Supplementary Materials of CAFE+

A Mathematical Proofs

A.1 Effect of HotSketch on Identifying Hot Features

Theorem 3.1. *Given a data stream with n features, and suppose their importance score vector is $a = \{a_1, a_2, \cdots, a_n\}$, where $a_1 \geqslant a_2 \geqslant \cdots \geqslant a_n$. Suppose that our HotSketch has w buckets, and each bucket contains c cells. Without distribution assumption, for a hot feature with a total score larger than $\gamma\|a\|_1$, it can be held in HotSketch with probability at least:* $\Pr > 1 - \frac{1-\gamma}{(c-1)\gamma w}$.

Proof. The expected score sum of the other features $\hat{f}$ entering the same bucket is

$$\mathbb{E}\left[\hat{f}\right] = \frac{(1-\gamma)\|a\|}{w}$$

By following the properties of Space-Saving algorithm, if the score $\hat{f}$ of the other features entering the bucket is no more than $(c-1)\gamma\|a\|_1$, then the feature must be held in the bucket.

Using Markov inequality, we have

$$\Pr\left(\hat{f} > (c-1)\gamma\|a\|_1\right) \leqslant \frac{1-\gamma}{(c-1)\gamma w}$$

which means that $\Pr > 1 - \frac{1-\gamma}{(c-1)\gamma w}$. □

Lemma 3.2. *Given a data stream with score vector $a = \{a_1, a_2, \cdots, a_n\}$, where $a_1 \geqslant a_2 \geqslant \cdots \geqslant a_n$. Suppose that a follows a Zipfian distribution with parameter z, meaning that $a_i = \frac{a_1}{i^z}$. Suppose our HotSketch has w buckets, and each bucket contains c cells. Suppose we would like to check whether the k' hottest features can be hashed into the buckets. Then the mathematical expectation of the score sum of the non-hot features entering each bucket is:* $\mathbb{E}[\hat{f}] \leqslant \frac{\|a\|_1 \cdot k'^{1-z}}{w}$ *with probability at least $3^{-\frac{k'}{w}}$ for $z > 1$ and $n \to +\infty$.*

Proof. The probability that the k' hottest features are not hashed into this bucket is

$$\left(1 - \frac{1}{w}\right)^{k'} = \left(\left(1 - \frac{1}{w}\right)^{w}\right)^{\frac{k'}{w}} > 3^{-\frac{k'}{w}}$$

When $w \geqslant 6$, $\left(1 - \frac{1}{w}\right)^{w}$ increases monotonically with w. The expected score sum of the non-hot features entering this bucket is:

$$\begin{aligned}
\mathbb{E}[\hat{f}] =& \frac{\sum_{i=k'+1}^{n} a_i}{w} = \frac{\sum_{i=k'+1}^{n} \frac{a_1}{i^z}}{w} = \frac{\|a\|_1}{w} \cdot \left(\sum_{i=k'+1}^{n} i^{-z}\right) \cdot \frac{1}{\sum_{i=1}^{n} i^{-z}} \\
\leqslant& \frac{\|a\|_1}{w} \cdot \left(\int_{k'}^{+\infty} x^{-z} dx\right) \cdot \left(\int_{1}^{+\infty} x^{-z} dx\right)^{-1} \\
\leqslant& \frac{\|a\|_1}{w} \cdot \frac{k'^{1-z}}{z-1} \cdot (z-1) = \frac{\|a\|_1 k'^{1-z}}{w}
\end{aligned}$$

for $z > 1$ and $n \to +\infty$. □

Theorem 3.3. *Given a data stream with score vector $a = \{a_1, a_2, \cdots, a_n\}$, where $a_1 \geqslant a_2 \geqslant \cdots \geqslant a_n$. Suppose that a follows a Zipfian distribution with parameter z. Suppose that our HotSketch has w buckets, and each bucket contains c cells. Let $k' = \eta w$. Then for a hot feature with a score larger than $\gamma\|a\|_1$, it can be held in the sketch with probability at least:* $\Pr > \sup_{\eta>0} \left(3^{-\eta} \cdot \left(1 - \frac{\eta}{(c-1)\gamma(\eta w)^z}\right)\right)$ *for $z > 1$ and $n \to +\infty$.*

Proof. The condition $\mathcal{C}$ that none of the k' hottest features collide with this item holds with probability at least $3^{-\frac{k'}{w}}$.

By following the properties of SpaceSaving algorithm, if the scores $\hat{f}$ of the other features entering the bucket is no more than $(c-1)\gamma\|a\|_1$, then the feature must be held in the bucket.

Using Markov inequality and Lemma 3.2, we have

$$\Pr\left(\hat{f} > (c-1)\gamma\|a\|_1 \mid \mathcal{C}\right) \leqslant \frac{\frac{\|a\|_1 \cdot k'^{1-z}}{w}}{(c-1)\gamma\|a\|_1} = \frac{k'^{1-z}}{(c-1)\gamma w}.$$

Then we have

$$\begin{aligned}\Pr\left(\hat{f} > (c-1)\gamma\|a\|_1\right) \leqslant & \Pr\left(\hat{f} > (c-1)\gamma\|a\|_1, \mathcal{C}\right) + \Pr\left(\neg\mathcal{C}\right) \\ \leqslant & 3^{-\frac{k'}{w}} \cdot \left(\frac{k'^{1-z}}{(c-1)\gamma w} - 1\right) + 1.\end{aligned}$$

Let $k' = \eta w$, we have

$$\Pr\left(\hat{f} > (c-1)\gamma\|a\|_1\right) \leqslant 3^{-\eta} \cdot \left(\frac{1}{\eta^{z-1}(c-1)\gamma w^z} - 1\right) + 1.$$

And we have the probability that this feature must be held greater than

$$\Pr > \sup_{\eta>0} \left(3^{-\eta} \cdot \left(1 - \frac{\eta}{(c-1)\gamma(\eta w)^z}\right)\right).$$

□

Corollary 3.4. *The larger the parameter c, w, z, and γ, the larger the probability that the feature with score larger than $\gamma\|a\|_1$ be held in the sketch. The larger c and w means the larger memory used by sketch, the larger z means the more skew the data stream is, and the larger γ means the hotter the feature is.*

Proof. The following formula monotonically decreases with parameter c, w, z, and γ: $\frac{\eta}{(c-1)\gamma(\eta w)^z}$. □

Corollary 3.5. *To let the feature with score larger than $\gamma\|a\|_1$ be held with maximum probability in a fixed memory budget, the more skew the data stream is, the less cells per bucket should be used. Specifically, we recommend to use $c = 1 + \frac{1}{z-1}$.*

Proof. With a fixed memory budget $M = cw$, to minimize $\frac{\eta}{(c-1)\gamma(\eta w)^z}$, we should maximize $(c-1)w^z = \left(\frac{M}{w} - 1\right) w^z$.

As it has a derivative function

$$\left[\left(\frac{M}{w} - 1\right) w^z\right]' = ((z-1)M - zw)w^{z-2},$$

the optimal w should be $\frac{z-1}{z}M$, and the optimal c^* should be

$$c^* = \frac{z}{z-1} = 1 + \frac{1}{z-1}.$$

□

Table 1: Symbols used in Section A.2.

Symbol	Meaning
f	Neural network
N	Number of training samples
$\boldsymbol{\theta}$	Learnable parameters
L	Lipschitz constant bounding gradient changes
σ_0	Bound on the expected norm of gradient
σ	Bound on the expected norm of gradient difference
α	Learning rate
$\boldsymbol{g}$	Standard gradient without compression
$\widetilde{\boldsymbol{g}}$	Gradient in compressed DLRM
T	Number of training iterations
ϵ_t	Deviation of embedding gradients

A.2 Convergence Analysis against Deviation

We study the following (non-convex) empirical risk minimization problem:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^D} f(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i}^{N} f_i(\boldsymbol{\theta}), \quad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha\widetilde{\boldsymbol{g}}_{\boldsymbol{i_t}}$$

where α is learning rate, $\boldsymbol{g}_{\boldsymbol{i_t}} = \nabla f_i(\boldsymbol{\theta}_{i_t})$ is the standard gradient without compression, $\widetilde{\boldsymbol{g}}_{\boldsymbol{i_t}}$ is the real gradient with compression.

Assumption 1. *For $\forall i \in \{1, 2, ..., N\},\ \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^D$, we make the following assumptions:*

(1. L-Lipschitz) $\|\nabla f_i(\boldsymbol{\theta}) - \nabla f_i(\boldsymbol{\theta}')\| < L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$;

(2. Bounded moment) $\mathbb{E}[\|\nabla f_i(\boldsymbol{\theta})\|] < \sigma_0,\ \mathbb{E}[\|\nabla f(\boldsymbol{\theta})\|] < \sigma_0$;

(3. Bounded variance) $\mathbb{E}[\|\nabla f_i(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})\|] < \sigma$;

(4. Existence of global minimum) $\exists f^*\ s.t. f(\boldsymbol{\theta}) \geqslant f^*$.

Theorem 3.6. *Suppose we run SGD optimization with CAFE+ on DLRMs satisfying the assumptions above, with $\epsilon_t = \|\widetilde{\boldsymbol{g}}_{\boldsymbol{i_t}} - \boldsymbol{g}_{\boldsymbol{i_t}}\|$ as the deviation of embedding gradients. Assume the learning rate α satisfies $\alpha < \frac{1}{L}$. After T steps, for $\overline{\boldsymbol{\theta}}_T$ which is randomly selected from $\{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{T-1}\}$, we have:*

$$\mathbb{E}[\|\nabla f(\overline{\boldsymbol{\theta}}_T)\|^2] \leqslant \frac{f(\boldsymbol{\theta}_0) - f^*}{T\alpha(1-\alpha L)} + \frac{\alpha(2L\sigma^2 + \sigma_0^2)}{2(1-\alpha L)} + \frac{(1+\alpha^2 L)\sum_{t=0}^{T-1}\mathbb{E}[\epsilon_t^2]}{2T\alpha(1-\alpha L)}$$

Proof. By Taylor's Expansion Formula with Lagrangian Remainder,

$$\begin{aligned}
f(\boldsymbol{\theta}_{t+1}) &= f(\boldsymbol{\theta}_t - \alpha\widetilde{\boldsymbol{g}}_{\boldsymbol{i_t}}) \\
&= f(\boldsymbol{\theta}_t - \alpha\boldsymbol{g}_{\boldsymbol{i_t}} + \alpha(\boldsymbol{g}_{\boldsymbol{i_t}} - \widetilde{\boldsymbol{g}}_{\boldsymbol{i_t}})) \\
&= f(\boldsymbol{\theta}_t - \alpha\boldsymbol{g}_{\boldsymbol{i_t}}) + \alpha(\boldsymbol{g}_{\boldsymbol{i_t}} - \widetilde{\boldsymbol{g}}_{\boldsymbol{i_t}})^T\nabla f(\boldsymbol{\theta}_t - \alpha\boldsymbol{g}_{\boldsymbol{i_t}}) + \frac{1}{2}\alpha^2(\boldsymbol{g}_{\boldsymbol{i_t}} - \widetilde{\boldsymbol{g}}_{\boldsymbol{i_t}})^T\nabla^2 f(\boldsymbol{\psi_t})(\boldsymbol{g}_{\boldsymbol{i_t}} - \widetilde{\boldsymbol{g}}_{\boldsymbol{i_t}}) \\
&\leqslant f(\boldsymbol{\theta}_t - \alpha\boldsymbol{g}_{\boldsymbol{i_t}}) + \frac{1}{2}(\|\boldsymbol{g}_{\boldsymbol{i_t}} - \widetilde{\boldsymbol{g}}_{\boldsymbol{i_t}}\|^2 + \alpha^2\|\nabla f(\boldsymbol{\theta}_t - \alpha\boldsymbol{g}_{\boldsymbol{i_t}})\|^2) + \frac{1}{2}\alpha^2 L\|\boldsymbol{g}_{\boldsymbol{i_t}} - \widetilde{\boldsymbol{g}}_{\boldsymbol{i_t}}\|^2 \\
&\leqslant f(\boldsymbol{\theta}_t - \alpha\boldsymbol{g}_{\boldsymbol{i_t}}) + \frac{1}{2}\alpha^2\sigma_0^2 + \frac{1}{2}(1+\alpha^2 L)\epsilon_t^2,
\end{aligned}$$

where the first inequality is due to GM-QM inequality $ab \leqslant \frac{1}{2}(a^2 + b^2)$ and the property of Lipschitz continuity, and the second inequality is due to the bounded momentum. Again, using Taylor's Expansion Formula with Lagrangian Remainder,

$$
\begin{aligned}
f(\boldsymbol{\theta}_t - \alpha \boldsymbol{g}_{\boldsymbol{i}_t}) &= f(\boldsymbol{\theta}_t) - \alpha \boldsymbol{g}_{\boldsymbol{i}_t}^T \nabla f(\boldsymbol{\theta}_t) + \frac{1}{2}\alpha^2 \boldsymbol{g}_{\boldsymbol{i}_t}^T \nabla^2 f(\boldsymbol{\psi}_t') \boldsymbol{g}_{\boldsymbol{i}_t} \\
&\leqslant f(\boldsymbol{\theta}_t) - \alpha \boldsymbol{g}_{\boldsymbol{i}_t}^T \nabla f(\boldsymbol{\theta}_t) + \frac{1}{2}\alpha^2 L \|\boldsymbol{g}_{\boldsymbol{i}_t}\|^2 \\
&= f(\boldsymbol{\theta}_t) - \alpha \boldsymbol{g}_{\boldsymbol{i}_t}^T \nabla f(\boldsymbol{\theta}_t) + \frac{1}{2}\alpha^2 L \|\nabla f(\boldsymbol{\theta}_t) + (\boldsymbol{g}_{\boldsymbol{i}_t} - \nabla f(\boldsymbol{\theta}_t))\|^2 \\
&\leqslant f(\boldsymbol{\theta}_t) - \alpha \boldsymbol{g}_{\boldsymbol{i}_t}^T \nabla f(\boldsymbol{\theta}_t) + \alpha^2 L (\|\nabla f(\boldsymbol{\theta}_t)\|^2 + \|(\boldsymbol{g}_{\boldsymbol{i}_t} - \nabla f(\boldsymbol{\theta}_t))\|^2),
\end{aligned}
$$

where the first inequality is still due to the property of Lipschitz continuity, and the second inequality is due to the AM-QM inequality $(a+b)^2 \leqslant 2(a^2+b^2)$

Notice that $\mathbb{E}[\boldsymbol{g}_{\boldsymbol{i}_t}] = \nabla f(\boldsymbol{\theta}_t)$, we combine the above inequalities and take the expectation on both sides,

$$
\begin{aligned}
\mathbb{E}[f(\boldsymbol{\theta}_{t+1})] &\leqslant \mathbb{E}[f(\boldsymbol{\theta}_t - \alpha \boldsymbol{g}_{\boldsymbol{i}_t})] + \frac{1}{2}\alpha^2 \sigma_0^2 + \frac{1}{2}(1+\alpha^2 L)\mathbb{E}[\epsilon_t^2] \\
&\leqslant \mathbb{E}[f(\boldsymbol{\theta}_t)] - (\alpha - \alpha^2 L)\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] + \alpha^2 L \sigma^2 + \frac{1}{2}\alpha^2 \sigma_0^2 + \frac{1}{2}(1+\alpha^2 L)\mathbb{E}[\epsilon_t^2].
\end{aligned}
$$

Rearranging the inequality and summing over t from 0 to $T-1$, we have

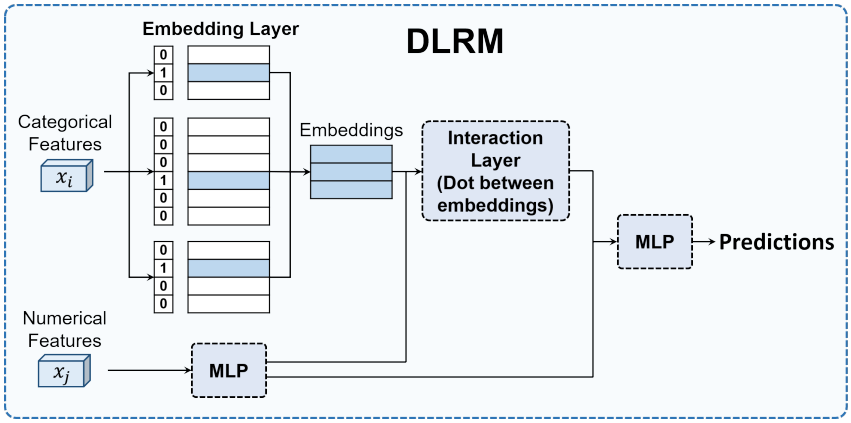
$$
\sum_{t=0}^{T-1}(\alpha - \alpha^2 L)\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \leqslant f(\boldsymbol{\theta}_0) - \mathbb{E}[f(\boldsymbol{\theta}_T)] + T\alpha^2(L\sigma^2 + \frac{1}{2}\sigma_0^2) + \frac{1}{2}(1+\alpha^2 L)\sum_{t=0}^{T-1}\mathbb{E}[\epsilon_t^2]
$$

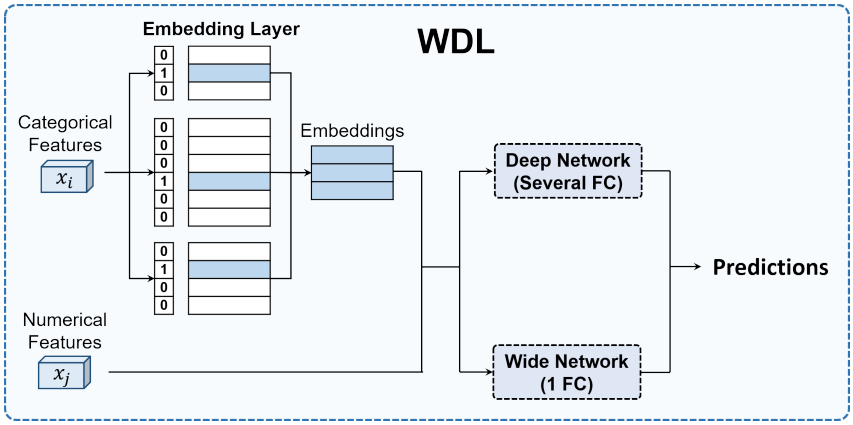
Therefore,

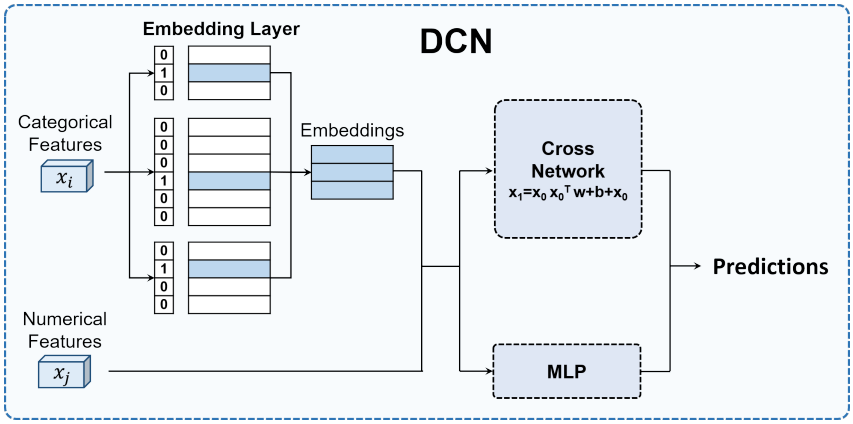
$$
\begin{aligned}
\mathbb{E}[\|\nabla f(\overline{\boldsymbol{\theta}})\|^2] &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \\
&\leqslant \frac{f(\boldsymbol{\theta}_0) - f^*}{T\alpha(1-\alpha L)} + \frac{\alpha(2L\sigma^2 + \sigma_0^2)}{2(1-\alpha L)} + \frac{(1+\alpha^2 L)\sum_{t=0}^{T-1}\mathbb{E}[\epsilon_t^2]}{2T\alpha(1-\alpha L)}
\end{aligned}
$$

□

B Model Structures







C Model Size

#Param (Emb,NN)	Avazu	Criteo	KDD12	CriteoTB
DLRM	(150M,250K)	(540M,480K)	(3.5B,160K)	(26B,540K)
WDL	(150M,160K)	(540M,180K)	(3.5B,250K)	(26B,920K)
DCN	(150M,220K)	(540M,240K)	(3.5B,320K)	(26B,1.0M)

Table 2: The number of parameters.

In each tuple, the left part is the size of the embedding table, while the right part is the size of the neural network part. The size of the neural network part is negligible, since the embedding table takes up more than 99.9% parameters in most cases.

D Throughput of AutoEncoder

The table lists the training throughput (sample per second) of each method. AutoEncoder has very low training throughput, because it has to update the entire decoder matrix, and the size of the decoder matrix scales linearly with the number of unique features.

Hash	Q-R Trick	AdaEmbed	MDE	CAFE	AutoEncoder
8015	4757	4571	5740	4589	159

Table 3: Training throughput on Criteo ($5\times$).