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

A PREPRINT

Alexander I. Bogdanov

bogdanov.ai@phystech.edu

Alexander N. Beznosikov

beznosikov.an@phystech.edu

ABSTRACT

In this paper, we address the challenges of solving a convex, non-smooth stochastic minimization problem subject to constraints, which often arises in modern machine learning applications. We propose a gradient-free momentum-based Frank-Wolfe algorithm that incorporates one-point and two-point feedback mechanisms, which have been shown to significantly improve convergence speed and robustness to noise, especially in the presence of large-scale and high-dimensional data. Specifically, our algorithm leverages the momentum term to accelerate convergence and overcome oscillations, while the feedback mechanism helps to adaptively adjust the step size, improving the overall performance. We provide theoretical analysis and numerical experiments to demonstrate the effectiveness and efficiency of our algorithm on various optimization problems, comparing it with existing algorithms such as the classical Frank-Wolfe and stochastic gradient descent methods.

Keywords Gradient-free methods · Zeroth-order methods · Stochastic optimization · Frank-Wolfe algorithms · Momentum-based method

1 Introduction

It will be rewritten because the topic has changed in the process.

2 Related work

It will be rewritten because the topic has changed in the process.

3 Main results

3.1 Jaguar gradient approximation. Non-stochastic case

In this section we consider non-stochastic optimization problem

$$\min_{x \in Q} \quad f(x). \tag{1}$$

where $Q \subseteq \mathbb{R}^d$ is arbitrary set.

We assume that we have access only to zero-order oracle, i.e. we can only get values of functions $f(x)$, not of the gradient $\nabla f(x)$. This means that we need to approximate gradient of the function $f(x)$.

In practice, however, we usually do not even have access to $f(x)$, but only to its noisy version, i.e., we assume that zero-order oracle the returns to us the noisy values of the function $f(x)$ based on the point x given to it, i.e. zero-order oracle returns $f_\delta(x) := f(x) + \delta(x)$.

We define such difference scheme that will be used in our algorithm of gradient approximation

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

where e_i is i-th basis vector from the standard basis in space $\mathbb{R}^d$, γ is a smoothing parameter.

How we can present algorithm of gradient approximation in the point x^k, where x^k is a point on a step k of any algorithm, that solves problem (1):

Algorithm 1 JAGUAR gradient approximation. Non-stochastic case

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

This approximation algorithm can be used for various iterative methods that, at each step k, obtain a new point x^k that converges to the solution x^* of the problem 1. Then we will also obtain the sequence h^k of Line 4, which serves in a sense as a memory of the gradients from the previous iterations.

We provide several assumptions required for the analysis:

Assumption 1 (Smoothness)**.** *The function $f(x)$ is L-smooth on a set Q, i.e.*

$$\forall x, y \in Q \hookrightarrow \|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|,$$

where $\|\cdot\|$ is the standard Euclidean norm. We will use this notation throughout the paper.

Because zero-order oracle returns to us noisy values of the function $f(x)$ we make common assumption on this noise.

Assumption 2 (Bounded oracle noise)**.**

$$\exists \Delta > 0 : \ \forall x \in Q \hookrightarrow |\delta(x)|^2 \leq \Delta^2$$

Now we start to analyze convergence of JAGUAR gradient approximation 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. However, first we need to introduce some auxiliary lemmas.

Lemma 1. *Let us introduce the auxiliary notation*

$$\widetilde{\nabla} f_\delta(x) := \sum_{i=1}^{d} \frac{f_\delta(x+\gamma e_i) - f_\delta(x-\gamma e_i)}{2\gamma} e_i. \tag{3}$$

Under Assumptions 1 and 2 the following inequality holds

$$\left\| \widetilde{\nabla} f_\delta(x) - \nabla f(x) \right\|^2 \leq dL^2\gamma^2 + \frac{2d\Delta^2}{\gamma^2}. \tag{4}$$

For a detailed proof of Lemma 1, see proof of Lemma 3 in Appendix C in case $\sigma_\nabla^2 = \sigma_f^2 = 0$ (see details in the Section 3.2). The $\widetilde{\nabla} f_\delta(x)$ is a more precise version of approximation (2) as it approximates the gradient in every coordinate while (2) approximates only one, however approximation (3) requires d times more zero-order oracle calls than (2). However, we will show that utilising memory from previous iterations in the form of introducing a variable h^k into our Algorithm 1 in line 4 will achieve the same accuracy as for $\widetilde{\nabla} f_\delta(x)$, but will only require two calls to the zero-order oracle at each iteration.

Lemma 2. *Under Assumptions 1 and 2 the following inequality holds*

$$\mathbb{E}\left[\left\|h^{k+1}-\nabla f(x^{k+1})\right\|^2\right] \leq \left(1-\frac{1}{2d}\right)\mathbb{E}\left[\left\|h^{k}-\nabla f(x^{k})\right\|^2\right] + 2dL^2\mathbb{E}\left[\left\|x^{k+1}-x^{k}\right\|^2\right] + L^2\gamma^2 + \frac{2\Delta^2}{\gamma^2} \tag{5}$$

For a detailed proof of Lemma 2, see proof of Lemma 4 in Appendix C in case $\sigma_\nabla^2 = \sigma_f^2 = 0$ (see details in the Section 3.2). Let us analyse formula (5), and show that this result in the same estimate as in (4). In many algorithms of optimization we can consider that

$$\left\|x^{k+1}-x^{k}\right\|^2 \leq \gamma_k^2 D^2, \tag{6}$$

where γ_k is an optimizer step and D^2 is a constant, which depends on the optimisation algorithm. In the next section, we consider the Frank-Wolfe algorithm, where the diameter of the set Q stands for D.

Theorem 1 (Step tuning for JAGUAR. Non-stochastic case)**.** *Consider Assumptions 3 and 4. If equation* (6) *is true, then we can take*

$$\gamma_k = \frac{4}{k+8d},$$

then we following convergence rate hold:

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

If $h_0 = \widetilde{\nabla} f_\delta(x^0) = \sum_{i=1}^d \frac{f_\delta(x+\gamma e_i) - f_\delta(x-\gamma e_i)}{2\gamma} e_i$ *we can obtain*

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

For detailed proof of Theorem 3 see Appendix B. From Theorem 1 we can conclude that after $\mathcal{O}\left(\frac{\sqrt{d}D}{\gamma} - d\right)$ steps we get exactly the same estimate as in equation (4), but at each step except $k=0$ in algorithm JAGUAR 1 we make two calls to the zero-order oracle, and to get estimate (4) we needed to make $2d$ oracle calls each step.

In the next section, we consider the more general stochastic problem (7). In this problem, we can no longer use h^k as an approximation of the gradient, since it is in some sense biased.

3.2 Jaguar gradient approximation. Stochastic case

In this section we consider stochastic optimization problem

$$\min_{x\in Q} \quad f(x) := \mathbb{E}_\xi\left[f(x,\xi)\right], \tag{7}$$

where $Q \subseteq \mathbb{R}^d$ is arbitrary set.

In this Section, we also assume that we do not have access to the true value of the gradient $\nabla f(x,\xi)$, we only have access to the zero-order oracle, which returns the noisy value of the function $f(x,\xi)$: $f_\delta(x,\xi) := f(x,\xi) + \delta(x,\xi)$.

In two point feedback (tpf) we define such gradient approximations of function $f(x)$:

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

where e_i is i-th basis vector from the standard basis in space $\mathbb{R}^d$, γ is a smoothing parameter. In one point feedback (opf) we define slightly different gradient approximations function $f(x)$:

$$\widetilde{\nabla}_i f_\delta(x, \xi^+, \xi^-) := \frac{f_\delta(x + \gamma e_i, \xi^+) - f_\delta(x - \gamma e_i, \xi^-)}{2\gamma} e_i \tag{9}$$

The key difference between approximations (8) and (9) is that Scheme (8) is more accurate, but it is difficult to implement in practice because we have to get the same realization of ξ at two different points $x + \gamma e$ and $x - \gamma e$, then Scheme (9) is more interesting from a practical point of view.

To simplify further, we will assume that in the case of tpf we will have the same inscription as in opf, but only $\xi^+ = \xi^- = \xi$.

How we can present algorithm of gradient approximation in the point x^k, where x^k is a point on a step k of any algorithm, that solves problem (7):

Algorithm 2 JAGUAR gradient approximation. Stochastic case

1: **Input:** $x, h, g \in \mathbb{R}^d$, $0 \leq \eta \leq 1$
2: Sample $i \in \overline{1, d}$ independently and uniform
3: Sample 2 realizations of ξ: ξ^+ and ξ^- independently (in tpf $\xi^+ = \xi^-$)
4: Compute $\widetilde{\nabla}_i f_\delta(x, \xi^+, \xi^-) = \frac{f_\delta(x+\gamma e_i, \xi^+) - f_\delta(x-\gamma e_i, \xi^-)}{2\gamma} e_i$
5: $h = h - \langle h, e_i \rangle e_i + \widetilde{\nabla}_i f_\delta(x, \xi^+, \xi^-)$
6: $\rho = h - d \cdot \langle h, e_i \rangle e_i + d \cdot \widetilde{\nabla}_i f_\delta(x, \xi^+, \xi^-)$
7: $g = (1 - \eta) g + \eta \rho$
8: **Output:** h and g

This Algorithm is similar to 1, but in Lines 6 and 7 we need to use SEGA and momentum parts in order to converge to the stochastic case. When applying this gradient approximation Algorithm 2 to various iterative methods, we will now have not one additional h^k sequence, but three: h^k, g^k, and η_k.

We provide several assumptions required for the analysis:

Assumption 3 (Smoothness)**.** *The functions $f(x, \xi)$ are $L(\xi)$-smooth on a set Q, i.e.*

$$\forall x, y \in Q \hookrightarrow \|\nabla f(x, \xi) - \nabla f(y, \xi)\| \leq L(\xi) \|x - y\|.$$

And exists constant L^2 such that

$$L^2 := \mathbb{E}\left[L(\xi)^2\right].$$

If Assumption 3 holds, then function $f(x)$ is L-smooth on a set Q, since for all $x, y \in Q$ holds that

$$\|\nabla f(x) - \nabla f(y)\|^2 = \|\mathbb{E}\left[\nabla f(x, \xi) - \nabla f(y, \xi)\right]\|^2 \leq \mathbb{E}\left[\|\nabla f(x, \xi) - \nabla f(y, \xi)\|^2\right] \leq L^2 \|x - y\|^2.$$

Because zero-order oracle returns to us noisy values of the function $f(x, \xi)$ we make common assumption on this noise.

Assumption 4 (Bounded oracle noise)**.**

$$\exists \Delta > 0 : \ \forall x \in Q \hookrightarrow \mathbb{E}\left[|\delta(x, \xi)|^2\right] \leq \Delta^2$$

If Assumption 4 holds, then if we define $\delta(x) := \mathbb{E}\left[\delta(x, \xi)\right]$, then it holds that $|\delta(x)|^2 \leq \Delta^2$, since

$$|\delta(x)|^2 = |\mathbb{E}\left[\delta(x, \xi)\right]|^2 \leq \mathbb{E}\left[|\delta(x, \xi)|^2\right] \leq \mathbb{E}\left[\Delta^2\right] = \Delta^2.$$

Assumptions 3 and 4 are similar to Assumptions 1 and 2, but we need to add random variable ξ since we consider stochastic problem (7). Now we present two assumptions that are needed only in stochastic case.

Assumption 5 (Bounded second moment of gradient)**.**

$$\exists \sigma_\nabla^2 : \mathbb{E}\left[\|\nabla f(x, \xi) - \nabla f(x)\|^2\right] \leq \sigma_\nabla^2$$

Assumption 6 (Bounded second moment of function)**.**

$$\exists \sigma_f^2 : \mathbb{E}\left[|f(x, \xi) - f(x)|^2\right] \leq \sigma_f^2$$

If two point feedback (8) we will not need Assumption 6, so for simplicity of future exposition we will assume that in the case of tpf Assumption 6 is fulfilled with $\sigma_f^2 = 0$.

Now we start to analyze convergence of JAGUAR gradient approximation 2 in stochastic case. First two lemmas of our analysis will be similar to Lemmas 1 and 2.

Lemma 3. *Let us introduce the auxiliary notation*

$$\widetilde{\nabla} f_\delta(x, \xi_1^+, \xi_1^-, ..., \xi_d^+, \xi_d^-) := \sum_{i=1}^{d} \frac{f_\delta(x + \gamma e_i, \xi_i^+) - f_\delta(x - \gamma e_i, \xi_i^-)}{2\gamma} e_i. \tag{10}$$

In two point feedback (8) $\xi_j^+ = \xi_j^-$.

Under Assumptions 3, 4, 5 and 6 in opf case (9) *the following inequality holds*

$$\mathbb{E}\left[\left\|\widetilde{\nabla} f_\delta(x, \xi_1^+, \xi_1^-, ..., \xi_d^+, \xi_d^-) - \nabla f(x)\right\|^2\right] \leq dL^2\gamma^2 + \frac{8d\sigma_f^2}{\gamma^2} + 2d\sigma_\nabla^2 + \frac{2d\Delta^2}{\gamma^2}. \tag{11}$$

In two point feedback (8) $\sigma_f^2 = 0$.

For a detailed proof of Lemma 3, see Appendix C.

Lemma 4. *Under Assumptions 3, 4, 5 and 6 in opf case* (9) *the following inequality holds*

$$\begin{aligned}\mathbb{E}\left[\left\|h^{k+1} - \nabla f(x^{k+1})\right\|^2\right] \leq \left(1 - \frac{1}{2d}\right)\mathbb{E}\left[\left\|h^k - \nabla f(x^k)\right\|^2\right] + 2dL^2\mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right] \\ + L^2\gamma^2 + \frac{8\sigma_f^2}{\gamma^2} + 2\sigma_\nabla^2 + \frac{2\Delta^2}{\gamma^2}\end{aligned} \tag{12}$$

In two point feedback (8) $\sigma_f^2 = 0$

For a detailed proof of Lemma 4, see Appendix C.

This Lemmas 3 and 4 are similar to Lemmas 1 and 2, but summands with σ_∇^2 and σ_f^2 (in opf), which are related to stochasticity, interfere with the convergence of our Algorithm 2. It is for this reason that we need to introduce SAGA and moment correction in the form of a variables ρ^k and g^k in the Algorithm 2 step on the lines 6 and 7.

Lemma 5. *Under Assumptions 3, 4, 5 and 6 in opf case* (9) *the following inequality holds*

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

In two point feedback (8) $\sigma_f^2 = 0$

For a detailed proof of Lemma 5, see Appendix C. As we can see, using SEGA line 6 deteriorates our estimates by a factor of d compared to using h^k as a gradient approximator, but in the stochastic case we care about the unbiased property, i.e.

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

Therefore we have to accept this factor.

Lemma 6. *Under Assumption 3 the following inequality holds*

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

For a detailed proof of Lemma 6, see Appendix C.

Theorem 2 (Step tuning for JAGUAR. Stochastic case)**.** *Consider Assumptions 3, 4, 5 and 6 in opf case* (9)*.*

$$\gamma_k = \frac{4}{k + 8d^{3/2}} \quad \textit{and} \quad \eta_k = \frac{4}{(k+8d^{3/2})^{2/3}},$$

If equation (6) *is true, then the following inequality holds:*

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

If $h^0 = g^0 = \widetilde{\nabla} f_\delta(x^0, \xi_1^+, \xi_1^-, ..., \xi_d^+, \xi_d^-) = \sum_{i=1}^d \frac{f_\delta(x+\gamma e_i, \xi_i^+) - f_\delta(x-\gamma e_i, \xi_i^-)}{2\gamma} e_i$ *we can obtain*

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

In two point feedback (8) $\sigma_f^2 = 0$*.*

For a detailed proof of Theorem 2, see Appendix C.

3.3 Frank-Wolfe via JAGUAR

In this section we consider the minimisation problem on the set Q. We make common assumptions required for analysis

Assumption 7 (Convex)**.** *The objective function $f(x)$ is convex on a set Q, i.e.*

$$\forall x, y \in Q \hookrightarrow f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$$

Assumption 8 (Compact domain)**.** *The set Q is compact, i.e.*

$$\exists D > 0 : \ \forall x, y \in Q \hookrightarrow \|x - y\| \leq D$$

Now we can introduce Frank-Wolfe algorithm using JAGUAR approximation of the gradient

Algorithm 3 FW via JAGUAR

Input, non-stochastic case: $x_0 \in Q$, $g^0 = \widetilde{\nabla} f(x^0)$
Input, stochastic case: $x_0 \in Q$, $h^0 = g^0 = \widetilde{\nabla} f(x^0, \xi_1^+, ..., \xi_d^-)$, $\{\eta_k\}_{k=0}^N \subset [0;1]$
for k = 0, 1, 2, ... , N **do**
  $g^{k+1} = \text{JAGUAR}\left(x^k, h^k\right)$ ▷ Non-stochastic case
  $h^{k+1}, g^{k+1} = \text{JAGUAR}\left(x^k, h^k, g^k, \eta_k\right)$ ▷ Stochastic case
  $s^k = \arg\min_{x \in Q} \left\{\left\langle s, g^{k+1}\right\rangle\right\}$
  $x^{k+1} = x^k + \gamma_k(s^k - x^k)$
end for
Output: x^{N+1}

In Lines 4 and 5 we use JAGUAR Algorithm 1 if we consider non-stochastic problem (1) and Algorithm 2 if we consider stochastic problem (7). We have denoted the variable returned by the non-stochastic JAGUAR approximation 1 in this algorithm as g^k, although in Section 3.1 we denoted it as h^k, this was done for the general form of the Line 6.

We now explore the convergence of Algorithm 3 in two problems (1) and (7).

Theorem 3 (Convergence rate of FW via JAGUAR 3. Non-Stochastic case)**.** *Consider Assumptions 1, 2, 7 and 8. If we take*

$$\gamma_k = \frac{4}{k+8d},$$

then we FW via JAGUAR Algorithm 3 in non-stochastic case (1) *has the following convergence rate*

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

For a detailed proof of Theorem 3, see Appendix D.

Corollary 1. *Let Assumptions from Theorem 3 be satisfied, then Algorithm 3 in non-stochastic case* (1) *has the following convergence rate*

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

where ε is desired accuracy, i.e. $\mathbb{E}\left[f(x^N) - f(x^*)\right] \leq \varepsilon$.

For a detailed proof of Corollary 1, see Appendix D.

Theorem 4 (Convergence rate of FW via JAGUAR 3. Stochastic case)**.** *Consider Assumptions 3, 4, 5, 7, 8 and 6 in opf case* (9)*. If we take*

$$\gamma_k = \frac{4}{k+8d^{3/2}} \quad \textit{and} \quad \eta_k = \frac{4}{(k+8d^{3/2})^{2/3}},$$

then we FW via JAGUAR Algorithm 3 has the following convergence rate

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

In two point feedback (8) *$\sigma_f^2 = 0$.*

For a detailed proof of Theorem 4, see Appendix D.

Corollary 2. *Let Assumptions from Theorem 4 be satisfied, then Algorithm 3 in stochastic case* (7) *has the following convergence rate*

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

where ε is desired accuracy, i.e. $\mathbb{E}\left[f(x^N) - f(x^*)\right] \leq \varepsilon$. In two point feedback (8) *$\sigma_f^2 = 0$ and convergence on N takes form*

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

For a detailed proof of Corollary 2, see Appendix D.

4 Experiments

It hasn't been written yet, because the code doesn't work.

References

[1] 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.

Appendix

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 7. *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 8 (Recursion Lemma)**.** *Suppose we have the following recurrence relation for variables $\{r_k\}_{k=0}^N \subset \mathbb{R}$*

$$r_{k+1} \leq \left(1 - \frac{\beta_0}{(k + k_0)^{\alpha_0}}\right) r_k + \sum_{i=1}^{m} \frac{\beta_i}{(k + k_0)^{\alpha_i}}, \tag{13}$$

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:*

$$r_k \leq 2 \cdot \sum_{i=1}^{m} \frac{Q_i}{(k+k_0)^{\alpha_i - \alpha_0}}, \tag{14}$$

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 (14) 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 (14) holds for some k. Now we will show that this condition will hold for $k+1$.

We start by fitting (14) into the original recurrence relation (13) 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

$$\frac{2Q_i}{(k+k_0)^{\alpha_i-\alpha_0}} - \frac{Q_i\beta_0}{(k+k_0)^{\alpha_i}} \leq \frac{2Q_i}{(k+k_0+1)^{\alpha_i-\alpha_0}}. \tag{15}$$

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 7 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 (15) 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:

$$\frac{2}{\beta_0}\max\{1,\alpha_i-\alpha_0\}\leq(k+k_0)^{1-\alpha_0}. \tag{16}$$

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 (16) desired inequality (15) will be fulfilled for all $i\in\overline{1,m}$ for all $\beta_0>0$.
- If $\alpha_0=1$, then inequality (16) 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 (16) desired inequality (15) will be fulfilled for all $i\in\overline{1,m}$ for all $k_0\in\mathbb{N}$.

This finishes the proof.

□

B Proof of converge rate of JAGUAR Algorithm 2. Non-stochastic case.

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

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

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

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

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

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

This finishes the proof. □

C Proof of converge rate of JAGUAR Algorithm 2. Stochastic case.

proof of Lemma 3. Let's start by writing out a definition of gradient approximation (10):

$$\begin{aligned}
\mathbb{E}\left[\left\|\widetilde{\nabla} f_\delta(x,\xi_1^+,\xi_1^-,...,\xi_d^+,\xi_d^-)-\nabla f(x)\right\|^2\right] &= \mathbb{E}\left[\left\|\sum_{i=1}^d \frac{f_\delta(x+\gamma e_i,\xi_i^+)-f_\delta(x-\gamma e_i,\xi_i^-)}{2\gamma}e_i-\nabla f(x)\right\|^2\right]\\
&= \mathbb{E}\left[\left\|\sum_{i=1}^d \left(\frac{f_\delta(x+\gamma e_i,\xi_i^+)-f_\delta(x-\gamma e_i,\xi_i^-)}{2\gamma}-\langle\nabla f(x),e_i\rangle\right)e_i\right\|^2\right]\\
&\overset{(\star)}{=} \sum_{i=1}^d \mathbb{E}\left[\left\|\left(\frac{f_\delta(x+\gamma e_i,\xi_i^+)-f_\delta(x-\gamma e_i,\xi_i^-)}{2\gamma}-\langle\nabla f(x),e_i\rangle\right)e_i\right\|^2\right]\\
&= \sum_{i=1}^d \mathbb{E}\left[\left|\frac{f_\delta(x+\gamma e_i,\xi_i^+)-f_\delta(x-\gamma e_i,\xi_i^-)}{2\gamma}-\langle\nabla f(x),e_i\rangle\right|^2\right]
\end{aligned}$$

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:

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

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

$$\begin{aligned}
\mathbb{E}\left[\left|f(x+\gamma e_i,\xi_i^+)-f(x-\gamma e_i,\xi_i^-)-\langle\nabla f(x),2\gamma e_i\rangle\right|^2\right] &\leq 4\mathbb{E}\left[\left|f(x+\gamma e_i,\xi_i^+)-f(x,\xi_i^+)-\left\langle\nabla f(x,\xi_i^+),\gamma e_i\right\rangle\right|^2\right]\\
&\quad +4\mathbb{E}\left[\left|-f(x-\gamma e_i,\xi_i^-)+f(x,\xi_i^-)+\left\langle\nabla f(x,\xi_i^-),-\gamma e_i\right\rangle\right|^2\right]\\
&\quad +4\mathbb{E}\left[\left|f(x,\xi_i^+)-f(x,\xi_i^-)\right|^2\right]\\
&\quad +4\mathbb{E}\left[\left|\left\langle\nabla f(x,\xi_i^+)+\nabla f(x,\xi_i^-)-2\nabla f(x),\gamma e_i\right\rangle\right|^2\right]
\end{aligned} \tag{17}$$

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:

$$\begin{aligned}
\left|f(x+\gamma e_i,\xi_i^+)-f(x,\xi_i^+)-\left\langle\nabla f(x,\xi_i^+),\gamma e_i\right\rangle\right| &\leq \frac{L(\xi_i^+)}{2}\gamma^2 \leq \frac{L}{2}\gamma^2\\
\left|-f(x-\gamma e_i,\xi_i^-)+f(x,\xi_i^-)+\left\langle\nabla f(x,\xi_i^-),-\gamma e_i\right\rangle\right| &\leq \frac{L(\xi_i^-)}{2}\gamma^2 \leq \frac{L}{2}\gamma^2
\end{aligned} \tag{18}$$

If we consider tpf approximation (8), then third term in (17) equals to zero, since $\xi_i^+ = \xi_i^-$, if we consider opf case (9), then we can obtain

$$\mathbb{E}\left[\left|f(x,\xi_i^+)-f(x,\xi_i^-)\right|^2\right] \le 2\mathbb{E}\left[\left|f(x,\xi_i^+)-f(x)\right|^2\right] + 2\mathbb{E}\left[\left|f(x,\xi_i^-)-f(x)\right|^2\right] \le 4\sigma_f^2 \tag{19}$$

Consider the last point in (17) and using Cauchy–Schwarz inequality A.2 we can obtain:

$$\mathbb{E}\left[\left|\left\langle \nabla f(x,\xi_i^+)-\nabla f(x), \gamma e_i\right\rangle\right|^2\right] \le \mathbb{E}\left[\left\|\nabla f(x,\xi_i^+)-\nabla f(x)\right\|^2\gamma^2\right] \le \sigma_\nabla^2\gamma^2 \tag{20}$$

Combining (18), (19) and (20) we obtain

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

In two point feedback (8) $\sigma_f^2 = 0$. □

proof of Lemma 4. Let's start by writing out a definition of h^k using line 5 of Algorithm 2

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

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

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

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

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

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:

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

Using Lemma 3 we obtain

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

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

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

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

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

□

Proof of Lemma 5. Let's start by writing out a definition of ρ^k using line 6 of Algorithm 2

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

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

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

This finishes the proof.

□

Proof of Lemma 6. We start by writing out a definition of g^k using line 7 of Algorithm 2

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

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

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

Fact that for $\widetilde{\nabla} f_\delta(x)$ Lemma 3 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 3 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.

□

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

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

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

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

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

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

Consider $\mathbb{E}\left[\left\|\rho^k - \nabla f(x^k)\right\|^2\right]$. Using Lemmas 5 and 3 we obtain

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

Consider $\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right]$. We write out result from Lemma 6 and setting up $\eta_k = \frac{4}{(k+8d^{3/2})^{2/3}}$:

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

Using Lemma 8 with $\alpha_0 = 2/3, \beta_0 = 4$; $\alpha_1 = 4/3, \beta_1 = 4L^2D^2$; $\alpha_2 = 4/3, \beta_2 = 4d^2L^2\gamma^2 + \frac{4d^2\sigma_f^2}{\gamma^2} + 4d^2\sigma_\nabla^2 + \frac{4d^2\Delta^2}{\gamma^2}$; $\alpha_3 = 10/3, \beta_3 = 4d^3 \max\{L^2D^2, d\left\|h^0 - \nabla f(x^0)\right\|^2\}$; $\alpha_4 = 2/3, \beta_4 = dL^2\gamma^2 + \frac{d\Delta^2}{\gamma^2}$ and $i^* = 2$ we get:

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

Since

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

we can simplify (25):

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

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

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

This finishes the proof. □

D Proof of converge rate of FW via JAGUAR Algorithm 3.

Proof of Theorem 3. We start by writing our the result of Lemma 2 from [1]. Under Assumptions 3, 7 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

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

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

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

In Lemma 8 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.

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

Therefore we need to take

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

□

Proof of Theorem 4. Again we write out result of Lemma 2 from [1]:

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

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 2 we can obtain

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

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

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

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

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

In Lemma 8 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:

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

This finishes the proof.

□

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

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

Therefore we need to take

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

$$\gamma = \mathcal{O}\left(\frac{\varepsilon}{\sqrt{d}LD}\right), \quad \Delta = \mathcal{O}\left(\frac{\varepsilon^2}{dLD^2}\right).$$

□