DL(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{z}_{k-1} - \boldsymbol{w}_{k-1}\|). \qquad (46)$$

Meanwhile, if there exists a $\boldsymbol{w}^*$ that satisfies Assumption 4, *i.e.*, $\forall \boldsymbol{w} \in \mathcal{W}$, $\langle F(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^*\rangle \ge \frac{\sigma}{\gamma}(V_{\boldsymbol{w}-\boldsymbol{w}_0}(\boldsymbol{w}^* - \boldsymbol{w}_0) + V_{\boldsymbol{w}^*-\boldsymbol{w}_0}(\boldsymbol{w} - \boldsymbol{w}_0))$ with $\sigma > 0$, then in (45), let $\boldsymbol{w} := \boldsymbol{w}^*$, and by the fact $V_{\boldsymbol{w}_k-\boldsymbol{w}_0}(\boldsymbol{w}^* - \boldsymbol{w}_0) \ge \frac{\gamma}{2}\|\boldsymbol{w}_k - \boldsymbol{w}^*\|^2$ and $V_{\boldsymbol{w}^*-\boldsymbol{w}_0}(\boldsymbol{w}_k - \boldsymbol{w}_0) \ge \frac{\gamma}{2}\|\boldsymbol{w}_k - \boldsymbol{w}^*\|^2$, we have

$$\begin{aligned}
\sigma\|\boldsymbol{w}_k - \boldsymbol{w}^*\|^2 \quad &\le \quad \frac{\sigma}{\gamma}(V_{\boldsymbol{w}_k-\boldsymbol{w}_0}(\boldsymbol{w}^* - \boldsymbol{w}_0) + V_{\boldsymbol{w}^*-\boldsymbol{w}_0}(\boldsymbol{w}_k - \boldsymbol{w}_0)) \le \langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{w}^*\rangle \\
&\le \quad \big(1 + \frac{\delta}{\alpha\gamma}\big)L(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{z}_{k-1} - \boldsymbol{w}_{k-1}\|)\|\boldsymbol{w}_k - \boldsymbol{w}^*\|. \qquad (47)
\end{aligned}$$

So it follows that

$$\|\boldsymbol{w}_k - \boldsymbol{w}^*\| \le \big(1 + \frac{\delta}{\alpha\gamma}\big)\frac{L}{\sigma}(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{z}_{k-1} - \boldsymbol{w}_{k-1}\|). \qquad (48)$$

Lemma 4 is proved. ■

C.4 Proof of Theorem 1

Proof. Firstly, by the setting $a_k = \frac{\alpha\gamma(1+\sigma A_{k-1})}{L}$ and $A_0 = 0, A_k = A_{k-1} + a_k$, we have: $\forall k \geq 0$,

- If $\sigma = 0$, then $A_k = \frac{\alpha\gamma k}{L}$.
- If $\sigma > 0$, then $A_k = \frac{1}{\sigma}\left(1 + \frac{\alpha\gamma\sigma}{L}\right)^k - \frac{1}{\sigma}$.

By Lemmas 2 and 3, we have

$$
\begin{aligned}
&\sum_{k=1}^{K} a_k \left(\langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{u} \rangle - \frac{\sigma}{\gamma} V_{\boldsymbol{w}_k - \boldsymbol{w}_0}(\boldsymbol{u} - \boldsymbol{w}_0) \right) \\
\leq &\sum_{k=1}^{K} E_{1k} + \frac{1}{2\gamma}\|\boldsymbol{u} - \boldsymbol{w}_0\|^2 \\
\leq &- \frac{L}{8\alpha} \sum_{k=1}^{K} a_{k-1}\big(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2\big) + \frac{1}{2\gamma}\|\boldsymbol{u} - \boldsymbol{w}_0\|^2. \qquad (49)
\end{aligned}
$$

Let $\boldsymbol{w}$ be the $\boldsymbol{w}^*$ in Assumption 3 if $\sigma = 0$ or the $\boldsymbol{w}^*$ in Assumption 4 if $\sigma > 0$. Then by the property of $\boldsymbol{w}^*$, we have $\langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{w}^* \rangle - \frac{\sigma}{\gamma} V_{\boldsymbol{w}_k - \boldsymbol{w}_0}(\boldsymbol{w}^* - \boldsymbol{w}_0) \geq 0$. So by (49), it follows that

$$
\begin{aligned}
&\frac{L}{16\alpha} \sum_{k=1}^{K} a_{k-1}\big(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|\big)^2 \\
\leq \quad &\frac{L}{8\alpha} \sum_{k=1}^{K} a_{k-1}\big(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2\big) \leq \frac{1}{2\gamma}\|\boldsymbol{w}^* - \boldsymbol{w}_0\|^2. \qquad (50)
\end{aligned}
$$

By the setting $A_k = A_{k-1} + a_k$ with $A_0 = 0$ in Algorithm 4, we have $A_k = \sum_{i=1}^{k} a_i$. Meanwhile, for convenience, we have set $a_0 = a_1$. So we have

$$
\frac{L}{16\alpha}(A_{K-1} + a_1) \min_{k\in[K]} (\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|)^2 \leq \frac{1}{2\gamma}\|\boldsymbol{w}^* - \boldsymbol{w}_0\|^2. \qquad (51)
$$

So for the so computed $\{\boldsymbol{w}_k, \boldsymbol{z}_{k-1}\}$, let $\tilde{k} := \arg\min_{k\in[K]}(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|)$ and $\tilde{\boldsymbol{w}}_K := \boldsymbol{w}_{\tilde{k}}$. Then combining (50) and (51), we have

$$
\|\boldsymbol{w}_{\tilde{k}} - \boldsymbol{z}_{\tilde{k}-1}\| + \|\boldsymbol{w}_{\tilde{k}-1} - \boldsymbol{z}_{\tilde{k}-1}\| \leq \sqrt{\frac{\|\boldsymbol{w}^* - \boldsymbol{w}_0\|^2}{L(A_{K-1} + a_1)} \frac{8\alpha}{\gamma}}. \qquad (52)
$$

So by (20) of Lemma 4 and (52), it follows that

$$
\begin{aligned}
&\sup_{\boldsymbol{w}\in\mathcal{W}, \|\tilde{\boldsymbol{w}}_K - \boldsymbol{w}\| \leq D} \langle F(\tilde{\boldsymbol{w}}_K), \tilde{\boldsymbol{w}}_K - \boldsymbol{w} \rangle \\
\leq \quad &(1 + \frac{\delta}{\alpha\gamma}) DL(\|\boldsymbol{w}_{\tilde{k}} - \boldsymbol{z}_{\tilde{k}-1}\| + \|\boldsymbol{w}_{\tilde{k}-1} - \boldsymbol{z}_{\tilde{k}-1}\|) \\
\leq \quad &(1 + \frac{\delta}{\alpha\gamma}) D\sqrt{\frac{L\|\boldsymbol{w}^* - \boldsymbol{w}_0\|^2}{(A_{K-1} + a_1)} \frac{8\alpha}{\gamma}}. \qquad (53)
\end{aligned}
$$

Similarly, if $\sigma > 0$, then by (21) of Lemma 4 and (52), we have

$$
\begin{aligned}
\|\tilde{\boldsymbol{w}}_K - \boldsymbol{w}^*\| \quad &\leq \quad (1 + \frac{\delta}{\alpha\gamma}) \frac{L}{\sigma} (\|\boldsymbol{w}_{\tilde{k}} - \boldsymbol{z}_{\tilde{k}-1}\| + \|\boldsymbol{w}_{\tilde{k}-1} - \boldsymbol{z}_{\tilde{k}-1}\|) \\
&\leq \quad \left(1 + \frac{\delta}{\alpha\gamma}\right) \frac{1}{\sigma} \sqrt{\frac{L\|\boldsymbol{w}^* - \boldsymbol{w}_0\|^2}{(A_{K-1} + a_1)} \frac{8\alpha}{\gamma}}. \qquad (54)
\end{aligned}
$$

Then by defining $C_0 := \left(1 + \frac{\delta}{\alpha\gamma}\right)\sqrt{\frac{8\alpha}{\gamma}}$, Theorem 1 is proved.

■

C.5 Proof of Proposition 1

Proof. The proof follows the same paradigm of Section C.4. Firstly, by the setting $a_k = \frac{\alpha\gamma(1+\sigma A_{k-1})}{L}$ and $A_0 = 0, A_k = A_{k-1} + a_k$, we have $\forall k \geq 0$,

$$a_k = \frac{\alpha\gamma}{L}\left(1 + \frac{\alpha\gamma\sigma}{L}\right)^{k-1}. \tag{55}$$

Then the (51) of Section C.4 is replaced by

$$\frac{L}{16\alpha} a_{K-1}(\|\boldsymbol{w}_K - \boldsymbol{z}_{K-1}\| + \|\boldsymbol{w}_{K-1} - \boldsymbol{z}_{K-1}\|)^2 \leq \frac{1}{2\gamma}\|\boldsymbol{w}^* - \boldsymbol{w}_0\|^2. \tag{56}$$

Then similar to (52) to (54), we obtain the last iterate convergence result as

$$\begin{aligned}\sup_{\boldsymbol{w}\in\mathcal{W}, \|\tilde{\boldsymbol{w}}_K - \boldsymbol{w}\| \leq D} \langle F(\boldsymbol{w}_K), \boldsymbol{w}_K - \boldsymbol{w}\rangle &\leq (1 + \frac{\delta}{\alpha\gamma})D\sqrt{\frac{L\|\boldsymbol{w}^* - \boldsymbol{w}_0\|^2}{a_{K-1}}\frac{8\alpha}{\gamma}}, \\ \|\boldsymbol{w}_K - \boldsymbol{w}^*\| &\leq \left(1 + \frac{\delta}{\alpha\gamma}\right)\frac{1}{\sigma}\sqrt{\frac{L\|\boldsymbol{w}^* - \boldsymbol{w}_0\|^2}{a_{K-1}}\frac{8\alpha}{\gamma}}.\end{aligned}$$

Thus by the definition of C_0 in Theorem 1, Proposition 1 is proved.

■

C.6 Proof of Lemma 1

Proof. By the definition of the Bregman divergence $V_{\boldsymbol{w}}(\boldsymbol{v})$, we have

$$V_{\boldsymbol{v}-\boldsymbol{w}_0}(\boldsymbol{w} - \boldsymbol{w}_0) = \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2 - \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{w}_0\|^2 - \left\langle \nabla_{\boldsymbol{v}}\frac{1}{2}\|\boldsymbol{v} - \boldsymbol{w}_0\|^2, \boldsymbol{w} - \boldsymbol{v}\right\rangle \tag{57}$$

$$V_{\boldsymbol{w}-\boldsymbol{w}_0}(\boldsymbol{v} - \boldsymbol{w}_0) = \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{w}_0\|^2 - \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2 - \left\langle \nabla_{\boldsymbol{w}}\frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2, \boldsymbol{v} - \boldsymbol{w}\right\rangle. \tag{58}$$

So combining (57) and (58), it follows that

$$\left\langle \nabla_{\boldsymbol{w}}\frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2 - \nabla_{\boldsymbol{v}}\frac{1}{2}\|\boldsymbol{v} - \boldsymbol{w}_0\|^2, \boldsymbol{w} - \boldsymbol{v}\right\rangle = V_{\boldsymbol{v}-\boldsymbol{w}_0}(\boldsymbol{w} - \boldsymbol{w}_0) + V_{\boldsymbol{w}-\boldsymbol{w}_0}(\boldsymbol{v} - \boldsymbol{w}_0). \tag{59}$$

So if $F(\boldsymbol{w})$ is monotone, then we have: $\forall \boldsymbol{w}_0, \boldsymbol{w}, \boldsymbol{v} \in \mathcal{W}$,

$$\begin{aligned}&\left\langle \left(F(\boldsymbol{w}) + \epsilon\nabla_{\boldsymbol{w}}\frac{1}{2\gamma}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2\right) - \left(F(\boldsymbol{v}) + \epsilon\nabla_{\boldsymbol{v}}\frac{1}{2\gamma}\|\boldsymbol{v} - \boldsymbol{w}_0\|^2\right), \boldsymbol{w} - \boldsymbol{v}\right\rangle \\ \geq\ & \frac{\epsilon}{\gamma}(V_{\boldsymbol{v}-\boldsymbol{w}_0}(\boldsymbol{w} - \boldsymbol{w}_0) + V_{\boldsymbol{w}-\boldsymbol{w}_0}(\boldsymbol{v} - \boldsymbol{w}_0)).\end{aligned} \tag{60}$$

As Assumption 4 includes the strongly monotone assumption, by (60), we know that the VIP($F + \epsilon\nabla\frac{1}{2\gamma}\|\cdot - \boldsymbol{w}_0\|^2, \mathcal{W}$) satisfies Assumption 4 with parameter $\sigma = \epsilon$.

Lemma 1 is proved. ■

C.7 Proof of Corollary 1

Proof. By Theorem 1 and Lemma 1, if we optimize the regularized problem VIP($F + \epsilon\nabla\frac{1}{2\gamma}\|\cdot - \boldsymbol{w}_0\|^2, \mathcal{W}$) by the ODE Algorithm 4, then after K iterations, we have

$$\begin{aligned}&\sup_{\boldsymbol{w}\in\mathcal{W}, \|\tilde{\boldsymbol{w}}_K - \boldsymbol{w}\| \leq D} \langle F(\tilde{\boldsymbol{w}}_K) + \epsilon\nabla_{\tilde{\boldsymbol{w}}_K}\frac{1}{2\gamma}\|\tilde{\boldsymbol{w}}_K - \boldsymbol{w}_0\|^2, \tilde{\boldsymbol{w}}_K - \boldsymbol{w}\rangle \\ &\leq C_0 D\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|\sqrt{\frac{L}{A_{K-1} + a_1}},\end{aligned} \tag{61}$$

where C_0 is defined in Theorem 1, $A_{K-1} = \frac{1}{\epsilon}\left(1 + \frac{\sqrt{\alpha\gamma}\epsilon}{L}\right)^{K-1} - \frac{1}{\epsilon}$.

Meanwhile, by the convexity of $\frac{1}{2\gamma}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2$, we have

$$\langle \nabla_{\tilde{\boldsymbol{w}}_K} \frac{1}{2\gamma}\|\tilde{\boldsymbol{w}}_K - \boldsymbol{w}_0\|^2, \boldsymbol{w} - \tilde{\boldsymbol{w}}_K\rangle \leq \frac{1}{2\gamma}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2 - \frac{1}{2\gamma}\|\tilde{\boldsymbol{w}}_K - \boldsymbol{w}_0\|^2 \leq \frac{1}{2\gamma}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2. \quad (62)$$

So combining (61) and (62), we have

$$\begin{aligned} &\sup_{\boldsymbol{w}\in\mathcal{W}, \|\tilde{\boldsymbol{w}}_K - \boldsymbol{w}\|\leq D, \|\boldsymbol{w}-\boldsymbol{w}_0\|\leq D} \langle F(\tilde{\boldsymbol{w}}_K), \tilde{\boldsymbol{w}}_K - \boldsymbol{w}\rangle \\ \leq & D\epsilon + DC_0\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|\sqrt{\frac{L\epsilon}{\left(1 + \frac{\alpha\gamma\epsilon}{L}\right)^{K-1} - 1 + \frac{\alpha\gamma}{L}}}. \end{aligned} \quad (63)$$

Corollary 1 is proved. ■

C.8 Proof of Corollary 2

Proof. By Proposition 1 and Lemma 1, if we optimize the regularized problem VIP$(F + \epsilon\nabla\frac{1}{2\gamma}\|\cdot - \boldsymbol{w}_0\|^2, \mathcal{W})$ by the OptDE Algorithm 4, then after K iterations, we have

$$\begin{aligned} &\sup_{\boldsymbol{w}\in\mathcal{W}, \|\boldsymbol{w}_K - \boldsymbol{w}\|\leq D} \langle F(\boldsymbol{w}_K) + \epsilon\nabla_{\boldsymbol{w}_K}\frac{1}{2\gamma}\|\boldsymbol{w}_K - \boldsymbol{w}_0\|^2, \boldsymbol{w}_K - \boldsymbol{w}\rangle \\ \leq & C_0 D\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|\sqrt{\frac{L}{a_{K-1}}}, \end{aligned} \quad (64)$$

where C_0 is defined in Theorem 1, $a_{K-1} = \frac{\alpha\gamma}{L}\left(1 + \frac{\alpha\gamma\sigma}{L}\right)^{K-2}$.

Meanwhile, by the convexity of $\frac{1}{2\gamma}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2$, we have

$$\langle \nabla_{\boldsymbol{w}_K} \frac{1}{2\gamma}\|\boldsymbol{w}_K - \boldsymbol{w}_0\|^2, \boldsymbol{w} - \boldsymbol{w}_K\rangle \leq \frac{1}{2\gamma}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2 - \frac{1}{2\gamma}\|\boldsymbol{w}_K - \boldsymbol{w}_0\|^2 \leq \frac{1}{2\gamma}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2. \quad (65)$$

So combining (64) and (65), we have

$$\begin{aligned} &\sup_{\boldsymbol{w}\in\mathcal{W}, \|\boldsymbol{w}_K - \boldsymbol{w}\|\leq D, \|\boldsymbol{w}-\boldsymbol{w}_0\|\leq D} \langle F(\boldsymbol{w}_K), \boldsymbol{w}_K - \boldsymbol{w}\rangle \\ \leq & D\epsilon + DC_0 L\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|\sqrt{\frac{1}{\alpha\gamma\left(1 + \frac{\alpha\gamma\epsilon}{L}\right)^{K-2}}}. \end{aligned}$$

Corollary 2 is proved. ■

D Proof of Section B

By the definition of proximal operator (8), we can equivalently reformulate the stochastic optimistic dual extrapolation (SODE) of the main body as below. Then based on the definition of $\boldsymbol{g}_k$ in Step 7 and the definition of the Bregman divergence $V_{\boldsymbol{w}}(\boldsymbol{u})$, we can verify that

$$\boldsymbol{z}_k = \arg\min_{\boldsymbol{z}\in\mathcal{W}} \left\{ \hat{\psi}_k(\boldsymbol{z}) := \sum_{i=1}^{k} a_i \left(\langle F(\boldsymbol{w}_i; \xi_i), \boldsymbol{z} - \boldsymbol{u}\rangle + \frac{\sigma}{\gamma} V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{z} - \boldsymbol{w}_0) \right) + \frac{1}{2\gamma}\|\boldsymbol{z} - \boldsymbol{w}_0\|^2 \right\}, \quad (66)$$

where $\boldsymbol{u}$ is an arbitrary vector in $\mathcal{W}$ and is irrelevant to the minimizer $\boldsymbol{z}_k$. In our context, $\hat{\psi}_k(\boldsymbol{z})$ plays the role of a "generalized estimation sequence" to help us conduct convergence analysis. By the γ-strong convexity of the Bregman divergence $V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{z} - \boldsymbol{w}_0)$, we know that $\hat{\psi}_k(\boldsymbol{z})$ is strongly convex with strong convexity parameter $1 + \sigma\sum_{i=1}^{k} a_i = 1 + \sigma A_k$.

Algorithm 5 Stochastic Optimistic Dual Extrapolation **(Reformulation)**

```
1: Input: Lipshitz constant $L > 0$ from Assumption 1, $\gamma, \delta > 0$ from Assumption 2. The
   VIP$(F, \mathcal{W})$ satisfying Assumption 3 ($\sigma = 0$) or Assumption 4 ($\sigma > 0$).
2: $A_0 = 0, \alpha = \min\{\frac{\gamma}{32}, \frac{1}{16}\}$.
3: $\boldsymbol{w}_0 = \boldsymbol{z}_0 \in \mathcal{W}, \boldsymbol{g}_0 = \mathbf{0}$.
4: for $k = 1, 2, 3, \ldots, K$ do
5:     $a_k = \frac{\alpha\gamma\sqrt{1+\sigma A_{k-1}}}{L}, A_k = A_{k-1} + a_k$.
6:     $\boldsymbol{w}_k = \arg\min_{\boldsymbol{w}\in\mathcal{W}} \Big\{\langle F(\boldsymbol{w}_{k-1}; \xi_{k-1}), \boldsymbol{w}\rangle + \frac{L^2 a_k}{2(\alpha\gamma)^2}\|\boldsymbol{w} - \boldsymbol{z}_{k-1}\|^2\Big\}$.
7:     $\boldsymbol{g}_k = \boldsymbol{g}_{k-1} + a_k\big(F(\boldsymbol{w}_k; \xi_k) - \frac{\sigma}{\gamma}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{w}_0\|^2\big)$.
8:     $\boldsymbol{z}_k = \arg\min_{\boldsymbol{z}\in\mathcal{W}} \Big\{\langle \boldsymbol{g}_k, \boldsymbol{z}\rangle + \frac{1+\sigma A_k}{2\gamma}\|\boldsymbol{z} - \boldsymbol{w}_0\|^2\Big\}$
9: end for
10: $\tilde{\boldsymbol{w}}_K = \boldsymbol{w}_k$, where $k$ is chosen at random with probability distribution $\{\frac{a_1}{A_K}, \frac{a_2}{A_K}, \ldots, \frac{a_K}{A_K}\}$.
11: return $\tilde{\boldsymbol{w}}_K$.
```

D.1 Proof of Lemma 5

Proof. Given the definition of the generalized estimation sequence $\hat{\psi}_k(\boldsymbol{z})$ in (66) and by the optimality condition of the minimizer $\boldsymbol{z}_k$ in the Step 6 of Algorithm 5, we have: $\forall \boldsymbol{u} \in \mathcal{W}$,

$$\Big\langle \sum_{i=1}^{k} a_i(F(\boldsymbol{w}_i; \xi_i) + \frac{\sigma}{\gamma}\nabla V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{z}_k - \boldsymbol{w}_0)) + \nabla_{\boldsymbol{z}_k}\frac{1}{2\gamma}\|\boldsymbol{z}_k - \boldsymbol{w}_0\|^2, \boldsymbol{u} - \boldsymbol{z}_k\Big\rangle \geq 0. \tag{67}$$

Then we have: $\forall k \in [K]$,

$$\begin{aligned}
\hat{\psi}_k(\boldsymbol{z}_k) \quad &= \quad \sum_{i=1}^{k} a_i \left(\langle F(\boldsymbol{w}_i; \xi_i), \boldsymbol{z}_k - \boldsymbol{u}\rangle + \frac{\sigma}{\gamma}V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{z}_k - \boldsymbol{w}_0)\right) + \frac{1}{2\gamma}\|\boldsymbol{z}_k - \boldsymbol{w}_0\|^2 \\
&\overset{(a)}{\leq} \quad \frac{\sigma}{\gamma}\sum_{i=1}^{k} a_i\left(\langle \nabla V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{z}_k - \boldsymbol{w}_0), \boldsymbol{u} - \boldsymbol{z}_k\rangle + V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{z}_k - \boldsymbol{w}_0)\right) \\
&\qquad + \Big\langle \nabla_{\boldsymbol{z}_k}\frac{1}{2\gamma}\|\boldsymbol{z}_k - \boldsymbol{w}_0\|^2, \boldsymbol{u} - \boldsymbol{z}_k\Big\rangle + \frac{1}{2\gamma}\|\boldsymbol{z}_k - \boldsymbol{w}_0\|^2 \\
&\overset{(b)}{\leq} \quad \frac{\sigma}{\gamma}\sum_{i=1}^{k} a_i V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{u} - \boldsymbol{w}_0) + \frac{1}{2\gamma}\|\boldsymbol{u} - \boldsymbol{w}_0\|^2,
\end{aligned} \tag{68}$$

where (a) is by the optimality condition (67) and (b) is by the convexity of $V_{\boldsymbol{w}_i - \boldsymbol{w}_0}(\boldsymbol{u} - \boldsymbol{w}_0)$ and $\frac{1}{2\gamma}\|\boldsymbol{u} - \boldsymbol{w}_0\|^2$.

Meanwhile $\forall k \in [K]$, we have

$$\begin{aligned}
\hat{\psi}_k(\boldsymbol{z}_k) &= \hat{\psi}_{k-1}(\boldsymbol{z}_k) + a_k\langle F(\boldsymbol{w}_k; \xi_k), \boldsymbol{z}_k - \boldsymbol{u}\rangle \\
&\overset{(a)}{\geq} \hat{\psi}_{k-1}(\boldsymbol{z}_{k-1}) + \frac{1 + \sigma A_{k-1}}{2}\|\boldsymbol{z}_k - \boldsymbol{z}_{k-1}\|^2 + a_k\langle F(\boldsymbol{w}_k; \xi_k), \boldsymbol{z}_k - \boldsymbol{u}\rangle \\
&= \hat{\psi}_{k-1}(\boldsymbol{z}_{k-1}) + \frac{1 + \sigma A_{k-1}}{2}\|\boldsymbol{z}_k - \boldsymbol{z}_{k-1}\|^2 \\
&\qquad + a_k\langle F(\boldsymbol{w}_k; \xi_k), \boldsymbol{z}_k - \boldsymbol{w}_k\rangle + a_k\langle F(\boldsymbol{w}_k; \xi_k), \boldsymbol{w}_k - \boldsymbol{u}\rangle,
\end{aligned} \tag{69}$$

where (a) is the $(1 + \sigma A_{k-1})$-strong convexity of $\hat{\psi}_{k-1}(\boldsymbol{z})$. Meanwhile, by the γ-strong convexity of $\frac{1}{2}\|\cdot\|^2$, we have

$$\begin{aligned}
&a_k\,\langle F(\boldsymbol{w}_k; \xi_k), \boldsymbol{w}_k - \boldsymbol{z}_k\rangle - \frac{1 + \sigma A_{k-1}}{2}\|\boldsymbol{z}_k - \boldsymbol{z}_{k-1}\|^2 \\
\leq &\Big\langle a_k F(\boldsymbol{w}_k; \xi_k) + (1 + \sigma A_{k-1})\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{z}_k\Big\rangle \\
&\quad - (1 + \sigma A_{k-1})\Big(\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{\gamma}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2\Big).
\end{aligned} \tag{70}$$

Then combining (69) and (70), we have

$$\begin{aligned}
& a_k\langle F(\boldsymbol{w}_k;\xi_k), \boldsymbol{w}_k-\boldsymbol{u}\rangle \\
\leq\ & \Big\langle a_k F(\boldsymbol{w}_k;\xi_k) + (1+\sigma A_{k-1})\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k-\boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k-\boldsymbol{z}_k\Big\rangle \\
& -(1+\sigma A_{k-1})\Big(\frac{1}{2}\|\boldsymbol{w}_k-\boldsymbol{z}_{k-1}\|^2+\frac{\gamma}{2}\|\boldsymbol{w}_k-\boldsymbol{z}_k\|^2\Big)+\hat{\psi}_k(\boldsymbol{z}_k)-\hat{\psi}_{k-1}(\boldsymbol{z}_{k-1}). \qquad (71)
\end{aligned}$$

Summing (71) from $k=1$ to K, we have

$$\begin{aligned}
& \sum_{k=1}^{K} a_k\langle F(\boldsymbol{w}_k;\xi_k), \boldsymbol{w}_k-\boldsymbol{u}\rangle \\
\leq\ & \sum_{k=1}^{K}\bigg(\Big\langle a_k F(\boldsymbol{w}_k;\xi_k) + (1+\sigma A_{k-1})\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k-\boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k-\boldsymbol{z}_k\Big\rangle \\
& -(1+\sigma A_{k-1})\Big(\frac{1}{2}\|\boldsymbol{w}_k-\boldsymbol{z}_{k-1}\|^2+\frac{\gamma}{2}\|\boldsymbol{w}_k-\boldsymbol{z}_k\|^2\Big)\bigg)+\hat{\psi}_K(\boldsymbol{z}_K)-\hat{\psi}_0(\boldsymbol{z}_0) \\
\overset{(a)}{\leq}\ & \sum_{k=1}^{K}\bigg(\Big\langle a_k F(\boldsymbol{w}_k;\xi_k) + (1+\sigma A_{k-1})\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k-\boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k-\boldsymbol{z}_k\Big\rangle \\
& -(1+\sigma A_{k-1})\Big(\frac{1}{2}\|\boldsymbol{w}_k-\boldsymbol{z}_{k-1}\|^2+\frac{\gamma}{2}\|\boldsymbol{w}_k-\boldsymbol{z}_k\|^2\Big)\bigg) \\
& +\frac{\sigma}{\gamma}\sum_{k=1}^{K} a_k V_{\boldsymbol{w}_k-\boldsymbol{w}_0}(\boldsymbol{u}-\boldsymbol{w}_0)+\frac{1}{2\gamma}\|\boldsymbol{u}-\boldsymbol{w}_0\|^2 \\
\overset{(b)}{=}\ & \sum_{k=1}^{K} a_k\bigg(\Big\langle F(\boldsymbol{w}_k;\xi_k) + \frac{L^2 a_k}{(\alpha\gamma)^2}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k-\boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k-\boldsymbol{z}_k\Big\rangle \\
& \qquad -\frac{L^2 a_k}{(\alpha\gamma)^2}\Big(\frac{1}{2}\|\boldsymbol{w}_k-\boldsymbol{z}_{k-1}\|^2+\frac{\gamma}{2}\|\boldsymbol{w}_k-\boldsymbol{z}_k\|^2\Big)\bigg) \\
& +\frac{\sigma}{\gamma}\sum_{k=1}^{K} a_k V_{\boldsymbol{w}_k-\boldsymbol{w}_0}(\boldsymbol{u}-\boldsymbol{w}_0)+\frac{1}{2\gamma}\|\boldsymbol{u}-\boldsymbol{w}_0\|^2, \qquad (72)
\end{aligned}$$

where (a) is by the fact $\hat{\psi}_0(\boldsymbol{z}_0)=0$, the upper bound of $\hat{\psi}_K(\boldsymbol{z}_K)$ in (68), (b) is by the setting $a_k^2=\frac{(\alpha\gamma)^2(1+\sigma A_{k-1})}{L^2}$ in Algorithm 5. Meanwhile, taking expectation on ξ_k, we have: $\forall\boldsymbol{u}\in\mathcal{W}$,

$$\begin{aligned}
\langle F(\boldsymbol{w}_k), \boldsymbol{w}_k-\boldsymbol{u}\rangle &= \mathbb{E}_{\xi_k}\Big[\langle F(\boldsymbol{w}_k)-F(\boldsymbol{w}_k;\xi_k), \boldsymbol{w}_k-\boldsymbol{u}\rangle\Big]+\mathbb{E}_{\xi_k}\Big[\langle F(\boldsymbol{w}_k;\xi_k), \boldsymbol{w}_k-\boldsymbol{u}\rangle\Big] \\
&= \mathbb{E}_{\xi_k}\Big[\langle F(\boldsymbol{w}_k;\xi_k), \boldsymbol{w}_k-\boldsymbol{u}\rangle\Big]. \qquad (73)
\end{aligned}$$

So taking expectation on the randomness of all the history for (72), and using (73) and the definition of $\{E_{2k}\}$ in Lemma 5, after simple arrangements, Lemma 5 is proved. ■

D.2 Proof of Lemma 6

Proof. By the definition of E_{2k} in Lemma 5, we have: $\forall k \in [K]$,

$$\begin{aligned} E_{2k} \;&=\; a_k\Big(\Big\langle F(\boldsymbol{w}_k;\xi_k) + \frac{L^2 a_k}{(\alpha\gamma)^2}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{z}_k\Big\rangle \\ &\qquad -\frac{L^2 a_k}{(\alpha\gamma)^2}\Big(\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{\gamma}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2\Big)\Big) \\ &\quad + a_k\langle F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k;\xi_k), \boldsymbol{w}_k - \boldsymbol{u}\rangle \\ \;&\le\; a_k\Big(\Big\langle F(\boldsymbol{w}_k;\xi_k) - F(\boldsymbol{w}_{k-1};\xi_{k-1}), \boldsymbol{w}_k - \boldsymbol{z}_k\Big\rangle \\ &\qquad + \Big\langle F(\boldsymbol{w}_{k-1};\xi_{k-1}) + \frac{L^2 a_k}{(\alpha\gamma)^2}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{z}_k\Big\rangle \\ &\qquad -\frac{L^2 a_k}{(\alpha\gamma)^2}\Big(\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{\gamma}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2\Big)\Big) \\ &\quad + a_k\langle F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k;\xi_k), \boldsymbol{w}_k - \boldsymbol{u}\rangle. \end{aligned} \tag{74}$$

Meanwhile, we have: for all $\alpha > 0$,

$$\begin{aligned} &\Big\langle F(\boldsymbol{w}_k;\xi_k) - F(\boldsymbol{w}_{k-1};\xi_{k-1}), \boldsymbol{w}_k - \boldsymbol{z}_k\Big\rangle \\ \overset{(a)}{\le}\;& \|F(\boldsymbol{w}_k;\xi_k) - F(\boldsymbol{w}_{k-1};\xi_{k-1})\|_*\|\boldsymbol{w}_k - \boldsymbol{z}_k\| \\ \overset{(b)}{\le}\;& (\|F(\boldsymbol{w}_k) - F(\boldsymbol{w}_{k-1})\|_* + \|F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k;\xi_k)\|_* \\ &\quad + \|F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1};\xi_{k-1})\|_*)\|\boldsymbol{w}_k - \boldsymbol{z}_k\| \\ \overset{(c)}{\le}\;& (L\|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\| + \|F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k;\xi_k)\|_* + \|F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1};\xi_{k-1})\|_*)\|\boldsymbol{w}_k - \boldsymbol{z}_k\| \\ \overset{(d)}{\le}\;& \frac{\alpha}{L^2 a_k}(L\|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\| + \|F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k;\xi_k)\|_* + \|F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1};\xi_{k-1})\|_*)^2 \\ &+ \frac{L^2 a_k}{4\alpha}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2. \\ \overset{(e)}{\le}\;& \frac{2\alpha}{a_k}\|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\|^2 + \frac{2\alpha}{L^2 a_k}(\|F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k;\xi_k)\|_* + \|F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1};\xi_k)\|_*)^2 \\ &+ \frac{L^2 a_k}{4\alpha}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2 \\ \overset{(f)}{\le}\;& \frac{2\alpha}{a_k}\|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\|^2 + \frac{4\alpha}{L^2 a_k}(\|F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k;\xi_k)\|_*^2 + \|F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1};\xi_k)\|_*^2) \\ &+ \frac{L^2 a_k}{4\alpha}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2 \\ \overset{(g)}{\le}\;& \frac{4\alpha}{a_k}(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \|\boldsymbol{z}_{k-1} - \boldsymbol{w}_{k-1}\|^2) \\ &+ \frac{4\alpha}{L^2 a_k}(\|F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k;\xi_k)\|_*^2 + \|F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1};\xi_k)\|_*^2) + \frac{L^2 a_k}{4\alpha}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2, \end{aligned} \tag{75}$$

where (a) is by the Cauchy-Schwarz inequality, (b) is by the triangle inequality of the norm $\|\cdot\|_*$, (c) is by the Lipschitz continuity of $F(\boldsymbol{w})$, (d) is by the fact $ab \le a^2 + \frac{b^2}{4}$, $(e), (f)$ and (g) is by the fact $(a+b)^2 \le 2(a^2+b^2)$.

Then by the optimality condition of $\boldsymbol{w}_k$ in Algorithm 5, we have: $\forall \boldsymbol{z} \in \mathcal{W}$,

$$\Big\langle F(\boldsymbol{w}_{k-1};\xi_{k-1}) + \frac{L^2 a_k}{(\alpha\gamma)^2}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{z}\Big\rangle \le 0. \tag{76}$$

Combining (74), (75) and (76) with $\boldsymbol{z} := \boldsymbol{z}_k$, we have

$$\begin{aligned} E_{2k} \quad &\leq \quad -\Big(\frac{L^2 a_k^2}{2(\alpha\gamma)^2} - 4\alpha\Big)\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 \\ &\quad + \frac{4\alpha}{L^2}(\|F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k;\xi_k)\|_*^2 + \|F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1};\xi_{k-1})\|_*^2) \\ &\quad + 4\alpha\|\boldsymbol{z}_{k-1} - \boldsymbol{w}_{k-1}\|^2 - \frac{L^2 a_k^2}{4\alpha}\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2 + a_k\langle F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k;\xi_k), \boldsymbol{w}_k - \boldsymbol{u}\rangle. \end{aligned}$$

For both the settings $\sigma = 0$ and $\sigma > 0$, by our setting, we have $a_k \geq a_1 = \frac{\alpha\gamma}{L}$ and $\alpha = \min\{\frac{\gamma}{32}, \frac{1}{16}\}$, so we have

$$\frac{L^2 a_k^2}{2(\alpha\gamma)^2} \geq \frac{1}{2} \geq 8\alpha, \quad \frac{L^2 a_k^2}{4\alpha} \geq 8\alpha. \tag{77}$$

Then it follows that

$$\begin{aligned} E_{2k} \quad &\leq \quad -4\alpha\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \frac{4\alpha}{L^2}(\|F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k;\xi_k)\|_*^2 + \|F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1};\xi_{k-1})\|_*^2) \\ &\quad + 4\alpha\|\boldsymbol{z}_{k-1} - \boldsymbol{w}_{k-1}\|^2 - 8\alpha\|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2 + \frac{\alpha\gamma}{L}\langle F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k;\xi_k), \boldsymbol{w}_k - \boldsymbol{u}\rangle. \end{aligned} \tag{78}$$

So summing (78) from $k = 1$ to K and by the fact $\mathbb{E}[\langle F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k;\xi_k), \boldsymbol{w}_k - \boldsymbol{u}\rangle] = 0$, we have

$$\begin{aligned} \mathbb{E}\Big[\sum_{k=1}^K E_{2k}\Big] \quad &\leq \quad \mathbb{E}\Big[-4\alpha\sum_{k=1}^K(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \|\boldsymbol{w}_k - \boldsymbol{z}_k\|^2) \\ &\quad + \frac{4\alpha}{L^2}\sum_{k=1}^K(\|F(\boldsymbol{w}_k) - F(\boldsymbol{w}_k;\xi_k)\|_*^2 + \|F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1};\xi_{k-1})\|_*^2)\Big] \\ &\overset{(a)}{\leq} \quad -\mathbb{E}\Big[4\alpha\sum_{k=1}^K(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2)\Big] - \mathbb{E}[4\alpha\|\boldsymbol{w}_K - \boldsymbol{z}_K\|^2] + \frac{8\alpha s^2 K}{L^2} \\ &\leq \quad -\mathbb{E}\Big[4\alpha\sum_{k=1}^K(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2)\Big] + \frac{8\alpha s^2 K}{L^2}, \end{aligned} \tag{79}$$

where (a) is by the condition $\boldsymbol{w}_0 = \boldsymbol{z}_0$ and Assumption 5. Lemma 6 is proved.

■

D.3 Proof of Lemma 7

It follows that: $\forall \boldsymbol{w} \in \mathcal{W}$

$$\begin{aligned} &\langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{w}\rangle \\ = \quad &\Big\langle F(\boldsymbol{w}_k) - \Big(F(\boldsymbol{w}_{k-1};\xi_{k-1}) + \frac{L^2 a_k}{(\alpha\gamma)^2}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2\Big), \boldsymbol{w}_k - \boldsymbol{w}\Big\rangle \\ &+ \Big\langle F(\boldsymbol{w}_{k-1};\xi_{k-1}) + \frac{L^2 a_k}{(\alpha\gamma)^2}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{w}\Big\rangle \\ \overset{(a)}{=} \quad &\Big\langle F(\boldsymbol{w}_k) - \Big(F(\boldsymbol{w}_{k-1};\xi_{k-1}) + \frac{L}{\alpha\gamma}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2\Big), \boldsymbol{w}_k - \boldsymbol{w}\Big\rangle \\ &+ \Big\langle F(\boldsymbol{w}_{k-1};\xi_{k-1}) + \frac{L}{\alpha\gamma}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{w}\Big\rangle \\ \overset{(b)}{\leq} \quad &\Big\langle F(\boldsymbol{w}_k) - \big(F(\boldsymbol{w}_{k-1};\xi_{k-1}) + \frac{L}{\alpha\gamma}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2\big), \boldsymbol{w}_k - \boldsymbol{w}\Big\rangle \\ \leq \quad &\langle F(\boldsymbol{w}_k) - F(\boldsymbol{w}_{k-1}), \boldsymbol{w}_k - \boldsymbol{w}\rangle + \langle F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1};\xi_{k-1}), \boldsymbol{w}_k - \boldsymbol{w}\rangle \\ &\quad + \Big\langle \frac{L}{\alpha\gamma}\nabla_{\boldsymbol{w}_k}\frac{1}{2}\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2, \boldsymbol{w}_k - \boldsymbol{w}\Big\rangle, \end{aligned}$$

where (a) is by the fact $a_k = \frac{\alpha\gamma}{L}$ when $\sigma = 0$, (b) is by the optimality condition of $\boldsymbol{w}_k$.

So it follows that

$$\begin{aligned}
& \langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{w}\rangle \\
\overset{(a)}{\leq} \; & \|F(\boldsymbol{w}_k) - F(\boldsymbol{w}_{k-1})\|_* \|\boldsymbol{w}_k - \boldsymbol{w}\| \\
& + \langle F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1}; \xi_{k-1}), \boldsymbol{w}_k - \boldsymbol{w}_{k-1}\rangle \\
& + \langle F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1}; \xi_{k-1}), \boldsymbol{w}_{k-1} - \boldsymbol{w}\rangle \\
& + \frac{L}{\alpha\gamma} \left\|\nabla_{\boldsymbol{w}_k} \frac{1}{2} \|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2\right\|_* \|\boldsymbol{w}_k - \boldsymbol{w}\| \\
\overset{(b)}{\leq} \; & L\|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\|\|\boldsymbol{w}_k - \boldsymbol{w}\| + \|F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1}; \xi_{k-1})\|_* \|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\| \\
& + \langle F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1}; \xi_{k-1}), \boldsymbol{w}_{k-1} - \boldsymbol{w}\rangle + \frac{L\delta}{\alpha\gamma} \|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|\|\boldsymbol{w}_k - \boldsymbol{w}\| \\
\leq \; & (1 + \frac{\delta}{\alpha\gamma}) L(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|)\|\boldsymbol{w}_k - \boldsymbol{w}\| \\
& + \|F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1}; \xi_{k-1})\|_* \|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\| \\
& + \langle F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1}; \xi_{k-1}), \boldsymbol{w}_{k-1} - \boldsymbol{w}\rangle,
\end{aligned} \tag{80}$$

(a) is by the Cauchy Schwarz inequality and simple arrangement, and (b) is by Assumption 1.

Then

$$\begin{aligned}
& \langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{w}\rangle \\
\overset{(a)}{\leq} \; & (1 + \frac{\delta}{\alpha\gamma}) L(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|)\|\boldsymbol{w}_k - \boldsymbol{w}\| \\
& + \frac{1}{2L^2} \|F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1}; \xi_{k-1})\|_*^2 \\
& + \frac{L^2}{2} \|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\|^2 + \langle F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1}; \xi_{k-1}), \boldsymbol{w}_{k-1} - \boldsymbol{w}\rangle,
\end{aligned} \tag{81}$$

where (a) is by the fact $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$.

So taking expectation on ξ_{k-1}, by Assumption 5, we have: $\forall \boldsymbol{w} \in \mathcal{W}$

$$\begin{aligned}
& \mathbb{E}_{\xi_{k-1}}[\langle F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1}; \xi_{k-1}), \boldsymbol{w}_{k-1} - \boldsymbol{w}\rangle] \\
= \; & \langle \mathbb{E}_{\xi_{k-1}}[F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1}; \xi_{k-1})], \boldsymbol{w}_{k-1} - \boldsymbol{w}\rangle \\
= \; & \langle F(\boldsymbol{w}_{k-1}) - F(\boldsymbol{w}_{k-1}), \boldsymbol{w}_{k-1} - \boldsymbol{w}\rangle \\
= \; & 0.
\end{aligned} \tag{82}$$

By Assumption 5, we have

$$\begin{aligned}
& \mathbb{E}_{\xi_{k-1}}\Big[\sup_{\boldsymbol{w}\in\mathcal{W}, \|\boldsymbol{w}_k - \boldsymbol{w}\| \leq D} \langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{w}\rangle\Big] \\
\leq & (1 + \frac{\delta}{\alpha\gamma}) LD\mathbb{E}_{\xi_{k-1}}[(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\| + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|)] + \frac{L^2}{2}\mathbb{E}_{\xi_{k-1}}[\|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\|^2] + \frac{s^2}{2L^2}.
\end{aligned}$$

Lemma 7 is proved.

D.4 Proof of Theorem 2

Proof. Firstly, by the setting $a_k = \frac{\alpha\gamma\sqrt{1+\sigma A_{k-1}}}{L}$ and $A_0 = 0$, $A_k = A_{k-1} + a_k$, we have

- If $\sigma = 0$, then $A_k = \frac{\alpha\gamma k}{L}$.
- If $\sigma > 0$, then $A_k = \left(\frac{\alpha\gamma}{4L}\right)^2 \sigma(k+1)^2$.

Then for both the setting $\sigma = 0$ ($i.e.$, Assumption 3 holds) and $\sigma > 0$ ($i.e.$, Assumption 4 holds), we have

$$\langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{w}^* \rangle \geq \frac{\sigma}{\gamma}(V_{\boldsymbol{w}_k - \boldsymbol{w}_0}(\boldsymbol{w}^* - \boldsymbol{w}_0) + V_{\boldsymbol{w}^* - \boldsymbol{w}_0}(\boldsymbol{w}_k - \boldsymbol{w}_0)). \tag{83}$$

So in Lemma 5, let $\boldsymbol{u} = \boldsymbol{w}^*$, we have

$$\begin{aligned} 0 &\leq \mathbb{E}\Big[\sum_{k=1}^{K} \frac{\sigma a_k}{2}\|\boldsymbol{w}_k - \boldsymbol{w}^*\|^2\Big] \\ &\overset{(a)}{\leq} \mathbb{E}\Big[\sum_{k=1}^{K} \frac{\sigma a_k}{\gamma} V_{\boldsymbol{w}^* - \boldsymbol{w}_0}(\boldsymbol{w}_k - \boldsymbol{w}_0)\Big] \\ &\overset{(b)}{\leq} \mathbb{E}\Big[\sum_{k=1}^{K} a_k \left(\langle F(\boldsymbol{w}_k), \boldsymbol{w}_k - \boldsymbol{w}^*\rangle - \frac{\sigma}{\gamma} V_{\boldsymbol{w}_k - \boldsymbol{w}_0}(\boldsymbol{w}^* - \boldsymbol{w}_0)\right)\Big] \\ &\overset{(c)}{\leq} \mathbb{E}\Big[\sum_{k=1}^{K} E_{2k} + \frac{1}{2\gamma}\|\boldsymbol{w}^* - \boldsymbol{w}_0\|^2\Big] \\ &\overset{(d)}{\leq} -\mathbb{E}\Big[4\alpha \sum_{k=1}^{K}(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2)\Big] + \frac{8\alpha s^2 K}{L^2} + \frac{1}{2\gamma}\|\boldsymbol{w}^* - \boldsymbol{w}_0\|^2, \end{aligned} \tag{84}$$

where (a) is by the γ-strong convexity Bregman divergence of $V_{\boldsymbol{w}^* - \boldsymbol{w}_0}(\boldsymbol{w}_k - \boldsymbol{w}_0)$, (b) is by the Assumption 3 ($\sigma = 0$) or the Assumption 4 ($\sigma > 0$), (c) is by Lemma 5, and (d) is by Lemma 6.

After a simple arrangement, we have

$$\mathbb{E}\Big[\sum_{k=1}^{K} \frac{1}{K}(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2)\Big] \leq \frac{\|\boldsymbol{w}^* - \boldsymbol{w}_0\|^2}{8\alpha K} + \frac{2s^2}{L^2}. \tag{85}$$

Then by randomly picking a $\tilde{k} \in [K]$ with probability distribution $\left\{\frac{a_1}{A_K}, \frac{a_2}{A_K}, \ldots, \frac{a_K}{A_K}\right\}$ and let the output $\tilde{\boldsymbol{w}}_K := \boldsymbol{w}_{\tilde{k}}$, then taking expectation on $\tilde{\boldsymbol{w}}_K$

$$\mathbb{E}_{\tilde{k}}[\|\boldsymbol{w}_{\tilde{k}} - \boldsymbol{z}_{\tilde{k}-1}\|^2 + \|\boldsymbol{w}_{\tilde{k}-1} - \boldsymbol{z}_{\tilde{k}-1}\|^2] = \sum_{k=1}^{K} \frac{a_k}{A_K}(\|\boldsymbol{w}_k - \boldsymbol{z}_{k-1}\|^2 + \|\boldsymbol{w}_{k-1} - \boldsymbol{z}_{k-1}\|^2). \tag{86}$$

So taking expectation on all the history, we have

$$\mathbb{E}[\|\boldsymbol{w}_{\tilde{k}} - \boldsymbol{z}_{\tilde{k}-1}\|^2 + \|\boldsymbol{w}_{\tilde{k}-1} - \boldsymbol{z}_{\tilde{k}-1}\|^2] \leq \frac{\|\boldsymbol{w}^* - \boldsymbol{w}_0\|^2}{8\alpha K} + \frac{2s^2}{L^2}. \tag{87}$$

Then taking expectation on all the history, we have

$$\begin{aligned} &\mathbb{E}[\|\boldsymbol{w}_{\tilde{k}} - \boldsymbol{z}_{\tilde{k}-1}\| + \|\boldsymbol{w}_{\tilde{k}-1} - \boldsymbol{z}_{\tilde{k}-1}\|] \\ \overset{(a)}{\leq}\; &(\mathbb{E}[(\|\boldsymbol{w}_{\tilde{k}} - \boldsymbol{z}_{\tilde{k}-1}\| + \|\boldsymbol{w}_{\tilde{k}-1} - \boldsymbol{z}_{\tilde{k}-1}\|)^2])^{1/2} \\ \overset{(b)}{\leq}\; &(\mathbb{E}[2(\|\boldsymbol{w}_{\tilde{k}} - \boldsymbol{z}_{\tilde{k}-1}\|^2 + \|\boldsymbol{w}_{\tilde{k}-1} - \boldsymbol{z}_{\tilde{k}-1}\|^2)])^{1/2} \\ \overset{(c)}{\leq}\; &\sqrt{2}\sqrt{\frac{\|\boldsymbol{w}^* - \boldsymbol{w}_0\|^2}{8\alpha K} + \frac{2s^2}{L^2}}, \end{aligned} \tag{88}$$

where (a) is by the Jensen inequality, (b) is by the fact that $(a+b)^2 \leq 2(a^2 + b^2)$ and (c) is by (87).

$$
\begin{aligned}
&\mathbb{E}\Big[\sup_{\boldsymbol{w}\in\mathcal{W},\|\boldsymbol{w}_{\tilde{k}}-\boldsymbol{w}\|\le D}\langle F(\boldsymbol{w}_{\tilde{k}}),\boldsymbol{w}_{\tilde{k}}-\boldsymbol{w}\rangle\Big]\\
\overset{(a)}{\le}\;& \mathbb{E}\Big[\big(1+\frac{\delta}{\alpha\gamma}\big)LD(\|\boldsymbol{w}_{\tilde{k}}-\boldsymbol{z}_{\tilde{k}-1}\|+\|\boldsymbol{w}_{\tilde{k}-1}-\boldsymbol{z}_{\tilde{k}-1}\|)\\
&+\frac{L^2}{2}\|\boldsymbol{w}_{\tilde{k}}-\boldsymbol{w}_{\tilde{k}-1}\|^2+\frac{1}{2L^2}\|F(\boldsymbol{w}_{\tilde{k}-1})-F(\boldsymbol{w}_{\tilde{k}-1};\xi_{\tilde{k}-1})\|_*^2\Big]\\
\overset{(b)}{\le}\;& \mathbb{E}\Big[\big(1+\frac{\delta}{\alpha\gamma}\big)LD(\|\boldsymbol{w}_{\tilde{k}}-\boldsymbol{z}_{\tilde{k}-1}\|+\|\boldsymbol{w}_{\tilde{k}-1}-\boldsymbol{z}_{\tilde{k}-1}\|)+\frac{L^2}{2}(\|\boldsymbol{w}_{\tilde{k}}-\boldsymbol{z}_{\tilde{k}-1}\|+\|\boldsymbol{w}_{\tilde{k}-1}-\boldsymbol{z}_{\tilde{k}-1}\|)^2\Big]\\
&+\frac{s^2}{2L^2}\\
\overset{(c)}{\le}\;& \mathbb{E}\Big[\big(1+\frac{\delta}{\alpha\gamma}\big)LD(\|\boldsymbol{w}_{\tilde{k}}-\boldsymbol{z}_{\tilde{k}-1}\|+\|\boldsymbol{w}_{\tilde{k}-1}-\boldsymbol{z}_{\tilde{k}-1}\|)+L^2(\|\boldsymbol{w}_{\tilde{k}}-\boldsymbol{z}_{\tilde{k}-1}\|+\|\boldsymbol{w}_{\tilde{k}-1}-\boldsymbol{z}_{\tilde{k}-1}\|)^2\Big]\\
&+\frac{s^2}{2L^2}\\
\overset{(d)}{\le}\;& \sqrt{2}\big(1+\frac{\delta}{\alpha\gamma}\big)LD\sqrt{\frac{\|\boldsymbol{w}^*-\boldsymbol{w}_0\|^2}{8\alpha K}+\frac{2s^2}{L^2}}+L^2\Big(\frac{\|\boldsymbol{w}^*-\boldsymbol{w}_0\|^2}{8\alpha K}+\frac{2s^2}{L^2}\Big)+\frac{s^2}{2L^2},
\end{aligned}
\tag{89}
$$

where (a) is by Lemma 7, (b) is by the triangle inequality of $\|\cdot\|$ and Assumption 5, (c) is by the triangle inequality of $\|\cdot\|$, (d) is by (87) and (88).

For $\sigma>0$, by (84) and (85), we have

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{k=1}^{K}\frac{a_k\sigma}{2}\|\boldsymbol{w}_k-\boldsymbol{w}^*\|^2\Big]\\
\le\;& -\mathbb{E}\Big[4\alpha\sum_{k=1}^{K}(\|\boldsymbol{w}_k-\boldsymbol{z}_{k-1}\|^2+\|\boldsymbol{w}_{k-1}-\boldsymbol{z}_{k-1}\|^2)\Big]+\frac{8\alpha s^2K}{L^2}+\frac{1}{2\gamma}\|\boldsymbol{w}^*-\boldsymbol{w}_0\|^2\\
\le\;& \frac{8\alpha s^2K}{L^2}+\frac{1}{2\gamma}\|\boldsymbol{w}^*-\boldsymbol{w}_0\|^2.
\end{aligned}
$$

So by the definition of $\tilde{\boldsymbol{w}}_K$, taking expectation on the randomness of all the history, we have

$$
\begin{aligned}
&\mathbb{E}[\|\boldsymbol{w}_{\tilde{k}}-\boldsymbol{w}^*\|^2]\\
\le\;& \mathbb{E}\Big[\sum_{k=1}^{K}\frac{a_k}{A_K}\|\boldsymbol{w}_k-\boldsymbol{w}^*\|^2\Big]\\
\le\;& \frac{2}{\sigma A_K}\Big(\frac{8\alpha s^2K}{L^2}+\frac{1}{2\gamma}\|\boldsymbol{w}^*-\boldsymbol{w}_0\|^2\Big)\\
\le\;& \frac{32L^2}{\sigma^2(\alpha\gamma)^2(K+1)^2}\Big(\frac{8\alpha s^2K}{L^2}+\frac{1}{2\gamma}\|\boldsymbol{w}^*-\boldsymbol{w}_0\|^2\Big).
\end{aligned}
$$

Theorem 2 is proved. ■