Supplementary file for CAFE

A Proof of Theorem 3.6

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

$$\begin{aligned}
f(\theta_{t+1}) &= f(\theta_t - \alpha\widetilde{g}_{i_t}) \\
&= f(\theta_t - \alpha g_{i_t} + \alpha(g_{i_t} - \widetilde{g}_{i_t})) \\
&= f(\theta_t - \alpha g_{i_t}) + \alpha(g_{i_t} - \widetilde{g}_{i_t})^T \nabla f(\theta_t - \alpha g_{i_t}) \\
&\quad + \frac{1}{2}\alpha^2 (g_{i_t} - \widetilde{g}_{i_t})^T \nabla^2 f(\psi_t)(g_{i_t} - \widetilde{g}_{i_t}) \\
&\leq f(\theta_t - \alpha g_{i_t}) + \frac{1}{2}(\|g_{i_t} - \widetilde{g}_{i_t}\|^2 + \alpha^2 \|\nabla f(\theta_t - \alpha g_{i_t})\|^2) \\
&\quad + \frac{1}{2}\alpha^2 L \|g_{i_t} - \widetilde{g}_{i_t}\|^2 \\
&\leq f(\theta_t - \alpha g_{i_t}) + \frac{1}{2}\alpha^2\sigma_0^2 + \frac{1}{2}(1 + \alpha^2 L)\epsilon_t^2,
\end{aligned}$$

where the first inequality is due to GM-QM inequality $ab \leq \frac{1}{2}(a^2 + b^2)$ 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,

$$\begin{aligned}
f(\theta_t - \alpha g_{i_t}) &= f(\theta_t) - \alpha g_{i_t}^T \nabla f(\theta_t) + \frac{1}{2}\alpha^2 g_{i_t}^T \nabla^2 f(\psi_t') g_{i_t} \\
&\leq f(\theta_t) - \alpha g_{i_t}^T \nabla f(\theta_t) + \frac{1}{2}\alpha^2 L \|g_{i_t}\|^2 \\
&= f(\theta_t) - \alpha g_{i_t}^T \nabla f(\theta_t) \\
&\quad + \frac{1}{2}\alpha^2 L \|\nabla f(\theta_t) + (g_{i_t} - \nabla f(\theta_t))\|^2 \\
&\leq f(\theta_t) - \alpha g_{i_t}^T \nabla f(\theta_t) \\
&\quad + \alpha^2 L(\|\nabla f(\theta_t)\|^2 + \|(g_{i_t} - \nabla f(\theta_t))\|^2),
\end{aligned}$$

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

Notice that $\mathbb{E}[g_{i_t}] = \nabla f(\theta_t)$, we combine the above inequalities and take the expectation on both sides,

$$\begin{aligned}
\mathbb{E}[f(\theta_{t+1})] &\leq \mathbb{E}[f(\theta_t - \alpha g_{i_t})] + \frac{1}{2}\alpha^2\sigma_0^2 + \frac{1}{2}(1 + \alpha^2 L)\mathbb{E}[\epsilon_t^2] \\
&\leq \mathbb{E}[f(\theta_t)] - (\alpha - \alpha^2 L)\mathbb{E}[\|\nabla f(\theta_t)\|^2] + \alpha^2 L\sigma^2 \\
&\quad + \frac{1}{2}\alpha^2\sigma_0^2 + \frac{1}{2}(1 + \alpha^2 L)\mathbb{E}[\epsilon_t^2].
\end{aligned}$$

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

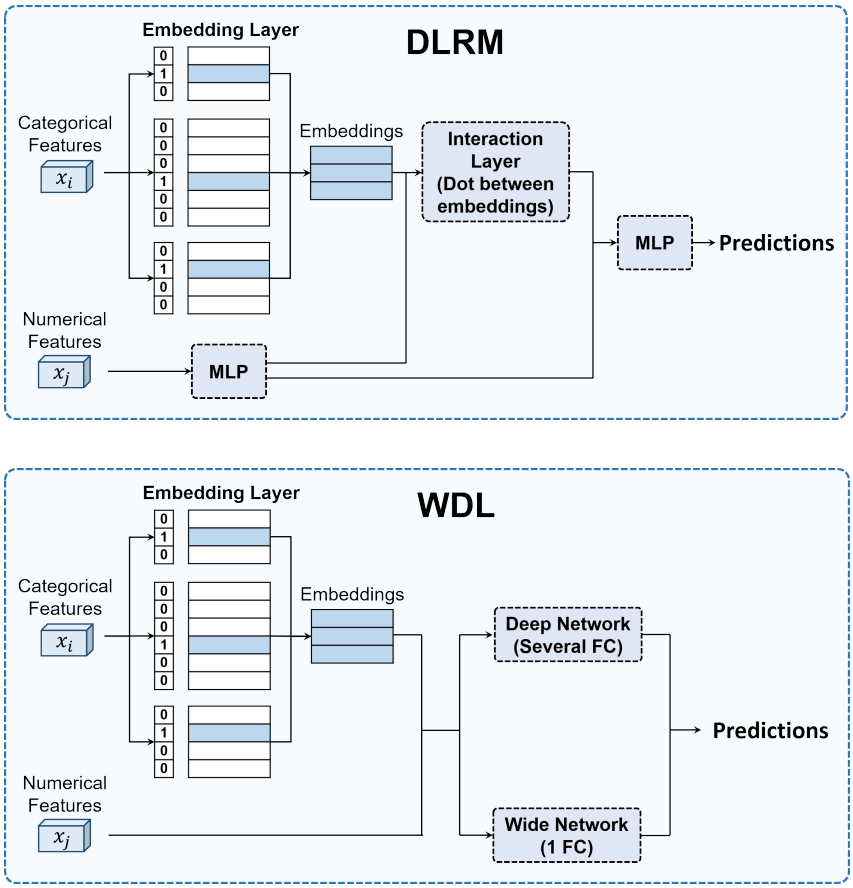
$$\begin{aligned}
\sum_{t=0}^{T-1}(\alpha - \alpha^2 L)\mathbb{E}[\|\nabla f(\theta_t)\|^2] &\leq f(\theta_0) - \mathbb{E}[f(\theta_T)] \\
&\quad + T\alpha^2(L\sigma^2 + \frac{1}{2}\sigma_0^2) + \frac{1}{2}(1 + \alpha^2 L)\sum_{t=0}^{T-1}\mathbb{E}[\epsilon_t^2]
\end{aligned}$$

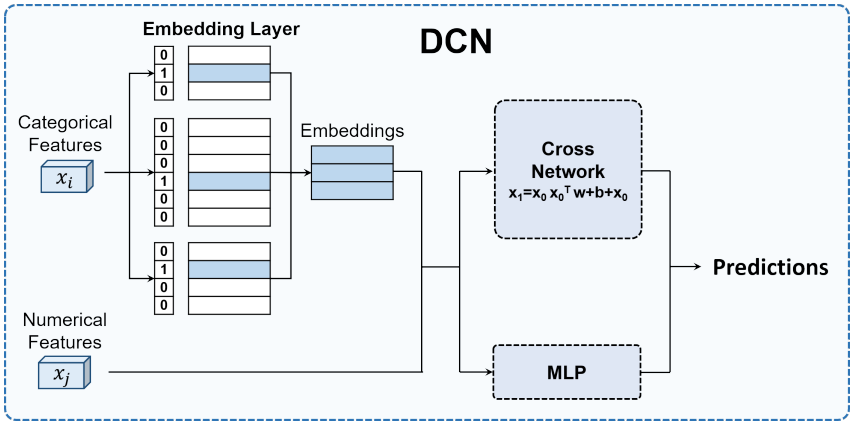
Therefore,

$$\mathbb{E}[\|\nabla f(\bar{\theta})\|^2] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\theta_t)\|^2]$$
$$\leq \frac{f(\theta_0) - f^*}{T\alpha(1-\alpha L)} + \frac{\alpha(2L\sigma^2 + \sigma_0^2)}{2(1-\alpha L)} + \frac{(1+\alpha^2 L)\sum_{t=0}^{T-1}\mathbb{E}[\epsilon_t^2]}{2T\alpha(1-\alpha L)}$$

□

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 1: 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

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

Table 2: Training throughput on Criteo ($5\times$).

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.