Bidirection Compression

In lectures "Distributed Optimization: Compressed Gradient Descent" and "Improving CGD with Variance Reduced Methods: DASHA Method," we consider a problem where communication **from workers to a server** is slow. In order to improve the communication bottleneck, we proposed to use compression:

$$x^{k+1} = x^k - \frac{\gamma}{n}\sum_{i=1}^{n} \mathcal{C}_i\left(\nabla f_i(x^k)\right). \tag{1}$$

The main problem is that the server still **sends non-compressed vectors x^k to the workers**. [2] proposed a new strategy called EF21-P to compress information from the server to the workers. In this project, you have to read [2], understand it, and implement the main methods.

1. Implement an environment (code) that would emulate the communication of workers with the server (see Figure 1 in [2]).
2. Implement CGD (1) from the lectures in this environment.
3. Implement EF21-P + DIANA and EF21-P + DCGD from [2]. **Notice that these methods work with two types of compressors: biased and unbiased. Try to understand the difference between them. For instance, RandK from the lectures is an unbiased compressor, but not biased! See [1] to understand the difference and find examples.**
4. Take 2-3 optimization problems (linear regression, logistic regression, (small) neural network with log loss on MNIST) gradients and compare all methods.
5. Try to play with different numbers of workers n, datasets, and different levels of compression.
6. Beside those three methods, try to find some other interesting methods.
7. For all experiments, plot the # iterations / # of sent bits vs. accuracy/log loss and draw conclusions. For each method, plot separately the number of bits/floats that the workers send to the server and the server sends to workers during the optimization process.
8. Any ideas on how one can improve these methods and faster convergence with less communication of information?

References

[1] A. Beznosikov, S. Horváth, P. Richtárik, and M. Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50, 2023.

[2] K. Gruntkowska, A. Tyurin, and P. Richtárik. EF21-P and friends: Improved theoretical communication complexity for distributed optimization with bidirectional compression. In *International Conference on Machine Learning*, pages 11761–11807. PMLR, 2023.