A UNIFIED THEORY OF ERROR FEEDBACK AND VARIANCE REDUCTION FOR NON-CONVEX OPTIMIZATION

Kai Yi
King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
kai.yi@kaust.edu.sa

ABSTRACT

In this project, we first extend EF-BV Condat et al. (2022) to non-convex setting and consider the special case under PL inequality.

CONTENTS

1 BACKGROUND

1.1 PROBLEM DEFINITION

We consider the standard federated optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

where n is the number of clients. f_i is the local optimization function at client i of the form

$$f_i(x) = \frac{1}{m}\sum_{j=1}^{m} f_{ij}(x), \tag{2}$$

where m is the number of datapoints at client i.

1.2 GENERAL COMPRESSORS

Based on bias-variance decomposition of the compression error (EF-BV Condat et al. (2022), Sec 2.3); for every $x \in \mathbb{R}^d$,

$$\mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] = \underbrace{\left\|\mathbb{E}[\mathcal{C}(x)] - x\right\|^2}_{\text{bias}} + \underbrace{\mathbb{E}\left[\left\|\mathcal{C}(x) - \mathbb{E}[\mathcal{C}(x)]\right\|^2\right]}_{\text{variance}}, \tag{3}$$

where the two terms at the right hand side satisfies

$$\begin{aligned} \|\mathbb{E}[\mathcal{C}(x)] - x\| &\le \eta\|x\|, \\ \mathbb{E}\left[\|\mathcal{C}(x) - \mathbb{E}[\mathcal{C}(x)]\|^2\right] &\le \omega\|x\|^2. \end{aligned} \tag{4}$$

η and ω is interpreted as the relative bias and variance controllers of the general compressor. Unbiased compressors and contractive biased compressors are all special case of this general compressor. More details could refer to EF-BV Condat et al. (2022), Sec 2.3.

1.3 AVERAGE VARIANCE OF COMPRESSORS

Given n compressors $\mathcal{C}_i$, the average relative variance Condat et al. (2022) ω_{av} is defined as: for every $x_i \in \mathbb{R}^d$,

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{C}_i\left(x_i\right) - \mathbb{E}\left[\mathcal{C}_i\left(x_i\right)\right]\right)\right\|^2\right] \le \frac{\omega_{\mathrm{av}}}{n}\sum_{i=1}^{n}\left\|x_i\right\|^2. \tag{5}$$

The property of this bound from EF-BF: $\omega_{\mathrm{av}} \le \omega$ and can be much smaller than ω. When $\mathcal{C}_i$ are mutually independent, $\omega_{\mathrm{av}} = \omega/n$.

1.4 EF-BV

The key part of EF-BV Condat et al. (2022) is that the gradient estimator is updated By

$$g_i^t = h_i^t + \nu\mathcal{C}(\nabla f_i(x^t) - h_i^t) \tag{6}$$

where the control variates are updated as

$$h_i^{t+1} = h_i^t + \lambda\mathcal{C}(\nabla f_i(x^t) - h_i^t). \tag{7}$$

More details could refer to EF-BV Condat et al. (2022) Sec. 3.

2 Convergence Analysis for Non-Convex Setting

2.1 Assumptions

Assumption 2.1 (Smoothness and Lower Bound)**.** Every f_i is L_i-smooth and f is L-smooth. $f^{\inf} := \inf_{x\in\mathbb{R}^d} f(x) > -\infty$.

Using Jensen's inequality, we have $L \le \frac{1}{n}\sum_i L_i$. Let $\tilde{L} := (\frac{1}{n}\sum_{i=1}^2)^{1/2}$. Using the arithmetic-quadratic mean inequality, $\frac{1}{n}\sum_i L_i \le \tilde{L}$.

Assumption 2.2 (Polyak-Łojasiewicz Condition)**.** There exists $\mu > 0$ such that

$$f(x) - f^{\inf} \le \frac{1}{2\mu}\|\nabla f(x)\|^2, \quad \forall x \in \mathbb{R}^d \tag{8}$$

2.2 Main Theorem

2.2.1 Definitions of Key Symbols

Here is the quick summary of defined key symbols.

$$\begin{aligned}
\lambda &:= \min\left(\frac{1-\eta}{(1-\eta)^2+\omega}, 1\right)\\
\nu &:= \min\left(\frac{1-\eta}{(1-\eta)^2+\omega_{\mathrm{av}}}, 1\right)\\
r &:= (1-\lambda+\lambda\eta)^2 + \lambda^2\omega\\
r_{\mathrm{av}} &:= (1-\nu+\nu\eta)^2 + \nu^2\omega_{\mathrm{av}}\\
s &:= \sqrt{\frac{1+r}{2r}} - 1\\
\theta &:= s(1+s)\frac{r}{r_{\mathrm{av}}}.
\end{aligned}$$

Theorem 2.3 (Linear convergence in the non-convex setting)**.** *Suppose Assumption 2.1 is satisfied. Suppose $\nu \in (0,1], \lambda \in (0,1]$ is such that $r < 1$, choosing the stepsize*

$$0 < \gamma \le \frac{1}{L + \tilde{L}\sqrt{\frac{r_{\mathrm{av}}}{r}}\frac{1}{s}}.$$

For every $t \ge 0$, define the Lyapunov function

$$\Psi^t := f\left(x^t\right) - f^{\inf} + \frac{\gamma}{2\theta}\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i\left(x^t\right) - h_i^t\right\|^2 \tag{9}$$

Fix $T \ge 1$ and let $\hat{x}^T$ be chosen from the iterates $x^0, x^1, \cdots, x^{T-1}$ uniformly at random. Then

$$\mathbb{E}\left[\left\|\nabla f\left(\hat{x}^T\right)\right\|^2\right] \le \frac{2\left(f\left(x^0\right) - f^{\inf}\right)}{\gamma T} + \frac{\mathbb{E}\left[G^0\right]}{\theta T}, \tag{10}$$

where $G^0 = \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(x^0) - h_i^0\right\|^2$.

Similar to EF-BV, we define the optimal values for the scaling parameters λ, ν:

$$\lambda^\star := \min\left(\frac{1-\eta}{(1-\eta)^2+\omega}, 1\right), \quad \nu^\star := \min\left(\frac{1-\eta}{(1-\eta)^2+\omega_{\mathrm{av}}}, 1\right).$$

Given $\lambda \in (0,1]$ and $\nu \in (0,1]$, we define for convenience

$$r := (1-\lambda+\lambda\eta)^2 + \lambda^2\omega, \quad r_{\text{av}} := (1-\nu+\nu\eta)^2 + \nu^2\omega_{\text{av}}.$$

as well as $s^\star := \sqrt{\frac{1+r}{2r}} - 1$ and $\theta^\star := s^\star\,(1+s^\star)\,\frac{r}{r_{\text{av}}}$.

Theorem 2.4 (Linear convergence under PL condition)**.** *Suppose Assumption 2.1 and PL assumption 2.2 are satisfied. Similarly, for every $t \geq 0$,*

$$\mathbb{E}\left[\Psi^t\right] \leq \left(\max\left(1-\gamma\mu, \frac{r+1}{2}\right)\right)^t \Psi^0. \tag{11}$$

3 Experiments

For our experiments on logistic regression and least squares, we use the dataset `mushrooms`, `phishing`, `a9a` and `w8a` from LibSVM Chang & Lin (2011). We randomly devide all datapoints into n different workers, where each worker has m datapoints. In our practice, we choose $m \in \{20, n_{\text{all}}\}$, where all is the total datapoints in a given dataset.

3.1 Logistic regression with nonconvex regularizer

We first consider the logistic regression optimization problem with nonconvex regularizer,

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp\left(-y_i a_i^\top x\right)\right) + \lambda \sum_{j=1}^{d} \frac{x_j^2}{1+x_j^2}, \tag{12}$$

where $a_i \in \mathbb{R}^d, y_i \in \{-1,1\}$ are the training data, and $\lambda > 0$ is the regularizer parameter. We used $\lambda = 0.1$ in all experiments. We present the results in Fig. 1.

3.2 Other Options

3.2.1 Least squares

Here we consider the function that satisfies the PL condition. We use the standard least squares objective function,

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \left(a_i^\top x - b_i\right)^2, \tag{13}$$

where $a_i \in \mathbb{R}^d, y_i \in \{-1,1\}$ are the training data.

We can extend to least squares if we want later.

3.2.2 Deep neural networks

We also consider the deep neural networks experiments. To make fair comparison, we choose CIFAR-10 as our target dataset, which contains 50, 000 images for training and 10,000 images for testing. We randomly split the whole training set and testing set into n=5 equal part. Similar to EF21 Richtárik et al. (2021), we consider models including ResNet18 and VGG11.

From EF21 Richtárik et al. (2021) Fig.13, 14 and 15, we found EF21 use tuned stepsizes. However, the advantage of EF-BV is that we have much larger stepsize under centain conditions. So here I pend the comparison with EF21 on deep neural networks.

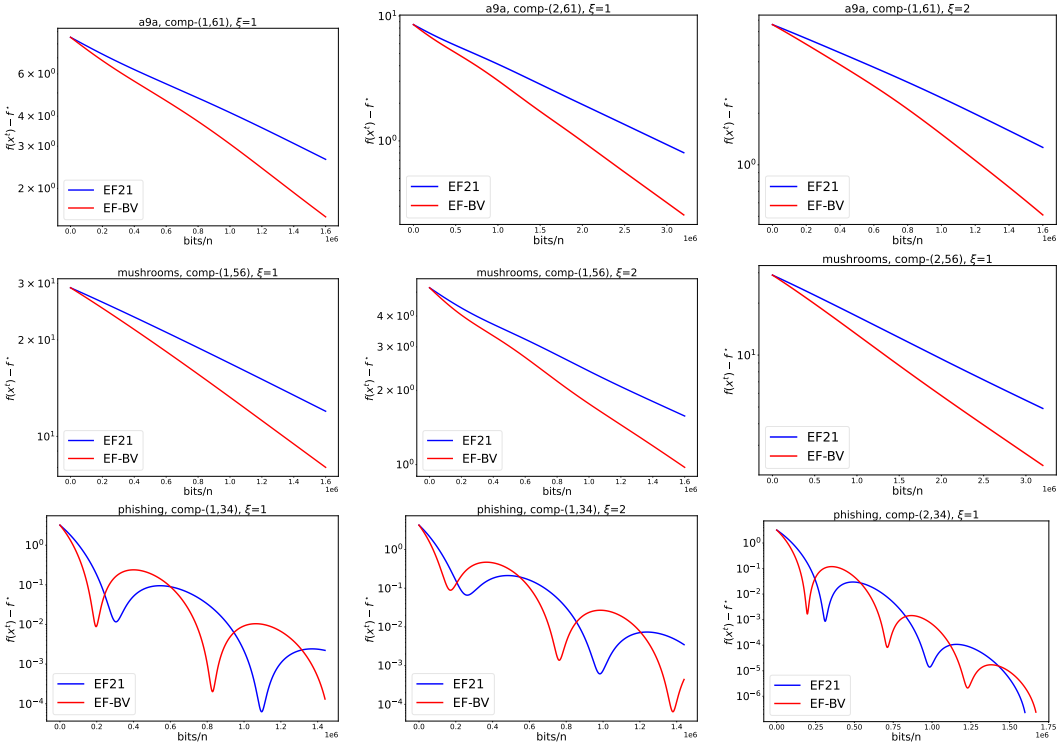

Figure 1: Comparison between EF21 and EF-BV in the non-convex setting. We see EF-BV outperforms EF21 on all datasets.

REFERENCES

Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3):1–27, 2011.

Laurent Condat, Kai Yi, and Peter Richtárik. Ef-bv: A unified theory of error feedback and variance reduction mechanisms for biased and unbiased compression in distributed optimization. arXiv preprint arXiv:2205.04180, 2022.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. In Proc. of 35th Conf. Neural Information Processing Systems (NeurIPS), 2021.

A APPENDIX

A.1 MISSING PROOFS

A.1.1 MISSING PROOF OF THEOREM 2.3

$g_i^t = h_i^t + \nu\mathcal{C}(\nabla f_i(x^t) - h_i^t); \quad h_i^{t+1} = h_i^t + \lambda\mathcal{C}(\nabla f_i(x^t) - h_i^t)$

Proof. We first bound $\mathbb{E}\left[\|g_i^{t+1} - \nabla f_i(x^t)\|^2\right]$.

Given any compressor $\mathcal{C}$, we can do a *bias-variance decomposition* of the compression error (see Eqn. 3). For every $t \geq 0$, define $W^t = \{x^t, h^t, (h_i^t)_{i=1}^n\}$,

$$\begin{aligned}
\mathbb{E}\left[\left\|g^{t+1} - \nabla f(x^t)\right\|^2 |W^t\right] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(h_i^t - \nabla f_i(x^t) + \nu\mathcal{C}_i^t\big(\nabla f_i(x^t) - h_i^t\big)\right)\right\|^2 |W^t\right] \\
&\overset{(3)}{=} \left\|\frac{1}{n}\sum_{i=1}^n \left(h_i^t - \nabla f_i(x^t) + \nu\mathbb{E}\big[\mathcal{C}_i^t\big(\nabla f_i(x^t) - h_i^t\big)\big]\right)|W^t\right\|^2 \\
&\quad + \nu^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\mathcal{C}_i^t\big(\nabla f_i(x^t) - h_i^t\big) - \mathbb{E}\big[\mathcal{C}_i^t\big(\nabla f_i(x^t) - h_i^t\big)\big]\right)\right\|^2 |W^t\right] \\
&\leq \left\|\frac{1}{n}\sum_{i=1}^n \left(h_i^t - \nabla f_i(x^t) + \nu\mathbb{E}\big[\mathcal{C}_i^t\big(\nabla f_i(x^t) - h_i^t\big)\big]\right)|W^t\right\|^2 \\
&\quad + \nu^2\frac{\omega_{\mathrm{av}}}{n}\sum_{i=1}^n \left\|\nabla f_i(x^t) - h_i^t\right\|^2,
\end{aligned} \tag{14}$$

where the last inequality follows from Eqn. 7 in Condat et al. (2022). In addition, using Jensen's inequality and the definition of general compressor (Eqn. 4),

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^n \left(h_i^t - \nabla f_i(x^t) + \nu\mathbb{E}\big[\mathcal{C}_i^t\big(\nabla f_i(x^t) - h_i^t\big)\big]\right)|W^t\right\| \\
&\leq \left\|\frac{1}{n}\sum_{i=1}^n \left(\nu\big(h_i^t - \nabla f_i(x^t)\big) + \nu\mathbb{E}\big[\mathcal{C}_i^t\big(\nabla f_i(x^t) - h_i^t\big)\big]\right)|W^t\right\| + (1-\nu)\left\|\frac{1}{n}\sum_{i=1}^n \big(h_i^t - \nabla f_i(x^t)\big)\right\| \\
&\leq \frac{\nu}{n}\sum_{i=1}^n \left\|h_i^t - \nabla f_i(x^t) + \mathbb{E}\big[\mathcal{C}_i^t\big(\nabla f_i(x^t) - h_i^t\big)\big]\,|W^t\right\| + \frac{1-\nu}{n}\sum_{i=1}^n \left\|h_i^t - \nabla f_i(x^t)\right\| \\
&\leq \frac{\nu\eta}{n}\sum_{i=1}^n \left\|\nabla f_i(x^t) - h_i^t\right\| + \frac{1-\nu}{n}\sum_{i=1}^n \left\|\nabla f_i(x^t) - h_i^t\right\| \\
&= \frac{1-\nu+\nu\eta}{n}\sum_{i=1}^n \left\|\nabla f_i(x^t) - h_i^t\right\|.
\end{aligned} \tag{15}$$

Therefore,

$$\left\|\frac{1}{n}\sum_{i=1}^n \left(h_i^t - \nabla f_i(x^t) + \nu\mathbb{E}\big[\mathcal{C}_i^t\big(\nabla f_i(x^t) - h_i^t\big)\big]\right)|W^t\right\|^2 \leq \frac{(1-\nu+\nu\eta)^2}{n}\sum_{i=1}^n \left\|\nabla f_i(x^t) - h_i^t\right\|^2.$$

Putting Eq. 15 into Eq. 14, we have

$$\mathbb{E}\left[\left\|g^{t+1} - \nabla f(x^t)\right\|^2 |W^t\right] \leq \left((1-\nu+\nu\eta)^2 + \nu^2\omega_{\mathrm{av}}\right)\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(x^t) - h_i^t\right\|^2.$$

Using the Tower property and we obtain the unconditioned term,

$$\mathbb{E}\Big[\big\|g^{t+1}-\nabla f(x^t)\big\|^2\Big] \leq \big((1-\nu+\nu\eta)^2+\nu^2\omega_{\mathrm{av}}\big)\,\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\big\|\nabla f_i(x^t)-h_i^t\big\|^2\Big].$$

Since we have the property (Richtárik et al., 2021, Lemma 4), for every $t \geq 0$,

$$f(x^{t+1})-f^{\inf} \leq f(x^t)-f^{\inf}-\frac{\gamma}{2}\big\|\nabla f(x^t)\big\|^2+\frac{\gamma}{2}\big\|g^{t+1}-\nabla f(x^t)\big\|^2+\left(\frac{L}{2}-\frac{1}{2\gamma}\right)\big\|x^{t+1}-x^t\big\|^2.$$

Thus, for every $t \geq 0$, conditionally on x^t, h^t and $(h_i^t)_{i=1}^n$,

$$\begin{aligned}\mathbb{E}\big[f(x^{t+1})-f^{\inf}\big] \leq \mathbb{E}\big[f(x^t)-f^{\inf}\big]&-\frac{\gamma}{2}\mathbb{E}\Big[\big\|\nabla f(x^t)\big\|^2\Big]\\&+\left(\frac{L}{2}-\frac{1}{2\gamma}\right)\mathbb{E}\Big[\big\|x^{t+1}-x^t\big\|^2\Big]+\frac{\gamma}{2}\mathbb{E}\Big[\big\|g^{t+1}-\nabla f(x^t)\big\|^2\Big].\end{aligned}$$

Now, let us study the control variates h_i^t. Let $s>0$. Using the Peter–Paul inequality $\|a+b\|^2 \leq (1+s)\|a\|^2+(1+s^{-1})\|b\|^2$, for any vectors a and b, we have, for every $t \geq 0$ and $i \in \mathcal{I}_n$,

$$\begin{aligned}\big\|\nabla f_i(x^{t+1})-h_i^{t+1}\big\|^2 &= \big\|h_i^t-\nabla f_i(x^{t+1})+\lambda\mathcal{C}_i^t\big(\nabla f_i(x^t)-h_i^t\big)\big\|^2\\&\leq (1+s)\big\|h_i^t-\nabla f_i(x^t)+\lambda\mathcal{C}_i^t\big(\nabla f_i(x^t)-h_i^t\big)\big\|^2\\&\quad+(1+s^{-1})\big\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\big\|^2\\&\leq (1+s)\big\|h_i^t-\nabla f_i(x^t)+\lambda\mathcal{C}_i^t\big(\nabla f_i(x^t)-h_i^t\big)\big\|^2\\&\quad+(1+s^{-1})L_i^2\big\|x^{t+1}-x^t\big\|^2.\end{aligned}$$

Moreover, conditionally on x^t, h^t and $(h_i^t)_{i=1}^n$,

$$\begin{aligned}\mathbb{E}\Big[\big\|h_i^t-\nabla f_i(x^t)+\lambda\mathcal{C}_i^t\big(\nabla f_i(x^t)-h_i^t\big)\big\|^2\Big] &= \big\|h_i^t-\nabla f_i(x^t)+\lambda\mathbb{E}\big[\mathcal{C}_i^t\big(\nabla f_i(x^t)-h_i^t\big)\big]\big\|^2\\&\quad+\lambda^2\mathbb{E}\Big[\big\|\mathcal{C}_i^t\big(\nabla f_i(x^t)-h_i^t\big)-\mathbb{E}\big[\mathcal{C}_i^t\big(\nabla f_i(x^t)-h_i^t\big)\big]\big\|^2\Big]\\&\leq \big\|h_i^t-\nabla f_i(x^t)+\lambda\mathbb{E}\big[\mathcal{C}_i^t\big(\nabla f_i(x^t)-h_i^t\big)\big]\big\|^2\\&\quad+\lambda^2\omega\big\|\nabla f_i(x^t)-h_i^t\big\|^2.\end{aligned}$$

In addition,

$$\begin{aligned}\big\|h_i^t-\nabla f_i(x^t)+\lambda\mathbb{E}\big[\mathcal{C}_i^t\big(\nabla f_i(x^t)-h_i^t\big)\big]\big\| &\leq \big\|\lambda\big(h_i^t-\nabla f_i(x^t)\big)+\lambda\mathbb{E}\big[\mathcal{C}_i^t\big(\nabla f_i(x^t)-h_i^t\big)\big]\big\|\\&\quad+(1-\lambda)\big\|h_i^t-\nabla f_i(x^t)\big\|\\&\leq \lambda\eta\big\|\nabla f_i(x^t)-h_i^t\big\|+(1-\lambda)\big\|\nabla f_i(x^t)-h_i^t\big\|\\&= (1-\lambda+\lambda\eta)\big\|\nabla f_i(x^t)-h_i^t\big\|.\end{aligned}$$

Therefore, conditionally on x^t, h^t and $(h_i^t)_{i=1}^n$,

$$\mathbb{E}\Big[\big\|h_i^t-\nabla f_i(x^t)+\lambda\mathcal{C}_i^t\big(\nabla f_i(x^t)-h_i^t\big)\big\|^2\Big] \leq \big((1-\lambda+\lambda\eta)^2+\lambda^2\omega\big)\big\|\nabla f_i(x^t)-h_i^t\big\|^2$$

and

$$\begin{aligned}\mathbb{E}\Big[\big\|\nabla f_i(x^{t+1})-h_i^{t+1}\big\|^2\Big] &\leq (1+s)\big((1-\lambda+\lambda\eta)^2+\lambda^2\omega\big)\big\|\nabla f_i(x^t)-h_i^t\big\|^2\\&\quad+(1+s^{-1})L_i^2\mathbb{E}\Big[\big\|x^{t+1}-x^t\big\|^2\Big],\end{aligned}$$

so that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\big\|\nabla f_i(x^{t+1})-h_i^{t+1}\big\|^2\right] &\leq (1+s)\big((1-\lambda+\lambda\eta)^2+\lambda^2\omega\big)\frac{1}{n}\sum_{i=1}^n\big\|\nabla f_i(x^t)-h_i^t\big\|^2\\&\quad+(1+s^{-1})\tilde{L}^2\mathbb{E}\Big[\big\|x^{t+1}-x^t\big\|^2\Big].\end{aligned}$$

Let $\theta > 0$; its value will be set to $\theta^\star$ later on. We introduce the Lyapunov function, for every $t \geq 0$,

$$\Psi^t \coloneqq f(x^t) - f^{\inf} + \frac{\gamma}{2\theta}\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(x^t) - h_i^t\right\|^2.$$

Hence, for every $t \geq 0$, conditionally on x^t, h^t and $(h_i^t)_{i=1}^n$, we have

$$\begin{aligned}
\mathbb{E}\left[\Psi^{t+1}\right] &\leq \mathbb{E}\left[f(x^t) - f^{\inf}\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \\
&\quad + \frac{\gamma}{2\theta}\Big(\theta\big((1-\nu+\nu\eta)^2 + \nu^2\omega_{\mathrm{av}}\big) + (1+s)\big((1-\lambda+\lambda\eta)^2 + \lambda^2\omega\big)\Big)\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(x^t) - h_i^t\right\|^2 \\
&\quad + \left(\frac{L}{2} - \frac{1}{2\gamma} + \frac{\gamma}{2\theta}(1+s^{-1})\tilde{L}^2\right)\mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right].
\end{aligned}$$

Let $r \coloneqq (1-\lambda+\lambda\eta)^2 + \lambda^2\omega, r_{\mathrm{av}} \coloneqq (1-\nu+\nu\eta)^2 + \nu^2\omega_{\mathrm{av}}$. Set $\theta = s(1+s)\frac{r}{r_{\mathrm{av}}}$, we can rewrite the above equation as:

$$\begin{aligned}
\mathbb{E}\left[\Psi^{t+1}\right] &\leq \mathbb{E}\left[f(x^t) - f^{\inf}\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + \frac{\gamma}{2\theta}\Big(\theta r_{\mathrm{av}} + (1+s)r\Big)\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(x^t) - h_i^t\right\|^2 \\
&\quad + \left(\frac{L}{2} - \frac{1}{2\gamma} + \frac{\gamma}{2\theta}(1+s^{-1})\tilde{L}^2\right)\mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right] \\
&= \mathbb{E}\left[f(x^t) - f^{\inf}\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + \frac{\gamma}{2\theta}(1+s)^2\frac{r}{n}\sum_{i=1}^n \left\|\nabla f_i(x^t) - h_i^t\right\|^2 \\
&\quad + \left(\frac{L}{2} - \frac{1}{2\gamma} + \frac{\gamma}{2s^2}\frac{r_{\mathrm{av}}}{r}\tilde{L}^2\right)\mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right].
\end{aligned}$$

We now choose γ small enough so that

$$L - \frac{1}{\gamma} + \frac{\gamma}{s^2}\frac{r_{\mathrm{av}}}{r}\tilde{L}^2 \leq 0. \tag{16}$$

A sufficient condition for equation 16 to hold is (Richtárik et al., 2021, Lemma 5):

$$0 < \gamma \leq \frac{1}{L + \tilde{L}\sqrt{\frac{r_{\mathrm{av}}}{r}}\frac{1}{s}}. \tag{17}$$

Then, assuming that equation 17 holds, we have, for every $t \geq 0$, conditionally on x^t, h^t and $(h_i^t)_{i=1}^n$,

$$\mathbb{E}\left[\Psi^{t+1}\right] \leq \mathbb{E}\left[f(x^t) - f^{\inf}\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + \frac{\gamma}{2\theta}(1+s)^2\frac{r}{n}\sum_{i=1}^n \left\|\nabla f_i(x^t) - h_i^t\right\|^2.$$

Since we choose λ, ν to make sure $r < 1$ and we choose $s = \sqrt{\frac{1+r}{2r}} - 1$, we have $(1+s)^2 r = \frac{1+r}{2} < 1$. Then using the Tower property, we have,

$$\mathbb{E}\left[\Psi^{t+1}\right] \leq \mathbb{E}\left[\Psi^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right].$$

By summing up inequalities for $t = 0, \cdots, T-1$, we get

$$0 \leq \mathbb{E}[\Psi(T)] \leq \mathbb{E}\left[\Psi^0\right] - \frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right].$$

Multiplying both sides by $\frac{2}{\gamma T}$, after rearranging we get

$$\sum_{t=0}^{T-1}\frac{1}{T}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \leq \frac{2}{\gamma T}\Psi(0),$$

where the left hand side can be interpreted as $\mathrm{E}\left[\left\|\nabla f\left(\hat{x}^T\right)\right\|^2\right]$, where $\hat{x}^T$ is chosen from $x^0, x^1, \ldots, x^{T-1}$ uniformly at random.

□

A.1.2 Missing proof of Theorem 2.4

This proof is under EF-BF Condat et al. (2022) Theorem.1.

B EF-BV Experiments Revisit

In thi section, we revisit the experiments in EF-BV on top of the refactored code. We fixed a few minor bugs.

For all experiments, we consider the logistic regression with convex regularizer,

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i a_i^\top x\right)\right) + \frac{\lambda}{2}\|x\|^2, \tag{18}$$

where $a_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$ are the training data, and $\lambda > 0$ is the regularizer parameter. We used $\lambda = 0.1$ in all experiments.

For baselines, since EF21 Richtárik et al. (2021) outperforms all other methods, we only compare our method with EF21.

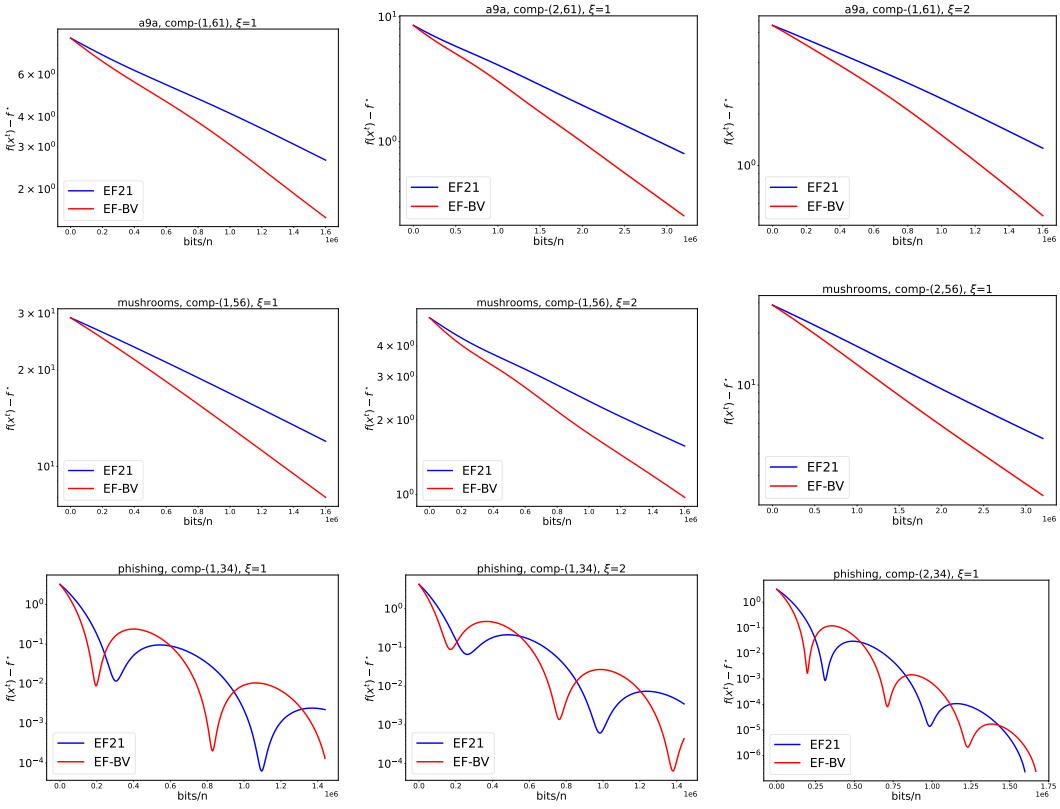

Figure 2: Comparison between EF21 and EF-BV in the convex setting. We see EF-BV outperforms EF21 on all datasets.

Next, we consider smaller number of workers to be 20 or 50. The results are shown in Fig. 5.

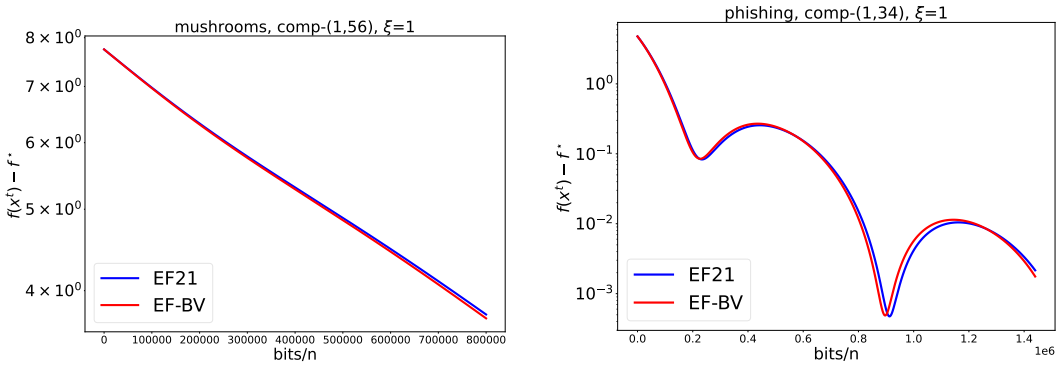

Figure 3: 20 workers.

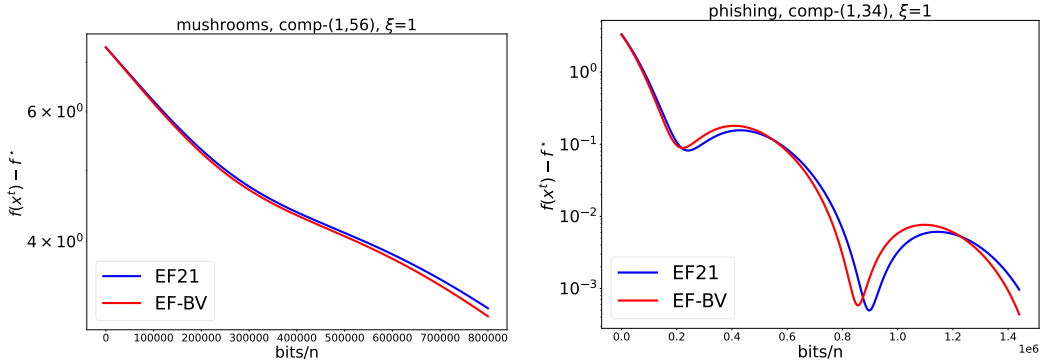

Figure 4: 50 workers.

Figure 5: Comparison between EF21 and EF-BV in the convex setting with smaller number of workers. We see the difference between EF-BV and EF21 is smaller, which is align with our theory.