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@s.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 $\sigma$-weak solution exists, the convergence guarantee is improved to the linear rate $O(\log 1/\epsilon)$. Along the way—as a byproduct of our inquiries into non-monotone variational inequalities—we provide the near-optimal $O(1/\epsilon \log 1/\epsilon)$ 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 $W \subset \mathbb{R}^d$ and an operator $F : W \to \mathbb{R}^d$, the variational inequality problem VIP($F, W$) aims to find a solution $w^* \in W$ such that:

$$\forall w \in W, \quad \langle F(w^*), w - w^* \rangle \geq 0,$$

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

$$\min_{x \in X} \max_{y \in Y} f(x, y),$$

let $W \equiv X \times Y, w \equiv [x, y], F(w) \equiv \begin{bmatrix} \nabla_x f(x, y) \\ -\nabla_y f(x, y) \end{bmatrix}$. 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.

Convex-Concave Minimax Problems. The operator $F(w)$ will be monotone if
\[ \forall w, v \in W, \langle F(w) - F(v), w - v \rangle \geq 0. \] (3)

VI with monotone operators has been well studied, which provides a concise and optimal framework for convex-concave minimax problems [29]. For monotone VIP($F, W$), it is well known that the strong solution satisfying (1) is also equivalent to the solution $w^* \in W$ satisfying:
\[ \forall w \in W, \langle F(w), w - w^* \rangle \geq 0, \] (4)
where $w^*$ is called a weak solution of VIP($F, W$). A classical result [29] under the monotone and Lipschitz continuous assumptions is that the Mirror-Prox algorithm [29] can converge to an $\epsilon$-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 $\epsilon$-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 condition3, 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?

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

3As 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.)

<table>
<thead>
<tr>
<th>Convergence measure</th>
<th>Merit function (Definition 1)</th>
<th>Distance $|w^*|^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weak solution exists</td>
<td>$O(1/\epsilon^2)$</td>
<td>$O(1/\epsilon^4)$</td>
</tr>
<tr>
<td>σ-weak solution exists</td>
<td>—</td>
<td>$O(\log \frac{1}{\epsilon})$</td>
</tr>
<tr>
<td>No. of gradient calls</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Convergence measure</th>
<th>Merit function (Definition 1)</th>
<th>Distance $\mathbb{E}[|w^*|^2]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>SEG [18]</td>
<td>SOptDE (this paper)</td>
</tr>
<tr>
<td>Weak solution exists</td>
<td>$O(1/\epsilon^4)$</td>
<td>$O(1/\epsilon^4)$</td>
</tr>
<tr>
<td>σ-weak solution exists</td>
<td>—</td>
<td>$O(1/\epsilon^2 \log \frac{1}{\epsilon})$</td>
</tr>
<tr>
<td>No. of gradient calls</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

**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** $w^*$, 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** $w^*$ 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 $\|w^*\|^2$ 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 $O(\frac{1}{\epsilon^2} \log \frac{1}{\epsilon})$ number of iterations. To our knowledge, the near-optimal result $O(\frac{1}{\epsilon^2} \log \frac{1}{\epsilon})$ 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 $\mathbf{x} \in \mathbb{R}^d$ 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 $\|y\|_* := \max_{\|x\| \leq 1} \langle x, y \rangle$. For $\mathbf{x} \in \mathbb{R}^d$ and $p \geq 1$, let $\|\mathbf{x}\|_p := \left( \sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}}$.

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

Assumption 1 (Existence of a weak solution) For the VIP($F, \mathcal{W}$) in (1), for all $\mathbf{w}, \mathbf{v} \in \mathcal{W}$, $\|F(\mathbf{w}) - F(\mathbf{v})\|_* \leq L \|\mathbf{w} - \mathbf{v}\|$, where $L > 0$ is the Lipschitz constant.

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

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

$$\left\| \nabla \frac{1}{2} \|\mathbf{w}\|^2 \right\|_* \leq \delta \|\mathbf{w}\|. \quad (6)$$

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

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

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

$$V_\nu(\mathbf{w}) := \frac{1}{2} \|\mathbf{w}\|^2 - \frac{1}{2} \|\mathbf{v}\|^2 - \langle \nabla \frac{1}{2} \|\mathbf{v}\|^2, \mathbf{w} - \mathbf{v} \rangle. \quad (9)$$

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

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 $\mathbf{w}^* \in \mathcal{W}$ such that for all $\mathbf{w} \in \mathcal{W}$, $\langle F(\mathbf{w}), \mathbf{w} - \mathbf{w}^* \rangle \geq 0$.

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


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: $A_0 = 0$, $0 < \alpha \leq \min \left\{ \frac{1}{4\sqrt{2}}, \frac{1}{3\sqrt{\gamma}} \right\}$.
3: $w_0 = z_0 \in \mathcal{W}, g_0 = 0$.
4: for $k = 1, 2, 3, \ldots, K$ do
5: $a_k = \frac{\alpha(1+\sigma A_k-1)}{L}, A_k = A_{k-1} + a_k$.
6: $w_k = P_{z_{k-1}} \left( \frac{1}{L} F(w_{k-1}) \right)$.
7: $g_k = g_{k-1} + a_k \left( F(w_k) - \frac{\gamma}{2} \nabla w_k \frac{1}{2} \|w_k - w_0\|^2 \right)$.
8: $z_k = P_{w_0} \left( \frac{1}{1+\sigma A_k} g_k \right)$.
9: end for
10: $\hat{w}_K^* = \arg \min_{w_k: k \in [K]} \left( \|w_k - z_{k-1}\| + \|w_{k-1} - z_{k-1}\| \right)$.
11: **return** $\hat{w}_K^*$.

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 $\min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} x^2 y^2$ 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 $\langle F(w), w - w^* \rangle \geq \sigma \|w - w^*\|^2_2$. 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 [3]. 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 $\alpha$ in Step 2. Then we initialize three vectors $w_0, z_0$ and $g_0$ 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 $\|w_k - z_{k-1}\| + \|w_{k-1} - z_{k-1}\|^8$.

---

4 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 $F(w_{k-1})$, not on $F(z_{k-1})$. 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 $\frac{1}{2} \| \cdot \|^2$ for $p \in (1, 2)$ may not have globally bounded Lipschitz continuous gradients.

In the following, we assume $w^*$ 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 $\hat{w}_K$ such that

$$
\sup_{w \in \mathcal{W}, \|\hat{w}_K - w\| \leq D} \langle F(\hat{w}_K), \hat{w}_K - w \rangle \leq C_0 D \|w_0 - w^*\| \sqrt{\frac{L}{A_{K-1} + a_1}},
$$

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

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

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

$$
\|\hat{w}_K - w^*\| \leq \frac{C_0}{\sigma} \|w_0 - w^*\| \sqrt{\frac{L}{A_{K-1} + a_1}}.
$$

**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), for all $K \geq 1$, after $K$ iterations, Algorithm 1 returns a $w_K$ such that

$$
\sup_{w \in \mathcal{W}, \|w_K - w\| \leq D} \langle F(w_K), w_K - w \rangle \leq C_0 D \|w_0 - w^*\| \sqrt{\frac{L}{a_{K-1}}},
$$

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

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

Meanwhile, we also have

$$
\|w_K - w^*\| \leq \frac{C_0}{\sigma} \|w_0 - w^*\| \sqrt{\frac{L}{a_{K-1}}}.
$$

6
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(ϵ−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, W) is monotone, then the regularized problem VIP(F + ϵ∇ 1/2 || · − w0 ||^2, W) satisfies Assumption 4 with σ = ϵ.

Proof. See Section C.6.

Due to Lemma 1, we can apply Theorem 1 and Proposition 1 to the regularized problem VIP(F + ϵ∇ 1/2 || · − w0 ||^2, W), and then obtain Corollaries 1 and 2 for the VIP(F, W), respectively.

Corollary 1 (Best iterate convergence in the monotone setting) Given w_0 ∈ W, let Assumptions 1 and 2 hold for the regularized problem VIP(F + ϵ∇ 1/2 || · − w0 ||^2, W). By optimizing the regularized problem by Algorithm 4, then the best iterate returned by Algorithm 4 satisfies

\[ \sup_{w \in W, ||w_0 - w|| \leq D, ||w - w_0|| \leq D} \langle F(\tilde{w}_K), \tilde{w}_K - w \rangle \leq \frac{L \epsilon}{(1 + \frac{\alpha \gamma}{L})^{K-1} - 1 + \frac{\alpha \gamma}{L}}, \]

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 ||w - w_0|| ≤ D in Corollary 1, which can be satisfied by choosing a large enough D. By Corollary 1, by choosing K = O(\frac{1}{2} log \frac{1}{\epsilon}), we will obtain an O(Dϵ)-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 w_0 ∈ W, let Assumptions 1 and 2 hold for the regularized problem VIP(F + ϵ∇ 1/2 || · − w0 ||^2, W). By optimizing the regularized problem by Algorithm 4, the last iterate of Algorithm 4 satisfies

\[ \sup_{w \in W, ||w_0 - w|| \leq D, ||w - w_0|| \leq D} \langle F(w_K), w_K - w \rangle \leq D \epsilon + DC_0 L ||w_0 - w^*|| \frac{1}{\alpha \gamma (1 + \frac{\alpha \gamma}{L})^{K-1}}, \]

where C_0 is defined in Theorem 1.
Algorithm 2 Stochastic Optimistic Dual Extrapolation

1: **Input:** Lipschitz constant $L > 0$ from Assumption 1, $\gamma, \delta > 0$ from Assumption 2. The VIP$(F, W)$ satisfying Assumption $3$ ($\sigma = 0$) or Assumption $4$ ($\sigma > 0$).
2: $A_0 = 0, \alpha = \min\{\frac{\gamma}{\delta}, \frac{1}{16}\}$.
3: $w_0 = z_0 \in W, g_0 = 0$.
4: for $k = 1, 2, 3, \ldots, K$ do
5: \hspace{1em} $a_k = \frac{\alpha \gamma \sqrt{1 + \sigma A_{k-1}}}{L}$, $A_k = A_{k-1} + a_k$.
6: \hspace{1em} $w_k = P_{z_{k-1}}\left(\frac{\alpha \gamma^2}{L \sigma a_k} F(w_{k-1}; \xi_{k-1})\right)$.
7: \hspace{1em} $g_k = g_{k-1} + a_k (F(w_k; \xi_k) - \frac{\sigma}{\delta} \nabla w_k \|w_k - w_0\|^2)$.
8: \hspace{1em} $z_k = P_{w_0}\left(\frac{1}{1 + \sigma a_k} g_k\right)$.
9: \hspace{1em} end for
10: $\tilde{w}_K = w_k$, 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** $\tilde{w}_K$.

Proof. See Section C.8.

Similar to Corollary 1 for best iterate, in Corollary 2, by choosing $K = O\left(\frac{1}{\epsilon} \log \frac{1}{\epsilon}\right)$, the last iterate will be an $O(1/\epsilon)$-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 $K = O\left(\frac{1}{\epsilon} \log \frac{1}{\epsilon}\right)$ with a prescribed accuracy parameter $\epsilon$. Thus our result does not contradict with the lower bound of last iterate [13].

Meanwhile, our proof only relies on the regularized problem VIP$(F + \epsilon \nabla \frac{1}{2\delta} \|w_0\|^2, W)$ satisfying Assumption $4$ with $\sigma = \epsilon$, which holds if the VIP$(F, W)$ is monotone. However, it is not necessary for the VIP$(F, W)$ to be monotone. For instance, if the VIP$(F, W)$ satisfies Assumption $3$ and $w_0 = w^*$, then the VIP$(F + \epsilon \nabla \frac{1}{2\delta} \|w_0\|^2, W)$ also satisfies Assumption $4$ with $\sigma = \epsilon$. Of course, letting $w_0 = w^*$ is impractical and we leave the more general setting of $w_0$ 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 $O\left(\frac{1}{\epsilon} \log \frac{1}{\epsilon}\right)$ 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 $\{F(w_k)\}$ by the unbiased stochastic estimations $\{F(w_k; \xi_k)\}$, 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. $\xi$ 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 \(\forall w \in W, F(w; \xi)\) is an unbiased estimation of $F(w)$ such that $\mathbb{E}_\xi[F(w; \xi)] = F(w)$; meanwhile the variance of $F(w; \xi)$ is bounded by $s^2$ such that $\mathbb{E}_\xi[\|F(w; \xi) - F(w)\|^2] \leq s^2$.

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 $\tilde{w}_K$ 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.

\[^9\]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 $w_K$ such that

$$
\mathbb{E}\left[ \sup_{w \in \mathcal{W}, \|\tilde{w}_K - w\| \leq D} (F(\tilde{w}_K), \tilde{w}_K - w) \right]
\leq \sqrt{2}(1 + \delta)LD \sqrt{\left( \frac{\|w^* - w_0\|^2}{8\alpha\gamma K} + \frac{s^2}{L^2} \right) + \frac{L^2}{2} \left( \frac{\|w^* - w_0\|^2}{8\alpha\gamma K} + \frac{s^2}{L^2} \right) + \frac{s^2}{2L^2}}.
$$

(16)

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

$$
\mathbb{E}[\|\tilde{w}_K - w^*\|^2] \leq \frac{32L^2}{\sigma^2(\alpha\gamma)^2(K + 1)^2} \left( \frac{8\alpha s^2 K}{L^2} + \frac{1}{2\gamma} \|w^* - w_0\|^2 \right).
$$

(17)

Proof. See Section D.4.

As shown 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 $O\left( \frac{1}{K} \right)$. Thus to guarantee the convergence of SOptDE, the variance should be $o(1)$, such as $s^2 = O\left( \frac{1}{K} \right)$. In the Euclidean setting that $\|\cdot\| := \|\cdot\|_2$, by the concentration inequality [39], to attain a variance of $O\left( \frac{1}{K} \right)$, we need $O(K)$ samples. Thus combining the setting $s^2 = O\left( \frac{1}{K} \right)$ 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 $\epsilon$-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, [18] has also considered variance reduction with a large batch size and used a “quadratic natural residual” (in our notation, it is $\mathbb{E}[\|w_k - z_{k-1}\|^2]$) 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 known 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” $\mathbb{E}[\|w_k - z_{k-1}\|^2]$ is no longer useful in deriving the theoretical guarantee by expected restricted strong merit function. As a result, we consider the term $\mathbb{E}[\|w_k - z_{k-1}\|^2 + \|w_{k-1} - z_{k-1}\|^2]$, 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 $O\left( \frac{1}{\epsilon^4} \right)$ 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 $O\left( \frac{1}{\epsilon^4} \right)$ result for strongly monotone variational inequalities, which can also be extended to the setting that $\sigma$-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**


