NEW ASPECTS OF BLACK BOX CONDITIONAL GRADIENT: VARIANCE REDUCTION AND ONE POINT FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper deals with the black-box optimization problem. In this setup, we do not have access to the gradient of the objective function, therefore, we need to estimate it somehow. We propose a new type of approximation JAGUAR, that memorizes information from previous iterations and requires $\mathcal{O}(1)$ oracle calls. We implement this approximation in the Frank-Wolfe and Gradient Descent algorithms and prove the convergence of these methods with different types of zero-order oracle. Our theoretical analysis covers scenarios of non-convex, convex and PL-condition cases. Also in this paper, we consider the stochastic minimization problem on the set Q with noise in the zero-order oracle; this setup is quite unpopular in the literature, but we prove that the JAGUAR approximation is robust not only in deterministic minimization problems, but also in the stochastic case. We perform experiments to compare our gradient estimator with those already known in the literature and confirm the dominance of our methods.

1 INTRODUCTION

The projection-free methods, such as Conditional Gradient, known as the Frank-Wolfe (FW) Frank & Wolfe (1956) algorithm, are widely used to solve various types of optimization problems. In the last decade, Conditional Gradient methods have attracted increasing interest in the machine learning community, since in many cases it is computationally cheaper to solve the linear minimization problem over the feasible convex set (e.g. l_p-balls or simplex Δ_d), than to do a projection into it LeBlanc et al. (1985); Jaggi (2011); Bubeck et al. (2015); Hazan et al. (2016); Goldfarb et al. (2017); Dadras et al. (2022); Freund et al. (2017).

In the original Frank-Wolfe paper Frank & Wolfe (1956), the authors used true gradient in their algorithm, but modern machine learning and artificial intelligence problems require the use of other gradient estimators due to the significant increase in dataset size and complexity of state-of-the-art models. Examples of such gradient estimators in FW-type algorithms include coordinate methods Lacoste-Julien et al. (2013); Wang et al. (2016); Osokin et al. (2016) and stochastic gradient approximation with batches Reddi et al. (2016); Zhang et al. (2020); Lu & Freund (2021).

But sometimes there are even more complicated setups where we can not compute the gradient in general because it is not available for various reasons, e.g. the target function is not differentiable or the computation of the gradient is computationally difficult Taskar et al. (2005); Chen et al. (2017); Nesterov & Spokoiny (2017). This setting is called black-box optimization Lian et al. (2015), and in this case we are forced to use zero-order (ZO) gradient estimation methods via finite differences of the objective function (sometimes with additional noise) to approximate the gradient Nesterov (2012); Duchi et al. (2012).

Over the last few years of research on the topic of black-box optimization, we can outline two main methods for approximating the gradient using finite differences. The first estimates the gradient in m coordinates Richtárik & Takáč (2014); Wright (2015); Nesterov & Stich (2017):

$$\frac{d}{m}\sum_{i\in I}\frac{f(x+\tau e_i)-f(x-\tau e_i)}{2\tau}e_i., \tag{1}$$

where $I \subset \overline{1,d} : |I| = m$, e_i is a standard basis vector in $\mathbb{R}^d$ and τ is a smoothing parameter.

This finite difference approximates the gradient in m coordinates and requires $\mathcal{O}(m)$ oracle calls. If m is small then this estimation would be inaccurate, if m is large then we need to make many zero-order oracle calls at each iteration. In the case of $m = d$ we call this method *full-approximation*.

Another finite difference does not use the standard basis, but random vectors e Duchi et al. (2012); Nesterov & Spokoiny (2017); Gasnikov et al. (2022b;a); Statkevich et al. (2024):

$$d\frac{f(x+\tau e)-f(x-\tau e)}{2\tau}e, \tag{2}$$

where e can be uniformly distributed on a l_p-sphere $RS_p^d(1)$, then this scheme is called l_p-*smoothing*. In recent papers, authors usually use $p = 1$ Gasnikov et al. (2016); Akhavan et al. (2022) or $p = 2$ Nemirovskij & Yudin (1983); Shamir (2017); Gorbunov et al. (2022). Alternatively, e could be sampled from normal distribution with zero mean and unit covariance matrix Nesterov & Spokoiny (2017).

The approximations (1) and (2) have a very large variance or require many calls to the zero-order oracle, therefore there is a necessity to somehow reduce the approximation error without increasing the number of oracle calls. In stochastic optimization, the method of remembering information from previous iterations is widely used, for example in SVRG Johnson & Zhang (2013), SAGA Defazio et al. (2014), SARAH Nguyen et al. (2017) and SEGA Hanzely et al. (2018) authors suggest memorizing the gradient from previous iterations for improve the convergence of the method. We decided to use this technique in a black-box optimization problem, memorizing the gradient approximations from previous iterations to reduce batch size without significant loss of accuracy.

In this concept paper, we seek to answer the following research questions:

• *Can we create a zero-order method that uses information from previous iterations and approximates the true gradient as accurately as the full-approximation* (1)*, but requires* $\mathcal{O}(1)$ *calls to the zero-order oracle?*

• *Can we implement this approximation method in the Frank-Wolfe algorithms for deterministic and stochastic settings of the minimization problems?*

• *Whether the convergence estimates of this method are better than for the difference schemes* (1) *and* (2)*?*

In a more realistic setting, the zero-order oracle returns a noisy value of the objective function, i.e. it yields not $f(x)$ but $f(x) + \delta(x)$. Different types of noises $\delta(\cdot)$ are considered in the literature: it can be stochastic Bach & Perchet (2016); Gorbunov et al. (2022); Akhavan et al. (2020); Gasnikov et al. (2022b) or deterministic Risteski & Li (2016); Bogolubsky et al. (2016); Lobanov et al. (2023b). This raises another research question:

• *How do different types of noise affect the theoretical guarantees and practical results of our proposed approaches?*

1.1 Our contributions

According to the research questions, our contributions can be summarized as follows:

• We present the JAGUAR method, which approximates a true gradient of the objective function $\nabla f(x)$ at the point x. By using the memory of previous iterations we achieve an accuracy close to the full-approximation (1), but JAGUAR does not require $\mathcal{O}(d)$, but $\mathcal{O}(1)$ calls to the zero-order oracle. l_p-smoothing (2) also requires $\mathcal{O}(1)$ oracle calls, but since there is no memory technique in it, this method has large variance and is not robust.

• We prove theoretical estimates for this method (see Section 2.1). We consider both deterministic and stochastic noises in the zero-order oracle. If the first setting is somehow obtained in the literature Lobanov et al. (2023a;b), the second is rarely considered by authors, therefore our method is suitable for various black-box optimization problems.

• We implement the JAGUAR approximation in the Frank-Wolfe algorithm for stochastic and deterministic minimization problems and prove its convergence in both cases (see Sections 2.2 and 2.4).

• We also implement JAGUAR in the Gradient Descent method and prove its convergence in the non-convex and PL-condition cases (see Section 2.3).

• We perform several computational experiments, comparing the JAGUAR approximation with l_2-smoothing (2) and full-approximation (1) on different minimization problems (see Section 3).

1.2 Related work

In this section, we compare the problem statements and approximation methods in the literature on zero-order methods in the Frank-Wolfe-based algorithms. Some authors consider coordinate methods Lacoste-Julien et al. (2013) these are also gradient approximations, but these methods use the true gradient of the observed function f, therefore we can not use them directly in black-box optimization. In general, the l_p-smoothing technique can approximate the gradient using $\mathcal{O}(1)$ oracle calls Dvinskikh et al. (2022), but it may not be robust in the Frank-Wolfe setting, since in Lobanov et al. (2023a) authors have to collect large batches of directions e to achieve convergence. Note that in Lobanov et al. (2023a) the non-stochastic noise is taken into account. The full-approximation is also used in the literature Sahu et al. (2019); Gao & Huang (2020); Akhtar & Rajawat (2022), but at each iteration we need to make $\mathcal{O}(d)$ oracle calls, and since d is huge in modern applications, this can be a problem. Also, this method requires the smoothness of the objective function f. We summarize and compare the problem statements, approximation methods and results for them in Table 1.

Table 1: Adjustment of the various zero-order and coordinate FW methods.

Method	Setting		Noise		Batch size	Approximation
	Smooth	Zero-order	Stochastic	Deterministic		
ZO-SCGS Lobanov et al. (2023a)	✗	✓	✗	✓	$\mathcal{O}\left(1/\varepsilon^2\right)$	l_2-smoothing (2)
FZFW Gao & Huang (2020)	✓	✓	✗	✗	$\mathcal{O}\left(\sqrt{d}\right)$	Full-approximation (1)
DZOFW Sahu et al. (2019)	✓	✓	✗	✗	$\mathcal{O}\left(d\right)$	Full-approximation (1)
MOST-FW Akhtar & Rajawat (2022)	✓	✓	✗	✗	$\mathcal{O}\left(d\right)$	Full-approximation (1)
BCFW Lacoste-Julien et al. (2013)	✓	✗	✗	✗	$\mathcal{O}\left(1\right)$	Coordinate
SSFW Beznosikov et al. (2023)	✓	✗	✗	✗	$\mathcal{O}\left(1\right)$	Coordinate
`FW via JAGUAR` (this paper)	✓	✓	✓	✓	$\mathcal{O}\left(1\right)$	`JAGUAR` (Algorithms 1 and 5)

2 Main results

In this paper, we consider the following optimization problem

$$f^* := \min_{x \in Q} \quad f(x). \tag{3}$$

We now provide several assumptions that are necessary for the analysis.

Assumption 1 (Compact domain)**.** *The set Q is compact and convex, i.e.* $\exists D > 0 : \ \forall x, y \in Q \hookrightarrow \|x - y\| \leq D$.

Assumption 2 (L-smoothness)**.** *The function $f(x)$ is L-smooth on the set Q, i.e.* $\exists\, L > 0 : \forall x, y \in Q \hookrightarrow \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.

Assumption 3 (Convex)**.** *The function $f(x)$ is convex on the set Q, i.e.* $\forall x, y \in Q \hookrightarrow f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.

We assume that we only have access to the zero-order oracle and that it returns the noisy value of the function $f(x)$: $f_\delta(x) := f(x) + \delta(x)$. Therefore, we make a common assumption about this noise Dvinskikh et al. (2022); Lobanov et al. (2023b). For stochastic case see Section 2.4.

Assumption 4 (Bounded oracle noise)**.** *The noise in the zero-order oracle is bounded by a constant $\Delta > 0$, i.e.* $\exists \Delta > 0 : \ \forall x \in Q \hookrightarrow |\delta(x)|^2 \leq \Delta^2$.

2.1 `JAGUAR` gradient approximation. Deterministic case

Above we reviewed the gradient approximation techniques using finite differences (1) and (2). In this section, we introduce the new gradient estimation technique `JAGUAR` (Algorithm 1), which is based on the already investigated methods and uses the memory of previous iterations.

Algorithm 1 `JAGUAR` gradient approximation. Deterministic case

1: **Input:** $x, h \in \mathbb{R}^d$
2: Sample $i \in \overline{1, d}$ independently and uniform
3: Compute $\widetilde{\nabla}_i f_\delta(x) = \frac{f_\delta(x+\tau e_i) - f_\delta(x-\tau e_i)}{2\tau} e_i$
4: $h = h - \langle h, e_i \rangle e_i + \widetilde{\nabla}_i f_\delta(x)$

The idea behind the `JAGUAR` method is similar to well-known variance reduction techniques such as SAGA Defazio et al. (2014) or SVRG Johnson & Zhang (2013). However, in zero-order

optimization we need to approximate the gradient, therefore we need to apply the variance reduction technique to the coordinates Hanzely et al. (2018). Consequently, the JAGUAR method uses the memory of some coordinates of the previous gradients instead of memorizing gradients by batches in past points.

There are already works in the literature that combine zero-order optimization and variance reduction, but the essence of these papers is that they change the gradient calculation to the gradient-free approximation (1) in the batch variance reduced algorithms such as SVRG or SPIDER Ji et al. (2019), rather than using the variance reduction technique for coordinates as in Algorithm 1.

JAGUAR approximation algorithm can be used with any iterative scheme that return a new point x^k at each step k. Using these points, we obtain the sequence h^k in line 4, which in a sense serves as a memory of the gradient components from past moments. Therefore, it makes sense to use h^k as an estimator of the true gradient $\nabla f(x^k)$ in incremental optimization methods. Using the following unified scheme, we can describe such an iterative algorithm that solves (3) (Algorithm 2).

In Algorithm 2, Proc(x^k, grad_est) is a sequence of actions that translates x^k to x^{k+1} using grad_estimator as the true gradient. Now we start to analyze JAGUAR gradient approximation (Algorithm 1). Our goal is to estimate the closeness of the true gradient $\nabla f(x^k)$ and the output of the JAGUAR algorithm h^k at step k.

Algorithm 2 Iterative algorithm using gradient estimator via JAGUAR

Input: same as for Proc and h^0
for $k = 0, 1, 2, ..., N$ **do**
  h^{k+1} = JAGUAR(x^k, h^k)
  x^{k+1} = Proc(x^k, grad_est = h^{k+1})
end for

Lemma 1. *For x^k and h^k, generated by Algorithm 2, the following inequality holds*

$$H_{k+1} \leq \Big(1 - \frac{1}{2d}\Big) H_k + 2d\mathbb{E}\Big[\Big\|\nabla f(x^{k+1}) - \nabla f(x^k)\Big\|^2\Big] + \mathbb{E}\Big[\Big\|\widetilde{\nabla} f_\delta(x^k) - \nabla f(x^k)\Big\|^2\Big], \quad (4)$$

where we use notations $H_k := \mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right]$ and

$$\widetilde{\nabla} f_\delta(x) := \sum_{i=1}^{d} \frac{f_\delta(x + \tau e_i) - f_\delta(x - \tau e_i)}{2\tau} e_i. \quad (5)$$

For a detailed proof of Lemma 1, see the proof of Lemma 5 in Appendix C.2 in the case of $\sigma_\nabla^2 = \sigma_f^2 = 0$ (see details in Section 2.4).

Discussion. We do not need to make any assumptions to satisfy Lemma 1, since in its proof we have only used the form of Algorithm 1. This means that the performance of the JAGUAR approximation depends only on the quality of the full-approximation $\widetilde{\nabla} f_\delta(x)$ and the closeness of the points x^{k+1} and x^k, generated by the Algorithm 2. According to Lemma 4 in Appendix C.1, we can estimate quality of the $\widetilde{\nabla} f_\delta(x)$: under Assumptions 2 and 4 for all $x \in Q$ it holds that $\|\widetilde{\nabla} f_\delta(x) - \nabla f(x)\|^2 \leq dL^2\tau^2 + (2d\Delta^2/\tau^2)$.

Let us analyse the formula (4), and show that using JAGUAR gradient approximation (Algorithm 1) gives us the same estimates as using $\widetilde{\nabla} f_\delta(x)$. Many optimization algorithms use a sufficiently small step size γ, i.e. Proc(x^k, grad_est) $\approx x^k$. Therefore we can assume that $\nabla f(x^{k+1}) \approx \nabla f(x^k)$ (for a specific choice of γ see Theorem 1). We can unroll (4) and, if we consider $k \gg d$, then we can obtain that $\mathbb{E}[\|h^k - \nabla f(x^k)\|^2] = \mathcal{O}(\|\widetilde{\nabla} f_\delta(x^k) - \nabla f(x^k)\|^2)$, i.e. it is the same estimate as for the full-approximation (5), however now we make $\mathcal{O}(1)$ oracle calls at each iteration.

2.2 Frank-Wolfe via JAGUAR

In this section, we introduce the Frank-Wolfe algorithm, witch solves the problem (3) using the JAGUAR gradient approximation (Algorithm 3). Using a certain form of the Proc function in Algorithm 3, we can unroll the results of Lemma 1 to carefully tune the step size γ_k.

Algorithm 3 FW via JAGUAR. Deterministic case

Input: $x^0 \in Q$, $h^0 = \widetilde{\nabla} f_\delta(x^0)$, γ_k, τ
for $k = 0, 1, 2, ..., N$ **do**
  $h^{k+1} = \text{JAGUAR}\left(x^k, h^k\right)$
  $s^k = \arg\min_{x \in Q} \left\langle s, h^{k+1} \right\rangle$
  $x^{k+1} = x^k + \gamma_k(s^k - x^k)$
end for

Theorem 1 (Step tuning for FW via JAGUAR (Algorithm 3). Deterministic case)**.** *Consider Assumptions 1 and 2. For h^k, generated by Algorithm 3, we can take $\gamma_k = 4/(k + 8d)$, then the following inequality holds:*

$$\mathbb{E}[\|h^k - \nabla f(x^k)\|^2] = \mathcal{O}\Big(\|\widetilde{\nabla} f_\delta(x^k) - \nabla f(x^k)\|^2 + \frac{d^2L^2D^2}{(k + 8d)^2}\Big).$$

From Theorem 1 we can conclude that after $\mathcal{O}\left(\sqrt{d}D/\tau\right)$ steps we get the same estimate as in the full-approximation (5). We now examine the convergence of Algorithm 3.

Theorem 2 (Convergence rate of `FW via JAGUAR` (Algorithm 3))**.** *Consider Assumptions 1, 2, 3, 4. If we take $\gamma_k = 4/(k+8d)$, then `FW via JAGUAR` (Algorithm 3) has the following convergence rate*

$$\mathbb{E}\left[f(x^N) - f^*\right] = \mathcal{O}\Big(\frac{d\max\{LD^2, F_0\}}{N+8d} + \sqrt{d}LD\tau + \frac{\sqrt{d}\Delta D}{\tau}\Big),$$

The results of Theorem 2 can be rewritten as an upper complexity bound on a number of iterations of Algorithm 3, using the appropriate smoothing parameter τ and noise bound Δ.

Corollary 1. *Under the conditions of Theorem 2, choosing $\gamma_k, \tau(\varepsilon), \Delta(\varepsilon)$ as in the Appendix B.1 in order to achieve an ε-approximate solution (in terms of $\mathbb{E}[f(x^N) - f^*] \leq \varepsilon$) it takes*

$$\mathcal{O}\Big(\frac{d\max\{LD^2, F_0\}}{\varepsilon}\Big) \text{ iterations of Algorithm 3.}$$

For a detailed proof of Theorems 1, 2 and Corollaries 1, 6, see Appendix D.1.

Discussion. The results of Theorem 2 match with results Frank & Wolfe (1956); Jaggi (2013) in which the authors used a true gradient and got the result of the form $\mathbb{E}\left[f(x^N) - f^*\right] = \mathcal{O}\left(\max\{LD^2; f(x^0) - f^*\}/N\right)$. In the zero-order case, terms of the form $\mathcal{O}\left(\text{poly}(\tau) + \text{poly}(\Delta/\tau)\right)$ are inevitable, since they are crucial for the approximation of the true gradient and always affect the convergence of zero-order methods Risteski & Li (2016); Sahu et al. (2018); Liu et al. (2018); Beznosikov et al. (2020a). The factor d, which appears in our theoretical estimates compared to the first-order result, is related to the zero-order structure of the method.

In Corollary 1, the dependence $\Delta(\varepsilon)$ was obtained, which may seem incorrect, because usually the maximum noise is given by nature and we cannot reduce it. In this case, we should rewrite the dependence in the form $\varepsilon = \varepsilon(\Delta)$ and accordingly τ and N start to depend on Δ, not on ε (see Appendix B.1). In the rest of the corollaries in this paper, we will write the dependence as $\Delta(\varepsilon)$ for ease of presentation, but they can always be rewritten in terms of $\varepsilon(\Delta)$.

2.3 Gradient Descent via JAGUAR

In this section, we consider problem of the form

$$f^* := \min_{x \in \mathbb{R}^d} f(x).$$

We can not use FW-type algorithms in this setting because $Q = \mathbb{R}^d$ is not a bounded set, therefore, we consider Gradient Descent with `JAGUAR` approximation. We now examine the convergence of Algorithm 4.

Algorithm 4 `GD via JAGUAR`

1: **Input:** $x_0 \in \mathbb{R}^d, h^0 = \widetilde{\nabla} f_\delta(x^0), \gamma, \tau$
2: **for** $k = 0, 1, 2, ..., N$ **do**
3: $\quad h^{k+1} = \text{JAGUAR}\left(x^k, h^k\right)$
4: $\quad x^{k+1} = x^k - \gamma h^{k+1}$
5: **end for**

Theorem 3 (Convergence rate of `GD via JAGUAR` (Algorithm 4). Non-convex case)**.** *Consider Assumptions 2 and 4. If we take $\gamma \equiv 1/(4dL)$, then `GD via JAGUAR` (Algorithm 4) in the non-convex case has the following convergence rate*

$$\mathbb{E}\left[\|\nabla f(\widehat{x}_N)\|^2\right] = \mathcal{O}\left(\frac{dL\Phi_0}{N+1} + dL^2\tau^2 + \frac{d\Delta^2}{\tau^2}\right),$$

where $\widehat{x}_N$ is chosen uniformly from $\{x^k\}_{k=0}^N$ and $\Phi_0 := f(x^0) - f^* + d\gamma\left\|h^0 - \nabla f(x^0)\right\|^2$.

Corollary 2. *Under the conditions of Theorem 3, choosing γ, τ, Δ as in Appendix B.2 in order to achieve an ε-approximate solution (in terms of $\mathbb{E}[\|\nabla f(\widehat{x}_N)\|^2] \leq \varepsilon^2$, where $\widehat{x}_N$ is chosen uniformly from $\{x^k\}_{k=0}^N$) it takes*

$$\mathcal{O}\left(\frac{dL\Phi_0}{\varepsilon^2}\right) \text{ iterations of Algorithm 4.}$$

Under the PL-condition we explore a better convergence rate.

Assumption 5 (PL condition)**.** *The function $f(x)$ satisfies PL-condition if $\exists\, \mu > 0 : \forall x \in \mathbb{R}^d \hookrightarrow \|\nabla f(x)\|_2^2 \geq 2\mu(f(x) - f^*)$.*

Theorem 4 (Convergence rate of `GD via JAGUAR` (Algorithm 4) PL-case)**.** *Consider Assumptions 2, 4 and 5. If we take $\gamma \equiv 1/(4dL)$, then `GD via JAGUAR` (Algorithm 4) in the PL-condition case (Assumption 5) has the following convergence rate*

$$\mathbb{E}\left[f(x^k) - f^*\right] = \mathcal{O}\left(F_0 \exp\left[-\frac{\mu N}{4dL}\right] + \frac{dL^2\tau^2 + d\Delta^2/\tau^2}{\mu}\right),$$

Corollary 3. *Under the conditions of Theorem 4, choosing γ, τ, Δ as in Appendix B.2 in order to achieve an ε-approximate solution (in terms of $\mathbb{E}[f(x^N) - f^*] \leq \varepsilon$) it takes*

$$\mathcal{O}\left(\frac{dL}{\mu}\log\left[\frac{F_0}{\varepsilon}\right]\right) \text{ iterations of Algorithm 4.}$$

For a detailed proof of Theorems 3, 4 and Corollaries 2, 3, see Appendix E.

Discussion. The results of Theorems 3 and 4 match with first-order results, but in such results $\gamma \equiv 1/L$, therefore, in the zero-order case the factor d reappears in theoretical estimates. We also obtain smoothing term of the form $\mathcal{O}\left(\text{poly}(\tau) + \text{poly}(\Delta/\tau)\right)$. The same estimates arise in previous works when studying Gradient Descent in the zero-order oracle setting Bogolubsky et al. (2016); Bayandina et al. (2018); Dvinskikh et al. (2022).

2.4 Frank-Wolfe via JAGUAR. Stochastic case

In this section, we consider the stochastic version of the problem (3):

$$f(x) := \mathbb{E}_{\xi\sim\pi}[f(x,\xi)], \tag{6}$$

where ξ is a random vector from a mostly unknown distribution π. For this problem we can not use the values of the function $f(x)$ in the difference schemes, since only $f(x,\xi)$ is available. Again, we assume that we do not have access to the true value of the gradient $\nabla f(x,\xi)$, and zero-order oracle returns the noisy value of the function $f(x,\xi)$: $f_\delta(x,\xi) := f(x,\xi) + \delta(x,\xi)$.

In the stochastic setup (6), there are two versions of the differences of the scheme (1) appear. The first one is called Two Point Feedback (TPF) Duchi et al. (2015); Shamir (2017); Gorbunov et al. (2022); Beznosikov et al. (2020b); Gasnikov et al. (2022b). In this case, we define such gradient approximations of the function $f(x)$:

$$\widetilde{\nabla}_i f_\delta(x,\xi) := \frac{f_\delta(x+\tau e_i,\xi) - f_\delta(x-\tau e_i,\xi)}{2\tau} e_i. \tag{7}$$

The second one is called One Point Feedback (OPF) Nemirovskij & Yudin (1983); Flaxman et al. (2004); Gasnikov et al. (2017); Akhavan et al. (2020); Beznosikov et al. (2021). In this case, we define a slightly different gradient approximation of the function $f(x)$:

$$\widetilde{\nabla}_i f_\delta(x,\xi^\pm) := \frac{f_\delta(x+\tau e_i,\xi^+) - f_\delta(x-\tau e_i,\xi^-)}{2\tau} e_i. \tag{8}$$

The main difference between the approximations (7) and (8) is that the scheme (7) is more accurate, but difficult to implement in practice, because we have to get the same realization of ξ at two different points $x+\tau e$ and $x-\tau e$, hence the scheme (8) is more interesting from a practical point of view. To further simplify, we consider that in the case of two point feedback (7) we have the same inscription as for the one point feedback (8), but only $\xi^+ = \xi^- = \xi$. We provide several assumptions required for the analysis.

Assumption 6 (Smoothness)**.** *The functions $f(x,\xi)$ are $L(\xi)$-smooth on the set Q, i.e. $\forall x, y \in Q \hookrightarrow \|\nabla f(x,\xi) - \nabla f(y,\xi)\| \leq L(\xi)\|x-y\|$. We also assume that exists constant L such that $L^2 := \mathbb{E}[L(\xi)^2]$. Then function $f(x)$ is L-smooth on the set Q.*

Assumption 7 (Bounded oracle noise)**.** *The noise in the oracle is bounded by some constant $\Delta > 0$, i.e. $\exists\, \Delta > 0 :\ \forall x \in Q \hookrightarrow \mathbb{E}[|\delta(x,\xi)|^2] \leq \Delta^2$. If we define $\delta(x) := \mathbb{E}[\delta(x,\xi)]$, then it holds that $|\delta(x)|^2 \leq \Delta^2$.*

Now we present two assumptions that are only needed in the stochastic case.

Assumption 8 (Bounded variance of gradient)**.** *The variance of the $\nabla f(x,\xi)$ is bounded, i.e. $\exists\, \sigma_\nabla \geq 0 :\ \forall x \in Q \hookrightarrow \mathbb{E}[\|\nabla f(x,\xi) - \nabla f(x)\|^2] \leq \sigma_\nabla^2$.*

Assumption 9 (Bounded variance of function)**.** *The variance of the $f(x,\xi)$ is bounded, i.e. $\exists\, \sigma_f \geq 0 :\ \forall x \in Q \hookrightarrow \mathbb{E}[\|f(x,\xi) - f(x)\|^2] \leq \sigma_f^2$.*

Assumptions 8 and 9 are classical in the stochastic optimization literature Agarwal et al. (2011); Bach & Perchet (2016); Akhavan et al. (2020); Dvurechensky et al. (2021). In the case of two point feedback (7), we do not need Assumption 9, therefore, for the sake of simplicity we will assume that in this case Assumption 9 is satisfied with $\sigma_f = 0$. Now we can present the Frank-Wolfe algorithm, witch solves the problem (3) + (6) using `JAGUAR` gradient approximation.

Algorithm 5 `FW via JAGUAR`. Stochastic case

1: **Input:** $\gamma_k, \eta_k, \tau, x^0 \in Q, h^0 = g^0 = 1/(2\tau)\sum_{i=1}^d (f_\delta(x^0 + \tau e_i, \xi_i^+) - f_\delta(x^0 - \tau e_i, \xi_i^-))e_i$
2: **for** $k = 0, 1, 2, ..., N$ **do**
3: Sample $i_k \in \overline{1, d}$ independently and uniform
4: Sample $\xi_k^+, \xi_k^- \sim \pi$ independently (in TPF $\xi_k^+ = \xi_k^-$)
5: Compute $\widetilde{\nabla}_{i_k} f_\delta(x^k, \xi_k^\pm) = 1/(2\tau)(f_\delta(x^k + \tau e_{i_k}, \xi_k^+) - f_\delta(x^k - \tau e_{i_k}, \xi_k^-))e_{i_k}$
6: $h^{k+1} = h^k - \langle h^k, e_{i_k} \rangle e_{i_k} + \widetilde{\nabla}_{i_k} f_\delta(x^k, \xi_k^+, \xi_k^-)$
7: $\rho^k = h^k - d \cdot \langle h^k, e_{i_k} \rangle e_{i_k} + d \cdot \widetilde{\nabla}_{i_k} f_\delta(x^k, \xi_k^+, \xi_k^-)$
8: $g^{k+1} = (1 - \eta_k) g^k + \eta_k \rho^k$
9: $s^k = \arg\min_{s \in Q} \langle s, g^{k+1} \rangle$
10: $x^{k+1} = x^k + \gamma_k (s^k - x^k)$
11: **end for**

Algorithm 5 is similar to `FW via JAGUAR` in the deterministic case (Algorithm 3), but in lines 7 and 8 we use SEGA Hanzely et al. (2018) and momentum Mokhtari et al. (2020) parts to converge in the stochastic case.

• We need the SEGA part Hanzely et al. (2018) ρ_k in Algorithm 5, because in the stochastic case we care about the "unbiased" property (see proof of Lemma 7 in Appendix C), i.e.

$$\mathbb{E}_k[\rho^k] = \widetilde{\nabla} f_\delta(x^k) := \sum_{i=1}^d \frac{f_\delta(x + \tau e_i) - f_\delta(x - \tau e_i)}{2\tau} e_i,$$

where $\mathbb{E}_k[\cdot]$ is a conditional mathematical expectation at step k. Using the SEGA part ρ^k degrades our estimates by a factor of d compared to using h^k as a gradient approximation (see Lemmas 5 and 6 in Appendix C), but we have to accept this factor.

• We need the momentum part Mokhtari et al. (2020) η_k in Algorithm 5, because when evaluating the expression of $\mathbb{E}[\|\widetilde{\nabla} f_\delta(x, \xi_{\overline{1,d}}^\pm) - \nabla f(x)\|^2]$ in the stochastic case, terms containing σ_∇^2 appear and σ_f^2 and interfere with convergence (see Lemma 4 in Appendix C). This is a common problem in the stochastic Frank-Wolfe-based methods (see Mokhtari et al. (2020)). We can provide a theorem similar to the one in Section 2.2 (Theorem 1), where we carefully choose the step sizes γ_k and η_k.

Theorem 5 (Step tuning for `FW via JAGUAR`. Stochastic case)**.** *Consider Assumptions 1, 6, 7, 8 and 9 in the one point feedback case. For x^k generated by Algorithm 5, we can take $\gamma_k = 4/(k + 8d^{3/2})$ and $\eta_k = 4/(k + 8d^{3/2})^{2/3}$, then, the following inequality holds:*

$$G_k = \mathcal{O}\Big(\Big\| \widetilde{\nabla} f_\delta(x^k) - \nabla f(x^k) \Big\|^2 + \frac{L^2 D^2 + d^2 \sigma_f^2 / \tau^2 + d^2 \sigma_\nabla^2}{(k + 8d^{3/2})^{2/3}} \Big),$$

where we use the notation $G_k := \mathbb{E}[\|g^k - \nabla f(x^k)\|^2]$. In the two point feedback case, $\sigma_f^2 = 0$.

We get worse estimates compared to the deterministic case in Theorem 1 because we consider a more complicated setup. We now examine the convergence of Algorithm 5.

Theorem 6 (Convergence rate of `FW via JAGUAR` (Algorithm 5). Stochastic case)**.** *Consider Assumptions 1, 3, 6, 7, 8 and 9 in the one point feedback case. We can take $\gamma_k = 4/(k + 8d^{3/2})$ and $\eta_k = 4/(k + 8d^{3/2})^{2/3}$, then we `FW via JAGUAR` (Algorithm 5) has the following convergence rate*

$$F_N = \mathcal{O}\Big(\frac{LD^2 + d\sigma_f D/\tau + d\sigma_\nabla D + \sqrt{d} F_0}{(N + 8d^{3/2})^{1/3}} + \sqrt{d} L D \tau + \frac{\sqrt{d} \Delta D}{\tau} \Big),$$

where we use the notation $F_N := \mathbb{E}[f(x^N) - f^*]$. In the two point feedback case, $\sigma_f^2 = 0$.

Corollary 4. *Under the conditions of Theorem 6, choosing $\gamma_k, \eta_k, \tau, \Delta$ as in Appendix B.3 in order to achieve an ε-approximate solution (in terms of $\mathbb{E}[f(x^N) - f^*] \leq \varepsilon$) it takes*

$$\mathcal{O}\left(\max\left\{ \left[\frac{LD^2 + d\sigma_\nabla D + \sqrt{d}(f(x^0) - f^*)}{\varepsilon} \right]^3 ; \frac{d^{9/2}\sigma_f^3 L^3 D^6}{\varepsilon^6} \right\} \right) \text{ iterations of Algorithm 5.}$$

In the two point feedback case, $\sigma_f^2 = 0$ and the last equation takes form

$$\mathcal{O}\left(\left[\frac{LD^2 + d\sigma_\nabla D + \sqrt{d}(f(x^0) - f^*)}{\varepsilon} \right]^3 \right) \text{ iterations of Algorithm 5.}$$

For a detailed proof of Theorems 5, 6 and Corollary 4, see Appendix D.2.

Discussion. Since we used SEGA and momentum parts in the JAGUAR approximation algorithm (Algorithm 5) we do not get the same convergence rate as in Theorems 1 and 2, even if we switch from stochastic to deterministic setups, i.e, when we set $\sigma_\Delta = \sigma_f = 0$ in Theorems 5 and 6. The same problems arise in the first-order case Mokhtari et al. (2020); Zhang et al. (2020), due to the difficulty of implementing the stochastic gradient in FW-type algorithms.

We can apply the deterministic JAGUAR method (Algorithm 1) to the stochastic problem (6) and obtain the same estimates as in Theorems 1 and 2, only the smoothed term of the form $\mathcal{O}\left(\text{poly}(\tau) + \text{poly}(\Delta/\tau)\right)$ will contain summands of the form $\mathcal{O}\left(\text{poly}(\sigma_\Delta^2) + \text{poly}(\sigma_f^2/\tau)\right)$ Therefore, if $\sigma_\Delta^2, \sigma_f^2 \sim \Delta$, then deterministic Algorithm 1 is suitable for the stochastic problem (6). However, this means that we need to use big batches, therefore, we forced to use SEGA and momentum parts in the JAGUAR approximation.

3 EXPERIMENTS

In this section, we present and discuss the results of our experiments on the application of JAGUAR gradient approximation to a variety of black-box optimization problems. Our results include optimization with Frank-Wolfe (FW) and Gradient Descent (GD) algorithms. In FW we consider both deterministic and stochastic cases, in GD we consider only the deterministic case.

3.1 EXPERIMENT SETUP

We perform optimization on classification tasks with the SVM model on the set Q of the form:

$$\min_{w \in Q, b \in \mathbb{R}} \left\{ f(w, b) = \frac{1}{m} \sum_{k=1}^{m} \left(1 - y_k[(Xw)_k - b]\right)_+ + \frac{1}{2C}\|w\|^2 \right\}. \tag{9}$$

We also consider a logistic regression model on the set Q of the form:

$$\min_{w \in Q} \left\{ f(w) = \frac{1}{m} \sum_{k=1}^{m} \log\left(1 + \exp\left[-y_k (Xw)_k\right]\right) + \frac{1}{2C}\|w\|^2 \right\}. \tag{10}$$

In both problems we use the regularization term $C = 10$. As minimization set Q we consider the simplex Δ_d and l_2-ball in the Frank-Wolfe algorithm and $\mathbb{R}^d$ in Gradient Descent algorithm. For the classification problem we use the classical datasets MNIST Deng (2012) and Mushrooms Chang & Lin (2011). We incorporate different approximation methods into optimization algorithms, solving (9) and (10) problems and show that the algorithm that uses the JAGUAR method (Algorithms 1 and 5) performs best. We consider l_2-smoothing (2) and full-approximation (1) as baseline estimators of the gradients. The results of the experiments are provided in the Apppendix **??**.

3.2 Deterministic Frank-Wolfe

In this section, we consider deterministic noise of the form $f_\delta(x) = \text{round}(f(x), 5)$, i.e. we round all values of the function f to the fifth decimal place. Figure 1 shows the convergence over zero-order oracle calls of the deterministic FW algorithm. FW via JAGUAR (3 algorithm) shows better results than the baseline algorithms. This observation confirms our theoretical findings.

3.3 Stochastic Frank-Wolfe

In this section, we consider stochastic noise of the form $f_\delta(x, \xi) = f(x) + \xi; \xi \sim \mathcal{N}(0, 0.1)$. Figure 2 shows the convergence over zero-order oracle calls of the stochastic FW algorithm. Our theoretical findings are supported by observations. FW via JAGUAR (5 algorithm) is robust to noise and outperforms the baseline algorithms.

3.4 Gradient descent

In this section, we again consider deterministic noise of the from $f_\delta(x) = \text{round}(f(x), 5)$. Figure 3 shows convergence over zero-order oracle calls of the deterministic GD algorithm. GD via JAGUAR (algorithm 4) outperforms the baseline algorithms, albeit by a small margin, and this observation confirms our theoretical findings.

4 Conclusion

This paper introduces the JAGUAR algorithm, a novel gradient approximation method designed for black-box optimization problems, utilizing memory from previous iterations to estimate the true gradient with high accuracy while requiring only $\mathcal{O}(1)$ oracle calls. The study provides rigorous theoretical proofs and extensive experimental validation, demonstrating JAGUAR superior performance in both deterministic and stochastic settings. Key contributions include convergence proofs for the Frank-Wolfe and Gradient Descent algorithms, with detailed theorems establishing convergence rates. Experimental results show that JAGUAR outperforms baseline methods in SVM and logistic regression optimization problems. The results highlight JAGUAR efficiency and accuracy, making it a promising approach for future research and applications in zero-order optimization.

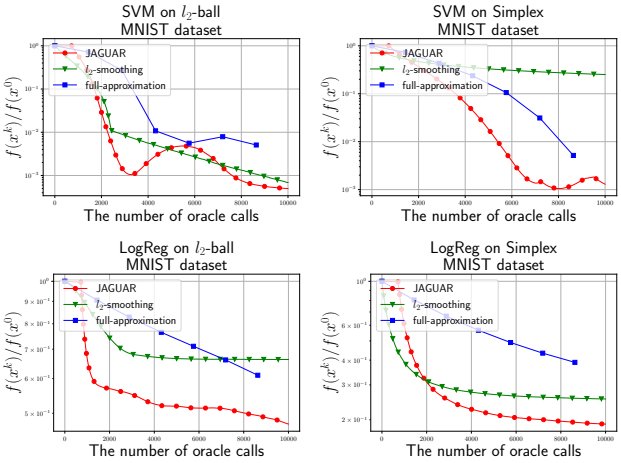

Figure 1: Deterministic FW Algorithm.

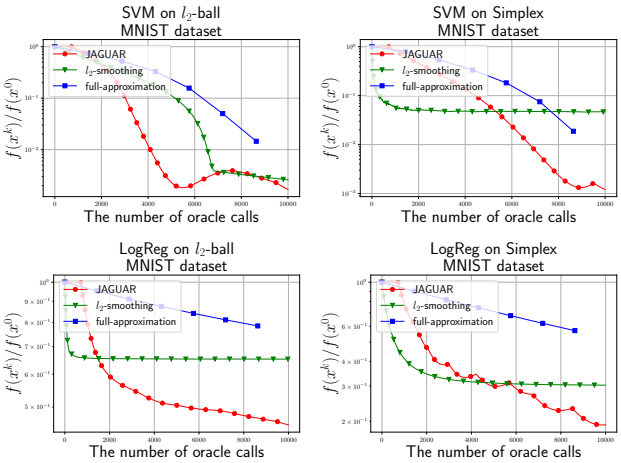

Figure 2: Stochastic FW Algorithm.

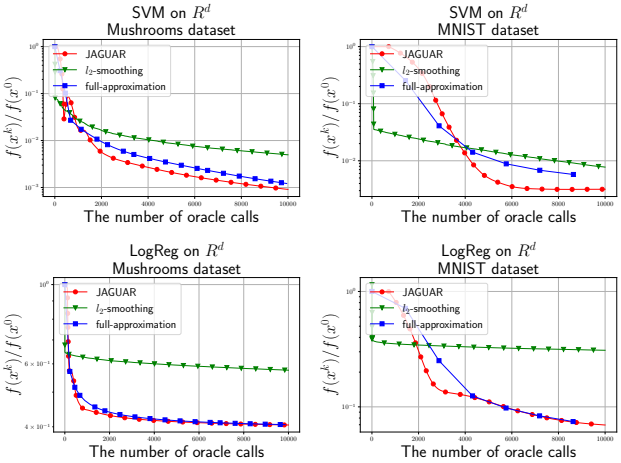

Figure 3: GD Algorithm.

REFERENCES

Alekh Agarwal, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. *Advances in Neural Information Processing Systems*, 24, 2011.

Arya Akhavan, Massimiliano Pontil, and Alexandre Tsybakov. Exploiting higher order smoothness in derivative-free optimization and continuous bandits. *Advances in Neural Information Processing Systems*, 33:9017–9027, 2020.

Arya Akhavan, Evgenii Chzhen, Massimiliano Pontil, and Alexandre Tsybakov. A gradient estimator via l1-randomization for online zero-order optimization with two point feedback. *Advances in Neural Information Processing Systems*, 35:7685–7696, 2022.

Zeeshan Akhtar and Ketan Rajawat. Zeroth and first order stochastic frank-wolfe algorithms for constrained optimization. *IEEE Transactions on Signal Processing*, 70:2119–2135, 2022.

Francis Bach and Vianney Perchet. Highly-smooth zero-th order online optimization. In *Conference on Learning Theory*, pp. 257–283. PMLR, 2016.

Anastasia Sergeevna Bayandina, Alexander V Gasnikov, and Anastasia A Lagunovskaya. Gradient-free two-point methods for solving stochastic nonsmooth convex optimization problems with small non-random noises. *Automation and Remote Control*, 79:1399–1408, 2018.

Aleksandr Beznosikov, Eduard Gorbunov, and Alexander Gasnikov. Derivative-free method for composite optimization with applications to decentralized distributed optimization. *IFAC-PapersOnLine*, 53(2):4038–4043, 2020a.

Aleksandr Beznosikov, Abdurakhmon Sadiev, and Alexander Gasnikov. Gradient-free methods with inexact oracle for convex-concave stochastic saddle-point problem. In *International Conference on Mathematical Optimization Theory and Operations Research*, pp. 105–119. Springer, 2020b.

Aleksandr Beznosikov, Vasilii Novitskii, and Alexander Gasnikov. One-point gradient-free methods for smooth and non-smooth saddle-point problems. In *Mathematical Optimization Theory and Operations Research: 20th International Conference, MOTOR 2021, Irkutsk, Russia, July 5–10, 2021, Proceedings 20*, pp. 144–158. Springer, 2021.

Aleksandr Beznosikov, David Dobre, and Gauthier Gidel. Sarah frank-wolfe: Methods for constrained optimization with best rates and practical features. *arXiv preprint arXiv:2304.11737*, 2023.

Lev Bogolubsky, Pavel Dvurechenskii, Alexander Gasnikov, Gleb Gusev, Yurii Nesterov, Andrei M Raigorodskii, Aleksey Tikhonov, and Maksim Zhukovskii. Learning supervised pagerank with gradient-based and gradient-free optimization methods. *Advances in neural information processing systems*, 29, 2016.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 15–26, 2017.

Ali Dadras, Karthik Prakhya, and Alp Yurtsever. Federated frank-wolfe algorithm. In *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*, 2022.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Darina Dvinskikh, Vladislav Tominin, Iaroslav Tominin, and Alexander Gasnikov. Noisy zeroth-order optimization for non-smooth saddle point problems. In *International Conference on Mathematical Optimization Theory and Operations Research*, pp. 18–33. Springer, 2022.

Pavel Dvurechensky, Eduard Gorbunov, and Alexander Gasnikov. An accelerated directional derivative method for smooth stochastic convex optimization. *European Journal of Operational Research*, 290(2):601–621, 2021.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Robert M Freund, Paul Grigas, and Rahul Mazumder. An extended frank–wolfe method with "in-face" directions, and its application to low-rank matrix completion. *SIAM Journal on optimization*, 27 (1):319–346, 2017.

Hongchang Gao and Heng Huang. Can stochastic zeroth-order frank-wolfe method converge faster for non-convex problems? In *International conference on machine learning*, pp. 3377–3386. PMLR, 2020.

Alexander Gasnikov, Anastasia Lagunovskaya, Ilnura Usmanova, and Fedor Fedorenko. Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex. *Automation and Remote Control*, 77:2018–2034, 2016.

Alexander Gasnikov, Darina Dvinskikh, Pavel Dvurechensky, Eduard Gorbunov, Aleksander Beznosikov, and Alexander Lobanov. Randomized gradient-free methods in convex optimization. *arXiv preprint arXiv:2211.13566*, 2022a.

Alexander Gasnikov, Anton Novitskii, Vasilii Novitskii, Farshed Abdukhakimov, Dmitry Kamzolov, Aleksandr Beznosikov, Martin Takac, Pavel Dvurechensky, and Bin Gu. The power of first-order smooth optimization for black-box non-smooth problems. In *International Conference on Machine Learning*, pp. 7241–7265. PMLR, 2022b.

Alexander V Gasnikov, Ekaterina A Krymova, Anastasia A Lagunovskaya, Ilnura N Usmanova, and Fedor A Fedorenko. Stochastic online optimization. single-point and multi-point non-linear multi-armed bandits. convex and strongly-convex case. *Automation and remote control*, 78:224–234, 2017.

Donald Goldfarb, Garud Iyengar, and Chaoxu Zhou. Linear convergence of stochastic frank wolfe variants. In *Artificial Intelligence and Statistics*, pp. 1066–1074. PMLR, 2017.

Eduard Gorbunov, Pavel Dvurechensky, and Alexander Gasnikov. An accelerated method for derivative-free smooth stochastic convex optimization. *SIAM Journal on Optimization*, 32(2):1210–1238, 2022. doi: 10.1137/19M1259225. URL `https://doi.org/10.1137/19M1259225`.

Filip Hanzely, Konstantin Mishchenko, and Peter Richtárik. Sega: Variance reduction via gradient sketching. *Advances in Neural Information Processing Systems*, 31, 2018.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Martin Jaggi. Sparse convex optimization methods for machine learning. 2011.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International conference on machine learning*, pp. 427–435. PMLR, 2013.

Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International conference on machine learning*, pp. 3100–3109. PMLR, 2019.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate frank-wolfe optimization for structural svms. In *International Conference on Machine Learning*, pp. 53–61. PMLR, 2013.

Larry J LeBlanc, Richard V Helgason, and David E Boyce. Improved efficiency of the frank-wolfe algorithm for convex network programs. *Transportation Science*, 19(4):445–462, 1985.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International conference on machine learning*, pp. 6286–6295. PMLR, 2021.

Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. *Advances in neural information processing systems*, 28, 2015.

Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Aleksandr Lobanov, Anton Anikin, Alexander Gasnikov, Alexander Gornov, and Sergey Chukanov. Zero-order stochastic conditional gradient sliding method for non-smooth convex optimization. *arXiv preprint arXiv:2303.02778*, 2023a.

Aleksandr Lobanov, Andrew Veprikov, Georgiy Konin, Aleksandr Beznosikov, Alexander Gasnikov, and Dmitry Kovalev. Non-smooth setting of stochastic decentralized convex optimization problem over time-varying graphs. *Computational Management Science*, 20(1):48, 2023b.

Haihao Lu and Robert M Freund. Generalized stochastic frank–wolfe algorithm with stochastic "substitute" gradient for structured convex optimization. *Mathematical Programming*, 187(1): 317–349, 2021.

Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *The Journal of Machine Learning Research*, 21(1):4232–4280, 2020.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Yu. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. doi: 10.1137/100802001. URL `https://doi.org/10.1137/100802001`.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

Yurii Nesterov and Sebastian U. Stich. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017. doi: 10.1137/16M1060182. URL `https://doi.org/10.1137/16M1060182`.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pp. 2613–2621. PMLR, 2017.

Anton Osokin, Jean-Baptiste Alayrac, Isabella Lukasewitz, Puneet Dokania, and Simon Lacoste-Julien. Minding the gaps for block frank-wolfe optimization of structured svms. In *international conference on machine learning*, pp. 593–602. PMLR, 2016.

Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic frank-wolfe methods for nonconvex optimization. In *2016 54th annual Allerton conference on communication, control, and computing (Allerton)*, pp. 1244–1251. IEEE, 2016.

Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1):1–38, 2014.

Andrej Risteski and Yuanzhi Li. Algorithms and matching lower bounds for approximately-convex optimization. *Advances in Neural Information Processing Systems*, 29, 2016.

Anit Kumar Sahu, Dusan Jakovetic, Dragana Bajovic, and Soummya Kar. Distributed zeroth order optimization over random networks: A kiefer-wolfowitz stochastic approximation approach. In *2018 IEEE Conference on Decision and Control (CDC)*, pp. 4951–4958. IEEE, 2018.

Anit Kumar Sahu, Manzil Zaheer, and Soummya Kar. Towards gradient free and projection free stochastic optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3468–3477. PMLR, 2019.

Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *The Journal of Machine Learning Research*, 18(1):1703–1713, 2017.

Ekaterina Statkevich, Sofiya Bondar, Darina Dvinskikh, Alexander Gasnikov, and Aleksandr Lobanov. Gradient-free algorithm for saddle point problems under overparametrization. *Chaos, Solitons & Fractals*, 185:115048, 2024.

Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd international conference on Machine learning*, pp. 896–903, 2005.

Yu-Xiang Wang, Veeranjaneyulu Sadhanala, Wei Dai, Willie Neiswanger, Suvrit Sra, and Eric Xing. Parallel and distributed block-coordinate frank-wolfe algorithms. In *International Conference on Machine Learning*, pp. 1548–1557. PMLR, 2016.

Stephen J Wright. Coordinate descent algorithms. *Mathematical programming*, 151(1):3–34, 2015.

Mingrui Zhang, Zebang Shen, Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. One sample stochastic frank-wolfe. In *International Conference on Artificial Intelligence and Statistics*, pp. 4012–4023. PMLR, 2020.

Supplementary Material

CONTENTS

B Full versions of Corollaries from Section 2 **17**

B.1 Full version of Corollary 1 17

B.2 Full version of Corollaries 2 and 3 18

B.3 Full version of Corollary 4 18

C Proof of converge rate of `JAGUAR`. Deterministic and stochastic cases. **19**

C.1 Result for the full-approximation (stochastic case) 19

C.2 Result for the `JAGUAR` approximation (Algorithm 1) (stochastic case) 20

C.3 Result for the SAGA approximation (line 7 of Algorithm 3) 22

C.4 Result for the `JAGUAR` approximation (Algorithm 3) 23

D Proof of converge rate of `FW via JAGUAR` **24**

D.1 Results for the deterministic case (Algorithm 3) 24

D.2 Results for the stochastic case (Algorithm 5) 25

E Proof of converge rate of `GD via JAGUAR` (Algorithm 4) **28**

E.1 Results for the non-convex case 28

E.2 Results for the PL-condition case 29

A Auxiliary Lemmas and Facts

In this section we list auxiliary facts and our results that we use several times in our proofs.

A.1 Squared norm of the sum

For all $x_1, ..., x_n \in \mathbb{R}^n$, where $n \in \{2, 4\}$

$$\|x_1 + x_2 + ... + x_n\|^2 \leq n \|x_1\|^2 + ... + n \|x_n\|^2 .$$

A.2 Cauchy–Schwarz inequality

For all $x, y \in \mathbb{R}^d$

$$\langle x, y \rangle \leq \|x\| \|y\| .$$

A.3 Fenchel-Young inequality

For all $x, y \in \mathbb{R}^d$ and $\beta > 0$

$$2 \langle x, y \rangle \leq \beta^{-1} \|x\|^2 + \beta \|y\|^2.$$

A.4 Recursion Lemma

Lemma 2. *For all $x \in [0; 1)$ consider a function*

$$\phi(x) := 1 - (1 - x)^\alpha - \max\{1, \alpha\} x.$$

Then for all $0 \leq x < 1$ and $\alpha \in \mathbb{R}$ we can obtain that $\phi(x) \leq 0$.

Proof. First consider the case of $\alpha \notin (0; 1)$. Then we can write out Bernoulli's inequality: for all $x < 1$ it holds that

$$(1 - x)^\alpha \geq 1 - \alpha x.$$

Therefore for $0 \leq x < 1$:

$$\phi(x) = 1 - (1 - x)^\alpha - \max\{1, \alpha\} x \leq 1 - (1 - x)^\alpha - \alpha x \leq 0.$$

Now we consider case $0 < \alpha < 1$, therefore $\phi(x)$ takes the form

$$\phi(x) = 1 - (1 - x)^\alpha - x.$$

Note that

$$\phi''(x) = \alpha(1 - \alpha)(1 - x)^{\alpha - 2} > 0.$$

Therefore $\phi(x)$ is convex on a segment $[0; 1]$ and $\psi(0) = \psi(1) = 0$, that means that $\phi(x) \leq 0$ for all $x \in [0; 1)$. This finishes the proof. □

Lemma 3 (Recursion Lemma)**.** *Suppose we have the following recurrence relation for variables $\{r_k\}_{k=0}^N \subset \mathbb{R}$*

$$r_{k+1} \leq \left(1 - \frac{\beta_0}{(k + k_0)^{\alpha_0}}\right) r_k + \sum_{i=1}^m \frac{\beta_i}{(k + k_0)^{\alpha_i}}, \tag{11}$$

where $\beta_i > 0 \ \ \forall i \in \overline{0,m}$, $0 \leq \alpha_0 \leq 1$, $\alpha_i \in \mathbb{R} \ \ \forall i \in \overline{1,m}$.

Then we can estimate the convergence of the sequence $\{r_k\}_{k=0}^N$ *to zero:*

$$r_k \leq 2 \cdot \sum_{i=1}^{m} \frac{Q_i}{(k+k_0)^{\alpha_i - \alpha_0}}, \tag{12}$$

where $Q_{i^*} = \max\{\beta_{i^*}/\beta_0, r_0 k_0^{\alpha_{i^*}-\alpha_0}\}$ *and* $Q_i = \beta_i/\beta_0$ *if* $i \neq i^*$, *where* i^* *we can choose arbitrarily from the set* $\overline{1,m}$, *and*

- *if* $0 \leq \alpha_0 < 1$:

$$k_0 \geq \left(\frac{2}{\beta_0}\max\{1, \max\{\alpha_i\} - \alpha_0\}\right)^{\frac{1}{1-\alpha_0}} \quad \textit{and} \quad \beta_0 > 0.$$

- *if* $\alpha_0 = 1$:

$$k_0 \in \mathbb{N} \quad \textit{and} \quad \beta_0 \geq 2\max\{1, \max\{\alpha_i\} - 1\}.$$

Proof. We prove the claim in (12) by induction. First, note that

$$r_0 = r_0 \cdot \left(\frac{k_0}{0+k_0}\right)^{\alpha_{i^*}-\alpha_0} \leq \frac{Q_{i^*}}{(0+k_0)^{\alpha_{i^*}-\alpha_0}} \leq 2 \cdot \sum_{i=0}^{m} \frac{Q_i}{(0+k_0)^{\alpha_i-\alpha_0}}.$$

and therefore the base step of the induction holds true.

Now assume that the condition in (12) holds for some k. Now we will show that this condition will hold for $k+1$.

We start by fitting (12) into the original recurrence relation (11) and using that $\beta_i \leq Q_i\beta_0$:

$$\begin{aligned} r_{k+1} &\leq \left(1 - \frac{\beta_0}{(k+k_0)^{\alpha_0}}\right) \cdot \left(2\sum_{i=1}^{m} \frac{Q_i}{(k+k_0)^{\alpha_i-\alpha_0}}\right) + \sum_{i=1}^{m} \frac{\beta_i}{(k+k_0)^{\alpha_i}} \\ &\leq 2\sum_{i=1}^{m} \frac{Q_i}{(k+k_0)^{\alpha_i-\alpha_0}} - \sum_{i=1}^{m} \frac{Q_i\beta_0}{(k+k_0)^{\alpha_i}} = \sum_{i=1}^{m} \left(\frac{2Q_i}{(k+k_0)^{\alpha_i-\alpha_0}} - \frac{Q_i\beta_0}{(k+k_0)^{\alpha_i}}\right). \end{aligned}$$

Our goal is to show that for all $i \in \overline{1,m}$ it holds that

$$\frac{2Q_i}{(k+k_0)^{\alpha_i-\alpha_0}} - \frac{Q_i\beta_0}{(k+k_0)^{\alpha_i}} \leq \frac{2Q_i}{(k+k_0+1)^{\alpha_i-\alpha_0}}. \tag{13}$$

Let us rewrite this inequality in such a way that it takes a more convenient form:

$$\frac{2}{\beta_0}\underbrace{\left[1 - \left(1 - \frac{1}{k+k_0+1}\right)^{\alpha_i-\alpha_0}\right]}_{①} \leq \left(\frac{1}{k+k_0}\right)^{\alpha_0}.$$

Using Lemma 2 with $x = (k+k_0+1)^{-1} \in [0;1)$ and $\alpha = \alpha_i - \alpha_0$ we can obtain that

$$① \leq \max\{1, \alpha_i - \alpha_0\}\frac{1}{k+k_0+1} \leq \max\{1, \alpha_i - \alpha_0\}\frac{1}{k+k_0}.$$

Now our desired inequality (13) takes form

$$\frac{2}{\beta_0}\max\{1,\alpha_i-\alpha_0\}\frac{1}{k+k_0}\leq\left(\frac{1}{k+k_0}\right)^{\alpha_0}.$$

Again, we rewrite it in a more convenient form:

$$\frac{2}{\beta_0}\max\{1,\alpha_i-\alpha_0\}\leq(k+k_0)^{1-\alpha_0}. \tag{14}$$

Now consider two cases

- If $0\leq\alpha_0<1$.
  In this case $(k+k_0)^{1-\alpha_0}\geq k_0^{1-\alpha_0}$ and if we take

  $$k_0\geq\left(\frac{2}{\beta_0}\max\{1,\max\{\alpha_i\}-\alpha_0\}\right)^{\frac{1}{1-\alpha_0}},$$

  then according to (14) desired inequality (13) will be fulfilled for all $i\in\overline{1,m}$ for all $\beta_0>0$.
- If $\alpha_0=1$, then inequality (14) takes form

  $$\frac{2}{\beta_0}\max\{1,\alpha_i-1\}\leq 1.$$

  Therefore if we take

  $$\beta_0\geq 2\max\{1,\max\{\alpha_i\}-1\},$$

  then again according to (14) desired inequality (13) will be fulfilled for all $i\in\overline{1,m}$ for all $k_0\in\mathbb{N}$.

This finishes the proof. □

B Full versions of Corollaries from Section 2

B.1 Full version of Corollary 1

Corollary 5. *Under the conditions of Theorem 2, choosing γ_k,τ,Δ as*

$$\gamma_k=\frac{4}{k+8d},\ \tau=\mathcal{O}\left(\frac{\varepsilon}{\sqrt{d}LD}\right),\ \Delta=\mathcal{O}\left(\frac{\varepsilon^2}{dLD^2}\right),$$

in order to achieve an ε-approximate solution (in terms of $\mathbb{E}\left[f(x^k)-f^*\right]\leq\varepsilon$) it takes

$$\mathcal{O}\left(\frac{d\max\{LD^2,F_0\}}{\varepsilon}\right)\ \textit{iterations of Algorithm 3.}$$

We can re-write this corollary in terms of $\varepsilon(\Delta)$.

Corollary 6. *Under the conditions of Theorem 2, choosing $\gamma_k, \tau, \varepsilon$ as*

$$\gamma_k = \frac{4}{k+8d},\ \tau = \mathcal{O}\left(\sqrt{\Delta/L}\right),\ \varepsilon = \mathcal{O}\left(\sqrt{dLD^2\Delta}\right),$$

in order to achieve an ε-approximate solution (in terms of $\mathbb{E}\left[f(x^k) - f^*\right] \leq \varepsilon$) it takes

$$\mathcal{O}\left(\frac{\sqrt{d}\max\{LD^2, F_0\}}{\sqrt{LD^2\Delta}}\right) \text{ iterations of Algorithm 3.}$$

B.2 Full version of Corollaries 2 and 3

Corollary 7 (Full version of Corollary 2)**.** *Under the conditions of Theorem 3, choosing γ, τ, Δ as*

$$\gamma \equiv \frac{1}{4dL},\ \tau = \mathcal{O}\left(\frac{\varepsilon}{\sqrt{d}L}\right),\ \Delta = \mathcal{O}\left(\frac{\varepsilon^2}{dL}\right),$$

in order to achieve an ε-approximate solution (in terms of $\mathbb{E}\left[\|\nabla f(\widehat{x}_N)\|^2\right] \leq \varepsilon^2$, where $\widehat{x}_N$ is chosen uniformly from $\{x^k\}_{k=0}^N$) it takes

$$\mathcal{O}\left(\frac{dL\Phi_0}{\varepsilon^2}\right) \text{ iterations of Algorithm 4.}$$

Corollary 8 (Full version of Corollary 3)**.** *Under the conditions of Theorem 4, choosing γ, τ, Δ as*

$$\gamma \equiv \frac{1}{4dL},\ \tau = \mathcal{O}\left(\frac{\sqrt{\varepsilon\mu}}{\sqrt{d}L}\right),\ \Delta = \mathcal{O}\left(\frac{\varepsilon\mu}{dL}\right).$$

in order to achieve an ε-approximate solution (in terms of $\mathbb{E}\left[f(x^N) - f^*\right] \leq \varepsilon$) it takes

$$\mathcal{O}\left(\frac{dL}{\mu}\log\left[\frac{F_0}{\varepsilon}\right]\right) \text{ iterations of Algorithm 4.}$$

B.3 Full version of Corollary 4

Corollary 9. *Under the conditions of Theorem 6, choosing $\gamma_k, \eta_k, \tau, \Delta$ as*

$$\gamma_k = \frac{4}{k+8d^{3/2}},\ \eta_k = \frac{4}{(k+8d^{3/2})^{2/3}},\ \tau = \mathcal{O}\left(\frac{\varepsilon}{\sqrt{d}LD}\right),\ \Delta = \mathcal{O}\left(\frac{\varepsilon^2}{dLD^2}\right),$$

in order to achieve an ε-approximate solution (in terms of $\mathbb{E}\left[f(x^N) - f^*\right] \leq \varepsilon$) it takes

$$\mathcal{O}\left(\max\left\{\left[\frac{LD^2 + d\sigma_\nabla D + \sqrt{d}(f(x^0) - f^*)}{\varepsilon}\right]^3; \frac{d^{9/2}\sigma_f^3 L^3 D^6}{\varepsilon^6}\right\}\right) \text{ iterations of Algorithm 5.}$$

In the two point feedback case, $\sigma_f^2 = 0$ and the last equation takes form

$$\mathcal{O}\left(\left[\frac{LD^2 + d\sigma_\nabla D + \sqrt{d}(f(x^0) - f^*)}{\varepsilon}\right]^3\right).$$

C PROOF OF CONVERGE RATE OF JAGUAR. DETERMINISTIC AND STOCHASTIC CASES.

C.1 RESULT FOR THE FULL-APPROXIMATION (STOCHASTIC CASE)

Lemma 4. *Under Assumptions 6, 7, 8 and 9 in the OPF case* (8) *the following inequality holds*

$$\mathbb{E}\left[\left\|\widetilde{\nabla} f_\delta(x, \xi_1^+, \xi_1^-, ..., \xi_d^+, \xi_d^-) - \nabla f(x)\right\|^2\right] \leq dL^2\tau^2 + \frac{8d\sigma_f^2}{\tau^2} + 2d\sigma_\nabla^2 + \frac{2d\Delta^2}{\tau^2}. \tag{15}$$

In the case of two point feedback $\sigma_f^2 = 0$ *and in the deterministic case* (3) $\sigma_\nabla^2 = \sigma_f^2 = 0$.

Proof. Let's start by writing out a definition of gradient approximation $\widetilde{\nabla} f_\delta(x, \xi_{1,d}^\pm)$:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{\nabla} f_\delta(x, \xi_{1,d}^\pm) - \nabla f(x)\right\|^2\right] = \mathbb{E}\left[\left\|\sum_{i=1}^d \frac{f_\delta(x+\tau e_i, \xi_i^+) - f_\delta(x-\tau e_i, \xi_i^-)}{2\tau} e_i - \nabla f(x)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\sum_{i=1}^d \left(\frac{f_\delta(x+\tau e_i, \xi_i^+) - f_\delta(x-\tau e_i, \xi_i^-)}{2\tau} - \langle \nabla f(x), e_i\rangle\right) e_i\right\|^2\right] \\
&\overset{(\star)}{=} \sum_{i=1}^d \mathbb{E}\left[\left\|\left(\frac{f_\delta(x+\tau e_i, \xi_i^+) - f_\delta(x-\tau e_i, \xi_i^-)}{2\tau} - \langle \nabla f(x), e_i\rangle\right) e_i\right\|^2\right] \\
&= \sum_{i=1}^d \mathbb{E}\left[\left|\frac{f_\delta(x+\tau e_i, \xi_i^+) - f_\delta(x-\tau e_i, \xi_i^-)}{2\tau} - \langle \nabla f(x), e_i\rangle\right|^2\right].
\end{aligned}$$

The $(\star)$ equality holds since $\langle e_i, e_j\rangle = 0$ if $i \neq j$. Now let's estimate the value under the summation:

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{f_\delta(x+\tau e_i, \xi_i^+) - f_\delta(x-\tau e_i, \xi_i^-)}{2\tau} - \langle \nabla f(x), e_i\rangle\right|^2\right] \\
&= \mathbb{E}\left[\left|\frac{f(x+\tau e_i, \xi_i^+) - f(x-\tau e_i, \xi_i^-)}{2\tau} - \langle \nabla f(x), e_i\rangle + \frac{\delta(x+\tau e_i, \xi_i^+) - \delta(x-\tau e_i, \xi_i^-)}{2\tau}\right|^2\right] \\
&\overset{A.1}{\leq} \frac{1}{2\tau^2} \underbrace{\mathbb{E}\left[\left|f(x+\tau e_i, \xi_i^+) - f(x-\tau e_i, \xi_i^-) - \langle \nabla f(x), 2\tau e_i\rangle\right|^2\right]}_{①} + \frac{2\Delta^2}{\tau^2}.
\end{aligned}$$

Last inequality holds since noise is bounded . Consider ①. Using A.1 with $n = 4$ we get:

$$\begin{aligned}
&\mathbb{E}\left[\left|f(x+\tau e_i, \xi_i^+) - f(x-\tau e_i, \xi_i^-) - \langle \nabla f(x), 2\tau e_i\rangle\right|^2\right] \\
&\leq 4\mathbb{E}\left[\left|f(x+\tau e_i, \xi_i^+) - f(x, \xi_i^+) - \left\langle \nabla f(x, \xi_i^+), \tau e_i\right\rangle\right|^2\right] \\
&+ 4\mathbb{E}\left[\left|-f(x-\tau e_i, \xi_i^-) + f(x, \xi_i^-) + \left\langle \nabla f(x, \xi_i^-), -\tau e_i\right\rangle\right|^2\right] \\
&+ 4\mathbb{E}\left[\left|f(x, \xi_i^+) - f(x, \xi_i^-)\right|^2\right] \\
&+ 4\mathbb{E}\left[\left|\left\langle \nabla f(x, \xi_i^+) + \nabla f(x, \xi_i^-) - 2\nabla f(x), \tau e_i\right\rangle\right|^2\right].
\end{aligned} \tag{16}$$

Let's evaluate all these four components separately. Since functions $f(x, \xi_i^+)$ and $f(x, \xi_i^-)$ are $L(\xi_i^\pm)$-smooth we have estimates for first and second:

$$\begin{aligned} &\left|f(x+\tau e_i, \xi_i^+) - f(x, \xi_i^+) - \left\langle \nabla f(x, \xi_i^+), \tau e_i \right\rangle\right| \leq \frac{L(\xi_i^+)}{2}\tau^2 \leq \frac{L}{2}\tau^2, \\ &\left|-f(x-\tau e_i, \xi_i^-) + f(x, \xi_i^-) + \left\langle \nabla f(x, \xi_i^-), -\tau e_i \right\rangle\right| \leq \frac{L(\xi_i^-)}{2}\tau^2 \leq \frac{L}{2}\tau^2. \end{aligned} \tag{17}$$

If we consider TPF approximation (7), then third term in (16) equals to zero, since $\xi_i^+ = \xi_i^-$, if we consider the OPF case (8), then we can obtain

$$\mathbb{E}\left[\left|f(x, \xi_i^+) - f(x, \xi_i^-)\right|^2\right] \leq 2\mathbb{E}\left[\left|f(x, \xi_i^+) - f(x)\right|^2\right] + 2\mathbb{E}\left[\left|f(x, \xi_i^-) - f(x)\right|^2\right] \leq 4\sigma_f^2. \tag{18}$$

Consider the last point in (16) and using Cauchy–Schwarz inequality A.2 we can obtain:

$$\mathbb{E}\left[\left|\left\langle \nabla f(x, \xi_i^+) - \nabla f(x), \tau e_i \right\rangle\right|^2\right] \leq \mathbb{E}\left[\left\|\nabla f(x, \xi_i^+) - \nabla f(x)\right\|^2 \tau^2\right] \leq \sigma_\nabla^2 \tau^2. \tag{19}$$

Combining (17), (18) and (19) we obtain

$$\mathbb{E}\left[\left\|\widetilde{\nabla} f_\delta(x, \xi_1^+, \xi_1^-, ..., \xi_d^+, \xi_d^-) - \nabla f(x)\right\|^2\right] \leq dL^2\tau^2 + \frac{8d\sigma_f^2}{\tau^2} + 2d\sigma_\nabla^2 + \frac{2d\Delta^2}{\tau^2}.$$

In the case of two point feedback $\sigma_f^2 = 0$ and in the deterministic case (3) $\sigma_\nabla^2 = \sigma_f^2 = 0$. This finishes the proof. □

C.2 Result for the JAGUAR approximation (Algorithm 1) (stochastic case)

Lemma 5. *Under Assumptions 6, 7, 8 and 9 in the OPF case* (8) *the following inequality holds*

$$\begin{aligned} \mathbb{E}\left[\left\|h^{k+1} - \nabla f(x^{k+1})\right\|^2\right] \leq \left(1 - \frac{1}{2d}\right)\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right] + 2dL^2\mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right] \\ + L^2\tau^2 + \frac{8\sigma_f^2}{\tau^2} + 2\sigma_\nabla^2 + \frac{2\Delta^2}{\tau^2}. \end{aligned} \tag{20}$$

In the case of two point feedback $\sigma_f^2 = 0$ and in the deterministic case (3) $\sigma_\nabla^2 = \sigma_f^2 = 0$.

Proof. Let us start by writing out a definition of h^k using line 6 of Algorithm 5

$$
\begin{aligned}
&\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right] = \mathbb{E}\left[\left\|h^{k-1} + \widetilde{\nabla}_i f_\delta(x^k, \xi^+, \xi^-) - \left\langle h^{k-1}, e_i\right\rangle e_i - \nabla f(x^k)\right\|^2\right] \\
&= \mathbb{E}\Bigg[\Bigg\| \left(I - e_i e_i^T\right)\left(h^{k-1} - \nabla f(x^{k-1})\right) \\
&\qquad + e_i e_i^T \left(\widetilde{\nabla} f_\delta(x^k, \xi^+, \xi^-, ..., \xi^+, \xi^-) - \nabla f(x^k)\right) \\
&\qquad - \left(I - e_i e_i^T\right)\left(\nabla f(x^k) - \nabla f(x^{k-1})\right)\Bigg\|^2\Bigg] \\
&= \underbrace{\mathbb{E}\left[\left\|\left(I - e_i e_i^T\right)\left(h^{k-1} - \nabla f(x^{k-1})\right)\right\|^2\right]}_{①} \\
&\quad + \underbrace{\mathbb{E}\left[\left\|e_i e_i^T \left(\widetilde{\nabla} f_\delta(x^k, \xi^+, \xi^-, ..., \xi^+, \xi^-) - \nabla f(x^k)\right)\right\|^2\right]}_{②} \\
&\quad + \underbrace{\mathbb{E}\left[\left\|\left(I - e_i e_i^T\right)\left(\nabla f(x^k) - \nabla f(x^{k-1})\right)\right\|^2\right]}_{③} \\
&\quad + \underbrace{\mathbb{E}\left[2\left\langle\left(I - e_i e_i^T\right)\left(h^{k-1} - \nabla f(x^{k-1})\right), \left(I - e_i e_i^T\right)\left(\nabla f(x^k) - \nabla f(x^{k-1})\right)\right\rangle^2\right]}_{④}.
\end{aligned}
$$

In the last equality the two remaining scalar products are zero, since $e_i e_i^T\left(I - e_i e_i^T\right) = e_i^T e_i - e_i^T e_i = 0$. Consider the ①. Using notation $v := h^{k-1} - \nabla f(x^{k-1})$ we obtain

$$
\begin{aligned}
\mathbb{E}\left[\left\|\left(I - e_i e_i^T\right)\left(h^{k-1} - \nabla f(x^{k-1})\right)\right\|^2\right] &= \mathbb{E}\left[v^T\left(I - e_i e_i^T\right)^T\left(I - e_i e_i^T\right)v\right] \\
&= \mathbb{E}\left[v^T\left(I - e_i e_i^T\right)v\right] = \mathbb{E}\left[\mathbb{E}_{k-1}\left[v^T\left(I - e_i e_i^T\right)v\right]\right],
\end{aligned}
$$

where $\mathbb{E}_{k-1}[\cdot]$ is the conditional expectation with fixed randomness of all steps up to $k-1$. Since at step k the vectors e_i are generated independently, we obtain

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{E}_{k-1}\left[v^T\left(I - e_i e_i^T\right)v\right]\right] &= \mathbb{E}\left[v^T \mathbb{E}_{k-1}\left[\left(I - e_i e_i^T\right)\right]v\right] \\
&= \left(1 - \frac{1}{d}\right)\mathbb{E}\left[\left\|h^{k-1} - \nabla f(x^{k-1})\right\|^2\right].
\end{aligned}
$$

Consider ②. Since we generate i independently, x^k is independent of the e_i generated at step k, then we can apply the same technique as in estimation ①:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|e_i e_i^T \left(\widetilde{\nabla} f_\delta(x^k, \xi^+, \xi^-, ..., \xi^+, \xi^-) - \nabla f(x^k)\right)\right\|^2\right] \\
&= \frac{1}{d}\mathbb{E}\left[\left\|\widetilde{\nabla} f_\delta(x^k, \xi^+, \xi^-, ..., \xi^+, \xi^-) - \nabla f(x^k)\right\|^2\right].
\end{aligned}
$$

Using Lemma 4 we obtain

$$
\frac{1}{d}\mathbb{E}\left[\left\|\widetilde{\nabla} f_\delta(x^k, \xi^+, \xi^-, ..., \xi^+, \xi^-) - \nabla f(x^k)\right\|^2\right] \leq L^2\tau^2 + \frac{8\sigma_f^2}{\tau^2} + 2\sigma_\nabla^2 + \frac{2\Delta^2}{\tau^2}.
$$

Consider ③. Using the same technique as in estimation ①:

$$\mathbb{E}\left[\left\|\left(I - e_i e_i^T\right)\left(\nabla f(x^k) - \nabla f(x^{k-1})\right)\right\|^2\right] \leq \left(1 - \frac{1}{d}\right) L^2 \mathbb{E}\left[\left\|x^k - x^{k-1}\right\|^2\right].$$

Consider ④. Using Fenchel-Young inequality A.3 with $\beta = 2d$ we obtain

$$④ \leq \left(1 - \frac{1}{d}\right)\left(\frac{1}{2d}\mathbb{E}\left[\left\|h^{k-1} - \nabla f(x^{k-1})\right\|^2\right] + 2dL^2\mathbb{E}\left[\left\|x^k - x^{k-1}\right\|^2\right]\right).$$

Therefore it holds that

$$\begin{aligned}\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right] \leq \left(1 - \frac{1}{2d}\right)\mathbb{E}\left[\left\|h^{k-1} - \nabla f(x^{k-1})\right\|^2\right] + 2dL^2\mathbb{E}\left[\left\|x^k - x^{k-1}\right\|^2\right] \\ + L^2\tau^2 + \frac{8\sigma_f^2}{\tau^2} + 2\sigma_\nabla^2 + \frac{2\Delta^2}{\tau^2}.\end{aligned}$$

This finishes the proof. □

C.3 Result for the SAGA approximation (line 7 of Algorithm 3)

Lemma 6. *Under Assumptions 6, 7, 8 and 9 in the OPF case* (8) *the following inequality holds*

$$\begin{aligned}\mathbb{E}\left[\left\|\rho^k - \nabla f(x^k)\right\|^2\right] \leq 4d\mathbb{E}\left[\left\|h^{k-1} - \nabla f(x^{k-1})\right\|\right] \\ + 4d^2\left(L^2\tau^2 + \frac{8\sigma_f^2}{\tau^2} + 2\sigma_\nabla^2 + \frac{2\Delta^2}{\tau^2}\right) + 2dL^2\mathbb{E}\left[\left\|x^k - x^{k-1}\right\|^2\right].\end{aligned}$$

In the case of two point feedback $\sigma_f^2 = 0$.

Proof. Let's start by writing out a definition of ρ^k using line 7 of Algorithm 5

$$\begin{aligned}&\mathbb{E}\left[\left\|\rho^k - \nabla f(x^k)\right\|^2\right] = \mathbb{E}\left[\left\|h^{k-1} + d\widetilde{\nabla}_i f_\delta(x^k, \xi^+, \xi^-) - d\left\langle h^{k-1}, e_i\right\rangle e_i - \nabla f(x^k)\right\|^2\right] \\ &= \mathbb{E}\left[\left\|\left(I - de_ie_i^T\right)\left(h^{k-1} - \nabla f(x^{k-1})\right) + de_ie_i^T\left(\widetilde{\nabla} f_\delta(x^k, \xi^+, \xi^-, ..., \xi^+, \xi^-) - \nabla f(x^k)\right)\right.\right. \\ &\qquad \left.\left. + \left(I - de_ie_i^T\right)\left(\nabla f(x^{k-1}) - \nabla f(x^k)\right)\right\|^2\right] \\ &\overset{\star}{\leq} 4(d-1)\mathbb{E}\left[\left\|h^{k-1} - \nabla f(x^{k-1}\right\|\right] + 4d\mathbb{E}\left[\left\|\widetilde{\nabla} f_\delta(x^k, \xi^+, \xi^-, ..., \xi^+, \xi^-) - \nabla f(x^k)\right\|^2\right] \\ &\quad + 2(d-1)\mathbb{E}\left[\left\|\nabla f(x^{k-1}) - \nabla f(x^k)\right\|^2\right].\end{aligned}$$

The $\star$ inequality is correct due to similar reasoning as in the proof of Lemma 5 and due to Fenchel-Young inequality A.3. Now we can estimate all three summands using Lemmas 5 and 4 and using Assumption 6:

$$\begin{aligned}\mathbb{E}\left[\left\|\rho^k - \nabla f(x^k)\right\|^2\right] \leq 4d\mathbb{E}\left[\left\|h^{k-1} - \nabla f(x^{k-1}\right\|\right] \\ + 4d^2\left(L^2\tau^2 + \frac{8\sigma_f^2}{\tau^2} + 2\sigma_\nabla^2 + \frac{2\Delta^2}{\tau^2}\right) + 2dL^2\mathbb{E}\left[\left\|x^k - x^{k-1}\right\|^2\right].\end{aligned}$$

This finishes the proof. □

C.4 RESULT FOR THE JAGUAR APPROXIMATION (ALGORITHM 3)

Lemma 7. *Under Assumption 6 the following inequality holds*

$$\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \leq (1-\eta_k)\,\mathbb{E}\left[\left\|\nabla f(x^{k-1}) - g^{k-1}\right\|^2\right] + \frac{4L^2}{\eta_k}\mathbb{E}\left[\left\|x^k - x^{k-1}\right\|^2\right] + \eta_k^2\mathbb{E}\left[\left\|\nabla f(x^k) - \rho^k\right\|^2\right] + 3\eta_k\mathbb{E}\left[\left\|\widetilde{\nabla} f_\delta(x^k) - \nabla f(x^k)\right\|^2\right].$$

Proof. We start by writing out a definition of g^k using line 8 of Algorithm 5

$$\begin{aligned}
&\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] = \mathbb{E}\left[\left\|\nabla f(x^{k-1}) - g^{k-1} + \nabla f(x^k) - \nabla f(x^{k-1}) - \left(g^k - g^{k-1}\right)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\nabla f(x^{k-1}) - g^{k-1} + \nabla f(x^k) - \nabla f(x^{k-1}) - \eta_k\left(\rho^k - g^{k-1}\right)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|(1-\eta_k)(\nabla f(x^{k-1}) - g^{k-1}) + (1-\eta_k)(\nabla f(x^k) - \nabla f(x^{k-1})) + \eta_k\left(\nabla f(x^k) - \rho^k\right)\right\|^2\right] \\
&= (1-\eta_k)^2\underbrace{\mathbb{E}\left[\left\|\nabla f(x^{k-1}) - g^{k-1}\right\|^2\right]}_{①} + (1-\eta_k)^2\underbrace{\mathbb{E}\left[\left\|\nabla f(x^k) - \nabla f(x^{k-1})\right\|^2\right]}_{②} \\
&\quad + \eta_k^2\underbrace{\mathbb{E}\left[\left\|\nabla f(x^k) - \rho^k\right\|^2\right]}_{③} + 2(1-\eta_k)^2\underbrace{\mathbb{E}\left[\left\langle\nabla f(x^{k-1}) - g^{k-1}, \nabla f(x^k) - \nabla f(x^{k-1})\right\rangle\right]}_{④} \\
&\quad + 2\eta_k(1-\eta_k)\underbrace{\mathbb{E}\left[\left\langle\nabla f(x^{k-1}) - g^{k-1}, \nabla f(x^k) - \rho^k\right\rangle\right]}_{⑤} \\
&\quad + 2\eta_k(1-\eta_k)\underbrace{\mathbb{E}\left[\left\langle\nabla f(x^k) - \nabla f(x^{k-1}), \nabla f(x^k) - \rho^k\right\rangle\right]}_{⑥}.
\end{aligned}$$

Consider ⑤. Since we generate ξ^+ and ξ^- independently, we obtain

$$⑤ = \mathbb{E}\left[\left\langle\nabla f(x^{k-1}) - g^{k-1}, \mathbb{E}_{k-1}\left[\nabla f(x^k) - \rho^k\right]\right\rangle\right],$$

where $\mathbb{E}_{k-1}[\cdot]$ is the conditional expectation with fixed randomness of all steps up to $k-1$. Using fact that

$$\mathbb{E}_{k-1}\left[\nabla f(x^k) - \rho^k\right] = \nabla f(x^k) - \widetilde{\nabla} f(x^k) = \nabla f(x^k) - \sum_{i=1}^{d}\frac{f(x+\tau e_i) - f(x-\tau e_i)}{2\tau}e_i.$$

Fact that for $\widetilde{\nabla} f_\delta(x)$ Lemma 4 holds true with $\sigma_f^2 = \sigma_\nabla^2 = 0$ and using Cauchy Schwarz inequality A.2 with $\beta = 2(1-\eta_k)$ we can assume

$$⑤ \leq \frac{1}{4(1-\eta_k)}\mathbb{E}\left[\left\|\nabla f(x^{k-1}) - g^{k-1}\right\|^2\right] + (1-\eta_k)\mathbb{E}\left[\left\|\widetilde{\nabla} f_\delta(x^k) - \nabla f(x^k)\right\|^2\right]. \tag{21}$$

Similarly it can be shown that

$$⑥ \leq \frac{1}{2(1-\eta_k)\eta_k^2}\mathbb{E}\left[\left\|\nabla f(x^k) - \nabla f(x^{k-1})\right\|^2\right] + \frac{(1-\eta_k)\eta_k^2}{2}\mathbb{E}\left[\left\|\widetilde{\nabla} f_\delta(x^k) - \nabla f(x^k)\right\|^2\right]. \tag{22}$$

Using Assumption 6 we can obtain that

$$② \leq L^2 \mathbb{E}\left[\left\|x^k - x^{k-1}\right\|^2\right]. \tag{23}$$

Consider ④. Using auchy Schwarz inequality A.2 with $\beta = 2\frac{(1-\eta_k)^2}{\eta_k}$ we can assume

$$④ \leq \frac{\eta_k}{4(1-\eta_k)^2}\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] + \frac{(1-\eta_k)^2}{\eta_k}L^2\mathbb{E}\left[\left\|x^k - x^{k-1}\right\|^2\right]. \tag{24}$$

Putting (21), (22), (23) and (24) all together and using the fact that $(1-\eta_k)^2 \leq 1-\eta_k$, we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \leq (1-\eta_k)\,\mathbb{E}\left[\left\|\nabla f(x^{k-1}) - g^{k-1}\right\|^2\right] + \frac{4L^2}{\eta_k}\mathbb{E}\left[\left\|x^k - x^{k-1}\right\|^2\right] \\ + \eta_k^2\mathbb{E}\left[\left\|\nabla f(x^k) - \rho^k\right\|^2\right] + 3\eta_k\mathbb{E}\left[\left\|\widetilde{\nabla} f_\delta(x^k) - \nabla f(x^k)\right\|^2\right].\end{aligned}$$

This finishes the proof. □

D PROOF OF CONVERGE RATE OF FW VIA JAGUAR

D.1 RESULTS FOR THE DETERMINISTIC CASE (ALGORITHM 3)

Proof of Theorem 1. We start by writing out result from Lemma 5 with $\sigma_f^2 = \sigma_\nabla^2 = 0$ and setting up $\gamma_k = \frac{4}{k+k_0}$:

$$\begin{aligned}\mathbb{E}\left[\left\|h^{k+1} - \nabla f(x^{k+1})\right\|^2\right] \leq \left(1 - \frac{1}{2d}\right)\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right] + \frac{32dL^2D^2}{(k+k_0)^2} \\ + L^2\tau^2 + \frac{2\Delta^2}{\tau^2}.\end{aligned}$$

Now we use Lemma 3 with $\alpha_0 = 0, \beta_0 = 1/2d$; $\alpha_1 = 2, \beta_1 = 32dL^2D^2$; $\alpha_2 = 0, \beta_2 = L^2\tau^2 + \frac{2\Delta^2}{\tau^2}$ and $i^* = 1$.

$$\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right] = \mathcal{O}\left(dL^2\tau^2 + \frac{d\Delta^2}{\tau^2} + \frac{\max\{d^2L^2D^2, \left\|h^0 - \nabla f(x^0)\right\|^2 \cdot k_0^2\}}{(k+k_0)^2}\right),$$

where $k_0 = (4d \cdot 2)^1 = 8d$. If $h_0 = \widetilde{\nabla} f_\delta(x^0)$ we can obtain

$$\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right] = \mathcal{O}\left(dL^2\tau^2 + \frac{d\Delta^2}{\tau^2} + \frac{d^2L^2D^2}{(k+8d)^2}\right).$$

This finishes the proof. □

Proof of Theorem 2. We start by writing our the result of Lemma 2 from Mokhtari et al. (2020). Under Assumptions 6, 3 the following inequality holds

$$\mathbb{E}\left[f(x^{k+1}) - f(x^*)\right] \leq (1-\gamma_k)\mathbb{E}\left[f(x^k) - f(x^*)\right] + \gamma_k D\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|\right] + \frac{LD^2\gamma_k^2}{2}.$$

We can evaluate $\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|\right]$ using Jensen's inequality:

$$\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|\right] \leq \sqrt{\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right]}.$$

Using result from Theorem 1 we can obtain

$$\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|\right] = \mathcal{O}\left(\frac{dLD}{k+8d} + \sqrt{d}L\tau + \frac{\sqrt{d}\Delta}{\tau}\right).$$

Using Lemma 3 with $\alpha_0 = 1, \beta_0 = 4, k_0 = 8d$; $\alpha_1 = 2, \beta_1 = 8LD^2 + dLD^2$; $\alpha_2 = 1, \beta_2 = \sqrt{d}L\tau D + \frac{\sqrt{d}\Delta D}{\tau}$ and $i^* = 1$, we get:

$$\mathbb{E}\left[f(x^k) - f(x^*)\right] = \mathcal{O}\left(\frac{d\max\{LD^2, f(x^0) - f(x^*)\}}{k+8d} + \sqrt{d}LD\tau + \frac{\sqrt{d}\Delta D}{\tau}\right).$$

In Lemma 3 if $\alpha_0 = 1$ we need to take $\beta_0 \geq 2 \cdot 1 = 2$, we take $\beta_0 = 4$. This finishes the proof. □

Proof of Corollary 1. We aim to achieve precision ε, i.e.

$$\mathbb{E}\left[f(x^N) - f(x^*)\right] = \mathcal{O}\left(\frac{d\max\{LD^2, f(x^0) - f(x^*)\}}{N+8d} + \sqrt{d}LD\tau + \frac{\sqrt{d}\Delta D}{\tau}\right) \leq \varepsilon.$$

Therefore we need to take

$$N = \mathcal{O}\left(\frac{d\max\{LD^2, f(x^0) - f(x^*)\}}{\varepsilon}\right),$$
$$\tau = \mathcal{O}\left(\frac{\varepsilon}{\sqrt{d}LD}\right), \quad \Delta = \mathcal{O}\left(\frac{\varepsilon\tau}{\sqrt{d}D}\right) = \mathcal{O}\left(\frac{\varepsilon^2}{dLD^2}\right).$$

This finishes the proof. □

D.2 Results for the stochastic case (Algorithm 5)

Proof of Theorem 5. Consider $\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right]$. We start by writing out result from Lemma 5 and setting up $\gamma_k = \frac{4}{k+k_0}$:

$$\mathbb{E}\left[\left\|h^{k+1} - \nabla f(x^{k+1})\right\|^2\right] \leq \left(1 - \frac{1}{2d}\right)\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right] + \frac{32dL^2D^2}{(k+k_0)^2}$$
$$+ L^2\tau^2 + \frac{8\sigma_f^2}{\tau^2} + 2\sigma_\nabla^2 + \frac{2\Delta^2}{\tau^2}.$$

Now we use Lemma 3 with $\alpha_0 = 0, \beta_0 = 1/2d$; $\alpha_1 = 2, \beta_1 = 32dL^2D^2$; $\alpha_2 = 0, \beta_2 = L^2\tau^2 + \frac{8\sigma_f^2}{\tau^2} + 2\sigma_\nabla^2 + \frac{2\Delta^2}{\tau^2}$ and $i^* = 1$.

$$\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right]$$
$$= \mathcal{O}\left(dL^2\tau^2 + \frac{d\sigma_f^2}{\tau^2} + d\sigma_\nabla^2 + \frac{d\Delta^2}{\tau^2} + \frac{\max\{d^2L^2D^2, \left\|h^0 - \nabla f(x^0)\right\|^2 \cdot k_0^2\}}{(k+k_0)^2}\right),$$

where $k_0 = (4d \cdot 2)^1 = 8d$. For simplicity of calculations further we take $k_0 = 8d^{3/2} > 8d$. If $h^0 = \widetilde{\nabla} f_\delta(x^0, \xi_1^+, \xi_1^-, ..., \xi_d^+, \xi_d^-)$ we can obtain

$$\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right] = \mathcal{O}\left(dL^2\tau^2 + \frac{d\sigma_f^2}{\tau^2} + d\sigma_\nabla^2 + \frac{d\Delta^2}{\tau^2} + \frac{d^2L^2D^2}{(k+8d^{3/2})^2}\right).$$

Consider $\mathbb{E}\left[\left\|\rho^k - \nabla f(x^k)\right\|^2\right]$. Using Lemmas 6 and 4 we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\rho^k - \nabla f(x^k)\right\|^2\right] \\
&= \mathcal{O}\left(d^2L^2\tau^2 + \frac{d^2\sigma_f^2}{\tau^2} + d^2\sigma_\nabla^2 + \frac{d^2\Delta^2}{\tau^2} + \frac{d^3\max\{L^2D^2, d\left\|h^0 - \nabla f(x^0)\right\|^2\}}{(k+8d^{3/2})^2}\right).
\end{aligned}$$

Consider $\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right]$. We write out result from Lemma 7 and setting up $\eta_k = \frac{4}{(k+8d^{3/2})^{2/3}}$:

$$\begin{aligned}
&\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \leq (1-\eta_k)\,\mathbb{E}\left[\left\|\nabla f(x^{k-1}) - g^{k-1}\right\|^2\right] + \frac{4L^2D^2}{(k+8d^{3/2})^{4/3}} \\
&+ \frac{4}{(k+8d^{3/2})^{4/3}}\mathcal{O}\left(d^2L^2\tau^2 + \frac{d^2(\sigma_f^2+\Delta^2)}{\tau^2} + d^2\sigma_\nabla^2 + \frac{d^3\max\{L^2D^2, d\left\|h^0 - \nabla f(x^0)\right\|^2\}}{(k+8d^{3/2})^2}\right) \\
&+ \frac{12}{(k+8d^{3/2})^{2/3}}\left(dL^2\tau^2 + \frac{d\Delta^2}{\tau^2}\right).
\end{aligned}$$

Using Lemma 3 with $\alpha_0 = 2/3, \beta_0 = 4$; $\alpha_1 = 4/3, \beta_1 = 4L^2D^2$; $\alpha_2 = 4/3, \beta_2 = 4d^2L^2\tau^2 + \frac{4d^2\sigma_f^2}{\tau^2} + 4d^2\sigma_\nabla^2 + \frac{4d^2\Delta^2}{\tau^2}$; $\alpha_3 = 10/3, \beta_3 = 4d^3\max\{L^2D^2, d\left\|h^0 - \nabla f(x^0)\right\|^2\}$; $\alpha_4 = 2/3, \beta_4 = dL^2\tau^2 + \frac{d\Delta^2}{\tau^2}$ and $i^* = 2$ we get:

$$\begin{aligned}
&\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \\
&= \mathcal{O}\Bigg(\frac{L^2D^2 + \max\{d^2L^2\tau^2 + d^2\sigma_f^2/\tau^2 + d^2\sigma_\nabla^2 + d^2\Delta^2/\tau^2, d\left\|g^0 - \nabla f(x^0)\right\|^2\}}{(k+8d^{3/2})^{2/3}} \\
&\qquad + \frac{d^3\max\{L^2D^2, d\left\|h^0 - \nabla f(x^0)\right\|^2\}}{(k+8d^{3/2})^{8/3}} + dL^2\tau^2 + \frac{d\Delta^2}{\tau^2}\Bigg).
\end{aligned} \tag{25}$$

Since

$$\frac{d^3L^2D^2}{(k+8d^{3/2})^{8/3}} \leq \frac{L^2D^2}{(k+8d^{3/2})^{2/3}} \quad \text{and} \quad \frac{d^2L^2\tau^2 + d^2\Delta^2/\tau^2}{(k+8d^{3/2})^{2/3}} \leq dL^2\tau^2 + \frac{d\Delta^2}{\tau^2},$$

we can simplify (25):

$$\begin{aligned}
\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] = \mathcal{O}\Bigg(&\frac{L^2D^2 + \max\{d^2\sigma_f^2/\tau^2 + d^2\sigma_\nabla^2, d\left\|g^0 - \nabla f(x^0)\right\|^2\}}{(k+8d^{3/2})^{2/3}} \\
&+ \frac{d^4\left\|h^0 - \nabla f(x^0)\right\|^2}{(k+8d^{3/2})^{8/3}} + dL^2\tau^2 + \frac{d\Delta^2}{\tau^2}\Bigg).
\end{aligned}$$

If $h^0 = g^0 = \widetilde{\nabla} f_\delta(x^0, \xi_1^+, \xi_1^-, ..., \xi_d^+, \xi_d^-)$ we can obtain

$$\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] = \mathcal{O}\left(\frac{L^2D^2 + d^2\sigma_f^2/\tau^2 + d^2\sigma_\nabla^2}{(k+8d^{3/2})^{2/3}} + dL^2\tau^2 + \frac{d\Delta^2}{\tau^2}\right).$$

This finishes the proof. □

Proof of Theorem 6. Again we write out result of Lemma 2 from Mokhtari et al. (2020):

$$\mathbb{E}\left[f(x^{k+1}) - f(x^*)\right] \leq (1-\gamma_k)\mathbb{E}\left[f(x^k) - f(x^*)\right] + \gamma_k D\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|\right] + \frac{LD^2\gamma_k^2}{2}. \quad (26)$$

We can evaluate $\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|\right]$ using Jensen's inequality:

$$\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|\right] \leq \sqrt{\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right]}.$$

Using result from Theorem 5 we can obtain

$$\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|\right] = \mathcal{O}\left(\frac{LD + d\sigma_f/\tau + d\sigma_\nabla}{(k+8d^{3/2})^{1/3}} + \sqrt{d}L\tau + \frac{\sqrt{d}\Delta}{\tau}\right).$$

Set up $\gamma_k = \frac{4}{k+8d^{3/2}}$ into (26):

$$\begin{aligned}\mathbb{E}\left[f(x^{k+1}) - f(x^*)\right] &\leq (1-\gamma_k)\mathbb{E}\left[f(x^k) - f(x^*)\right] + \frac{8LD^2}{(k+8d^{3/2})^2} \\ &\quad + \frac{4D}{k+8d^{3/2}}\mathcal{O}\left(\frac{LD + d\sigma_f/\tau + d\sigma_\nabla}{(k+8d^{3/2})^{1/3}} + \sqrt{d}L\tau + \frac{\sqrt{d}\Delta}{\tau}\right).\end{aligned}$$

Using Lemma 3 with $\alpha_0 = 1, \beta_0 = 4, k_0 = 8d^{3/2}$; $\alpha_1 = 2, \beta_1 = 8LD^2$; $\alpha_2 = 4/3; \beta_2 = LD + d\sigma_f/\tau + d\sigma_\nabla$; $\alpha_3 = 1, \beta_3 = \sqrt{d}L\tau + \frac{\sqrt{d}\Delta}{\tau}$ and $i^* = 2$, we get:

$$\begin{aligned}&\mathbb{E}\left[f(x^k) - f(x^*)\right] \\ &= \mathcal{O}\Bigg(\frac{LD^2}{k+8d^{3/2}} + \frac{\max\{LD^2 + d\sigma_f D/\tau + d\sigma_\nabla D, \sqrt{d}(f(x^0) - f(x^*))\}}{(k+8d^{3/2})^{1/3}} \\ &\qquad + \sqrt{d}LD\tau + \frac{\sqrt{d}\Delta D}{\tau}\Bigg).\end{aligned}$$

In Lemma 3 if $\alpha_0 = 1$ we need to take $\beta_0 \geq 2 \cdot 1 = 2$, we take $\beta_0 = 4$. Since $k + 8d^{3/2} > (k+8d^{3/2})^{1/3}$, we can obtain:

$$\begin{aligned}&\mathbb{E}\left[f(x^k) - f(x^*)\right] \\ &= \mathcal{O}\left(\frac{LD^2 + d\sigma_f D/\tau + d\sigma_\nabla D + \sqrt{d}(f(x^0) - f(x^*))}{(k+8d^{3/2})^{1/3}} + \sqrt{d}LD\tau + \frac{\sqrt{d}\Delta D}{\tau}\right).\end{aligned}$$

This finishes the proof. □

Proof of Corollary 4. We aim to achieve precision ε, i.e.

$$\begin{aligned}
&\mathbb{E}\left[f(x^k) - f(x^*)\right] \\
&= \mathcal{O}\left(\frac{LD^2 + d\sigma_f D/\tau + d\sigma_\nabla D + \sqrt{d}(f(x^0) - f(x^*))}{(k + 8d^{3/2})^{1/3}} + \sqrt{d}LD\tau + \frac{\sqrt{d}\Delta D}{\tau}\right) \leq \varepsilon.
\end{aligned}$$

Therefore we need to take

$$N = \mathcal{O}\left(\max\left\{\left[\frac{LD^2 + d\sigma_\nabla D + \sqrt{d}(f(x^0) - f(x^*))}{\varepsilon}\right]^3, \frac{d^{9/2}\sigma_f^3 L^3 D^6}{\varepsilon^6}\right\}\right),$$
$$\tau = \mathcal{O}\left(\frac{\varepsilon}{\sqrt{d}LD}\right), \quad \Delta = \mathcal{O}\left(\frac{\varepsilon^2}{dLD^2}\right).$$

This finishes the proof. □

E PROOF OF CONVERGE RATE OF GD VIA JAGUAR (ALGORITHM 4)

E.1 RESULTS FOR THE NON-CONVEX CASE

Proof of Theorem 3. We start by writing our the result of Lemma 2 from Li et al. (2021). Under Assumption 6 the following inequality holds

$$\begin{aligned}
\mathbb{E}\left[f(x^{k+1}) - f(x^*)\right] \leq \mathbb{E}\left[f(x^k) - f(x^*)\right] - \frac{\gamma_k}{2}\mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2\right] \\
- \left(\frac{1}{2\gamma_k} - \frac{L}{2}\right)\mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right] + \frac{\gamma_k}{2}\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right].
\end{aligned} \tag{27}$$

Using results from Lemma 1:

$$\begin{aligned}
\mathbb{E}\left[\left\|h^{k+1} - \nabla f(x^{k+1})\right\|^2\right] \leq \left(1 - \frac{1}{2d}\right)\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right] + 2dL^2\mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right] \\
+ L^2\tau^2 + \frac{2\Delta^2}{\tau^2}.
\end{aligned} \tag{28}$$

Let us introduce notations for shortness:

$$r_k := \mathbb{E}\left[f(x^k) - f(x^*)\right], \;\; \psi_k := \mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right],$$
$$q_k := \mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2\right], \;\; p_k := \mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right], \;\; c := L^2\tau^2 + \frac{2\Delta^2}{\tau^2}.$$

Now equations (27) and (28) take form

$$\begin{aligned}
r_{k+1} &\leq r_k - \frac{\gamma_k}{2}q_k - \left(\frac{1}{2\gamma_k} - \frac{L}{2}\right)p_k + \frac{\gamma_k}{2}\psi_k, \\
\psi_{k+1} &\leq \left(1 - \frac{1}{2d}\right)\psi_k + 2dL^2 p_k + c.
\end{aligned}$$

Summarizing these two inequalities by multiplying the second by $\gamma_k d$ and introducing the notation $\Phi_k := r_k + \gamma_k d\psi_k$, we obtain:

$$\Phi_{k+1} \leq \Phi_k - \left(\frac{1}{2\gamma_k} - \frac{L}{2} - 2\gamma_k d^2 L^2\right) p_k - \frac{\gamma_k}{2} q_k + \gamma_k dc. \tag{29}$$

If we consider $\gamma_k \equiv \gamma_0 = \frac{1}{4dL}$, when

$$\frac{1}{2\gamma_k} - \frac{L}{2} - 2\gamma_k d^2 L^2 = 2dL - \frac{L}{2} - \frac{dL}{2} > 0.$$

Therefore equation (29) takes form

$$\Phi_{k+1} \leq \Phi_k - \frac{\gamma_0}{2} q_k + \gamma_0 dc. \tag{30}$$

Summing (30) from $k = 0$ to $k = N$ we obtain

$$0 \leq \Phi_{N+1} \leq \Phi_0 - \frac{\gamma_0}{2} \sum_{k=0}^{N} q_k + \gamma_0 dc(N+1).$$

Therefore

$$\sum_{k=0}^{N} \mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2\right] \leq \frac{2\Phi_0}{\gamma_0} + dc(N+1).$$

Since we choose $\widehat{x}_N$ chosen uniformly from $\{x^k\}_{k=0}^N$ we can obtain

$$\mathbb{E}\left[\|\nabla f(\widehat{x}_N)\|^2\right] \leq \frac{8Ld\Phi_0}{N+1} + dL^2\tau^2 + \frac{2d\Delta^2}{\tau^2}.$$

This finishes the proof. □

Proof of Corollary 2. We aim to achieve precision ε, i.e.

$$\mathbb{E}\left[\|\nabla f(\widehat{x}_N)\|^2\right] \leq \frac{8Ld\Phi_0}{N+1} + dL^2\tau^2 + \frac{2d\Delta^2}{\tau^2} \leq \varepsilon^2.$$

Therefore we need to take

$$N = \mathcal{O}\left(\frac{Ld\Phi_0}{\varepsilon^2}\right), \quad \tau = \mathcal{O}\left(\frac{\varepsilon}{\sqrt{d}L}\right), \quad \Delta = \mathcal{O}\left(\frac{\varepsilon\tau}{\sqrt{d}}\right) = \mathcal{O}\left(\frac{\varepsilon^2}{dL}\right).$$

This finishes the proof. □

E.2 Results for the PL-condition case

Proof of Theorem 4. We start by writing our the result of Lemma 5 from Li et al. (2021). Under Assumption 6 the following inequality holds

$$\begin{aligned}\mathbb{E}\left[f(x^{k+1}) - f(x^*)\right] \leq (1-\mu\gamma_k)\mathbb{E}\left[f(x^k) - f(x^*)\right] \\ - \left(\frac{1}{2\gamma_k} - \frac{L}{2}\right)\mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right] + \frac{\gamma_k}{2}\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right].\end{aligned} \tag{31}$$

Using results from Lemma 1 and introducing same notations as in proof of Theorem 3 equations (31) and (28) take form

$$
\begin{aligned}
r_{k+1} &\leq (1-\mu\gamma_k)r_k - \left(\frac{1}{2\gamma_k} - \frac{L}{2}\right)p_k + \frac{\gamma_k}{2}\psi_k, \\
\psi_{k+1} &\leq \left(1 - \frac{1}{2d}\right)\psi_k + 2dL^2 p_k + c.
\end{aligned}
$$

Summarizing these two inequalities by multiplying the second by $2\gamma_k d$ and introducing the notation $\Phi_k := r_k + 2\gamma_k d\psi_k$, we obtain:

$$
\Phi_{k+1} \leq (1-\mu\gamma_k)r_k + \left(1 - \frac{1}{4d}\right)2\gamma_k d\psi_k - \left(\frac{1}{2\gamma_k} - \frac{L}{2} - 4\gamma_k d^2 L^2\right)p_k + 2\gamma_k dc. \tag{32}
$$

If we consider $\gamma_k \equiv \gamma_0 = \frac{1}{4dL}$, when $1 - \frac{1}{4d} \leq 1 - \mu\gamma_0$, since $L \geq \mu$ and

$$
\frac{1}{2\gamma_0} - \frac{L}{2} - 4\gamma_0 d^2 L^2 = 2dL - \frac{L}{2} - dL > 0.
$$

Therefore equation (29) takes form

$$
\Phi_{k+1} \leq (1-\mu\gamma_0)\Phi_k + 2\gamma_0 dc. \tag{33}
$$

Using recursion and inequality (33) , we can obtain:

$$
\Phi_N = \mathcal{O}\left(\Phi_0 \exp\left[-\frac{\mu N}{4dL}\right] + \frac{dL^2\tau^2 + d\Delta^2/\tau^2}{\mu}\right).
$$

Using the fact that $\Phi_k \geq \mathbb{E}\left[f(x^k) - f(x^*)\right]$ and $h^0 = \widetilde{\nabla} f_\delta(x^0)$ we get the desired inequality:

$$
\mathbb{E}\left[f(x^N) - f(x^*)\right] = \mathcal{O}\left(r_0 \exp\left[-\frac{\mu N}{4dL}\right] + \frac{dL^2\tau^2 + d\Delta^2/\tau^2}{\mu}\right),
$$

where $r_0 = f(x^0) - f(x^*)$. This finishes the proof. □

Proof of Corollary 3. We aim to achieve precision ε, i.e.

$$
\mathbb{E}\left[f(x^N) - f(x^*)\right] = \mathcal{O}\left(r_0 \exp\left[-\frac{\mu N}{4dL}\right] + \frac{dL^2\tau^2 + d\Delta^2/\tau^2}{\mu}\right) \leq \varepsilon.
$$

Therefore we need to take

$$
N = \mathcal{O}\left(\frac{Ld}{\mu}\log\left[\frac{f(x^0) - f(x^*)}{\varepsilon}\right]\right), \quad \tau = \mathcal{O}\left(\frac{\sqrt{\varepsilon\mu}}{\sqrt{d}L}\right), \quad \Delta = \mathcal{O}\left(\frac{\sqrt{\varepsilon\mu}\tau}{\sqrt{d}}\right) = \mathcal{O}\left(\frac{\varepsilon\mu}{dL}\right).
$$

This finishes the proof. □