APPENDIX

CONTENTS

1 Introduction **1**

1.1 The expressive power of variational inequalities 1

1.2 Training of supervised models via distributed optimization 2

1.3 Two classes of compression operators 2

1.4 Towards communication-efficient distributed methods for VIs 2

2 Summary of Contributions **2**

2.1 Two distributed problems: deterministic and stochastic 3

2.2 Two new methods with compressed communication: MASHA1 and MASHA2 3

2.3 Theoretical complexity results 4

3 Problem Formulation and Assumptions **4**

3.1 Problem formulation 4

3.2 Assumptions 5

4 MASHA1: Handling Unbiased Compressors **5**

4.1 The Algorithm 5

4.2 Theory 6

5 Experiments **8**

5.1 Bilinear Saddle Point Problem 8

5.2 Adversarial Training of Transformers 8

A Proof of Theorem 1 **14**

A.1 A lemma 14

A.2 Deterministic case: Theorem 1 14

A.3 Strongly monotone/convex case 15

A.4 Monotone/convex case ($\mu_h = 0$, $\mu_F = 0$) 18

A.5 Stochastic case 23

B MASHA2: Handling Contractive Compressors **23**

C Proof of Theorem 2 **23**

C.1 Strongly monotone case 25

D Motivating Examples **29**

A PROOF OF THEOREM 1

A.1 A LEMMA

Lemma 1 *Let h be μ_h - strongly convex and $z^+ = prox_{\gamma h}(z)$. Then for all $x \in \mathbb{R}^d$ the following iniquility holds:*

$$\langle z^+ - z, x - z^+ \rangle \geq \gamma \left(h(z^+) - h(x) + \frac{\mu_h}{2}\|z^+ - x\|^2 \right). \tag{14}$$

Proof: We use $\gamma\mu$-strong convexity of the function γh (8):

$$\gamma \left(h(x) - h(z^+) \right) - \langle \gamma \nabla h(z^+), x - z^+ \rangle \geq \frac{\gamma\mu_h}{2}\|x - z^+\|^2.$$

With definition of prox and necessary optimality condition: $\gamma\nabla h(z^+) = z - z^+$, it completes the proof.

□

In the next theorem we will use the following notation:

$$g^k = F(w^k), \quad g^{k+1/2} = Q^{\text{serv}}\left[\frac{1}{M}\sum_{m=1}^{M} Q_m^{\text{dev}}(F_m(z^{k+1/2}) - F_m(w^k)) \right] + F(w^k).$$

A.2 DETERMINISTIC CASE: THEOREM 1

.

Proof of Theorem 1: By Lemma 1 for $z^{k+1/2} = \text{prox}_{\gamma h}(\bar{z}^k - \gamma g^k)$ and $z^{k+1} = \text{prox}_{\gamma h}(\bar{z}^k - \gamma g^{k+1/2})$ with $x = u$ we get

$$\langle z^{k+1} - \bar{z}^k + \gamma g^{k+1/2}, u - z^{k+1} \rangle \geq \gamma \left(h(z^{k+1}) - h(u) + \frac{\mu_h}{2}\|z^{k+1} - u\|^2 \right),$$
$$\langle z^{k+1/2} - \bar{z}^k + \gamma g^k, z^{k+1} - z^{k+1/2} \rangle \geq \gamma \left(h(z^{k+1/2}) - h(z^{k+1}) + \frac{\mu_h}{2}\|z^{k+1} - z^{k+1/2}\|^2 \right).$$

Then we sum two inequalities and make some rearrangement:

$$\begin{aligned}
&\langle z^{k+1} - \bar{z}^k, u - z^{k+1} \rangle + \langle z^{k+1/2} - \bar{z}^k, z^{k+1} - z^{k+1/2} \rangle \\
&+ \gamma \langle g^{k+1/2} - g^k, z^{k+1/2} - z^{k+1} \rangle + \gamma \langle g^{k+1/2}, u - z^{k+1/2} \rangle \\
&\qquad \geq \gamma \left(h(z^{k+1/2}) - h(u) + \frac{\mu_h}{2}\|z^{k+1} - z^{k+1/2}\|^2 + \frac{\mu_h}{2}\|z^{k+1} - u\|^2 \right).
\end{aligned}$$

Multiplying by 2 and using definition of $\bar{z}^k$, we have

$$\begin{aligned}
&2\tau \langle z^{k+1} - z^k, u - z^{k+1} \rangle + 2(1-\tau)\langle z^{k+1} - w^k, u - z^{k+1} \rangle \\
&+ 2\tau \langle z^{k+1/2} - z^k, z^{k+1} - z^{k+1/2} \rangle + 2(1-\tau)\langle z^{k+1/2} - w^k, z^{k+1} - z^{k+1/2} \rangle \\
&+ 2\gamma \langle g^{k+1/2} - g^k, z^{k+1/2} - z^{k+1} \rangle + 2\gamma \langle g^{k+1/2}, u - z^{k+1/2} \rangle \\
&\qquad \geq 2\gamma \left(h(z^{k+1/2}) - h(u) + \frac{\mu_h}{2}\|z^{k+1} - z^{k+1/2}\|^2 + \frac{\mu_h}{2}\|z^{k+1} - u\|^2 \right).
\end{aligned}$$

For the first and second lines we use identity $2\langle a, b \rangle = \|a + b\|^2 - \|a\|^2 - \|b\|^2$, and get

$$\begin{aligned}
&\tau \left(\|z^k - u\|^2 - \|z^{k+1} - z^k\|^2 - \|z^{k+1} - u\|^2 \right) \\
&+ (1-\tau) \left(\|w^k - u\|^2 - \|z^{k+1} - w^k\|^2 - \|z^{k+1} - u\|^2 \right) \\
&+ \tau(\|z^{k+1} - z^k\|^2 - \|z^{k+1/2} - z^k\|^2 - \|z^{k+1} - z^{k+1/2}\|^2) \\
&+ (1-\tau)(\|z^{k+1} - w^k\|^2 - \|z^{k+1/2} - w^k\|^2 - \|z^{k+1} - z^{k+1/2}\|^2) \\
&+ 2\gamma \langle g^{k+1/2} - g^k, z^{k+1/2} - z^{k+1} \rangle + 2\gamma \langle g^{k+1/2}, u - z^{k+1/2} \rangle \\
&\qquad \geq 2\gamma \left(h(z^{k+1/2}) - h(u) + \frac{\mu_h}{2}\|z^{k+1} - z^{k+1/2}\|^2 + \frac{\mu_h}{2}\|z^{k+1} - u\|^2 \right).
\end{aligned}$$

A small rearrangement gives

$$
\begin{aligned}
(1+\gamma\mu_h)\|z^{k+1}-u\|^2 &\leq \tau\|z^k-u\|^2+(1-\tau)\|w^k-u\|^2\\
&-\tau\|z^{k+1/2}-z^k\|^2-(1-\tau)\|z^{k+1/2}-w^k\|^2-(1+\gamma\mu_h)\|z^{k+1}-z^{k+1/2}\|^2\\
&+2\gamma\langle g^{k+1/2}-g^k, z^{k+1/2}-z^{k+1}\rangle-2\gamma\langle g^{k+1/2}, z^{k+1/2}-u\rangle-2\gamma\left(h(z^{k+1/2})-h(u)\right).
\end{aligned}
$$

By simple fact: $2\langle a,b\rangle \leq \eta\|a\|^2+\frac{1}{\eta}\|b\|^2$ with $a=g^{k+1/2}-g^k, b=z^{k+1/2}-z^{k+1}, \eta=2\gamma$, we get

$$
\begin{aligned}
(1+\gamma\mu_h)\|z^{k+1}-u\|^2 &\leq \tau\|z^k-u\|^2+(1-\tau)\|w^k-u\|^2\\
&-\tau\|z^{k+1/2}-z^k\|^2-(1-\tau)\|z^{k+1/2}-w^k\|^2-\left(\frac{1}{2}+\gamma\mu_h\right)\|z^{k+1}-z^{k+1/2}\|^2\\
&+2\gamma^2\|g^{k+1/2}-g^k\|^2-2\gamma\langle g^{k+1/2}, z^{k+1/2}-u\rangle-2\gamma\left(h(z^{k+1/2})-h(u)\right). \qquad (15)
\end{aligned}
$$

We now consider the two cases of the theorem separately.

A.3 Strongly monotone/convex case

Let substitute $u=z^*$, take full mathematical expectation and get

$$
\begin{aligned}
(1+\gamma\mu_h)\mathbb{E}\left[\|z^{k+1}-z^*\|^2\right] &\leq \tau\mathbb{E}\left[\|z^k-z^*\|^2\right]+(1-\tau)\mathbb{E}\left[\|w^k-z^*\|^2\right]\\
&-\tau\mathbb{E}\left[\|z^{k+1/2}-z^k\|^2\right]-(1-\tau)\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right]-\left(\frac{1}{2}+\gamma\mu_h\right)\mathbb{E}\left[\|z^{k+1}-z^{k+1/2}\|^2\right]\\
&+2\gamma^2\mathbb{E}\left[\|g^{k+1/2}-g^k\|^2\right]-2\gamma\mathbb{E}\left[\langle g^{k+1/2}, z^{k+1/2}-z^*\rangle+h(z^{k+1/2})-h(z^*)\right]\\
&=\tau\mathbb{E}\left[\|z^k-z^*\|^2\right]+(1-\tau)\mathbb{E}\left[\|w^k-z^*\|^2\right]\\
&-\tau\mathbb{E}\left[\|z^{k+1/2}-z^k\|^2\right]-(1-\tau)\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right]-\left(\frac{1}{2}+\gamma\mu_h\right)\mathbb{E}\left[\|z^{k+1}-z^{k+1/2}\|^2\right]\\
&+2\gamma^2\mathbb{E}\left[\|g^{k+1/2}-g^k\|^2\right]-2\gamma\mathbb{E}\left[\langle\mathbb{E}\left[g^{k+1/2}\mid z^{k+1/2}\right], z^{k+1/2}-z^*\rangle+h(z^{k+1/2})-h(z^*)\right]. \qquad (16)
\end{aligned}
$$

Let us work with $\mathbb{E}\left[\|g^{k+1/2}-g^k\|^2\right]$, with (1) we get

$$
\begin{aligned}
\mathbb{E}\left[\|g^{k+1/2}-g^k\|^2\right] &= \mathbb{E}\left[\left\|Q^{\text{serv}}\left[\frac{1}{M}\sum_{m=1}^M Q_m^{\text{dev}}(F_m(z^{k+1/2})-F_m(w^k))\right]\right\|^2\right]\\
&\leq \frac{q^{\text{serv}}}{M^2}\mathbb{E}\left[\left\|\sum_{m=1}^M Q_m^{\text{dev}}(F_m(z^{k+1/2})-F_m(w^k))\right\|^2\right]\\
&= \frac{q^{\text{serv}}}{M^2}\sum_{m=1}^M\mathbb{E}\left[\left\|Q_m^{\text{dev}}(F_m(z^{k+1/2})-F_m(w^k))\right\|^2\right]\\
&\quad+\frac{q^{\text{serv}}}{M^2}\sum_{m\neq l}\mathbb{E}\left[\langle Q_m^{\text{dev}}(F_m(z^{k+1/2})-F_m(w^k)); Q_l^{\text{dev}}(F_l(z^{k+1/2})-F_l(w^k))\rangle\right]
\end{aligned}
$$

Next we apply (1) and Assumption 2 for the first term and independence and unbiasedness of Q for the second term:

$$\begin{aligned}
\mathbb{E}\left[\|g^{k+1/2}-g^k\|^2\right] &\leq \frac{q^{\text{serv}}}{M^2}\sum_{m=1}^{M} q_m^{\text{dev}} L_m^2 \mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right] \\
&\quad + \frac{q^{\text{serv}}}{M^2}\sum_{m\neq l} \mathbb{E}\left[\langle F_m(z^{k+1/2})-F_m(w^k); F_l(z^{k+1/2})-F_l(w^k)\rangle\right] \\
&\leq \frac{q^{\text{serv}}}{M^2}\sum_{m=1}^{M} q_m^{\text{dev}} L_m^2 \mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right] \\
&\quad + \frac{q^{\text{serv}}}{2M^2}\sum_{m\neq l} \mathbb{E}\left[\|F_m(z^{k+1/2})-F_m(w^k)\|^2 + \|F_l(z^{k+1/2})-F_l(w^k)\|^2\right] \\
&\leq \frac{q^{\text{serv}}}{M^2}\sum_{m=1}^{M} q_m^{\text{dev}} L_m^2 \mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right] \\
&\quad + \frac{q^{\text{serv}}}{2M^2}\sum_{m\neq l} \mathbb{E}\left[L_m^2\|z^{k+1/2}-w^k\|^2 + L_l^2\|z^{k+1/2}-w^k\|^2\right] \\
&= \frac{q^{\text{serv}}}{M^2}\sum_{m=1}^{M} q_m^{\text{dev}} L_m^2 \mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right] + \frac{q^{\text{serv}}(M-1)}{M^2}\sum_{m=1}^{M} L_m^2 \mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right] \\
&= \frac{q^{\text{serv}}}{M^2}\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right]\cdot\sum_{m=1}^{M}(q_m^{\text{dev}}+M-1)L_m^2 \qquad (17)
\end{aligned}$$

Let us define new constant $C_q = \sqrt{\frac{q^{\text{serv}}}{M^2}\sum_{m=1}^{M}(q_m^{\text{dev}}+M-1)L_m^2}$ and then connect (16) and (17):

$$\begin{aligned}
(1+\gamma\mu_h)\mathbb{E}\left[\|z^{k+1}-z^*\|^2\right] &\leq \tau\mathbb{E}\left[\|z^k-z^*\|^2\right] + (1-\tau)\mathbb{E}\left[\|w^k-z^*\|^2\right] \\
&\quad - \tau\mathbb{E}\left[\|z^{k+1/2}-z^k\|^2\right] - (1-\tau)\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right] - \left(\frac{1}{2}+\gamma\mu_h\right)\mathbb{E}\left[\|z^{k+1}-z^{k+1/2}\|^2\right] \\
&\quad + 2\gamma^2 C_q^2\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right] - 2\gamma\mathbb{E}\left[\langle F(z^{k+1/2}), z^{k+1/2}-z^*\rangle + h(z^{k+1/2}) - h(z^*)\right] \\
&= \tau\mathbb{E}\left[\|z^k-z^*\|^2\right] + (1-\tau)\,\mathbb{E}\left[\|w^k-z^*\|^2\right] - \tau\mathbb{E}\left[\|z^{k+1/2}-z^k\|^2\right] \\
&\quad - \left(\frac{1}{2}+\gamma\mu_h\right)\mathbb{E}\left[\|z^{k+1}-z^{k+1/2}\|^2\right] - \left((1-\tau)-2\gamma^2C_q^2\right)\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right] \\
&\quad - 2\gamma\mathbb{E}\left[\langle F(z^{k+1/2}), z^{k+1/2}-z^*\rangle + h(z^{k+1/2}) - h(z^*)\right].
\end{aligned}$$

The property of the solution (3) gives

$$\begin{aligned}
(1+\gamma\mu_h)\mathbb{E}\left[\|z^{k+1}-z^*\|^2\right] &\leq \tau\mathbb{E}\left[\|z^k-z^*\|^2\right] + (1-\tau)\,\mathbb{E}\left[\|w^k-z^*\|^2\right] - \tau\mathbb{E}\left[\|z^{k+1/2}-z^k\|^2\right] \\
&\quad - \left(\frac{1}{2}+\gamma\mu_h\right)\mathbb{E}\left[\|z^{k+1}-z^{k+1/2}\|^2\right] - \left((1-\tau)-2\gamma^2C_q^2\right)\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right] \\
&\quad - 2\gamma\mathbb{E}\left[\langle F(z^{k+1/2})-F(z^*), z^{k+1/2}-z^*\rangle\right].
\end{aligned}$$

And by Assumption 2 in strong monotone case we have

$$\begin{aligned}
(1+\gamma\mu_h)\mathbb{E}\left[\|z^{k+1}-z^*\|^2\right] &\leq \tau\mathbb{E}\left[\|z^k-z^*\|^2\right] + (1-\tau)\,\mathbb{E}\left[\|w^k-z^*\|^2\right] - \tau\mathbb{E}\left[\|z^{k+1/2}-z^k\|^2\right] \\
&\quad - \left(\frac{1}{2}+\gamma\mu_h\right)\mathbb{E}\left[\|z^{k+1}-z^{k+1/2}\|^2\right] - \left((1-\tau)-2\gamma^2C_q^2\right)\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right] \\
&\quad - 2\gamma\mu_F\mathbb{E}\left[\|z^{k+1/2}-z^*\|^2\right].
\end{aligned}$$

One the other hand we get

$$\mathbb{E}\left[\|w^{k+1}-z^*\|^2\right] = (1-\tau)\mathbb{E}\left[\|z^{k+1}-z^*\|^2\right] + \tau\mathbb{E}\left[\|w^k-z^*\|^2\right].$$

Summing two previous inequalities:

$$\begin{aligned}
\tau\mathbb{E}\left[\|z^{k+1}-z^*\|^2\right] + \mathbb{E}\left[\|w^{k+1}-z^*\|^2\right] &\leq \tau\mathbb{E}\left[\|z^k-z^*\|^2\right] + \mathbb{E}\left[\|w^k-z^*\|^2\right] \\
&- \tau\mathbb{E}\left[\|z^{k+1/2}-z^k\|^2\right] - \gamma\mu_h\mathbb{E}\left[\|z^{k+1}-z^*\|^2\right] - \left(\frac{1}{2}+\gamma\mu_h\right)\mathbb{E}\left[\|z^{k+1}-z^{k+1/2}\|^2\right] \\
&- \left((1-\tau)-2\gamma^2C_q^2\right)\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right] - 2\gamma\mu_F\mathbb{E}\left[\|z^{k+1/2}-z^*\|^2\right].
\end{aligned}$$

We have Lyapunov function in the left side:

$$\begin{aligned}
\mathbb{E}\left[V_{k+1}\right] &= \tau\mathbb{E}\left[\|z^{k+1}-z^*\|^2\right] + \mathbb{E}\left[\|w^{k+1}-z^*\|^2\right] \\
&\leq \tau\mathbb{E}\left[\|z^k-z^*\|^2\right] + \mathbb{E}\left[\|w^k-z^*\|^2\right] - \tau\mathbb{E}\left[\|z^{k+1/2}-z^k\|^2\right] \\
&- 2\gamma\mu_F\mathbb{E}\left[\|z^{k+1/2}-z^*\|^2\right] - \gamma\mu_h\mathbb{E}\left[\|z^{k+1}-z^*\|^2\right] - \left(\frac{1}{2}+\gamma\mu_h\right)\mathbb{E}\left[\|z^{k+1}-z^{k+1/2}\|^2\right] \\
&- \left((1-\tau)-2\gamma^2C_q^2\right)\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right].
\end{aligned}$$

With $-\|z^{k+1}-z^*\|^2 \leq -\frac{1}{2}\|z^{k+1/2}-z^*\|^2 + \|z^{k+1}-z^{k+1/2}\|^2$ we deduce:

$$\begin{aligned}
\mathbb{E}\left[V_{k+1}\right] &\leq \tau\mathbb{E}\left[\|z^k-z^*\|^2\right] + \mathbb{E}\left[\|w^k-z^*\|^2\right] - \tau\mathbb{E}\left[\|z^{k+1/2}-z^k\|^2\right] \\
&- \left((1-\tau)-2\gamma^2C_q^2\right)\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right] - \left(\frac{1}{2}+\gamma\mu_h\right)\mathbb{E}\left[\|z^{k+1}-z^{k+1/2}\|^2\right] \\
&- \gamma\left(2\mu_F+\frac{\mu_h}{2}\right)\tau\mathbb{E}\left[\|z^{k+1/2}-z^*\|^2\right] - \gamma\left(2\mu_F+\frac{\mu_h}{2}\right)\cdot(1-\tau)\mathbb{E}\left[\|z^{k+1/2}-z^*\|^2\right].
\end{aligned}$$

A simple facts: $\|z^{k+1/2}-z^*\|^2 \geq \frac{1}{2}\|z^k-z^*\|^2 - \|z^{k+1/2}-z^k\|^2$ and $\|z^{k+1/2}-z^*\|^2 \geq \frac{1}{2}\|w^k-z^*\|^2 - \|z^{k+1/2}-w^k\|^2$, gives

$$\begin{aligned}
\mathbb{E}\left[V_{k+1}\right] &\leq \tau\mathbb{E}\left[\|z^k-z^*\|^2\right] + \mathbb{E}\left[\|w^k-z^*\|^2\right] \\
&- \left((1-\tau)-2\gamma^2C_q^2-\gamma\left(2\mu_F+\frac{\mu_h}{2}\right)\cdot(1-\tau)\right)\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right] \\
&- \left(\frac{1}{2}+\gamma\mu_h\right)\mathbb{E}\left[\|z^{k+1}-z^{k+1/2}\|^2\right] - \left(1-\gamma\left(2\mu_F+\frac{\mu_h}{2}\right)\right)\tau\mathbb{E}\left[\|z^{k+1/2}-z^k\|^2\right] \\
&- \gamma\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\|z^k-z^*\|^2\right] - \gamma\left(\mu_F+\frac{\mu_h}{4}\right)\cdot(1-\tau)\mathbb{E}\left[\|w^k-z^*\|^2\right]. \qquad (18)
\end{aligned}$$

Next we work with the last line of (18):

$$\begin{aligned}
&-\gamma\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\|z^k-z^*\|^2\right] - \gamma\left(\mu_F+\frac{\mu_h}{4}\right)\cdot(1-\tau)\mathbb{E}\left[\|w^k-z^*\|^2\right] \\
&\quad= -\frac{\gamma}{2}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\|z^k-z^*\|^2\right] - \frac{\gamma}{2}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\|z^k-z^*\|^2\right] \\
&\quad\quad- \gamma\left(\mu_F+\frac{\mu_h}{4}\right)\cdot(1-\tau)\mathbb{E}\left[\|w^k-z^*\|^2\right] \\
&\quad\leq -\frac{\gamma}{2}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\|z^k-z^*\|^2\right] - \frac{\gamma}{4}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\|w^k-z^*\|^2\right] \\
&\quad\quad+ \frac{\gamma}{2}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\|z^k-w^k\|^2\right] - \gamma\left(\mu_F+\frac{\mu_h}{4}\right)\cdot(1-\tau)\mathbb{E}\left[\|w^k-z^*\|^2\right] \\
&\quad\leq -\frac{\gamma}{4}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\|z^k-z^*\|^2\right] - \frac{\gamma}{4}\left(\mu_F+\frac{\mu_h}{4}\right)\mathbb{E}\left[\|w^k-z^*\|^2\right] \\
&\quad\quad+ \frac{\gamma}{2}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\|z^k-w^k\|^2\right] \\
&\quad\leq -\frac{\gamma}{4}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\|z^k-z^*\|^2\right] - \frac{\gamma}{4}\left(\mu_F+\frac{\mu_h}{4}\right)\mathbb{E}\left[\|w^k-z^*\|^2\right] \\
&\quad\quad+ \gamma\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\|z^{k+1/2}-z^k\|^2\right] + \gamma\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right].
\end{aligned}$$

Substituting this into (18), we get

$$\begin{aligned}\mathbb{E}\left[V_{k+1}\right] \leq \tau\mathbb{E}\left[\|z^k - z^*\|^2\right] + \mathbb{E}\left[\|w^k - z^*\|^2\right] \\ - \left((1-\tau) - 2\gamma^2 C_q^2 - \gamma\left(2\mu_F + \frac{\mu_h}{2}\right)\right)\mathbb{E}\left[\|z^{k+1/2} - w^k\|^2\right] \\ - \left(\frac{1}{2} + \gamma\mu_h\right)\mathbb{E}\left[\|z^{k+1} - z^{k+1/2}\|^2\right] - \left(1 - 3\gamma\left(\mu_F + \frac{\mu_h}{4}\right)\right)\tau\mathbb{E}\left[\|z^{k+1/2} - z^k\|^2\right] \\ - \frac{\gamma}{4}\left(\mu_F + \frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\|z^k - z^*\|^2\right] - \frac{\gamma}{4}\left(\mu_F + \frac{\mu_h}{4}\right)\mathbb{E}\left[\|w^k - z^*\|^2\right]. \end{aligned} \tag{19}$$

It remains only to choose $\gamma \leq \min\left\{\frac{\sqrt{1-\tau}}{4C_q}; \frac{1-\tau}{4(\mu_F+\mu_h)}\right\}$ and get

$$\mathbb{E}\left[V_{k+1}\right] \leq \left(1 - \gamma \cdot \frac{\mu_F + \mu_h}{16}\right) \cdot \mathbb{E}\left[V_k\right].$$

Running the recursion completes the proof.

A.4 Monotone/convex case ($\mu_h = 0$, $\mu_F = 0$)

We start from (15) with additional notation $\text{gap}(z^{k+1/2}, u) = \langle F(z^{k+1/2}), z^{k+1/2} - u\rangle + h(z^{k+1/2}) - h(u)$:

$$\begin{aligned}2\gamma \cdot \text{gap}(z^{k+1/2}, u) + \|z^{k+1} - u\|^2 \leq \tau\|z^k - u\|^2 + (1-\tau)\|w^k - u\|^2 \\ - \tau\|z^{k+1/2} - z^k\|^2 - (1-\tau)\|z^{k+1/2} - w^k\|^2 + 2\gamma^2\|g^{k+1/2} - g^k\|^2 \\ - 2\gamma\langle g^{k+1/2} - F(z^{k+1/2}), z^{k+1/2} - u\rangle.\end{aligned}$$

Adding both sides $\|w^{k+1} - u\|^2$ and making small rearrangement we have

$$\begin{aligned}2\gamma \cdot \text{gap}(z^{k+1/2}, u) \leq \left[\tau\|z^k - u\|^2 + \|w^k - u\|^2\right] - \left[\tau\|z^{k+1} - u\|^2 + \|w^{k+1} - u\|^2\right] \\ - \tau\|w^k - u\|^2 - (1-\tau)\|z^{k+1} - u\|^2 + \|w^{k+1} - u\|^2 \\ - \tau\|z^{k+1/2} - z^k\|^2 - (1-\tau)\|z^{k+1/2} - w^k\|^2 + 2\gamma^2\|g^{k+1/2} - g^k\|^2 \\ - 2\gamma\langle g^{k+1/2} - F(z^{k+1/2}), z^{k+1/2} - u\rangle.\end{aligned}$$

Then we sum up over $k = 0, \ldots, K-1$, take maximum of both sides over $z \in \mathcal{C}$, after take expectation and get

$$\begin{aligned}2\gamma \cdot \mathbb{E}\left[\max_{u\in\mathcal{C}} \sum_{k=0}^{K-1} \text{gap}(z^{k+1/2}, u)\right] \leq \max_{u\in\mathcal{C}}\left[\tau\|z^0 - u\|^2 + \|w^0 - u\|^2\right] \\ + \mathbb{E}\left[\max_{u\in\mathcal{C}} \sum_{k=0}^{K-1}\left[-\tau\|w^k - u\|^2 - (1-\tau)\|z^{k+1} - u\|^2 + \|w^{k+1} - u\|^2\right]\right] \\ - \sum_{k=0}^{K-1}\left[\tau\mathbb{E}\left[\|z^{k+1/2} - z^k\|^2\right] + (1-\tau)\mathbb{E}\left[\|z^{k+1/2} - w^k\|^2\right] - 2\gamma^2\mathbb{E}\left[\|g^{k+1/2} - g^k\|^2\right]\right] \\ + 2\gamma\mathbb{E}\left[\max_{u\in\mathcal{C}} \sum_{k=0}^{K-1}\left[\langle g^{k+1/2} - F(z^{k+1/2}), u - z^{k+1/2}\rangle\right]\right].\end{aligned}$$

Applying (17) for $\mathbb{E}\left[\|g^{k+1/2}-g^k\|^2\right]$, we get

$$
\begin{aligned}
2\gamma\cdot\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\text{gap}(z^{k+1/2},u)\right] &\leq \max_{u\in\mathcal{C}}\left[\tau\|z^0-u\|^2+\|w^0-u\|^2\right]\\
&+\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\left[-\tau\|w^k-u\|^2-(1-\tau)\|z^{k+1}-u\|^2+\|w^{k+1}-u\|^2\right]\right]\\
&-\sum_{k=0}^{K-1}\left[\tau\mathbb{E}\left[\|z^{k+1/2}-z^k\|^2\right]+(1-\tau)\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right]\right.\\
&\left.-2\gamma^2C_q^2\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right]\right]\\
&+2\gamma\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\left[\langle g^{k+1/2}-F(z^{k+1/2}),u-z^{k+1/2}\rangle\right]\right].
\end{aligned}
$$

With $\gamma\leq\frac{\sqrt{1-\tau}}{2C_q}$ we get

$$
\begin{aligned}
2\gamma\cdot\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\text{gap}(z^{k+1/2},u)\right] &\leq \max_{u\in\mathcal{C}}\left[\tau\|z^0-u\|^2+\|w^0-u\|^2\right]\\
&+\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\left[-\tau\|w^k-u\|^2-(1-\tau)\|z^{k+1}-u\|^2+\|w^{k+1}-u\|^2\right]\right]\\
&+2\gamma\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\left[\langle g^{k+1/2}-F(z^{k+1/2}),u-z^{k+1/2}\rangle\right]\right]. \qquad (20)
\end{aligned}
$$

To finish the proof we need to estimate terms in two last lines. We begin with $\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\langle F(z^{k+1/2})-g^{k+1/2},z^{k+1/2}-u\rangle\right]$. Let define sequence v: $v^0=z^0$, $v^{k+1}=\text{prox}_{\gamma h}(v^k-\gamma\delta_k)$ with $\delta^k=F(z^{k+1/2})-g^{k+1/2}$. Then we have

$$
\sum_{k=0}^{K-1}\langle\delta^k,z^{k+1/2}-u\rangle=\sum_{k=0}^{K-1}\langle\delta^k,z^{k+1/2}-v^k\rangle+\sum_{k=0}^{K-1}\langle\delta^k,v^k-u\rangle. \qquad (21)
$$

By the definition of v^{k+1} (property of prox), we have for all $z\in\mathcal{Z}$

$$
\langle v^{k+1}-v^k+\gamma\delta^k,z-v^{k+1}\rangle\geq 0.
$$

Rewriting this inequality, we get

$$
\begin{aligned}
\langle\gamma\delta^k,v^k-z\rangle &\leq \langle\gamma\delta^k,v^k-v^{k+1}\rangle+\langle v^{k+1}-v^k,z-v^{k+1}\rangle\\
&\leq \langle\gamma\delta^k,v^k-v^{k+1}\rangle+\frac{1}{2}\|v^k-z\|^2-\frac{1}{2}\|v^{k+1}-z\|^2-\frac{1}{2}\|v^k-v^{k+1}\|^2\\
&\leq \frac{\gamma^2}{2}\|\delta^k\|^2+\frac{1}{2}\|v^k-v^{k+1}\|^2+\frac{1}{2}\|v^k-z\|^2-\frac{1}{2}\|v^{k+1}-z\|^2-\frac{1}{2}\|v^k-v^{k+1}\|^2\\
&= \frac{\gamma^2}{2}\|\delta^k\|^2+\frac{1}{2}\|v^k-z\|^2-\frac{1}{2}\|v^{k+1}-z\|^2.
\end{aligned}
$$

With (21) it gives

$$
\begin{aligned}
\sum_{k=0}^{K-1}\langle\delta^k,z^{k+1/2}-u\rangle &\leq \sum_{k=0}^{K-1}\langle\delta^k,z^{k+1/2}-v^k\rangle+\frac{1}{\gamma}\sum_{k=0}^{K-1}\left(\frac{\gamma^2}{2}\|\delta^k\|^2+\frac{1}{2}\|v^k-u\|^2-\frac{1}{2}\|v^{k+1}-u\|^2\right)\\
&\leq \sum_{k=0}^{K-1}\langle\delta^k,z^{k+1/2}-v^k\rangle+\frac{\gamma}{2}\sum_{k=0}^{K-1}\|\delta^k\|^2+\frac{1}{2\gamma}\|v^0-u\|^2.
\end{aligned}
$$

We take the maximum on u and get

$$\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\langle\delta^k, z^{k+1/2}-u\rangle \leq \quad \sum_{k=0}^{K-1}\langle\delta^k, z^{k+1/2}-v^k\rangle + \frac{\gamma}{2}\sum_{k=0}^{K-1}\|F(z^{k+1/2})-g^{k+1/2}\|^2 + \frac{1}{2\gamma}\max_{u\in\mathcal{C}}\|v^0-u\|^2.$$

Taking the full expectation, we get

$$\begin{aligned}
\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\langle\delta^k, z^{k+1/2}-u\rangle\right] &\leq \mathbb{E}\left[\sum_{k=0}^{K-1}\langle\delta^k, z^{k+1/2}-v^k\rangle\right] \\
&\quad + \frac{\gamma}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\|F(z^{k+1/2})-g^{k+1/2}\|^2\right] + \frac{1}{2\gamma}\max_{u\in\mathcal{C}}\|v^0-u\|^2 \\
&= \mathbb{E}\left[\sum_{k=0}^{K-1}\langle\mathbb{E}\left[F(z^{k+1/2})-g^{k+1/2}\mid z^{k+1/2}-v^k\right], z^{k+1/2}-v^k\rangle\right] \\
&\quad + \frac{\gamma}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\|F(z^{k+1/2})-g^{k+1/2}\|^2\right] + \frac{1}{2\gamma}\max_{u\in\mathcal{C}}\|v^0-u\|^2 \\
&= \frac{\gamma}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\|F(z^{k+1/2})-g^{k+1/2}\|^2\right] + \frac{1}{2\gamma}\max_{u\in\mathcal{C}}\|v^0-u\|^2. \qquad (22)
\end{aligned}$$

Now let us estimate $\mathbb{E}\left[\max\limits_{u\in\mathcal{C}}\sum\limits_{k=0}^{K-1}\left[-\tau\|w^k-u\|^2-(1-\tau)\|z^{k+1}+u\|^2+\|w^{k+1}-u\|^2\right]\right]$, for this we note that

$$\begin{aligned}
&\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\left[-\tau\|w^k-u\|^2-(1-\tau)\|z^{k+1}-u\|^2+\|w^{k+1}-u\|^2\right]\right] \\
&= \mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\left[-2\langle(1-\tau)z^{k+1}+\tau w^k-w^{k+1}, u\rangle-(1-\tau)\|z^{k+1}\|^2-\tau\|w^k\|^2+\|w^{k+1}\|^2\right]\right] \\
&= \mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\left[-2\langle(1-\tau)z^{k+1}+\tau w^k-w^{k+1}, u\rangle\right]\right] \\
&\quad + \mathbb{E}\left[\sum_{k=0}^{K-1}-(1-\tau)\|z^{k+1}\|^2-\tau\|w^k\|^2+\|w^{k+1}\|^2\right].
\end{aligned}$$

One can note that by definition w^{k+1}: $\mathbb{E}\left[(1-\tau)\|z^{k+1}\|^2+\tau\|w^k\|^2-\|w^{k+1}\|^2\right]=0$, then

$$\begin{aligned}
&\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\left[-\tau\|w^k-u\|^2-(1-\tau)\|z^{k+1}-u\|^2+\|w^{k+1}-u\|^2\right]\right] \\
&\qquad = 2\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\langle(1-\tau)z^{k+1}+\tau w^k-w^{k+1}, -u\rangle\right] \\
&\qquad = 2\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\langle(1-\tau)z^{k+1}+\tau w^k-w^{k+1}, u\rangle\right].
\end{aligned}$$

Further, one can carry out the reasoning similarly to chain for (22):

$$\begin{aligned}
&\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\left[\tau\|w^k-u\|^2+(1-\tau)\|z^{k+1}-u\|^2-\|w^{k+1}-u\|^2\right]\right]\\
&\leq\sum_{k=0}^{K-1}\mathbb{E}\left[\|(1-\tau)z^{k+1}+\tau w^k-w^{k+1}\|^2\right]+\max_{u\in\mathcal{C}}\|v^0-u\|^2\\
&=\sum_{k=0}^{K-1}\mathbb{E}\left[\|\mathbb{E}_{w^{k+1}}[w^{k+1}]-w^{k+1}\|^2\right]+\max_{u\in\mathcal{C}}\|v^0-u\|^2\\
&=\sum_{k=0}^{K-1}\mathbb{E}\left[-\|\mathbb{E}_{w^{k+1}}[w^{k+1}]\|^2+\mathbb{E}_{w^{k+1}}\|w^{k+1}\|^2\right]+\max_{u\in\mathcal{C}}\|v^0-u\|^2\\
&=\sum_{k=0}^{K-1}\mathbb{E}\left[-\|(1-\tau)z^{k+1}+\tau w^k\|^2+(1-\tau)\|z^{k+1}\|^2+\tau\|w^k\|^2\right]+\max_{u\in\mathcal{C}}\|v^0-u\|^2\\
&=\sum_{k=0}^{K-1}\tau(1-\tau)\mathbb{E}\left[\|z^{k+1}-w^k\|^2\right]+\max_{u\in\mathcal{C}}\|v^0-u\|^2. \qquad (23)
\end{aligned}$$

Substituting (22) and (23) in (20) we get

$$\begin{aligned}
2\gamma\cdot\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\text{gap}(z^{k+1/2},u)\right]&\leq\max_{u\in\mathcal{C}}\left[(2+\tau)\|z^0-u\|^2+\|w^0-u\|^2\right]\\
&+\sum_{k=0}^{K-1}\left[\tau(1-\tau)\mathbb{E}\left[\|z^{k+1}-w^k\|^2\right]+\gamma^2\mathbb{E}\left[\|F(z^{k+1/2})-g^{k+1/2}\|^2\right]\right]. \qquad (24)
\end{aligned}$$

Next we work separately with $\mathbb{E}\left[\|F(z^{k+1/2})-g^{k+1/2}\|^2\right]$:

$$\begin{aligned}
&\mathbb{E}\left[\|F(z^{k+1/2})-g^{k+1/2}\|^2\right]\\
&=\mathbb{E}\left[\left\|Q^{\text{serv}}\left[\frac{1}{M}\sum_{m=1}^{M}Q_m^{\text{dev}}(F_m(z^{k+1/2})-F_m(w^k))\right]+F(w^k)-F(z^{k+1/2})\right\|^2\right]\\
&=\mathbb{E}\left[\left\|Q^{\text{serv}}\left[\frac{1}{M}\sum_{m=1}^{M}Q_m^{\text{dev}}(F_m(z^{k+1/2})-F_m(w^k))\right]\right\|^2\right]+\mathbb{E}\left[\|F(z^{k+1/2})-F(w^k)\|^2\right]\\
&\quad+\mathbb{E}\left[\langle Q^{\text{serv}}\left[\frac{1}{M}\sum_{m=1}^{M}Q_m^{\text{dev}}(F_m(z^{k+1/2})-F_m(w^k))\right];F(z^{k+1/2})-F(w^k)\rangle\right].
\end{aligned}$$

With (17) we get

$$\begin{aligned}
\mathbb{E}\left[\|F(z^{k+1/2})-g^{k+1/2}\|^2\right]&\leq C_q^2\mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right]+\mathbb{E}\left[\|F(z^{k+1/2})-F(w^k)\|^2\right]\\
&\quad+\mathbb{E}\left[\langle\frac{1}{M}\sum_{m=1}^{M}Q_m^{\text{dev}}(F_m(z^{k+1/2})-F_m(w^k));F(z^{k+1/2})-F(w^k)\rangle\right]\\
&=C_q^2\mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right]+2\mathbb{E}\left[\|F(z^{k+1/2})-F(w^k)\|^2\right]\\
&\leq C_q^2\mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right]+\frac{2}{M}\sum_{m=1}^{M}L_m^2\cdot\mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right]. \qquad (25)
\end{aligned}$$

With new notation $L^2 = \frac{1}{M}\sum\limits_{m=1}^{M} L_m^2$ from (24) and (25) we have

$$2\gamma\cdot\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\text{gap}(z^{k+1/2},u)\right] \leq \max_{u\in\mathcal{C}}\left[(2+\tau)\|z^0-u\|^2+\|w^0-u\|^2\right] + \sum_{k=0}^{K-1}\left[\tau(1-\tau)\mathbb{E}\left[\|z^{k+1}-w^k\|^2\right]+\gamma^2(C_q^2+2L^2)\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right]\right].$$

With $\gamma \leq \frac{\sqrt{1-\tau}}{2\sqrt{C_q^2+2L^2}}$ we deduce to

$$\begin{aligned}2\gamma\cdot\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\text{gap}(z^{k+1/2},u)\right] &\leq \max_{u\in\mathcal{C}}\left[(2+\tau)\|z^0-u\|^2+\|w^0-u\|^2\right] \\ &+ (1-\tau)\sum_{k=0}^{K-1}\left[\mathbb{E}\left[\|z^{k+1}-w^k\|^2\right]+\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right]\right] \\ &\leq \max_{u\in\mathcal{C}}\left[(2+\tau)\|z^0-u\|^2+\|w^0-u\|^2\right] \\ &+ 3(1-\tau)\sum_{k=0}^{K-1}\left[\mathbb{E}\left[\|z^{k+1}-z^{k+1/2}\|^2\right]+\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right]\right].\end{aligned}$$

Let us go back to (19) with $\mu_h = 0$, $\mu_F = 0$, $\gamma \leq \frac{\sqrt{1-\tau}}{4C_q}$ and get that

$$\begin{aligned}\mathbb{E}\left[V_{k+1}\right] &\leq \mathbb{E}\left[V_k\right] - \left((1-\tau)-2\gamma^2C_q^2\right)\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right] \\ &- \frac{1}{2}\mathbb{E}\left[\|z^{k+1}-z^{k+1/2}\|^2\right] \\ &\leq \mathbb{E}\left[V_k\right] - \frac{(1-\tau)}{2}\mathbb{E}\left[\|z^{k+1/2}-w^k\|^2\right] \\ &- \frac{(1-\tau)}{2}\mathbb{E}\left[\|z^{k+1}-z^{k+1/2}\|^2\right].\end{aligned}$$

Hence substituting this we go to the end of the proof:

$$\begin{aligned}2\gamma\cdot\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\text{gap}(z^{k+1/2},u)\right] &\leq \max_{u\in\mathcal{C}}\left[(2+\tau)\|z^0-u\|^2+\|w^0-u\|^2\right] + 6\sum_{k=0}^{K-1}\left[\mathbb{E}\left[V_k\right]-\mathbb{E}\left[V_{k+1}\right]\right] \\ &\leq \max_{u\in\mathcal{C}}\left[(2+7\tau)\|z^0-u\|^2+7\|w^0-u\|^2\right] \\ &\leq \max_{u\in\mathcal{C}}\left[16\|z^0-u\|^2\right].\end{aligned}$$

It remains to slightly correct the convergence criterion by monotonicity of F and Jensen's inequality for convex functions:

$$\begin{aligned}\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\text{gap}(z^{k+1/2},u)\right] &= \mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\left[\langle F(z^{k+1/2}),z^{k+1/2}-u\rangle+h(z^{k+1/2})-h(u)\right]\right] \\ &\geq \mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\left[\langle F(u),z^{k+1/2}-u\rangle+h(z^{k+1/2})-h(u)\right]\right] \\ &\geq \mathbb{E}\left[K\cdot\max_{u\in\mathcal{C}}\left[\langle F(u),\bar{z}^K-u\rangle+h(\bar{z}^K)-h(u)\right]\right] \\ &= K\cdot\mathbb{E}\left[\text{Gap}(\bar{z}^K)\right],\end{aligned}$$

where we additionally use $\bar{z}^K = \frac{1}{K}\sum\limits_{k=0}^{K-1} z^{k+1/2}$. This brings us to

$$\mathbb{E}\left[\text{Gap}(\bar{z}^K)\right] \leq \frac{8\max_{u\in\mathcal{C}}\left[\|z^0-u\|^2\right]}{\gamma K}.$$

Theorem 1 is completely proved for deterministic case.

□

A.5 Stochastic case

The case of a finite sum (stochastic) is proved in a similar way. We need to replace F_m with F_{m,π_m^k}.

B MASHA2: Handling Contractive Compressors

Now we present a method for working with biased compression operators – MASHA2. Algorithm 1 and Algorithm 2 are similar. The main and key difference is the error feedback technique Karimireddy et al. (2019), which is classic for working with biased compressors. To do this, we need to introduce additional sequences the sequence e_m. The purpose of these sequences is to accumulate error - something that is not communicated in previous iterations. Additionally, for Algorithm 2, we consider a simpler setting than for Algorithm 1, namely $\mathcal{Z} = \mathbb{R}^d$ and $h = 0$. In this case $\text{prox}_{\gamma h}(z) = z$. Also, in the case of comps, we compress the information in one direction - the server makes a full broadcast.

Similarly to Theorem 1 we use Lyapunov function

$$\hat{V}_k = \tau\|\hat{z}^k - z^*\|^2 + \|\hat{w}^k - z^*\|^2,$$

where $\hat{z}^k = z^k - \frac{1}{M}\sum\limits_{m=1}^{M} e_m^k$, $\hat{w}^k = w^k - \frac{1}{M}\sum\limits_{m=1}^{M} e_m^k$.

Theorem 2 *Let Assumptions 1 and 2 (SM) be satisfied. Then, if* $\gamma \leq \min\left(\frac{1-\tau}{4\mu_F}, \frac{\sqrt{1-\tau}}{60\delta L}, \frac{\mu_F(1-\tau)}{10^5\tau^2\delta^2L^2}\right)$, *the following estimates holds*

$$\mathbb{E}\left[\hat{V}_K\right] \leq \left(1-\gamma\cdot\tfrac{\mu_F}{4}\right)^{K-1}\cdot\hat{V}_0.$$

The proof of the above theorem can be found in Appendix C.

C Proof of Theorem 2

We first introduce useful notation:

$$\hat{z}^k = z^k - \frac{1}{M}\sum_{m=1}^{M} e_m^k, \quad \hat{z}^{k+1/2} = z^{k+1/2} - \frac{1}{M}\sum_{m=1}^{M} e_m^k, \quad \hat{w}^k = w^k - \frac{1}{M}\sum_{m=1}^{M} e_m^k.$$

It is easy to verify that

$$\begin{aligned}
\hat{z}^{k+1} &= z^{k+1} - \frac{1}{M}\sum_{m=1}^{M} e_m^{k+1} \\
&= z^{k+1/2} - \frac{1}{M}\sum_{m=1}^{M} C_m^{\text{dev}}(\gamma\cdot F_m(z^{k+1/2}) - \gamma\cdot F_m(w^k) + e_m^k) \\
&\quad - \frac{1}{M}\sum_{m=1}^{M}\left[e_m^k + \gamma\cdot F_m(z^{k+1/2}) - \gamma\cdot F_m(w^k) - C_m^{\text{dev}}(\gamma\cdot F_m(z^{k+1/2}) - \gamma\cdot F_m(w^k) + e_m^k)\right] \\
&= z^{k+1/2} - \frac{1}{M}\sum_{m=1}^{M} e_m^k - \gamma\cdot F(z^{k+1/2}) \\
&= \hat{z}^{k+1/2} - \gamma\cdot(F(z^{k+1/2}) - F(w^k)).
\end{aligned} \tag{26}$$

Algorithm 2 MASHA2 (handling contractive compressors)

Parameters: Stepsize $\gamma > 0$, number f iterations K.
Initialization: Choose $z^0 = w^0 \in \mathcal{Z}$, $e_m^0 = 0$.
Server sends to devices $z^0 = w^0$ and devices compute $F_m(w^0)$ and send to server and get $F(w^0)$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  for all devices in parallel **do**
    $\bar{z}^k = \tau z^k + (1-\tau) w^k$
    $z^{k+1/2} = \bar{z}_k - \gamma \cdot F(w^k)$
    Generate π_m^k from $\{1, \ldots, r\}$ independently
    Compute $F_m(z^{k+1/2})$ and send to server $C_m^{\text{dev}}(\gamma \cdot F_m(z^{k+1/2}) - \gamma \cdot F_m(w^k) + e_m^k)$
    $e_m^{k+1} = e_m^k + \gamma \cdot F_m(z^{k+1/2}) - \gamma \cdot F_m(w^k) - C_m^{\text{dev}}(\gamma \cdot F_m(z^{k+1/2}) - \gamma \cdot F_m(w^k) + e_m^k)$
    Compute $F_{m,\pi_m^k}(z^{k+1/2})$ and send to server $C_m^{\text{dev}}(\gamma \cdot F_{m,\pi_m^k}(z^{k+1/2}) - \gamma \cdot F_{m,\pi_m^k}(w^k) + e_m^k)$
    $e_m^{k+1} = e_m^k + \gamma \cdot F_{m,\pi_m^k}(z^{k+1/2}) - \gamma \cdot F_{m,\pi_m^k}(w^k) - C_m^{\text{dev}}(\gamma \cdot F_{m,\pi_m^k}(z^{k+1/2}) - \gamma \cdot F_{m,\pi_m^k}(w^k) + e_m^k)$
  end for
  for server **do**
    Compute $Q^{\text{serv}}\left[\frac{1}{M}\sum_{m=1}^M C_m^{\text{dev}}(\gamma F_m(z^{k+1/2}) - \gamma F_m(w^k) + e_m^k)\right]$ & send to devices
    Compute $Q^{\text{serv}}\left[\frac{1}{M}\sum_{m=1}^M C_m^{\text{dev}}(\gamma F_{m,\pi_m^k}(z^{k+1/2}) - \gamma F_{m,\pi_m^k}(w^k) + e_m^k)\right]$ & send to devices
    Sends to devices one bit b_k: 1 with probability $1-\tau$, 0 with with probability τ
  end for
  for all devices in parallel **do**
    $z^{k+1} = z^{k+1/2} - \frac{1}{M}\sum_{m=1}^M C_m^{\text{dev}}(\gamma \cdot F_m(z^{k+1/2}) - \gamma \cdot F_m(w^k) + e_m^k)$
    $z^{k+1} = z^{k+1/2} - \frac{1}{M}\sum_{m=1}^M C_m^{\text{dev}}(\gamma \cdot F_{m,\pi_m^k}(z^{k+1/2}) - \gamma \cdot F_{m,\pi_m^k}(w^k) + e_m^k)$
    if $b_k = 1$ **then**
      $w^{k+1} = z^{k+1}$
      Compute $F_m(w^{k+1})$ and it send to server; and get $F(w^{k+1})$
    else
      $w^{k+1} = w^k$
    end if
  end for
end for

Because of such a beautiful property for hat sequences, we will use them in the proof.

Proof of Theorem 2: We start from these two equalities:

$$\|\hat{z}^{k+1} - z^*\|^2 = \|z^{k+1/2} - z^*\|^2 + 2\langle \hat{z}^{k+1} - z^{k+1/2}, z^{k+1/2} - z^* \rangle + \|\hat{z}^{k+1} - z^{k+1/2}\|^2,$$

$$\|z^{k+1/2} - z^*\|^2 = \|\hat{z}^k - z^*\|^2 + 2\langle z^{k+1/2} - \hat{z}^k, z^{k+1/2} - z^* \rangle - \|z^{k+1/2} - \hat{z}^k\|^2.$$

Summing up, we obtain

$$\|\hat{z}^{k+1} - z^*\|^2 = \|\hat{z}^k - z^*\|^2 + 2\langle \hat{z}^{k+1} - \hat{z}^k, z^{k+1/2} - z^* \rangle + \|\hat{z}^{k+1} - z^{k+1/2}\|^2 - \|z^{k+1/2} - \hat{z}^k\|^2. \tag{27}$$

Using that $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and (26), we get

$$\begin{aligned}
\|\hat{z}^{k+1} - z^{k+1/2}\|^2 &\leq 2\|\hat{z}^{k+1} - \hat{z}^{k+1/2}\|^2 + 2\|\hat{z}^{k+1/2} - z^{k+1/2}\|^2 \\
&\leq 2\gamma^2 L^2 \cdot \|z^{k+1/2} - w^k\|^2 + \frac{2}{M}\sum_{m=1}^M \left\|e_m^k\right\|^2 \\
&\leq 2\gamma^2 L^2 \cdot \|z^{k+1/2} - w^k\|^2 + \frac{2}{M}\sum_{m=1}^M \left\|e_m^k\right\|^2. \qquad (28)
\end{aligned}$$

Additionally, here we use that F is L Lipschitz, where $L^2 = \frac{1}{M}\sum_{m=1}^{M} L_m^2$.

Next, (27) with (28) gives

$$\|\hat{z}^{k+1} - z^*\|^2 \leq \|\hat{z}^k - z^*\|^2 + 2\langle \hat{z}^{k+1} - \hat{z}^k, z^{k+1/2} - z^* \rangle + 2\gamma^2 L^2 \cdot \|z^{k+1/2} - w^k\|^2 + \frac{2}{M}\sum_{m=1}^{M} \left\|e_m^k\right\|^2 - \|z^{k+1/2} - \hat{z}^k\|^2. \tag{29}$$

Now we consider the next inner product $\langle \hat{z}^{k+1} - \hat{z}^k, z^{k+1/2} - z^* \rangle$. Using that

$$\begin{aligned}\hat{z}^{k+1} - \hat{z}^k &= \hat{z}^{k+1} - \hat{z}^{k+1/2} + \hat{z}^{k+1/2} - \hat{z}^k = -\gamma \cdot (F(z^{k+1/2}) - F(w^k)) + z^{k+1/2} - z^k \\ &= -\gamma \cdot F(z^{k+1/2}) + \bar{z}^k - z^k.\end{aligned}$$

and optimality condition $\langle F(z^*), z^{k+1/2} - z^* \rangle \geq 0$, we get

$$\begin{aligned}2\langle \hat{z}^{k+1} - \hat{z}^k, z^{k+1/2} - z^* \rangle &= 2\langle -\gamma \cdot F(z^{k+1/2}) + \bar{z}^k - z^k, z^{k+1/2} - z^* \rangle \\ &\leq 2\langle \gamma \cdot (F(z^*) - F(z^{k+1/2})), z^{k+1/2} - z^* \rangle \\ &\quad + 2\langle \bar{z}^k - z^k, z^{k+1/2} - z^* \rangle.\end{aligned}$$

C.1 Strongly monotone case

With Assumption 2 (SM) we obtain

$$\begin{aligned}2\langle \hat{z}^{k+1} - \hat{z}^k, z^{k+1/2} - z^* \rangle &\leq 2\langle \gamma \cdot (F(z^*) - F(z^{k+1/2})), z^{k+1/2} - z^* \rangle + 2\langle \bar{z}^k - z^k, z^{k+1/2} - z^* \rangle \\ &\leq -2\mu_F \gamma \|z^{k+1/2} - z^*\|^2 + 2(1-\tau)\langle \hat{w}^k - \hat{z}^k, z^{k+1/2} - z^* \rangle. \end{aligned} \tag{30}$$

Additionally we use here definition of $\bar{z}^k$ and fact that $\hat{w}^k - \hat{z}^k = w^k - z^k$. Combining the obtained inequalities (29) and (30), we have

$$\begin{aligned}\|\hat{z}^{k+1} - z^*\|^2 \leq \quad & \|\hat{z}^k - z^*\|^2 - 2\mu_F\gamma\|z^{k+1/2} - z^*\|^2 + 2(1-\tau)\langle \hat{w}^k - \hat{z}^k, z^{k+1/2} - z^* \rangle \\ & 2\gamma^2 L^2 \cdot \|z^{k+1/2} - w^k\|^2 + \frac{2}{M}\sum_{m=1}^{M} \left\|e_m^k\right\|^2 - \|z^{k+1/2} - \hat{z}^k\|^2.\end{aligned}$$

The inequality $2\langle a, b\rangle = \|a+b\|^2 - \|a\|^2 - \|b\|^2$ gives

$$\begin{aligned}\|\hat{z}^{k+1} - z^*\|^2 &\leq \|\hat{z}^k - z^*\|^2 - 2\mu_F\gamma\|z^{k+1/2} - z^*\|^2 \\ &\quad + 2(1-\tau)\langle \hat{w}^k - z^{k+1/2}, z^{k+1/2} - z^* \rangle \\ &\quad + 2(1-\tau)\langle z^{k+1/2} - \hat{z}^k, z^{k+1/2} - z^* \rangle \\ &\quad + 2\gamma^2 L^2 \cdot \|z^{k+1/2} - w^k\|^2 + \frac{2}{M}\sum_{m=1}^{M} \left\|e_m^k\right\|^2 - \|z^{k+1/2} - \hat{z}^k\|^2 \\ &= \|\hat{z}^k - z^*\|^2 - 2\mu_F\gamma\|z^{k+1/2} - z^*\|^2 \\ &\quad + (1-\tau)\|\hat{w}^k - z^*\|^2 - (1-\tau)\|\hat{w}^k - z^{k+1/2}\|^2 - (1-\tau)\|z^{k+1/2} - z^*\|^2 \\ &\quad + (1-\tau)\|z^{k+1/2} - \hat{z}^k\|^2 + (1-\tau)\|z^{k+1/2} - z^*\|^2 - (1-\tau)\|\hat{z}^k - z^*\|^2 \\ &\quad + 2\gamma^2 L^2 \cdot \|z^{k+1/2} - w^k\|^2 + \frac{2}{M}\sum_{m=1}^{M} \left\|e_m^k\right\|^2 - \|z^{k+1/2} - \hat{z}^k\|^2 \\ &= \tau\|\hat{z}^k - z^*\|^2 + (1-\tau)\|\hat{w}^k - z^*\|^2 - 2\mu_F\gamma\|z^{k+1/2} - z^*\|^2 - (1-\tau)\|\hat{w}^k - z^{k+1/2}\|^2 \\ &\quad + 4\gamma^2 L^2\|\hat{w}^k - z^{k+1/2}\|^2 + 4\gamma^2 L^2\|w^k - \hat{w}^k\|^2 + \frac{2}{M}\sum_{m=1}^{M} \left\|e_m^k\right\|^2 - \tau\|z^{k+1/2} - \hat{z}^k\|^2 \\ &= \tau\|\hat{z}^k - z^*\|^2 + (1-\tau)\|\hat{w}^k - z^*\|^2 - 2\mu_F\gamma\|z^{k+1/2} - z^*\|^2 \\ &\quad - (1-\tau-4\gamma^2L^2)\|\hat{w}^k - z^{k+1/2}\|^2 + (2+4\gamma^2L^2)\frac{1}{M}\sum_{m=1}^{M} \left\|e_m^k\right\|^2 - \tau\|z^{k+1/2} - \hat{z}^k\|^2.\end{aligned} \tag{31}$$

We can weight (31) by p^k and get

$$\sum_{k=0}^{K} p^k \|\hat{z}^{k+1} - z^*\|^2 \leq \tau \sum_{k=0}^{K} p^k \|\hat{z}^k - z^*\|^2 + (1-\tau) \sum_{k=0}^{K} p^k \|\hat{w}^k - z^*\|^2 - 2\mu\gamma \sum_{k=0}^{K} p^k \|z^{k+1/2} - z^*\|^2 \\ - (1 - \tau - 4\gamma^2 L^2) \sum_{k=0}^{K} p^k \|\hat{w}^k - z^{k+1/2}\|^2 \\ + (2 + 4\gamma^2 L^2) \cdot \sum_{k=0}^{K} p^k \frac{1}{M} \sum_{m=1}^{M} \left\|e_m^k\right\|^2 - \tau \sum_{k=0}^{K} p^k \|z^{k+1/2} - \hat{z}^k\|^2. \quad (32)$$

Next we will take an full expectation from the both side of previous inequality. Since w^{k+1} is chosen,

$$\mathbb{E}\left[\|\hat{w}^{k+1} - z^*\|^2\right] = \tau \mathbb{E}\left[\left\|\hat{w}^k + \frac{1}{M}\sum_{m=1}^{M}(e_m^{k+1} - e_m^k) - z^*\right\|^2\right] + (1-\tau)\mathbb{E}\left[\|\hat{z}^{k+1} - z^*\|^2\right] \\ \leq (1+\eta)\tau \mathbb{E}\left[\left\|\hat{w}^k - z^*\right\|^2\right] + (1 + 1/\eta)\tau \mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}(e_m^{k+1} - e_m^k)\right\|^2\right] \\ + (1-\tau)\mathbb{E}\left[\|\hat{z}^{k+1} - z^*\|^2\right],$$

with weighs it gives

$$\sum_{k=0}^{K} p^k \mathbb{E}\left[\|\hat{w}^{k+1} - z^*\|^2\right] \leq (1+\eta)\tau \sum_{k=0}^{K} p^k \mathbb{E}\left[\left\|\hat{w}^k - z^*\right\|^2\right] + (1-\tau)\sum_{k=0}^{K} p^k \mathbb{E}\left[\|\hat{z}^{k+1} - z^*\|^2\right] \\ + (1 + 1/\eta)\tau \sum_{k=0}^{K} p^k \mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}(e_m^{k+1} - e_m^k)\right\|^2\right] \\ \leq (1+\eta)\tau \sum_{k=0}^{K} p^k \mathbb{E}\left[\left\|\hat{w}^k - z^*\right\|^2\right] + (1-\tau)\sum_{k=0}^{K} p^k \mathbb{E}\left[\|\hat{z}^{k+1} - z^*\|^2\right] \\ + 2(1 + 1/\eta)\tau \sum_{k=0}^{K} p^k \cdot \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[\left\|e_m^{k+1}\right\|^2\right] \\ + 2(1 + 1/\eta)\tau \sum_{k=0}^{K} p^k \cdot \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[\left\|e_m^{k}\right\|^2\right]$$

Summing this one and (32), we get

$$
\begin{aligned}
&\sum_{k=0}^{K} p^k \left(\tau\|\hat{z}^{k+1} - z^*\|^2 + \|\hat{w}^{k+1} - z^*\|^2\right) \\
&\quad\leq \sum_{k=0}^{K} p^k \left(\tau\|\hat{z}^{k} - z^*\|^2 + (1+\tau\eta)\|\hat{w}^{k} - z^*\|^2\right) - 2\mu_F\gamma \sum_{k=0}^{K} p^k \|z^{k+1/2} - z^*\|^2 \\
&\qquad - (1 - \tau - 4\gamma^2 L^2) \sum_{k=0}^{K} p^k \|\hat{w}^k - z^{k+1/2}\|^2 + \frac{2(1+1/\eta)\tau}{p} \sum_{k=1}^{K+1} p^k \frac{1}{M} \sum_{m=1}^{M} \left\|e_m^k\right\|^2 \\
&\qquad + (2 + 4\gamma^2 L^2 + 2(1+1/\eta)\tau) \cdot \sum_{k=0}^{K} p^k \frac{1}{M} \sum_{m=1}^{M} \left\|e_m^k\right\|^2 - \tau \sum_{k=0}^{K} p^k \|z^{k+1/2} - \hat{z}^k\|^2 \\
&\quad\leq \sum_{k=0}^{K} p^k \left(\tau\|\hat{z}^{k} - z^*\|^2 + (1+\tau\eta)\|\hat{w}^{k} - z^*\|^2\right) - 2\mu_F\gamma \sum_{k=0}^{K} p^k \|z^{k+1/2} - z^*\|^2 \\
&\qquad - (1 - \tau - 4\gamma^2 L^2) \sum_{k=0}^{K} p^k \|\hat{w}^k - z^{k+1/2}\|^2 - \tau \sum_{k=0}^{K} p^k \|z^{k+1/2} - \hat{z}^k\|^2 \\
&\qquad + \left(2 + 4\gamma^2 L^2 + 2(1+1/\eta)\tau + \frac{2(1+1/\eta)\tau}{p}\right) \cdot \sum_{k=0}^{K} p^k \frac{1}{M} \sum_{m=1}^{M} \left\|e_m^k\right\|^2.
\end{aligned}
$$

Next we will estimate error term:

$$
\begin{aligned}
\frac{1}{M}\sum_{m=1}^{M} \left\|e_m^{k+1}\right\|^2 &= \frac{1}{M}\sum_{m=1}^{M} \left\|e_m^k + \gamma \cdot F_m(z^{k+1/2}) - \gamma \cdot F_m(w^k) - C_m^{\text{dev}}(\gamma \cdot F_m(z^{k+1/2}) - \gamma \cdot F_m(w^k) + e_m^k)\right\|^2 \\
&\leq \frac{1}{M}\sum_{m=1}^{M} \left\|e_m^k + \gamma \cdot F_m(z^{k+1/2}) - \gamma \cdot F_m(w^k) - C_m^{\text{dev}}(\gamma \cdot F_m(z^{k+1/2}) - \gamma \cdot F_m(w^k) + e_m^k)\right\|^2 \\
&\leq \frac{(1-1/\delta)}{M}\sum_{m=1}^{M} \left\|e_m^k + \gamma \cdot F_m(z^{k+1/2}) - \gamma \cdot F_m(w^k)\right\|^2 \\
&\leq \frac{(1-1/\delta)}{M}\sum_{m=1}^{M} (1+\xi)\left\|e_m^k\right\|^2 + (1+1/\xi)\gamma^2 \cdot \left\|F_m(z^{k+1/2}) - F_m(w^k)\right\|^2.
\end{aligned}
$$

Here we use definition of biased compression and inequality $\|a+b\|^2 \leq (1+\xi)\|a\|^2 + (1+1/\xi)\|b\|^2$ (for $\xi > 0$). With $\xi = \frac{1}{2(\delta-1)}$ and $\gamma \leq \frac{1}{4\delta L}$, we get

$$
\begin{aligned}
\frac{1}{M}\sum_{m=1}^{M} \left\|e_m^{k+1}\right\|^2 &\leq \frac{1}{M}\sum_{m=1}^{M} (1-1/2\delta)\left\|e_m^k\right\|^2 + 2\delta\gamma^2 \cdot \left\|F_m(z^{k+1/2}) - F_m(w^k)\right\|^2 \\
&\leq (1-1/2\delta)\frac{1}{M}\sum_{m=1}^{M} \left\|e_m^k\right\|^2 + 2\delta\gamma^2 L^2 \cdot \left\|z^{k+1/2} - w^k\right\|^2 \\
&\leq (1-1/2\delta)\frac{1}{M}\sum_{m=1}^{M} \left\|e_m^k\right\|^2 + 4\delta\gamma^2 L^2 \cdot \left\|z^{k+1/2} - \hat{w}^k\right\|^2 + 4\delta\gamma^2 L^2 \cdot \frac{1}{M}\sum_{m=1}^{M} \left\|e_m^k\right\|^2 \\
&\leq (1-1/2\delta+4\delta\gamma^2 L^2)\frac{1}{M}\sum_{m=1}^{M} \left\|e_m^k\right\|^2 + 4\delta\gamma^2 L^2 \cdot \left\|z^{k+1/2} - \hat{w}^k\right\|^2 \\
&\leq (1-1/4\delta)\frac{1}{M}\sum_{m=1}^{M} \left\|e_m^k\right\|^2 + 4\delta\gamma^2 L^2 \cdot \left\|z^{k+1/2} - \hat{w}^k\right\|^2 \\
&\leq 4\delta\gamma^2 L^2 \sum_{j=0}^{k} (1-1/4\delta)^{k-j} \cdot \left\|z^{j+1/2} - \hat{w}^j\right\|^2.
\end{aligned}
$$

We weigh the sequence as follows $\sum\limits_{k=0}^{K} p^k \cdot \frac{1}{M} \sum\limits_{m=1}^{M} \left\|e_m^k\right\|^2$, where p such as $p^k \leq p^j(1+1/8\delta)^{k-j}$. Then

$$\begin{aligned}
\sum_{k=0}^{K} p^k \cdot \frac{1}{M} \sum_{m=1}^{M} \left\|e_m^k\right\|^2 &\leq 4\delta\gamma^2 L^2 \sum_{k=0}^{K} p^k \sum_{j=0}^{k-1} (1-1/4\delta)^{k-j-1} \cdot \left\|z^{j+1/2} - \hat{w}^j\right\|^2 \\
&\leq \frac{4\delta\gamma^2 L^2}{(1-1/4\delta)} \sum_{k=0}^{K} \sum_{j=0}^{k-1} p^j (1+1/8\delta)^{k-j} (1-1/4\delta)^{k-j} \cdot \left\|z^{j+1/2} - \hat{w}^j\right\|^2 \\
&\leq \frac{4\delta\gamma^2 L^2}{(1-1/4\delta)} \sum_{k=0}^{K} \sum_{j=0}^{k-1} p^j (1-1/8\delta)^{k-j} \cdot \left\|z^{j+1/2} - \hat{w}^j\right\|^2 \\
&\leq \frac{4\delta\gamma^2 L^2}{(1-1/4\delta)} \sum_{k=0}^{K} p^k \cdot \left\|z^{k+1/2} - \hat{w}^k\right\|^2 \cdot \sum_{j=0}^{\infty} (1-1/8\delta)^j \\
&\leq 128\delta^2\gamma^2 L^2 \sum_{k=0}^{K} p^k \cdot \left\|z^{k+1/2} - \hat{w}^k\right\|^2. \qquad (33)
\end{aligned}$$

(33) us to the finish line of proof:

$$\begin{aligned}
&\sum_{k=0}^{K} p^k \left(\tau\|\hat{z}^{k+1} - z^*\|^2 + \|\hat{w}^{k+1} - z^*\|^2\right) \\
&\quad \leq \sum_{k=0}^{K} p^k \left(\tau\|\hat{z}^k - z^*\|^2 + (1+\tau\eta)\|\hat{w}^k - z^*\|^2\right) - 2\mu_F\gamma \sum_{k=0}^{K} p^k \|z^{k+1/2} - z^*\|^2 \\
&\quad\quad - (1 - \tau - 4\gamma^2 L^2) \sum_{k=0}^{K} p^k \|\hat{w}^k - z^{k+1/2}\|^2 - \tau \sum_{k=0}^{K} p^k \|z^{k+1/2} - \hat{z}^k\|^2 \\
&\quad\quad + \left(2 + 4\gamma^2 L^2 + 2(1+1/\eta)\tau + \frac{2(1+1/\eta)\tau}{p}\right) \cdot 128\delta^2\gamma^2 L^2 \sum_{k=0}^{K} p^k \cdot \left\|z^{k+1/2} - \hat{w}^k\right\|^2.
\end{aligned}$$

With $\eta = \frac{\mu_F\gamma}{4\tau}$, $\gamma \leq \frac{1}{\mu_F}$ and $p \geq 1$ we have

$$\begin{aligned}
&\sum_{k=0}^{K} p^k \left(\tau\|\hat{z}^{k+1} - z^*\|^2 + \|\hat{w}^{k+1} - z^*\|^2\right) \\
&\quad \leq (1 - \mu_F\gamma/4) \sum_{k=0}^{K} p^k \left(\tau\|\hat{z}^k - z^*\|^2 + \|\hat{w}^k - z^*\|^2\right) \\
&\quad\quad - \left(1 - \tau - \mu_F\gamma - 800\delta^2\gamma^2 L^2 - 512\delta^2\gamma^4 L^4 \delta^2 - \frac{2048\delta^2\tau^2\gamma L^2}{\mu_F}\right) \cdot \sum_{k=0}^{K} p^k \cdot \left\|z^{k+1/2} - \hat{w}^k\right\|^2.
\end{aligned}$$

Choice $\gamma \leq \min\left(\frac{1-\tau}{4\mu_F}, \frac{\sqrt{1-\tau}}{60\delta L}, \frac{\mu_F(1-\tau)}{10^5\tau^2\delta^2 L^2}\right)$ gives

$$\sum_{k=0}^{K} p^k \left(\tau\|\hat{z}^{k+1} - z^*\|^2 + \|\hat{w}^{k+1} - z^*\|^2\right) \leq (1 - \mu_F\gamma/4) \sum_{k=0}^{K} p^k \left(\tau\|\hat{z}^k - z^*\|^2 + \|\hat{w}^k - z^*\|^2\right).$$

Then we just need to take $p = 1/(1 - \mu_F\gamma/4)$ (easy to check that $p^k \leq p^j(1+1/8\delta)^{k-j}$ works) and get

$$\left(\tau\|\hat{z}^{K+1} - z^*\|^2 + \|\hat{w}^{K+1} - z^*\|^2\right) \leq (1 - \mu_F\gamma/4)^{K+1} \left(\tau\|\hat{z}^0 - z^*\|^2 + \|\hat{w}^0 - z^*\|^2\right).$$

This ends our proof.

□

D MOTIVATING EXAMPLES

Let us motivate the utility of considering VIs in machine learning on a handful of examples.

Lagrangian multipliers and SVM. Lagrange multipliers are a standard approach to solving constrained optimization problems. This technique reduces the original problem to a saddle point problem. This approach is one of the basic and classic for SVM (Shalev-Shwartz & Ben-David, 2014):

$$\min_{w,b} \max_{\lambda} \frac{1}{N} \sum_{n=1}^{N} \lambda_n (y_n (\langle w, x_n \rangle + b) - 1) + \frac{\beta}{2} \|w\|^2, \tag{34}$$

where w are the weights of the model, b – some number, $\{(x_n, y_n)\}_{n=1}^N$ are pairs of the training data and labels, and $\beta \geq 0$ is a regularization parameter.

GANs. A simple GAN setup consists of two parts: the discriminator D aimed at distinguishing real samples x from adversarial ones by giving probability that a sample is real, and the generator G trying to fool the discriminator by generating realistic samples from random noise z. Following Goodfellow et al. (2014), the value function $V(G, D)$ used in such a minimax game can be expressed in a saddle point form as

$$\min_{G} \max_{D} V(D, G) := \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})} [\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})} [\log(1 - D(G(\boldsymbol{z})))]. \tag{35}$$

Adversarial loss. To force a model to be more stable and robust, it can be trained in a constructive way, for example, by introducing adversarial noise (Madry et al., 2017; Nouiehed et al., 2019). For example, the approach of Liu et al. (2020); Zhu et al. (2019) works well in NLP. From the point of view of theory, this latter approach reduces to the saddle point problem

$$\min_{w} \max_{\|\rho_1\| \leq e, \ldots \|\rho_N\| \leq e} \frac{1}{N} \sum_{n=1}^{N} l(f(w, x_n + \rho_n, y_n)^2 + \frac{\lambda}{2} \|w\|^2 - \frac{\beta}{2} \|\rho\|^2, \tag{36}$$

where w are the weights of the model, $\{(x_n, y_n)\}_{n=1}^N$ are pairs of the training data and labels, ρ is the so-called adversarial noise which introduces a perturbation in the data, and $\lambda > 0$ and $\beta > 0$ are the regularization parameters. The main difference from a standard approach is in explicit training of ρ so that the noise from it is harmful, and for w to adapt to this noise.

Online transport and Wasserstein barycenters Online transport or Wasserstein Barycenter (WB) problem can be rewritten as a saddle point problem (Dvinskikh & Tiapkin, 2021). This representation comes from the dual view on the transportation polytope.