SUPPLEMENTARY MATERIAL

Our unified analysis of ASYNC-RED is based on the monotone operator theory (Ryu & Boyd, 2016). In Supplement A, we first clarify our setting for the access of the shared memory. In Supplement B, we present the proof of Theorem 1 and Theorem 2, proving the fixed-point convergence of ASYNC-RED to $\mathsf{zer}(\mathsf{G})$ in both batch and stochastic settings. In Supplement C, we provide a brief review of the related knowledge on monotone operators. In Supplement D, we include additional technical details and experiments omitted from the main paper due to space.

A MEMORY ACCESS WITHOUT GLOBAL LOCK

In the setting of ASYNC-RED, multiple cores may simultaneously read and update the blocks $\boldsymbol{x}_i$ in shared memory. We coordinate the memory access of different cores by imposing certain *local* locks. For example, consider one work cycle of core c_i for updating the block $\boldsymbol{x}_i$. First, a local *read* lock is imposed to $\boldsymbol{x}_i$ such that only read operations (by c_i or others) can be performed on $\boldsymbol{x}_i$. If, at the same time, other cores want to write $\boldsymbol{x}_i$, then they have to wait until the read lock is released by the last one who finishes reading the block. However, if they want to write other blocks, their operations will not be blocked. Secondly, core c_i evaluates the RED update on $\boldsymbol{x}_i$, while other cores continuously update $\boldsymbol{x}$. Here, we assume that the number of updates by cores other than c_i is bounded by some positive integer, which is exactly what Assumption 1 refers to. After the evaluation finishes, core c_i imposes a local *write* lock, which prevents both read and write by other cores, on $\boldsymbol{x}_i$ and write the block with the computed update. Similarly, other cores have to wait until the lock is released before operating on $\boldsymbol{x}_i$. Finally, when the update finishes, the local lock will be released and core c_i will restart a new cycle. Note that $\boldsymbol{x}$ is never locked *globally* during the full update cycle, and the reads of each block are always consistent.

In order to ensure the consistent read of $\boldsymbol{x}$, we leverage the dual-memory strategy for block coordinate settings proposed in (Peng et al., 2016) (see section 1.2.1 *'Block coordinate'*). Its key idea is that, before every write to a block $\boldsymbol{x}_i$, a copy of the old version of the block is kept for reading. In this way, there always exists some state of $\boldsymbol{x}$ in the memory for the cores to access.

B PROOF OF ANALYSIS

In this section, we first present the proof of Theorem 1, then followed by the proof of Theorem 2. For a review of monotone operators, we refer to Supplement C.

Throughout the proof, we consider the probability space $(\Omega, \mathcal{F}, P)$, where Ω denotes the sample space, $\mathcal{F}$ the σ-algebra, and P the probability measure. $\boldsymbol{x}^k$ is a random variable defined in $\mathbb{R}^n$. We use $\|\cdot\|$ to denote the ℓ_2-norm. We define the sequence of sub σ-algebra $\{\mathcal{X}^k\}_{k\in\mathbb{N}}$ of $\mathcal{F}$ as

$$\mathcal{X}^k \coloneqq \sigma(\boldsymbol{x}^0, ..., \boldsymbol{x}^k, \Delta_0, ..., \Delta_k),$$

where σ generates the filtration (smallest σ-algebra) from $\boldsymbol{x}^0, ..., \boldsymbol{x}^k$, and $\Delta_0, ..., \Delta_k$. Note that the sequence $\{\mathcal{X}^k\}_{k\in\mathbb{N}}$ is such that $\mathcal{X}^k \subset \mathcal{X}^{k+1}$ for any $k \in \mathbb{N}$. We use $\boldsymbol{x}^*$ to denote some fixed point in the set $\mathsf{zer}(\mathsf{G})$.

B.1 PROOF OF THEOREM 1

Our proof needs the following lemma on the RED operator.

Lemma 1. *Let Assumption 3 and 4 hold for g and D_σ. The composite operator G is $1/(L+2\tau)$-cocoercive, that is*

$$(\mathsf{G}(\boldsymbol{x}) - \mathsf{G}(\boldsymbol{y}))^\mathsf{T}(\boldsymbol{x} - \boldsymbol{y}) \geq \frac{1}{L+2\tau}\|\mathsf{G}(\boldsymbol{x}) - \mathsf{G}(\boldsymbol{y})\|^2.$$

Proof. This lemma is adapted from Lemma 3 in (Sun et al., 2019a). Consider the following decomposition

$$\mathsf{I} - \frac{2}{L+2\tau}\mathsf{G} = \left(\frac{2}{L+2\tau}\cdot\frac{L}{2}\right)\left[\mathsf{I} - \frac{2}{L}\nabla g\right] + \left(\frac{2}{L+2\tau}\cdot\frac{2\tau}{2}\right)\left[\mathsf{I} - \frac{1}{\tau}\mathsf{H}\right], \tag{16}$$

where we recall $\mathsf{H} = \tau(\mathsf{I} - \mathsf{D}_\sigma)$. According to Assumption 3, g is convex and ∇g is L-Lipschitz continuous. By Proposition 1 in Supplement C, ∇g is $1/L$-cocoercive. Hence, by Proposition 2 in Supplement C, $\mathsf{I} - (2/L)\nabla g$ is nonexpansive. Since $\mathsf{D}_\sigma = \mathsf{I} - (1/\tau)\mathsf{H}$, this means that $\mathsf{I} - (1/\tau)\mathsf{H}$ is nonexpansive. From Proposition 3 in Supplement C, we know that the convex combination of two nonexpansive operators is nonexpansive. Thus, $\mathsf{I} - (2/(L+2\tau))\mathsf{G}$ is nonexpansive, which also means that G is $1/(L+2\tau)$-cocoercive according to Proposition 2 in Supplement C. □

Now we can start the main proof. Under the fixed stepsize $\gamma > 0$, we begin with the following equations regarding the fixed point $\boldsymbol{x}^* \in \mathsf{zer}(\mathsf{G})$

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2 | \mathcal{X}^k\right] \\
&= \mathbb{E}\left[\|\boldsymbol{x}^k - \gamma \mathsf{G}_i(\widetilde{\boldsymbol{x}}^k) - \boldsymbol{x}^*\|^2 | \mathcal{X}^k\right] \\
&= \mathbb{E}\left[\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 | \mathcal{X}^k\right] + \gamma^2 \mathbb{E}\left[\|\mathsf{G}_i(\widetilde{\boldsymbol{x}}^k)\|^2 | \mathcal{X}^k\right] + 2\gamma \mathbb{E}\left[(\mathsf{G}_i(\widetilde{\boldsymbol{x}}^k))^\mathsf{T}(\boldsymbol{x}^* - \boldsymbol{x}^k) | \mathcal{X}^k\right]
\end{aligned} \tag{17}$$

Since $\mathsf{G}_i : \mathbb{R}^n \to \mathbb{R}^n$ is evaluated on a random block of $\boldsymbol{x}_i$, we have the following conditional expectations

$$\mathbb{E}\left[(\mathsf{G}_i(\widetilde{\boldsymbol{x}}^k))^\mathsf{T}(\boldsymbol{x}^* - \boldsymbol{x}^k) | \mathcal{X}^k\right] = \frac{1}{b}\sum_{i=1}^{b}(\mathsf{G}_i(\widetilde{\boldsymbol{x}}^k))^\mathsf{T}(\boldsymbol{x}^* - \boldsymbol{x}^k) = \frac{1}{b}(\mathsf{G}(\widetilde{\boldsymbol{x}}^k))^\mathsf{T}(\boldsymbol{x}^* - \boldsymbol{x}^k) \tag{18}$$

and

$$\mathbb{E}\left[\|\mathsf{G}_i(\widetilde{\boldsymbol{x}}^k)\|^2 | \mathcal{X}^k\right] = \frac{1}{b}\sum_{i=1}^{b}\|\mathsf{G}_i(\widetilde{\boldsymbol{x}}^k)\|^2 = \frac{1}{b}\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2. \tag{19}$$

Thus, plugging the above results into (17)

$$\mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2 | \mathcal{X}^k\right] \leq \|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 + \frac{\gamma^2}{b}\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2 + \underbrace{\frac{2\gamma}{b}(\mathsf{G}(\widetilde{\boldsymbol{x}}^k))^\mathsf{T}(\boldsymbol{x}^* - \boldsymbol{x}^k)}_{(\dagger)}. \tag{20}$$

The term $(\dagger)$ can be expressed as

$$\begin{aligned}
&\frac{2\gamma}{b}(\mathsf{G}(\widetilde{\boldsymbol{x}}^k))^\mathsf{T}(\boldsymbol{x}^* - \boldsymbol{x}^k) \\
&= \frac{2\gamma}{b}(\mathsf{G}(\widetilde{\boldsymbol{x}}^k))^\mathsf{T}(\boldsymbol{x}^* - \widetilde{\boldsymbol{x}}^k + \sum_{s=k-\Delta_k}^{k-1}(\boldsymbol{x}^s - \boldsymbol{x}^{s+1})) \\
&= \frac{2\gamma}{b}(\mathsf{G}(\widetilde{\boldsymbol{x}}^k) - \mathsf{G}(\boldsymbol{x}^*))^\mathsf{T}(\boldsymbol{x}^* - \widetilde{\boldsymbol{x}}^k) + \frac{2\gamma}{b}(\mathsf{G}(\widetilde{\boldsymbol{x}}^k))^\mathsf{T}(\sum_{s=k-\Delta_k}^{k-1}(\boldsymbol{x}^s - \boldsymbol{x}^{s+1})) \\
&= \frac{2\gamma}{b}(\mathsf{G}(\widetilde{\boldsymbol{x}}^k) - \mathsf{G}(\boldsymbol{x}^*))^\mathsf{T}(\boldsymbol{x}^* - \widetilde{\boldsymbol{x}}^k) + \frac{2\gamma^2}{b}\sum_{s=k-\Delta_k}^{k-1}\mathsf{G}(\widetilde{\boldsymbol{x}}^k)^\mathsf{T}\mathsf{G}_{i_s}(\widetilde{\boldsymbol{x}}^s),
\end{aligned} \tag{21}$$

where in the second line we used the definition of the stale iterate $\boldsymbol{x}^{s+1} = \boldsymbol{x}^s - \gamma \mathsf{G}_{i_s}(\widetilde{\boldsymbol{x}}^k)$, and in the third line the fact that $\mathsf{G}(\boldsymbol{x}^*) = \boldsymbol{0}$. By using Lemma 1, we obtain the upper bound for the first term in equation (21)

$$\frac{2\gamma}{b}(\mathsf{G}(\widetilde{\boldsymbol{x}}^k) - \mathsf{G}(\boldsymbol{x}^*))^\mathsf{T}(\boldsymbol{x}^* - \widetilde{\boldsymbol{x}}^k) \leq -\frac{2\gamma\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2}{b(L+2\tau)}. \tag{22}$$

For the second term in (21), we have

$$\begin{aligned}
\frac{2\gamma^2}{b}\sum_{s=k-\Delta_k}^{k-1}\mathsf{G}(\widetilde{\boldsymbol{x}}^k)^\mathsf{T}\mathsf{G}_{i_s}(\widetilde{\boldsymbol{x}}^s) &\leq \frac{\lambda\gamma^2\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2}{b} + \sum_{s=k-\Delta_k}^{k-1}\frac{\gamma^2\|\mathsf{G}_{i_s}(\widetilde{\boldsymbol{x}}^s)\|^2}{b}, \\
&\leq \frac{\lambda\gamma^2\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2}{b} + \sum_{s=k-\lambda}^{k-1}\frac{\gamma^2\|\mathsf{G}(\widetilde{\boldsymbol{x}}^s)\|^2}{b},
\end{aligned} \tag{23}$$

where in the first inequality we used the *Young's inequality*

$$\boldsymbol{x}_1^{\mathsf{T}}\boldsymbol{x}_2 \leq \frac{1}{2}\left[\|\boldsymbol{x}_1\|^2 + \|\boldsymbol{x}_2\|^2\right], \tag{24}$$

and in the second inequality we use

$$\sum_{s=k-\Delta k}^{k-1} \gamma^2 \|\mathsf{G}_{i_s}(\widetilde{\boldsymbol{x}}^s)\|^2 = \sum_{s=k-\Delta k}^{k-1} \|\boldsymbol{x}^s - \boldsymbol{x}^{s+1}\|_2^2 \leq \sum_{s=k-\lambda}^{k-1} \|\boldsymbol{x}^s - \boldsymbol{x}^{s+1}\|_2^2 = \sum_{s=k-\lambda}^{k-1} \gamma^2 \|\mathsf{G}(\widetilde{\boldsymbol{x}}^s)\|^2.$$

Applying (22) and (23) in (21) yields the overall upper bound for the term (†)

$$\frac{2\gamma}{b}(\mathsf{G}(\widetilde{\boldsymbol{x}}^k))^{\mathsf{T}}(\boldsymbol{x}^* - \boldsymbol{x}^k) \leq \frac{(L+2\tau)\lambda\gamma^2 - 2\gamma}{(L+2\tau)b}\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2 + \sum_{s=k-\lambda}^{k-1} \frac{\gamma^2\|\mathsf{G}(\widetilde{\boldsymbol{x}}^s)\|^2}{b}. \tag{25}$$

Next, by plugging (25) into (17) and re-arranging the terms, we obtain the following inequality

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2 | \mathcal{X}^k\right] \\
&\leq \|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 + \sum_{s=k-\lambda}^{k-1} \frac{\gamma^2\|\mathsf{G}(\widetilde{\boldsymbol{x}}^s)\|^2}{b} + \frac{(L+2\tau)(1+\lambda)\gamma^2 - 2\gamma}{(L+2\tau)b}\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2.
\end{aligned} \tag{26}$$

Taking the total expectation of equation (26) and re-arranging the terms yields that

$$\begin{aligned}
&\frac{2\gamma - (L+2\tau)(1+\lambda)\gamma^2}{(L+2\tau)b}\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right] \\
&\leq \mathbb{E}\left[\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2\right] - \mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2\right] + \gamma^2 \sum_{s=k-\lambda}^{k-1} \frac{\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^s)\|^2\right]}{b}
\end{aligned} \tag{27}$$

We then telescope-sum equation (27) over $t > 0$ iterations to have

$$\begin{aligned}
&\sum_{k=0}^{t-1} \frac{2\gamma - (L+2\tau)(1+\lambda)\gamma^2}{(L+2\tau)b}\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right] \\
&\leq \mathbb{E}\left[\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2\right] - \mathbb{E}\left[\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2\right] + \gamma^2 \sum_{k=0}^{t-1}\sum_{s=k-\lambda}^{k-1} \frac{\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^s)\|^2\right]}{b}
\end{aligned} \tag{28}$$

where the index s always start at 0. Under the assumption of consistent read, it is true that

$$\sum_{k=0}^{t-1}\sum_{s=k-\lambda}^{k-1} \frac{\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^s)\|^2\right]}{b} \leq \lambda \sum_{k=0}^{t-1} \frac{\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right]}{b}. \tag{29}$$

In the case of inconsistent read, the above inequality does not always hold. We refer to Peng et al. (2016) for a comprehensive analysis for asynchronous block-coordinate methods with inconsistent reads. Now, we rewrite equation (28) as

$$\sum_{k=0}^{t-1} \frac{2\gamma - (L+2\tau)(1+2\lambda)\gamma^2}{(L+2\tau)b}\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right] \leq \mathbb{E}\left[\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2\right] - \mathbb{E}\left[\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2\right]. \tag{30}$$

In order to ensure the convergence, we need the coefficient of $\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right]$ to be positive. From basic algebra, one feasible range for the stepsize γ is

$$0 < \gamma \leq \frac{1}{(L+2\tau)(1+2\lambda)},$$

which directly implies that

$$0 < \frac{\gamma}{(L+2\tau)b} \leq \frac{2\gamma - (L+2\tau)(1+2\lambda)\gamma^2}{(L+2\tau)b}.$$

By simplifying (30) with the above result and dropping the negative term, we can derive the following bound for the $\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right]$ averaged over t iterations

$$\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right] \leq \frac{(L+2\tau)b}{\gamma t}\mathbb{E}\left[\|\boldsymbol{x}^0-\boldsymbol{x}^*\|^2\right] \leq \frac{(L+2\tau)b}{\gamma t}R_0^2. \tag{31}$$

The above inequality establishes that the change of the stale iterate $\widetilde{\boldsymbol{x}}^k$ converges to zero as t increases. Next, we will use the bound to establish the similar result for the actual iterate $\boldsymbol{x}^k$. We know that $\|\mathsf{G}(\boldsymbol{x}^k)\|^2$ can be bounded by

$$\begin{aligned}
\|\mathsf{G}(\boldsymbol{x}^k)\|^2 &\leq (\|\mathsf{G}(\boldsymbol{x}^k)-\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|+\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|)^2\\
&= \|\mathsf{G}(\boldsymbol{x}^k)-\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2+\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2+2\|\mathsf{G}(\boldsymbol{x}^k)-\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|\\
&\leq 2\|\mathsf{G}(\boldsymbol{x}^k)-\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2+2\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\\
&\leq 2(L+2\tau)^2\|\boldsymbol{x}^k-\widetilde{\boldsymbol{x}}^k\|^2+2\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2
\end{aligned} \tag{32}$$

where in the second inequality we used the Young's inequality (24), and in the third inequality we used the following result implied by Lemma 1

$$(L+2\tau)\|\boldsymbol{x}-\boldsymbol{y}\| \geq \|\mathsf{G}(\boldsymbol{x})-\mathsf{G}(\boldsymbol{y})\|.$$

By expressing the stale iterate $\widetilde{\boldsymbol{x}}^k$, we can write equation (32) as

$$\begin{aligned}
\|\mathsf{G}(\boldsymbol{x}^k)\|^2 &\leq 2(L+2\tau)^2\|\sum_{s=k-\lambda}^{k-1}\gamma\mathsf{G}_{i_s}(\widetilde{\boldsymbol{x}}^s)\|^2+2\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2.\\
&\leq 2\lambda(L+2\tau)^2\sum_{s=k-\lambda}^{k-1}\gamma^2\|\mathsf{G}_{i_s}(\widetilde{\boldsymbol{x}}^s)\|^2+2\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2.
\end{aligned} \tag{33}$$

where we use the fact

$$\|\sum_{i=1}^n\boldsymbol{x}_i\|^2=\sum_{i=1}^n\|\boldsymbol{x}_i\|^2+\sum_{a\neq b}\boldsymbol{x}_a^\mathsf{T}\boldsymbol{x}_b \leq \sum_{i=1}^n\|\boldsymbol{x}_i\|^2+\frac{1}{2}\sum_{a\neq b}\left[\|\boldsymbol{x}_a\|^2+\|\boldsymbol{x}_b\|^2\right]=n\sum_{i=1}^n\|\boldsymbol{x}_i\|^2$$

Taking the expectation of equation (33) leads to

$$\begin{aligned}
&\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right]\\
&\leq 2\lambda(L+2\tau)^2\sum_{s=k-\lambda}^{k-1}\gamma^2\mathbb{E}\left[\|\mathsf{G}_{i_s}(\widetilde{\boldsymbol{x}}^s)\|^2\right]+2\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right]\\
&\leq 2\lambda(L+2\tau)^2\sum_{s=k-\lambda}^{k-1}\frac{\gamma^2\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^s)\|^2\right]}{b}+2\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right],
\end{aligned} \tag{34}$$

By averaging (34) over $t>0$ iterations, we obtain that

$$\begin{aligned}
&\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right]\\
&\leq \frac{2\lambda(L+2\tau)^2}{t}\sum_{k=0}^{t-1}\sum_{s=k-\lambda}^{k-1}\frac{\gamma^2\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^s)\|^2\right]}{b}+\frac{2}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right]\\
&\leq \frac{2\lambda^2(L+2\tau)^2}{t}\sum_{k=0}^{t-1}\frac{\gamma^2\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right]}{b}+\frac{2}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right]
\end{aligned} \tag{35}$$

where we again used result in (29) in the last inequality. Re-arranging the terms in (35) yields

$$\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right] \leq \left[\frac{2\lambda^2(L+2\tau)^2}{b}\gamma^2+2\right]\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right] \tag{36}$$

We plug the result in (31) into (36) and obtain

$$\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right] \leq \left[\frac{2\lambda^2(L+2\tau)^2}{b}\gamma^2+2\right]\frac{(L+2\tau)b}{\gamma t}R_0^2, \tag{37}$$

Since it is always true that

$$\gamma \leq \frac{1}{(L+2\tau)(1+2\lambda)} \leq \frac{1}{(L+2\tau)(1+\lambda)}.$$

we can simplify the bound by using the above inequality related to the stepsize γ

$$\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right] \leq \left[\frac{2\lambda^2}{(1+\lambda)^2 b}+2\right]\frac{(L+2\tau)b}{\gamma t}R_0^2. \tag{38}$$

Let $D = 2\lambda^2/(1+\lambda)^2$, and we derive the desired result.

$$\min_{0\leq k\leq t-1}\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right] \leq \frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right] \leq \left[\frac{D}{b}+2\right]\frac{(L+2\tau)b}{\gamma t}R_0^2. \tag{39}$$

B.2 Proof of Theorem 2

We prove Theorem 2 by following the procedure in the proof of Theorem 1 with the adaptation to the block stochastic operator $\widehat{\mathsf{G}}_i$. In the key steps, we will highlight the difference between the two proofs. In addition to Lemma 1, our second proof requires the following lemma related to the statistical properties of $\widehat{\mathsf{G}}$.

Lemma 2. *Let Assumption 3 and 4 hold for g and D_σ. Then, we can establish the following statements for operator $\widehat{\mathsf{G}}$*

$$\mathbb{E}\left[\widehat{\mathsf{G}}(\boldsymbol{x})\right] = \mathsf{G}(\boldsymbol{x}), \quad \mathbb{E}\left[\|\widehat{\mathsf{G}}(\boldsymbol{x})-\mathsf{G}(\boldsymbol{x})\|^2\right] \leq \frac{\nu^2}{w},$$

which further implies that

$$\mathbb{E}\left[\|\widehat{\mathsf{G}}(\boldsymbol{x})\|^2\right] \leq \frac{\nu^2}{w}+\|\mathsf{G}(\boldsymbol{x})\|^2.$$

Proof. Since the the stochasticity happens only in the evaluation of the gradient, it is straightforward to see that

$$\mathbb{E}\left[\widehat{\mathsf{G}}(\boldsymbol{x})\right] = \mathbb{E}[\widehat{\nabla} g(\boldsymbol{x})]+\mathsf{D}_\sigma(\boldsymbol{x}) = \mathsf{G}(\boldsymbol{x}),$$

Similarly, we have that

$$\mathbb{E}\left[\|\widehat{\mathsf{G}}(\boldsymbol{x})-\mathsf{G}(\boldsymbol{x})\|_2^2\right] = \mathbb{E}\left[\|\widehat{\nabla} g(\boldsymbol{x})-\nabla g(\boldsymbol{x})\|_2^2\right] \leq \frac{\nu^2}{w}$$

Given that $\mathrm{Tr}(\mathbb{E}\left[X^\mathsf{T}X\right]) = \mathrm{Tr}(\mathrm{Cov}\left[X\right]) + \mathrm{Tr}(\mathbb{E}\left[X\right]^2)$, we obtain that

$$\mathbb{E}\left[\|\widehat{\mathsf{G}}(\boldsymbol{x})\|^2\right] = \mathbb{E}\left[\|\widehat{\mathsf{G}}(\boldsymbol{x})-\mathsf{G}(\boldsymbol{x})\|^2\right]+\mathbb{E}\left[\widehat{\mathsf{G}}(\boldsymbol{x})\right]^2 \leq \frac{\nu^2}{w}+\|\mathsf{G}(\boldsymbol{x})\|^2,$$

where we let $\mathbb{E}\left[\widehat{\mathsf{G}}(\boldsymbol{x})\right]^2 := \mathbb{E}\left[\widehat{\mathsf{G}}(\boldsymbol{x})\right]^\mathsf{T}\mathbb{E}\left[\widehat{\mathsf{G}}(\boldsymbol{x})\right]$. Note that $\mathrm{Tr}(\cdot)$ and $\mathrm{Cov}(\cdot)$ denote the computation of the trace and covariance of a matrix and a vector, respectively. □

Now we start the proof. Similar as (17), we write that

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^*\|^2|\mathcal{X}^k\right]\\
&=\mathbb{E}\left[\|\boldsymbol{x}^k-\gamma\widehat{\mathsf{G}}_i(\widetilde{\boldsymbol{x}}^k)-\boldsymbol{x}^*\|^2|\mathcal{X}^k\right]\\
&=\mathbb{E}\left[\|\boldsymbol{x}^k-\boldsymbol{x}^*\|^2|\mathcal{X}^k\right]+\gamma^2\mathbb{E}\left[\|\widehat{\mathsf{G}}_i(\widetilde{\boldsymbol{x}}^k)\|^2|\mathcal{X}^k\right]+2\gamma\mathbb{E}\left[(\widehat{\mathsf{G}}_i(\widetilde{\boldsymbol{x}}^k))^\mathsf{T}(\boldsymbol{x}^*-\boldsymbol{x}^k)|\mathcal{X}^k\right]
\end{aligned} \tag{40}$$

Here, the conditional expectation is taken for $\widehat{\mathsf{G}}_i(\boldsymbol{x}) = \mathsf{U}_i\mathsf{U}_i^\mathsf{T}\widehat{\mathsf{G}}(\boldsymbol{x})$. By using Lemma 2, we can compute conditional expectations as

$$\mathbb{E}\left[(\widehat{\mathsf{G}}_i(\widetilde{\boldsymbol{x}}^k))^\mathsf{T}(\boldsymbol{x}^* - \boldsymbol{x}^k)|\mathcal{X}^k\right] = \frac{1}{b}\mathbb{E}\left[(\widehat{\mathsf{G}}(\widetilde{\boldsymbol{x}}^k))^\mathsf{T}(\boldsymbol{x}^* - \boldsymbol{x}^k)|\mathcal{X}^k\right] = \frac{1}{b}(\mathsf{G}(\widetilde{\boldsymbol{x}}^k))^\mathsf{T}(\boldsymbol{x}^* - \boldsymbol{x}^k) \tag{41}$$

and

$$\mathbb{E}\left[\|\widehat{\mathsf{G}}_i(\widetilde{\boldsymbol{x}}^k)\|^2|\mathcal{X}^k\right] = \frac{1}{b}\mathbb{E}\left[\|\widehat{\mathsf{G}}(\widetilde{\boldsymbol{x}}^k)\|^2|\mathcal{X}^k\right] \leq \frac{\nu^2}{wb} + \frac{\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2}{b}. \tag{42}$$

where we first compute the expectation corresponding to the randomized block and then the expectation for the stochastic measurements. We note that the expectation of the cross term (41) remains the same as the result in (18), while the expectation in (42) has one extra term related to the norm variance of the stochastic operator compared with (19). As we shall see in the future steps, the difference in the expectation of the operator's squared norm leads to the most modifications. Using the above results in equation (40) yields that

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2|\mathcal{X}^k\right]\\ &\leq \|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 + \frac{\gamma^2}{b}\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2 + \frac{\gamma^2\nu^2}{wb} + \underbrace{\frac{2\gamma}{b}(\mathsf{G}(\widetilde{\boldsymbol{x}}^k))^\mathsf{T}(\boldsymbol{x}^* - \boldsymbol{x}^k)}_{(\dagger)}.\end{aligned} \tag{43}$$

By following (21), we can express the term $(\dagger)$ as

$$\begin{aligned}&\frac{2\gamma}{b}(\mathsf{G}(\widetilde{\boldsymbol{x}}^k))^\mathsf{T}(\boldsymbol{x}^* - \boldsymbol{x}^k)\\ &= \frac{2\gamma}{b}(\mathsf{G}(\widetilde{\boldsymbol{x}}^k) - \mathsf{G}(\boldsymbol{x}^*))^\mathsf{T}(\boldsymbol{x}^* - \widetilde{\boldsymbol{x}}^k) + \frac{2\gamma^2}{b}\sum_{s=k-\Delta_k}^{k-1}\mathsf{G}(\widetilde{\boldsymbol{x}}^k)^\mathsf{T}\widehat{\mathsf{G}}_{i_s}(\widetilde{\boldsymbol{x}}^s),\end{aligned} \tag{44}$$

The upper bound of the first term is the same as shown in (22), which is

$$\frac{2\gamma}{b}(\mathsf{G}(\widetilde{\boldsymbol{x}}^k) - \mathsf{G}(\boldsymbol{x}^*))^\mathsf{T}(\boldsymbol{x}^* - \widetilde{\boldsymbol{x}}^k) \leq -\frac{2\gamma\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2}{b(L+2\tau)}. \tag{45}$$

Similarly, our second term is bounded by

$$\frac{2\gamma^2}{b}\sum_{s=k-\Delta_k}^{k-1}\mathsf{G}(\widetilde{\boldsymbol{x}}^k)^\mathsf{T}\widehat{\mathsf{G}}_{i_s}(\widetilde{\boldsymbol{x}}^s) \leq \frac{\lambda\gamma^2\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2}{b} + \sum_{s=k-\lambda}^{k-1}\frac{\gamma^2\|\widehat{\mathsf{G}}(\widetilde{\boldsymbol{x}}^s)\|^2}{b}, \tag{46}$$

where we used the Young's inequality (24) together with the fact that

$$\sum_{s=k-\Delta k}^{k-1}\|\widehat{\mathsf{G}}_{i_s}(\widetilde{\boldsymbol{x}}^k)\|^2 \leq \sum_{s=k-\lambda}^{k-1}\|\widehat{\mathsf{G}}_{i_s}(\widetilde{\boldsymbol{x}}^k)\|^2 \leq \sum_{s=k-\lambda}^{k-1}\|\widehat{\mathsf{G}}(\widetilde{\boldsymbol{x}}^k)\|^2.$$

Equation (45) and (46) together establish the overall upper bound for the term $(\dagger)$

$$\frac{2\gamma}{b}(\mathsf{G}(\widetilde{\boldsymbol{x}}^k))^\mathsf{T}(\boldsymbol{x}^* - \boldsymbol{x}^k) \leq \frac{(L+2\tau)\lambda\gamma^2 - 2\gamma}{(L+2\tau)b}\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2 + \sum_{s=k-\lambda}^{k-1}\frac{\gamma^2\|\widehat{\mathsf{G}}(\widetilde{\boldsymbol{x}}^s)\|^2}{b}. \tag{47}$$

By plugging (47) into (40) and re-arranging the terms, we obtain that

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2|\mathcal{X}^k\right]\\ &\leq \|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 + \frac{\gamma^2\nu^2}{wb} + \sum_{s=k-\lambda}^{k-1}\frac{\gamma^2\|\widehat{\mathsf{G}}(\widetilde{\boldsymbol{x}}^s)\|^2}{b} + \frac{(L+2\tau)(1+\lambda)\gamma^2 - 2\gamma}{(L+2\tau)b}\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2.\end{aligned} \tag{48}$$

Taking the total expectation of equation (48) and re-arranging the terms yields that

$$\begin{aligned}&\frac{2\gamma - (L+2\tau)(1+\lambda)\gamma^2}{(L+2\tau)b}\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right]\\ &\leq \mathbb{E}\left[\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2\right] - \mathbb{E}\left[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2\right] + \frac{\gamma^2\nu^2}{wb} + \gamma^2\sum_{s=k-\lambda}^{k-1}\left[\frac{\nu^2}{wb} + \frac{\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^s)\|^2\right]}{b}\right]\end{aligned} \tag{49}$$

where we use the following inequality derived by using the law of total expectation and Lemma 2

$$\mathbb{E}\left[\|\widehat{\mathsf{G}}(\widetilde{\boldsymbol{x}}^s)\|^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|\widehat{\mathsf{G}}(\widetilde{\boldsymbol{x}}^s)\|^2|\mathcal{X}^s\right]\right] \leq \frac{\nu^2}{w} + \mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^s)\|^2\right]. \tag{50}$$

We telescope-sum equation (49) over $t > 0$ iterations to obtain

$$\begin{aligned}
&\sum_{k=0}^{t-1} \frac{2\gamma - (L+2\tau)(1+\lambda)\gamma^2}{(L+2\tau)b} \mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right] \\
&\leq \mathbb{E}\left[\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2\right] - \mathbb{E}\left[\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2\right] + \sum_{k=0}^{t-1} \frac{\gamma^2\nu^2}{wb} + \gamma^2 \sum_{k=0}^{t-1} \sum_{s=k-\lambda}^{k-1} \left[\frac{\nu^2}{wb} + \frac{\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^s)\|^2\right]}{b}\right]
\end{aligned} \tag{51}$$

By applying the same relaxation trick in (29) to (51)

$$\sum_{k=0}^{t-1} \sum_{s=k-\lambda}^{k-1} \left[\frac{\nu^2}{wb} + \frac{\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^s)\|^2\right]}{b}\right] \leq \lambda \sum_{k=0}^{t-1} \left[\frac{\nu^2}{wb} + \frac{\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right]}{b}\right], \tag{52}$$

we then have that

$$\sum_{k=0}^{t-1} \frac{2\gamma - (L+2\tau)(1+2\lambda)\gamma^2}{(L+2\tau)b} \mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right] \leq \mathbb{E}\left[\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2\right] + \frac{(1+\lambda)\gamma^2\nu^2}{wb} \cdot t, \tag{53}$$

where we dropped the negative term. Recall that if γ is in the range $\gamma \in (0, 1/((L+2\tau)(1+2\lambda))]$, we have the inequality

$$\frac{\gamma}{(L+2\tau)b} \leq \frac{2\gamma - (L+2\tau)(1+2\lambda)\gamma^2}{(L+2\tau)b}.$$

By relaxing the coefficient in the lefthand side, dividing the inequality by t, and re-arranging the terms, we obtain the convergence in terms of the stale iterate $\widetilde{\boldsymbol{x}}^k$

$$\begin{aligned}
\frac{1}{t}\sum_{k=0}^{t-1} \mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right] &\leq \frac{(L+2\tau)b}{\gamma t}\left[\mathbb{E}\left[\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2\right] + \frac{(1+\lambda)\gamma^2\nu^2}{wb} \cdot t\right] \\
&\leq \frac{(L+2\tau)b}{\gamma t} R_0^2 + \frac{\gamma}{w} C
\end{aligned} \tag{54}$$

where we used Assumption 2 and let $C = (L+2\tau)(1+\lambda)\nu^2$. Compared with the result in equation (31),equation (54) has the extra term related to the variance of $\widehat{\mathsf{G}}_i(\boldsymbol{x})$. Next, we establish the convergence in terms of actual iterate $\boldsymbol{x}^k$. Following the steps from (32) to (34), we directly obtain the inequality related to $\widehat{\mathsf{G}}_i(\widetilde{\boldsymbol{x}})$

$$\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right] \leq 2\lambda(L+2\tau)^2 \sum_{s=k-\lambda}^{k-1} \gamma^2 \mathbb{E}\left[\|\widehat{\mathsf{G}}_{i_s}(\widetilde{\boldsymbol{x}}^s)\|^2\right] + 2\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right] \tag{55}$$

By using the the result in (50), we derive from (55) that

$$\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right] \leq 2\lambda(L+2\tau)^2 \sum_{s=k-\lambda}^{k-1} \gamma^2 \left[\frac{\nu^2}{wb} + \frac{\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^s)\|^2\right]}{b}\right] + 2\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right]. \tag{56}$$

By averaging (55) over $t > 0$ iterations, we obtain that

$$\begin{aligned}
&\frac{1}{t}\sum_{k=0}^{t-1} \mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right] \\
&\leq \frac{2\lambda(L+2\tau)^2}{t} \sum_{k=0}^{t-1} \sum_{s=k-\lambda}^{k-1} \gamma^2 \left[\frac{\nu^2}{wb} + \frac{\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^s)\|^2\right]}{b}\right] + \frac{2}{t}\sum_{k=0}^{t-1} \mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right] \\
&\leq \frac{2\lambda^2(L+2\tau)^2}{t} \sum_{k=0}^{t-1} \gamma^2 \left[\frac{\nu^2}{wb} + \frac{\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right]}{b}\right] + \frac{2}{t}\sum_{k=0}^{t-1} \mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right]
\end{aligned} \tag{57}$$

where we again used the relaxation (52) in the last inequality. Re-arranging the terms in (57) yields

$$\begin{aligned}&\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right]\\&\leq\frac{2\lambda^2(L+2\tau)^2\cdot\nu^2}{wb}\gamma^2+\left[\frac{2\lambda^2(L+2\tau)^2}{b}\gamma^2+2\right]\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\mathsf{G}(\widetilde{\boldsymbol{x}}^k)\|^2\right]\end{aligned}\tag{58}$$

We plug the result in (54) into (58) and obtain

$$\begin{aligned}&\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right]\\&\leq\frac{2\lambda^2(L+2\tau)^2\cdot\nu^2}{wb}\gamma^2+\left[\frac{2\lambda^2(L+2\tau)^2}{b}\gamma^2+2\right]\left[\frac{(L+2\tau)b}{\gamma t}R_0^2+\frac{\gamma}{w}C\right]\end{aligned}\tag{59}$$

Similarly, we can use the fact

$$\gamma\leq\frac{1}{(L+2\tau)(1+\lambda)}.$$

to simplify the bound in (59)

$$\begin{aligned}&\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right]\\&\leq\frac{2\lambda^2(L+2\tau)^2\cdot\nu^2}{wb}\cdot\frac{1}{(L+2\tau)(1+\lambda)}\cdot\gamma+\left[\frac{2\lambda^2}{(1+\lambda)^2b}+2\right]\left[\frac{(L+2\tau)b}{\gamma t}R_0^2+\frac{\gamma}{w}C\right]\\&=\frac{2\lambda^2}{(1+\lambda)^2b}\cdot\frac{(L+2\tau)(1+\lambda)\nu^2}{w}\cdot\gamma+\left[\frac{2\lambda^2}{(1+\lambda)^2b}+2\right]\left[\frac{(L+2\tau)b}{\gamma t}R_0^2+\frac{\gamma}{w}C\right]\\&=\frac{2\lambda^2}{(1+\lambda)^2b}\cdot\frac{C}{w}\gamma+\left[\frac{2\lambda^2}{(1+\lambda)^2b}+2\right]\left[\frac{(L+2\tau)b}{\gamma t}R_0^2+\frac{\gamma}{w}C\right]\end{aligned}\tag{60}$$

where we recall $C=(L+2\tau)(1+\lambda)\nu^2$. Let $D=2\lambda^2/(1+\lambda)^2$ and we can derive the result of Theorem 2

$$\min_{0\leq k\leq t-1}\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right]\leq\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right]\leq\left[\frac{D}{b}+2\right]\frac{(L+2\tau)b}{\gamma t}R_0^2+\left[\frac{2D}{b}+2\right]\frac{\gamma}{w}C,\tag{61}$$

which immediately implies the result in remark 1 by setting $\gamma=1/\sqrt{wt}$

$$\min_{0\leq k\leq t-1}\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right]\leq\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\mathsf{G}(\boldsymbol{x}^k)\|^2\right]\leq\left[\frac{D}{b}+2\right]\frac{(L+2\tau)b}{\sqrt{wt}}R_0^2+\left[\frac{2D}{b}+2\right]\frac{C}{\sqrt{wt}}.\tag{62}$$

From basic algebra, we can derive the condition for λ

$$\frac{1}{\sqrt{wt}}\leq\frac{1}{(L+2\tau)(1+2\lambda)}\quad\Rightarrow\quad\lambda\leq\frac{1}{2}\left[\frac{\sqrt{wt}}{L+2\tau}-1\right].$$

C Background on Monotone Operators

The results in our review can be found in different forms in standard textbooks (Rockafellar & Wets, 1998; Boyd & Vandenberghe, 2004; Nesterov, 2004; Bauschke & Combettes, 2017), and we include these results for completeness.

Definition 1. *An operator* T *is Lipschitz continuous with constant* $L>0$ *if*

$$\|\mathsf{T}\boldsymbol{x}-\mathsf{T}\boldsymbol{y}\|\leq L\|\boldsymbol{x}-\boldsymbol{y}\|,\quad\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n.$$

When $L=1$*, we say that* T *is nonexpansive. When* $L<1$*, we say that* T *is a contraction.*

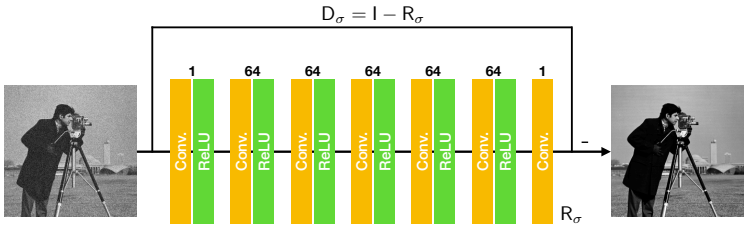

Figure 5: Illustration of the architecture of DnCNN used in all experiments. The neural net is trained to remove the AWGN from its noisy input image. We also constrains the Lipschitz constant of R_σ to be smaller than 2 by using the spectral normalization technique in (Sedghi et al., 2019). This provides a necessary condition for the satisfaction of Assumption 4.

Figure 6: Six test images used in the experiments on CS. From the left to right, there are *cameraman*, *house*, *pepper*, *starfish*, *butterfly*, and *jet*.

Definition 2. T *is monotone if*

$$(\mathsf{T}(\boldsymbol{x}) - \mathsf{T}(\boldsymbol{y}))^\mathsf{T}(\boldsymbol{x} - \boldsymbol{y}) \geq 0, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

We say that it is strongly monotone or coercive with parameter $\mu > 0$ *if*

$$(\mathsf{T}(\boldsymbol{x}) - \mathsf{T}(\boldsymbol{y}))^\mathsf{T}(\boldsymbol{x} - \boldsymbol{y}) \geq \mu\|\boldsymbol{x} - \boldsymbol{y}\|^2, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

Definition 3. T *is cocoercive with constant* $\beta > 0$ *if*

$$(\mathsf{T}(\boldsymbol{x}) - \mathsf{T}(\boldsymbol{y}))^\mathsf{T}(\boldsymbol{x} - \boldsymbol{y}) \geq \beta\|\mathsf{T}\boldsymbol{x} - \mathsf{T}\boldsymbol{y}\|^2, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

When $\beta = 1$*, we say that* T *is firmly nonexpansive.*

The following results are derived from the definition above.

Proposition 1. *For a convex and continuously differentiable function* f*, we have*

$$\nabla f \text{ is } L\text{-Lipschitz continuous} \quad \Leftrightarrow \quad \nabla f \text{ is } (1/L)\text{-cocoercive}.$$

Proof. The proof is a minor variation of the one presented as Theorem 2.1.5 in Section 2.1 of (Nesterov, 2004). □

Proposition 2. *Consider* $\mathsf{T} : \mathbb{R}^n \to \mathbb{R}^n$ *and* $\beta > 0$*. Then, the following are equivalent*

$$\mathsf{T} \text{ is } \beta\text{-cocoercive} \quad \Leftrightarrow \quad \mathsf{I} - 2\beta\mathsf{T} \text{ is nonexpansive.}$$

Proof. Let $\mathsf{R} \coloneqq \mathsf{I} - 2\beta\mathsf{T}$, then $\mathsf{T} = 1/(2\beta)(\mathsf{I} - \mathsf{R})$. First suppose that T is β-cocoercive. Let $\boldsymbol{h} \coloneqq \boldsymbol{x} - \boldsymbol{y}$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. We then have

$$\beta\|\mathsf{T}(\boldsymbol{x}) - \mathsf{T}(\boldsymbol{y})\|^2 \leq (\mathsf{T}(\boldsymbol{x}) - \mathsf{T}(\boldsymbol{y}))^\mathsf{T}\boldsymbol{h} = \frac{1}{2\beta}\|\boldsymbol{h}\|^2 - \frac{1}{2\beta}(\mathsf{R}(\boldsymbol{x}) - \mathsf{R}(\boldsymbol{y}))^\mathsf{T}\boldsymbol{h}.$$

We also have that

$$\beta\|\mathsf{T}(\boldsymbol{x}) - \mathsf{T}(\boldsymbol{y})\|^2 = \frac{1}{4\beta}\|\boldsymbol{h}\|^2 - \frac{1}{2\beta}(\mathsf{R}(\boldsymbol{x}) - \mathsf{R}(\boldsymbol{y}))^\mathsf{T}\boldsymbol{h} + \frac{1}{4\beta}\|\mathsf{R}(\boldsymbol{x}) - \mathsf{R}(\boldsymbol{y})\|^2.$$

By combining these two and simplifying the expression

$$\|\mathsf{R}(\boldsymbol{x}) - \mathsf{R}(\boldsymbol{y})\| \leq \|\boldsymbol{h}\|.$$

The converse can be proved by following this logic in reverse.

□

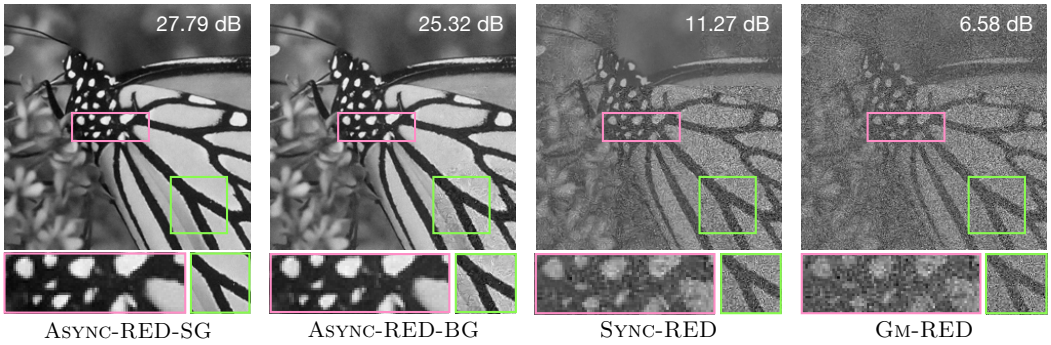

Figure 7: Visualization of the recovered images from the compressed measurements by ASYNC-RED-BG/SG, SYNC-RED, and GM-RED. Each algorithm is run with a time budget of 700 seconds.

The following characterization is also convenient.

Proposition 3. *For nonexpansive operators* T_1 *and* T_2 *with a constant* $\alpha \in (0,1)$*, then the convex combination of the two operators* $(1-\alpha)\mathsf{T}_1 + \alpha\mathsf{T}_2$ *is nonexpansive.*

Proof. Let $\mathsf{T} := (1-\alpha)\mathsf{T}_1 + \alpha\mathsf{T}_2$. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we can write

$$\|\mathsf{T}(\boldsymbol{x}) - \mathsf{T}(\boldsymbol{y})\| \leq (1-\alpha)\|\mathsf{T}_1(\boldsymbol{x}) - \mathsf{T}_1(\boldsymbol{y})\| + \alpha\|\mathsf{T}_2(\boldsymbol{x}) - \mathsf{T}_2(\boldsymbol{y})\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|$$

□

D ADDITIONAL TECHNICAL DETAILS

This section presents several technical details that were omitted from the main paper for space. Section D.1 presents the architecture and training of our DnCNN prior. Section D.2 provides extra details and validations that compliment the experiments in Section 5 of the main paper.

D.1 ARCHITECTURE AND TRAINING OF THE DNCNN PRIOR

Our denoiser follows the standard architecture of DnCNN (Zhang et al., 2017a). Fig. 5 visualizes the architectural details of the DnCNN prior used in our experiments. Similar priors are extensively used in various PnP and RED algorithms (Zhang et al., 2017b; Ryu et al., 2019; Sun et al., 2019a). In total, the network contains 7 layers, of which the first 6 layers consist of a convolutional layer and a rectified linear unit (ReLU), while the last layer contains only a convolution operation. A skip connection from the input to the output is used to enforce the residual network R_σ to predict the noise residual. The output images of the first 6 layers have 64 feature maps, while that of the last layer is a single-channel image. We set all convolutional kernels to be 3×3 with stride 1, which indicates that intermediate images have the same spatial size as the input image. We generated 44700 training examples by adding AWGN to 400 images from the BSD400 dataset (Martin et al., 2001) and extracting small patches of 128×128 pixels with stride 30. Our DnCNN denoiser is trained to optimize the *mean squared error* by using the Adam optimizer (Kingma & Ba, 2015).

Different approaches have been used to constrain the Lipschitz constant (LC) of the denoising prior (Ryu et al., 2019; Sun et al., 2019a). We adopt the spectral normalization technique in (Sedghi et al., 2019) to control the LC of our DnCNN prior. In the training, we constrain the residual network R_σ such that its LC is smaller than 2. Since the non-expansiveness of D_σ implies that R_σ has LC ≤ 2, this provides a *necessary* condition for D_σ to satisfy Assumption 4 (Sun et al., 2019a).

D.2 EXTRA DETAILS AND VALIDATIONS

All experiments are run on the server equipped with 32 Intel(R) Xeon(R) CPU E5-2620 v4 processors of 3.2 GHz and 264 GBs of DDR memory. We trained all neural nets using NVIDIA RTX 2080

Table 1: SNR values obtained by ASYNC-RED-BG using different ***block sizes*** on CS task.

Block size	cameraman	house	pepper	starfish	butterfly	jet	Average
120	27.77	30.92	29.60	28.23	28.89	28.90	**29.05**
80	27.75	30.95	29.58	28.28	28.78	28.76	**29.01**
60	27.74	30.95	29.65	28.12	28.85	28.71	**28.99**

Table 2: SNR values obtained by ASYNC-RED-SG using different ***minibatch sizes*** on CS task.

minibatch size	cameraman	house	pepper	starfish	butterfly	jet	Average
1120	27.00	30.56	28.99	26.98	27.83	27.09	**28.03**
2240	27.31	30.78	29.01	27.64	28.29	28.03	**28.51**
3360	27.56	30.87	29.53	28.24	28.72	28.71	**28.93**

GPUs. We define the SNR (dB) used in the experiments as

$$\mathrm{SNR}(\hat{\boldsymbol{x}}, \boldsymbol{x}) \triangleq 20 \log_{10} \left(\frac{\|\boldsymbol{x}\|_2}{\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2} \right)$$

where $\hat{\boldsymbol{x}}$ represents the reconstructed image and $\boldsymbol{x}$ denotes the ground truth. Note that our experimental setup satisfies Assumption 1-3 but provides a necessary condition for Assmption 4.

Fig. 6 shows the six test images used in the experiments of CS. They are resized to the size of 240×240 pixels by using the Matlab function `imresize`. As demonstrated in the middle figure in Fig. 3, ASYNC-RED-SG converges faster than ASYNC-RED-BG given a fixed amount of time. This is further visualized in Fig. 7, where each algorithm is run for roughly 700 seconds. Since ASYNC-RED-SG uses only one-fourth of the total measurements, the per-iteration complexity is lower than ASYNC-RED-BG, leading to the faster convergence speed. In particular, the final SNR value obtained by ASYNC-RED-SG is roughly 2 dB higher than ASYNC-RED-BG. Additionally, both ASYNC-RED-BG/SG achieves significantly better results than SYNC-RED and GM-RED due to their adoption of asynchronous updates.

Table 1 and 2 illustrate the evolution of the reconstruction performance as the block size b and minibatch size w changes, respectively. Table 1 summarizes the SNR values obtained for three block sizes $b \in \{60, 80, 120\}$. Async-RED-BG achieves almost the same SNR values under these settings. Table 2 summarizes the SNR values for three minibatch sizes $w \in \{1120, 2240, 3360\}$, which corresponds to 1/4, 1/2, and 3/4 of the full batch. As w increases, the final SNR performance improves, which is consistent with our theory. On the other hand, the error term due to stochastic processing in Theorem 2 is also proportional to the step size γ, which means that by using smaller γ, Async-RED-SG can approximate GM-RED as accurately as desired. However, a reduction in γ would also lead to slower convergence. One thus needs to tradeoff the desired accuracy against the desired speed to select a suitable configuration for Async-RED-SG.

The benefit of ASYNC-RED is fully explored when the denoiser acts like a *block-wise denoiser*, which means that it can perform denoising on blocks as effective as on the full image. A simple strategy for making denoisers block-effective is to include additional neighboring pixels at the input, but use the exact block size at the output. Table 3 reports the results of experimenting with the idea of *input padding* for DnCNN. The results indicate that by including a small number of pixels around each block at the input of DnCNN, one can match the performance of using the full image at the input of DnCNN.

The test image used in the experiment of CT is selected from the dataset of human protein atlas (Williams et al., 2017). We download 51 images that have the size of 3000×3000 pixels. We select one image for test, which is cropped to 800×800 pixels. We extract 39000 patches

Table 3: SNR values obtained by ASYNC-RED-BG using different ***pad size*** on CS task.

Pad size	cameraman	house	pepper	starfish	butterfly	jet	Average
0	27.64	30.86	29.44	28.09	28.64	28.69	**28.89**
10	27.72	30.99	29.56	28.25	28.79	28.73	**29.00**
20	27.72	30.99	29.57	28.27	28.79	28.75	**29.01**
full	27.75	30.95	29.58	28.28	28.78	28.76	**29.01**

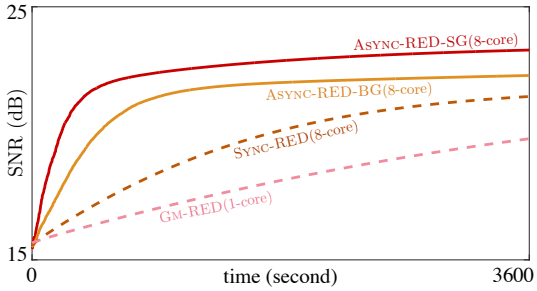

Figure 8: Convergence Illustration of ASYNC-RED-BG/SG and GM-RED for CT reconstruction with a time budget of 1 hour. Here, ASYNC-RED-SG randomly uses one-third of the total measurements at every iteration.

from the rest 50 images to train five specific DnCNN denoisers for the removal of AWGN with $\sigma \in \{5, 10, 15, 20, 25\}$. We report the result that has the highest SNR values. The Radon matrix used in the experiments corresponds to 180 angles with 1131 detectors. We synthesize the measurements by multiplying the Radom matrix with the vectorized image and add AWGN corresponding to 70 dB input SNR. In all tests, ASYNC-RED-SG randomly uses the measurements of 60 angles at each iteration, while ASYNC-RED-BG uses the entire measurement set. Fig. 8 plots SNR against the iteration number for Async-RED-BG/SG, Async-RED, and Gm-RED. Due to the lower per-iteration complexity, Async-RED-SG achieves the highest SNR value within the time budget of 1 hour. Fig. 9 provides a corresponding visual comparison between these methods. As reference, we also include the proximal gradient method with total variation regularizer (PGM-TV). The visual result of each method is obtained by running the algorithm with a time budget of 1 hour. Specifically, the per-iteration time cost of ASYNC-RED-BG/SG, SYNC-RED, GM-RED, PGM-TV are 5.23, 3.21, and 13.13, 19.19, and 44.74 seconds, respectively. The results clearly demonstrate that ASYNC-RED are indeed effective and efficient for a realistic, nontrivial imaging task on a large-scale image.

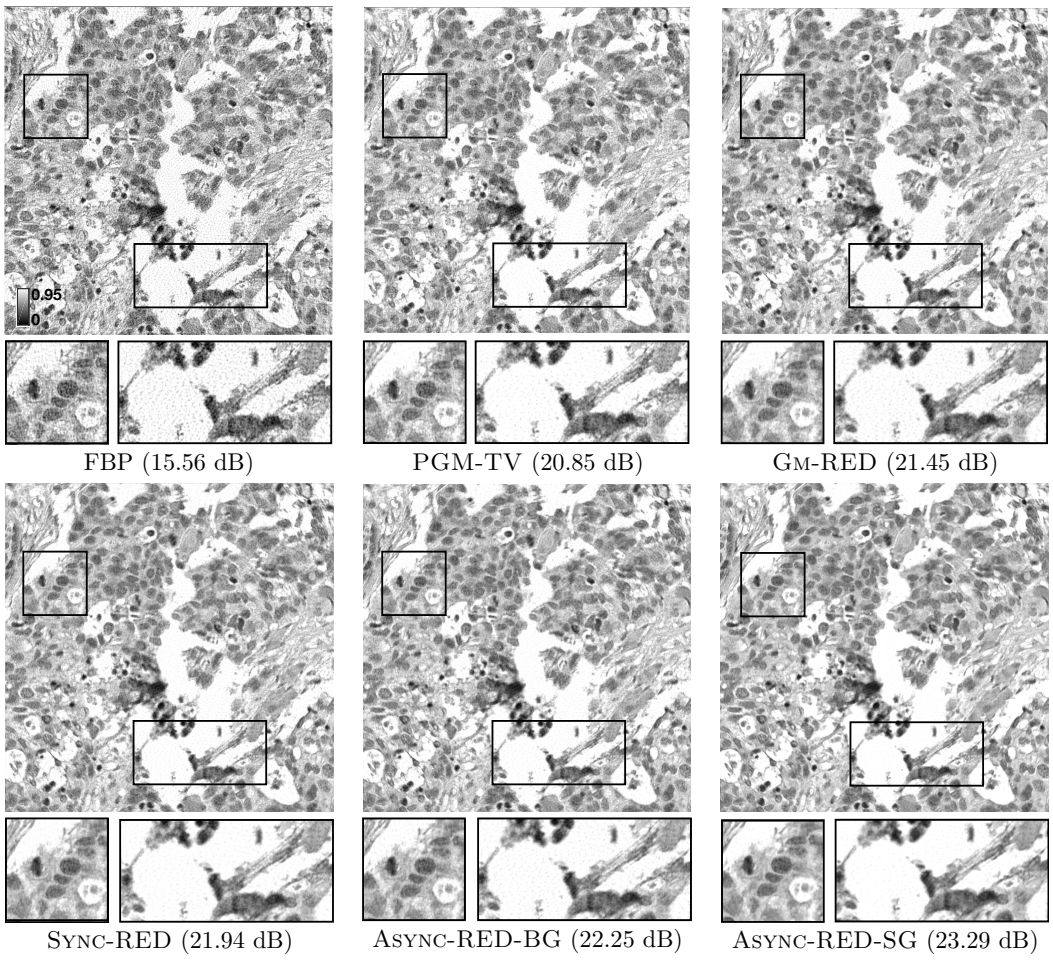

Figure 9: Visualization of the reconstructed CT images by PGM-TV, GM-RED, SYNC-RED, and ASYNC-RED-BG/SG. Each algorithm is run with a time budget of 1 hour. The colormap is adjusted for the best visual quality.