A PROOFS

A.1 Assumptions

Before our analysis, we make the following commonly used assumptions:

Assumption 1. *We claim the following standard properties:*

- ***Lower Bound****:* $\forall \mathbf{x} \in \mathbb{R}^d, f(\mathbf{x}) \geq f_{inf}$.
- ***Lipschitz continuity****:* $\forall$ $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,
$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|.$$

Assumption 2. *For the the j-th worker, $\forall \mathbf{x}^j \in \mathbb{R}^d$, the gradient of i-th model component is $G_i^j(\mathbf{x}^j)$, which satisfies:*

- ***Unbiased estimation****:* $\mathbb{E}_\xi[G_i^j(\mathbf{x}^j)] = \nabla_i f(\mathbf{x}^j)$,
- ***Bounded variance****:* $\mathbb{E}_\xi[\|G_i^j(\mathbf{x}^j)\|^2] \leq B^2$.

We remark that both the above assumptions have been used in many related studies [31, 32, 48] as well. Mathematically, the convergence property is formulated with Theorem 1.

A.2 Explanations on Per-embedding Clock Bounded Consistency

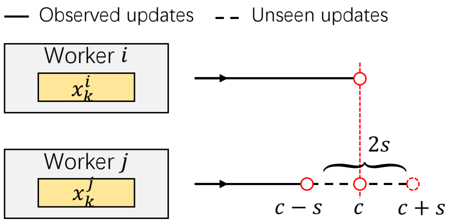


Figure 10: Illustration of per-embedding clock bounded consistency. Red circles represent cache synchronization.

Figure 10 illustrates a running example to explain the per-embedding clock bounded consistency guarantees. Considering for an embedding $\mathbf{x}_k$ at a certain iteration, $\mathbf{x}_k^i$ just finished the cache synchronization with server. We suppose the worst case that there exists $\mathbf{x}_k^j$ on worker j at clock c, which also synchronizes its updates on the embedding immediately after i's synchronization. Since the stale-write mechanism, worker i will not see the accumulated updates from j at clock c. But the updates before $c - s$ could be observed because they have already been written to the global embedding table and could be fetched by worker i during the synchronization. After clock c, due to the cache valid conditions, worker j could continue go forward at most s clocks (i.e., s updates) before i's next synchronization. Therefore, worker i could observe all updates from worker j older than $2s$ clocks, i.e., those in the range of $[0, \mathbf{x}_k^j.c_c - 2s]$.

Then we revisit the relationship between the embedding clocks and the iteration number t. Considering the case of a single worker, in our design, the embedding clock $\mathbf{x}_i.c_c$ represents the number of updates on $\mathbf{x}_i$. For dense models, it equals to the iteration number t. However, different embeddings in a model could have different popularity. Therefore, to bridge the gap between the clocks and t, we need to formally characterize the embedding access pattern. Suppose the input data are distributed to each worker in a unbiased manner, we define a *per-embedding existing probability* p_i to represent the probability of embedding x_i existing in each mini-batch of input data. Therefore, we can infer the following lemma to motivate our following proofs:

Lemma 2. *Given embedding per-embedding existing probability p_i for x_i, it needs around s/p_i iterations to perform s updates.*

Note that, our following results could also be proved similarly even on biased data distribution based methods by characterize the embedding existing probabilities on different workers.

A.3 Proof of Theorem 1

Our key tool for the convergence analysis is to define a reference model sequence $\mathbf{x}(t)$ without staleness, for each embedding $\mathbf{x}_i(t)$ in $\mathbf{x}(t)$, suppose the existing probability of $\mathbf{x}_i$ is p_i, we have:

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) - \frac{\eta}{N}\sum_{n=0}^{N-1} G_i^n(\mathbf{x}^n(t))p_i. \tag{4}$$

Intuitively, $\mathbf{x}(t)$ would like to follow the "clean" SGD iteration, by including all the gradients generated by the end of round t. Without loss of generality, $\mathbf{x}(0) = \mathbf{x}^j(0)$ for any worker $0 \leq j \leq N$. Then we will prove that the "true" sequence is not too far from the models on the workers:

Lemma 3. *For any $t \geq 0$ and worker j, we have:*
$$\mathbb{E}[\|\mathbf{x}(t) - \mathbf{x}^j(t)\|^2] \leq 4s^2\eta^2M^2B^2.$$

Proof. Based on the Cauchy-Schwarz inequality, we have:

$$\|\mathbf{x}(t) - \mathbf{x}^j(t)\|^2 \leq M\sum_{i=1}^{M}\|\mathbf{x}_i(t) - \mathbf{x}_i^j(t)\|^2. \tag{5}$$

Then for each embedding $\mathbf{x}_i$, due to Lemma 2 we have:

$$\begin{aligned}
\|\mathbf{x}_i(t) - \mathbf{x}_i^j(t)\|^2 &\leq \|\sum_{k=t-\frac{2s}{p_i}}^{t} \eta\sum_{n=1}^{N} G_i^n(\mathbf{x}^n(k))p_i/N\|^2 &(6)\\
&\leq (\eta p_i/N)^2\|\sum_{k=t-\frac{2s}{p_i}}^{t}\sum_{n=1}^{N} G_i^n(\mathbf{x}^n(k))\|^2 &(7)\\
&\leq (\eta p_i/N)^2\frac{2s}{p_i}N\sum_{k=t-\frac{2s}{p_i}}^{t}\sum_{n=1}^{N}\|G_i^n(\mathbf{x}^n(k))\|^2 &(8)\\
&\leq 4s^2\eta^2B^2, &(9)
\end{aligned}$$

where we have used the properties stated in Lemma 1 (in particular the Staleness Bound), and the triangle inequality. Next, we notice that the expected squared norm of each of the missing gradients is bounded by B^2 (by the second moment bound). Combing Eq. (9) with Eq. (5) finally implies the claimed inequality.

□

Proofs on Theorem 1.

Proof. We begin from the definition of $\mathbf{x}(t)$ in Eq.(4). We will first prove the above statement for the iterate $\mathbf{x}(t)$, and then will extend the proof for $\mathbf{x}^j(t)$. For simplicity, let us denote $G(t) = \sum_{n=0}^{N-1} G^n(\mathbf{x}^n(t))$. We can use the Taylor expansion of $f(\mathbf{x}(t+1))$ around $\mathbf{x}(t)$ and the smoothness condition to obtain the following inequality that $f(\mathbf{x}(t+1))$

$$
\begin{aligned}
&\leq f(\mathbf{x}(t)) + (\mathbf{x}(t+1) - \mathbf{x}(t))^T \nabla f(\mathbf{x}(t)) + \frac{L}{2}\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 \\
&= f(\mathbf{x}(t)) - \eta \nabla f(\mathbf{x}(t))^T \nabla f(\mathbf{x}(t)) + \frac{\eta^2 L}{2N^2}\|G(t)\|^2 + \\
&+ \eta(\nabla f(\mathbf{x}(t)) - G(t)/N)^T \nabla f(\mathbf{x}(t)).
\end{aligned}
$$

We can therefore apply the expectation with respect to the random sampling at step t, the second moment bound assumption:

$$
\begin{aligned}
\mathbb{E}\left[f(\mathbf{x}(t+1))|\mathbf{x}(t)\right] \leq & f(\mathbf{x}(t)) - \eta\|\nabla f(\mathbf{x}(t))\|^2 + \frac{\eta^2 L}{2} M^2 B^2 \\
& + \eta(\nabla f(\mathbf{x}(t)) - \nabla f(\mathbf{x}^j(t)))^T \nabla f(\mathbf{x}(t)).
\end{aligned}
$$

To bound the last term, we can now apply the Cauchy-Schwarz inequality and the fact that the gradients are L-Lipschitz:

$$
\begin{aligned}
\mathbb{E}\left[f(\mathbf{x}(t+1))|\mathbf{x}(t)\right] \leq & f(\mathbf{x}(t)) - \eta\|\nabla f(\mathbf{x}(t))\|^2 + \frac{\eta^2 L}{2} M^2 B^2 \\
& + \eta L\|\mathbf{x}(t) - \mathbf{x}^j(t)\|\|\nabla f(\mathbf{x}(t))\|.
\end{aligned}
$$

To further bound the last term, we can apply the classic inequality $a^2 + b^2 \geq 2ab$ together with Lemma 3 to obtain:

$$
\begin{aligned}
\mathbb{E}\left[f(\mathbf{x}(t+1))|\mathbf{x}(t)\right] \leq & \ f(\mathbf{x}(t)) - \eta\|\nabla f(\mathbf{x}(t))\|^2 + \frac{\eta^2 L}{2} M^2 B^2 \\
& + \eta\|\nabla f(\mathbf{x}(t))\|^2/2 + 2s^2\eta^3 L^2 M^2 B^2.
\end{aligned}
$$

Rearranging terms and taking total expectation:

$$
\begin{aligned}
\mathbb{E}\left[\|\nabla f(\mathbf{x}(t))\|^2\right] \leq & \frac{2\mathbb{E}\left[f(\mathbf{x}(t)) - f(\mathbf{x}(t+1))\right]}{\eta} + \eta M^2 B^2 L \\
& + 4s^2\eta^2 L^2 M^2 B^2.
\end{aligned}
$$

Summing across all t and dividing by T, we get:

$$
\begin{aligned}
\min_{1\leq t\leq T} \mathbb{E}\left[\|\nabla f(\mathbf{x}(t))\|^2\right] &\leq \frac{1}{T}\sum_t \mathbb{E}\left[\|\nabla f(\mathbf{x}(t))\|^2\right] \\
&\leq \frac{2\left(f(\mathbf{x}(0)) - f_{\text{inf}}\right)}{\eta T} + \eta M^2 B^2 L + 4s^2\eta^2 L^2 M^2 B^2. \qquad (10)
\end{aligned}
$$

We now study the set of conditions for each of the three RHS terms to be less than $\epsilon/12$. We have that it is sufficient for the following three conditions to hold:

(1) $T \geq \frac{24(f(x(0) - f_{\text{inf}})}{\eta\epsilon}$;

(2) $\eta \leq \frac{\epsilon}{12M^2B^2L}$;

(3) $\eta \leq \frac{\sqrt{\epsilon}}{4\sqrt{3}sLMB}$.

All these conditions hold by assumption from the theorem statement. We have therefore obtained that there exists $t^\star$ such that $\|\nabla f(\mathbf{x}(t^\star))\|^2 \leq \epsilon/4$. However, by smoothness and Lemma 3 we know that

$$\mathbb{E}\|\nabla f(\mathbf{x}(t^\star)) - \nabla f(\mathbf{x}^j(t^\star))\|^2 \leq 4Ls^2\eta^2M^2B^2.$$

To prove our final target, we need to prove

$$\mathbb{E}\|\nabla f(\mathbf{x}(t^\star)) - \nabla f(\mathbf{x}^j(t^\star))\|^2 \leq 4Ls^2\eta^2M^2B^2 \leq \epsilon/4.$$

Based on this inequality, we have the second bound of the learning rate:

$$\eta \leq \frac{\sqrt{\epsilon}}{4\sqrt{L}sMB},$$

which has been clarified as an assumption in the theorem statement on the upper bound on η.

Finally, we can apply the classic inequality $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$ to obtain that

$$\mathbb{E}\|\nabla f(\mathbf{x}^j(t^\star))\|^2 \leq \epsilon.$$

□

Corollary 1. *Under Theorem 1, with a proper learning rate $\eta = \frac{1}{L\sqrt{T}}$, we have:*

$$\frac{1}{T}\sum_t \mathbb{E}\left[\|\nabla f(\mathbf{x}(t))\|^2\right] \leq \frac{2L\left(f(\mathbf{x}(0)) - f_{inf}\right) + M^2B^2}{\sqrt{T}} + \frac{4s^2M^2B^2}{T}. \quad (11)$$

We make the following observations regarding the bound. First, we prove that, for a given sequence of learning rates, the algorithm will converge to a point of negligible gradient. Corollary 1 implies that when the learning rate is $\eta = 1/(L\sqrt{T})$ and T is sufficiently large, we derive that the convergence rate is around $O(1/\sqrt{T})$, which is consistent with that of BSP. Second, we note that first RHS term in Eq. (11) is the standard SGD error and the third term comes from the staleness. This suggests to trade off the additional performance cost of synchronization with the slower convergence due to delayed gradient information in practice.

Table 3: Model Architectures

Model	Layer#	Hidden	Embedding Size	Dense Parameter
WDL	3	256	128	137.98K
DFM	3	256	128	917.77K
DCN	3	256	128	1.01M
GraphSage	2	128	128	2.40k-19.48k

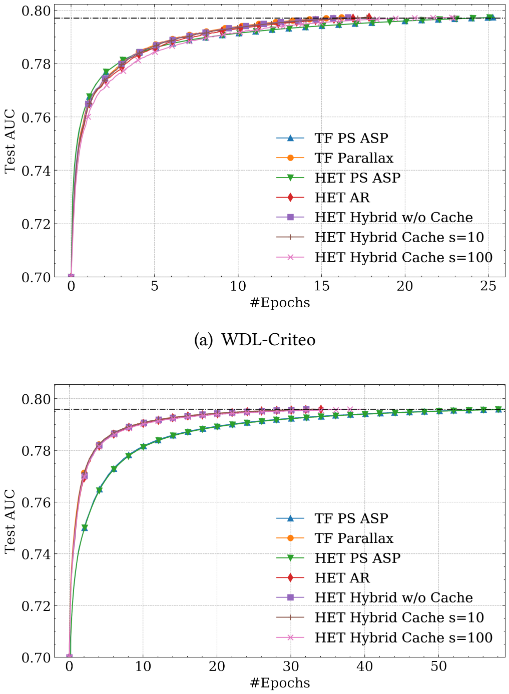
Table 4: Statistics of Datasets

Datasets	Embedding#	Sample#	Feature
Criteo	33.76M	45.84M	39
Amazon	2.79M	2.45M	17
Reddit	233.96K	232.96K	1
ogbn-mag	1.93M	736.39K	21

B MODELS AND DATASETS

We introduce more details about the model configurations in Table 3, and also some statistics of datasets in Table 4. In order to elucidate the effectiveness of our embedding learning system, we replace dense features in Amazon, Reddit and ogbn-mag with a trainable node-id embedding. In addition, we include each product's word of bag feature as sparse input in Amazon. And in ogbn-mag, we include each paper's author and field of interest as additional sparse features. We use the train/eval split provided in the original papers of these datasets. Note that, the embedding size is set to be 128 for almost of our experiments except for the last model scalability in Figure 9(c). The largest model in our experiments contains more than one trillion embedding parameters (i.e., 33.76M×4096).

C EXPERIMENT RESULTS



(a) WDL-Criteo

(b) DFM-Criteo

Figure 11: Convergence performance comparison.

We illustrate the convergence performance on WDL-Criteo and DFM-Criteo with the number of epochs as the x-axis. As shown in Figure 11, TF PS and HET PS have the same statistical efficiency and perform worse than the other methods due to asynchronous updates. But our proposed methods perform very close to the fully synchronous baselines including TF Parallax and HET Hybrid w/o Cache.