On the linear convergence analysis of the Determinstic Federated Learning algorithm

February 1, 2023

Abstract

In this paper, we present a mathematical proof for the linear convergence of the Deterministic Federated Learning algorithm under appropriately chosen parameters. Our analytical tool is to recast the algorithm as the well-known inexact Uzawa framework, apply the regularized convex functional to overcome the non-linearity. We provide the linear convergence for federated learning algorithm, based on both exact local solve, i.e., Gauss-Seidel method, and inexact local solve, i.e., $n-$step Gradient descent methods for sufficiently large n. We also provide the lower bound of n for the convergence in terms of the condition number of the objective functional. Some discussion on the optimal number of steps in GD is made as well, similar to what is discussed for the stochastic algorithm [12].

1 Introduction

Data becomes increasingly decentralized and the privacy of individual data is an utmost importance in the digital age [5, 1, 2, 11, 15]. Unlike standard machine learning approaches, *Federated learning* (FL) encourages each client to have a local training data set, which will not be stored to the server and to update the local correction of the current global model maintained by the main server via the local data and local gradient descent method. Federated learning has been used successfully in many different areas, which include Internet of Things (IoT) applications [6, 3]. Federated learning can be formulated as the optimization problem with a certain consensus constraint of distributed local optimization problem as discussed in [12]. More precisely, let N be the number of local clusters or clients, then FL can be given as follows:

$$\min_{\substack{X\in\mathbb{R}^{N_x}\\ X^1=X^2=\cdots=X^N}} \left\{ F(X) := \frac{1}{N}\sum_{i=1}^{N} f_i(X^i) \right\}, \qquad X = (X^1,\cdots,X^N)^T, \tag{1}$$

where $N_x = N \times d$ and $f_i : \mathbb{R}^d \to \mathbb{R}$ is the objective functional obtained based on the data owned on i^{th} client for $i \in [N] := \{1, 2, \cdots, N\}$. The problem is stated in the consensus formulation. Namely, the parameter X stores the cloned parameters $X^k \in \mathbb{R}^d$ for each client, $k = 1, \cdots, N$ such that $X^1 = \cdots = X^N$ or the problem (1) obtains the consensus, namely, the optimizer denoted by $X_* \in \mathbb{R}^{N_x}$ such that $X_* = (z_*, \cdots, z_*)^T$ for some $z_* \in \mathbb{R}^d$. Note that the local objective function f_k depends on the data of the k^{th} worker, but not on those of the other clients. The standard Federated Avgerage (*FebAvg*) algorithm consists of the following three steps:

1. the central server broadcasts the latest model X_t, to all the clients;
2. every worker, say i^{th} worker, lets $X_t^i = X_t$ and then performs one or few local updates with learning rate γ

$$X_{t+1}^i \leftarrow X_t^i - \gamma \nabla f_i(X_t^i), \tag{2}$$

3. the server then aggregates the local models, $X_{t+1}^1, \cdots X_{t+1}^N$, to produce the new global model X_{t+1} [9].

The main bottleneck in the Federated learning algorithm lies in step 3, which is generally orders of magnitude more expensive than the local computations. Furthermore, communications make the algorithm vulnerable to cybersecurity. Recent methods, therefore, aim at enhancing the privacy of FL by using a reduced model or even sacrificing system efficiency. However, providing privacy has to be carefully balanced with system efficiency [10]. One recent algorithm, called Scaffnew or ProxSkip, is shown to achieve the best communication efficiency, without sacrificing the convergence property, until today [12]. Among others, Scaffnew reformulates the step 2, (2) as follows: for $\ell = 0, 1, \cdots, n-1$,

$$X^i_{t+\frac{\ell+1}{n}} \leftarrow X^i_{t+\frac{\ell}{n}} - \gamma\left(\nabla f_i\left(X^i_{t+\frac{\ell}{n}}\right) - H_t^i\right), \tag{3}$$

where a certain shift, H_t^i is introduced for each client, and finite steps of the gradient descent method are used. Let $K = \mathbf{1} \otimes I \in \mathbb{R}^{N_x \times d}$, where I is the $d \times d$ identity matrix and $H_t = (H_t^1, \cdots, H_t^N)^T$, the complete algorithm is then given as in the following Algorithm 1. Note that ProxSkip introduces the probability p to choose to apply the local step. Thus, the Determinstic ProxSkip can be understood that $p = 1/n$, where n is the number of GD steps. This is summarized as the following algorithm 1. Note that the speed can be optimized if $p = 1/\sqrt{\kappa}$, where κ is the condition number of F.

Algorithm 1 Deterministic ProxSkip or SCAFFOLD [8]

Given a stepsize $\gamma > 0$, n, the number of GD steps, a stepsize for H update, $\omega = \frac{1}{n\gamma}$, initial iterate $X_0 = (X_0^1, \ldots, X_0^N)$, we perform the following until the convergence:

for $t = 0, 1, 2, \cdots$ **do**
  $X_t = Z_t$
  for $\ell = 0, 1, \ldots, n-1$ **do**
    $X_{t+\frac{\ell+1}{n}} = X_{t+\frac{\ell}{n}} - \gamma(\nabla F(X_{t+\frac{\ell}{n}}) - H_t)$
  end for
  $Z_{t+1} = K z_{t+1}$, with $z_{t+1} = \frac{1}{N}\sum_{i=1}^N X_{t+1}^i$
  $H_{t+1} = H_t + \frac{1}{n\gamma}(Z_{t+1} - X_{t+1})$
end for

This method is demonstrated to be convergent and popularly used in the Federated learning community. However, its convergence analysis is open until today. Note that the stochastic version of Federated algorithm, for example, ProxSkip algorithm is shown to converge linearly in [12]. The main result in this paper is to provide a mathematical proof of the linear convergence of the Algorithm 1. Furthermore, we shall also establish the linear convergence when the multiple GD scheme is replaced by exact solver. The rest of the paper is organized as follows: §2.1 discusses the formulation of the Federated learning algorithm and a couple of discussion on convex functional.

Throughout the paper, we shall denote $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ by the standard Euclidean norm and inner product, respectively. In particular, we shall use for $\alpha > 0$, the following scaled norm:

$$\|x\|_\alpha = \alpha\|x\|. \tag{4}$$

We also use the standard tensor product $\otimes$. For any convex functional F, which is twice continuously differential, the Hessian of F can be well-defined and it will be denoted by $\mathcal{H}_F$. By $\sigma(\mathcal{H}_F)$, we mean the spectrum of $\mathcal{H}_F$. We shall also denote $\rho(A)$ by the spectral radius of the matrix A. The notation $\text{Null}(A)$ denotes the null space of the matrix A.

2 Inexact Uzawa formulation of Federated Learning

In this section, we shall formulate the Federated learning as an inexact Uzawa method or Alternating Direction Method of Multiplier, [14]. We begin with a couple of preliminaries to further discuss the Algorithm 1.

2.1 Convex Functions and Some Preliminaries

In this section, we shall formally define problem and introduce a couple of important facts about convex functional. Let $F : \mathbb{R}^{N_x} \mapsto \mathbb{R}$ be proper, lower semi-continuous and a strongly convex functional. We consider the following optimization problem:

$$\operatorname*{arg\,min}_{\substack{X\in\mathbb{R}^{N_x} \\ Kz-X=0,\ z\in\mathbb{R}^d}} \quad F(X), \tag{5}$$

where Kz is N_c copies of z. We shall assume that F is λ_F-strongly convex and L_F-smooth. Namely,

$$F(X) - F(Y) - \langle \nabla F(Y), X - Y\rangle \geq \frac{\lambda_F}{2}\|X - Y\|^2, \quad \forall X, Y \in \mathbb{R}^{N_x}. \tag{6}$$

Furthermore, we have

$$\|\nabla F(X) - \nabla F(Y)\| \leq L_F\|X - Y\|, \quad \forall X, Y \in \mathbb{R}^{N_x}, \tag{7}$$

which is equivalent to

$$|F(X) - F(Y) - \langle \nabla F(Y), X - Y\rangle| \leq \frac{L_F}{2}\|X - Y\|^2, \quad \forall X, Y \in \mathbb{R}^{N_x}. \tag{8}$$

We note that for the constraint $Kz = X$, we can introduce the Lagrange multiplier H and reformulate the problem (5) in the following form:

$$\min_{X \in \mathbb{R}^{N_x}, z \in \mathbb{R}^d} \max_{H \in \mathbb{R}^{N_x}} F(X) + \langle H, Kz - X \rangle, \tag{9}$$

which can be further modified to be an Augmented Lagrangian method given as follows:

$$\min_{X \in \mathbb{R}^{N_x}, z \in \mathbb{R}^d} \max_{H \in \mathbb{R}^{N_x}} L_r(X, z, H), \tag{10}$$

where with $r > 0$ being a positive parameter,

$$L_r(X, z, H) = F(X) + \langle H, Kz - X \rangle + \frac{r}{2} \|Kz - X\|^2. \tag{11}$$

Note that the optimality conditions for Lagrangian and the Augmented Lagrangian are given as follows.

$$\begin{cases} \nabla F(X_*) - H_* = 0, \\ K^T H_* = 0, \\ Kz_* - X_* = 0. \end{cases} \tag{12}$$

On the other hand, the optimality condition for the Augmented Lagrangian formulation is given as follows:

$$\begin{cases} \nabla F(X_*) - H_* - r(Kz_* - X_*) = 0, \\ K^T H_* + rK^T(Kz_* - X_*) = 0, \\ Kz_* - X_* = 0 \end{cases} \tag{13}$$

Throughout the paper, we shall denote $A = \nabla F$ and among others, it is important to keep in mind that it holds

$$A(X_*) = H_*. \tag{14}$$

We shall interpret the Scaffnew algorithm as to find (X_*, z_*, H_*) for the system describing the optimality condition of the Augmented Lagrangian. Namely, we are intereted in solving the following nonlinear system:

$$\begin{pmatrix} \nabla F(\cdot) + rI & -rK & -I \\ -rK^T & rK^T K & K^T \\ -I & K & 0 \end{pmatrix} \begin{pmatrix} X_* \\ z_* \\ H_* \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{15}$$

We shall now introduce a couple of important fact about convex functional, some of which can be found at [7]. First of all, we introduce the Legendre-Fenchel convex conjugate of the functional F, denoted by $F^* : \mathbb{R}^{N_x} \to \mathbb{R}$, which is defined by the following:

$$F^*(S) = \max_{X \in \mathbb{R}^{N_x}} \{\langle S, X \rangle - F(X)\}. \tag{16}$$

Among others, we recall the following important lemma, stated for the dual of F:

Lemma 1. *Let $F : \mathbb{R}^{N_x} \to \mathbb{R}$ be a proper, convex, and lower semi-continuous. Then, F is λ_F-strongly convex if and only if F^* is $\frac{1}{\lambda_F}-$smooth.*

We also hereby note that the gradient of F^* is basically the inverse of ∇F. Namely, we have that

$$\nabla F^*(\nabla F(X)) = X \quad \text{and} \quad \nabla F(\nabla F^*(S)) = S. \tag{17}$$

We have only assumed that F is L_F-smooth and thus it may not be twice differentiable. On the other hand, its Hessian is well-defined almost everywhere due to Rademacher theorem [13]. λ_F convexity and L_F smoothness is equivalent to the fact that the Hessian of F has the following property:

$$\lambda_F I \le \mathcal{H}_F \le L_F I. \tag{18}$$

The mean value theorem will be useful in our analysis. Thus, we shall invoke this in many different places. On the other hand, F is not sufficiently smooth, i.e., not twice continuously differentiable, then, we shall use the smoothed version of F, denoted by F_δ, which is obtained by the mollifier applied to F. Namely,

$$F_\delta = F \star e_\delta, \tag{19}$$

where e_δ is the standard mollifying function [4]. Note that it will be only used for analysis. Note that it holds true that

$$\partial_j F_\delta(x) := \int_{\mathbb{R}^{N_x}} \partial_j F(x-y) e_\delta(y)\, dy, \quad \forall j = 1, \cdots, N_x. \tag{20}$$

In practice, it will not be needed at all. We shall list a couple of important fact about F_δ.

Lemma 2. *Let F be λ_F-strongly convex and L_F-smooth. Then, F_δ satisfies the following three properties:*

(a) F_δ is convex

(b) F_δ inherits λ_F-strong convexity as well as L_F-smoothness independent of δ.

(c) ∇F_δ converges to ∇F as $\delta \to 0$.

As a consequence of the lemma, we have that H_{F_δ} is symmetric positive definite with the following property that for all $\delta > 0$, we have

$$\lambda_F \le \min\left\{\sigma(\mathcal{H}_{F_\delta}(x)) : x \in \mathbb{R}^{N_x}\right\} \quad \text{and} \quad \max\left\{\sigma(\mathcal{H}_{F_\delta}(x)) : x \in \mathbb{R}^{N_x}\right\} \le L_F. \tag{21}$$

Note that it is well-known that if F is λ_F-strongly convex and L_F smooth, then for $\gamma \le 1/L_F$, the functional $G(x) = \|x\|^2/2 - \gamma F(x)$ is $1-\gamma\lambda_F$ smooth and $1-\gamma L_F$ strongly convex. We shall demonstrate the usefulness of the mollifier to prove some important fact, which will be used frequently later in this paper.

Lemma 3. *Let $w_k = x_k - \omega\nabla F(x_k)$ and $w = x - \omega\nabla F(x)$, where F is λ_F-strongly convex and L_F smooth. Then it holds that if ω is such that $0 < \omega \le 1/L_F$, then it holds:*

$$\|w_k - w\| \le (1-\omega\lambda_F)\|x_k - x\|. \tag{22}$$

On the other hand, if $\omega = \frac{2}{\lambda_F + L_F}$, then we have

$$\|w_k - w\| \le \frac{\kappa(F)-1}{\kappa(F)+1}\|x_k - x\|, \tag{23}$$

where $\kappa(F) = \frac{L_F}{\lambda_F}$.

Proof. Since F is λ_F-strongly convex and L_F smooth, we have that

$$\lambda_F \|x-y\| \leq \|\nabla F(x) - \nabla F(y)\| \leq L_F \|x-y\|, \quad \forall x, y \in \mathbb{R}^{N_x}. \tag{24}$$

We consider F_δ and define $w_\delta = x - \omega \nabla F_\delta(x)$ and $w_{k,\delta} = x_k - \omega \nabla F_\delta(x_k)$. We then see that

$$\begin{aligned}
\|w_{k,\delta} - w_\delta\| &= \|x_k - x - \omega(\nabla F_\delta(x_k) - \nabla F_\delta(x))\| \\
&= \|(I - \omega \nabla F_\delta)(x_k) - (I - \omega \nabla F_\delta)(x)\| \\
&\leq \sup_\xi \|I - \omega \mathcal{H}_{F_\delta}(\xi)\| \|x_k - x\| \leq \frac{\kappa(F_\delta) - 1}{\kappa(F_\delta) + 1} \|x_k - x\| \\
&\leq \frac{\kappa(F) - 1}{\kappa(F) + 1} \|x_k - x\|.
\end{aligned}$$

By taking limit $\delta \to 0$, we complete the first part of the result. We have also used the fact that the following function when $t \geq 1$, takes the maximum when t is the largest, i.e.,

$$\frac{L_F}{\lambda_F} = \arg\max_{t \geq 1} \left\{ \frac{t-1}{t+1} \right\}. \tag{25}$$

The first result can be done similarly, and easier. Thus, we skip the proof. This completes the proof. □

2.2 Augmented Lagrangian and Inexact Uzawa method

In this section, we shall now present the general inexact Uzawa iterative framework of the Algorithm 1. We recall that the Lagrangian is given as follows for $r > 0$,

$$L_r(X, z, H) = F(X) + \langle H, Kz - X \rangle + \frac{r}{2} \|Kz - X\|^2. \tag{26}$$

The method then reads as follows:

$$X_{k+1} = \arg\min_X L_r(X, z_k, H_k) \tag{27a}$$

$$z_{k+1} = \arg\min_z L_r(X_{k+1}, z, H_k) \tag{27b}$$

$$H_{k+1} = H_k + \omega(Kz_{k+1} - X_{k+1}). \tag{27c}$$

Generally, X_{k+1} can also be defined approximately:

$$X_{k+1} \approx \arg\min_X L_r(X, z_k, H_k), \tag{28a}$$

while no changes are made for z_{k+1} and H_{k+1}. This shall induce different solution technique. To describe the general case in one stroke. We first observe that X_{k+1} obtained from (156) can be characterized as follows:

$$(A + rI)(X_{k+1}) = \nabla F(X_{k+1}) + rX_{k+1} = H_k + rKz_k \tag{29}$$

or equivalently, we have

$$X_{k+1} = A_r^*(H_k + rKz_k), \tag{30}$$

where $A_r = A + rI = \nabla G$ with $G = \frac{r}{2}\|\cdot\|^2 + F(\cdot)$ and $A_r^* = \nabla G^*$. We shall denote D_r^* by the approximate A_r^* such that it satisfies the following relationship:

Assumption 2.1. The approximate inverse of A_r satisfies the following identity:

$$D_r^*(H_* + rKz_*) = X_*. \tag{31}$$

We remark that for the operator A_r, it holds true that

$$A_r(X_*) = \nabla F(X_*) + rX_* = H_* + rKz_*. \tag{32}$$

Thus, A_r^* satisfies the Assumption 2.1. For the inexact solve, to achieve the Assumption 2.1, it is notable that the initial iterate has to be set as rKz_*. Therefore, we shall assume that when D_r^* requires the initial iterate, what we mean by $D_r^*(H_k + rKz_k)$ is that it uses $X_k = Kz_k$ as a start. Under this condition on the initial iterate, we note that the gradient descent method that is discussed in §3.2 also satisfies the Assumption 2.1. The general version of the Algorithm 1 can be written as the following Algorithm 2.

Algorithm 2 Inexact Uzawa formulation of FL

Given H_0 such that $K^T H_0 = 0$, updates are obtained as follows:

for $k = 0, 1, 2, \cdots, K-1$ **do**

X_{k+1} update: (with $X_k = Kz_k$),

$$X_{k+1} = D_r^*(H_k + rKz_k), \tag{33}$$

z_{t+1} update:

$$K^T H_k + rK^T(Kz_{k+1} - X_{k+1}) = 0, \tag{34}$$

Update the Lagrange multiplier:

$$H_{k+1} = H_k + \omega(Kz_{k+1} - X_{k+1}). \tag{35}$$

end for

Lemma 4. *If $K^T H = 0$, then*

$$K^T H_{k+1} = K^T H_k + K^T \omega(Kz_{k+1} - X_{k+1}) = 0. \tag{36}$$

Proof. We note that A is the block matrix and thus, we have that

$$K^T \omega = \begin{pmatrix} I & \cdots & I \end{pmatrix} \begin{pmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & 0 & \cdots \end{pmatrix} = \tag{37}$$

□

We remark that the Algorithm 1 can, therefore, be interpreted as an inexact Uzawa method. In particular, if $D_r = A_r$, then we obtain the Uzawa method with Gauss-Seidel block solver. Before we analyze the convergence of the algorithm, we shall list important facts about the algorithm.

Lemma 5. *If H_0 is given so that $K^T H_0$, then the Algorithm 2 produces H_k such that $K^T H_k = 0$ for all $k = 1, 2, \cdots$.*

Proof. Multiplying K^T to (184) and adding it to (157), we have $K^T H_{k+1} = 0$ for all $k = 0, 1, \cdots$. This completes the proof. □

We now define $P_Z = K(K^T K)^{-1} K^T$ and $Q_Z = I - P_Z$. First we note that P_Z has the spectral radius less than or equal to one.

Lemma 6. *Let $P_Z = K(K^T K)^{-1} K^T$. Then it holds that*

$$\|P_Z X\| \leq \|X\|, \quad \forall X \quad \textit{or equivalently} \quad \lambda_{max}(P_Z) = 1. \tag{38}$$

Proof. We consider the Rayleigh Quotient. Using the Jenssen's inequality, we obtain that

$$\sup_{\|X\| \neq 0} \frac{\langle K(K^T K)^{-1} K^T X, X \rangle}{\langle X, X \rangle} = \sup_{\|X\| \neq 0} \frac{\frac{1}{N} \left\langle \sum_{i=1}^N X_i, \sum_{i=1}^N X_i \right\rangle}{\sum_{i=1}^N \|X_i\|^2} \leq 1.$$

This completes the proof. □

Secondly, we note that both P_Z and Q_Z are symmetric. Furthermore, one can see for all $k = 0, 1, \cdots$,

$$\begin{aligned}
P_Z H_k &= 0 && \text{(39a)}\\
P_Z H_* &= 0, && \text{(39b)}\\
Q_Z H_k &= H_k, && \text{(39c)}\\
Q_Z H_* &= H_*, && \text{(39d)}\\
P_Z (Kz) &= Kz, && \text{(39e)}\\
Q_Z (Kz) &= 0. && \text{(39f)}
\end{aligned}$$

In particular, we have that

Lemma 7. *The following holds:*

$$\langle Kz_* - Kz_i, H_* - H_j \rangle = 0, \quad \forall i, j \geq 0. \tag{40}$$

and

$$\langle X_{k+1} - X_*, H_{k+1} - H_* \rangle = \frac{1}{\omega} \langle H_k - H_{k+1}, H_{k+1} - H_* \rangle \tag{41}$$

Proof. Note that

$$\langle Kz_* - Kz_i, H_* - H_j \rangle = \langle Kz_* - Kz_i, Q_Z(H_* - H_j) \rangle = \langle Q_Z(Kz_* - Kz_i), H_* - H_j \rangle = 0.$$

We now notice that since $H_{k+1} - H_k = -\omega Q_Z X_{k+1}$ and $Q_Z X_* = 0$. Thus, we have $H_{k+1} - H_k = -\omega Q_Z(X_{k+1} - X_*)$. This gives

$$\begin{aligned}\langle X_{k+1} - X_*, H_{k+1} - H_* \rangle &= \langle Q_Z(X_{k+1} - X_*), H_{k+1} - H_* \rangle \\ &= \frac{1}{\omega}\langle H_k - H_{k+1}, H_{k+1} - H_* \rangle.\end{aligned}$$

This completes the proof. □

Lastly, we show the nonexpansiveness as stated below. More precisely, we have the following result:

Lemma 8. *The following nonexpansiveness holds true:*

$$\|P_Z X\|^2 + \|Q_Z X\|^2 = \|X\|^2, \quad \forall X \in \mathbb{R}^{N_x}. \tag{42}$$

Proof. We observe that

$$\begin{aligned}\|P_Z X\|^2 + \|Q_Z X\|^2 &= \|P_Z X\|^2 + \|P_Z X\|^2 - 2\langle P_Z X, X \rangle + \|X\|^2 \\ &= \|P_Z X\|^2 + \|P_Z X\|^2 - 2\langle P_Z X, P_Z X \rangle + \|X\|^2 \\ &= \|X\|^2\end{aligned}$$

This completes the proof. □

We shall use the standard notation that for all $k \geq 0$ to discuss the convergence:

$$\begin{aligned}E_k^X &= X_* - X_k \\ E_k^Z &= Kz_* - Kz_k \\ E_k^H &= H_* - H_k.\end{aligned}$$

In the following two sections, we shall present convergence analysis. The first section will deal with the Gauss-Seidel method in which $D_r = A_r$. The second section shall deal with D_r^* being the $n-$step Gradient Descent method. Algorithmic details are presented in each subsection.

In passing to each subsection, we shall present an important result from the nonexpansiveness:

Lemma 9. *The Algorithm 2 produces iterate (X_k, z_k, H_k), for which the following error bound holds true:*

$$\left\|E_{k+1}^H\right\|^2 + \omega^2 \left\|E_{k+1}^Z\right\|^2 = \left\|E_k^H - \omega(A_r^*(H_* + rKz_*) - D_r^*(H_k + rKz_k))\right\|^2.$$

Proof. The Algorithm 2 leads to iterates, given as follows:

$$\begin{aligned}X_{k+1} &= D_r^*(H_k + rKz_k) \\ Kz_{k+1} &= K(rK^TK)^{-1}(rK^TX_{k+1} - K^TH_k) \\ H_{k+1} &= H_k + \omega(-X_{k+1} + Kz_{k+1}),\end{aligned}$$

where D_r^* is an approximate of A_r^*, the Fenchel-dual conjugate of A_r. We first notice that if $K^T H_0 = 0$, then $K^T H_k = 0$ and also $K^T H_* = 0$. This is due to the proximal operator P_Z. Therefore, we have

$$\begin{aligned} X_{k+1} &= D_r^*(H_k + rKz_k) \\ Kz_{k+1} &= K(rK^TK)^{-1}(rK^TD_r^*(H_k + rKz_k)) = P_Z[D_r^*(H_k + rKz_k)] \\ H_{k+1} &= H_k + \omega(-X_{k+1} + Kz_{k+1}) \end{aligned}$$

On the other hand, we have that

$$\begin{aligned} X_* &= A_r^*(H_* + rKz_*) \\ Kz_* &= P_Z[A_r^*(H_* + rKz_*)] \\ H_* &= H_* + \omega(-X_* + Kz_*). \end{aligned}$$

Therefore, we have the following error equation:

$$\begin{aligned} E_{k+1}^X &= A_r^*(H_* + rKz_*) - D_r^*(H_k + rKz_k) \\ E_{k+1}^Z &= P_Z[A_r^*(H_* + rKz_*) - D_r^*(H_k + rKz_k)] \\ E_{k+1}^H &= H_* - H_k + \omega(-X_* + X_{k+1} + Kz_* - Kz_{k+1}) \end{aligned}$$

Rearranging the error in H variable, we have

$$E_{k+1}^H - \omega E_{k+1}^Z = E_k^H - \omega E_{k+1}^X = E_k^H - \omega\left(A_r^*(H_* + rKz_*) - D_r^*(H_k + rKz_k)\right). \tag{43}$$

Taking the squared norm on both sides of the equation (113), and using the orthogonality, we have

$$\|E_{k+1}^H\|^2 + \omega^2\|E_{k+1}^Z\|^2 = \|E_k^H - \omega(A_r^*(H_* + rKz_*) - D_r^*(H_k + rKz_k))\|^2.$$

This completes the proof.

$\square$

3 Convergence analysis of Algorithm 2

In this section, we shall present the convergence of the Algorithm 2. This section consists of two parts. One part is the convergence when $D_r^* = A_r^*$ and the other part is the convergence when D_r^* is the $n-$step GD method.

3.1 Linear Convergence of Algorithm 2 for $D_r = A_r$

In this section, we shall establish the linear convergence of the exact Uzawa method with $D_r = A_r$.

Theorem 1. *The Algorithm 2 with GS as local solve, and $\omega = \frac{2}{\lambda_{G^*} + L_{G^*}}$ has the convergence rate given as follows:*

$$\|E_{k+1}^H\|^2 + \|E_{k+1}^Z\|_\omega^2 \le \rho_{GS}^2(r, L_F, \lambda_F)\left(\|E_k^H\|^2 + \|E_k^Z\|_\omega^2\right), \tag{44}$$

where with $\kappa(G) = \frac{r+L_F}{r+\lambda_F}$,

$$\rho_{GS}^2(r, L_F, \lambda_F) = \left(\frac{\kappa(G)-1}{\kappa(G)+1}\right)^2 + \left(\frac{r}{r+\lambda_F}\right)^2 = \left(\frac{L_F-\lambda_F}{2r+L_F+\lambda_F}\right)^2 + \left(\frac{r}{r+\lambda_F}\right)^2 \qquad (45)$$

We also have that

$$\|E_{k+1}^X\|^2 \leq \frac{1}{(r+\lambda_F)^2}\left(\|E_k^H\|^2 + \|E_k^Z\|_\omega^2\right).$$

Furthermore, there exists an interval around 0 *such that if* $r \in [0, r_{\rm opt}]$, *then the convergence can be achieved. For larger* $r > L_F^2/(8\lambda_F)$, *the convergence is also guaranteed for any* L_F *and* λ_F.

Proof. The Algorithm 2 produces iterates given as follows:

$$\begin{aligned} X_{k+1} &= A_r^*(H_k + rKz_k) \\ Kz_{k+1} &= K(rK^TK)^{-1}(rK^TX_{k+1} - K^TH_k) \\ H_{k+1} &= H_k + \omega(-X_{k+1} + Kz_{k+1}), \end{aligned}$$

where A_r^* is the Fenchel-dual conjugate of A_r. We first notice that if $K^TH_0 = 0$, then $K^TH_k = 0$ and also $K^TH_* = 0$. Now due to Lemma 9, we have that with $D_r = A_r$,

$$\|E_{k+1}^H\|^2 + \omega^2\|E_{k+1}^Z\|^2 = \|E_k^H - \omega(A_r^*(H_* + rKz_*) - A_r^*(H_k + rKz_k))\|^2$$

On the other hand, we have that

$$\begin{aligned} X_* &= A_r^*(H_* + rKz_*) \\ Kz_* &= P_Z[A_r^*(H_* + rKz_*)] \\ H_* &= H_* + \omega(-X_* + Kz_*). \end{aligned}$$

Therefore, we have the following error equation:

$$\begin{aligned} X_* - X_{k+1} &= A_r^*(H_* + rKz_*) - A_r^*(H_k + rKz_k) \\ Kz_* - Kz_{k+1} &= P_Z[A_r^*(H_* + rKz_*) - A_r^*(H_k + rKz_k)]. \end{aligned}$$

The trick is to multiply $-\omega$ for E_{k+1}^Z error term and to obtain

$$-\omega\left(Kz_* - Kz_{k+1}\right) = -\omega\left(P_Z[A_r^*(H_* + rKz_*) - A_r^*(H_k + rKz_k)]\right).$$

Lastly, for H, we have

$$\begin{aligned} H_* - H_{k+1} &= H_* - H_k + \omega(-X_* + X_{k+1} + Kz_* - Kz_{k+1}) \\ &= H_* - H_k - \omega[X_* - X_{k+1} - (Kz_* - Kz_{k+1})] \\ &= H_* - H_k - \omega[A_r^*(H_* + rKz_*) - A_r^*(H_k + rKz_k) \\ &\quad - P_Z[A_r^*(H_* + rKz_*) - A_r^*(H_k + rKz_k)]] \\ &= H_* - H_k - \omega Q_Z(A_r^*(H_* + rKz_*) - A_r^*(H_k + rKz_k)) \\ &= Q_Z[H_* - H_k - \omega(A_r^*(H_* + rKz_*) - A_r^*(H_k + rKz_k))] \end{aligned}$$

Thus, we have that

$$H_* - H_{k+1} - \omega(Kz_* - Kz_{k+1}) = H_* - H_k - \omega(A_r^*(H_* + rKz_*) - A_r^*(H_k + rKz_k)).$$

We now define two important quantities:

$$\begin{aligned} A^{Z_*}_{H_*,H_k} &:= A_r^*(H_* + rKz_*) - A_r^*(H_k + rKz_*) && (46)\\ A^{H_k}_{Z_*,Z_k} &= A_r^*(H_k + rKz_*) - A_r^*(H_k + rKz_k). && (47)\end{aligned}$$

Then, we have that

$$\begin{aligned} \|E^H_{k+1} - \omega E^Z_{k+1}\|^2 &= \|E^H_k - \omega(A_r^*(H_* + rKz_*) - A_r^*(H_k + rKz_k))\|^2\\ &= \|E^H_k - \omega(A^{Z_*}_{H_*,H_k} + A^{H_k}_{Z_*,Z_k})\|^2,\\ &\le \|E^H_k - \omega A^{Z_*}_{H_*,H_k}\|^2 + \omega^2\|A^{H_k}_{Z_*,Z_k}\|^2\\ &\quad -2\omega\left\langle E^H_k - \omega A^{Z_*}_{H_*,H_k}, A^{H_k}_{Z_*,Z_k}\right\rangle.\end{aligned}$$

Since $\lambda_{G^*} = 1/(r + L_F)$ and $L_{G^*} = 1/(r + \lambda_F)$, we have that

$$\omega = \frac{2}{\lambda_{G^*} + L_{G^*}} = \frac{2}{\frac{1}{r+L_F} + \frac{1}{r+\lambda_F}} = \frac{2(r + \lambda_F)(r + L_F)}{2r + L_F + \lambda_F} \tag{48}$$

and

$$\left\langle E^H_k - \omega A^{Z_*}_{H_*,H_k}, A^{H_k}_{Z_*,Z_k}\right\rangle \le \left(\frac{\kappa(G) - 1}{\kappa(G) + 1}\right)\frac{r}{r + \lambda_F}\|E^H_k\|\|E^z_k\|.$$

Thus, we have that

$$\begin{aligned} -2\omega\langle E^H_k - \omega A^{Z_*}_{H_*,H_k}, A^{H_k}_{Z_*,Z_k}\rangle &\le 2\left(\frac{\kappa(G) - 1}{\kappa(G) + 1}\right)\frac{r}{r + \lambda_F}\|E^H_k\|\|E^Z_k\|_\omega && (49)\\ &\le \left(\frac{r}{r + \lambda_F}\right)^2\|E^H_k\|^2 + \left(\frac{\kappa(G) - 1}{\kappa(G) + 1}\right)^2\|E^Z_k\|^2_\omega. && (50)\end{aligned}$$

Again with $\omega = 2/(\lambda_{G^*} + L_{G^*})$, we have that

$$\begin{aligned} &\|E^H_k - \omega(A_r^*(H_* + rKz_*) - A_r^*(H_k + rKz_*))\|^2\\ &\quad\le \|(H_* + rKz_*) - (H_k + rKz_*) - \omega(A_r^*(H_* + rKz_*) - A_r^*(H_k + rKz_*))\|^2\\ &\quad\le \left(\frac{\kappa(G) - 1}{\kappa(G) + 1}\right)^2\|E^H_k\|^2.\end{aligned}$$

On the other hand, we have that

$$\omega^2\|A_r^*(H_k + rKz_*) - A_r^*(H_k + rKz_k)\|^2 \le \left(\frac{r}{r + \lambda_F}\right)^2\|E^Z_k\|^2_\omega.$$

Therefore, we obtain that

$$\|E^H_{k+1}\|^2 + \|E^Z_{k+1}\|^2_\omega \le \left\{\left(\frac{\kappa(G) - 1}{\kappa(G) + 1}\right)^2 + \left(\frac{r}{r + \lambda_F}\right)^2\right\}\left(\|E^H_k\|^2 + \|E^Z_k\|^2_\omega\right).$$

This provides the convergence rate. Finally, we notice that for all $r \geq 0$,

$$\frac{r^2}{\omega^2} = \frac{r^2(2r + L_F + \lambda_F)^2}{4(r + \lambda_F)^2(r + L_F)^2} < 1.$$

Thus, we obtain that due to the orthogonality,

$$\begin{aligned}
\|E_{k+1}^X\|^2 &= \|A_r^*(H_* + rKz_*) - A_r^*(H_k + rKz_k)\|^2 \\
&\leq \frac{1}{(r+\lambda_F)^2}\|E_k^H - rE_k^Z\|^2 = \frac{1}{(r+\lambda_F)^2}\left(\|E_k^H\|^2 + r^2\|E_k^Z\|^2\right) \\
&= \frac{1}{(r+\lambda_F)^2}\left(\|E_k^H\|^2 + \frac{r^2}{\omega^2}\|E_k^Z\|_\omega^2\right) \leq \frac{1}{(r+\lambda_F)^2}\left(\|E_k^H\|^2 + \|E_k^Z\|_\omega^2\right)
\end{aligned}$$

We now discuss the convergence factor denoted by ρ_{GS}^2 and given by

$$\begin{aligned}
f(r) &= \rho_{GS}^2(r, L_F, \lambda_F) = \left(\frac{\kappa(G) - 1}{\kappa(G) + 1}\right)^2 + \left(\frac{r}{r + \lambda_F}\right)^2 \\
&= \left(\frac{L_F - \lambda_F}{2r + L_F + \lambda_F}\right)^2 + \left(\frac{r}{r + \lambda_F}\right)^2 \\
&= \frac{4r^4 + 4r^3(L_F + \lambda_F) + r^2((L_F + \lambda_F)^2 + (L_F - \lambda_F)^2) + 2r\lambda_F(L_F - \lambda_F)^2 + \lambda_F^2(L_F - \lambda_F)^2}{4r^4 + 4r^3(L_F + 3\lambda_F) + r^2((L_F + \lambda_F)^2 + 8\lambda_F(L_F + \lambda_F) + 4\lambda_F^2) + r(4\lambda_F^2(L_F + \lambda_F) + 2\lambda_F(L_F + \lambda_F)^2) + \lambda_F^2(L_F - \lambda_F)^2}.
\end{aligned}$$

Clearly, if $r = 0$, then $f(r) = \left(\frac{L_F - \lambda_F}{L_F + \lambda_F}\right)^2$. Thus, the convergence is guaranteed with the rate similar to the Gradient descent method. On the other hand, for all $L_F, \lambda_F \geq 0$, we have the convergence can be achieved under the sufficient condition that

$$r > \frac{L_F^2}{8\lambda_F}. \tag{51}$$

We also note that a simple calculation shows that the derivative of $f(r)$ is given as follows:

$$f'(r) = \frac{-4(L_F - \lambda_F)^2}{(2r + L_F + \lambda_F)^3} + \frac{2r\lambda_F}{(r + \lambda_F)^3}. \tag{52}$$

Therefore, for small r, it takes the negative sign. Thus, but it changes its sign for larger r and continues to be positive. Thus, there exists a single critical point, which gives the optimal $r_{\rm opt}$. On the other hand, the convergence rate is smaller than one at $r = 0$. Thus, the optimal $r_{\rm opt}$ is such that

$$f'(r_{\rm opt}) = 0. \tag{53}$$

Readers refer to the Figure 1). We can conclude that it converges linearly for all fixed $0 \leq r \leq r_{\rm opt}$. This completes the proof. □

Remark 1. *We remark that it is more natural to write the convergence rate in terms of $\kappa(G^*)$, the condition number of G^* since the choice of ω is made for solving the system relevant to G^*. However, it is also fine to use $\kappa(G)$ since it is more relvant to the problem to be solved and we have that*

$$\kappa(G^*) = \kappa(G) \quad \rightarrow \quad \frac{\kappa(G) - 1}{\kappa(G) + 1} = \frac{\kappa(G^*) - 1}{\kappa(G^*) + 1}. \tag{54}$$

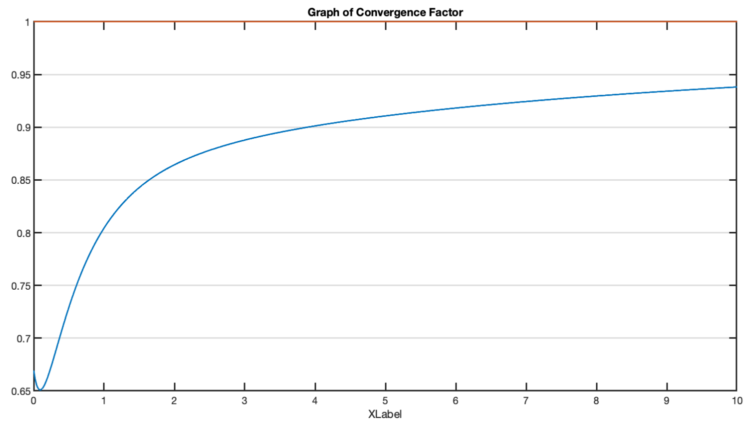

Figure 1: Graphs of the convergence factor as a function of r for $L_F = 5$ and $\lambda_F = 0.5$

3.2 Convergence analysis of Algorithm 2 for $D_r^* \neq A_r^*$

In this section, our goal is to establish the convergence of inexact Uzawa with Gauss-Seidel method, $D_r = A_r$, is replaced by inexact local solve for X-variable. This includes the $n-$step of Gradient Descent iteration. We recall that the Gauss-Seidel method can be interpreted as to solve the following optimization exactly in X-update (156) of the Algorithm 2:

$$\min_X L_r(X, z_k, H_k). \tag{55}$$

We now let $G(X) = L_r(X, z_k, H_k)$, i.e.,

$$G(X) = F(X) + \langle H_k, Kz_k - X\rangle + \frac{r}{2}\|Kz_k - X\|^2 \tag{56}$$

and let $Y_* = \arg\min_X G(X)$. Note that we reserve X_{k+1} as the result of $N-$step of GD method. Then the following statements hold true:

1. $Y_* = A_r^*(H_k + rKz_k)$
2. $\nabla F(Y_*) + rY_* = H_k + rKz_k$
3. $\nabla G(Y_*) = 0.$

Note that it is easy to show that G is $r + L_F$ smooth and $r + \lambda_F$ strongly convex. Our discussion is for general inexact local solve and the outcome will be denoted by X_{k+1}, i.e.,

$$X_{k+1} = D_r^*(H_k + rKz_k). \tag{57}$$

For the convergence analysis, the exact Gauss-Seidel case will be used as an intermediate step. Thus, naturally, the convergence estimate could be made to be sharper. We begin our discussion by introducing two quantities:

$$\begin{aligned} \mathsf{E}_1 &:= E_k^H - \omega\{A_r^*(H_* + rKz_*) - A_r^*(H_k + rKz_k)\} \\ \mathsf{E}_2 &:= A_r^*(H_k + rKz_k) - D_r^*(H_k + rKz_k). \end{aligned}$$

It is evident that E_1 is for the error of the Algorithm 2 with $D_r^* = A_r^*$ and E_2 represents the difference between two iterates, one from the inexact solve for $X-$variable and the other from the exact solve. We shall now see that Lemma 9 can lead to the following estimate easily.

$$\begin{aligned}\left\|E_{k+1}^H\right\|^2 + \omega^2 \left\|E_{k+1}^Z\right\|^2 &= \left\|E_k^H - \omega\{A_r^*(H_* + rKz_*) - D_r^*(H_k + rKz_k)\}\right\|^2 \\ &\le \|\mathsf{E}_1\|^2 + \omega^2\|\mathsf{E}_2\|^2 + 2\omega\|\mathsf{E}_1\|\|\mathsf{E}_2\|.\end{aligned}$$

The first term is relevant to the Algorithm 2 with $D_r = A_r$. We shall assume that there exists $\delta < 1$ such that

$$\|\mathsf{E}_2\|_\omega^2 \le \delta^2 \left(\|E_k^H\|^2 + \|E_k^Z\|_\omega^2\right). \tag{58}$$

Namely, the inexact solve provides a reasonably close solution to Y_*. Under this assumption, we shall establish the convergence of Algorithm 2. How small δ can be, for still allowing the convergence is determined by the close investigation of the convergence of Algorithm 2 with exact local solve after incorporating δ. We note that δ is provided in Lemma 11 below.

We are now in a position to provide a main instrumental lemma in this section.

Theorem 2. *Let δ_{GS} be the convergence rate for the Algorithm 2, when $D_r = A_r$, as given in equation* (44), *$\omega = \frac{2}{\lambda_{G^*} + L_{G^*}}$ and $D_r^*(H_k + rKz_k)$ be such that the equation* (58) *hold. Then, we have the following estimate:*

$$\left\|E_{k+1}^H\right\|^2 + \left\|E_{k+1}^Z\right\|_\omega^2 \le (\delta_{GS} + \delta)^2 \left(\|E_k^H\|^2 + \|E_k^Z\|_\omega^2\right).$$

Proof. We shall let $E^2 = \|E_k^H\|^2 + \|E_k^Z\|_\omega^2$. Due to the equation (58), we have that

$$\|\mathsf{E}_2\|_\omega^2 \le \delta^2 E^2. \tag{59}$$

We then obtain that

$$\begin{aligned}\left\|E_{k+1}^H\right\|^2 + \left\|E_{k+1}^Z\right\|_\omega^2 &\le \|\mathsf{E}_1\|^2 + \|\mathsf{E}_2\|_\omega^2 + 2\|\mathsf{E}_1\|\|\mathsf{E}_2\|_\omega \\ &= \delta_{GS}^2 E^2 + \delta^2 E^2 + 2\delta_{GS}\delta E^2 = (\delta + \delta_{GS})^2 E^2.\end{aligned}$$

This completes the proof. □

We also note that the $n-$step of GD method is basically, given as follows: for a given step size $\gamma > 0$ and $X_k = Kz_k$,

$$\begin{aligned}X_{k+\frac{1}{n}} &= X_k - \gamma\nabla G(X_k) && \text{(60a)}\\ X_{k+\frac{2}{n}} &= X_{k+\frac{1}{n}} - \gamma\nabla G(X_{k+\frac{1}{n}}) && \text{(60b)}\\ &\vdots && \text{(60c)}\\ X_{k+\frac{n-1}{n}} &= X_{k+\frac{n-2}{n}} - \gamma\nabla G(X_{k+\frac{n-2}{n}}) && \text{(60d)}\\ X_{k+\frac{n}{n}} &= X_{k+\frac{n-1}{n}} - \gamma\nabla G(X_{k+\frac{n-1}{n}}), && \text{(60e)}\end{aligned}$$

where ∇G is given as follows:

$$\nabla G(X) = \nabla F(X) + rX - H_k - rKz_k. \tag{61}$$

We shall now present a simple but important lemma:

Lemma 10. *Let* $\gamma = \frac{2}{\lambda_G + L_G}$ *and* $D_r^*(H_k + rKz_k)$ *be the n-step GD, as given in the equation* (60)*, then it holds true that*

$$\|A_r^*(H_k + rKz_k) - D_r^*(H_k + rKz_k)\|^2 \leq \left(\frac{\kappa(G) - 1}{\kappa(G) + 1}\right)^{2n} \|Y_* - X_k\|^2.$$

Proof. Let $Y_* = \arg\min_X G(X)$. On the other hand, X_{k+1} is obtained by the following iteration:

$$\begin{aligned} X_{k+\frac{1}{n}} &= X_k - \gamma \nabla G(X_k) \\ X_{k+\frac{2}{n}} &= X_{k+\frac{1}{n}} - \gamma \nabla G(X_{k+\frac{1}{n}}) \\ &\vdots \\ X_{k+\frac{n-1}{n}} &= X_{k+\frac{n-2}{n}} - \gamma \nabla G(X_{k+\frac{n-2}{n}}) \\ X_{k+\frac{n}{n}} &= X_{k+\frac{n-1}{n}} - \gamma \nabla G(X_{k+\frac{n-1}{n}}). \end{aligned}$$

Since G is $r + L_F$ smooth and $r + \lambda_F$ strongly convex, we see that by Lemma 3, we have that

$$\|Y_* - X_{k+1}\|^2 \leq \left(\frac{\kappa(G) - 1}{\kappa(G) + 1}\right)^{2n} \|Y_* - X_k\|^2.$$

This completes the proof. □

Remark 2. *The choice of* $\gamma = \frac{2}{\lambda_G + L_G}$ *is optimal for GD and the convergence rate can be calculated. Since* $\kappa(G) = \frac{r + L_F}{r + \lambda_F}$*, we have that*

$$\frac{\kappa(G) - 1}{\kappa(G) + 1} = \frac{L_F - \lambda_F}{2r + L_F + \lambda_F}. \tag{62}$$

The convergence is even faster for large r and large n.

We shall now see that the norm of E_2 can be made to be quite small.

Lemma 11. *Let* $\gamma = \frac{2}{\lambda_G + L_G}$, $\omega = \frac{2}{\lambda_{G^*} + L_{G^*}}$ *and* $D_r^*(H_k + rKz_k)$ *be the n-step GD, as given in the equation* (60)*, then it holds true that*

$$\|\mathsf{E}_2\|_\omega^2 \leq 4 \left(\frac{\kappa(G) - 1}{\kappa(G) + 1}\right)^{2n} \left(\|E_k^H\|^2 + \|E_k^Z\|_\omega^2\right).$$

Proof. By Lemma 10, we have that

$$\begin{aligned} \|\mathsf{E}_2\|_\omega^2 &= \omega^2 \|A_r^*(H_k + rKz_k) - D_r^*(H_k + rKz_k)\|^2 \\ &\leq \left(\frac{\kappa(G) - 1}{\kappa(G) + 1}\right)^{2n} \omega^2 \|Y_* - X_k\|^2. \end{aligned}$$

On the other hand, we have that since $X_k = Kz_k$ and due to Theorem 1,

$$\begin{aligned}
\omega^2\|Y_* - X_k\|^2 &\leq 2\omega^2\|Y_* - X_*\|^2 + 2\|X_* - X_k\|_\omega^2 \\
&\leq \frac{2\omega^2}{(r+\lambda_F)^2}\left(\|E_k^H\|^2 + \|E_k^Z\|_\omega^2\right) + 2\|X_* - X_k\|_\omega^2 \\
&\leq \frac{8(r+L_F)^2}{(2r+L_F+\lambda_F)^2}\left(\|E_k^H\|^2 + \|E_k^Z\|_\omega^2\right) + 2\|E_k^Z\|_\omega^2 \\
&\leq 4\left(\|E_k^H\|^2 + \|E_k^Z\|_\omega^2\right)
\end{aligned}$$

The last inequality is due to the fact that $X_k = Kz_k$ and

$$\omega = \frac{2(r+\lambda_F)(r+L_F)}{(2r+L_F+\lambda_F)}, \tag{63}$$

thus,

$$\frac{2\omega^2}{(r+\lambda_F)^2} = \frac{8(r+L_F)^2}{(2r+L_F+\lambda_F)^2} \leq 2. \tag{64}$$

This completes the proof. □

This result gives that if n is large enough, then we can obtain the convergence of $n-$step GD based FL.

Theorem 3. *The convergence rate for the Algorithm 2, when n-step GD is used and $\omega = \frac{2}{\lambda_{G^*}+L_{G^*}}$ can be estimated as follows:*

$$\left\|E_{k+1}^H\right\|^2 + \left\|E_{k+1}^Z\right\|_\omega^2 \leq \rho^2\left(\|E_k^H\|^2 + \|E_k^Z\|_\omega^2\right),$$

where

$$\rho = \sqrt{\left(\frac{r}{r+\lambda_F}\right)^2 + \left(\frac{L_F-\lambda_F}{2r+L_F+\lambda_F}\right)^2} + 2\left(\frac{L_F-\lambda_F}{2r+L_F+\lambda_F}\right)^n. \tag{65}$$

Remark 3. *In case $r = 0$, we apply n sufficiently large, i.e., the local solve is done accurate enough, then the overall convergence rate reduces to*

$$\rho^2 \approx \left(\frac{L_F-\lambda_F}{L_F+\lambda_F}\right)^2. \tag{66}$$

We note that if r is chosen to be large, then, n can be made to be small enough so that the overall convergence rate can reduce to that of GS. However, the larger r leads to the regime where the GS rate can be larger than one, depending on L_F and λ_F, thus the convergence can not be made. Furthermore, we can show that if n satisfies the following inequality:

$$n > \frac{\log(\kappa(G)+1)}{\log[(\kappa(G)+1)/(\kappa(G)-1)]}. \tag{67}$$

then the convergence can be obtained.

Corollary 1. *We assume that $r = 0$ and $\omega = \frac{r+\lambda_F}{n}$. Then, we have that*

$$\rho = \left(\frac{n-1}{n}\right) + 2\left(\frac{L_F - \lambda_F}{L_F + \lambda_F}\right)^n. \tag{68}$$

Remark 4. *Under the assumption that ρ behaves like $(n-1)/n$, i.e., $\kappa(F)$ is not too large, if we choose $n = \sqrt{\kappa(F)}$, then we arrive at the argument in [12]. The discussion made for the stochastic case is solely for the dominant rate among two rates.*

References

[1] Tongan Cai, Haomiao Ni, Mingli Yu, Xiaolei Huang, Kelvin Wonḡ, John Volpi, James Z Wang, and Stephen TC Wong. Deepstroke: An efficient stroke screening framework for emergency room s with multimodal adversarial deep learning. *Medical Image Analysis, arXiv preprint arXiv:2109.12065*, 2022, under revision for 2nd-round review.

[2] Yukun Chen, Zhuomin Zhang, Chenyan Wu, Dolzodm̃aa Davaasuren, Jeffery A Goldstein, Alison D Gernand, and James Z Wang. Ai-plax: Ai-based placental assessment and examination using photos. *Computerized Medical Imaging and Graphics*, 84:101744, 2020.

[3] Mohamed Amine Ferrag, Othmane Friha, Leandros Maglaras, Helge Janicke, and Lei Shu. Federated deep learning for cyber security in the internet of things: Concepts, applications, and experimental analysis. *IEEE Access*, 9:138509–138542, 2021.

[4] Mohammad Ghomi. The problem of optimal smoothing for convex functions. *Proceedings of the American Mathematical Society*, 130(8):2255–2259, 2002.

[5] White House. Consumer data privacy in a networked world: A framework for protecting privacy and promoting innovation in the global digital economy. *White House, Washington, DC*, pages 1–62, 2012.

[6] Yong Ho Hwang. Iot security & privacy: threats and challenges. In *Proceedings of the 1st ACM workshop on IoT privacy, trust, and security*, pages 1–1, 2015.

[7] Sham Kakade, Shai Shalev-Shwartz, Ambuj Tewari, et al. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09. pdf*, 2(1):35, 2009.

[8] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

[9] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[10] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[11] Yu Luo, Jianbo Ye, Reginald B Adams, Jia Li, M̃ichelle G Newman, and James Z Wang. Arbee: Towards automated recognition of bodily expression of emotion i n the wild. *International journal of computer vision*, 128(1):1–25, 2020.

[12] Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. ProxSkip: Yes! Local Gradient Steps Provably Lead to Communication Acceleration! Finally! *arXiv preprint arXiv:2202.09357*, 2022.

[13] Mícheál O'Searcoid. *Metric spaces*. Springer Science & Business Media, 2006.

[14] Wei Shi, Qing Ling, Kun Yuan, Gang Wu, and Wotao Yin. On the linear convergence of the ADMM in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

[15] James Z. Wang, Norman Badler, Nadia Berthouze, Rick O. Gilmore, Kerri L. Johnson, Agata Lapedriza, Xin Lu, and Nikolaus Troje. Bodily expressed emotion understanding research: A multidisciplinary perspective. In *Proceedings of the First International Workshop on Bodily Expressed Emotion Understanding, in conjunction with the European Computer Vision Conference*, pages 1–14, 2020.

3.3 Scaffold

Using our notation, we now introduce the ProxSkip algorithm. See algorithm 3.

Algorithm 3 ProxSkip

Given a stepsize $\gamma > 0$, initial iterate $Z_0 = X_0 = (x_0, \dots, x_0) \in \mathbb{R}^{dn}$, h_0, number of iterations $T \geq 1$, we perform the following:

for $t = 0, 1, 2, \cdots, T-1$ **do**
  $X_{t+1} = Z_t - \gamma(\nabla F(Z_t) - h_t)$
  Flip a coin θ_t, $P(\theta_t = 1) = p$
  if $\theta_t = 1$ **then**
    $Z_{t+1} = \mathrm{prox}_{\frac{\gamma}{p}\psi}\left(X_{t+1} - \frac{\gamma}{p}h_t\right)$
  else
    $Z_{t+1} = X_{t+1}$
  end if
  $h_{t+1} = h_t + \frac{p}{\gamma}(Z_{t+1} - X_{t+1})$
end for

A deterministic version of ProxSkip is given in algorithm 4.

Algorithm 4 ProxSkip Deterministic (SCAFFOLD)

Given a stepsize $\gamma > 0$, initial iterate $Z_0 = X_0 = (x_0, \ldots, x_0) \in \mathbb{R}^{dn}$, number of iterations $T \geq 1$, we perform the following:

for $t = 0, 1, 2, \cdots, T-1$ **do**
  $X_t = Z_t$
  for $k = 0, 1, \ldots, N-1$ **do**
    $X_{t+\frac{k+1}{N}} = X_{t+\frac{k}{N}} - \gamma(\nabla F(X_{t+\frac{k}{N}}) - H_t)$
  end for
  $Z_{t+1} = \mathrm{prox}_{N\gamma\psi}(X_{t+1} - N\gamma H_t)$
  $H_{t+1} = H_t + \frac{1}{N\gamma}(Z_{t+1} - X_{t+1})$
end for

4 Federated Learning for F being a quadratic functional; Linear Case

We restrict our discussion of federated learning algorithm for the linear case only.

4.1 Problem Description

We want to solve the following problem:

$$\begin{pmatrix} A & 0 & -I \\ 0 & 0 & K^T \\ -I & K & 0 \end{pmatrix} \begin{pmatrix} X_* \\ z_* \\ H_* \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ 0 \end{pmatrix}, \tag{69}$$

where

$$K = \mathbf{1} \otimes I. \tag{70}$$

We introduce some notation:

$$\mathcal{A} = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} -I & K \end{pmatrix}, \quad f = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}, \quad \text{and} \quad U = \begin{pmatrix} X \\ z \end{pmatrix}. \tag{71}$$

This problem can be written as follows:

$$\mathcal{A}U_* + \mathcal{B}^T H_* = f, \tag{72a}$$

$$\mathcal{B}U_* = 0. \tag{72b}$$

The Augmented Lagrangian Uzawa is based on the addition of the penalty term:

$$(\mathcal{A} + r\mathcal{B}^T\mathcal{B})U_* + \mathcal{B}^T H_* = f \tag{73a}$$

$$\mathcal{B}U_* = 0. \tag{73b}$$

We shall denote $\mathcal{A}_r = \mathcal{A} + r\mathcal{B}^T\mathcal{B}$ and $A_r = A + rI$.

We note that in a full matrix notation, it reads as follows:

$$\begin{pmatrix} A_r & -rK & -I \\ -rK^T & rK^TK & K^T \\ -I & K & 0 \end{pmatrix} \begin{pmatrix} X_* \\ z_* \\ H_* \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ g \end{pmatrix}, \tag{74}$$

4.2 Augmented Lagrangian Uzawa method and its convergence

By removing U, we get the following system for H:

$$\mathcal{S}_r H_* = \mathcal{B}\mathcal{A}_r^{-1}\mathcal{B}^T H_* = \mathcal{B}\mathcal{A}_r^{-1} f. \tag{75}$$

Therefore, the system can be given as follows:

$$\begin{aligned} (\mathcal{A} + r\mathcal{B}^T\mathcal{B})U_* + \mathcal{B}^T H_* &= f && \text{(76a)} \\ \mathcal{S}_r H_* &= \mathcal{B}\mathcal{A}_r^{-1} f. && \text{(76b)} \end{aligned}$$

We can apply the Richardson method to update H, whose parameter will be denoted by ω. Thus, we arrive at the following Augmented Lagrangian Uzawa method:

$$\begin{aligned} (\mathcal{A} + r\mathcal{B}^T\mathcal{B})U_{k+1} &= f - \mathcal{B}^T H_k \\ H_{k+1} &= H_k + \omega(\mathcal{B}\mathcal{A}_r^{-1} f - \mathcal{S}_r H_k) \\ &= H_k + \omega\mathcal{B}U_{k+1}. \end{aligned}$$

This is summarized as an algorithm.

Algorithm 5 Augmented Lagrangian Uzawa

---

for $k = 0, 1, 2, \cdots$ **do**

Update of U_{k+1}:

$$(\mathcal{A} + r\mathcal{B}^T\mathcal{B})U_{k+1} = f - \mathcal{B}^T H_k \tag{77}$$

Update of H_{k+1}:

$$H_{k+1} = H_k + \omega\mathcal{B}U_{k+1}. \tag{78}$$

end for

Theorem 4. *Let $\sigma(A) \in [\lambda_F, L_F]$. Then, the algorithm 5 converges with the convergence rate given as follows:*

1. *For $0 < \omega < 2/\rho(\mathcal{S}_r)$, we have the following convergence:*

$$\|H_* - H_k\| \le \rho(I - \omega\mathcal{S}_r)^k \|H_* - H_0\|. \tag{79}$$

Furthermore, we have that

$$\|X_* - X_k\|_A = |U_* - U_k|_{\mathcal{A}} \le \sqrt{1/r}\,\rho(I - \omega\mathcal{S}_r)^k \|H_* - H_0\|. \tag{80}$$

2. *For $\omega = r$, we have the convergence rate given as follows:*

$$\|H_* - H_k\| \le \left(\frac{1}{1 + r/L_F}\right)^k \|H_* - H_0\| \tag{81}$$

and

$$\|X_* - X_k\|_A = |U_* - U_k|_{\mathcal{A}} \le \sqrt{1/r}\left(\frac{1}{1 + r/L_F}\right)^k \|H_* - H_0\|, \tag{82}$$

where $\mu_0 = 1/L_F$ is the smallest eigenvalue of $\mathcal{B}\mathcal{A}^\dagger\mathcal{B}^T$.

3. *For* $\omega = \frac{2}{\frac{\lambda_F}{1+\lambda_F r} + \frac{L_F}{1+L_F r}}$, *we have*

$$\|H_* - H_k\| \leq \left(\frac{L_F - \lambda_F}{L_F + \lambda_F + 2L_F\lambda_F r}\right)^k \|H_* - H_0\|. \tag{83}$$

Furthermore, we have that

$$\|X_* - X_k\|_A = |U_* - U_k|_{\mathcal{A}} \leq \sqrt{1/r}\left(\frac{L_F - \lambda_F}{L_F + \lambda_F + 2L_F\lambda_F r}\right)^k \|H_* - H_0\|. \tag{84}$$

Proof. The convergence of the Augmented Lagrangian Uzawa relies on the spectrum of the Schur complement operator. This leads to the choice of paramters ω. The Schur compliment operator is given as follows:

$$\mathcal{S}_r = \mathcal{B}(\mathcal{A} + r\mathcal{B}^T\mathcal{B})^{-1}\mathcal{B}^T. \tag{85}$$

We note that

$$\mathrm{Null}(\mathcal{A}) \cap \mathrm{Null}(\mathcal{B}) = \{0\}. \tag{86}$$

Further, we can show that $\mathcal{B}\mathcal{A}^\dagger\mathcal{B}^T$ is symmetric positive definite. We note that

$$\mathcal{A}^\dagger = \begin{pmatrix} A^{-1} & 0 \\ 0 & 0 \end{pmatrix} \tag{87}$$

and thus

$$\mathcal{B}\mathcal{A}^\dagger\mathcal{B}^T = \begin{pmatrix} -I & K \end{pmatrix}\begin{pmatrix} A^{-1} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} -I \\ K^T \end{pmatrix} = \begin{pmatrix} -I & K \end{pmatrix}\begin{pmatrix} -(\nabla^2 F)^{-1} \\ 0 \end{pmatrix} = \nabla^2 F^{-1}. \tag{88}$$

This means, that

$$\frac{1}{L_F} I \leq \mathcal{B}\mathcal{A}^\dagger\mathcal{B}^T \leq \frac{1}{\lambda_F} I. \tag{89}$$

By applying the Sherman-Morrison-Woodbury formula, we have

$$\mathcal{S}_r^{-1} = rI + (\mathcal{B}\mathcal{A}^\dagger\mathcal{B}^T)^{-1} = rI + A = A_r. \tag{90}$$

Thus, the spectrum of $\mathcal{S}_r$ is given by

$$\sigma(\mathcal{S}_r) = \left\{\frac{\eta}{1 + r\eta} : \eta \in \sigma(\mathcal{B}\mathcal{A}^\dagger\mathcal{B}^T)\right\}. \tag{91}$$

Thus, the spectral radius of $\mathcal{S}_r$ has the upper bound i.e., $\rho(\mathcal{S}_r) < 1/r$ for any $r > 0$. Therefore, since the convergence of Richardson method will be guaranteed if $0 < \omega < 2/\rho(\mathcal{S}_r) = 2r$, a simple choice could be $\omega = r$ for the convergence. While it is not the optimal choice, the convergence can be shown as follows: The Augmented Lagrangian Uzawa can be shown to behave as follows:

$$\|H_* - H_k\| \leq \left(\frac{1}{1 + r/L_F}\right)^k \|H_* - H_0\| \tag{92}$$

and

$$|U_* - U_k|_{\mathcal{A}} \leq \sqrt{1/r}\left(\frac{1}{1+r/L_F}\right)^k \|H_* - H_0\|, \tag{93}$$

where $\mu_0 = 1/L_F$ is the smallest eigenvalue of $\mathcal{B}\mathcal{A}^\dagger\mathcal{B}^T$. Now, we shall consider more detailed discussion on optimal choice of ω. We note that

$$\sigma(\mathcal{S}_r) \in \left[\frac{1}{\frac{1}{\lambda_F}+r}, \frac{1}{\frac{1}{L_F}+r}\right] \tag{94}$$

Then, the optimum convergence rate is given as follows for $\omega = \frac{2}{\lambda_{min}(\mathcal{S}_r)+\lambda_{max}(\mathcal{S}_r)}$:

$$\frac{\kappa(\mathcal{S}_r)-1}{\kappa(\mathcal{S}_r)+1} = \frac{L-\lambda}{L+\lambda+2L\lambda r}. \tag{95}$$

This completes the proof. □

To clarify the discussion, we assume that $N_c = 1$ and the original 3×3 system can be written as

$$\begin{pmatrix} \nabla^2 G & B^T \\ B & 0 \end{pmatrix}\begin{pmatrix} U_* \\ H_* \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}, \qquad B^T = \begin{pmatrix} -I \\ K^T \end{pmatrix}. \tag{96}$$

Lemma 12. *The matrix $\nabla^2 G(X,z)$ is symmetric positive definite and the spectrum is given by*

$$\sigma(H(G)) \subset \left\{\min\left\{r+\lambda_F, \frac{r\lambda_F}{r+\lambda_F}\right\}, \max\left\{r+L_F, \frac{rL_F}{r+L_F}\right\}\right\}. \tag{97}$$

Proof. We begin with the spectrally equivalent matrix to $H(G)$, given as follows:

$$\begin{pmatrix} A_r & 0 \\ 0 & rK^TK - rK^TA_r^{-1}rK \end{pmatrix} = \begin{pmatrix} A_r & 0 \\ 0 & rK^T(I - rA_r^{-1})K \end{pmatrix}$$

Due to the λ_F-strong convexity and L_F-smoothness of F, it is easy to see

$$\begin{aligned} &\sigma(A_r) \subset \{r+\lambda_F, L+\lambda_F\} \\ &\sigma\left(I - rA_r^{-1}\right) \subset \left\{1 - \frac{r}{r+\lambda_F}, 1 - \frac{r}{r+L_F}\right\} = \left\{\frac{\lambda_F}{r+\lambda_F}, \frac{L_F}{r+L_F}\right\} \end{aligned}$$

Thus, we have that

$$\sigma\left(rI - r^2A_r^{-1}\right) \subset \left\{\frac{r\lambda_F}{r+\lambda_F}, \frac{rL_F}{r+L_F}\right\}$$

and

$$\sigma(H(G)) \subset \left\{\min\left\{r+\lambda_F, \frac{r\lambda_F}{r+\lambda_F}\right\}, \max\left\{r+L_F, \frac{rL_F}{r+L_F}\right\}\right\}. \tag{98}$$

This completes the proof. □

4.3 Convergence of Exact and Inexact Block Gauss-Seidel Method for the U system

In this section, we discuss the solution by Gauss-Seidel for the following system:

$$\mathcal{A}_r \begin{pmatrix} X_* \\ z_* \end{pmatrix} = \begin{pmatrix} A + rI & -rK \\ -rK^T & rK^T K \end{pmatrix} \begin{pmatrix} X_* \\ z_* \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}. \tag{99}$$

We shall consider both exact block Gauss-Seidel method and inexact block Gauss-Seidel method. To handle these at the same time, it would be good to introduce a general framework. We denote R by the modification of D^{-1}. Then, the modified block Gauss-Seidel method is given as follows:

$$U_{k+1} = U_k + (R^{-1} + L)^{-1}(f - \mathcal{A}_r U_k), \quad k = 1, 2, \cdots . \tag{100}$$

Note that

$$D = \begin{pmatrix} A + rI & 0 \\ 0 & rNI \end{pmatrix} \tag{101a}$$

$$L = \begin{pmatrix} 0 & 0 \\ -rK^T & 0 \end{pmatrix}. \tag{101b}$$

We note that the modified block Gauss-Seidel method converges if

$$\overline{R} = R^T + R - R^T DR > 0. \tag{102}$$

Furthermore, the convergence rate can be obtained as follows:

$$\|I - (R^{-1} + L)^{-1}\mathcal{A}_r\|_{\mathcal{A}_r}^2 = 1 - \frac{1}{1 + c_0(r)}, \tag{103}$$

where

$$c_0(r) = \sup_{\|v\|_{\mathcal{A}_r}=1} \langle \overline{R}^{-1} R^T (D + U - R^{-1})v, R^T (D + U - R^{-1})v \rangle. \tag{104}$$

Remark 5. *The above framework can handle the case when $n-$step Gradient descent method is used to handle $A_r = A + rI$ block. Note that in such an occasion, the $n-$step Gradient descent requires the initial guess and thus, we can set it as X_k. The question arises here if we can use the total of $m-$step iteration to define U_{k+1}. Namely, we can apply the total of $n-$step inner iteration while we use $m-$step outer iteration.*

In the next two sections, we shall consider the case $R = D^{-1}$ and R is an approximate block solve for $A_r = A + rI$. We begin with the first case.

5 Appendix

5.1 Study of Convergence for the algorithm for the Total System

In this section, we discuss the convergence of the iterative method based on inexact Block Gauss-Seidel for U block and Richardson for H block. The Algorithm can be written as given in the Algorithm 6. We note that the action of the operator $G_{n,r}$ depends

Algorithm 6 Federated Learning formulation of FL

Given H_0 such that $K^T H_0 = 0$, updates are obtained as follows:

for $k = 0, 1, 2, \cdots, K-1$ **do**

X_{k+1} update: (with $X_k = Kz_k$),

$$X_{k+1} = G_{n,r}(X_k; H_k + rKz_k), \tag{105}$$

z_{t+1} update:

$$K^T H_k + rK^T(Kz_{k+1} - X_{k+1}) = 0, \tag{106}$$

Update the Lagrange multiplier:

$$H_{k+1} = H_k + \omega(Kz_{k+1} - X_{k+1}). \tag{107}$$

end for

on X_k. The case when $G_{n,r}$ is the standard $n-$step GD, the algorithm can be written in a very standard way as given in Algorithm 7.

Algorithm 7 The case that $G_{n,r}$ is the Standard GD

Given H_k and z_k with $K^T H_k = 0$, update for X_{k+1} is obtained as follows:

for $\ell = 1, 2, \cdots, n$ **do**

X_{k+1} update: (with $X_k = Kz_k$ and $b_{H,rKz} = H_k + rKz_k$),

$$X_{k+\ell/n} = X_{k+(\ell-1)/n} + \gamma(b_{H,rKz} - A_r X_{k+(\ell-1)/n}) \tag{108}$$

end for

In this section, we shall discuss the convergence of Algorithm 1. We first begin our discussion for the standard GD case. The standard $n-$step GD method to solve the following system for X:

$$A_r X = H_k + rKz_k. \tag{109}$$

will be given as follows: with $X_k = Kz_k$,

$$
\begin{aligned}
X_{k+\frac{1}{n}} &= X_k + \gamma(H_k + rKz_k - A_r X_k) \\
X_{k+\frac{2}{n}} &= X_{k+\frac{1}{n}} + \gamma(H_k + rKz_k - A_r(X_{k+\frac{1}{n}})) \\
&\vdots \\
X_{k+\frac{n-1}{n}} &= X_{k+\frac{n-2}{n}} + \gamma(H_k + rKz_k - A_r(X_{k+\frac{n-2}{n}})) \\
X_{k+\frac{n}{n}} &= X_{k+\frac{n-1}{n}} + \gamma(H_k + rKz_k - A_r(X_{k+\frac{n-1}{n}})).
\end{aligned} \tag{110a}
$$

The Algorithm 7 satisfies the following identity for $n \to \infty$, i.e., Y_* such that

$$
A(Y_*) + rY_* = H_k + rKz_k. \tag{111}
$$

In passing to the next section, we shall make a simple remark. In case $H_k = H_*$ and $z_k = z_*$, we see that both schemes lead to $X_{k+1} = X_*$ in a single iteration. Thus, we observe that for all $r \geq 0$,

$$
X_* = G_{n,r}(X_*; H_* + rKz_*) = A_r^{-1}(H_* + rKz_*). \tag{112}
$$

5.1.1 General framework of convergence analysis for Algorithm 6

In this section, we shall discuss the basic framework to analyze the convergence of the Algorithm 6. We shall use the standard notation that for all $k \geq 0$ to discuss the convergence:

$$
\begin{aligned}
E_k^X &= X_* - X_k \\
E_k^Z &= Kz_* - Kz_k \\
E_k^H &= H_* - H_k.
\end{aligned}
$$

The following is the main result in this section.

Theorem 5. *The Algorithm 6 with the inexact or exact solve, the Algorithm 7, produces iterate (X_k, z_k, H_k), for which the following error bound holds true:*

$$
\left\|E_{k+1}^H\right\|^2 + \omega^2 \left\|E_{k+1}^Z\right\|^2 = \left\|E_k^H - \omega(A_r^{-1}(H_* + rKz_*) - G_{n,r}(Kz_k; H_k + rKz_k))\right\|^2.
$$

Proof. The Algorithm 6 leads to iterates, given as follows:

$$
\begin{aligned}
X_{k+1} &= G_{n,r}(Kz_k; H_k + rKz_k) \\
Kz_{k+1} &= K(rK^TK)^{-1}(rK^TX_{k+1} - K^TH_k) \\
H_{k+1} &= H_k + \omega(-X_{k+1} + Kz_{k+1}),
\end{aligned}
$$

where $G_{n,r}$ is an approximate of A_r^{-1}. We first notice that if $K^TH_0 = 0$, then $K^TH_k = 0$ and also $K^TH_* = 0$. This is due to the proximal operator $P_Z = K(K^TK)^{-1}K$. Therefore, we have

$$
\begin{aligned}
X_{k+1} &= G_{n,r}(Kz_k; H_k + rKz_k) \\
Kz_{k+1} &= K(rK^TK)^{-1}(rK^TG_{n,r}(Kz_k; H_k + rKz_k)) = P_Z[G_{n,r}(Kz_k; H_k + rKz_k)] \\
H_{k+1} &= H_k + \omega(-X_{k+1} + Kz_{k+1})
\end{aligned}
$$

On the other hand, we have that

$$\begin{aligned} X_* &= A_r^{-1}(H_* + rKz_*) \\ Kz_* &= P_Z[A_r^{-1}(H_* + rKz_*)] \\ H_* &= H_* + \omega(-X_* + Kz_*). \end{aligned}$$

Therefore, we have the following error equation:

$$\begin{aligned} E_{k+1}^X &= A_r^{-1}(H_* + rKz_*) - G_{n,r}(Kz_k; H_k + rKz_k) \\ E_{k+1}^Z &= P_Z[A_r^{-1}(H_* + rKz_*) - G_{n,r}(Kz_k; H_k + rKz_k)] \\ E_{k+1}^H &= H_* - H_k + \omega(-X_* + X_{k+1} + Kz_* - Kz_{k+1}) \end{aligned}$$

Rearranging the error in H variable, we have

$$E_{k+1}^H - \omega E_{k+1}^Z = E_k^H - \omega E_{k+1}^X = E_k^H - \omega\left(A_r^{-1}(H_* + rKz_*) - G_{n,r}(Kz_k; H_k + rKz_k)\right). \quad (113)$$

Taking the squared norm on both sides of the equation (113), and using the orthogonality between E_i^H and E_j^Z for all i, j, we have

$$\|E_{k+1}^H\|^2 + \omega^2\|E_{k+1}^Z\|^2 = \|E_k^H - \omega(A_r^{-1}(H_* + rKz_*) - G_{n,r}(Kz_k; H_k + rKz_k))\|^2.$$

This completes the proof.

□

5.2 Convergence analysis of FL Algorithm 6 with GS

In this section, we shall establish that the following holds:

Theorem 6. *Given*

$$\omega \leq r + \lambda_F, \quad (114)$$

the Algorithm 6 with GS, produces iterate (X_k, z_k, H_k), for which the following convergence rate is valid: for n sufficiently large, we have

$$\left\|E_{k+1}^H\right\|^2 + \left\|E_{k+1}^Z\right\|_\omega^2 \leq \left(\left(1 - \frac{1}{\kappa(A_r)}\right)^2 + \left(\frac{r}{r + \lambda_F}\right)^2\right)\left(\left\|E_k^H\right\|^2 + \left\|E_k^Z\right\|_\omega^2\right).$$

Proof. We observe that

$$\begin{aligned} \left\|E_{k+1}^H\right\|^2 + \left\|E_{k+1}^Z\right\|_\omega^2 &\leq \left\|E_k^H - \omega(A_r^{-1}(H_* + rKz_*) - A_r^{-1}(H_k + rKz_k))\right\|^2 \\ &= \left\|E_k^H - \omega(A_r^{-1}(H_*) + A_r^{-1}(rKz_*) - A_r^{-1}(H_k) - A_r^{-1}(rKz_k)\right\|^2 \\ &= \left\|(I - \omega A_r^{-1})(E_k^H) - \omega r A_r^{-1}(E_k^Z)\ \right\|^2 \\ &\leq \|(I - \omega A_r^{-1})E_k^H\|^2 + \|\omega r A_r^{-1}E_k^Z\|^2 - 2\langle(I - \omega A_r^{-1})E_k^H, \omega r A_r^{-1}E_k^Z\rangle. \end{aligned}$$

We therefore, note that

$$\begin{aligned}
\|(I-\omega A_r^{-1})E_k^H\|^2 &\le \left(1-\frac{1}{\kappa(A_r)}\right)^2\|E_k^H\|^2\\
\|\omega r A_r^{-1}E_k^Z\|^2 &\le \left(\frac{r}{r+\lambda_F}\right)^2\|E_k^Z\|_\omega^2\\
-2\langle (I-\omega A_r^{-1})E_k^H, \omega r A_r^{-1}E_k^Z\rangle &\le \left(\frac{r}{r+\lambda_F}\right)^2\|E_k^H\|^2+\left(1-\frac{1}{\kappa(A_r)}\right)^2\|E_k^Z\|_\omega^2.
\end{aligned}$$

This completes the proof for GS case. □

5.3 Convergence analysis of FL Algorithm 6 with GS and variable ω

In this section, we shall establish that the following holds:

Theorem 7. *Given*

$$\omega = A_r \tag{115}$$

the Algorithm 6 with GS, produces iterate (X_k, z_k, H_k), for which the following convergence rate is valid: for n sufficiently large, we have

$$\left\|E_{k+1}^H\right\|^2+\left\|E_{k+1}^Z\right\|_\omega^2 \le \left(\frac{r}{r+\lambda_F}\right)^2\left(\left\|E_k^H\right\|^2+\left\|E_k^Z\right\|_\omega^2\right).$$

Proof. We observe that

$$\begin{aligned}
\left\|E_{k+1}^H\right\|^2+\left\|E_{k+1}^Z\right\|_\omega^2 &\le \left\|E_k^H-\omega(A_r^{-1}(H_*+rKz_*)-A_r^{-1}(H_k+rKz_k))\right\|^2\\
&= \left\|E_k^H-\omega(A_r^{-1}(H_*)+A_r^{-1}(rKz_*)-A_r^{-1}(H_k)-A_r^{-1}(rKz_k)\right\|^2\\
&= \left\|(I-\omega A_r^{-1})(E_k^H)-\omega r A_r^{-1}(E_k^Z)\ \right\|^2\\
&\le \|(I-\omega A_r^{-1})E_k^H\|^2+\|\omega r A_r^{-1}E_k^Z\|^2-2\langle (I-\omega A_r^{-1})E_k^H, \omega r A_r^{-1}E_k^Z\rangle\\
&= \|(I-A_rA_r^{-1})E_k^H\|^2+\|A_r r A_r^{-1}E_k^Z\|^2-2\langle (I-A_rA_r^{-1})E_k^H, A_r r A_r^{-1}E_k^Z\rangle\\
&= r^2\|E_k^Z\|^2 = r^2\|A_r^{-1}A_rE_k^Z\|^2=\left(\frac{r}{r+\lambda_F}\right)^2\|E_k^Z\|_\omega^2.
\end{aligned}$$

This completes the proof for GS case. □

5.4 Convergence analysis of FL Algorithm 6 with GD

In this section, we shall establish that the following holds:

Theorem 8. *Given*

$$\omega \le r+\lambda_F, \tag{116}$$

the Algorithm 6 with GD, produces iterate (X_k, z_k, H_k), for which the following convergence rate is valid: for n sufficiently large, we have

Proof. We observe that

$$
\begin{aligned}
\left\|E_{k+1}^H\right\|^2 + \left\|E_{k+1}^Z\right\|_\omega^2 &\leq \left\|E_k^H - \omega(A_r^{-1}(H_* + rKz_*) - G_{n,r}(H_k + rKz_k))\right\|^2 \\
&= \left\|E_k^H - \omega(A_r^{-1}(H_* + rKz_*) - A_r^{-1}(H_k + rKz_*) + A_r^{-1}(H_k + rKz_*) - G_{n,r}(H_k + rKz_k))\right\|^2 \\
&= \left\|(I - \omega A_r^{-1})(E_k^H) - \omega(A_r^{-1}(H_k + rKz_*) - G_{n,r}(H_k + rKz_k))\ \right\|^2 \\
&= \|E_1 - E_2\|^2
\end{aligned}
$$

We therefore, note that

$$
\begin{aligned}
&\|(I - \omega A_r^{-1})E_k^H\|^2 \leq \left(1 - \frac{1}{\kappa(A_r)}\right)^2 \|E_k^H\|^2 \\
&\|\omega(A_r^{-1}(H_k + rKz_*) - G_{n,r}(H_k + rKz_k))\|^2 \\
&\qquad \leq \|\omega(A_r^{-1}(H_k + rKz_*) - A_r^{-1}(H_k + rKz_k) + A_r^{-1}(H_k + rKz_*) - G_{n,r}(H_k + rKz_k))\|^2 \\
&\qquad \leq 2\|\omega(A_r^{-1}(H_k + rKz_*) - A_r^{-1}(H_k + rKz_k)\|^2 + 2\|\omega(A_r^{-1}(H_k + rKz_k) - G_{n,r}(H_k + rKz_k))\|^2 \\
&\qquad \leq 2\omega^2 \left(\frac{r}{r + \lambda_F}\right)^2 \|Kz_* - Kz_k\|^2 + 2\omega^2\delta^{2n}\|Y_* - Kz_k\|^2 \\
&\qquad \leq 2\omega^2 \left(\frac{r}{r + \lambda_F}\right)^2 \|Kz_* - Kz_k\|^2 + 2\frac{\omega^2}{(r + \lambda_F)^2}\delta^{2n}\left(\|E_k^H\|^2 + \|E_k^Z\|_\omega^2\right)
\end{aligned}
$$

This completes the proof for GS case. $\square$

We note that these two main theorems produce identity for the convergence estimate. The key is now to obtain the estimate of the right hand side with an appropriate choice of ω and γ.

5.4.1 Convergence analysis of Algorithm 6 with GD given in Algorithm 7

Throughout this section, we shall set

$$\delta = \frac{\kappa(A_r) - 1}{\kappa(A_r)}. \tag{117}$$

In this section, we shall establish that the following holds:

Theorem 9. *Given $\gamma = \frac{1}{r+L_F}$ and*

$$\omega = \frac{2}{\frac{1-\delta^n}{r+\lambda_F} + \frac{1}{r+L_F}}, \tag{118}$$

the Algorithm 6 with GD given in the Algorithm 7, produces iterate (X_k, z_k, H_k), for which the following convergence rate is valid: for n sufficiently large, we have

$$\left\|E_{k+1}^H\right\|^2 + \left\|E_{k+1}^Z\right\|_\omega^2 \le \left(\left(\delta^n + (1-\delta^n)\frac{r}{r+\lambda_F}\right)^2 + \left(\frac{\kappa(S_r)-1}{\kappa(S_r)+1}\right)^2\right)\left(\left\|E_k^H\right\|^2 + \left\|E_k^Z\right\|_\omega^2\right),$$

where $\kappa(S_r) = \frac{(r+L_F)}{(r+\lambda_F)}(1-\delta^n)$. Furthermore, we know that E_k^Z is orthogonal to E_k^H. If we can show that these can be decomposed into eigenvectors of A or A-orthogonal to each other, then we have that under the assumption that n is chosen so that

$$1 - \frac{1}{\sqrt{\kappa}} = \delta^n. \tag{119}$$

We have

$$\left\|E_{k+1}^H\right\|^2 + \left\|E_{k+1}^Z\right\|_\omega^2 \le \left(1 - \frac{1}{\sqrt{\kappa}}\right)^2 \left(\left\|E_k^H\right\|^2 + \left\|E_k^Z\right\|_\omega^2\right),$$

We choose $1 - \delta^n = \frac{1}{\sqrt{\kappa}}$, then we have

$$1 - \frac{1}{\sqrt{\kappa}} = \delta^n. \tag{120}$$

Furthermore, we have

$$\kappa(S_r) = \frac{L_F}{\lambda_F}(1-\delta^n) = \sqrt{\kappa}. \tag{121}$$

Remark 6. *For $r = 0$, we have that*

$$\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}} = \delta^n. \tag{122}$$

Taking log, we have

$$n \log\delta = \log\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}}\right). \tag{123}$$

Therefore, we see that

$$n = \frac{\log\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}}\right)}{\log\delta}. \tag{124}$$

Proof. We first recall that the following identity holds: with $\mathcal{H}_\gamma = I - \gamma A_r$,

$$\begin{aligned} G_{n,r}(X; H_k + rKz_k) &= (I - \gamma A_r)^n X + (I - (I - \gamma A_r)^n) A_r^{-1}(H_k + rKz_k) \\ &= \mathcal{H}_\gamma^n X + (I - \mathcal{H}_\gamma^n) A_r^{-1}(H_k + rKz_k). \end{aligned}$$

With an observation that

$$A_r^{-1}(H_* + rKz_*) = G_{n,r}(Kz_*; H_* + rKz_*). \tag{125}$$

According to Theorem, we shall need to estimate the following:

$$\begin{aligned} &\left\| E_k^H - \omega(G_{n,r}(Kz_*; H_* + rKz_*) - G_{n,r}(Kz_k; H_k + rKz_k)) \right\|^2 \\ &\quad = \left\| E_k^H - \omega(G_{n,r}(Kz_*; H_* + rKz_*) - G_{n,r}(Kz_k; H_k + rKz_k)) \right\|^2 \\ &\quad = \Big\| E_k^H - \omega(G_{n,r}(Kz_*; H_* + rKz_*) - G_{n,r}(Kz_*; H_k + rKz_*)) \\ &\qquad - \omega(G_{n,r}(Kz_*; H_k + rKz_*) - G_{n,r}(Kz_k; H_k + rKz_k)) \Big\|^2 . \end{aligned}$$

We shall now investigate the following:

$$\begin{aligned} E_1 &= E_k^H - \omega(G_{n,r}(Kz_*; H_* + rKz_*) - G_{n,r}(Kz_*; H_k + rKz_*)) \\ &= E_k^H - \omega(I - \mathcal{H}_\gamma^n) A_r^{-1}(H_* - H_k) \\ &= (I - \omega(I - \mathcal{H}_\gamma^n) A_r^{-1}) E_k^H \\ E_2 &= G_{n,r}(Kz_*; H_k + rKz_*) - G_{n,r}(Kz_k; H_k + rKz_k) \\ &= \mathcal{H}_\gamma^n K(z_* - z_k) + (I - \mathcal{H}_\gamma^n) A_r^{-1} r(Kz_* - Kz_k) \\ &= \left(\mathcal{H}_\gamma^n + r(I - \mathcal{H}_\gamma^n) A_r^{-1} \right) K(z_* - z_k). \end{aligned}$$

This leads that

$$\begin{aligned} &\left\| E_k^H - \omega(G_{n,r}(Kz_*; H_* + rKz_*) - G_{n,r}(Kz_k; H_k + rKz_k)) \right\|^2 \\ &\quad = \|E_1 - \omega E_2\|^2 \\ &\quad = \|E_1\|^2 - 2\langle E_1, \omega E_2 \rangle + \|E_2\|_\omega^2. \end{aligned}$$

A simple choice of γ would be $\gamma = \frac{1}{r + L_F}$. Then, we see that, in an increasing order:

$$\begin{aligned} \sigma(\mathcal{H}_\gamma^n) &= \{(1 - \gamma(r + L_F)^n, \cdots, (1 - \gamma(r + \lambda_F)^n\} \\ \sigma(I - \mathcal{H}_\gamma^n) &= \{1 - (1 - \gamma(r + \lambda_F))^n, \cdots, 1 - (1 - \gamma(r + L_F))^n\} \\ \rho((I - \mathcal{H}_\gamma^n) A_r^{-1}) &= \max\left\{ \frac{(1 - (1 - \gamma(r + \lambda_F))^n)}{r + \lambda_F}, \frac{1 - (1 - \gamma(r + L_F))^n}{r + L_F} \right\} \\ \rho(\mathcal{H}_\gamma^n + (I - \mathcal{H}_\gamma^n) A_r^{-1} r) &= \max\Big\{ (1 - \gamma(r + \lambda_F))^n + (1 - (1 - \gamma(r + \lambda_F))^n)) \frac{r}{r + \lambda_F}, \\ &\qquad (1 - \gamma(r + L_F))^n + (1 - (1 - \gamma(r + L_F))^n) \frac{r}{r + L_F} \Big\}. \end{aligned}$$

Thus, we have that

$$\begin{aligned}
\sigma(\mathcal{H}_\gamma^n) &= \{(1-\gamma(r+L_F))^n, \cdots, (1-\gamma(r+\lambda_F))^n\} \\
\sigma(I-\mathcal{H}_\gamma^n) &= \{1-(1-\gamma(r+\lambda_F))^n, \cdots, 1-(1-\gamma(r+L_F))^n\} \\
\rho((I-\mathcal{H}_\gamma^n)A_r^{-1}) &= \max\left\{\frac{(1-\delta^n)}{r+\lambda_F}, \frac{1}{r+L_F}\right\} \\
\rho(\mathcal{H}_\gamma^n + (I-\mathcal{H}_\gamma^n)A_r^{-1}r) &= \max\left\{\delta^n + (1-\delta^n)\frac{r}{r+\lambda_F}, \frac{r}{r+L_F}\right\}.
\end{aligned}$$

The easy bound would be for E_2. We note that

$$\|E_2\| \le \max\left\{\delta^n + (1-\delta^n)\frac{r}{r+\lambda_F}, \frac{r}{r+L_F}\right\}\|Kz_* - Kz_k\|. \tag{126}$$

On the other hand, we have that with $S_r = (I-\mathcal{H}_\gamma^n)A_r^{-1}$,

$$\lambda_{min}(S_r) = \min\left\{\frac{(1-\delta^n)}{r+\lambda_F}, \frac{1}{r+L_F}\right\} \quad \text{and} \quad \lambda_{max}(S_r) = \max\left\{\frac{(1-\delta^n)}{r+\lambda_F}, \frac{1}{r+L_F}\right\}.$$

Thus, the choice of ω given as follows:

$$\omega = \frac{2}{\frac{1-\delta^n}{r+\lambda_F} + \frac{1}{r+L_F}}. \tag{127}$$

With this choice, we obtain the convergence rate given as follows:

$$\|E_1\| \le \left(\frac{\kappa(S_r)-1}{\kappa(S_r)+1}\right)\|H_* - H_k\|. \tag{128}$$

We shall now make it clear by choosing n small and n large. First, for $n \gg 1$, we have that

$$\|E_2\| \le \left(\delta^n + (1-\delta^n)\frac{r}{r+\lambda_F}\right)\|Kz_* - Kz_k\| \tag{129}$$

and

$$\kappa(S_r) = \frac{(r+L_F)}{(r+\lambda_F)}(1-\delta^n). \tag{130}$$

On the other hand, if $n = O(1)$, then it is unclear since there are many factors. This completes the proof. □

Remark 7. *Note that $\mathcal{H}_\gamma$ takes the spectral radius for the vector*

$$(I-\gamma A)^n\phi = \left(1-\frac{\lambda}{L}\right)^n\phi = \left(1-\frac{1}{\kappa}\right)^n\phi, \tag{131}$$

namely, for eigenvector that corresponds to the smallest eigenvalue. On the other hand, the spectral radius for ϕ, which is also corresponding to the smallest eigenvalue of A as well, i.e.,

$$I - \omega(I-(I-\gamma A)^n)A^{-1} \tag{132}$$

happens for ϕ as well. In any case, two operators $(I-\gamma A)^n$ and $(I-\omega(I-(I-\gamma A)^n)A^{-1}$ share eigenvectors with A or A^{-1}.

We now consider the GD case. In this case, we shall need to estimate the following:

$$\begin{aligned}
&\left\|E_k^H - \omega(A_r^{-1}(H_* + rKz_*) - G_{n,r}(H_k + rKz_k))\right\|^2 \\
&\quad = \left\|E_k^H - \omega(A_r^{-1}(H_* + rKz_*) - A_r^{-1}(H_k + rKz_*))\right\|^2 \\
&\quad -\omega(A_r^{-1}(H_k + rKz_*) - G_{n,r}(Kz_k; H_k + rKz_k))\right\|^2 .
\end{aligned}$$

We shall now investigate the following:

$$\begin{aligned}
E_1 &= E_k^H - \omega(A_r^{-1}(H_* + rKz_*) - A_r^{-1}(H_k + rKz_*)) \\
&= E_k^H - \omega A_r^{-1}(H_* - H_k) \\
E_2 &= A_r^{-1}(H_k + rKz_*) - G_{n,r}(Kz_k; H_k + rKz_k).
\end{aligned}$$

This leads that

$$\begin{aligned}
&\left\|E_k^H - \omega(A_r^{-1}(H_* + rKz_*) - G_{n,r}(Kz_k; H_k + rKz_k))\right\|^2 \\
&\quad = \|E_1 - \omega E_2\|^2 \\
&\quad = \|E_1\|^2 - 2\langle E_1, \omega E_2\rangle + \|E_2\|_\omega^2.
\end{aligned}$$

We then have the following: for $\omega \leq r + \lambda_F$,

$$\begin{aligned}
\|E_1\|_{1/\omega} &= \|E_k^H - \omega(A_r^{-1}(H_* + rKz_*) - A_r^{-1}(H_k + rKz_*))\|_{1/\omega} \\
&= \|(I - \omega A_r^{-1})E_k^H\|_{1/\omega} \leq \rho(I - \omega A_r^{-1})\|E_k^H\|_{1/\omega} = \left(1 - \frac{1}{\kappa(A_r)}\right)\|E_k^H\|_{1/\omega}.
\end{aligned}$$

where $\rho(I - \omega A_r^{-1})$ is given as follows:

$$\sigma(I - \omega A_r^{-1}) = \left\{1 - \frac{r + \lambda_F}{r + \lambda} : \lambda \in \sigma(A_r^{-1})\right\}. \tag{133}$$

On the other hand, for E_2, we have that

$$\begin{aligned}
\|E_2\|_\omega &= \|\omega(A_r^{-1}(H_k + rKz_*) - G_{n,r}(Kz_k; H_k + rKz_k))\| \\
&\leq \|(A_r^{-1}(H_k + rKz_*) - G_{n,r}^{-1}(H_k + rKz_k))A_r\| \\
&= \|H_k + rKz_* - G_{n,r}^{-1}(H_k + rKz_k)A_r\| \\
&= \frac{r}{\sqrt{r + \lambda_F}}\|E_k^Z\|_{A_r} +
\end{aligned}$$

With this choice, we obtain the convergence rate given as follows:

$$\|E_1\| \leq \left(\frac{\kappa(S_r) - 1}{\kappa(S_r) + 1}\right)\|H_* - H_k\|. \tag{134}$$

We shall now make it clear by choosing n small and n large. First, for $n \gg 1$, we have that

$$\|E_2\| \leq \left(\delta^n + (1 - \delta^n)\frac{r}{r + \lambda_F}\right)\|Kz_* - Kz_k\| \tag{135}$$

and

$$\kappa(S_r) = \frac{(r + L_F)}{(r + \lambda_F)}(1 - \delta^n). \tag{136}$$

On the other hand, if $n = O(1)$, then it is unclear since there are many factors. This completes the proof.

5.4.2 Convergence analysis of Algorithm 6 with GD given in Algorithm ??

Throughout this section, we shall set

$$\delta = \frac{\kappa(A) - 1}{\kappa(A)}. \tag{137}$$

In this section, we shall establish that the following holds:

Theorem 10. *Given $\gamma = \frac{1}{L_F}$ and*

$$\omega = \frac{2}{\frac{1-\delta^n}{\lambda_F} + \frac{1}{L_F}}, \tag{138}$$

the Algorithm 6 with GD given in the Algorithm 7, produces iterate (X_k, z_k, H_k), for which the following convergence rate is valid: for n sufficiently large, we have

$$\left\|E^H_{k+1}\right\|^2 + \left\|E^Z_{k+1}\right\|^2_\omega \le \left(\delta^{2n} + \left(\frac{\kappa(S) - 1}{\kappa(S) + 1}\right)^2\right)\left(\left\|E^H_k\right\|^2 + \left\|E^Z_k\right\|^2_\omega\right),$$

where $\kappa(S) = \frac{L_F}{\lambda_F}(1 - \delta^n)$.

Proof. We first recall that the following identity holds: with $\mathcal{H}_\gamma = I - \gamma A$,

$$\begin{aligned}
G_{n,0}(Kz_k; H_k) &= (I - \gamma A)^n K z_k + (I - (I - \gamma A)^n) A^{-1}(H_k) \\
&= \mathcal{H}^n_\gamma K z_k + (I - \mathcal{H}^n_\gamma) A^{-1}(H_k).
\end{aligned}$$

With an observation that

$$A^{-1}(H_*) = G_{n,0}(Kz_*; H_*). \tag{139}$$

According to Theorem, we shall need to estimate the following:

$$\begin{aligned}
&\left\|E^H_k - \omega(G_{n,0}(Kz_*; H_*) - G_{n,0}(Kz_k; H_k))\right\|^2 \\
&\quad = \left\|E^H_k - \omega(G_{n,0}(Kz_*; H_*) - G_{n,0}(Kz_k; H_k))\right\|^2 \\
&\quad = \left\|E^H_k - \omega(G_{n,0}(Kz_*; H_*) - G_{n,0}(Kz_*; H_k)) - \omega(G_{n,0}(Kz_*; H_k) - G_{n,0}(Kz_k; H_k))\right\|^2.
\end{aligned}$$

We shall now investigate the following:

$$\begin{aligned}
E_1 &= E^H_k - \omega(G_{n,0}(Kz_*; H_*) - G_{n,0}(Kz_*; H_k)) \\
&= E^H_k - \omega(I - \mathcal{H}^n_\gamma) A^{-1}(H_* - H_k) \\
E_2 &= G_{n,0}(Kz_*; H_k) - G_{n,0}(Kz_k; H_k) \\
&= \mathcal{H}^n_\gamma K(z_* - z_k)
\end{aligned}$$

This leads that

$$\begin{aligned}
&\left\|E^H_k - \omega(G_{n,0}(Kz_*; H_*) - G_{n,0}(Kz_k; H_k))\right\|^2 \\
&\quad = \|E_1 - \omega E_2\|^2 \\
&\quad = \|E_1\|^2 - 2\langle E_1, \omega E_2\rangle + \|E_2\|^2_\omega.
\end{aligned}$$

A simple choice of γ would be $\gamma = \frac{1}{L_F}$. Then, we see that, in an increasing order:

$$\begin{aligned}
\sigma(\mathcal{H}_\gamma^n) &= \{(1-\gamma L_F)^n, \cdots, (1-\gamma\lambda_F)^n\} \\
\sigma(I-\mathcal{H}_\gamma^n) &= \{1-(1-\gamma\lambda_F)^n, \cdots, 1-(1-\gamma L_F)^n\} \\
\rho((I-\mathcal{H}_\gamma^n)A^{-1}) &= \max\left\{\frac{(1-(1-\gamma\lambda_F)^n)}{\lambda_F}, \frac{1-(1-\gamma L_F)^n}{L_F}\right\} \\
\rho(\mathcal{H}_\gamma^n) &= \delta^n.
\end{aligned}$$

The easy bound would be for E_2. We note that

$$\|E_2\| \leq \delta^n \|Kz_* - Kz_k\|. \tag{140}$$

On the other hand, we have that with $S = (I-\mathcal{H}_\gamma^n)A^{-1}$,

$$\lambda_{min}(S) = \min\left\{\frac{(1-\delta^n)}{\lambda_F}, \frac{1}{L_F}\right\} \quad \text{and} \quad \lambda_{max}(S) = \max\left\{\frac{(1-\delta^n)}{\lambda_F}, \frac{1}{L_F}\right\}.$$

Thus, the choice of ω given as follows:

$$\omega = \frac{2}{\frac{1-\delta^n}{\lambda_F} + \frac{1}{L_F}}. \tag{141}$$

With this choice, we obtain the convergence rate given as follows:

$$\|E_1\| \leq \left(\frac{\kappa(S)-1}{\kappa(S)+1}\right)\|H_* - H_k\|. \tag{142}$$

We shall now make it clear by choosing n small and n large. First, for $n \gg 1$, we have that

$$\|E_2\| \leq \delta^n \|Kz_* - Kz_k\| \tag{143}$$

and

$$\kappa(S) = \frac{L_F}{\lambda_F}(1-\delta^n). \tag{144}$$

On the other hand, if $n = O(1)$, then it is unclear since there are many factors. This completes the proof. $\square$

Remark 8. *We remark that the best choice for γ would be $\gamma = \frac{2}{\lambda_F + L_F}$, then we have*

$$\|E_2\| \leq \left(\frac{\kappa(A)-1}{\kappa(A)+1}\right)^n \|Kz_* - Kz_k\|. \tag{145}$$

On the other hand, such a choice makes $\mathcal{H}_\gamma$ indefinite. Thus, the analysis of H error becomes nontrivial.

Remark 9. *We shall now discuss the choice of ω in Algorithm 1. The choice of ω was given as follows:*

$$\omega = \frac{1}{n\gamma}, \tag{146}$$

where n is the GD steps with $\gamma = 1/L_F$. Especially, for the linear problem, the convergence is guaranteed for $\omega < 2L_F$ assuming that n is sufficiently large since if not, it may happen that

$$\rho(S) = \frac{1-\delta^n}{\lambda_F}. \tag{147}$$

Thus, assuming that n is sufficiently large and

$$\omega = \frac{L_F}{n}, \tag{148}$$

we see that the following holds:

$$\rho(I - \omega S) = 1 - \frac{\omega}{L_F} = 1 - \frac{1}{n} = \frac{n-1}{n}, \tag{149}$$

which is because

$$1 - \frac{\omega}{L_F} > 1 - \frac{\omega(1-\delta^n)}{\lambda_F}. \tag{150}$$

This means, the convergence is much deteriorate when n gets larger. Furthermore, if we choose $n = \sqrt{\kappa(A)}$, then we get an optimal convergence rate just like the conjugate gradient method. This argument agrees with what is observed in [12] for the stochastic case. On the other hand, for the case when $r \neq 0$, we have to analyze the spectral radius of $S_r = (I - (I - \gamma A_r)^n)A_r^{-1}$. With $\gamma = \frac{1}{r+L_F}$, we have that

$$\lambda_{min}(S_r) = (1-\delta^n)\frac{1}{r+\lambda_F} \quad \text{and} \quad \lambda_{max}(S_r) = \frac{1}{r+L_F}. \tag{151}$$

Therefore, in a similar manner, if we choose $\omega = \frac{r+L_F}{n}$, then we have that

$$\rho(I - \omega S_r) = \frac{n-1}{n}. \tag{152}$$

Therefore, we can choose $n = \sqrt{\kappa(A_r)}$. This can give faster convergence. Therefore, we can apply larger r, which will reduce the number of steps for GD, still leading to the faster convergence. On the other hand, we notice that there is an additional factor, dependent on $\frac{r}{r+\lambda_F}$, which deteriorates as r increases.

Corollary 2. *Let $n = 1$. Given $\gamma = \frac{1}{L_F}$ and $\omega = L_F$, the Algorithm 6 with GD given in the Algorithm* **??***, produces iterate (X_k, z_k, H_k), for which the following convergence rate is valid:*

$$\left\|E^H_{k+1}\right\|^2 + \left\|E^Z_{k+1}\right\|^2_\omega \leq \left(\frac{\kappa(A)-1}{\kappa(A)}\right)^2 \left(\left\|E^H_k\right\|^2 + \left\|E^Z_k\right\|^2_\omega\right).$$

Proof. We first recall that the following identity holds:

$$G_{1,0}(Kz_k; H_k) = (I - \gamma A)Kz_k + \gamma H_k = \mathcal{H}_\gamma K z_k + \gamma H_k. \tag{153}$$

We observe that with $\omega = 1/\gamma$, we have

$$E_1 = E^H_k - \omega(G_{1,0}(Kz_*; H_*) - G_{1,0}(Kz_*; H_k)) = E^H_k - \omega\gamma(H_* - H_k) = 0.$$

Thus, we only need to estimate E_2 given as follows:

$$\begin{aligned} E_2 &= G_{1,0}(Kz_*; H_k) - G_{1,0}(Kz_k; H_k) \\ &= \mathcal{H}_\gamma K(z_* - z_k). \end{aligned}$$

This leads that

$$\left\| E_k^H - \omega(G_{1,0}(Kz_*; H_*) - G_{1,0}(Kz_k; H_k)) \right\|^2 = \|E_2\|_\omega^2.$$

Note that the choice of $\gamma = 1/L_F$ gives

$$0 \le \mathcal{H}_\gamma = (I - \gamma A) \le \left(1 - \frac{\lambda_F}{L_F}\right) = \left(1 - \frac{1}{\kappa(A)}\right) = \left(\frac{\kappa(A) - 1}{\kappa(A)}\right). \tag{154}$$

Thus, we have that

$$\|E_2\| \le \left(\frac{\kappa(A) - 1}{\kappa(A)}\right) \|Kz_* - Kz_k\|. \tag{155}$$

This completes the proof. □

6 Alternative Proof

We now present the convergence analysis based on $A-$ norm of E_k^X and the convexity of A due to [14]. The federated learning algorithm is given as in Algorithm 8. We shall

---

Algorithm 8 Federated Learning formulation of FL

Given H_0 such that $K^T H_0 = 0$, updates are obtained as follows:

for $k = 0, 1, 2, \cdots, K - 1$ **do**

X_{k+1} update: (with $X_k = Kz_k$),

$$AX_{k+1} + rX_{k+1} = H_k + rKz_k, \tag{156}$$

z_{t+1} update:

$$K^T H_k + rK^T(Kz_{k+1} - X_{k+1}) = 0, \tag{157}$$

Update the Lagrange multiplier:

$$H_{k+1} = H_k + r(Kz_{k+1} - X_{k+1}). \tag{158}$$

end for

---

adapt the proof originated from [14] to establish the result of linear convergence.

Lemma 13. *The Algorithm 8 produces iterates (X_k, z_k, H_k) such that the following identity holds: for all $k = 0, 1, 2 \cdots$,*

$$\begin{aligned} A(X_{k+1}) - H_{k+1} + r(Kz_{k+1} - Kz_k) &= 0, \\ H_{k+1} - H_k + r(I - P_Z)X_{k+1} &= 0, \\ Kz_{k+1} - P_Z X_{k+1} &= 0. \end{aligned}$$

Proof. First of all, we note that $K^T H = 0$ with $H = H_k$ or $H = H_*$ and also

$$Kz = P_Z X. \tag{159}$$

This is then easy to see that

$$\begin{aligned} A(X_{k+1}) - H_{k+1} + r(Kz_{k+1} - Kz_k) &= 0, \\ K^T H_{k+1} &= 0, \\ H_{k+1} - H_k - r(Kz_{k+1} - X_{k+1}) &= 0. \end{aligned}$$

Multiplying the third equation, by K^T, we obtain the desired equation. This completes the proof. □

We shall obtain a simple but important lemma:

Lemma 14. *We have the following identities:*

$$\begin{aligned} A(X_{k+1}) - A(X_*) &= r(Kz_k - Kz_{k+1}) + H_{k+1} - H_*, && (160) \\ H_{k+1} - H_k &= -rQ_Z(X_{k+1} - X_*), && (161) \\ Kz_{k+1} - Kz_* &= P_Z(X_{k+1} - X_*). && (162) \end{aligned}$$

Proof. We recall the optimality condition, which can be given as follows:

$$\begin{aligned} A(X_*) - H_* &= 0 \\ K^T H_* &= 0 \\ Kz_* - X_* &= 0 \end{aligned}$$

However, using the property of Q_Z and P_Z, the optimality condition implies that it holds true

$$\begin{aligned} 0 &= P_Z(Kz_* - X_*) = Kz_* - P_Z X_* && (163a) \\ 0 &= Q_Z(Kz_* - X_*) = -Q_Z X_*. && (163b) \end{aligned}$$

Subtracting the optimality conditions (163) from (160), completes the proof. □

The main theorem considers the convergence of a vector U that combines the primal variable Kz and the dual variable H,

$$U = \begin{pmatrix} Kz \\ H \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} rI & 0 \\ 0 & \frac{1}{r}I \end{pmatrix} \tag{164}$$

We also define a $C-$norm on $U = (Kz, H)^T$ by

$$\begin{aligned} \|U\|_C^2 &= r\|Kz\|^2 + \frac{1}{r}\|H\|^2 \\ &= r(Kz, Kz) + \frac{1}{r}(H, H) = (rK^T Kz, z) + \frac{1}{r}(H, H). \end{aligned}$$

We shall need the following identity:

Lemma 15. *We have the following identity:*

$$2\langle Y_k - Y_{k+1}, Y_{k+1} - Y_* \rangle = \|Y_k - Y_*\|^2 - \|Y_{k+1} - Y_*\|^2 - \|Y_{k+1} - Y_k\|^2.$$

Proof.

$$\begin{aligned}
\langle Y_k - Y_{k+1}, Y_{k+1} - Y_* \rangle &= \langle Y_k - Y_* - (Y_{k+1} - Y_*), Y_{k+1} - Y_* \rangle \\
&= \langle Y_k - Y_* - (Y_{k+1} - Y_*), Y_{k+1} - Y_* \rangle \\
&= \langle Y_k - Y_*, Y_{k+1} - Y_* \rangle - \|Y_{k+1} - Y_*\|^2
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
\langle Y_k - Y_{k+1}, Y_{k+1} - Y_* \rangle &= \langle Y_k - Y_{k+1}, Y_{k+1} - Y_k + Y_k - Y_* \rangle \\
&= -\|Y_{k+1} - Y_k\|^2 + \langle Y_k - Y_{k+1}, Y_k - Y_* \rangle.
\end{aligned}$$

By adding these two identities, we obtain the result. This completes the proof. □

Theorem 11. *Assume that $K^T H_0 = 0$. Then the ADMM iterations produces $U_k = [Kz_k; H_k]$ that is linearly convergent to the optimal solution $U_* = [Kz_*; H_*]$ in the C-norm, defined by*

$$\|U_{k+1} - U_*\|_C^2 \le \frac{1}{1+\delta}\|U_k - U_*\|_C^2, \tag{165}$$

where δ is some positive parameter. Furthermore, X_k is linearly convergent to the optimal solution X_* in the following form

$$\|X_{k+1} - X_*\|^2 \le \frac{1}{2\lambda_F}\|U_k - U_*\|_C^2 \tag{166}$$

Proof. We begin with using the λ_F−strongly convexity condition for F as follows:

$$\begin{aligned}
\lambda_F\|X_{k+1} - X_*\|^2 &\le \langle X_{k+1} - X_*, A(X_{k+1}) - A(X_*) \rangle \\
&\le \langle X_{k+1} - X_*, rK(z_k - z_{k+1}) \rangle + \langle X_{k+1} - X_*, H_{k+1} - H_* \rangle \\
&= r\langle X_{k+1} - X_*, P_Z(K(z_k - z_{k+1})) \rangle + \langle X_{k+1} - X_*, Q_Z(H_{k+1} - H_*) \rangle \\
&= r\langle P_Z(X_{k+1} - X_*), Kz_k - Kz_{k+1} \rangle + \langle Q_Z(X_{k+1} - X_*), H_{k+1} - H_* \rangle \\
&= r\langle Kz_k - Kz_{k+1}, Kz_{k+1} - Kz_* \rangle + \frac{1}{r}\langle H_k - H_{k+1}, H_{k+1} - H_* \rangle \\
&= (U_k - U_{k+1})^T C (U_{k+1} - U_*),
\end{aligned}$$

where

$$U_k = \begin{pmatrix} Kz_k \\ H_k \end{pmatrix}, \quad U_{k+1} = \begin{pmatrix} Kz_{k+1} \\ H_{k+1} \end{pmatrix}, \quad U_* = \begin{pmatrix} Kz_* \\ H_* \end{pmatrix} \tag{167}$$

and the matrix

$$C = \begin{pmatrix} rI & 0 \\ 0 & \frac{1}{r}I \end{pmatrix} \tag{168}$$

This implies

$$\lambda_F\|X_{k+1} - X_*\|^2 \le \frac{1}{2}\|U_k - U_*\|_C^2 - \frac{1}{2}\|U_{k+1} - U_*\|_C^2 - \frac{1}{2}\|U_k - U_{k+1}\|_C^2. \tag{169}$$

Now, by rearranging terms, we have

$$2\lambda_F\|X_{k+1} - X_*\|^2 + \|U_k - U_{k+1}\|_C^2 + \|U_{k+1} - U_*\|_C^2 \le \|U_k - U_*\|_C^2. \tag{170}$$

This immediately, leads to

$$\|X_{k+1} - X_*\|^2 \le \frac{1}{2\lambda_F}\|U_k - U_*\|_C^2. \tag{171}$$

Having (170), it suffices to show for some $\delta > 0$, we have

$$\delta\|U_{k+1} - U_*\|_C^2 \le 2\lambda_F\|X_{k+1} - X_*\|^2 + \|U_k - U_{k+1}\|_C^2, \tag{172}$$

or equivalently,

$$\begin{aligned}\delta\left(r\|Kz_{k+1} - Kz_*\|^2 + \frac{1}{r}\|H_{k+1} - H_*\|^2\right) \quad &\le \quad 2\lambda_F\|X_{k+1} - X_*\|^2 \\ &+ \quad r\|Kz_k - Kz_{k+1}\|^2 + \frac{1}{r}\|H_k - H_{k+1}\|^2,\end{aligned}$$

which will imply the desired inequality:

$$\|U_{k+1} - U_*\|_C^2 \le \frac{1}{1+\delta}\|U_k - U_*\|_C^2. \tag{173}$$

To prove the inequality (172), first, we observe that the following holds true:

$$\|Kz_{k+1} - Kz_*\|^2 \le \|X_{k+1} - X_*\|^2. \tag{174}$$

Further, from (160) and using L_F-smoothness of F, we have

$$\|H_{k+1} - H_*\| \le r\|Kz_k - Kz_{k+1}\| + L_F\|X_{k+1} - X_*\| \tag{175}$$

This implies

$$\begin{aligned}\|H_{k+1} - H_*\|^2 &\le (r\|Kz_k - Kz_{k+1}\| + L_F\|X_{k+1} - X_*\|)^2 \\ &\le 2\left(r^2\|Kz_k - Kz_{k+1}\|^2 + L_F^2\|X_{k+1} - X_*\|^2\right).\end{aligned} \tag{176}$$

Thus, we have

$$\frac{1}{r}\|H_{k+1} - H_*\|^2 \le 2\left(r\|Kz_k - Kz_{k+1}\|^2 + \frac{L_F^2}{r}\|X_{k+1} - X_*\|^2\right). \tag{177}$$

Substituting (174) and (176) into left hand side of (173) and rearranging, we have for

$$\delta = \min\left\{\frac{1}{2}, \frac{2\lambda_F}{r + \frac{2L_F^2}{r}}\right\}, \tag{178}$$

we have

$$\begin{aligned}\delta\left(r\|Kz_{k+1}-Kz_*\|^2+\frac{1}{r}\|H_{k+1}-H_*\|^2\right) &\le \delta\left(r+\frac{2L_F^2}{r}\right)\|X_{k+1}-X_*\|^2+2\delta r\|Kz_k-Kz_{k+1}\|^2 \\ &\le 2\lambda_F\|X_{k+1}-X_*\|^2+r\|Kz_k-Kz_{k+1}\|^2 \\ &\quad +\frac{1}{r}\|H_k-H_{k+1}\|^2,\end{aligned}$$

by making δ sufficiently small such that

$$\delta \le \frac{2\lambda_F}{r+\frac{2L_F^2}{r}} = \frac{2r\lambda_F}{r^2+2L_F^2} \quad \textit{and} \quad \delta \le \frac{1}{2}. \tag{179}$$

The convergence rate is then given as follows:

$$\frac{1}{1+\frac{2r\lambda_F}{r^2+2L_F^2}} = \frac{r^2+2L_F^2}{r^2+2L_F^2+2r\lambda_F} \quad \approx \quad \frac{r}{r+\lambda_F} \quad \text{for} \quad r \gg 1. \tag{180}$$

This completes the proof. □

Remark 10. *This approach is difficult to apply for inexact GS case even when F is a quadratic functional or when $\omega \neq r$. Also, it is different from the proposed approach, since it can not show the convergence for $r = 0$. For $r = 0$, we have that $\rho = 1$. Thus, the convergence can not be attained.*

7 On the Linear Convergence of ProxSkip

For the convergence measure, we introduce the so-called Lyapunov function:

$$\Psi_k := \|Kz_k - X_*\|^2 + \frac{\gamma^2}{p^2}\|H_k - H_*\|^2. \tag{185}$$

We further define

$$\begin{aligned} w_k &= Kz_k - A(Kz_k) && (186a) \\ w_* &= Kz_* - A(X_*). && (186b) \end{aligned}$$

We consider to choose

$$\omega = \frac{1}{\gamma n} \ \text{ and } \ p = \frac{1}{n}. \tag{187}$$

Therefore, $\omega^2 = p^2/\gamma^2$. Thus, the weights in the Lyapunov function is nothing else than,

$$\omega^2\Psi_k := \omega^2\|Kz_k - X_*\|^2 + \|H_k - H_*\|^2. \tag{188}$$

Under these settings, we shall then show that ProxSkip generates iterates $\{Kz_k\}_{k=1,\cdots}$ converges linearly in the sense that

$$\mathbb{E}(\Psi_T) \le (1-\zeta)^T \Psi_0, \tag{189}$$

where $\zeta = \kappa(F) = L_F/\lambda_F$.

Algorithm 9 ProxSkip

Given $\gamma > 0$ and probability $p > 0$, initial iterate Kz_0 and initial control variate, such that $K^T H_0 = 0$, updates are obtained as follows:

for $k = 0, 1, 2, \cdots, T$ **do**

X_{k+1} update: (with $X_k = Kz_k$),

$$X_{k+1} = X_k - \gamma(A(X_k) - H_k), \tag{181}$$

Flip a coin $\theta_k \in \{0, 1\}$ where $\mathrm{Prob}(\theta_k = 1) = p$

if $\theta_k = 1$ **then**

$$K^T H_k + rK^T(Kz_{k+1} - X_{k+1}) = 0, \tag{182}$$

else

$$Kz_{k+1} = X_{k+1}, \tag{183}$$

end if

$$H_{k+1} = H_k + \frac{p}{\gamma}(Kz_{k+1} - X_{k+1}). \tag{184}$$

end for

Theorem 12. *For $\gamma = 1/L_F > 0$ and $0 < p \leq 1$, we have* 644

$$\mathbb{E}(\Psi_{k+1}) \leq (1 - \min\{\gamma\lambda_F, p^2\})\Psi_k. \tag{190}$$

where the expectation is taken over the θ_k in the Algorithm 1. 645

Proof. We let $Kz = P_Z(X)$. 646

$$X := X_{k+1} - \frac{\gamma}{p}H_k \quad \text{and} \quad Y = X_* - \frac{\gamma}{p}H_*. \tag{191}$$

Then since $P_Z H_k = 0$ and $P_Z H_* = 0$, we have 647

$$X_* = P_Z(Y). \tag{192}$$

The method reads as follows: 648

$$Kz_{k+1} = \begin{cases} P_Z(X) & \text{with probability } p \\ X_{k+1} & \text{with probability } 1 - p \end{cases} \tag{193}$$

Furthermore, we have that 649

$$H_{k+1} = H_k + \frac{p}{\gamma}(Kz_{k+1} - X_{k+1}) = \begin{cases} H_k + \frac{p}{\gamma}(P_Z(X_{k+1}) - X_{k+1}) & \text{with } p \\ H_k & \text{with } 1 - p \end{cases} \tag{194}$$

Now, we compute the expected value of the Lyapunov function 650

$$\Psi_k := \|Kz_k - X_*\|^2 + \frac{\gamma^2}{p^2}\|H_k - H_*\|^2 \tag{195}$$

at the time step $k+1$, with respect to the coin toss at iteration k, which is

$$\mathbb{E}(\Psi_{k+1}) = p\left(\|P_Z(X_{k+1}) - X_*\|^2 + \frac{\gamma^2}{p^2}\left\|H_k + \frac{p}{\gamma}(P_Z(X_{k+1}) - X_{k+1}) - H_*\right\|^2\right) \tag{196}$$

$$+(1-p)\left(\|X_{k+1} - X_*\|^2 + \frac{\gamma^2}{p^2}\|H_k - H_*\|^2\right) \tag{197}$$

$$= p\left(\|P_Z(X) - P_Z(Y)\|^2 + \left\|\frac{\gamma}{p}H_k + (P_Z(X) - X_{k+1}) - \frac{\gamma}{p}H_*\right\|^2\right) \tag{198}$$

$$+(1-p)\left(\|X_{k+1} - X_*\|^2 + \frac{\gamma^2}{p^2}\|H_k - H_*\|^2\right) \tag{199}$$

$$= p\left(\|P_Z(X) - P_Z(Y)\|^2 + \left\|P_Z(X) - \left(X_{k+1} - \frac{\gamma}{p}H_k\right) + (Y - X_*)\right\|^2\right) \tag{200}$$

$$+(1-p)\left(\|X_{k+1} - X_*\|^2 + \frac{\gamma^2}{p^2}\|H_k - H_*\|^2\right) \tag{201}$$

$$= p\left(\|P_Z(X) - P_Z(Y)\|^2 + \|Q_Z(X) - Q_Z(Y)\|^2\right) \tag{202}$$

$$+(1-p)\left(\|X_{k+1} - X_*\|^2 + \frac{\gamma^2}{p^2}\|H_k - H_*\|^2\right) \tag{203}$$

$$\leq p\left\|\left(X_{k+1} - \frac{\gamma}{p}H_k\right) - \left(X_* - \frac{\gamma}{p}H_*\right)\right\|^2 \tag{204}$$

$$+(1-p)\left(\|X_{k+1} - X_*\|^2 + \frac{\gamma^2}{p^2}\|H_k - H_*\|^2\right) \tag{205}$$

$$= \|X_{k+1} - X_*\|^2 + \frac{\gamma^2}{p^2}\|H_k - H_*\|^2 - 2\gamma\langle X_{k+1} - X_*, H_k - H_*\rangle \tag{206}$$

$$= \|X_k - \gamma AX_k + \gamma H_k - (X_* - \gamma AX_* + \gamma H_*)\|^2 + \frac{\gamma^2}{p^2}\|H_k - H_*\|^2 \tag{207}$$

$$-2\gamma\langle X_k - \gamma AX_k + \gamma H_k - X_* - \gamma AX_* + \gamma H_*, H_k - H_*\rangle \tag{208}$$

$$= \|(X_k - \gamma AX_k - (X_* - \gamma AX_*) + \gamma(H_k - H_*)\|^2 + \frac{\gamma^2}{p^2}\|H_k - H_*\|^2 \tag{209}$$

$$-2\gamma\langle(X_k - \gamma AX_k - (X_* - \gamma AX_*) + \gamma(H_k - H_*), H_k - H_*\rangle \tag{210}$$

$$= \|(X_k - \gamma AX_k) - (X_* - \gamma AX_*)\|^2 + 2\gamma\langle(X_k - \gamma AX_k) - (X_* - \gamma AX_*), (H_k - H_*)\rangle \tag{211}$$

$$+\gamma^2\|H_k - H_*\|^2 + \frac{\gamma^2}{p^2}\|H_k - H_*\|^2 \tag{212}$$

$$-2\gamma\langle(X_k - \gamma AX_k) - (X_* - \gamma AX_*) + \gamma(H_k - H_*), H_k - H_*\rangle \tag{213}$$

$$= \|(X_k - \gamma AX_k) - (X_* - \gamma AX_*)\|^2 - \gamma^2\|H_k - H_*\|^2 \tag{214}$$

$$+\frac{\gamma^2}{p^2}\|H_k - H_*\|^2 \tag{215}$$

$$\leq (1 - \gamma\lambda_F)\|X_K - X_*\|^2 + (1 - p^2)\frac{\gamma^2}{p^2}\|H_k - H_*\|^2 \tag{216}$$

$$\leq \max\{(1 - \gamma\lambda_F), (1 - p^2)\}\Psi_k \tag{217}$$

$$= (1 - \min\{\gamma\lambda_F, p^2\})\Psi_k. \tag{218}$$

This completes the proof. □

7.1 U block analysis in standard way

We consider to solve the following system:

$$\begin{aligned} A_r X - rKz &= f \\ -rK^T X + rK^T Kz &= 0. \end{aligned}$$

The Gauss-Seidel method leads

$$\begin{aligned} A_r X_{k+1} - rKz_k &= f \\ -rK^T X_{k+1} + rK^T Kz_{k+1} &= 0. \end{aligned}$$

This is equivalent to say that

$$\begin{aligned} X_{k+1} &= f + rKz_k \\ Kz_{k+1} &= K(rK^T K)^{-1} rK^T X_{k+1}. \end{aligned}$$

This can be rewritten as follows: with $B_r = rK^T K$,

$$\begin{aligned} X_{k+1} &= X_k + A_r^{-1}(f + rKz_k - A_r X_k) \\ Kz_{k+1} &= Kz_k + KB_r^{-1}(rK^T X_{k+1} - B_r z_k), \end{aligned}$$

where $B_r = rK^T K$. On the other hand, we have the optimality condition that

$$\begin{aligned} X_* &= X_* + A_r^{-1}(f + rKz_* - A_r X_*) \\ Kz_* &= Kz_* + KB_r^{-1}(rK^T X_* - B_r z_*), \end{aligned}$$

Therefore, the error analysis is given as follows:

$$\begin{aligned} E^X_{k+1} &= E^X_k + R_r(rE^Z_k - A_r E^X_k) = (I - R_r A_r)E^X_k + rR_r E^Z_k \\ E^Z_{k+1} &= E^Z_k + KB_r^{-1}(rK^T X_* - B_r z_*) \\ &= E^Z_k + KB_r^{-1} rK^T E^X_{k+1} - E^Z_k = KB_r^{-1} rK^T E^X_{k+1} \\ &= P_Z E^X_{k+1} = P_Z((I - R_r A_r)E^X_k + rR_r E^Z_k) \end{aligned}$$

Note that $P_Z = K(K^T K)^{-1} K^T$. In a matrix form, we have that

$$\begin{pmatrix} E^X_{k+1} \\ E^Z_{k+1} \end{pmatrix} = \begin{pmatrix} I - R_r A_r & rR_r \\ P_Z(I - R_r A_r) & P_Z(rR_r) \end{pmatrix} \begin{pmatrix} E^X_k \\ E^Z_k \end{pmatrix}.$$

$$R_r \quad := \quad (I - (I - \gamma A_r)^n)A_r^{-1} = (I - (I - \gamma A_r)^n)A_r^{-1}.$$

For n large enough, we shall consider two operators: with $\gamma = 1/(r + L_F)$,

$$\begin{aligned} I - R_r A_r &= (I - \gamma A_r)^n = \rho^n \\ \sigma(R_r) &= \left(1 - \left(1 - \frac{r + \lambda}{r + L_F}\right)^n\right) \frac{1}{r + \lambda} \\ \rho(R_r) &\leq (1 - \delta^n) \frac{1}{r + \lambda_F}. \end{aligned}$$

Thus, we have that

$$\begin{pmatrix} \|E^X_{k+1}\| \\ \|E^Z_{k+1}\| \end{pmatrix} = \begin{pmatrix} \delta^n & (1 - \delta^n)\frac{r}{r+\lambda_F} \\ \delta^n & (1 - \delta^n)\frac{r}{r+\lambda_F} \end{pmatrix} \begin{pmatrix} \|E^X_k\| \\ \|E^Z_k\| \end{pmatrix}.$$