Supplementary Materials for Elastic Aggregation for Federated Optimization

Dengsheng Chen[1*], Jie Hu[2,3*], Vince Junkai Tan[4], Xiaoming Wei[1], Enhua Wu[2,3,5†]
[1] Meituan [2] State Key Laboratory of Computer Science, ISCAS
[3] University of Chinese Academy of Sciences [4] Bytedance Inc. [5] University of Macau
{chendengsheng,weixiaoming}@meituan.com, hujie@ios.ac.cn, vince.tan.jun.kai@gmail.com, ehwu@um.edu.mo

1. More details about parameter sensitivity

An intuitive approach to compute sensitivity Ω^i for the i^{th} parameter is using the average method over batches:

$$\Omega^i = \sum_{x \in D_k} |g(\theta^i; x)|/|D_k|, \tag{1}$$

However, this average method may not be suitable for a real scenario in federated learning, for that training data may be collected randomly. The momentum approach described in paper is more flexible for us to do an online accumulating of parameter sensitivity. Also, we find that the momentum approach can achieve a better performance and is more robust for different τ under different tasks compared with the average approach. We also conduct an experiment on synthetic federated dataset of CIFAR-10. We show the inspired percentage of parameters during training in Fig. 1. It indicates that momentum approach prefers to keep more parameters to be restricted from client drifting. However, the average approach prefers to inspire more parameters to explore a better distribution. In federated learning, we are more eager to solve the client-drift problem caused by non-IID-ness via restricting parameters. The final accuracy in Tab. 1 also indicates a better performance of momentum approach.

	naive	elastic(average)	elastic(momentum)
Train Acc(%)	55.39	55.49	**58.74**
Test Acc(%)	61.22	61.11	**61.45**

Table 1. Performance with different parameter sensitivity computation approach.

2. Data distribution

Generating synthetic federated datasets Different distribution has a very large influence on the final performance of federated optimization. The Dirichlet distribution is used on the label ratios to ensure uneven label distributions among clients for non-IID splits, as in [12]. This can generate nonIIDness with an unbalanced sample number on each label. The Dirichlet distribution is a density over a K dimensional vector p whose K components are positive and sum to 1. Dirichlet can support the probabilities of a K-way categorical event. In federated learning, we can view K clients' sample numbers obeying the Dirichlet distribution. You can check here[1] for more details of the Dirichlet distribution. To generate unbalanced data, we sample the number of data points from a log-normal distribution. Controlling the variance of log-normal distribution gives unbalanced data.

We use the above-introduced approach to generate synthetic federated datasets for MNIST, CIFAR-10, and CINIC-10 in our paper.

*Equal contribution.
†Corresponding author. This work is supported in part by NSFC Grants (62072449).
[1]https://en.wikipedia.org/wiki/Dirichlet_distribution

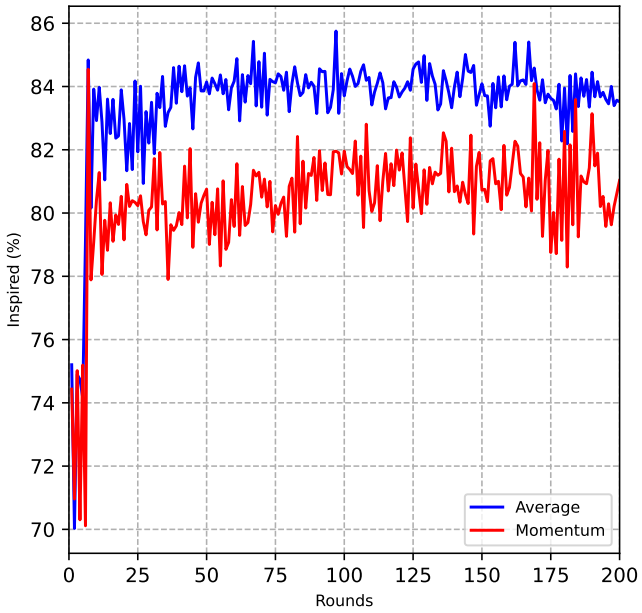

Figure 1. Percentage of parameters boosted during training.

Fed-CIFAR100 The dataset is derived from the CIFAR-100 dataset[2]. The training and testing examples are partitioned across 500 and 100 clients (respectively). No clients share any data samples, so it is a true partition of CIFAR-100. The train clients have string client IDs in the range [0-499], while the test clients have string client IDs in the range [0-99]. The train clients form a true partition of the CIFAR-100 training split, while the test clients form a true partition of the CIFAR-100 testing split. The data partitioning is done using a hierarchical Latent Dirichlet Allocation (LDA) process, referred to as the Pachinko Allocation Method [6]. This method uses a two-stage LDA process, where each client has an associated multinomial distribution over the coarse labels of CIFAR-100, and a coarse-to-fine label multinomial distribution for that coarse label over the labels under that coarse label. The coarse label multinomial is drawn from a symmetric Dirichlet with parameter 0.1, and each coarse-to-fine multinomial distribution is drawn from a symmetric Dirichlet with parameter 10. Each client has 100 samples. To generate a sample for the client, we first select a coarse label by drawing from the coarse label multinomial distribution and then draw a fine label using the coarse-to-fine multinomial distribution. We then randomly draw a sample from CIFAR-100 with that label (without replacement). If this exhausts the set of samples with this label, we remove the label from the coarse-to-fine multinomial and renormalize the multinomial distribution.

Fed-EMNIST This dataset is derived from the Leaf [1] repository[3] pre-processing of the Extended MNIST dataset, grouping examples by the writer. This dataset does not include some additional preprocessing that MNIST includes, such as size normalization and centering. In the Federated EMNIST data, the value of 1.0 corresponds to the background, and 0.0 corresponds to the color of the digits themselves. It contains 3,400 users, 62 label classes, and 671,585 training examples, 77,483 testing examples. Rather than holding out specific users, each user's examples are split across train and test so that all users have at least one example in the train and one example in the test. Writers that had less than 2 examples are excluded from the data set.

3. Federated optimizer with elastic aggregation

Federated Average with Momentum (FedAvgM) and Elastic Aggregation have been presented in Algorithms. 1. FedProx with Elastic Aggregation has been presented in Algorithms. 2. FedAvgM is an enhancement of FedAvg on the server side, and FedProx is an enhancement of FedAvg on the client side. Elastic aggregation can work well with other complementary approaches designed for the client-side or server-side. In Fig. 2, we show the convergence speed with different optimizers.

[2]https://www.cs.toronto.edu/ kriz/cifar.html
[3]https://github.com/TalwalkarLab/leaf

Algorithm 1: FedAvg with Momentum [9] and Elastic Aggregation

A variable with a superscript i indicates the i^{th} element of the variable. A variable with a subscript k indicates the variable from k^{th} client. η, η' are learning rates of server and clients respectively. μ, μ', τ are the hyper-parameters. $\theta, \theta_k \in \mathbb{R}^n$ are the server's and the k^{th} client's parameters respectively. $\Omega \in \mathbb{R}^n$ is the aggregated parameter sensitivity. $\Omega_k \in \mathbb{R}^n$ is the parameter sensitivity on the k^{th} client. $m \in \mathbb{R}^n$ is the momentum vector.

Initialize θ
Initialize $m \leftarrow 0$
$B_k \leftarrow$ Sample a subset of training data D_k.
$D_k \leftarrow$ Drop the samples of B_k from D_k.
for *each round* **do**
  for *each activated client* k **do**
    Initialize Ω_k as zeros.
    for *each batch data* $x \in B_k$ **do**
      $g = \nabla ||F(\theta; x)||_2^2$
      for $i \in [1, \cdots, n]$ **do**
        $\Omega_k^i \leftarrow \mu \Omega_k^i + (1 - \mu)|g^i|$
    $\theta_k \leftarrow \theta$
    for *each epoch* **do**
      for *each batch data* $x \in D_k$ **do**
        $\theta_k \leftarrow \theta_k - \eta' \nabla \ell_k(F(\theta_k; x))$
    $\Delta_k = \theta_k - \theta$
  $w_k \leftarrow |D_k| / \sum_k |D_k|; \Omega = \sum_k (w_k \cdot \Omega_k); \Omega' = \max(\Omega)$
  for $i \in [1, \cdots, n]$ **do**
    $\zeta^i = 1 + \tau - \Omega^i / \Omega'$
    $\Delta^i = \zeta^i \cdot \sum_k (w_k \cdot \Delta_k^i)$
    $m^i \leftarrow \mu' m^i + (1 - \mu') \Delta^i$
    $\theta^i \leftarrow \theta^i - \eta \cdot m^i$

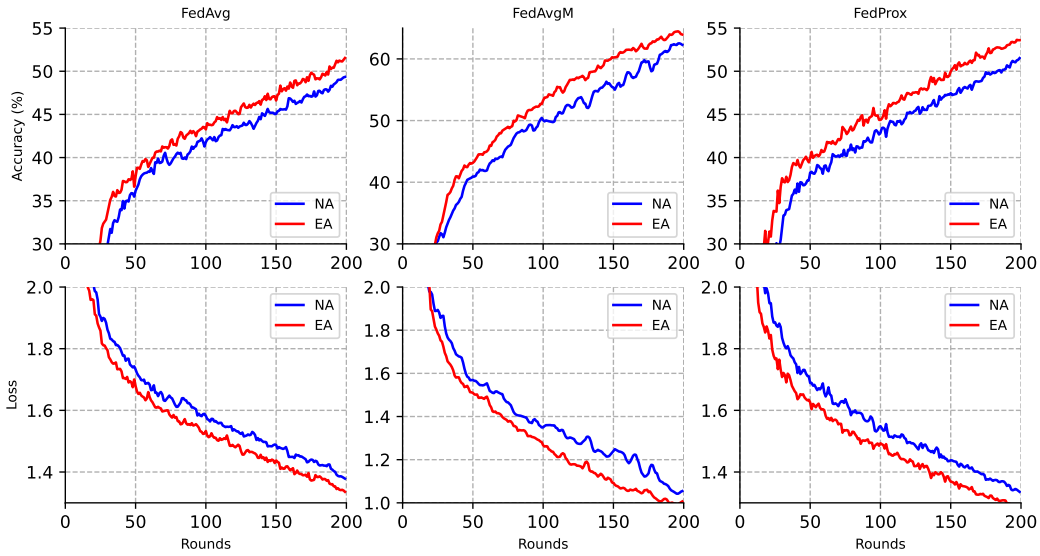

Figure 2. Elastic aggregation can be easily integrated with different federated optimizers, achieving performance improvements.

Algorithm 2: FedProx [5] with Elastic Aggregation

A variable with a superscript i indicates the i^{th} element of the variable. A variable with a subscript k indicates the variable from k^{th} client. η, η' are learning rates of server and clients respectively. μ, τ are the hyper-parameters. ρ is penalty coefficient of FedProx. $\theta, \theta_k \in \mathbb{R}^n$ are the server's and the k^{th} client's parameters respectively. $\Omega \in \mathbb{R}^n$ is the aggregated parameter sensitivity. $\Omega_k \in \mathbb{R}^n$ is the parameter sensitivity on the k^{th} client.

Initialize θ
$B_k \leftarrow$ Sample a subset of training data D_k.
$D_k \leftarrow$ Drop the samples of B_k from D_k.
for *each round* **do**
  for *each activated client* k **do**
    Initialize Ω_k as zeros.
    for *each batch data* $x \in B_k$ **do**
      $g = \nabla ||F(\theta; x)||_2^2$
      for $i \in [1, \cdots, n]$ **do**
        $\Omega_k^i \leftarrow \mu \Omega_k^i + (1 - \mu)|g^i|$
    $\theta_k \leftarrow \theta$
    for *each epoch* **do**
      for *each batch data* $x \in D_k$ **do**
        $\theta_k \leftarrow \theta_k - \eta' \nabla \ell_k(F(\theta_k; x)) + \rho(\theta - \theta^k)$
    $\Delta_k = \theta_k - \theta$
  $w_k \leftarrow |D_k| / \sum_k |D_k|$; $\Omega = \sum_k (w_k \cdot \Omega_k)$; $\Omega' = \max(\Omega)$
  for $i \in [1, \cdots, n]$ **do**
    $\zeta^i = 1 + \tau - \Omega^i / \Omega'$
    $\Delta^i = \zeta^i \cdot \sum_k (w_k \cdot \Delta_k^i)$
    $\theta^i \leftarrow \theta^i - \eta \cdot \Delta^i$

4. The computational overhead of parameter sensitivity

This computational overhead can be neglectable in terms of total computational cost in the training phase. Moreover, parameter sensitivity is not required in the inference phase.

In a training task, the parameter sensitivity is only calculated once for each round. Suppose that each round contains e epochs, and several backward perform on a small fraction μ of training instances, which is enough to precisely estimate the parameter sensitivity. Thus, the additional computational cost can be roughly given by $\frac{\mu}{2 \times e}$. Empirically, we set $\mu = 10\%$ and $e = 10$. The extra cost only takes 0.5% against the total training cost.

5. Communication overhead

As for the communication budget, we introduce no overhead for downloading the global model but require an extra communication overhead for uploading the parameter sensitivities. And this overhead seems inevitable. Such as the well-known related method mentioned in Table 5, the SCAFFOLD2019 [3] also introduces such communication overhead. We list the additional communication overhead using FedAvg [7] as baseline in Tab. 2 (downloading parameters notes as 1x and uploading parameters also notes as 1x, so the FedAvg is 2x in total):

FedAvg	FedAvgM	FedProx	SCAFFOLD	AdaOpt	PFNM	Ours
2x	2x	2x	4x	2x	2x	3x

Table 2. Communication overhead.

	Pros	Cons
FedAvg	Efficient, Robust	Slow convergence, Limited upper performance
FedProx [5]	Fast convergence, Better performance	Light extra overhead
SCAFFOLD [3]	Fast convergence	Not robust
AdaOpt [9]	Fast convergence, Excellent performance	Considerable extra overhead
PFNM [12]	Alleviate client drift, Excellent performance, Fast convergence	Complex implementation, Considerable extra overhead
Ours	Alleviate client drift, Excellent performance, Robust, Fast convergence	Light extra overhead

Table 3. Compare to prior works in pros/cons.

6. Compare to prior works in pros/cons

Here we list the pros/cons of several related works in Tab. 3. From the table, our proposed Elastic Aggregation basically incorporates the advantages of other methods without any defect except a light extra computational overhead.

7. Technicalities

We formalize the problem as minimizing a sum of stochastic functions like [3], with only access to stochastic samples:

$$\min_{x\in\mathbb{R}^d}\{f(\boldsymbol{x}) := \frac{1}{N}\sum_{i=1}^{N}(f_i(\boldsymbol{x}) := \mathbb{E}_{\zeta_i}[f_i(\boldsymbol{x};\zeta_i)])\}. \tag{2}$$

The functions f_i represents the loss function on client i. All our results can be easily extended to the weighted case.

We assume that f is bounded from below by $f^\star$ and f_i is β-smooth. Further, we assume $g_i(\boldsymbol{x}) := \nabla f_i(\boldsymbol{x};\zeta_i)$ is an unbiased stochastic gradient of f_i with variance bounded by σ^2. For some results, we assume $\mu \geq 0$ (strong) convexity. Note that σ only bounds the variance within clients.

Now, we examine some additional definitions and introduce some technical lemmas.

7.1. Additional definitions

We make precise a few definitions and explain some of their implications.

A1 There exists constants $G \geq 0$ and $B \geq 1$ such that

$$\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(\boldsymbol{x})\|^2 \leq G^2 + B^2\|\nabla f(\boldsymbol{x})\|^2, \forall \boldsymbol{x} \tag{3}$$

A2 f_i is μ-convex for $\mu \geq 0$ and satisfies:

$$\langle \nabla f_i(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle \geq -(f_i(\boldsymbol{x}) - f_i(\boldsymbol{y}) + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2), \forall i, \boldsymbol{x}, \boldsymbol{y}. \tag{4}$$

Here, we allow that $\mu = 0$ (we refer to this case as the general convex case as opposed to strongly convex). It is also possible to generalize all proofs here to the weaker notion of PL-strong convexity [2].

A3 $g_i(\boldsymbol{x}) := \nabla f_i(\boldsymbol{x}; \zeta_i)$ is unbiased stochastic gradient of f_i with bounded variance

$$\mathbb{E}_{\zeta_i}[\|g_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x})\|^2] \leq \sigma^2, \forall i, \boldsymbol{x}. \tag{5}$$

Note that (A3) only bounds the variance within the same client, but not the variance across the clients.

A4 $\{f_i\}$ are β-smooth and satisfy:

$$\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\| \leq \beta \|\boldsymbol{x} - \boldsymbol{y}\|, \forall i, \boldsymbol{x}, \boldsymbol{y}. \tag{6}$$

The assumption (A4) also implies the following quadratic upper bound on f_i

$$f_i(\boldsymbol{y}) \leq f_i(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\beta}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

If additionally the function $\{f_i\}$ are convex and $x^\star$ is an optimum of f, (A4) implies

$$\frac{1}{2\beta N} \sum_{i=1}^{N} \|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}^\star)\|^2 \leq f(\boldsymbol{x}) - f^\star.$$

Further, if f_i is twice-differentiable, (A4) implies that

$$\|\nabla^2 f_i(\boldsymbol{x})\| \leq \beta, \forall \boldsymbol{x}.$$

7.2. Some technical lemmas

Now we cover some technical lemmas which are useful for computations later on. The two lemmas below are useful to unroll recursions and derive convergence rates. The first one is a slightly improved (and simplified) version of ([10], Theorem 2). It is straightforward to remove the additional logarithmic terms if we use a varying step-size ([4], Lemma 13).

Lemma1 (linear convergence rate).
For every non-negative sequence $\{d_{r-1}\}_{r\geq 1}$ and any parameters $\mu > 0, \eta_{\max} \in (0, \frac{1}{\mu}], c \geq 0, R \geq \frac{1}{2\eta_{\max}\mu}$, there exists a constant step-size $\eta \leq \eta_{\max}$ and weights $w_r := (1-\mu\eta)^{1-r}$ such that for $W_R := \sum_{r=1}^{R+1} w_r$,

$$\Psi_R := \frac{1}{W_R} \sum_{r=1}^{R+1} \left(\frac{w_r}{\eta}(1-\mu\eta)d_{r-1} - \frac{w_r}{\eta} d_r + c\eta w_r\right) = \tilde{\mathcal{O}}\left(\mu d_0 \exp(-\mu\eta_{\max} R) + \frac{c}{\mu R}\right). \tag{7}$$

Proof. By substituting the value of w_r, we observe that we end up with a telescoping sum and estimate

$$\Psi_R = \frac{1}{\eta W_R} \sum_{r=1}^{R+1} (w_{r-1} d_{r-1} - w_r d_r) + \frac{c\eta}{W_R} \sum_{r=1}^{R+1} w_r \leq \frac{d_0}{\eta W_R} + c\eta.$$

When $R > \frac{1}{2\mu\eta}$, $(1-\mu\eta)^R \leq \exp(-\mu\eta R) \leq \frac{2}{3}$. For such an R, we can lower bound ηW_R using

$$\eta W_R = \eta(1-\mu\eta)^{-R} \sum_{r=0}^{R} (1-\mu\eta)^r = \eta(1-\mu\eta)^{-R} \frac{1-(1-\mu\eta)^R}{\mu\eta} \geq (1-\mu\eta)^{-R} \frac{1}{3\mu}.$$

This proves that for all $R \geq \frac{1}{2\mu\eta}$,

$$\Psi_R \leq 3\mu d_0 (1-\mu\eta)^R + c\eta \leq 3\mu d_0 \exp(-\mu\eta R) + c\eta.$$

The lemma now follows by carefully tuning η. Consider the following two cases depending on the magnitude of R and $\eta_{\max}$:

- Suppose $\frac{1}{2\mu R} \leq \eta_{\max} \leq \frac{\log(\max(1,\mu^2 R d_0/c))}{\mu R}$. Then we can choose $\eta = \eta_{\max}$,

$$\Psi_R \leq 3\mu d_0 \exp[-\mu\eta_{\max} R] + c\eta_{\max} \leq 3\mu d_0 \exp[-\mu\eta_{\max} R] + \tilde{\mathcal{O}}(\frac{c}{\mu R}).$$

- Instead if $\eta_{\max} > \frac{\log(\max(1,\mu^2 R d_0/c))}{\mu R}$, we pick $\eta = \frac{\log(\max(1,\mu^2 R d_0/c))}{\mu R}$ to claim that

$$\Psi_R \leq 3\mu d_0 \exp[-\log(\max(1,\mu^2 R d_0/c))] + \tilde{\mathcal{O}}(\frac{c}{\mu R}) \leq \tilde{\mathcal{O}}(\frac{c}{\mu R}).$$

The next lemma is useful to derive convergence rates for general convex functions ($\mu = 0$) and non-convex functions.

Lemma 2 (sub-linear convergence rate).
For every non-negative sequence $\{d_{r-1}\}_{r\geq 1}$ any parameters $\eta_{\max} \geq 0, c \geq 0, R \geq 0$, there exists a constant step-size $\eta \leq \eta_{\max}$ and weights $w_r = 1$ such that,

$$\Psi_R := \frac{1}{R+1}\sum_{r=1}^{R+1}(\frac{d_{r-1}}{\eta} - \frac{d_r}{\eta} + c_1\eta + c_2\eta^2) \leq \frac{d_0}{\eta_{\max}(R+1)} + \frac{2\sqrt{c_1 d_0}}{\sqrt{R+1}} + 2(\frac{d_0}{R+1})^{\frac{2}{3}} c_2^{\frac{1}{3}}. \tag{8}$$

Proof. Unrolling the sum, we can simplify

$$\Psi_R \leq \frac{d_0}{\eta(R+1)} + c_1\eta + c_2\eta^2.$$

Similar to the strongly convex case (Lemma 1), we distinguish the following cases:

- When $R+1 \leq \frac{d_0}{c_1\eta_{\max}^2}$, and $R+1 \leq \frac{d_0}{c_2\eta_{\max}^3}$, we pick $\eta = \eta_{\max}$ to claim

$$\Psi_R \leq \frac{d_0}{\eta_{\max}(R+1)} + c_1\eta_{\max} + c_2\eta_{\max}^2 \leq \frac{d_0}{\eta_{\max}(R+1)} + \frac{\sqrt{c_1 d_0}}{\sqrt{R+1}} + (\frac{d_0}{R+1})^{\frac{2}{3}} c_2^{\frac{1}{2}}.$$

- In the other case, we have $\eta_{\max}^2 \geq \frac{d_0}{c_1(R+1)}$ or $\eta_{\max}^2 \geq \frac{d_0}{c_2(R+1)}$. We choose $\eta = \min\{\sqrt{\frac{d_0}{c_1(R+1)}}, \sqrt[3]{\frac{d_0}{c_2(R+1)}}\}$ to prove

$$\Psi_R \leq \frac{d_0}{\eta(R+1)} + c\eta = \frac{2\sqrt{c_1 d_0}}{\sqrt{R+1}} + 2\sqrt[3]{\frac{d_0^2 c_2}{(R+1)^2}}.$$

Next, we state a relaxed triangle inequality true for the squared ℓ_2 norm.

Lemma 3 (relaxed triangle inequality).
Let $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_\tau\}$ be τ vectors in $\mathbb{R}^d$. The the following are true:

$$\begin{cases} \|\boldsymbol{v}_i + \boldsymbol{v}_j\|^2 \leq (1+a)\|\boldsymbol{v}_i\|^2 + (1+\frac{1}{a})\|\boldsymbol{v}_j\|^2, & \forall a > 0, \\ \|\sum_{i=1}^{\tau} \boldsymbol{v}_i\|^2 \leq \tau \sum_{i=1}^{\tau} \|\boldsymbol{v}_i\|^2. \end{cases} \tag{9}$$

Proof. The proof of the first statement for any $a > 0$ follows from the identity:

$$\|\boldsymbol{v}_i + \boldsymbol{v}_j\|^2 = (1+a)\|\boldsymbol{v}_i\|^2 + (1+\frac{1}{a})\|\boldsymbol{v}_j\|^2 - \|\sqrt{a}\boldsymbol{v}_i + \frac{1}{\sqrt{a}}\boldsymbol{v}_j\|^2.$$

For the second inequality, we use the convexity of $\boldsymbol{x} \to \|\boldsymbol{x}\|^2$ and Jensen's inequality

$$\|\frac{1}{\tau}\sum_{i=1}^{\tau} \boldsymbol{v}_i\|^2 \leq \frac{1}{\tau}\sum_{i=1}^{\tau} \|\boldsymbol{v}_i\|^2.$$

Next we state an elementary lemma about expectations of norms of random vectors.

Lemma 4 (separating mean and variance).

Let $\Xi_1, \cdots, \Xi_\tau$ be τ random variables in $\mathbb{R}^d$ which are not necessarily independent. First suppose that their mean is $\mathbb{E}[\Xi_i] = \xi_i$ and variance is bounded as $\mathbb{E}[\|\Xi_i - \xi_i\|^2] \leq \sigma^2$. Then, the following holds

$$\mathbb{E}[\|\sum_{i=1}^{\tau} \Xi_i\|^2] \leq \|\sum_{i=1}^{\tau} \xi_i\|^2 + \tau^2\sigma^2. \tag{10}$$

Now instead suppose that their conditional mean is $\mathbb{E}[\Xi_i|\Xi_{i-1}, \cdots, \Xi_1] = \xi_i$, i.e. the variables $\{\Xi_i - \xi_i\}$ form a martingale difference sequences, and the variance is bounded by $\mathbb{E}[\|\Xi_i - \xi_i\|^2] \leq \sigma^2$ as before. Then we can show the tighter bound

$$\mathbb{E}[\|\sum_{i=1}^{\tau} \Xi_i\|^2] \leq 2\|\sum_{i=1}^{\tau} \xi_i\|^2 + 2\tau\sigma^2. \tag{11}$$

Proof. For any random variable $X, \mathbb{E}[X^2] = (\mathbb{E}[X - \mathbb{E}[X]])^2 + (\mathbb{E}[X])^2$ implying

$$\mathbb{E}[\|\sum_{i=1}^{\tau} \Xi_i\|^2] = \|\sum_{i=1}^{\tau} \xi_i\|^2 + \mathbb{E}[\|\sum_{i=1}^{\tau} \Xi_i - \xi_i\|^2].$$

Expanding the above expression using relaxed triangle inequality (Lemma 3) proves the first claim:

$$\mathbb{E}[\|\sum_{i=1}^{\tau} \Xi_i - \xi_i\|^2] \leq \tau \sum_{i=1}^{\tau} \mathbb{E}[\|\Xi_i - \xi_i\|^2] \leq \tau^2\sigma^2.$$

For the second statement, ξ_i is not deterministic and depends on $\Xi_{i-1}, \cdots, \Xi_1$. Hence we have to resort to the cruder relaxed triangle inequality to claim

$$\mathbb{E}[\|\sum_{i=1}^{\tau} \Xi_i\|^2] \leq 2\|\sum_{i=1}^{\tau} \xi_i\|^2 + 2\mathbb{E}[\|\sum_{i=1}^{\tau} \Xi_i - \xi_i\|^2]$$

and then use the tighter expansion of the second term:

$$\mathbb{E}[\|\sum_{i=1}^{\tau} \Xi_i - \xi_i\|^2] = \sum_{i,j} \mathbb{E}[(\Xi_i - \xi_i)^T(\Xi_j - \xi_j)] = \sum_i \mathbb{E}[\|\Xi_i - \xi_i\|^2] \leq \tau\sigma^2.$$

The cross terms in the above expression have zero mean since $\{\Xi_i - \xi_i\}$ form a martingale difference sequence.

8. Properties of convex functions

We now study two lemmas which hold for any smooth and strongly-convex functions. The first is a generalization of the standard strong convexity inequality (A2), but can handle gradients computed at slightly perturbed points.

Lemma 5 (perturbed strong convexity).

The following holds for any β-smooth and μ-strongly convex function h:

$$\langle \nabla h(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{y} \rangle \geq h(\boldsymbol{z}) - h(\boldsymbol{y}) + \frac{\mu}{4}\|\boldsymbol{y} - \boldsymbol{z}\|^2 - \beta\|\boldsymbol{z} - \boldsymbol{x}\|^2, \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in h \tag{12}$$

Proof. Given any $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z}$, we get the following two inequalities using smoothness and strong convexity of h:

$$\langle \nabla h(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{x} \rangle \geq h(\boldsymbol{z}) - h(\boldsymbol{x}) - \frac{\beta}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2,$$
$$\langle \nabla h(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq h(\boldsymbol{x}) - h(\boldsymbol{y}) + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

Further, applying the relaxed triangle inequality gives

$$\frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \geq \frac{\mu}{4}\|\boldsymbol{y} - \boldsymbol{x}\|^2 = \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2.$$

Combining all the inequalities together we have

$$\langle \nabla h(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{y} \rangle \geq h(\boldsymbol{z}) - h(\boldsymbol{y}) + \frac{\mu}{4}\|\boldsymbol{y} - \boldsymbol{z}\|^2 - \frac{\beta + \mu}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2.$$

The lemma follows since $\beta \geq \mu$.

Lemma 6 (contractive mapping).
For any β-smooth and μ -strongly convex function h and step-size $\eta \leq \frac{1}{\beta}$, the following is true

$$\|\boldsymbol{x} - \eta \nabla h(\boldsymbol{x}) - \boldsymbol{y} + \eta \nabla h(\boldsymbol{y})\|^2 \leq (1 - \mu\eta)\|\boldsymbol{x} - \boldsymbol{y}\|^2, \forall \boldsymbol{x}, \boldsymbol{y} \in h. \tag{13}$$

Proof.

$$\begin{aligned}\|\boldsymbol{x} - \eta \nabla h(\boldsymbol{x}) - \boldsymbol{y} + \eta \nabla h(\boldsymbol{y})\|^2 &= \|\boldsymbol{x} - \boldsymbol{y}\|^2 + \eta^2 \|\nabla h(\boldsymbol{x}) - \nabla h(\boldsymbol{y})\|^2 - 2\eta \langle \nabla h(\boldsymbol{x}) - \nabla h(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \\ &\leq \|\boldsymbol{x} - \boldsymbol{u}\|^2 + (\eta^2 \beta - 2\eta)\langle \nabla h(\boldsymbol{x}) - \nabla h(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle.\end{aligned}$$

Recall our bound on the step-size $\eta \leq \frac{1}{\beta}$ which implies that $(\eta^2 \beta - 2\eta) \leq -\eta$. Finally, apply the μ-strong convexity of h to get

$$-\eta \langle \nabla h(\boldsymbol{x}) - \nabla h(\boldsymbol{y}), (x) - \boldsymbol{y} \rangle \leq -\eta\mu \|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

9. Convergence of elastic aggregation

Here we will give a general convergence rate for elastic aggregation and in the next section, we will use it to analyze the ideal convergence rate for our proposed elastic aggregation.

9.1. elastic aggregation

We outline the general aggregation method in Algorithm 3. In round r we sample $\mathcal{S}^r \subseteq [N]$ clients with $|\mathcal{S}^r| = S$ and then perform the following updates:
Step 1: Starting from the shared global parameters $\boldsymbol{y}_{i,0}^r = \boldsymbol{x}^{r-1}$, we update the local parameters for $k \in [K]$

$$\boldsymbol{y}_{i,k}^r = \boldsymbol{y}_{i,k-1}^r - \eta_l g_i(\boldsymbol{y}_{i,k-1}^r).$$

Step 2: Compute the new global parameters using only updates from the clients $i \in \mathcal{S}^r$ and a global step-size η_g:

$$\boldsymbol{x}^r = \boldsymbol{x}^{r-1} + \frac{1}{S}\phi\eta_g \sum_{i \in \mathcal{S}^r} (\boldsymbol{y}_{i,K}^r - \boldsymbol{x}^{r-1}). \tag{14}$$

where ϕ is the parameter sensitivities respect to $\boldsymbol{x}^{r-1}$. Finally, for some weights $\{w_r\}$, we output $\bar{\boldsymbol{x}}^R = \boldsymbol{x}^{r-1}$ with probability $\frac{w_r}{\sum_\tau w_\tau}$ for $r \in \{1, \cdots, R+1\}$.

9.2. Bounding heterogeneity

Recall our bound on the gradient dissimilarity:

$$\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(\boldsymbol{x})\|^2 \leq G^2 + B^2 \|\nabla f(\boldsymbol{x})\|^2.$$

If $\{f_i\}$ are convex, we can relax the assumption to

$$\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(\boldsymbol{x})\|^2 \leq G^2 + 2\beta B^2 (f(\boldsymbol{x}) - f^\star).$$

We defined two variants of the bounds on the heterogeneity depending of whether the functions are convex or not. Suppose that the functions f is indeed convex as in (7.1) and β-smooth as in (7.1), then it is straightforward to see that (9.2) implies (9.2). Suppose that the functions $\{f_1, \cdots, f_N\}$ are convex and β-smooth. Then (9.2) is satisfied with $B^2 = 2$ since

Algorithm 3: Simplified elastic aggregation

For the convenience of representation, we simplify or omit some extra hyper parameters (i.e. τ, μ) and processes that will not affect the convergence analysis.

server input: initial $\boldsymbol{x}$, and global step-size η_g
client's input: local step-size η_l
for *each round* $r = 1, \cdots, R$ **do**
  sample clients $\mathcal{S} \subseteq \{1, \cdots, N\}$
  communicate $\boldsymbol{x}$ to all clients $i \in \mathcal{S}$
  for *client* $i \in \mathcal{S}$ **do**
    initialize local model $\boldsymbol{y}_i \leftarrow \boldsymbol{x}$
    accumulate local parameter sensitivities $\phi_i \leftarrow \Phi(\boldsymbol{x})$
    for $k = 1, \cdots, K$ **do**
      compute mini-batch gradient $g_i(\boldsymbol{y}_i)$
      $\boldsymbol{y}_i \leftarrow \boldsymbol{y}_i - \eta_l g_i(\boldsymbol{y}_i)$
    communicate $\Delta \boldsymbol{y}_i \leftarrow \boldsymbol{y}_i - \boldsymbol{x}$ and ϕ_i
  $\Delta x \leftarrow \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \Delta \boldsymbol{y}_i$
  $\phi \leftarrow \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \phi_i$
  $\boldsymbol{x} \leftarrow \boldsymbol{x} + \phi \eta_g \Delta \boldsymbol{x}$

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(\boldsymbol{x})\|^2 &\leq \frac{2}{N}\sum_{i=1}^{N}\|\nabla f_i(\boldsymbol{x}^\star)\|^2 + \frac{2}{N}\sum_{i=1}^{N}\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}^\star)\|^2 \\ &\leq \underbrace{\frac{2}{N}\sum_{i=1}^{N}\|\nabla f_i(\boldsymbol{x}^\star)\|^2}_{=:\sigma_f^2} + 4\beta(f(\boldsymbol{x}) - f^\star).\end{aligned}$$

Thus, (9.2) is equivalent to the heterogeneity assumption of ([8]) with $G^2 = \sigma_f^2$. Instead, if the functions are possibly non-convex, then $G = \epsilon$ corresponds to the local dissimilarity defined in ([5]). Note that assuming G is negligible is quite strong and corresponds to the strong-growth condition ([11]).

9.3. Rates of convergence

Theorem I. *Suppose that the functions $\{f_i\}$ satisfies assumptions A1, A3 and A4. Then, in each of the following cases, there exist weights $\{w_r\}$ and local step-sizes η_l such that for any $\phi\eta_g \geq 1$ the output of general aggregation $\bar{\boldsymbol{x}}^R$ satisfies*

Strongly convex: f_i satisfies (A2) for $\mu > 0, \eta_l \leq \frac{1}{8(1+B^2)\beta K \phi \eta_g}, R \geq \frac{8(1+B^2)\beta}{\mu}$ then

$$\mathbb{E}[f(\bar{\boldsymbol{x}}^R)] - f(\boldsymbol{x}^\star) \leq \tilde{\mathcal{O}}\left(\frac{M^2}{\mu RKS} + \frac{\beta G^2}{\mu^2 R^2} + \mu D^2 \exp\left(-\frac{\mu}{16(1+B^2)\beta}R\right)\right), \tag{15}$$

General convex: f_i satisfies (A2) for $\mu = 0, \eta_l \leq \frac{1}{(1+B^2)8\beta K \phi \eta_g}, R \geq 1$ then

$$\mathbb{E}[f(\bar{\boldsymbol{x}}^R)] - f(\boldsymbol{x}^\star) \leq \mathcal{O}\left(\frac{MD}{\sqrt{RKS}} + \frac{D^{4/3}(\beta G^2)^{1/3}}{(R+1)^{2/3}} + \frac{B^2\beta D^2}{R}\right), \tag{16}$$

Non-convex: f_i satisfies (A1) and $\eta_l \leq \frac{1}{(1+B^2)8\beta K \phi \eta_g}$, then

$$\mathbb{E}[\|\nabla f(\bar{\boldsymbol{x}}^R)\|^2] \leq \mathcal{O}\left(\frac{\beta M\sqrt{F}}{\sqrt{RKS}} + \frac{F^{2/3}(\beta G^2)^{1/3}}{(R+1)^{2/3}} + \frac{B^2\beta F}{R}\right), \tag{17}$$

where $M^2 := \sigma^2(1 + \frac{S}{\phi^2\eta_g^2}) + K(1 - \frac{S}{N})G^2, D := \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2$, *and* $F := f(\boldsymbol{x}^0) - f(\boldsymbol{x}^\star)$.

9.4. Proof of convergence

We will only prove the rate of convergence for convex functions here. The corresponding rates for non-convex functions are easy to derive following the techniques in the rest of the paper.

Lemma 7. (one round progress) *Suppose our functions satisfies assumptions (A1) and (A2)-(A4). For any step-size satisfying* $\eta_l \leq \frac{1}{(1+B^2)8\beta K\eta_g}$ *and effective step-size* $\tilde{\eta} := K\phi\eta_g\eta_l$, the updates of general aggregation satisfy

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{x}^r - \boldsymbol{x}^\star]\|^2 \leq &(1 - \frac{\mu\tilde{\eta}}{2})\mathbb{E}[\|\boldsymbol{x}^{r-1} - \boldsymbol{x}^\star\|^2] + (\frac{1}{KS})\tilde{\eta}^2\sigma^2 \\ &+ (1 - \frac{S}{N})\frac{4\tilde{\eta}^2}{S}G^2 - \tilde{\eta}(\mathbb{E}[f(\boldsymbol{x}^{r-1})] - f(\boldsymbol{x}^\star)) + 3\beta\tilde{\eta}\mathcal{E}_r,\end{aligned}$$

where $\mathcal{E}_r$ *is the drift caused by the local updates on the clients defined to be*

$$\mathcal{E}_r := \frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}_r[\|\boldsymbol{y}_{i,k}^r - \boldsymbol{x}^{r-1}\|^2].$$

Proof. We start with the observation that the updates (10) and (11) imply that the server update in round r can be written as below (dropping the superscripts everywhere)

$$\begin{cases} \Delta\boldsymbol{x} = -\frac{\tilde{\eta}}{KS}\sum_{k,i\in\mathcal{S}} g_i(\boldsymbol{y}_{i,k-1}), \\ \mathbb{E}[\Delta\boldsymbol{x}] = -\frac{\tilde{\eta}}{KN}\sum_{k,i}\mathbb{E}[\nabla f_i(\boldsymbol{y}_{i,k-1})]. \end{cases} \tag{18}$$

We adopt the convention that summations are always over $k \in [K]$ or $i \in [N]$ unless otherwise stated. Expanding using above observing, we proceed as[4]

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{x} + \Delta\boldsymbol{x} - \boldsymbol{x}^\star\|^2] =&\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2 - \frac{2\tilde{\eta}}{KN}\sum_{k,i}\langle\nabla f_i(\boldsymbol{y}_{i,k-1}), \boldsymbol{x} - \boldsymbol{x}^\star\rangle \\ &+ \tilde{\eta}^2\mathbb{E}_{r-1}[\|\frac{1}{KS}\sum_{k,i\in\mathcal{S}} g_i(\boldsymbol{y_{(i,k-1)}})\|^2] \\ \leq&\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2 \underbrace{-\frac{2\tilde{\eta}}{KN}\sum_{k,i}\langle\nabla f_i(\boldsymbol{y}_{i,k-1}), \boldsymbol{x} - \boldsymbol{x}^\star\rangle}_{\mathcal{A}_1} \\ &\underbrace{+ \tilde{\eta}^2\mathbb{E}_{r-1}[\|\frac{1}{KS}\sum_{k,i\in\mathcal{S}}\nabla f_i(\boldsymbol{y}_{i,k-1})\|^2]}_{\mathcal{A}_2} + \frac{\tilde{\eta}^2\sigma^2}{KS}.\end{aligned}$$

We can directly apply Lemma 5 with $h = f_i, \boldsymbol{x} = \boldsymbol{y}_{i,k-1}, \boldsymbol{y} = \boldsymbol{x}^\star$ and $\boldsymbol{z} = \boldsymbol{x}$ to the first term $\mathcal{A}_1$

$$\begin{aligned}\mathcal{A}_1 &= \frac{2\tilde{\eta}}{KN}\sum_{k,i}\langle\nabla f_i(\boldsymbol{y}_{i,k-1}), \boldsymbol{x}^\star - \boldsymbol{x}\rangle \\ &\leq \frac{2\tilde{\eta}}{KN}\sum_{k,i}(f_i(\boldsymbol{x}^\star) - f_i(\boldsymbol{x}) + \beta\|\boldsymbol{y}_{i,k-1} - \boldsymbol{x}\|^2 - \frac{\mu}{4}\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2) \\ &= -2\tilde{\eta}(f(\boldsymbol{x}) - f(\boldsymbol{x}^{star}) + \frac{\mu}{4}\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2) + 2\beta\tilde{\eta}\mathcal{E}.\end{aligned}$$

[4] We use the notation $\mathbb{E}_{r-1}[.]$ to mean conditioned on filtration r i.e. on all the randomness generated prior to round r.

For the second term $\mathcal{A}_2$, we repeatedly apply the relaxed triangle inequality (Lemma 4)

$$\begin{aligned}
\mathcal{A}_2 &= \tilde{\eta}^2\mathbb{E}_{r-1}[\|\frac{1}{KS}\sum_{k,i\in\mathcal{S}}\nabla f_i(\boldsymbol{y}_{i,k-1}) - \nabla f_i(\boldsymbol{x}) + \nabla f_i(\boldsymbol{x})\|^2] \\
&\leq 2\tilde{\eta}^2\mathbb{E}_{r-1}[\frac{1}{KS}\sum_{k,i\in\mathcal{S}}\nabla f_i(\boldsymbol{y}_{i,k-1} - \nabla f_i(\boldsymbol{x}))\|^2] + 2\tilde{\eta}^2\mathbb{E}_{r-1}[\|\frac{1}{S}\sum_{i\in\mathcal{S}}\nabla f_i(x)\|^2] \\
&\leq \frac{2\tilde{\eta}^2}{KN}\sum_{i,k}\mathbb{E}_{r-1}[\|\nabla f_i(\boldsymbol{y}_{i,k-1}) - \nabla f_i(\boldsymbol{x})\|^2] + 2\tilde{\eta}^2\mathbb{E}_{r-1}[\|\frac{1}{S}\sum_{i\in\mathcal{S}}\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2] \\
&\leq \frac{2\tilde{\eta}^2\beta^2}{KN}\sum_{i,k}\mathbb{E}_{r-1}[\|\boldsymbol{y}_{i,k-1} - \boldsymbol{x}\|^2] + 2\tilde{\eta}^2\|\nabla f(\boldsymbol{x})\|^2 + (1-\frac{S}{N})4\tilde{\eta}^2\frac{1}{SN}\sum_i\|\nabla f_i(\boldsymbol{x})\|^2 \\
&\leq 2\tilde{\eta}^2\beta^2\mathcal{E} + 8\tilde{\eta}^2\beta(B^2+1)(f(\boldsymbol{x}) - f(\boldsymbol{x}^\star)) + (1-\frac{S}{N})\frac{4\tilde{\eta}^2}{S}G^2
\end{aligned}$$

The last step used Assumption (G,B)-BGD assumption (14) that $\frac{1}{N}\sum_{i=1}^N\|\nabla f_i(\boldsymbol{x})\|^2 \leq G^2 + 2\beta B^2(f(\boldsymbol{x}) - f^\star)$. The extra $(1-\frac{S}{N})$ improvement we get is due to sampling the functions $\{f_i\}$ without replacement. Plugging back the bounds on $\mathcal{A}_1$ and $\mathcal{A}_2$,

$$\begin{aligned}
\mathbb{E}_{r-1}[\|\boldsymbol{x} + \Delta\boldsymbol{x} - \boldsymbol{x}^\star\|^2] \leq &(1-\frac{\mu\tilde{\eta}}{2})\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2 - (2\tilde{\eta} - 8\beta\tilde{\eta}^2(B^2+1))(f(\boldsymbol{x}) - f(\boldsymbol{x}^\star)) \\
&+ (1+\tilde{\eta}\beta)2\beta\tilde{\eta}\mathcal{E} + \frac{1}{KS}\tilde{\eta}^2\sigma^2 + (1-\frac{S}{N})\frac{4\tilde{\eta}^2}{S}G^2.
\end{aligned}$$

The lemma now follows by observing that $8\beta\tilde{\eta}(B^2+1) \leq 1$ and that $B \geq 0$.

Lemma 8 (bounded drift). *Suppose our functions satisfies assumptions (A1) and (A2)–(A4). Then the updates of general aggregation for any step-size satisfying $\eta_l \leq \frac{1}{(1+B^2)8\beta K\phi\eta_g}$ have bounded drift:*

$$3\beta\tilde{\eta}\mathcal{E}_r \leq \frac{2\tilde{\eta}}{3}(\mathbb{E}[f(\boldsymbol{x}^{r-1})]) - f(\boldsymbol{x}^\star) + \frac{\tilde{\eta}^2\sigma^2}{2K\eta_g^2} + 18\beta\tilde{\eta}^3G^2. \tag{19}$$

Proof. If $K = 1$, the lemma trivially holds since $\boldsymbol{y}_{i,0} = \boldsymbol{x}$ for all $i \in [N]$ and $\mathcal{E}_r = 0$. Assume $K \geq 2$ here on. Recall that the local update made on client i is $\boldsymbol{y}_{i,k} = \boldsymbol{y}_{i,k-1} - \eta_l g_i(\boldsymbol{y}_{i,k-1})$. Then,

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{y}_{i,k} - \boldsymbol{x}\|^2] =&\mathbb{E}[\|\boldsymbol{y}_{i,k-1} - \boldsymbol{x} - \eta_l g_i(\boldsymbol{y}_{i,k-1})\|^2] \\
\leq&\mathbb{E}[\|\boldsymbol{y}_{i,k-1} - \boldsymbol{x} - \eta_l\nabla f_i(\boldsymbol{y}_{i,k-1})\|^2] + \eta_l^2\sigma^2 \\
\leq&(1-\frac{1}{K-1})\mathbb{E}[\|\boldsymbol{y}_{i,k-1} - \boldsymbol{x}\|^2] + K\eta_l^2\|\nabla f_i(\boldsymbol{y}_{i,k-1})\|^2 + \eta_l^2\sigma^2 \\
=&(1-\frac{1}{K-1})\mathbb{E}[\|\boldsymbol{y}_{i,k-1} - \boldsymbol{x}\|^2] + \frac{\tilde{\eta}^2}{\phi\eta_g K}\|\nabla f_i(\boldsymbol{y}_{i,k-1})\|^2 + \frac{\tilde{\eta}^2\sigma^2}{K^2\phi^2\eta_g^2} \\
\leq&(1-\frac{1}{K-1})\mathbb{E}[\|\boldsymbol{y}_{i,k-1} - \boldsymbol{x}\|^2] + \frac{2\tilde{\eta}^2}{\phi\eta_g K}\|\nabla f_i(\boldsymbol{y}_{i,k-1}) - \nabla f_i(\boldsymbol{x})\|^2 \\
&+ \frac{2\tilde{\eta}^2}{\phi\eta_g K}\|\nabla f_i(\boldsymbol{x})\|^2 + \frac{\tilde{\eta}^2\sigma^2}{K^2\phi^2\eta_g^2} \\
\leq&(1-\frac{1}{K-1} + \frac{2\tilde{\eta}^2\beta^2}{\phi\eta_g K})\mathbb{E}[\|\boldsymbol{y}_{i,k-1} - \boldsymbol{x}\|^2] + \frac{2\tilde{\eta}^2}{\phi\eta_g K}\|\nabla f_i(\boldsymbol{x})\|^2 + \frac{\tilde{\eta}^2\sigma^2}{K^2\phi^2\eta_g^2} \\
\leq&(1-\frac{1}{2(K-1)})\mathbb{E}[\|\boldsymbol{y}_{i,k-1} - \boldsymbol{x}\|^2] + \frac{2\tilde{\eta}^2}{\phi\eta_g K}\|\nabla f_i(\boldsymbol{x})\|^2 + \frac{\tilde{\eta}^2\sigma^2}{K^2\phi^2\eta_g^2}.
\end{aligned}$$

In the above proof we separated the mean and the variance in the first inequality, then used the relaxed triangle inequality with $a = \frac{1}{K-1}$ in the next inequality. Next equality uses the definition of $\tilde{\eta}$, and the rest follow from the Lipschitzness of the gradient. Unrolling the recursion above,

$$\mathbb{E}[\|\boldsymbol{y}_{i,k} - \boldsymbol{x}\|^2] \leq \sum_{\tau=1}^{k-1} (\frac{2\tilde{\eta}^2}{\phi\eta_g K}\|\nabla f_i(\boldsymbol{x})\|^2 + \frac{\tilde{\eta}^2\sigma^2}{K^2\phi^2\eta_g^2})(1 - \frac{1}{2(K-1)})^\tau \leq (\frac{2\tilde{\eta}^2}{\phi\eta_g K}\|\nabla f_i(\boldsymbol{x})\|^2 + \frac{\tilde{\eta}^2\sigma^2}{K^2\phi^2\eta_g^2}3K.$$

Averaging over i and k, multiplying by $3\beta\tilde{\eta}$ and then using Assumption A1,

$$3\beta\tilde{\eta}\mathcal{E}_r \leq \frac{1}{N}\sum_i 18\beta\tilde{\eta}^3\|\nabla f_i(\boldsymbol{x})\|^2 + \frac{3\beta\tilde{\eta}^3\sigma^2}{K\phi^2\eta_g^2} \leq 18\beta\tilde{\eta}^3 G^2 + \frac{3\beta\tilde{\eta}^3\sigma^2}{K\phi^2\eta_g^2} + 36\beta^2\tilde{\eta}^3 B^2(f(\boldsymbol{x}) - f(\boldsymbol{x}^\star))$$

The lemma now follows from our assumption that $8(B^2+1)\beta\tilde{\eta} \leq 1$.

Proof of Theorem I Adding the statements of Lemmas 7 and Lemmas 8, we get

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{x} + \Delta\boldsymbol{x}^\star\|^2] \leq &(1 - \frac{\mu\tilde{\eta}}{2})\mathbb{E}[\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2] + \frac{1}{KS}\tilde{\eta}^2\sigma^2 + (1 - \frac{S}{N})\frac{4\tilde{\eta}^2}{S}G^2 - \tilde{\eta}(\mathbb{E}[f(\boldsymbol{x})] - f(\boldsymbol{x}^\star)) \\ &+ \frac{2\tilde{\eta}}{3}(\mathbb{E}[f(\boldsymbol{x})] - f(\boldsymbol{x}^\star)) + \tilde{\eta}^2(\frac{\sigma^2}{KS}(1 + \frac{S}{\phi^2\eta_g^2}) + \frac{4G^2}{S}(1 - \frac{S}{N}) + 18\beta\tilde{\eta}G^2).\end{aligned}$$

Moving the $f(\boldsymbol{x}) - f(\boldsymbol{x}^\star)$ term and dividing throughout by $\frac{\tilde{\eta}}{3}$, we get the following bound for any $\tilde{\eta} \leq \frac{1}{8(1+B^2)\beta}$

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{x}^{r-1})] - f(\boldsymbol{x}^\star) \leq &\frac{3}{\tilde{\eta}}(1 - \frac{\mu\tilde{\eta}}{2})\|\boldsymbol{x}^{r-1} - \boldsymbol{x}^\star\|^2 - \frac{3}{\tilde{\eta}}\|\boldsymbol{x}^r - \boldsymbol{x}^\star\|^2 \\ &+ 3\tilde{\eta}(\frac{\sigma^2}{KS}(1 + \frac{S}{\phi^2\eta_g^2}) + \frac{4G^2}{S}(1 - \frac{S}{N}) + 18\beta\tilde{\eta}G^2).\end{aligned}$$

If $\mu = 0$ (general convex), we can directly apply Lemma 2. Otherwise, by averaging using weights $w_r = (1 - \frac{\mu\tilde{\eta}}{2})^{1-r}$. and using the same weights to pick output $\bar{\boldsymbol{x}}^R$, we can simplify the above recursive bound to prove that for any $\tilde{\eta}$ satisfying $\frac{1}{\mu R} \leq \tilde{\eta} \leq \frac{1}{8(1+B^2)\beta}$

$$\begin{aligned}\mathbb{E}[f\bar{\boldsymbol{x}}^R] - f(\boldsymbol{x}^\star) \leq &\underbrace{3\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2}_{=:d}\mu\exp(-\frac{\tilde{\eta}}{2}\mu R) \\ &+ \tilde{\eta}\underbrace{(\frac{2\sigma^2}{KS}(1 + \frac{S}{\phi^2\eta_g^2}) + \frac{8G^2}{S}(1 - \frac{S}{N}))}_{=:c_1} \\ &+ \tilde{\eta}^2\underbrace{(36\beta G^2)}_{=:c_2}\end{aligned}$$

Now, the choice of $\tilde{\eta} = \min\{\frac{\log(\max(1,\mu^2 Rd/c_1))}{\mu R}, \frac{1}{(1+B^2)8\beta}\}$ yields the desired rate. The proof of the non-convex case is very similar and also relies on Lemma 2.

9.5. Lower bound for general aggregation

We first formalize the class of algorithms we look at before proving out lower bound.

A6 We assume that general aggregation is run with $\eta_g = 1, K > 1$, and arbitrary possibly adaptive step-sizes $\{\eta_1, \cdots \eta_R\}$ are used with $\eta_r \leq \frac{1}{\mu}$ and fixed within a round for all clients. Further, the server update is a convex combination of the client updates with non-adaptive weights.

Note that we only prove the lower bound here for $\eta_g = 1$. In fact, bu taking η_g infinitely large and scaling $\eta_l \propto \frac{1}{K\eta_g}$ such that the effective step size $\tilde{\eta} = \eta_l \eta_g K$ remains constant, general aggregation reduces to the simple large batch SGD method. Hence, proving a lower bound for arbitrary η_g is not possible, but also is of questionable relevance. Further, note that when $\sigma^2 = 0$, the upper bound in Theorem V uses $\eta_g = 1$ and hence the lower bound serves to show that our analysis is tight.

Below we state a more formal version of Theorem II.

Theorem II. *For any positive constants G, μ, there exists μ -strongly convex functions satisfying A1 for which that the output of general aggregation satisfying A6 has the error for any $r \geq 1$:*

$$f(\boldsymbol{x}^r) - f(\boldsymbol{x}^\star) \geq \Omega(\min(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^\star), \frac{G^2}{\mu R^2})). \tag{20}$$

Proof. Consider the following simple one-dimensional functions for any given μ and G:

$$\begin{cases} f_1(x) & := \mu x^2 + Gx, \\ f_2(x) & := -Gx, \end{cases}$$

with $f(x) = \frac{1}{2}(f_1(x) + f_2(x)) = \frac{\mu}{2}x^2$ and optimum at $x = 0$. Clearly f is μ-strongly convex and further f_1 and f_2 satisfy A1 with $B = 3$. Note that we chose f_2 to be a linear function (not strongly convex) to simplify computations. The calculations make here can be extended with slightly more work for ($\tilde{f}_2 = \frac{\mu}{2}x^2 - Gx$).

Let us start general aggregation from $x^0 > 0$. A single local update for f_1 and f_2 in round $r \geq 1$ is respectively

$$\begin{cases} y_1 & = y_1 - \eta_r(2\mu x + G) \\ y_2 & = y_2 + \eta_r G \end{cases}$$

Then, straightforward computations show that the update at the end of round r is of the following form for some averaging weight $\alpha \in [0, 1]$

$$x^r = x^{r-1}((1-\alpha)(1-2\mu\eta_r)^K + \alpha) + \eta_r G \sum_{\tau=0}^{K-1}(\alpha - (1-\alpha)(1-2\mu\eta_r)^\tau).$$

Since α was picked obliviously, we can assume that $\alpha \leq 0.5$. If indeed $\alpha > 0.5$, we can swap the definitions of f_1 and f_2 and the sign of x^0. With this, we can simplify as

$$\begin{aligned} x^r &\geq x^{r-1}\frac{(1-2\mu\eta_r)^K + 1}{2} + \frac{\eta_r G}{2}\sum_{\tau=0}^{K-1}(1 - (1-2\mu\eta_r)^\tau) \\ &\geq x^{r-1}(1-2\mu\eta_r)^K + \frac{\eta_r G}{2}\sum_{\tau=0}^{K-1}(1 - (1-2\mu\eta_r)^\tau). \end{aligned}$$

Observe that in the above expression, the right hand side is increasing with η_r – this represents the effect of the client drift and increases the error as the step-size increases. The left hand side decreases with η_r – this is the usual convergence observed due to taking gradient steps. The rest of the proof is to show that even with a careful balancing of the two terms, the effect of G cannot be removed. Lemma 9 performs exactly such a computation to prove that for any $r \geq 1$,

$$x^r \geq c\min(x_0 . \frac{G}{\mu R}).$$

We finish the proof by noting that $f(x^r) = \frac{\mu}{2}(x^r)^2$.

Lemma 9. *Suppose that for all $r \geq 1, \eta_r \leq \frac{1}{\mu}$ and the following is true:*

$$x^r \geq x^{r-1}(1-2\mu\eta_r)^K + \frac{\eta_r G}{2}\sum_{\tau=0}^{K-1}(1 - (1-2\mu\eta_r)^\tau). \tag{21}$$

Then, there exists a constants $c > 0$ such that for any sequence of step-size $\{\eta^\tau\}$:

$$x^r \geq c \min(x_0, \frac{G}{\mu R})$$

Proof. Define $\gamma_r = \mu \eta_r R(K-1)$. Such a γ_r exists and is positive since $K \geq 2$. Then, γ_r satisfies

$$(1 - 2\mu\eta_r)^{\frac{K-1}{2}} = (1 - \frac{2\gamma_r}{R(K-1)})^{\frac{K-1}{2}} \leq \exp(-\frac{\gamma_r}{R}).$$

we then have

$$\begin{aligned} x^r &\geq x^{r-1}(1-2\mu\eta_r)^K + \frac{\eta_r G}{2} \sum_{\tau=0}^{K-1} (1 - (1-2\mu\eta_r)^\tau) \\ &\geq x^{r-1}(1-2\mu\eta_r)^K + \frac{\eta_r G}{2} \sum_{\tau=(K-1)/2}^{K-1} (1 - (1-2\mu\eta_r)^\tau) \\ &\geq x^{r-1}(1-2\mu\eta_r)^K + \frac{\gamma_r G}{4\mu}(1 - \exp(-\frac{\gamma_r}{R})). \end{aligned}$$

The second inequality follows because $\eta_r \leq \frac{1}{\mu}$ implies that $(1 - (1-2\mu\eta_r)^\tau)$ is always positive. If $\gamma_r \geq \frac{R}{8}$, then we have a constant $c_1 \in (0, \frac{1}{32})$ which satisfies

$$x^r \geq \frac{c_1 G}{\mu}.$$

On the other hand, if $\gamma_r < \frac{R}{8}$, we have a tighter inequality

$$(1 - 2\mu\eta_r)^{\frac{K-1}{2}} = (1 - \frac{2\gamma_r}{R(K-1)})^{\frac{K-1}{2}} \leq 1 - \frac{\gamma_r}{R},$$

implying that

$$x^r \geq x^{r-1}(1 - \frac{2\gamma_r}{R(K-1)})^K + \frac{\gamma_r^2 G}{4R\mu} \geq x^{r-1}(1 - \frac{4\gamma_r}{R}) + \frac{\gamma_r^2 G}{4\mu R}. \tag{22}$$

The last step used Bernoulli's inequality and the fact that $K - 1 \leq K/2$ for $K \geq 2$. Observe that in the above expression, the right hand side is increasing with γ_r –this represents the effect of the client drift and increases the error as the step-size increases. The left hand side decreases with γ_r –this is the usual convergence observed due to taking gradient steps. The rest of the proof is to show that even with a careful balancing of the two terms, the effect of G cannot be removed.

Suppose that all rounds after $r_0 \geq 0$ have a small step-size i.e. $\gamma_r \leq R/8$ for all $r > r_0$ and hence satisfies (22). Then we will prove via induction that

$$x^r \geq \min(\underbrace{(1 - \frac{1}{2R})^{r-r_0}}_{=:c_r} x^{r_0}, \frac{G}{256\mu R})$$

For $r = r_0$, (9.5) is trivially satisfied. Now for $r > r_0$,

$$x^r \geq x^{r-1}(1 - \frac{4\gamma_r}{R}) + \frac{\gamma_r^2 G}{4\mu R} \geq \min(x^{r-1}(1 - \frac{1}{2R}), \frac{G}{256\mu R}) = \min(c_r x^{r_0}, \frac{G}{256\mu R}).$$

The first step is because of (22) and the last step uses the induction hypothesis. The second step considers two cases for γ_r: either $\gamma_r \leq \frac{1}{8}$ and $(1 - \frac{1}{2R}) \geq (1 - \frac{1}{2R})$, or $\gamma_r^2 \geq \frac{1}{64}$. Finally note that $c^r \geq \frac{1}{2}$ using Bernoulli's inequality. We have hence proved

$$x^R \geq \min(\frac{1}{2} x^{r_0}, \frac{G}{256\mu R})$$

Now suppose $\gamma_{r_0} > \frac{R}{8}$. Then (9.5) implies that $x^R \geq \frac{cG}{\mu R}$ for some constant $c > 0$. If instead no such $r_0 \geq 1$ exists, μR then we can set $r_0 = 0$. Now finally observe that the previous proof did not make any assumption on R, and in fact the inequality stated above holds for all $r \geq 1$.

References

[1] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

[2] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-ojasiewicz condition. *arXiv e-prints*, 2016.

[3] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

[4] A. Kulunchakov and J. Mairal. Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. 2019.

[5] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

[6] Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584, 2006.

[7] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[8] Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019.

[9] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

[10] S. U. Stich and S. P. Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. 2019.

[11] S. Vaswani. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. 2018.

[12] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pages 7252–7261. PMLR, 2019.