Very nice report, well done
A few typing mistakes
Can be published in Intl. conference with about 2-3 months of work

One Sample Stochastic Subspace Frank-Wolfe Algorithm

Ritik Saxena (180617) || Anubhav Kumar (180121)

Abstract

The advantage of using projected gradient descent is its simple projecting operator which solves the minimizing problem off-grid. It becomes quite stable and fast in the stochastic setting which has been widely used in machine learning applications. However, projection-free algorithms are faster than Projection Gradient descent particularly Frank Wolfe Algorithm as it solves the problem linearly. Recently it has also been shown that even in a stochastic setting Frank-Wolfe can get optimal results in just one stochastic sample (1). On a different note, there have been works in improving per iterate time complexity as in the Subspace Stochastic Descent(2) by carrying out updates in a lower dimension. Our aim in this paper is to incorporate this idea of update in the subspace into One sample stochastic Frank-Wolfe. By doing this, the per iterate time complexity will improve as the bottleneck step of Frank Wolfe will be carried out in a lower dimension making it faster. In this paper, we propose **two** different algorithms that can be used in an oblivious setting.
The first algorithm recovers the optimality convergence rate of $\mathcal{O}(\frac{1}{\sqrt{T}})$ with the same oracle i.e. $\nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)$. In the second algorithm proposed, the optimality rate can be recovered by making a reasonable choice of dimensionality reduction factor which has been shown in the experiments. The Oracle in the second algorithm used is reduced to $\mathbf{P}^T\nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)$ (where P is a Haar matrix of **d** x **l**). Thus both of them recover the convergence rate as in One Sample Stochastic Frank-Wolfe in oblivious setting for convex problems in constrained sets.

1 Introduction

Many strongly convex optimization problems are solved faster with Frank-Wolfe than projected gradient descent. To extend such methods in the stochastic setting is a very challenging task because of its high sensitivity in calculating gradient. Now by carrying out the bottleneck step in lower-dimensional subspace in One Sample Stochastic Frank-Wolfe improves per iterate time complexity. There is perturbance in the optimality gap in the second algorithm. However the proper choice of dimension, the perturbances can be kept low and one can reap benefits of both the worlds (optimal convergence rate and lesser computational cost.)
A one-sample stochastic Frank Wolfe Method has been shown to achieve optimal complexity bounds in both oblivious and non-oblivious setting(1). In this paper, we will try to answer the question.

Can the optimal complexity bounds for a stochastic variant of Sub-space Frank-Wolfe be achieved while using a single stochastic sample per iteration?

In this paper, we have studied the problem in an oblivious setting and **proposed two algorithms** and successfully proved convergence for Algorithm A and analysed Algorithm B both seconded by promising experimental results.

2 Related Works

Several works have been done before in projection free methods. Many of them are used in machine learning applications and signal processing. Some of the works that are related to our algorithms are as follows:

2.1 One Sample Stochastic Frank Wolfe(1)

This algorithm is given according to the reference (1). We have shown the algorithm for oblivious case below :
Input: Step sizes $\rho \in (0,1), \eta \in (0,1)$, initial point $\mathbf{x} \in \mathcal{X}$, total number of iterations T
Output: $\mathbf{x}_{T+1}$

```
for $t = 1, 2, 3, ..., T$ do
        Sample a point $\mathbf{z}_1 \sim p(\mathbf{z})$
    if $t = 1$ then
        Compute $\mathbf{d}_1 = \nabla\tilde{F}(\mathbf{x}_1; \mathbf{z}_1)$
    else
        $\tilde{\Delta}_t = \nabla\tilde{F}(\mathbf{x}_t; \mathbf{z}_t) - \nabla\tilde{F}(\mathbf{x}_{t-1}; \mathbf{z}_t)$
        $\mathbf{d}_t = (1 - \rho_t)(\mathbf{d}_{t-1} + \tilde{\Delta}_t) + \rho_t \nabla\tilde{F}(\mathbf{x}_t; \mathbf{z}_t)$
    end if
    $\mathbf{u}_t = argmin_{\mathbf{x} \in \mathcal{X}_2} \langle \mathbf{d}_t, \mathbf{u} \rangle$
    $\mathbf{x}_{t+1} = (1 - \eta_t)\mathbf{x}_t + \eta_t \mathbf{u}_t$
    return $\mathbf{x}_{T+1}$
```

The paper (1) on One-Sample Stochastic Frank-Wolfe is the main reference where we got the idea to incorporate subspace. They successfully showed that we can even use just One Sample to get optimal convergence. It is very fast and stable. We will compare our algorithm with this as a benchmark in the experiments. **We have tried to make our Algorithm of One-Sample Subspace Frank Wolfe better than One sample Stochastic Frank Wolfe by carrying out the *arg* minimization in lower dimension and still maintaining the optimal convergence rate of $\mathcal{O}(\frac{1}{\sqrt{T}})$. We will see in the experiments that our algorithm takes same number of iterations in order to converge as the One-Sample Stochastic Frank Wolfe.**

2.2 Stochastic Subspace Descent(2)

This is one of the simplest algorithms in the subspace domain. It uses a randomly Haar generated matrix(2) which multiplied with its transpose has an expectation value of $\mathbb{I}$. The update for the algorithm is given below:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{P}_t \mathbf{P}_t^T \nabla F(\mathbf{x}_t)$$

where $\mathbf{P}$ is randomly generated Haar Matrix generated as given in subroutine of Section 4. This algorithm has consistency with the simple gradient descent algorithm. It is possible because of the property of the Haar matrix *i.e.* :

$$\mathbb{E}[\mathbf{P}_t \mathbf{P}_t^T \nabla F(\mathbf{x}_t] = \nabla F(\mathbf{x}_t)$$

This allows the change of oracle to $\mathbf{P}^T \nabla F(\mathbf{x})$ and the need for a d dimensional gradient gets removed. Now the oracle is giving a **l** dimensional projection of gradient. This fascinated us and we thought if this could be used in Frank Wolfe settings and the algorithm stays convergent with **same order complexity**, then the total time complexity gets **automatically improved** as bottleneck is carried out in **l** dimensional subspace rather than **d** dimensional subspace and **l<d**.

3 Problem Formulation

The stochastic optimization problem is below:

$$min_{x\in\mathcal{X}}F(\mathbf{x}) = min_{x\in\mathcal{X}}\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[\tilde{F}(\mathbf{x};\mathbf{z})] \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^d, \mathbf{x} \subseteq \mathcal{X}$ is the convex constraint set, and the objective function $F : \mathbb{R}^d \to \mathbb{R}$ is defined as expectation over a set of functions $\tilde{F} : \mathbb{R}^d \times \mathbb{Z} \to \mathbb{R}$ and $\mathbf{z} \sim p(\mathbf{z})$ where $\mathbf{z} \in \mathbb{Z}$.

3.1 Assumptions

The stochastic function $\tilde{F}(\mathbf{x};\mathbf{z})$ satisfies the following:

Assumption A1. *The constraint set $\mathcal{X} \subseteq \mathbb{R}^d$ is D compact i.e.*

$$D = max_{\mathbf{x},\mathbf{y}\in\mathcal{X}}\|\mathbf{x}-\mathbf{y}\| \tag{2}$$

Assumption A2. *The stochastic function $\tilde{F}(\mathbf{x};\mathbf{z})$ has uniformly bound function value ,i.e., $\forall\, \mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{Z}$*

$$\tilde{F}(\mathbf{x};\mathbf{z}) \le B \tag{3}$$

Assumption A3. *The stochastic function $\tilde{F}(\mathbf{x};\mathbf{z})$ has uniformly bound gradients ,i.e, $\forall\, \mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{Z}$*

$$\|\nabla\tilde{F}(\mathbf{x};\mathbf{z})\| \le G \tag{4}$$

Moreover, the function $\tilde{F}$ is $\tilde{L}$ smooth ,i.e., $\forall\, \mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{Z}$

$$\|\nabla\tilde{F}(\mathbf{x};\mathbf{z}) - \nabla\tilde{F}(\mathbf{y};\mathbf{z})\| \le \tilde{L}\|\mathbf{x}-\mathbf{y}\| \tag{5}$$

And, the objective function is L smooth ,i.e., $\forall\, \mathbf{x} \in \mathcal{X}$

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \le L\|\mathbf{x}-\mathbf{y}\| \tag{6}$$

$\mathbf{P} \in \mathbb{R}^{d\times l}$ is a random Haar distributed matrix (2) that projects the gradient onto l dimensional subspace and satisfies the following assumptions:

Assumption B1.

$$\mathbb{E}[\mathbf{P}\mathbf{P}^T] = \mathbf{I}_d \tag{7}$$

Assumption B2.

$$\mathbf{P}^T\mathbf{P} = \frac{d}{l}\mathbf{I}_l \tag{8}$$

3.2 Oblivious setting

In oblivious setting, we have the probability density function of $\mathbf{z}$ is independent of $\mathbf{x}$. Thus we have, $\mathbf{z} \sim p(\mathbf{z})$ and not $\mathbf{z} \sim p(\mathbf{z};\mathbf{x})$. We use $\tilde{\Delta_t} = \nabla\tilde{F}(\mathbf{x}_t;\mathbf{z}) - \nabla\tilde{F}(\mathbf{x}_{t-1};\mathbf{z})$ as the unbiased estimator for the gradient variation. We do the minimization in subspace with $l < d$ to reduce computational cost.

3.3 Non Oblivious setting

In this setting, the probability density function is dependent on $\mathbf{x}$. So the simple $\tilde{\Delta}_t$ is not an unbiased estimator of the gradient variation. We leave exploration of this setting with the incorporation of subspace idea as a possible future work.

4 Proposed Algorithms

Main Algorithm 1 One-Sample Subspace Frank Wolfe A **for oblivious setting**

Input: Step sizes $\rho \in (0,1), \eta \in (0,1)$, initial point $\mathbf{x} \in \mathcal{X}$, total number of iterations T
Output: $\mathbf{x}_{T+1}$

```
for $t = 1, 2, 3, ..., T$ do
    Compute $\mathbf{P}_t$ using sub routine
    if $t = 1$ then
        Sample a point $\mathbf{z}_1 \sim p(\mathbf{z})$
        Compute $\mathbf{d}_1 = \nabla\tilde{F}(\mathbf{x}_1; \mathbf{z}_1)$
    else
        Sample a point $\mathbf{z}_t \sim p(\mathbf{z})$
        $\tilde{\Delta}_t = \nabla\tilde{F}(\mathbf{x}_t; \mathbf{z}_t) - \nabla\tilde{F}(\mathbf{x}_{t-1}; \mathbf{z}_t)$
        $\mathbf{d}_t = (1 - \rho_t)(\mathbf{d}_{t-1} + \tilde{\Delta}_t) + \rho_t \nabla\tilde{F}(\mathbf{x}_t; \mathbf{z}_t)$
    end if
    $\mathbf{u}_t = argmin_{\mathbf{x} \in \mathcal{X}_2} \langle \mathbf{P}_t^T \mathbf{d}_t, \mathbf{u} \rangle$
    $\mathbf{y}_t = \mathbf{P}_t \mathbf{u}_t$
    $\mathbf{x}_{t+1} = (1 - \eta_t)\mathbf{x}_t + \eta_t \mathbf{y}_t$
    return $\mathbf{x}_{T+1}$
```

Main Algorithm 2 One-Sample Subspace Frank Wolfe B **for oblivious setting**

Input: Step sizes $\rho \in (0,1), \eta \in (0,1)$, initial point $\mathbf{x} \in \mathcal{X}$, total number of iterations T
Output: $\mathbf{x}_{T+1}$

```
for $t = 1, 2, 3, ..., T$ do
    Compute $\mathbf{P}_t$ using sub routine
    if $t = 1$ then
        Sample a point $\mathbf{z}_1 \sim p(\mathbf{z})$
        Compute $\mathbf{d}_1 = \mathbf{P}_1^T \nabla\tilde{F}(\mathbf{x}_1; \mathbf{z}_1)$
    else
        Sample a point $\mathbf{z}_t \sim p(\mathbf{z})$
        $\tilde{\Delta}_t = \mathbf{P}_t^T \nabla\tilde{F}(\mathbf{x}_t; \mathbf{z}_t) - \mathbf{P}_t^T \nabla\tilde{F}(\mathbf{x}_{t-1}; \mathbf{z}_t)$
        $\mathbf{d}_t = (1 - \rho_t)(\mathbf{P}_t^T \mathbf{P}_{t-1} \mathbf{d}_{t-1} + \tilde{\Delta}_t) + \rho_t \mathbf{P}_t^T \nabla\tilde{F}(\mathbf{x}_t; \mathbf{z}_t)$
    end if
    $\mathbf{u}_t = argmin_{\mathbf{x} \in \mathcal{X}_2} \langle \mathbf{d}_t, \mathbf{u} \rangle$
    $\mathbf{y}_t = \mathbf{P}_t \mathbf{u}_t$
    $\mathbf{x}_{t+1} = (1 - \eta_t)\mathbf{x}_t + \eta_t \mathbf{y}_t$
    return $\mathbf{x}_{T+1}$
```

```
Sub Routine Generate a scaled, random Haar distributed matrix P
Input: $l, d$
Output: P matrix satisfying B1 and B2 of 3
  1. Initialize $\mathbf{X} \in \mathbb{R}^{d\times l}$
  2. Set $\mathrm{x}_{i,j} \sim \mathcal{N}(0,1)$
  3. Calculate thin QR decomposition of X i.e. $\mathbf{X} = \mathbf{QR}$ to find $\Lambda$
     $\Lambda = \begin{bmatrix} \frac{R_{1,1}}{|R_{1,1}|} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{R_{l,l}}{|R_{l,l}|} \end{bmatrix}$
  4. Return $\mathbf{P} = \sqrt{\frac{d}{l}}\mathbf{Q}\Lambda$
```

5 Convergence Analysis

5.1 Convergence for Algorithm A

We begin by showing that the gradient estimator $\mathbb{E}[\mathbf{d}_t]$ is an unbiased estimator of $\nabla F(\mathbf{x}_t)$. We show this by using simple induction.

$$\mathbb{E}[\mathbf{d}_1] = \mathbb{E}[\nabla\tilde{F}(\mathbf{x}_1, \mathbf{z}_1)] = \nabla F(\mathbf{x}_1)$$

Assume that $\mathbb{E}[\mathbf{d}_{t-1}] = \nabla F(\mathbf{x}_{t-1})$
The update equation for $\mathbf{d}_t$ is given by,

$$\mathbf{d}_t = (1-\rho_t)(\mathbf{d}_{t-1} + \tilde{\Delta}_t) + \rho_t\nabla\tilde{F}(\mathbf{x}_t; \mathbf{z}_t)$$

Taking expectation of above equation, we have

$$\mathbb{E}[\mathbf{d}_t] = (1-\rho_t)(\mathbb{E}[\mathbf{d}_{t-1}] + \mathbb{E}[\tilde{\Delta}_t]) + \rho_t\mathbb{E}[\nabla\tilde{F}(\mathbf{x}_t; \mathbf{z}_t)]$$

$$\mathbb{E}[\mathbf{d}_t] = (1-\rho_t)(\nabla F(\mathbf{x}_{t-1}) + \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})) + \rho_t\nabla F(\mathbf{x}_t)$$

Thus, we complete induction and show that

$$\mathbb{E}[\mathbf{d}_t] = \nabla F(\mathbf{x}_t)$$

Update equation is given by,

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta_t(\mathbf{y}_t - \mathbf{x}_t) \tag{9}$$

Now, since F is L smooth, using Quadratic Upper Bound, we have

$$\implies F(\mathbf{x}_{t+1}) \le F(\mathbf{x}_t) + \langle\nabla F(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2$$
$$\implies F(\mathbf{x}_{t+1}) \le F(\mathbf{x}_t) + \eta_t\langle\nabla F(\mathbf{x}_t), \mathbf{y}_t - \mathbf{x}_t\rangle + \frac{L}{2}\eta_t^2\|\mathbf{y}_t - \mathbf{x}_t\|^2$$

Using compactness of F, we have:

$$\implies F(\mathbf{x}_{t+1}) \le F(\mathbf{x}_t) + \eta_t\langle\nabla F(\mathbf{x}_t), \mathbf{y}_t - \mathbf{x}_t\rangle + \frac{L}{2}\eta_t^2 D^2$$
$$\implies F(\mathbf{x}_{t+1}) \le F(\mathbf{x}_t) + \eta_t\langle\nabla F(\mathbf{x}_t) - \mathbf{d}_t + \mathbf{d}_t, \mathbf{y}_t - \mathbf{x}_t\rangle + \frac{L}{2}\eta_t^2 D^2$$

$$\implies F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \eta_t\langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{y}_t - \mathbf{x}_t\rangle + \eta_t\langle \mathbf{d}_t, \mathbf{y}_t - \mathbf{x}_t\rangle + \frac{L}{2}\eta_t^2 D^2 \tag{10}$$

From the way $\mathbf{u}_t$ is defined, we have

$$\begin{aligned}
\mathbf{u}_t &= \arg\min_{\mathbf{u}\in\mathcal{X}_2}\langle \mathbf{P}_t^T\mathbf{d}_t, \mathbf{u}\rangle \\
\implies &\langle \mathbf{P}_t^T\mathbf{d}_t, \mathbf{u}_t\rangle \leq \langle \mathbf{P}_t^T\mathbf{d}_t, \mathbf{u}^*\rangle \\
\implies &\langle \mathbf{d}_t, \mathbf{P}_t\mathbf{u}_t\rangle \leq \langle \mathbf{P}_t^T\mathbf{d}_t, \mathbf{P}_t^T\mathbf{x}^*\rangle \\
\implies &\langle \mathbf{d}_t, \mathbf{P}_t\mathbf{u}_t\rangle \leq \langle \mathbf{d}_t, \mathbf{P}_t\mathbf{P}_t^T\mathbf{x}^*\rangle
\end{aligned}$$

Now using $\mathbf{y}_t = \mathbf{P}_t\mathbf{u}_t$, we have

$$\langle \mathbf{d}_t, \mathbf{y}_t\rangle \leq \langle \mathbf{d}_t, \mathbf{P}_t\mathbf{P}_t^T\mathbf{x}^*\rangle \tag{11}$$

Now multiplying by η_t and adding to (10), we have

$$\implies F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \eta_t\langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{y}_t - \mathbf{x}_t\rangle + \eta_t\langle \mathbf{d}_t, \mathbf{P}_t\mathbf{P}_t^T\mathbf{x}^* - \mathbf{x}_t\rangle + \frac{L}{2}\eta_t^2 D^2$$

$$\implies F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \eta_t\langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{y}_t - \mathbf{P}_t\mathbf{P}_t^T\mathbf{x}^*\rangle + \eta_t\langle \nabla F(\mathbf{x}_t), \mathbf{P}_t\mathbf{P}_t^T\mathbf{x}^* - \mathbf{x}_t\rangle + \frac{L}{2}\eta_t^2 D^2$$

Applying Cauchy-Schwartz Inequality on the second term and using compactness again, we have

$$\implies F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) \leq F(\mathbf{x}_t) - F(\mathbf{x}^*) + \eta_t D\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\| + \eta_t\langle \nabla F(\mathbf{x}_t), \mathbf{P}_t\mathbf{P}_t^T\mathbf{x}^* - \mathbf{x}_t\rangle + \frac{L}{2}\eta_t^2 D^2$$

Take expectation and use $\mathbb{E}[\mathbf{P}_t\mathbf{P}_t^T] = \mathbf{I}_d$

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)] \leq \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] + \eta_t D\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|] + \eta_t\langle \nabla F(\mathbf{x}_t), \mathbf{x}^* - \mathbf{x}_t\rangle + \frac{L}{2}\eta_t^2 D^2$$

Now apply convexity on the inner product to get back exactly same one step inequality of the One Sample Frank Wolfe Method

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)] \leq (1-\eta_t)\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] + \eta_t D\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|] + \frac{L}{2}\eta_t^2 D^2 \tag{12}$$

Since $\mathbf{d}_t$ is an unbiased estimator of $\nabla F(\mathbf{x}_t)$ and the update equation of $\mathbf{d}_t$ is **exactly the same** as One Sample Frank Wolfe, we can directly take the result on bound of gradient error from **lemma 2** of **Mokhtari et al.[2019](1)** which proves that for ρ_t to be $(t-1)^{-\alpha}$ and η_t to be $(t^{-\alpha})$

$$\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2] \leq Ct^{-\alpha} \tag{13}$$

where C is a constant dependent on the parameters $L, \tilde{L}, G, D, \alpha$ in the paper Mokhtari et al.[2019](1)
Using Jensen's Inequality we can write that,

$$\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|] \leq \sqrt{\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2]} \leq \sqrt{C}t^{-\frac{\alpha}{2}} \tag{14}$$

Substituting $\alpha = 1$ and using the above result, we have

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)] \leq \frac{t-1}{t}\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] + \frac{1}{t\sqrt{t}}D\sqrt{C} + \frac{L}{2}\frac{1}{t^2}D^2$$

$$t\,\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)] \leq (t-1)\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] + \frac{1}{\sqrt{t}}D\sqrt{C} + \frac{L}{2}\frac{1}{t}D^2$$

Applying telescopic sum now, we have

$$T\,\mathbb{E}[F(\mathbf{x}_{T+1}) - F(\mathbf{x}^*)] \leq D\sqrt{C}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} + \sum_{t=1}^{T}\frac{L}{2}\frac{1}{t}D^2$$

$$\mathbb{E}[F(\mathbf{x}_{T+1}) - F(\mathbf{x}^*)] \leq \frac{D\sqrt{C}}{T}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} + \frac{LD^2}{2T}\sum_{t=1}^{T}\frac{1}{t}$$

$$\mathbb{E}[F(\mathbf{x}_{T+1}) - F(\mathbf{x}^*)] \leq \frac{D\sqrt{C}}{T}2\sqrt{T} + \frac{LD^2}{2}(1+\ln(T))$$

$$\mathbb{E}[F(\mathbf{x}_{T+1}) - F(\mathbf{x}^*)] \leq \frac{2D\sqrt{C}}{\sqrt{T}} + \frac{LD^2}{2T}(1+\ln(T)) \tag{15}$$

So, we recover the optimal order complexity bound of $\mathcal{O}\big(\frac{1}{\sqrt{T}}\big)$.

5.2 Convergence for Algorithm B

First we observe that due to independence of Haar distributed matrices with $\mathbf{z}$ and the assumed oblivious setting, we have

$$\mathbb{E}[\mathbf{P}_t\tilde{\Delta}_t] = \mathbb{E}[\mathbf{P}_t\mathbf{P}_t^T(\nabla\tilde{F}(\mathbf{x}_t,\mathbf{z}_t) - \nabla\tilde{F}\mathbf{x}_{t-1},\mathbf{z}_t)] = (\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})$$

Then we proceed to showing that the gradient estimator $\mathbb{E}[\mathbf{P}_t\mathbf{d}_t]$ is indeed an unbiased estimator of $\nabla F(\mathbf{x}_t)$.
We again show this by using simple induction.

$$\mathbb{E}[\mathbf{P}_1\mathbf{d}_1] = \mathbb{E}[\mathbf{P}_1\mathbf{P}_1^T\nabla\tilde{F}(\mathbf{x}_1,\mathbf{z}_1)] = \mathbb{E}[\mathbb{E}[\mathbf{P}_1\mathbf{P}_1^T]\mathbb{E}[\nabla\tilde{F}(\mathbf{x}_1,\mathbf{z}_1)]] = \nabla F(\mathbf{x}_1)$$

Assume that $\mathbb{E}[\mathbf{P}_{t-1}\mathbf{d}_{t-1}] = \nabla F(\mathbf{x}_{t-1})$
The update equation for $\mathbf{d}_t$ is given by,

$$\mathbf{d}_t = (1-\rho_t)(\mathbf{P}_t^T\mathbf{P}_{t-1}\mathbf{d}_{t-1} + \tilde{\Delta}_t) + \rho_t\mathbf{P}_t^T\nabla\tilde{F}(\mathbf{x}_t;\mathbf{z}_t)$$

$$\mathbf{P}_t\mathbf{d}_t = (1-\rho_t)(\mathbf{P}_t\mathbf{P}_t^T\mathbf{P}_{t-1}\mathbf{d}_{t-1} + \mathbf{P}_t\tilde{\Delta}_t) + \rho_t\mathbf{P}_t\mathbf{P}_t^T\nabla\tilde{F}(\mathbf{x}_t;\mathbf{z}_t)$$

Taking expectation of above equation,

$$\mathbb{E}[\mathbf{P}_t\mathbf{d}_t] = (1-\rho_t)(\mathbb{E}[\mathbf{P}_t\mathbf{P}_t^T\mathbf{P}_{t-1}\mathbf{d}_{t-1}] + \mathbb{E}[\mathbf{P}_t\tilde{\Delta}_t]) + \rho_t\mathbb{E}[\mathbf{P}_t\mathbf{P}_t^T\nabla\tilde{F}(\mathbf{x}_t;\mathbf{z}_t)]$$

$$\mathbb{E}[\mathbf{P}_t\mathbf{d}_t] = (1-\rho_t)(\nabla F(\mathbf{x}_{t-1}) + \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})) + \rho_t\nabla F(\mathbf{x}_t)$$

Thus, we complete induction and show that

$$\mathbb{E}[\mathbf{P}_t\mathbf{d}_t] = \nabla F(\mathbf{x}_t)$$

Now we write the update equation for $\mathbf{x}$

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta_t(\mathbf{y}_t - \mathbf{x}_t) \tag{16}$$

Now, since F is L smooth, using Quadratic Upper Bound, we have

$$\implies F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \langle\nabla F(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2$$

$$\implies F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \eta_t\langle\nabla F(\mathbf{x}_t), \mathbf{y}_t - \mathbf{x}_t\rangle + \frac{L}{2}\eta_t^2\|\mathbf{y}_t - \mathbf{x}_t\|^2$$

Using compactness of F, we have:

$$\implies F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \eta_t \langle \nabla F(\mathbf{x}_t), \mathbf{y}_t - \mathbf{x}_t \rangle + \frac{L}{2}\eta_t^2 D^2$$
$$\implies F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \eta_t \langle \nabla F(\mathbf{x}_t) - \mathbf{P}_t\mathbf{d}_t + \mathbf{P}_t\mathbf{d}_t, \mathbf{y}_t - \mathbf{x}_t \rangle + \frac{L}{2}\eta_t^2 D^2$$

$$\implies F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \eta_t \langle \nabla F(\mathbf{x}_t) - \mathbf{P}_t\mathbf{d}_t, \mathbf{y}_t - \mathbf{x}_t \rangle + \eta_t \langle \mathbf{P}_t\mathbf{d}_t, \mathbf{y}_t - \mathbf{x}_t \rangle + \frac{L}{2}\eta_t^2 D^2 \tag{17}$$

From the way $\mathbf{u}_t$ is defined, we have

$$\mathbf{u}_t = \arg\min_{\mathbf{u} \in \mathcal{X}_2} \langle \mathbf{d}_t, \mathbf{u} \rangle$$
$$\implies \langle \mathbf{d}_t, \mathbf{u}_t \rangle \leq \langle \mathbf{d}_t, \mathbf{P}_t^T x^* \rangle$$
$$\implies \langle \mathbf{d}_t, \mathbf{u}_t \rangle \leq \langle \mathbf{d}_t, \mathbf{P}_t^T \mathbf{x}^* \rangle$$
$$\implies \langle \mathbf{d}_t, \mathbf{u}_t \rangle \leq \langle \mathbf{P}_t\mathbf{d}_t, \mathbf{x}^* \rangle$$

Now using $\mathbf{y}_t = \mathbf{P}_t\mathbf{u}_t$ and $\mathbf{P}_t^T\mathbf{P}_t = \frac{d}{l}\mathbb{I}_l$, we have

$$\langle \mathbf{P}_t\mathbf{d}_t, \mathbf{y}_t \rangle = \langle \mathbf{P}_t\mathbf{d}_t, \mathbf{P}_t\mathbf{u}_t \rangle = \langle \mathbf{P}_t^T\mathbf{P}_t\mathbf{d}_t, \mathbf{u}_t \rangle = \frac{d}{l}\langle \mathbf{d}_t, \mathbf{u}_t \rangle \tag{18}$$

Now multiplying by η_t and adding to (17), we have

$$\implies F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \eta_t \langle \nabla F(\mathbf{x}_t) - \mathbf{P}_t\mathbf{d}_t, \mathbf{y}_t - \mathbf{x}_t \rangle + \frac{d}{l}\eta_t \langle \mathbf{d}_t, \mathbf{u}_t \rangle - \eta_t \langle \mathbf{P}_t\mathbf{d}_t, \mathbf{x}_t \rangle + \frac{L}{2}\eta_t^2 D^2$$
$$\implies F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \eta_t \langle \nabla F(\mathbf{x}_t) - \mathbf{P}_t\mathbf{d}_t, \mathbf{y}_t - \mathbf{x}_t \rangle + \frac{d}{l}\eta_t \langle \mathbf{P}_t\mathbf{d}_t, \mathbf{x}^* \rangle - \eta_t \langle \mathbf{P}_t\mathbf{d}_t, \mathbf{x}_t \rangle + \frac{L}{2}\eta_t^2 D^2$$

Applying Cauchy-Schwartz Inequality on the second term and using compactness again, we have

$$\implies F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) \leq F(\mathbf{x}_t) - F(\mathbf{x}^*) + \eta_t D \|\nabla F(\mathbf{x}_t) - \mathbf{P}_t\mathbf{d}_t\| + \eta_t \langle \mathbf{P}_t\mathbf{d}_t, \mathbf{x}^* - \mathbf{x}_t \rangle + (\frac{d}{l} - 1)\eta_t \langle \mathbf{P}_t\mathbf{d}_t, \mathbf{x}^* \rangle + \frac{L}{2}\eta_t^2 D^2$$

Take expectation and use $\mathbb{E}[\mathbf{P}_t\mathbf{d}_t] = \nabla F(\mathbf{x}_t)$

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)] \leq \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] + \eta_t D \mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{P}_t\mathbf{d}_t\|]$$
$$+ \eta_t \langle \nabla F(\mathbf{x}_t), \mathbf{x}^* - \mathbf{x}_t \rangle + (\frac{d}{l} - 1)\eta_t \langle \mathbf{P}_t\mathbf{d}_t, \mathbf{x}^* \rangle + \frac{L}{2}\eta_t^2 D^2$$

Now apply convexity on the inner product to get a one step inequality

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)] \leq (1 - \eta_t)\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] + \eta_t D \mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{P}_t\mathbf{d}_t\|] + (\frac{d}{l} - 1)\eta_t \langle \nabla F(\mathbf{x}_t), \mathbf{x}^* \rangle + \frac{L}{2}\eta_t^2 D^2$$

Using our lemma 1, for ρ_t to be $(t-1)^{-\alpha}$ and η_t to be $t^{-\alpha}$, we have

$$\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{P}_t\mathbf{d}_t\|^2] \leq Ct^{-\alpha} \tag{19}$$

where C is a constant dependent on the paramters $L, \tilde{L}, G, D, \alpha$
Using Jensen's Inequality we can write that,

$$\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{P}_t\mathbf{d}_t\|] \leq \sqrt{\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{P}_t\mathbf{d}_t\|^2]} \leq \sqrt{C}t^{-\frac{\alpha}{2}} \tag{20}$$

Substituting $\alpha = 1$ and using the above result, we have

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)] \leq \frac{t-1}{t}\mathbb{E}[F(\mathbf{x}_t) - F(x^*)] + \frac{1}{t\sqrt{t}}D\sqrt{C} + \frac{L}{2}\frac{1}{t^2}D^2 + \frac{1}{t}(\frac{d}{l} - 1)\langle \nabla F(\mathbf{x}_t), \mathbf{x}^* \rangle$$

The last term brings out the impact of dimensional reduction into the picture. This error term initially creates perturbations. If factor $(\frac{d}{l} - 1)$ is large ,i.e, $l << d$ then initial perturbations can significantly impact initially and cause the algorithm to run suboptimally. However, if the factor $(\frac{d}{l} - 1)$ is small then the perturbation is also small and we can do some more analysis. If we look a little more carefully into the last term since in expectation the gradient error is decaying at an inverse square root rate, we are moving in expectation in the descent direction, and thus the norm of $\nabla F(\mathbf{x}_t)$ tends to decrease. The $\frac{1}{t}$ on multiplication further reduces the term, therefore, we can consider a case where the dimensional reduction is not very large and there exists a time t' early enough after which the term becomes very small in comparison to other terms so that we can ignore it. Then we can recover optimal complexity bound of $\mathcal{O}\big(\frac{1}{\sqrt{T}}\big)$ by performing telescopic sum after the t' step.
The analysis after this assumption follows similar.

$$t\,\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)] \leq (t-1)\mathbb{E}[F(\mathbf{x}_t) - F(x^*)] + \frac{1}{\sqrt{t}}D\sqrt{C} + \frac{L}{2}\frac{1}{t}D^2$$

$$T\,\mathbb{E}[F(\mathbf{x}_{T+1}) - F(\mathbf{x}^*)] \leq (t'-1)\mathbb{E}[F(\mathbf{x}_{t'}) - F(\mathbf{x}^*)] + D\sqrt{C}\sum_{t=t'}^{T}\frac{1}{\sqrt{t}} + \sum_{t=t'}^{T}\frac{L}{2}\frac{1}{t}D^2$$

$$\mathbb{E}[F(\mathbf{x}_{T+1}) - F(\mathbf{x}^*)] \leq \frac{1}{T}(t'-1)\mathbb{E}[F(\mathbf{x}_{t'}) - F(\mathbf{x}^*)] + \frac{D\sqrt{C}}{T}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} + \frac{LD^2}{2T}\sum_{t=1}^{T}\frac{1}{t}$$

$$\mathbb{E}[F(\mathbf{x}_{T+1}) - F(\mathbf{x}^*)] \leq \frac{R}{T} + \frac{D\sqrt{C}}{T}2\sqrt{T} + \frac{LD^2}{2}(1 + \ln(T))$$

$$\mathbb{E}[F(\mathbf{x}_{T+1}) - F(\mathbf{x}^*)] \leq \frac{R}{T} + \frac{2D\sqrt{C}}{\sqrt{T}} + \frac{LD^2}{2T}(1 + \ln(T)) \tag{21}$$

The optimal convergence bound $\mathcal{O}\big(\frac{1}{\sqrt{T}}\big)$ is attainable with the assumption. We believe this assumption can be shown feasible for small $\frac{d}{l}$ i.e if the dimensionality. reduction is not kept too large. The experiments ahead are in favour of the above made analysis.

6 Experimental Observations

6.1 Experiments with Algorithm A

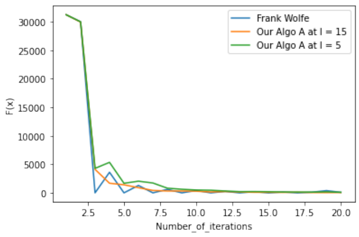
We compared convergence rates of Our Algorithm A with the one sample stochastic Frank Wolfe for various constraint sets with different values of l and d.First we tried to optimize the following objective:

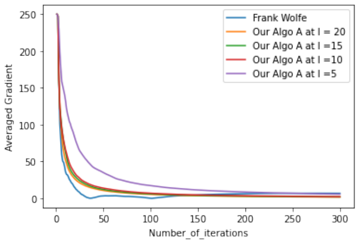
$$\tilde{F}(\mathbf{x}, \mathbf{z}) = z * \|x\|^2$$

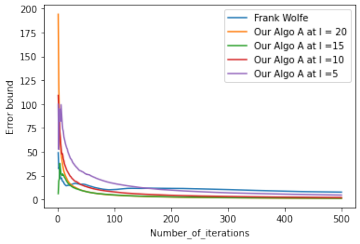
$$\min F(\mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(0,1)}[\tilde{F}(\mathbf{x}, \mathbf{z})]$$

$$s.t. \mathcal{X} = Box[-50, 50]$$

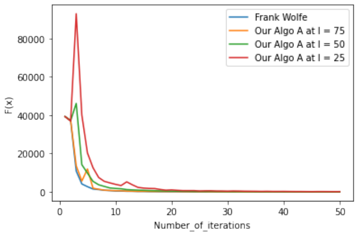
We kept d=25 and tried with different values of *i.e.* l=20,15,10 & 5. The averaged gradient and the error bound in gradient converges to zero at same rate as Frank Wolfe and the optimal value is obtained in same iteration complexity as Frank Wolfe as can be seen from below:

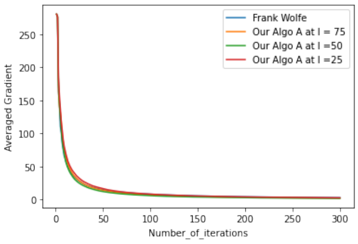


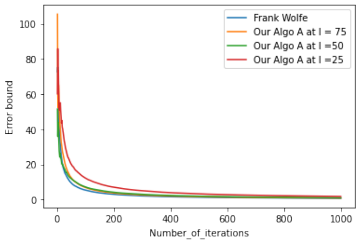




Moving further, we increased the dimension d to 100 keeping the function and constraints same. We checked for various l and here we are showing at l = 75,50,25.







We can clearly see that our algorithm is working as expected from the convergence analysis.

6.2 Experiments with Algorithm B

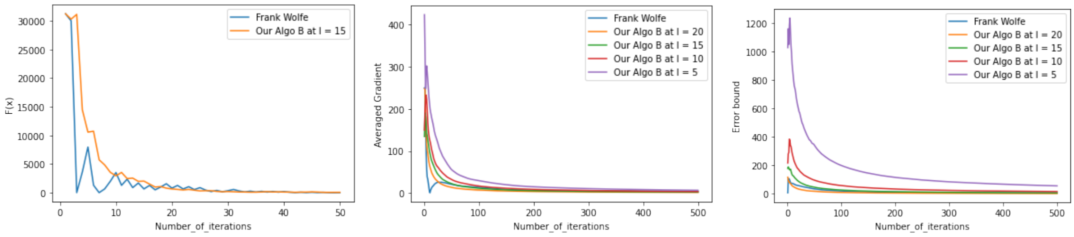
We compared convergence rates of Our Algorithm B with the one sample stochastic Frank Wolfe for various constraint sets with different values of l and d. First we tried to optimize the following objective:

$$\tilde{F}(\mathbf{x}, \mathbf{z}) = z * \|x - 100\|^2$$

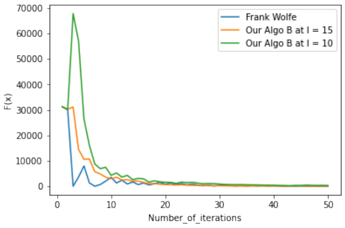
$$\min F(\mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(0,1)}[\tilde{F}(\mathbf{x}, \mathbf{z})]$$

$$s.t. \mathcal{X} = Box[50, 150]$$

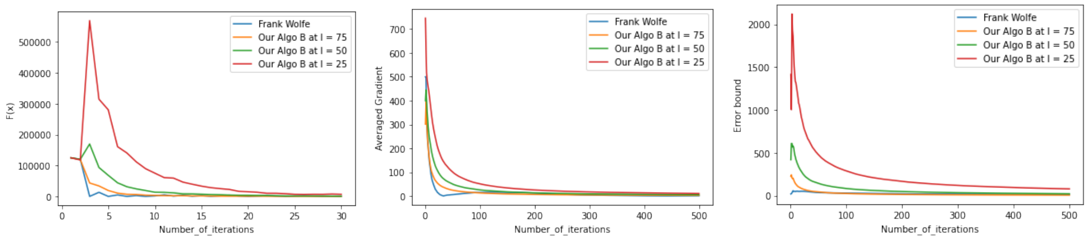
We kept d=25 and l=20,15,10,5 for our Algorithm B. We found optimistic results. It is shown below:



As expected with the Algorithm B, there is initial perturbance when dimensional reduction increases. Although later the algorithm converges back quickly at the expected optimal rate.



Next, we increased the dimension d to 100 and then checked for various l. Here we are showing at l = 75,50,25.



We derived same insights again. Below l = 50, the algorithm started shooting off a lot initially evident from l = 25 graph. However in all these cases , there existed a time t' after which the rate of decrease of gradient brought back the convergence. **Thus, by not keeping dimensional reduction too large, one can still reap benefits of faster total time complexity from our algorithm B.**

7 Challenges faced

We started with the algorithm which worked as follows:

1. $\tilde{\Delta}_t = \mathbf{P}_t^T \nabla \tilde{F}(\mathbf{x}_t; \mathbf{z}_t) - \mathbf{P}_t^T \nabla \tilde{F}(\mathbf{x}_{t-1}; \mathbf{z}_t)$
2. $\mathbf{d}_t = (1-\rho_t)(\mathbf{d}_{t-1} + \tilde{\Delta}_t) + \rho_t \mathbf{P}_t^T \nabla \tilde{F}(\mathbf{x}_t; \mathbf{z}_t)$
3. $\mathbf{u}_t = argmin_{\mathbf{x} \in \mathcal{X}_2} \langle \mathbf{d}_t, \mathbf{u} \rangle$
4. $\mathbf{y}_t = \mathbf{P}_t \mathbf{u}_t$
5. $\mathbf{x}_{t+1} = (1-\eta_t)\mathbf{x}_t + \eta_t \mathbf{y}_t$

However, in this the gradient was not getting properly estimated.

So we simply added the subspace optimization to atleast improve the bottleneck.

1. $\tilde{\Delta}_t = \nabla \tilde{F}(\mathbf{x}_t; \mathbf{z}_t) - \nabla \tilde{F}(\mathbf{x}_{t-1}; \mathbf{z}_t)$
2. $\mathbf{d}_t = (1-\rho_t)(\mathbf{d}_{t-1} + \tilde{\Delta}_t) + \rho_t \mathbf{P}_t^T \nabla \tilde{F}(\mathbf{x}_t; \mathbf{z}_t)$
3. $\mathbf{u}_t = argmin_{\mathbf{x} \in \mathcal{X}_2} \langle \mathbf{P}_t^T \mathbf{d}_t, \mathbf{u} \rangle$
4. $\mathbf{y}_t = \mathbf{P}_t \mathbf{u}_t$
5. $\mathbf{x}_{t+1} = (1-\eta_t)\mathbf{x}_t + \eta_t \mathbf{y}_t$

This algorithm **maintains the optimal convergence rate**, the oracle is still $\nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)$.

Then pondering further, we successfully figured out a way to recover the unbiased estimator.

1. $\tilde{\Delta}_t = \mathbf{P}_t^T \nabla \tilde{F}(\mathbf{x}_t; \mathbf{z}_t) - \mathbf{P}_t^T \nabla \tilde{F}(\mathbf{x}_{t-1}; \mathbf{z}_t)$
2. $\mathbf{d}_t = (1-\rho_t)(\mathbf{P}_t^T \mathbf{P}_{t-1} \mathbf{d}_{t-1} + \tilde{\Delta}_t) + \rho_t \mathbf{P}_t^T \nabla \tilde{F}(\mathbf{x}_t; \mathbf{z}_t)$
3. $\mathbf{u}_t = argmin_{\mathbf{x} \in \mathcal{X}_2} \langle \mathbf{d}_t, \mathbf{u} \rangle$
4. $\mathbf{y}_t = \mathbf{P}_t \mathbf{u}_t$
5. $\mathbf{x}_{t+1} = (1-\eta_t)\mathbf{x}_t + \eta_t \mathbf{y}_t$
6. return $\mathbf{x}_{T+1}$

and carried out detailed analysis and reported some interesting findings.

The subspace mapping reduces the information about gradient and in a setting where we are already using a single sample, the dimensional reduction factor plays an important role in adding variance. Also, we believe that there can be found a limit on the dimensional reduction up to which the **Algorithm B** can continue to achieve optimal bound.

8 Conclusion

1. Our proposed **Algorithm A** achieves the same iteration complexity as the **Frank Wolfe One Sample Stochastic** (1) Method which is $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$. The per iterate time complexity is reduced as the minimization is carried in **l** dimensional space than in **d** dimensional space.

2. The **algorithm A** still needs one full **d** dimensional sample of gradient $\nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)$

3. Our proposed **Algorithm B** achieves the same iteration complexity if the **l** is not made absurdly small. There is initial bouncing away which diminishes as analyzed in the convergence analysis.

4. The experiments and simulations confirmed the investigations into the algorithms. There is a possibility of getting an exact relation between the dimensional reduction factor $(\frac{\mathbf{d}}{\mathbf{l}})$ and its impact on the order time complexity in the **Algorithm B**. We have not yet been able to find it. We are ending the report herewith that as our future work.

9 Appendix

9.1 Proof of Lemma Used in Convergence Analysis of A

Let $A_t = \|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|$ By definition, we have

$$A_t = \|\nabla F(\mathbf{x}_{t-1}) - \mathbf{x}_{t-1} + \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1}) - (\mathbf{x}_t - \mathbf{x}_{t-1})\|^2$$

Note that

$$\mathbf{d}_t - \mathbf{d}_{t-1} = -\rho_t \mathbf{d}_{t-1} + \rho_t \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t) + (1 - \rho_t)\tilde{\Delta}_t$$

and define $\Delta_t = \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})$, we have

$$A_t = \|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1} + \Delta_t - (1 - \rho_t)\tilde{\Delta}_t - \rho_t \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t) + \rho_t \mathbf{d}_{t-1}\|^2$$

$$= \|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1} + (1 - \rho_t)(\Delta_t - \tilde{\Delta}_t) + \rho_t(\nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t) + \rho_t(\mathbf{d}_{t-1} - \nabla F(\mathbf{x}_{t-1}))\|^2$$

$$= \|(1 - \rho_t)(\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}) + (1 - \rho_t)(\Delta_t - \tilde{\Delta}_t) + \rho_t(\nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\|^2$$

Since $\tilde{\Delta}_t$ is an unbiased estimator of Δ_t, $\mathbb{E}[A_t]$ can be decomposed as

$$\mathbb{E}[A_t] = (1-\rho_t)^2 \mathbb{E}[\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2] + (1-\rho_t)^2 \mathbb{E}[\|\Delta_t - \tilde{\Delta}_t\|^2] + \rho_t^2 \mathbb{E}[(\|\nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\|^2)]$$
$$+ 2\rho_t(1 - \rho_t)\mathbb{E}[\langle \Delta_t - \tilde{\Delta}_t, \nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\rangle]$$

Now,

$$\mathbb{E}[\|\tilde{\Delta}_t - \Delta_t\|^2] = \mathbb{E}[\|\nabla_t^2(\mathbf{x}_t - \mathbf{x}_{t-1}) - (\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})\|^2] \leq \mathbb{E}[\|\nabla_t^2(\mathbf{x}_t - \mathbf{x}_{t-1})\|^2] + L^2\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2$$

$$\mathbb{E}[\|\tilde{\Delta}_t - \Delta_t\|^2] \leq \eta_{t-1}^2 D^2 \tilde{L}^2 + \eta_{t-1}^2 D^2 L^2$$

By Jensen's inequality, we have

$$\mathbb{E}[\|\tilde{\Delta}_t - \Delta_t\|^2] \leq \sqrt{\mathbb{E}[\|\tilde{\Delta}_t - \Delta_t\|]} \leq \eta_{t-1} D L'$$

where $L' = \sqrt{\tilde{L}^2 + L^2}$

Finally, by Assumption 3, we have $\|\nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\| \leq 2G$. Thus,

$$\rho_t^2 \|\nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\| \leq 4\rho_t^2 G^2$$

and,

$$\mathbb{E}[2\rho_t(1 - \rho_t)\langle \Delta_t - \tilde{\Delta}_t, \nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\rangle] \leq \mathbb{E}[2\rho_t(1 - \rho_t)\|\Delta_t - \tilde{\Delta}_t\|.\|\nabla F(\mathbf{x}_t) - \nabla \tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\|]$$

$$\leq 4\eta_{t-1}\rho_t(1 - \rho_t) G D \tilde{L}$$

Using all of above, we have

$$\mathbb{E}[A_t] \leq (1 - \rho_t)^2 \mathbb{E}[A_{t-1}] + (1 - \rho_t)^2 \eta_{t-1}^2 D^2 L^2 + 4\rho^2 G^2 + 4\eta_{t-1}\rho_t(1 - \rho_{t-1}) G D \tilde{L}$$

The recursive relation in the One Sample Frank Wolfe is given for general non oblivious settings and it is as follows:

$$\mathbb{E}[A_t] \leq (1 - \rho_t)^2 \mathbb{E}[A_{t-1}] + (1 - \rho_t)^2 \eta_{t-1}^2 D^2 L^2 + 4\rho^2 G^2$$
$$+ 4\eta_{t-1}\rho_t(1 - \rho_{t-1}) G D \tilde{L} + 2\eta_{t-1}\rho_{t-1}(1 - \rho_{t-1})\sqrt{\mathbb{E}[A_{t-1}]} D L'$$

Since our inequality is less than the one that comes in the Mokhtari et al.[2019] (1), their further analysis can be directly carried over. Basically induction after this step which holds exactly the same because the parameters ρ_t and η_t are kept same as their paper. Thus, we get the relation with $\alpha = 1$.

$$\mathbb{E}[A_t] \leq C t^{-\alpha}$$

9.2 Proof of Lemma Used in Convergence Analysis of B

Let $\mathbf{h}_t = \mathbf{P}_t\mathbf{d}_t$

$$A_t = \|\nabla F(\mathbf{x}_t) - \mathbf{P}_t\mathbf{d}_t\|^2 = \|\nabla F(x_t) - \mathbf{h}_t\|^2$$

Let $\Delta_t = \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})$

$$A_t = \|\nabla F(\mathbf{x}_{t-1}) - \mathbf{h}_{t-1} + \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1}) - (\mathbf{h}_t - \mathbf{h}_{t-1})\|^2$$

Only for this proof, We make a slight change in notation to keep things similar as of **lemma 2** of Mokhtari et al. [2019] (1), we define

$$\tilde{\Delta}_t = \nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t) - \nabla\tilde{F}(\mathbf{x}_{t-1}, \mathbf{z}_t)$$

Multiplying $\mathbf{P}_t$ in update equation for $\mathbf{d}_t$ to get,

$$\mathbf{P}_t\mathbf{d}_t = (1-\rho_t)(\mathbf{P}_t\mathbf{P}_t^T\mathbf{P}_{t-1}\mathbf{d}_{t-1} + \mathbf{P}_t\mathbf{P}_t^T\tilde{\Delta}_t) + \rho_t\mathbf{P}_t\mathbf{P}_t^T\nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)$$

Now rewriting in form of $\mathbf{h}_t$,

$$\mathbf{h}_t = (1-\rho_t)(\mathbf{P}_t\mathbf{P}_t^T\mathbf{h}_{t-1} + \mathbf{P}_t\mathbf{P}_t^T\tilde{\Delta}_t) + \rho_t\mathbf{P}_t\mathbf{P}_t^T\nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t) \tag{22}$$

$$A_t = \|\nabla F(\mathbf{x}_{t-1}) - \mathbf{h}_{t-1} + \Delta_t - (\mathbf{h}_t - \mathbf{h}_{t-1})\|^2$$

Replace $\mathbf{h}_t$ and regroup terms to get,

$$\begin{aligned} A_t = \|&(1-\rho_t)(\nabla F(\mathbf{x}_{t-1}) - \mathbf{h}_{t-1}) + (1-\rho_t)(\Delta_t - \tilde{\Delta}_t) + \rho_t(\nabla F(\mathbf{x}_t) - \nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)) \\ &+ (1-\rho_t)(\mathbb{I} - \mathbf{P}_t\mathbf{P}_t^T)\tilde{\Delta}_t + \rho_t(\mathbb{I} - \mathbf{P}_t\mathbf{P}_t^T)\nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t) + (1-\rho_t)(\mathbb{I} - \mathbf{P}_t\mathbf{P}_t^T)\mathbf{h}_{t-1}\|^2 \end{aligned}$$

We see that the **first three terms** are **same** as that during proof of **lemma 2** in (1). After expanding, we observe

The cross terms of inner product between first three terms and second three terms become $\mathbf{0}$ due to $\mathbb{E}[\mathbb{I} - \mathbf{P}_t^T\mathbf{P}_t] = 0$ Out of the inner products among first three terms, two become $\mathbf{0}$ and only the following remains:

$$\langle \Delta_t - \tilde{\Delta}_t, \nabla F(\mathbf{x}_t) - \nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\rangle$$

as $\mathbb{E}(\Delta_t - \tilde{\Delta}_t) = 0$, since Δ_t is the unbiased estimator of $\tilde{\Delta}_t$. Moreover, $\mathbb{E}(\nabla F(\mathbf{x}_t) - \nabla\tilde{F}(\mathbf{x}_t))$ $= 0$ since $\nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)$ is the unbiased estimator of $\nabla F(\mathbf{x}_t)$.

The inner product of next three terms can all be bounded in similar fashion as the one shown below.

$$\begin{aligned} \langle(\mathbb{I} - \mathbf{P}_t^T\mathbf{P}_t)\tilde{\Delta}_t, (\mathbb{I} - \mathbf{P}_t^T\mathbf{P}_t)\nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\rangle &\leq \|(\mathbb{I} - \mathbf{P}_t^T\mathbf{P}_t)\tilde{\Delta}_t\|\|(\mathbb{I} - \mathbf{P}_t^T\mathbf{P}_t)\nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\| \\ &\leq \|(\mathbb{I} - \mathbf{P}_t^T\mathbf{P}_t)\|\|\tilde{\Delta}_t\|\|(\mathbb{I} - \mathbf{P}_t^T\mathbf{P}_t)\|\|\nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\| \end{aligned}$$

$$\mathbb{E}[\langle(\mathbb{I} - \mathbf{P}_t^T\mathbf{P}_t)\tilde{\Delta}_t, (\mathbb{I} - \mathbf{P}_t^T\mathbf{P}_t)\nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\rangle] \leq \mathbb{E}[\|(\mathbb{I} - \mathbf{P}_t^T\mathbf{P}_t)\|\|\tilde{\Delta}_t\|\|(\mathbb{I} - \mathbf{P}_t^T\mathbf{P}_t)\|\|\nabla\tilde{F}(\mathbf{x}_t, \mathbf{z}_t)\|]$$

$$\leq \sigma^2 \eta_{t-1} D L' G$$

Also, the similar relations get carried forward,

$$\mathbb{E}[\|\tilde{\Delta}_t - \Delta_t\|^2] = \mathbb{E}[\|\nabla_t^2(\mathbf{x}_t - \mathbf{x}_{t-1}) - (\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})\|^2] \leq \mathbb{E}[\|\nabla_t^2(\mathbf{x}_t - \mathbf{x}_{t-1})\|^2] + L^2\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2$$

$$\mathbb{E}[\|\tilde{\Delta}_t - \Delta_t\|^2] \leq \eta_{t-1}^2 D^2 \tilde{L}^2 + \eta_{t-1}^2 D^2 L^2$$

Thus, after putting all the above equations in the recursive relation of A_t, when we rearrange we get back the same recursive relation with difference only in constants. Thus, the analysis gets repeated and the same result of previous lemma holds true as we again keep the parameters ρ_{t-1} and η_{t-1} same, the equation achieved is same and this relation holds with $alpha = 1$, i.e.,

$$\mathbb{E}[A_t] \leq C t^{-\alpha}$$

References

[1] Mingrui Zhang, Zebang Shen, Aryan Mokhtari, Hamed Hassani, Amin Karbasi: One Sample Stochastic Frank-Wolfe

[2] David Kozak, Stephen Becker, Alizera Doostan, Luis Tenorio. Stochatic Subspace Descent