A. Proof of Lower Bounds

In this section, we provide further details on the proofs of Theorems 1 and 2.

Theorem 1. *Consider any (possibly randomized) optimization method of the form described in the previous paragraph, i.e. where access to the objective is by evaluating $f(w_t, z_t)$ and $\nabla f(w_t, z_t)$ on semi-cyclic samples (4) and where w_t is chosen based on $\{(f(w_s, z_s), \nabla f(w_s, z_s)), s < t\}$ and the output $\hat{w}$ based on all iterates[4]. For any B, n, K and $m > 1$ there exists a 1-Lipschitz convex problem over high enough dimension such that $\mathbb{E}[F(\hat{w})] \geq F(w_\star) + \Omega(B/K)$, where the expectation is over z_t and any randomization in the method.*

Proof. Let $P(z = 1|i < m/2) = 1$ and $P(z = 2|i \geq m/2) = 1$, with the following functions taken from Woodworth et al. (2018), which in turn is based on the constructions in Arjevani & Shamir (2015); Woodworth & Srebro (2017); Carmon et al. (2017); Woodworth & Srebro (2017):

$$f(w, 1) = \frac{\eta}{8}\Big(-2a\langle v_1, w\rangle + \phi(\langle v_{4K}, w\rangle) + \sum_{k=1}^{2K-1} \phi(\langle v_{2k} - v_{2k+1}, w\rangle)\Big) \tag{23}$$

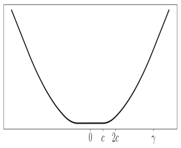
$$f(w, 2) = \frac{\eta}{8}\left(\sum_{k=1}^{2K} \phi(\langle v_{2k-1} - v_{2k}, w\rangle)\right)$$

where v_r are orthogonal vectors, $\eta = 4BK, \gamma = 2B/(\eta\sqrt{K}), a = 1/\sqrt{64K^3}$, and for now consider $\phi(x) = 2\gamma\,|x|$. The main observation is that each vector v_{2k+i} is only revealed after w includes a component in direction v_{2k+i-1} (more formally: it is not revealed if $w \in \operatorname{span}\{v_1, \ldots, v_{2k+i-2}\}$), and only when $f(w, i)$ is queried (Woodworth et al., 2018, Lemma 9). That is, each cycle will reveal at most two vectors, v_{2k+1} for queries on the first half of the blocks, and v_{2k+2} for queries on the second half. After K cycles, the method would only encounter vectors in the span of the first $2K$ vectors $v_1, \ldots, v_{2K}$. But for $\hat{w} \in \operatorname{span}\{v_1, \ldots, v_{2K}\}$, we have $F(\hat{w}) \geq F(w_\star) + \frac{B}{96K}$ (Woodworth et al., 2018, Lemma 8). These arguments apply if the method does not leave the span of gradients returned so far. Intuitively, in high enough dimensions, it is futile to investigate new directions aimlessly. More formally, to ensure that trying out new directions over $T = Kmn$ queries wouldn't help, following appendix C of (Woodworth et al., 2018), we can choose v_r randomly in $R^{\tilde{O}(K^5 n^2 m^2)}$ and use a piecewise quadratic $\phi(x)$ that is 0 for $|z| \leq a/2$ and is equal to $\phi(x) = 2\gamma\,|x| - \gamma^2 - a^2/2$ for $|x| \geq \gamma$. □

Theorem 2. *Under the same setup as in Theorem 1, for any B, n, K and $m > 1$ there exists a 1-Lipschitz convex problem where the gradient $\nabla_w f(w, z)$ is also 1-Lipschitz, such that $\mathbb{E}[F(\hat{w})] \geq F(w_\star) + \Omega(B^2/K^2)$.*

Proof. Use the same construction as in Theorem 1, but with $\eta = B^2$ and

$$\phi(z) = \begin{cases} 0 & |z| \leq a/2 \\ 2(|z| - a/2)^2 & a/2 < |z| \leq a \\ z^2 - a^2/2 & a < |z| \leq \gamma \\ 2\gamma|z| - \gamma^2 - a^2/2 & |z| > \gamma \end{cases}$$



The objective is smooth (Woodworth et al., 2018, Lemma 7), and the same arguments as in the proof of Theorem 1 hold except that now $\hat{w}$ spanned by $v_1, \ldots, v_{2K}$ has $F(\hat{w}) \geq F(w_\star) + \frac{B^2}{256K}$ (Woodworth et al., 2018, Lemma 8). □

B. Deferred Proofs from Section 6

In this section, we give a proof of Lemma 1. Such a result has been previously shown in Even-Dar et al. (2008); Sani et al. (2014), building on a lemma of Cesa-Bianchi et al. (2007). We give a full proof for completeness. We start with a simple Lemma.

Lemma 2. *For any $z > -\frac{1}{2}$,*

$$z - z^2 \leq \ln(1 + z) \leq z.$$

Proof. The upper bound on $\ln(1+z)$ is standard, and follows e.g. by the concavity of the log function. For the lower bound, write

$$f(z) = \ln(1 + z) - (z - z^2).$$

Then $f'(z) = \frac{1}{1+z} - 1 + 2z = \frac{z(1+2z)}{(1+z)}$. Thus f is decreasing in $(-\frac{1}{2}, 0)$ and increasing in $(0, \infty)$. Thus in the range $(-\frac{1}{2}, \infty)$, $f(z) \geq f(0) = 0$. □

Proof. Let $P(z = 1|i < m/2) = 1$ and $P(z = 2|i \geq m/2) = 1$, with the following functions taken from Woodworth et al. (2018), which in turn is based on the constructions in Arjevani & Shamir (2015); Woodworth & Srebro (2017); Carmon et al. (2017); Woodworth & Srebro (2017):

$$f(w, 1) = \frac{\eta}{8}\Big(-2a\langle v_1, w\rangle + \phi(\langle v_{4K}, w\rangle) + \sum_{k=1}^{2K-1} \phi(\langle v_{2k} - v_{2k+1}, w\rangle)\Big) \tag{24}$$

$$f(w, 2) = \frac{\eta}{8}\left(\sum_{k=1}^{2K} \phi(\langle v_{2k-1} - v_{2k}, w\rangle)\right)$$

where v_r are orthogonal vectors, $\eta = 4BK, \gamma = 2B/(\eta\sqrt{K}), a = 1/\sqrt{64K^3}$, and for now consider $\phi(x) = 2\gamma\,|x|$. The main observation is that each vector v_{2k+i}

[4] This theorem, as well as Theorem 2, holds even if the method is allowed "prox queries" of the form $\arg\min_w f(w, z_t) + \lambda_t \|w - w_t\|^2$.

is only revealed after w includes a component in direction v_{2k+i-1} (more formally: it is not revealed if $w \in \operatorname{span}\{v_1, \dots, v_{2k+i-2}\}$), and only when $f(w, i)$ is queried (Woodworth et al., 2018, Lemma 9). That is, each cycle will reveal at most two vectors, v_{2k+1} for queries on the first half of the blocks, and v_{2k+2} for queries on the second half. After K cycles, the method would only encounter vectors in the span of the first $2K$ vectors $v_1, \dots, v_{2K}$. But for $\hat{w} \in \operatorname{span}\{v_1, \dots, v_{2K}\}$, we have $F(\hat{w}) \geq F(w_\star) + \frac{B}{96K}$ (Woodworth et al., 2018, Lemma 8). These arguments apply if the method does not leave the span of gradients returned so far. Intuitively, in high enough dimensions, it is futile to investigate new directions aimlessly. More formally, to ensure that trying out new directions over $T = Kmn$ queries wouldn't help, following appendix C of (Woodworth et al., 2018), we can choose v_r randomly in $R^{\tilde{O}(K^5 n^2 m^2)}$ and use a piecewise quadratic $\phi(x)$ that is 0 for $|z| \leq a/2$ and is equal to $\phi(x) = 2\gamma\,|x| - \gamma^2 - a^2/2$ for $|x| \geq \gamma$. □

We consider the more general case of $K+1$ experts with losses in $[-M, M]$, and a chosen expert 0, with respect to which we want constant regret. We consider the PROD Algorithm that starts out with initial weights:

$$q_1^0 = 1 - \eta; \qquad q_1^i = \eta/K \quad \forall i = 1..K.$$

At time step t, it picks an expert j_t with probability proportional to q_t^j:

$$p_t^i = q_t^i \Big/ \sum_{j=0}^{K} q_t^j.$$

Finally, on receiving the loss function ℓ_t, it updates the weights according to the multiplicative update

$$q_{t+1}^i = q_t^i \cdot \big(1 + \eta\big(\ell_t(0) - \ell_t(i)\big)\big) \qquad \forall i = 0..K$$

Lemma 3. *Assume that* $0 < \eta \leq 1/(4M)$. *Then this* PROD *algorithm achieves*

$$\sum_{t=1}^{T} p_t^{j_t} \ell_t(j_t) - \sum_{t=1}^{T} \ell_t(j) \leq 4\eta M^2 T + \frac{1}{\eta} \ln \frac{K}{\eta}$$

for all $j = 1, \dots, K$, *and*

$$\sum_{t=1}^{T} p_t^{j_t} \ell_t(j_t) - \sum_{t=1}^{T} \ell_t(0) \leq 1 + \eta.$$

Proof. Let $Q_t = \sum_{j=0}^{K} q_t^j$ and let $\Delta_t^j = \ell_t(0) - \ell_t(j)$ denote the gap between the chosen expert and expert j at step t. Note that $|\Delta_t^j| \leq 2M$.

On the one hand,

$$\begin{aligned}
\ln \frac{Q_{T+1}}{Q_1} &= \sum_{t=1}^{T} \ln \frac{Q_{t+1}}{Q_t} \\
&= \sum_{t=1}^{T} \ln \left(\frac{1}{Q_t} \sum_{j=0}^{K} q_t^j (1 + \eta \Delta_t^j) \right) \\
&= \sum_{t=1}^{T} \ln \sum_{j=0}^{K} p_t^j (1 + \eta \Delta_t^j) \\
&= \sum_{t=1}^{T} \ln \left(1 + \eta \sum_{j=0}^{K} p_t^j \Delta_t^j \right) \\
&\leq \sum_{t=1}^{T} \eta \sum_{j=0}^{K} p_t^j \Delta_t^j \\
&= \eta \sum_{t=1}^{T} \ell_t(0) - \eta \sum_{t=1}^{T} p_t^{j_t} \ell_t(j_t).
\end{aligned}$$

On the other hand, for any j,

$$\begin{aligned}
\ln \frac{Q_{T+1}}{Q_1} &\geq \ln \frac{q_{T+1}^j}{q_1^j} + \ln \frac{q_1^j}{Q_1} \\
&= \sum_{t=1}^{T} \ln \frac{q_{t+1}^j}{q_t^j} + \ln \frac{q_1^j}{Q_1} \\
&= \sum_{t=1}^{T} \ln(1 + \eta \Delta_t^j) + \ln \frac{q_1^j}{Q_1} \\
&\geq \sum_{t=1}^{T} (\eta \Delta_t^j - (\eta \Delta_t^j)^2) + \ln \frac{q_1^j}{Q_1} \\
&\geq \eta \sum_{t=1}^{T} (\ell_t(0) - \ell_t(j)) - 4\eta^2 M^2 T + \ln \frac{q_1^j}{Q_1},
\end{aligned}$$

where the middle inequality holds since $|\eta \Delta_t^j| \leq 1/2$. It follows that for $j \neq 0$,

$$\sum_{t=1}^{T} p_t^{j_t} \ell_t(j_t) - \sum_{t=1}^{T} \ell_t(j) \leq 4\eta M^2 T + \frac{1}{\eta} \ln \frac{K}{\eta}.$$

Moreover, since q_t^0 does not change during the algorithm, we also have, using the lower bound in Lemma 2, that

$$\ln \frac{Q_{T+1}}{Q_1} \geq \ln \frac{q_1^0}{Q_1} = \ln(1 - \eta) \geq -\eta - \eta^2.$$

This implies that

$$\sum_{t=1}^{T} p_t^{j_t} \ell_t(j_t) - \sum_{t=1}^{T} \ell_t(0) \leq 1 + \eta. \qquad \square$$

Optimizing parameters, we get the corollary:

Corollary 1. *Set $\eta = \frac{1}{2M}\sqrt{\ln(KMT)/T}$ and assume that that $M \geq 1$ and that T is large enough so that $\eta \leq 1/(4M)$. Then the algorithm achieves the following regret bounds:*

$$\sum_{t=1}^{T} p_t^{j_t} \ell_t(j_t) \leq \sum_{t=1}^{T} \ell_t(j) + 4M\sqrt{T\ln(KMT)}$$

for all $j = 1, \ldots, K$, and

$$\sum_{t=1}^{T} p_t^{j_t} \ell_t(j_t) \leq \sum_{t=1}^{T} \ell_t(0) + 2.$$

Lemma 1 follows from the $K = 1$ version of this corollary, where the two experts are the algorithms w_t^j and w_t.

C. Experimental Details

The source code used for data preprocessing, training, evaluation, and plotting results will be made available at `https://github.com/tensorflow/federated/tree/master/tensorflow_federated/python/research/semi_cyclic_sgd`

C.1. Dataset

The sentiment140 dataset set (Go et al., 2009) was collected by querying Twitter (a popular social network) for posts (a.k.a. Tweets) containing positive and negative emoticons, and labeling the retrieved posts (with emoticons removed) as positive and negative sentiment, respectively.

The data sets used for the above scenarios are created by first shuffling the data randomly and splitting it into a training (90%, or $1,440,000$ examples) and test set (10%, or $160,000$ examples). This data set is used as-is for training and evaluating the *idealized i.i.d.* model. For the other scenarios trained on block-cyclic data, we group the shuffled training set and test set into $m = 6$ blocks each by the time of day of the post (e.g. midnight block: posts from 12am - 4am; noon block: posts from 12pm - 4pm). This results in blocks of varying sizes, on average $1,440,000/6 = 240,000$ (training) and $160,000/6 = 24,000$ (testing) examples, respectively.

We simulate $K = 10$ cycles (days). Observing that one pass (epoch) over the entire i.i.d. data set was sufficient for convergence of our relatively small model, this results in $mn = 1,440,000/10$ training examples per day, or $n = 1,440,000/10/6 = 24,000$ training examples per day per block.

C.2. Artificially balanced labels

The raw data grouped by time of day exhibits some block-cyclic characteristics; for instance, positive tweets are slightly more likely at night time hours than day time hours (see Figure 3). However, we believe this dataset has an artificially balanced label distribution, which is not ideal to illustrate semi-cyclic behavior (Go et al., 2009). In particular, the data collection process separately queried the Twitter API every 2 minutes for positive tweets (defined to be those containing the **:)** emoticon), and simultaneously for negative sentiment via **:(**. Since only up to 100 results are returned via each API query, this will generally produce an (artificially) balanced label distribution, as in Fig. 3. Due to this fact, because large diurnal variations are likely in practice in Federated Learning (e.g., differences in the use of English language between the US and India), and because it better illustrates our theoretical results, we adjust the positive-sentiment rate as a function of time as described in section 8.

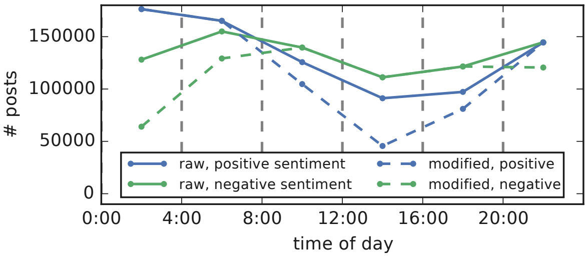


Figure 3. Sentiment bias as a function of day time in the Sentiment140 dataset. For experiments in this paper, we introduced additional time-of-day dependent label skew to allow for a clearer illustration of how pluralistic approaches differ.

C.3. Details of evaluation methodology

For the *block-cyclic consensus* model, picking a random iteration of the form $t(k, i, n)$ ensures we evaluate a set of models that have the same expected number of iterations as for the single-chain pluralistic approach, without using block-specific models. In the implementation, we compute the expectation of this quantity by evaluating all m iterates against all m $\hat{F}_i$s, and averaging these m^2 values.

This same $m \times m$ set of evaluation results is used to evaluate the pluralistic single SGD chain approach, but instead of averaging all m^2 accuracies, we only consider the diagonal, where the model most recently trained on data from component i is evaluated (only) on $\hat{F}_i$.