A Summary of Appendix

We provide more details of theoretical analysis (Appendix B and D), experiment results and settings (Appendix C) and a brief introduction to our new implementation of LightGBM CUDA version (Appendix E).

B Theoretical Analysis

For the simplicity of notations, we will use s in place of I_s to represent the indices of data instances in leaf s in this section.

B.1 Existence of $\gamma_s > 0$

Theorem B.1.1 With constant hessian value h, if leaf s has a split gain $\Delta\mathcal{L}_{s\to s_1,s_2} > 0$, then with weights $|g_i|$ and labels $\text{sign}(g_i)$, there exists $\hat{\gamma}_s > 0$ such that the split $s \to s_1, s_2$ has a weighted classification error rate $\frac{1}{2} - \hat{\gamma}_s < \frac{1}{2}$ for $\mathcal{D}_s$.

Proof of Theorem B.1.1 Since

$$\Delta\mathcal{L}_{s\to s_1,s_2} = \frac{G_{s_1}^2}{2H_{s_1}} + \frac{G_{s_2}^2}{2H_{s_2}} - \frac{G_s^2}{2H_s} = \frac{G_{s_1}^2}{2hn_{s_1}} + \frac{G_{s_2}^2}{2hn_{s_2}} - \frac{G_s^2}{2hn_s} > 0, \tag{10}$$

we have $|G_{s_1}| > 0$ or $|G_{s_2}| > 0$. W.L.O.G., suppose $|G_{s_1}| > 0$. Denote s_1^+ as the set of indices such that $\forall i \in s_1^+, \text{sign}(g_i)$ equals the weighted majority of $\text{sign}(g_i)$ for all $i \in s_1$ (weighted by $|g_i|$), and $s_1^- = s_1 - s_1^+$. Then

$$|G_{s_1}| = \left|\sum_{i\in s_1} g_i\right| = \sum_{i\in s_1^+} |g_i| - \sum_{i\in s_1^-} |g_i| > 0. \tag{11}$$

Similarly we define s_2^+ and s_2^-. Then by definition

$$|G_{s_2}| = \left|\sum_{i\in s_1} g_i\right| = \sum_{i\in s_2^+} |g_i| - \sum_{i\in s_2^-} |g_i| \geq 0. \tag{12}$$

Thus the weighted error rate

$$\frac{\sum_{i\in s_1^-} |g_i| + \sum_{i\in s_2^-} |g_i|}{\sum_{i\in s_1} |g_i| + \sum_{i\in s_2} |g_i|} = \frac{1}{2} - \frac{\sum_{i\in s_1^+} |g_i| + \sum_{i\in s_2^+} |g_i| - \sum_{i\in s_1^-} |g_i| - \sum_{i\in s_2^-} |g_i|}{\sum_{i\in s_1} |g_i| + \sum_{i\in s_2} |g_i|}. \tag{13}$$

Setting

$$\hat{\gamma}_s = \frac{\sum_{i\in s_1^+} |g_i| + \sum_{i\in s_2^+} |g_i| - \sum_{i\in s_1^-} |g_i| - \sum_{i\in s_2^-} |g_i|}{\sum_{i\in s_1} |g_i| + \sum_{i\in s_2} |g_i|} > 0 \tag{14}$$

completes the proof.

B.2 Proof of Theorem 5.3

Definition 5.1 (Weak Learnability of Stumps) Given a binary classification dataset $\mathcal{D} = \{(\mathbf{x}_i, c(\mathbf{x}_i))\}_{i=1}^N$ where $c(\mathbf{x}_i) \in \{-1, 1\}$, weights $\{w_i\}_{i=1}^N$, $w_i \geq 0$ and $\sum_i w_i > 0$, there exists $\gamma > 0$ and a two-leaf decision tree with leaf values in $\{-1, 1\}$ s.t. the weighted classification error rate on $\mathcal{D}$ is $\frac{1}{2} - \gamma$. Then the dataset $\mathcal{D}$ is γ-empirically weakly learnable by stumps w.r.t. c and $\{w_i\}_{i=1}^N$.

Assumption 5.2 Let $\text{sign}(\cdot)$ be the sign function (with $\text{sign}(0) = 1$). For data subset $\mathcal{D}_s \subset \mathcal{D}$ in leaf s, there exists a stump and a $\gamma_s > 0$ s.t. $\mathcal{D}_s$ is γ_s-empirically weakly learnable by stumps, w.r.t. concept $c(\mathbf{x}_i) = \text{sign}(g_i)$ and weights $w_i = |g_i|$, where $i \in I_s$.

Theorem 5.3 For loss functions with constant hessian value $h > 0$, if Assumption 5.2 holds for the subset $\mathcal{D}_s$ in leaf s for some $\gamma_s > 0$, then with stochastic rounding and leaf-value refitting, for any $\epsilon > 0$, and $\delta > 0$, at least one of the following conclusions holds:

1. With any split of leaf s and its descendants, the resultant average of absolute values of prediction values by the tree in current boosting iteration for data in $\mathcal{D}_s$ is no greater than ϵ/h.

2. For any split $s \to s_1, s_2$ of leaf s, with a probability of at least $1-\delta$,

$$\frac{\left|\widetilde{\mathcal{G}}_{s\to s_1,s_2} - \mathcal{G}_{s\to s_1,s_2}\right|}{\mathcal{G}_s^*} \leq \frac{\max\limits_{i\in[N]}|g_i|\sqrt{2\ln\frac{4}{\delta}}}{\gamma_s^2\epsilon\cdot 2^{B-1}}\left(\sqrt{\frac{1}{n_{s_1}}}+\sqrt{\frac{1}{n_{s_2}}}\right) + \frac{\left(\max\limits_{i\in[N]}|g_i|\right)^2\ln\frac{4}{\delta}}{\gamma_s^2\epsilon^2 n_s\cdot 4^{B-2}}. \tag{15}$$

Proof of Theorem 5.3 By leaf-wise weak learnability (Assumption 5.2), there exists a split $s \to s_L, s_R$ and $\gamma_s > 0$ for s s.t. for data in $\mathcal{D}_s$, with binary labels $c(\mathbf{x}_i) = \text{sign}(g_i)$ and weights $w_i = |g_i|$, the split results in a stump with weighted binary-classification error rate is $\frac{1}{2} - \gamma_s$. Suppose that in s_L, s_L^+ is the set of weighted majority samples, and s_L^- is the set of weighted minority samples (thus $\text{sign}(g_i) = +1, \forall i \in s_L^+$ and $\text{sign}(g_i) = -1, \forall i \in s_L^-$, or $\text{sign}(g_i) = -1, \forall i \in s_L^+$ and $\text{sign}(g_i) = +1, \forall i \in s_L^-$) such that $\sum_{i\in s_L^+}|g_i| \geq \sum_{i\in s_L^-}|g_i|$. Similarly, we define s_R^+ and s_R^-. Then we have the weighted error rate

$$\overline{\text{err}} = \frac{\sum_{i\in s_L^-}|g_i| + \sum_{i\in s_R^-}|g_i|}{\sum_{i\in s}|g_i|} = \frac{1}{2} - \gamma_s. \tag{16}$$

Thus

$$\frac{\sum_{i\in s_L^+}|g_i| + \sum_{i\in s_R^+}|g_i|}{\sum_{i\in s}|g_i|} = 1 - \overline{\text{err}} = \frac{1}{2} + \gamma_s. \tag{17}$$

Since $\mathcal{G}_s^*$ is for the optimal split in leaf s, we have

$$\begin{aligned}
\mathcal{G}_s^* \geq \mathcal{G}_{s\to s_L,s_R} &= \frac{\left(\sum_{i\in s_L} g_i\right)^2}{2hn_{s_L}} + \frac{\left(\sum_{i\in s_R} g_i\right)^2}{2hn_{s_R}} \\
&\geq \frac{\left(\left|\sum_{i\in s_L} g_i\right| + \left|\sum_{i\in s_R} g_i\right|\right)^2}{2h\left(n_{s_L}+n_{s_R}\right)} \\
&= \frac{\left(\sum_{i\in s_L^+}|g_i| - \sum_{i\in s_L^-}|g_i| + \sum_{i\in s_R^+}|g_i| - \sum_{i\in s_R^-}|g_i|\right)^2}{2h\left(n_{s_L}+n_{s_R}\right)} \\
&= \frac{\left(\sum_{i\in s_L^+\cup s_R^+}|g_i| - \sum_{i\in s_L^-\cup s_R^-}|g_i|\right)^2}{2h\left(n_{s_L}+n_{s_R}\right)} \\
&= \frac{\left(\left(\frac{1}{2}+\gamma_s\right)\sum_{i\in s}|g_i| - \left(\frac{1}{2}-\gamma_s\right)\sum_{i\in s}|g_i|\right)^2}{2h\left(n_{s_L}+n_{s_R}\right)} \\
&= \frac{2\gamma_s^2\left(\sum_{i\in s}|g_i|\right)^2}{h\left(n_{s_L}+n_{s_R}\right)} = \frac{2\gamma_s^2\left(\sum_{i\in s}|g_i|\right)^2}{hn_s}
\end{aligned} \tag{18}$$

If $\frac{\sum_{i\in s}|g_i|}{n_s} \leq \epsilon$, then suppose $s_1', ..., s_m'$ are all descendant leaves of s, then the average prediction values by current iteration for data in $\mathcal{D}_s$ in current tree is

$$\frac{\sum_{i=1}^m n_{s_i'}\left|w_{s_i'}^*\right|}{\sum_{i=1}^m n_{s_i'}} = \frac{\sum_{i=1}^m\left|\sum_{i\in s_i'} g_i\right|}{\sum_{i=1}^m n_{s_i'}h} \leq \frac{\sum_{i\in s}|g_i|}{n_s h} \leq \frac{\epsilon}{h}, \tag{19}$$

which guarantees that the first conclusion holds.

If $\frac{\sum_{i\in s}|g_i|}{n_s} > \epsilon$, let $\epsilon_i = \delta_g\widetilde{g}_i - g_i$, thus $|\epsilon_i| \leq \delta_g$ and $\mathbb{E}[\epsilon_i] = 0$. Then

$$\begin{aligned}
\left|\widetilde{\mathcal{L}}_{s_1}^* - \mathcal{L}_{s_1}^*\right| &= \left|\frac{\left(\sum_{i\in s_1} g_i + \epsilon_i\right)^2}{2hn_{s_1}} - \frac{\left(\sum_{i\in s_1} g_i\right)^2}{2hn_{s_1}}\right| \\
&= \frac{1}{2hn_{s_1}}\left|\sum_{i\in s_1} 2g_i + \epsilon_i\right|\left|\sum_{i\in s_1}\epsilon_i\right| \leq \frac{1}{2hn_{s_1}}\left(\left|\sum_{i\in s_1} 2g_i\right|\left|\sum_{i\in s_1}\epsilon_i\right| + \left|\sum_{i\in s_1}\epsilon_i\right|^2\right)
\end{aligned} \tag{20}$$

Note that ϵ_i's are independent variables. Let $t_{s_1} = \delta_g \sqrt{2n_{s_1} \ln \frac{4}{\delta}}$, then by Hoeffding's inequality,

$$P\left(\left|\sum_{i \in n_{s_1}} \epsilon_i\right| \geq t_{s_1}\right) \leq 2\exp\left(-\frac{2t_{s_1}^2}{n_{s_1}\left(2\delta_g\right)^2}\right) = \frac{\delta}{2}. \tag{21}$$

Then with a probability of at least $1 - \frac{\delta}{2}$

$$\begin{aligned}\left|\widetilde{\mathcal{L}}_{s_1}^* - \mathcal{L}_{s_1}^*\right| &\leq \frac{1}{hn_{s_1}}\left(\left|\sum_{i \in s_1} g_i\right| \cdot \delta_g \sqrt{2n_{s_1} \ln \frac{4}{\delta}} + \delta_g^2 n_{s_1} \ln \frac{4}{\delta}\right) \\ &= \frac{\left|\sum_{i \in s_1} g_i\right| \cdot \delta_g \sqrt{\frac{2 \ln \frac{4}{\delta}}{n_{s_1}}}}{h} + \frac{\delta_g^2 \ln \frac{4}{\delta}}{h}.\end{aligned} \tag{22}$$

We have

$$\begin{aligned}\frac{\left|\sum_{i \in s_1} g_i\right| \cdot \delta_g \sqrt{\frac{2 \ln \frac{4}{\delta}}{n_{s_1}}}}{h\mathcal{G}_s^*} &\leq \frac{\sum_{i \in s_1} |g_i| \cdot \delta_g \sqrt{\frac{2 \ln \frac{4}{\delta}}{n_{s_1}}}}{\frac{2\gamma_s^2\left(\sum_{i \in s} |g_i|\right)^2}{n_s}} \leq \frac{\delta_g \sqrt{\frac{2 \ln \frac{4}{\delta}}{n_{s_1}}}}{\frac{2\gamma_s^2\left(\sum_{i \in s} |g_i|\right)}{n_s}} \\ &\leq \frac{\delta_g \sqrt{\frac{2 \ln \frac{4}{\delta}}{n_{s_1}}}}{2\gamma_s^2 \epsilon} = \frac{\max_{i \in [N]} |g_i|}{2\gamma_s^2 \epsilon \left(2^{B-1} - 1\right)} \sqrt{\frac{2 \ln \frac{4}{\delta}}{n_{s_1}}}.\end{aligned} \tag{23}$$

And since

$$\mathcal{G}_s^* \geq \frac{2\gamma_s^2 \left(\sum_{i \in s} |g_i|\right)^2}{hn_s} > \frac{2\gamma_s^2 \epsilon^2 n_s}{h}, \tag{24}$$

we have

$$\frac{\delta_g^2 \ln \frac{4}{\delta}}{h\mathcal{G}_s^*} \leq \frac{\delta_g^2 \ln \frac{4}{\delta}}{2\gamma_s^2 \epsilon^2 n_s} = \frac{\left(\max_{i \in [N]} |g_i|\right)^2 \ln \frac{4}{\delta}}{2\gamma_s^2 \epsilon^2 \left(2^{B-1} - 1\right)^2 n_s}. \tag{25}$$

Thus with a probability of at least $1 - \frac{\delta}{2}$,

$$\begin{aligned}\frac{\left|\widetilde{\mathcal{L}}_{s_1}^* - \mathcal{L}_{s_1}^*\right|}{\mathcal{G}_s^*} &\leq \frac{\max_{i \in [N]} |g_i|}{2\gamma_s^2 \epsilon \left(2^{B-1} - 1\right)} \sqrt{\frac{2 \ln \frac{4}{\delta}}{n_{s_1}}} + \frac{\left(\max_{i \in [N]} |g_i|\right)^2 \ln \frac{4}{\delta}}{2\gamma_s^2 \epsilon^2 \left(2^{B-1} - 1\right)^2 n_s} \\ &\leq \frac{\max_{i \in [N]} |g_i|}{\gamma_s^2 \epsilon \cdot 2^{B-1}} \sqrt{\frac{2 \ln \frac{4}{\delta}}{n_{s_1}}} + \frac{\left(\max_{i \in [N]} |g_i|\right)^2 \ln \frac{4}{\delta}}{2\gamma_s^2 \epsilon^2 n_s \cdot 4^{B-2}}\end{aligned} \tag{26}$$

Similarly, with a probability of at least $1 - \frac{\delta}{2}$

$$\frac{\left|\widetilde{\mathcal{L}}_{s_2}^* - \mathcal{L}_{s_2}^*\right|}{\mathcal{G}_s^*} \leq \frac{\max_{i \in [N]} |g_i|}{\gamma_s^2 \epsilon \cdot 2^{B-1}} \sqrt{\frac{2 \ln \frac{4}{\delta}}{n_{s_2}}} + \frac{\left(\max_{i \in [N]} |g_i|\right)^2 \ln \frac{4}{\delta}}{2\gamma_s^2 \epsilon^2 n_s \cdot 4^{B-2}} \tag{27}$$

By union bound, with a probability of at least $1 - \delta$

$$\begin{aligned}\frac{\left|\widetilde{\mathcal{G}}_{s \to s_1, s_2} - \mathcal{G}_{s \to s_1, s_2}\right|}{\mathcal{G}_s^*} &\leq \frac{\left|\widetilde{\mathcal{L}}_{s_1}^* - \mathcal{L}_{s_1}^*\right| + \left|\widetilde{\mathcal{L}}_{s_2}^* - \mathcal{L}_{s_2}^*\right|}{\mathcal{G}_s^*} \\ \leq \frac{\max_{i \in [N]} |g_i| \sqrt{2 \ln \frac{4}{\delta}}}{\gamma_s^2 \epsilon \cdot 2^{B-1}} &\left(\sqrt{\frac{1}{n_{s_1}}} + \sqrt{\frac{1}{n_{s_2}}}\right) + \frac{\left(\max_{i \in [N]} |g_i|\right)^2 \ln \frac{4}{\delta}}{\gamma_s^2 \epsilon^2 n_s \cdot 4^{B-2}}\end{aligned} \tag{28}$$

B.3 Loss Functions with Non-constant Hessians

Commonly used loss functions for binary-classification, multi-classification, and ranking have non-constant hessian values. Note that all these loss functions have non-negative hessian values. We analyze the error caused by quantization for these functions in this section. Denote $\overline{h}_s = \frac{\sum_{i \in s} h_i}{n_s}$ to be the average of hessian values in leaf s. We have the following theorem.

Theorem B.3.1 For loss functions with non-constant hessian values, if Assumption 5.2 holds for the subset $\mathcal{D}_s$ in leaf s for some $\gamma_s > 0$, then with stochastic rounding and leaf-value refitting, for any $\epsilon > 0$ and $\delta > 0$, at least one of the following conclusions holds:

1. With any split of leaf s and its descendants, the resultant weighted average (weighted by h_i) of absolute values of prediction values by the tree in current boosting iteration for data in $\mathcal{D}_s$ is no greater than ϵ.

2. For any split $s \to s_1, s_2$ of leaf s, if $n_{s_1} \geq \frac{8\delta_h^2 \ln 8/\delta}{\overline{h}_{s_1}^2}$ and $n_{s_2} \geq \frac{8\delta_h^2 \ln 8/\delta}{\overline{h}_{s_2}^2}$, i.e. $n_{s_1} \leq \frac{\left(\sum_{i \in s_1} h_i\right)^2}{8\delta_h^2 \ln 8/\delta}$ and $n_{s_2} \leq \frac{\left(\sum_{i \in s_2} h_i\right)^2}{8\delta_h^2 \ln 8/\delta}$, then with a probability of at least $1 - \delta$

$$\begin{aligned}\frac{\left|\widetilde{\mathcal{G}}_{s \to s_1, s_2} - \mathcal{G}_{s \to s_1, s_2}\right|}{\mathcal{G}_s^*} \leq & \frac{\delta_g \sqrt{2 \ln 8/\delta}}{\gamma_s^2 \epsilon} \left(\frac{1}{\overline{h}_{s_1} \sqrt{n_{s_1}}} + \frac{1}{\overline{h}_{s_2} \sqrt{n_{s_2}}}\right) \\ & + \frac{\overline{h}_s \delta_h \sqrt{2 \ln 8/\delta}}{2\gamma_s^2} \left(\frac{n_s}{\overline{h}_{s_1}^2 n_{s_1} \sqrt{n_{s_1}}} + \frac{n_s}{\overline{h}_{s_2}^2 n_{s_2} \sqrt{n_{s_2}}}\right) \\ & + \frac{\delta_g^2 \ln 8/\delta}{\gamma_s^2 \overline{h}_s \epsilon^2} \left(\frac{1}{\overline{h}_{s_1} n_s} + \frac{1}{\overline{h}_{s_2} n_s}\right)\end{aligned} \tag{29}$$

Proof of Theorem B.3.1 By leaf-wise weak learnability (Assumption 5.2), there exists a split $s \to s_L, s_R$ and $\gamma_s > 0$ for s s.t. for data in $\mathcal{D}_s$, with binary labels $c(\mathbf{x}_i) = \text{sign}(g_i)$ and weights $w_i = |g_i|$, the split results in a stump with weighted binary-classification error is $\frac{1}{2} - \gamma_s$. Similar to the case of loss functions with constant hessian, we define s_L^+, s_L^-, s_R^+ and s_R^-, and first derive a lower bound for $\mathcal{G}_s^*$,

$$\begin{aligned}\mathcal{G}_s^* \geq \mathcal{G}_{s \to s_L, s_R} &= \frac{\left(\sum_{i \in s_L} g_i\right)^2}{2\sum_{i \in s_L} h_i} + \frac{\left(\sum_{i \in s_R} g_i\right)^2}{2\sum_{i \in s_R} h_i} \\ &\geq \frac{\left(\left|\sum_{i \in s_L} g_i\right| + \left|\sum_{i \in s_R} g_i\right|\right)^2}{2\sum_{i \in s} h_i} \\ &= \frac{\left(\sum_{i \in s_L^+} |g_i| - \sum_{i \in s_L^-} |g_i| + \sum_{i \in s_R^+} |g_i| - \sum_{i \in s_R^-} |g_i|\right)^2}{2\sum_{i \in s} h_i} \\ &= \frac{\left(\sum_{i \in s_L^+ \cup s_R^+} |g_i| - \sum_{i \in s_L^- \cup s_R^-} |g_i|\right)^2}{2\sum_{i \in s} h_i} \\ &= \frac{\left(\left(\frac{1}{2} + \gamma_s\right) \sum_{i \in s} |g_i| - \left(\frac{1}{2} - \gamma_s\right) \sum_{i \in s} |g_i|\right)^2}{2\sum_{i \in s} h_i} \\ &= \frac{2\gamma_s^2 \left(\sum_{i \in s} |g_i|\right)^2}{\sum_{i \in s} h_i}\end{aligned} \tag{30}$$

Let $\xi_i = \delta_h \widetilde{h}_i - h_i$ and $\epsilon_i = \delta_g \widetilde{g}_i - g_i$, thus $|\xi_i| \leq \delta_h$, $|\epsilon_i| \leq \delta_g$, $\mathbb{E}[\xi_i] = 0$ and $\mathbb{E}[\epsilon_i] = 0$. We then bound the error of $\sum_{i \in s_1} h_i$,

$$\left|\sum_{i \in s_1} (h_i + \xi_i) - \sum_{i \in s_1} h_i\right| = \left|\sum_{i \in s_1} \xi_i\right|. \tag{31}$$

Note that ξ_i's and ϵ_i's are independent variables. Let $t'_{s_1} = \delta_h \sqrt{2n_{s_1} \ln \frac{8}{\delta}}$, then by Hoeffding's inequality,

$$P\left(\left|\sum_{i\in s_1} \xi_i\right| \geq t'_{s_1}\right) \leq 2\exp\left(-\frac{2{t'_{s_1}}^2}{n_{s_1}(2\delta_h)^2}\right) = \frac{\delta}{4} \tag{32}$$

Similarly, let $t''_{s_1} = \delta_g \sqrt{2n_{s_1} \ln \frac{8}{\delta}}$, then by Hoeffding's inequality we can bound the error of $\sum_{i\in s_1} g_i$,

$$P\left(\left|\sum_{i\in s_1} \epsilon_i\right| \geq t''_{s_1}\right) \leq 2\exp\left(-\frac{2{t''_{s_1}}^2}{n_{s_1}(2\delta_g)^2}\right) = \frac{\delta}{4} \tag{33}$$

If $\frac{\sum_{i\in s}|g_i|}{\sum_{i\in s} h_i} \leq \epsilon$, then suppose $s'_1, ..., s'_m$ are all descendant leaves of s, then the average (weighted by h_i's) prediction value for data in $\mathcal{D}_s$ in current tree is

$$\frac{\sum_{i=1}^m \sum_{i\in s'_i} h_i \left|w^*_{s'_i}\right|}{\sum_{i=1}^m \sum_{i\in s'_i} h_i} = \frac{\sum_{i=1}^m \left|\sum_{i\in s'_i} g_i\right|}{\sum_{i=1}^m \sum_{i\in s'_i} h_i} \leq \frac{\sum_{i\in s}|g_i|}{\sum_{i\in s} h_i} \leq \epsilon \tag{34}$$

which guarantees that the first conclusion holds.

If $\frac{\sum_{i\in s}|g_i|}{\sum_{i\in s} h_i} > \epsilon$, then we denote $\overline{h}_{s_1} = \frac{\sum_{i\in s_1} h_i}{n_{s_1}}$. If $n_{s_1} \geq \frac{8\delta_h^2 \ln 8/\delta}{\overline{h}_{s_1}^2}$, i.e. $n_{s_1} \leq \frac{\left(\sum_{i\in s_1} h_i\right)^2}{8\delta_h^2 \ln 8/\delta}$, then we have $\left|\sum_{i\in s_1} \xi_i\right| \leq \frac{\sum_i h_i}{2}$ with a probability of at least $1 - \frac{\delta}{4}$ by equation (32). Thus with a probability of at least $1 - \frac{\delta}{2}$,

$$\begin{aligned}
\left|\widetilde{\mathcal{L}}^*_{s_1} - \mathcal{L}^*_{s_1}\right| &= \frac{1}{2}\left|\frac{\left(\sum_{i\in s_1} g_i + \epsilon_i\right)^2}{\sum_{i\in s_1} h_i + \xi_i} - \frac{\left(\sum_{i\in s_1} g_i\right)^2}{\sum_{i\in s_1} h_i}\right| \\
&= \frac{1}{2}\left|\frac{\left(\sum_{i\in s_1} g_i + \epsilon_i\right)^2 \left(\sum_{i\in s_1} h_i\right) - \left(\sum_{i\in s_1} h_i + \xi_i\right)\left(\sum_{i\in s_1} g_i\right)^2}{\left(\sum_{i\in s_1} h_i + \xi_i\right)\left(\sum_{i\in s_1} h_i\right)}\right| \\
&\leq \frac{\left|\left(\sum_{i\in s_1} g_i + \epsilon_i\right)^2 \left(\sum_{i\in s_1} h_i\right) - \left(\sum_{i\in s_1} g_i\right)^2 \left(\sum_{i\in s_1} h_i\right)\right|}{2\left(\sum_{i\in s_1} h_i + \xi_i\right)\left(\sum_{i\in s_1} h_i\right)} \\
&+ \frac{\left|\left(\sum_{i\in s_1} g_i\right)^2 \left(\sum_{i\in s_1} h_i\right) - \left(\sum_{i\in s_1} g_i\right)^2 \left(\sum_{i\in s_1} h_i + \xi_i\right)\right|}{2\left(\sum_{i\in s_1} h_i + \xi_i\right)\left(\sum_{i\in s_1} h_i\right)} \\
&\leq \frac{\left|\sum_{i\in s_1} \epsilon_i\right|\left|\sum_{i\in s_1} g_i\right|}{\sum_{i\in s_1} h_i + \xi_i} + \frac{\left|\sum_{i\in s_1} \epsilon_i\right|^2}{2\left(\sum_{i\in s_1} h_i + \xi_i\right)} + \frac{\left(\sum_{i\in s_1} g_i\right)^2 \left|\sum_{i\in s_1} \xi_i\right|}{2\left(\sum_{i\in s_1} h_i + \xi_i\right)\left(\sum_{i\in s_1} h_i\right)} \\
&\leq \frac{2\left|\sum_{i\in s_1} \epsilon_i\right|\left|\sum_{i\in s_1} g_i\right|}{\sum_{i\in s_1} h_i} + \frac{\left|\sum_{i\in s_1} \epsilon_i\right|^2}{\sum_{i\in s_1} h_i} + \frac{\left(\sum_{i\in s_1} g_i\right)^2 \left|\sum_{i\in s_1} \xi_i\right|}{\left(\sum_{i\in s_1} h_i\right)^2}
\end{aligned} \tag{35}$$

Because

$$\frac{2\left|\sum_{i\in s_1} \epsilon_i\right|\left|\sum_{i\in s_1} g_i\right|}{\mathcal{G}^*_s \sum_{i\in s_1} h_i} \leq \frac{\left(\sum_{i\in s} h_i\right)\left|\sum_{i\in s_1} \epsilon_i\right|\left|\sum_{i\in s_1} g_i\right|}{\gamma_s^2 \left(\sum_{i\in s}|g_i|\right)^2 \sum_{i\in s_1} h_i} \leq \frac{\left|\sum_{i\in s_1} \epsilon_i\right|}{\gamma_s^2 \overline{h}_{s_1} \epsilon n_{s_1}} \leq \frac{\delta_g \sqrt{2\ln 8/\delta}}{\gamma_s^2 \overline{h}_{s_1} \epsilon \sqrt{n_{s_1}}}, \tag{36}$$

$$\frac{\left|\sum_{i\in s_1} \epsilon_i\right|^2}{\mathcal{G}^*_s \sum_{i\in s_1} h_i} \leq \frac{\left(\sum_{i\in s} h_i\right)\left|\sum_{i\in s_1} \epsilon_i\right|^2}{2\gamma_s^2 \left(\sum_{i\in s}|g_i|\right)^2 \sum_{i\in s_1} h_i} \leq \frac{\left|\sum_{i\in s_1} \epsilon_i\right|^2}{2\gamma_s^2 \overline{h}_s \overline{h}_{s_1} \epsilon^2 n_s n_{s_1}} \leq \frac{\delta_g^2 \ln 8/\delta}{\gamma_s^2 \overline{h}_s \overline{h}_{s_1} \epsilon^2 n_s}, \tag{37}$$

$$\frac{\left(\sum_{i\in s_1} g_i\right)^2 \left|\sum_{i\in s_1} \xi_i\right|}{\mathcal{G}^*_s \left(\sum_{i\in s_1} h_i\right)^2} \leq \frac{n_s \overline{h}_s \left|\sum_{i\in s_1} \xi_i\right|}{2\gamma_s^2 \left(\sum_{i\in s_1} h_i\right)^2} \leq \frac{n_s \overline{h}_s \delta_h \sqrt{2\ln 8/\delta}}{2\gamma_s^2 \overline{h}_{s_1}^2 n_{s_1} \sqrt{n_{s_1}}}. \tag{38}$$

Thus with probability at least $1-\frac{\delta}{2}$ we have

$$\frac{\left|\widetilde{\mathcal{L}}^*_{s_1}-\mathcal{L}^*_{s_1}\right|}{\mathcal{G}^*_s} \leq \frac{\sqrt{2\ln 8/\delta}}{\gamma_s^2}\left(\frac{\delta_g}{\overline{h}_{s_1}\epsilon\sqrt{n_{s_1}}}+\frac{\overline{h}_s\delta_h n_s}{2\overline{h}^2_{s_1}n_{s_1}\sqrt{n_{s_1}}}\right)+\frac{\delta_g^2\ln 8/\delta}{\gamma_s^2\overline{h}_s\overline{h}_{s_1}\epsilon^2 n_s}. \tag{39}$$

Similarly with probability at least $1-\frac{\delta}{2}$ we have

$$\frac{\left|\widetilde{\mathcal{L}}^*_{s_2}-\mathcal{L}^*_{s_2}\right|}{\mathcal{G}^*_s} \leq \frac{\sqrt{2\ln 8/\delta}}{\gamma_s^2}\left(\frac{\delta_g}{\overline{h}_{s_2}\epsilon\sqrt{n_{s_2}}}+\frac{\overline{h}_s\delta_h n_s}{2\overline{h}^2_{s_2}n_{s_2}\sqrt{n_{s_2}}}\right)+\frac{\delta_g^2\ln 8/\delta}{\gamma_s^2\overline{h}_s\overline{h}_{s_2}\epsilon^2 n_s}. \tag{40}$$

And finally, with probability at least $1-\delta$,

$$\begin{aligned}\frac{\left|\widetilde{\mathcal{G}}_{s\to s_1,s_2}-\mathcal{G}_{s\to s_1,s_2}\right|}{\mathcal{G}^*_s} \leq & \frac{\delta_g\sqrt{2\ln 8/\delta}}{\gamma_s^2\epsilon}\left(\frac{1}{\overline{h}_{s_1}\sqrt{n_{s_1}}}+\frac{1}{\overline{h}_{s_2}\sqrt{n_{s_2}}}\right)\\ &+\frac{\overline{h}_s\delta_h\sqrt{2\ln 8/\delta}}{2\gamma_s^2}\left(\frac{n_s}{\overline{h}^2_{s_1}n_{s_1}\sqrt{n_{s_1}}}+\frac{n_s}{\overline{h}^2_{s_2}n_{s_2}\sqrt{n_{s_2}}}\right)\\ &+\frac{\delta_g^2\ln 8/\delta}{\gamma_s^2\overline{h}_s\epsilon^2}\left(\frac{1}{\overline{h}_{s_1}n_s}+\frac{1}{\overline{h}_{s_2}n_s}\right).\end{aligned} \tag{41}$$

C Experiment Details

In this section, we provide more details about the results, experiment environments, and hyperparameter settings.

C.1 Variance of Quantized Training

In Table 2 the metric values are averaged over 5 random seeds. The seeds are used to generate random numbers for stochastic rounding. We omit the standard deviation in Table 2 due to limited space. Here we provide a full table with standard deviation listed in Table 4. Note that we report the metric on the best iteration in the test sets. As we can see the variance caused by stochastic rounding in quantization is small, and quantized training is quite stable with different random seeds.

C.2 Accuracy of Quantized Training on GPU

Table 5 shows the accuracy of quantized training on GPU, averaged over 5 random seeds for stochastic rounding. For the GPU version, we run up to 5 bits for gradient discretization. As we can see, for most datasets a comparable performance is achieved with quantized training. Note that we report the metric on the best iteration in the test sets.

C.3 Time for Histogram Construction

Table 6 shows the histogram construction time. The number of bits does not influence the histogram construction time. This indicates that more acceleration can be achieved for low-bitwidth gradients like 2-bit or 3-bit, with better hardware support for operations of low-bitwidth integers.

C.4 Experiment Environments

Table 7 lists the experiment environments used in this paper for standalone machines. For CPU clusters in distributed experiments, we use 16 nodes each with one Intel(R) Xeon(R) CPU E5-2673 v4 or Intel(R) Xeon(R) Platinum 8171M CPU. The nodes are connected by a network of bandwidth between $7 \sim 8$Gbps (tested with iperf[9]).

[9] https://iperf.fr/

Table 4: Accuracy with standard variance, w.r.t. different quantized bits (CPU version).

Bitwidth	Binary Classification					Regression	Ranking	
	Higgs↑	Epsilon↑	Kitsune↑	Criteo↑	Bosch↑	Year↓	Yahoo LTR↑	LETOR↑
32-bit	0.845694 ±.000162	0.950203 ±.000144	0.950561 ±.000638	0.803791 ±.000052	0.703101 ±.000544	8.956278 ±.004803	**0.793857** ±.000198	0.524265 ±.000622
2-bit SR_{refit}	0.845587 ±.000042	0.949472 ±.000166	0.952703 ±.000729	0.803293 ±.000091	0.700322 ±.001102	8.953388 ±.012679	0.788579 ±.000357	0.519268 ±.000628
3-bit SR_{refit}	0.845725 ±.000193	0.949884 ±.000095	0.951309 ±.001352	0.803768 ±.000050	0.702756 ±.001704	**8.937374** ±.007487	0.791077 ±.000894	0.522220 ±.000721
4-bit SR_{refit}	0.845507 ±.000127	0.950049 ±.000072	0.950911 ±.001557	0.803783 ±.000073	0.703315 ±.000712	8.942898 ±.008327	0.792664 ±.000467	0.523796 ±.000514
5-bit SR_{refit}	0.845706 ±.000171	0.950298 ±.000108	0.949229 ±.001964	0.803766 ±.000053	0.702971 ±.001020	8.948542 ±.003732	0.793166 ±.000487	**0.524673** ±.000413
2-bit $SR_{no\ refit}$	**0.846713** ±.000184	0.944509 ±.000174	0.952974 ±.001024	0.803750 ±.000068	0.700900 ±.001108	9.112302 ±.014516	0.764862 ±.000858	0.486193 ±.001789
3-bit $SR_{no\ refit}$	0.846040 ±.000178	0.949593 ±.000119	0.951385 ±.001158	**0.803922** ±.000064	0.702501 ±.000751	8.990034 ±.009847	0.780041 ±.000618	0.507689 ±.001126
4-bit $SR_{no\ refit}$	0.845816 ±.000304	0.950127 ±.000172	0.951197 ±.001067	0.803812 ±.000074	0.703327 ±.000655	8.955256 ±.003074	0.787575 ±.001173	0.515767 ±.000448
5-bit $SR_{no\ refit}$	0.845842 ±.000119	0.950275 ±.000234	0.949794 ±.002275	0.803790 ±.000096	0.703226 ±.001484	8.952768 ±.009403	0.791631 ±.000590	0.520900 ±.001087
2-bit RN_{refit}	0.795991 ±.000582	0.889149 ±.000856	0.962201 ±.000820	0.779906 ±.000323	0.686617 ±.000405	9.429014 ±.017197	0.765103 ±.000918	0.454894 ±.005287
3-bit RN_{refit}	0.830506 ±.000495	0.944329 ±.000319	**0.966606** ±.001074	0.782732 ±.000210	0.688899 ±.000285	9.062854 ±.014744	0.772364 ±.000822	0.476726 ±.001458
4-bit RN_{refit}	0.840747 ±.000241	0.949946 ±.000207	0.961938 ±.001970	0.795803 ±.000099	0.691469 ±.000432	8.968694 ±.005092	0.777347 ±.000969	0.487256 ±.003072
5-bit RN_{refit}	0.843820 ±.000073	**0.950457** ±.000071	0.962427 ±.001150	0.802438 ±.000083	0.698954 ±.000541	8.952418 ±.003649	0.784333 ±.000612	0.494951 ±.001611
2-bit $RN_{no\ refit}$	0.836683 ±.000468	0.925220 ±.001545	0.946016 ±.005072	0.768338 ±.000202	0.704445 ±.002635	10.685840 ±.001819	0.632058 ±.005683	0.203732 ±.005507
3-bit $RN_{no\ refit}$	0.843482 ±.000306	0.946850 ±.000399	0.940961 ±.006586	0.791709 ±.000379	**0.708724** ±.000945	9.377560 ±.042545	0.732487 ±.001121	0.350127 ±.004163
4-bit $RN_{no\ refit}$	0.845788 ±.000176	0.949676 ±.000126	0.949228 ±.002973	0.802689 ±.000096	0.703718 ±.000580	8.969828 ±.005646	0.765432 ±.000426	0.437317 ±.001514
5-bit $RN_{no\ refit}$	0.845765 ±.000248	0.950307 ±.000150	0.952420 ±.003838	0.803645 ±.000102	0.698419 ±.000502	8.965400 ±.002101	0.782608 ±.000514	0.485752 ±.001105

Table 5: Accuracy with standard variance, w.r.t. different quantized bits (GPU version, SR_{refit} mode).

Bitwidth	Binary-Class					Regression	Ranking	
	Higgs↑	Epsilon↑	Kitsune↑	Criteo↑	Bosch↑	Year↓	Yahoo LTR↑	LETOR↑
32-bit	0.845729 ±.000081	0.950233 ±.000119	0.955709 ±.001021	0.803792 ±.000065	0.702893 ±.000152	8.956202 ±.004722	0.795476 ±.000387	0.526287 ±.000337
2-bit	0.846582 ±.000159	0.945205 ±.000255	0.952898 ±.001554	0.803594 ±.000077	0.701775 ±.001065	9.107948 ±.007273	0.769852 ±.000425	0.492308 ±.001299
3-bit	0.845877 ±.000255	0.949494 ±.000277	0.951672 ±.002186	0.803847 ±.000059	0.703032 ±.000939	8.980230 ±.006827	0.784374 ±.000371	0.512684 ±.000529
4-bit	0.845872 ±.000199	0.950176 ±.000066	0.951918 ±.001138	0.803799 ±.000089	0.703067 ±.000952	8.962148 ±.016629	0.791226 ±.000566	0.519651 ±.001012
5-bit	0.845849 ±.000238	0.950177 ±.000174	0.950538 ±.000354	0.803827 ±.000095	0.703823 ±.001013	8.953900 ±.004574	0.793799 ±.000479	0.524211 ±.000409

C.5 Hyperparameter Settings

For all accuracy and training time evaluations in this paper, we use the hyperparameters of LightGBM listed in Table 8, except for the Bosch dataset. For the Bosch dataset, we use learing_rate 0.015 and keep other hyperparameters the same as Table 8. For training time comparison with XGBoost and CatBoost, we use the hyperparameters listed in Table 9 and 10, except for Bosch. For the Bosch dataset, we use learing_rate 0.015 for CatBoost and eta 0.015 for XGBoost, max_leaves 45 for XGBoost, and keep other hyperparameters the same as in the tables. We found that the post-pruning

Table 6: Time for histogram construction with different number of bits (seconds).

	Algorithm	Bosch	Criteo	Epsilon	Higgs	Kitsune	Year	Yahoo LTR	LETOR
GPU Histogram time	LightGBM+	17	70	46	11	54	9	11	17
	LightGBM+ 2-bit	**8**	**21**	**11**	**4**	**16**	**4**	**8**	**10**
	LightGBM+ 3-bit	**8**	**21**	12	**4**	**16**	**4**	**8**	**10**
	LightGBM+ 4-bit	**8**	**21**	12	**4**	**16**	**4**	**8**	**10**
	LightGBM+ 5-bit	**8**	**21**	13	**4**	**16**	**4**	**8**	**10**
CPU Histogram time	LightGBM	98	629	737	94	339	12	108	109
	LightGBM 2-bit	**72**	458	708	68	203	10	**67**	**68**
	LightGBM 3-bit	75	437	**676**	**62**	180	10	73	69
	LightGBM 4-bit	76	**426**	680	65	**177**	9	74	73
	LightGBM 5-bit	73	399	681	63	206	**8**	78	72

Table 7: Experiment Environments

CPU	2 x Intel(R) Xeon(R) CPU E5-2673 v4
GPU	1 x NVIDIA V100
OS	Ubuntu 18.04

Table 8: Hyperparameters of LightGBM

boosting_type	gbdt
learning_rate	0.1
min_child_weight	100
num_leaves	255
max_bin	255
num_iterations	500
num_threads	16

strategy of XGBoost slows down the training much with max_leaves 255 on Bosch. Thus, we adjust the max_leaves to 45 which is close to the tree size after the pruning, for faster training speed. The hyperparameters are chosen so that all these algorithms have similar tree sizes for a fair comparison of training time.

The git commit used for CatBoost is 35552cf8057447262eedd9671f66fd715af34946. And for XGBoost it is fe4ce920b250d39133a7f6b1128f80da0d4018c6. For LightGBM, we use the version provided in our Github link https://github.com/Quantized-GBDT/Quantized-GBDT.

C.6 Data Split and Preprocessing

For most datasets (Higgs, Epsilon, Yahoo, LETOR, Year, Bosch) we use the convention in previous works or the default split [32, 23, 3], without additional preprocessing. For Criteo, we encode the categorical features in the original dataset with target and count encoding. We use the `train.txt` file of the Kaggle version of the Criteo dataset, with the first $41,256,555$ rows as the training set and the last $4,584,061$ rows as the test set. For Kitsune, we select the first 80% packets in each attack method to form the training set, and the final 20% packets to form the test set. The datasets can be freely downloaded from https://pretrain.blob.core.windows.net/quantized-gbdt/dataset.zip.

D Discussion on Loss Functions with Non-constant Hessians

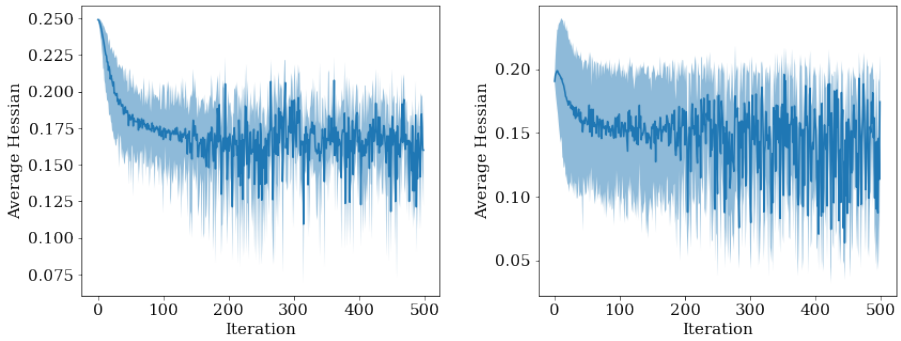
Appendix B.3 provides the theoretical analysis and proof for the error caused by quantization for loss functions with non-constant hessians. The assumption is a little bit stronger than constant hessian loss functions in that we are expecting the average hessian values per leaf won't be too small, so that $n_{s_1} \geq \frac{8\delta_h^2 \ln 8/\delta}{\overline{h}_{s_1}^2}$ and $n_{s_2} \geq \frac{8\delta_h^2 \ln 8/\delta}{\overline{h}_{s_2}^2}$ hold. Figure 7 shows the average hessian values in each iteration with 3-bit gradients. We first calculate the average hessian values for all leaves in each iteration. Then we plot the mean of the average hessian values over the leaves in each iteration in solid blue curves, with the shadow area indicating the range between 10% and 90% percentiles over the leaves in each iteration. For most leaves, the average hessian values are not too small. And it is easy to meet the condition $n_s \geq \frac{8\delta_h^2 \ln 8/\delta}{\overline{h}_s^2}$ with enough training data. For example, suppose $\overline{h}_s = 0.01$, then for binary classification, with 3-bit gradients, $\delta_h = \frac{0.25}{6} = \frac{1}{24}$. Let $\delta = 0.01$, then $\frac{8\delta_h^2 \ln 8/\delta}{\overline{h}_s^2} \approx 928$.

Table 9: Hyperparameters of XGBoost

tree_method	hist/gpu_hist
eta	0.1
max_depth	0
max_leaves	255
num_round	500
min_child_weight	100
nthread	16
gamma	0
lambda	0
alpha	0

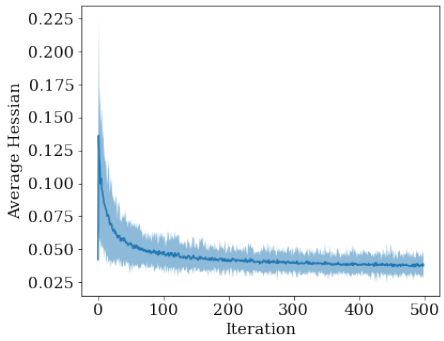
Table 10: Hyperparameters of CatBoost

thread_count	16
border_count	255
iterations	500
learning_rate	0.1
grow_policy	Lossguide
boosting_type	Plain
max_leaves	255
depth	256
min_data_in_leaf	400

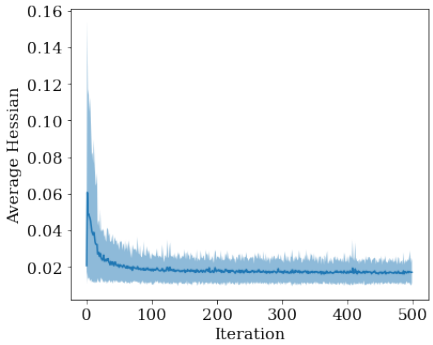


(a) Higgs

(b) Criteo



(c) Yahoo



(d) LETOR

Figure 7: Average Hessian Values by Iteration with 3-Bit Gradients

In addition, in the first conclusion of Theorem B.3.1 we consider weighted prediction values by h_i. Since with second-order approximation of the loss function, h_i influences how much a training sample contributes to the approximated loss by second-order Taylor expansion [5]. Thus, considering the weighted prediction values by h_i is meaningful.

Finally, the upper bound in Theorem B.3.1 requires a balanced split to be small. In other words, the data sizes in child nodes n_{s_1}, n_{s_2} shouldn't be significantly smaller than that in parent node n_s, so that the terms $\frac{n_s}{n_{s_1}\sqrt{n_{s_1}}}$ and $\frac{n_s}{n_{s_2}\sqrt{n_{s_2}}}$ can be bounded by a small value.

E New CUDA Framework of LightGBM

We implement a new CUDA version for LightGBM. Previous GPU versions of LightGBM only run histogram construction on GPUs. Our new implementation performs the whole training process including boosting (calculation of gradients and hessians) and tree learning on GPUs. We denote this new GPU version of LightGBM as LightGBM+ in our paper.