A EXPERIMENT DETAILS

A.1 HET-GMP

The hyper-parameters α, β and γ in E.q. (4) are set to be $100N/|\xi|$, $100N/|V|$ and 1e-2, respectively. We set $s = 100$ for HET-GMP in the scalability experiments. We set T to 5 in our Algorithm 2 and show the graph partitioning time costs in Table 3. As we can see, the graph partition is a pre-processing step on CPU and only takes a few minutes, which is negligible compared to the distributed training process.

Table 3: Partition metrics of the three datasets

Dataset	#Samples	#Features	Partition time (s)
Avazu	40,428,967	9,449,445	273
Criteo	45,840,617	33,762,577	453
Company	35,682,429	66,102,027	795

A.2 TensorFlow, Parallax

Different from HET-GMP, TensorFlow and Parallax both contain PS to store and update embedding parameters. In the experiments we use one server in each machine, while other configurations are the same with HET-GMP.

A.3 HugeCTR

The Embedding layer type we used in the HugeCTR experiments is `LocalizedSlotSparseEmbeddingHash`. We chose it because it performs better than `DistributedSlotSparseEmbeddingHash` in our experiments settings. The `LocalizedSlotSparseEmbeddingOneHot` option requires special NVSwitch hardware which we don't have.

B PROOFS

We define the global model $\mathbf{x}$ as the combination of all latest embeddings (i.e., master) for our analysis. We denote x_i and $\nabla_i f(\mathbf{x})$ as the i-th embedding and its gradients on $\nabla f(\mathbf{x})$, respectively. Clearly, $\mathbf{x} = (x_0, x_1, \ldots, x_m)$ and $\nabla f = (\nabla_0 f, \nabla_1 f, \cdots, \nabla_m f)$, where $m = |S|$. For simplicity, we suppose each worker only has one master embedding so that the master of the i-th embedding is at worker i. Note that, it is easily to extend our results to multiple master embeddings by combining them as a model component at worker i and making integral analysis.

Consider a global clock shared by all workers and denote T_i^j the set of active clocks when worker j takes an update on x_i, and $\mathbb{I}_{t\in T_i^j}$ as the indicator function of the event $t \in T_i^j$. Formally, the t-th iteration on worker j can be written as:

$$x_i^j(t+1) = x_i^j(t) - \mathbb{I}_{t\in T_i^j}\eta\nabla_i f(\mathbf{x}^j(t)), \forall i \tag{10}$$

$$\mathbf{x}^j(t) = \left(x_0^j(\tau_0^j(t)),\ \ldots\ ,\ x_m^j(\tau_m^j(t))\right) \tag{11}$$

where $0 \le \tau_j^i(t) \le t$ models the delay of mirrors. When worker j performs the t-th update, it only has access to $x_i^j(\tau_i^j(t))$, a delayed version of embedding x_i on the j-th worker. Here we use $x_i(t)$ to represent the latest x_i among all workers:

$$\mathbf{x}(t) = \left(x_1(t),\ \ldots\ ,\ x_m(t)\right). \tag{12}$$

We make the following standard assumptions for our convergence analysis:

Assumption 1 (Common SGD).

(1) The function f is bounded below.
(2) The function f is differentiable and the gradient ∇f of f is L-Lipschitz continuous:

Based on the intra-and-inter bounded asynchrony provided by the stale `Read` interface in HET-GMP, we make the following bounded delay assumption:

Assumption 2 (Bounded Delay). *The delay and active clocks satisfy:*

(1) $\forall i, \forall j, \forall t,\ 0 \le t - \tau_j^i(t) \le s,\ \tau_i^i(t) \equiv t$;
(2) $\forall i, \forall t, T_i \cap \{t, t+1, \cdots, t+s\} \neq \emptyset$.

Assumption 2.1 guarantees at the t-th iteration, the j-th embedding from the i-th worker is always not too obsolete in two aspects: 1) bounded asynchrony at **intra-embedding**: for the j-th embedding, we guarantee that its replica on any worker i is not obsolete compared to the latest version (bounded by at most s clocks apart); 2) bounded asynchrony at **inter-embedding**: for the i-th worker, we guarantee the any local embedding j is not obsolete compared to the other local embeddings (bounded by at most s clocks apart). The assumption $\tau_i^i(t) \equiv t$ is natural since the i-th worker is maintaining x_i hence would always have the latest copy.

Assumption 2.2 requires each machine to update at least once in every $s+1$ iterations, for otherwise some x_i may not be updated at all. We remark that Assumption 2.2 is very natural and have been widely adopted in previous works [18]. Clearly, when $s = 0$ (i.e., no delay), our algorithm reduces to the fully synchronous parallel training. Inspired by some previous studies [40], we involve the sufficient decrease assumption to control the sequence and sequence of objective value within a certain range.

Assumption 3 (Sufficient decrease). *There exists $\alpha > 0$ such that for all large t,*

$$F(\mathbf{x}(t+1)) \le F(\mathbf{x}(t)) - \alpha\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2. \tag{13}$$

Based on these assumotions, we clarify the following theorem:

Theorem 1. *Let Assumption 1, 2 and 3 hold, and let F satisfy the KŁ property in [11, Lemma 6]. Then, with step size $\eta \in \left(0, \frac{1}{L(1+2\sqrt{ps})}\right)$, every bounded sequence $\{\mathbf{x}(t)\}$ generated by satisfies*

$$\sum_{t=0}^{\infty}\|\mathbf{x}(t+1) - \mathbf{x}(t)\| < \infty, \tag{14}$$

$$\forall i = 1, \ldots, p,\ \sum_{t=0}^{\infty}\|\mathbf{x}^i(t+1) - \mathbf{x}^i(t)\| < \infty. \tag{15}$$

Furthermore, $\{\mathbf{x}(t)\}$ and $\{\mathbf{x}^i(t)\}_{i=1}^p$ converge to the same critical point of F. Recall from [29], we have the following convergence rate:

$$F(\frac{1}{t}\sum_{k=1}^{t}\mathbf{x}(k)) - F_{inf} \le O(\frac{1}{t}). \tag{16}$$

Note that, the notation p in Theorem 1 is a typo in the main body of our paper, which should be the size of embedding table m and we use m instead in the following analysis. Before we prove Theorem 1, we first introduce some necessary lemmas.

B.1 Preliminaries

Theorem 2. *Let Assumption 1 and 2 hold. If the step size* $\eta \in \left(0, \frac{1}{L(1+2\sqrt{ms})}\right)$, *then the sequence generated by HET-GMP is square summable, i.e.*

$$\sum_{t=0}^{\infty} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 < \infty. \tag{17}$$

In particular, $\lim_{t\to\infty} \|\mathbf{x}(t+1) - \mathbf{x}(t)\| = 0$ *and* $\lim_{t\to\infty} \|\mathbf{x}(t) - \mathbf{x}^i(t)\| = 0$.

Proof. We start from providing a technical tool to control the inconsistency between the local models $\mathbf{x}^i(t)$ and the global model $\mathbf{x}(t)$. We define $(t)_+ = \max\{t, 0\}$ to represent the positive part of t. At iteration t, we have:

$$\begin{aligned}
\|\mathbf{x}(t) - \mathbf{x}^i(t)\| &= \sqrt{\sum_{j=1}^{m} \|x_j(t) - x_j(\tau_j^i(t))\|^2} \\
&\le \sqrt{\sum_{j=1}^{m} \left(\sum_{k=\tau_j^i(t)}^{t-1} \|x_j(k+1) - x_j(k)\| \right)^2} \\
&\le \sqrt{\sum_{j=1}^{m} \left(\sum_{k=(t-s)_+}^{t-1} \|x_j(k+1) - x_j(k)\| \right)^2} \\
&= \left\| \left(\sum_{k=(t-s)_+}^{t-1} \|x_j(k+1) - x_j(k)\|, \cdots, \sum_{k=(t-s)_+}^{t-1} \|x_j(k+1) - x_j(k)\| \right) \right\| \\
&= \left\| \sum_{k=(t-s)_+}^{t-1} (\|x_j(k+1) - x_j(k)\|, \cdots, \sum_{k=(t-s)_+}^{t-1} \|x_j(k+1) - x_j(k)\|) \right\| \\
&\le \sum_{k=(t-s)_+}^{t-1} \left\| (\|x_j(k+1) - x_j(k)\|, \cdots, \sum_{k=(t-s)_+}^{t-1} \|x_j(k+1) - x_j(k)\|) \right\| \\
&= \sum_{k=(t-s)_+}^{t-1} \|\mathbf{x}(k+1) - \mathbf{x}(k)\|.
\end{aligned} \tag{18}$$

The above (18) bounds the inconsistency between the global model and the local models. And it will be repeatedly used in the following proofs.

Next we bound the progress of the global model $\mathbf{x}(t)$. Based on the model update rule, we have:

$$\frac{1}{2\eta} \|x_i(t+1) - x_i(t)\|^2 \le -\left\langle \nabla_i f(\mathbf{x}^i(t)), x_i(t+1) - x_i(t) \right\rangle. \tag{19}$$

Adding (20) for all i, we have

$$\frac{1}{2\eta} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 \le -\sum_{i=1}^{m} \left\langle \nabla_i f(\mathbf{x}^i(t)), x_i(t+1) - x_i(t) \right\rangle. \tag{20}$$

On the other hand, Assumption 1.2 implies:

$$\begin{aligned}
&f(\mathbf{x}(t+1)) - f(\mathbf{x}(t)) \\
&\le \langle \mathbf{x}(t+1) - \mathbf{x}(t), \nabla f(\mathbf{x}(t)) \rangle + \frac{L}{2} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2.
\end{aligned} \tag{21}$$

Combining (20) and (21), we have:

$$\begin{aligned}
&F(\mathbf{x}(t+1)) - F(\mathbf{x}(t)) - \tfrac{1}{2}(L - 1/\eta)\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 \\
&\le \sum_{i=1}^{m} \left\langle x_i(t+1) - x_i(t), \nabla_i f(\mathbf{x}(t)) - \nabla_i f(\mathbf{x}^i(t)) \right\rangle \\
&\le \sum_{i=1}^{m} \|x_i(t+1) - x_i(t)\| \cdot \|\nabla_i f(\mathbf{x}(t)) - \nabla_i f(\mathbf{x}^i(t))\| \\
&\overset{\text{(i)}}{\le} \sum_{i=1}^{m} \|x_i(t+1) - x_i(t)\| \cdot L\|\mathbf{x}(t) - \mathbf{x}^i(t)\| \qquad (22) \\
&\overset{\text{(ii)}}{\le} L \cdot \sum_{i=1}^{m} \|x_i(t+1) - x_i(t)\| \cdot \sum_{k=(t-s)_+}^{t-1} \|\mathbf{x}(k+1) - \mathbf{x}(k)\| \\
&\overset{\text{(iii)}}{\le} \sqrt{m}L\|\mathbf{x}(t+1) - \mathbf{x}(t)\| \cdot \sum_{k=(t-s)_+}^{t-1} \|\mathbf{x}(k+1) - \mathbf{x}(k)\| \qquad (23) \\
&\overset{\text{(iv)}}{\le} \frac{\sqrt{m}L}{2} \sum_{k=(t-s)_+}^{t-1} \left[\|\mathbf{x}(k+1) - \mathbf{x}(k)\|^2 + \|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 \right] \\
&\le \frac{\sqrt{m}Ls}{2} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 + \frac{\sqrt{m}L}{2} \sum_{k=(t-s)_+}^{t-1} \|\mathbf{x}(k+1) - \mathbf{x}(k)\|^2,
\end{aligned} \tag{24}$$

where (i) is due to the L-Lipschitz continuity of ∇f, (ii) follows from (18), (iii) is the Cauchy-Schwarz inequality, and (iv) follows from the elementary inequality $ab \le \frac{a^2+b^2}{2}$. Summing the above inequality over t from 0 to $n-1$ and rearranging we obtain

$$\begin{aligned}
F(\mathbf{x}(n)) - F(\mathbf{x}(0)) &\le \frac{1}{2}(L + \sqrt{m}Ls - 1/\eta) \sum_{t=0}^{n-1} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 \\
&\quad + \frac{L}{2} \sum_{t=0}^{n-1} \sum_{k=(t-s)_+}^{t-1} \|\mathbf{x}(k+1) - \mathbf{x}(k)\|^2 \\
&\le \frac{1}{2}(L + 2\sqrt{m}Ls - 1/\eta) \sum_{t=0}^{n-1} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2.
\end{aligned}$$

Therefore, if we choose $0 < \eta < \frac{1}{L(1+2\sqrt{ms})}$, then let $n \to \infty$ we deduce

$$\sum_{t=0}^{\infty} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 \le \frac{2}{1/\eta - L - 2\sqrt{m}Ls} [F(\mathbf{x}(0)) - \inf_{\mathbf{z}} F(\mathbf{z})]. \tag{25}$$

By Assumption 1.1, F is bounded from below, hence the right-hand side is finite. □

The first assertion of the above theorem states that the global sequence $\mathbf{x}(t)$ has square summable successive differences, while the second assertion implies that both the successive difference of the global sequence and the inconsistency between the local sequences

and the global sequence diminish as the number of iterations grows. These two conclusions provide a prelimenary stability guarantee for HET-GMP.

Next, we prove that the limit points (if exist) of the sequences $\mathbf{x}(t)$ and $\mathbf{x}^i(t), i = 1, \dots, m$ coincide, and they are critical points of F. Currently, no convexity assumption is imposed on f.

Theorem 3. *Under the same setting as in Theorem 2, the sequences* $\{\mathbf{x}(t)\}$ *and* $\{\mathbf{x}^i(t)\}, i = 1, \dots, m$ *share the same set of limit points, which is a subset of* crit F.

Proof. It is clear from Theorem 2 that $\{\mathbf{x}(t)\}$ and $\{\mathbf{x}^i(t)\}, i = 1, \dots, m$, share the same set of limit points, and we need to show that any limit point of $\{\mathbf{x}(t)\}$ is also a critical point of F. In the following, we need to prove the gradients to zero. Consider worker i and the iterations $\hat{t} \in T_i$, we have

$$\begin{aligned}
&\|\nabla_i f(\mathbf{x}(\hat{t}+1))\| \\
&\le \|\nabla_i f(\mathbf{x}(\hat{t}))\| + \|\nabla_i f(\mathbf{x}(\hat{t}+1)) - \nabla_i f(\mathbf{x}(\hat{t}))\| \\
&\overset{\text{(i)}}{\le} \left\|\tfrac{1}{\eta}\left[x_i(\hat{t}+1) - x_i(\hat{t})\right] + \nabla_i f\big(\mathbf{x}^i(\hat{t})\big) - \nabla_i f\big(\mathbf{x}(\hat{t})\big)\right\| + L\|\mathbf{x}(\hat{t}+1) - \mathbf{x}(\hat{t})\| \\
&\overset{\text{(ii)}}{\le} \tfrac{1}{\eta}\|x_i(\hat{t}+1) - x_i(\hat{t})\| + L\|\mathbf{x}^i(\hat{t}) - \mathbf{x}(\hat{t})\| + L\|\mathbf{x}(\hat{t}+1) - \mathbf{x}(\hat{t})\| \\
&\overset{\text{(iii)}}{\le} \tfrac{1}{\eta}\|x_i(\hat{t}+1) - x_i(\hat{t})\| + L\sum_{k=(\hat{t}-s)_+}^{\hat{t}} \|\mathbf{x}(k+1) - \mathbf{x}(k)\|, \qquad (26)
\end{aligned}$$

where (i) and (ii) are due to the L-Lipschitz continuity of ∇f, and (iii) follows from (18). Next, consider any other $t \notin T_i$, we denote $\hat{t}$ as the *largest* element in the set $\{k \le t : k \in T_i\}$. By Assumption 2.1, $\hat{t}$ always exists and $t - \hat{t} \le s$.

$$\begin{aligned}
&\|\nabla_i f(\mathbf{x}(t+1)) - \nabla_i f(\mathbf{x}(\hat{t}+1))\| \\
&\le \sum_{k=\hat{t}+1}^{t} \|\nabla_i f(\mathbf{x}(k+1)) - \nabla_i f(\mathbf{x}(k))\| \\
&\le \sum_{k=(t-s+1)_+}^{t} \|\nabla_i f(\mathbf{x}(k+1)) - \nabla_i f(\mathbf{x}(k))\| \\
&\le \sum_{k=(t-s+1)_+}^{t} L\|\mathbf{x}(k+1) - \mathbf{x}(k)\|. \qquad (27)
\end{aligned}$$

Combining the two cases in (26) and (27) we have for all t:

$$\|\mathbf{u}(t+1) + \nabla f(\mathbf{x}(t+1))\| \le (\sqrt{m}/\eta + 2L) \sum_{k=(t-2s)_+}^{t} \|\mathbf{x}(k+1) - \mathbf{x}(k)\|, \qquad (28)$$

where the factor $\sqrt{m} \ge 1$ is artificially introduced for the convenience of subsequent analysis. Therefore, by 28 and 2 we deduce

$$\lim_{t\to\infty} \operatorname{dist}_{\partial F(\mathbf{x}(t+1))}(\mathbf{0}) \le \lim_{t\to\infty} \|\mathbf{u}(t+1) + \nabla f(\mathbf{x}(t+1))\| = 0. \qquad (29)$$

□

B.2 Proof of theorem 2

Proof. We first show that (14) implies (15). Indeed, recall from (18):

$$\|\mathbf{x}^i(t+1) - \mathbf{x}^i(t)\| \le \sum_{k=(t-s)_+}^{t} \|\mathbf{x}(k+1) - \mathbf{x}(k)\|.$$

Therefore, summing for $t = 0, 1, \cdots, n$ gives

$$\begin{aligned}
\sum_{t=0}^{n} \|\mathbf{x}^i(t+1) - \mathbf{x}^i(t)\| &\le \sum_{t=0}^{n} \sum_{k=(t-s)_+}^{t} \|\mathbf{x}(k+1) - \mathbf{x}(k)\| \\
&\le (2s+1) \sum_{t=0}^{n} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|.
\end{aligned}$$

The claim then follows by letting n tend to infinity.

By Theorem 2, the limit points of $\{\mathbf{x}(t)\}$ and $\{\mathbf{x}^i(t)\}_{i=1}^m$ coincide and are critical points of F. Thus, the only thing left to prove is the finite length property in 14. By Theorem 3 and 1, the objective value $F(\mathbf{x}(t))$ decreases to a finite limit F^*. For all $\mathbf{x}^* \in \Omega$, we have $F(\mathbf{x}^*) = F^*$. Now fix $\varepsilon > 0$. Since Ω is compact, for t sufficiently large we have $\operatorname{dist}_\Omega(\mathbf{x}(t)) \le \varepsilon$. We now have all ingredients to apply the KŁ inequality for all sufficiently large t,

$$\varphi'\big(F(\mathbf{x}(t)) - F^*\big) \cdot \operatorname{dist}_{\partial F(\mathbf{x}(t))}(\mathbf{0}) \ge 1. \qquad (30)$$

Since φ is concave, we obtain

$$\begin{aligned}
\Delta_{t,t+1} &:= \varphi\big(F(\mathbf{x}(t)) - F^*\big) - \varphi\big(F(\mathbf{x}(t+1)) - F^*\big) \\
&\ge \varphi'\big(F(\mathbf{x}(t)) - F^*\big)\big(F(\mathbf{x}(t)) - F(\mathbf{x}(t+1))\big) \\
&\overset{\text{(i)}}{\ge} \frac{\alpha\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2}{\operatorname{dist}_{\partial F(\mathbf{x}(t))}(\mathbf{0})}, \qquad (31)
\end{aligned}$$

where (i) follows from Assumption 3 and (30). It is clear that the function φ (composed with F) serves as a Lyapunov function. Using the elementary inequality $2\sqrt{ab} \le a + b$ we obtain from 31 that for t sufficiently large,

$$2\|\mathbf{x}(t+1) - \mathbf{x}(t)\| \le \tfrac{\delta}{\alpha}\Delta_{t,t+1} + \tfrac{1}{\delta}\operatorname{dist}_{\partial F(\mathbf{x}(t))}(\mathbf{0}),$$

where $\delta > 0$ will be specified later. Recalling the bound for $\partial F(\mathbf{x}(t))$ in 28, and summing over t from $\hat{n}$ (sufficiently large) to n gives:

$$\begin{aligned}
2\sum_{t=\hat{n}}^{n} \|\mathbf{x}(t+1) - \mathbf{x}(t)\| &\le \sum_{t=\hat{n}}^{n} \frac{\delta}{\alpha}\Delta_{t,t+1} + \sum_{t=\hat{n}}^{n} \frac{1}{\delta}\operatorname{dist}_{\partial F(\mathbf{x}(t))}(\mathbf{0}) \\
&\overset{\text{(i)}}{\le} \frac{\delta}{\alpha}\varphi\big(F(\mathbf{x}(\hat{n})) - F^*\big) + \sum_{t=\hat{n}}^{n} \frac{\sqrt{m}/\eta + 2L}{\delta} \sum_{k=(t-2s)_+}^{t} \|\mathbf{x}(k+1) - \mathbf{x}(k)\| \\
&\le \frac{\delta}{\alpha}\varphi\big(F(\mathbf{x}(\hat{n})) - F^*\big) + \frac{(2s+1)(\sqrt{m}/\eta + 2L)}{\delta} \sum_{k=(\hat{n}-2s)_+}^{\hat{n}-1} \|\mathbf{x}(k+1) - \mathbf{x}(k)\| \\
&\quad + \frac{(2s+1)(\sqrt{m}/\eta + 2L)}{\delta} \sum_{t=\hat{n}}^{n} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|,
\end{aligned}$$

where (i) is due to (28). Setting $\delta = (2s+1)(\sqrt{m}/\eta + 2L)$ and rearranging gives

$$\sum_{t=\hat{n}}^{n} \|\mathbf{x}(t+1) - \mathbf{x}(t)\| \leq \frac{(2s+1)(\sqrt{m}/\eta + 2L)}{\alpha} \varphi\big(F(\mathbf{x}(\hat{n})) - F^*\big) + \sum_{k=(\hat{n}-2s)_+}^{\hat{n}-1} \|\mathbf{x}(k+1) - \mathbf{x}(k)\|.$$

Since the right-hand side is finite, let n tend to infinity completes the proof for (14). Then we prove the global convergence rate. We define that for any n:

$$\begin{aligned}\sum_{t=0}^{n} \|\mathbf{h}(t)\| &= \eta \sum_{t=0}^{n} \|(\nabla_1 f(\mathbf{x}^1(t)) - \nabla_1 f(\mathbf{x}(t)), \ldots, \\ &\quad (\nabla_m f(\mathbf{x}^m(t)) - \nabla_m f(\mathbf{x}(t))\| \\ &\leq \eta \sum_{t=0}^{n} \sum_{i=1}^{m} L\|\mathbf{x}(t) - \mathbf{x}^i(t)\| \\ &\leq \eta m L \sum_{t=0}^{n} \sum_{k=(t-s)_+}^{t-1} \|\mathbf{x}(k+1) - \mathbf{x}(k)\| \\ &\leq \eta s m L \sum_{t=0}^{n-1} \|\mathbf{x}(k+1) - \mathbf{x}(k)\|.\end{aligned}$$

Let n to be infinity, we have:

$$\sum_{t=0}^{\infty} \|\mathbf{h}(t)\| \leq \eta s m L \sum_{t=0}^{\infty} \|\mathbf{x}(k+1) - \mathbf{x}(k)\| < \infty. \tag{32}$$

Recall from [29], for a convex F, we have the following convergence rate:

$$F\Big(\frac{1}{t}\sum_{k=1}^{t} \mathbf{x}(k)\Big) - F_{\text{inf}} \leq \frac{(\|\mathbf{x}(0) - \mathbf{x}^*\| + 2\sum_{k=0}^{t} \|\mathbf{h}(k)\|)^2}{2t\eta}. \tag{33}$$

□