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 extend EF-BV to non-convex setting.

CONTENTS

REFERENCES

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

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

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

Given any compressor $\mathcal{C}$, we can do a *bias-variance decomposition* of the compression error. That means, for every $x \in \mathbb{R}^d$,

$$\mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] = \underbrace{\left\|\mathbb{E}[\mathcal{C}(x)] - x\right\|^2}_{\text{bias}} + \underbrace{\mathbb{E}\Big[\left\|\mathcal{C}(x) - \mathbb{E}[\mathcal{C}(x)]\right\|^2\Big]}_{\text{variance}}. \tag{1}$$

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

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

where the last inequality follows from Eqn. 7 in Condat et al. (2022). In addition,

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

where the last step is obtained by using the definition of the general compressor.

Therefore,

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

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

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

Using the Tower property and we obtain the unconditioned term,

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

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

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

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

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

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 $i \in \mathcal{I}_n$,

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

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

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

In addition,

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

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

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

and

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

so that

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

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

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

Making use of r and r_{av} and setting $\theta = s(1+s)\frac{r}{r_{\mathrm{av}}}$, we can rewrite the above equation as:

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

We now choose γ small enough so that

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

A sufficient condition for equation 2 to hold is (Richtárik et al., 2021, Lemma 5):

$$0 < \gamma \leq \frac{1}{L + \tilde{L}\sqrt{\frac{r_{\mathrm{av}}}{r}}\frac{1}{s}}. \tag{3}$$

Then, assuming that equation 3 holds, we have, for every $t \geq 0$, conditionally on x^t, h^t and $(h_i^t)_{i=1}^n$,

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

We see that s must be small enough so that $(1+s)^2 r < 1$; this is the case with $s = s^\star$, so that $(1+s^\star)^2 r = \frac{r+1}{2} < 1$. Therefore, we set $s = s^\star$, and, accordingly, $\theta = \theta^\star$. Then, for every $t \geq 0$, conditionally on x^t, h^t and $(h_i^t)_{i=1}^n$,

$$\mathbb{E}\big[\Psi^{t+1}\big] \leq \max\Big(1-\gamma\mu, \frac{r+1}{2}\Big)\Psi^t.$$

□