Quantized Training of Gradient Boosting Decision Trees

Yu Shi[1*] **Guolin Ke**[2*†],
Zhuoming Chen[3‡] **Shuxin Zheng**[1] **Tie-Yan Liu**[1]
[1]Microsoft Research [2]DP Technology [3]Tsinghua University
yushi2@microsoft.com, kegl@dp.tech
chen-zm19@mails.tsinghua.edu.cn, {shuxin.zheng, tyliu}@microsoft.com

Abstract

Recent years have witnessed significant success in Gradient Boosting Decision Trees (GBDT) for a wide range of machine learning applications. Generally, a consensus about GBDT's training algorithms is gradients and statistics are computed based on high-precision floating points. In this paper, we investigate an essentially important question which has been largely ignored by the previous literature - *how many bits are needed for representing gradients in training GBDT?* To solve this mystery, we propose to quantize all the high-precision gradients in a very simple yet effective way in the GBDT's training algorithm. Surprisingly, both our theoretical analysis and empirical studies show that the necessary precisions of gradients without hurting any performance can be quite low, e.g., 2 or 3 bits. With low-precision gradients, most arithmetic operations in GBDT training can be replaced by integer operations of 8, 16, or 32 bits. Promisingly, these findings may pave the way for much more efficient training of GBDT from several aspects: (1) speeding up the computation of gradient statistics in histograms; (2) compressing the communication cost of high-precision statistical information during distributed training; (3) the inspiration of utilization and development of hardware architectures which well support low-precision computation for GBDT training. Benchmarked on CPU, GPU, and distributed clusters, we observe up to $2\times$ speedup of our simple quantization strategy compared with SOTA GBDT systems on extensive datasets, demonstrating the effectiveness and potential of the low-precision training of GBDT. The code will be released to the official repository of LightGBM[1] .

1 Introduction

Gradient Boosting Decision Trees (GBDT) is a powerful machine learning algorithm. Despite the success of deep learning in recent years, GBDT is one of the best off-the-shelf choices of machine learning algorithms in many real-world tasks, including online advertising[28], search ranking[22, 12], finance prediction[24], etc. Along with carefully designed tools, including XGBoost[5], LightGBM[15], and CatBoost[20], GBDT has shown outstanding performance in various data science competitions[31] and industrial applications[27].

Despite the success of GBDT, we found there is room for GBDT algorithms to fully exploit modern computation resources. First, the training of GBDT requires arithmetic operations of high-precision FP (Floating Point) numbers, which hinders the usage of low-precision computation resources.

[*]Corresponding authors.
[†]Work primarily done at Microsoft Research
[‡]Work done during internship at Microsoft Research
[1]https://github.com/Microsoft/LightGBM

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Low-precision training has become a standard technique to significantly accelerate the training of neural networks[17]. The design of new processors for machine learning tends to encourage high throughput for low-precision computation. This results in a gap between GBDT training algorithms and the hardware architectures. Second, distributed training of GBDT is required for large datasets. The communication cost for distributed training, however, is huge due to high-precision statistic information passing among machines. The cost is especially high with a large number of features in the dataset, which hurts the scalability of GBDT in distributed systems.

In this paper, we propose a low-precision training algorithm for GBDT, based on stochastic gradient quantization. The main computation cost of training GBDT is the arithmetic operations over gradients. Before training each decision tree, we quantize the gradients into very low-precision (e.g., 2-bit or 3-bit) integers. Thus, we replace most of the FP arithmetic operations with integer arithmetic operations, which reduces computation time. In addition, low-bitwidth gradients result in a smaller memory footprint and better cache utilization. Techniques are proposed to preserve the accuracy of the model, including stochastic rounding in gradient quantization, and leaf value refitting with original gradient values. We show both empirically and theoretically that quantized training of GBDT with low-bitwidth gradients is almost lossless in terms of model accuracy. Thus, we empower GBDT to utilize low-precision computation resources.

Distributed training of GBDT relies on the synchronization of gradient statistics. The gradient statistics are summarized into a histogram for each feature. With quantized gradients, only integer histogram values are needed in our algorithm. The communication among hosts or GPUs becomes sending and reducing of these integer values. And the size of these histograms is at least half of the original histograms with FP values. So quantized training of GBDT has an advantage in communication cost by nature and improves the scalability of GBDT on distributed systems.

We implement our system on both CPUs and GPUs. The results show that our methods accelerate the training of GBDT in various settings, including a single process on CPUs, a single process on a GPU, and distributed training over CPUs. This validates that our algorithm is general for different types of computation resources. Besides acceleration on existing hardware architectures, our algorithm also lay the foundation for GBDT to exploit new hardware architectures with more flexible support for low-precision operations in the future. Based on LightGBM, we implement a quantized GBDT training system. With quantized training, we achieve up to $2\times$ speedup compared with SOTA GBDT tools on both CPUs and GPUs. Experiments also show that quantized GBDT training scales better in distributed systems.

Our algorithm also verifies huge redundancy in the information contained by gradient values in GBDT. Moreover, we reveal both theoretically and empirically how such redundancy can be properly reduced, to keep good model performance and training efficiency. And we show how to implement the quantized training of GBDT efficiently across different computation resources. We believe that these interesting findings will inspire new understandings and improvements for GBDT in the future.

2 Related Work

Trials to use low-precision training in deep neural networks (NN) are abundant. Using lower bitwidth numbers in NN training can reduce memory access and accelerate training [11, 13, 30, 17]. In addition to efficiency, the effectiveness of quantized NN training has also been discussed[7, 4]. Besides quantization during the forward and backward calculations, quantization of gradients has also been applied in distributed training of neural networks[25] to reduce communication cost. However, the possibility of low-precision training for GBDT is rarely discussed in existing literature. RatBoost[19] is a boosting algorithm based on re-weighting samples. In RatBoost[19], quantization is applied to the weights of training samples during boosting, while the application of quantization in GBDT is not discussed. Recently gradient quantization is applied in federated training of GBDT[6] to transform FP gradients into large integers (e.g., 53-bit integers) for the sake of encryption. But quantization towards very small number of bits is not explored. Accelerating GBDT training with quantized gradients was exploited by BitBoost[8]. And BitBoost shows empirically that it is possible to achieve comparable accuracy with quantized gradients of low bit-width. However, BitBoost quantizes gradients in a deterministic way. In this paper, we provide a theoretical guarantee of quantized training based on stochastic quantization. And we show that stochastic quantization is crucial to preserve good model accuracy. Besides, BitBoost focuses on accelerating GBDT with bit operations on a single CPU core, and the implementation is a single-thread version. In this paper, we

discuss important techniques in system implementation that enables significant acceleration across different computation platforms, including multi-core CPUs, GPUs, and distributed cluster.

Distributed training of GBDT is necessary for large-scale datasets. Training data are partitioned either by rows (data samples) or by columns (features) across different machines. We call the former strategy data-distributed training and the later feature-distributed training. The communication cost for feature-distributed training grows linearly with the number of training samples[1, 16]. On the other hand, the communication cost for data-distributed training is proportional with the number of features[1, 16]. The two strategies can be combined with each machine getting only part of features and samples[9]. Communication complexity of these strategies are analysis in detail[26], with empirical studies[9]. Since GBDT often summarize the gradient statistics into a histogram for each feature, the cost for data-distributed training is reduced with the more concise histogram information[5]. Thus, data-distributed training is more favorable with large-scale datasets. With a large number of features, feature-distributed or the combination of two strategies can be faster[9].

Based on these distributed training strategies, some efforts have been made to reduce the communication cost. Meng et al[16] proposed Parallel-Voting Tree (PV-Tree) to further reduce the cost of data-distributed training. However, PV-Tree guarantees good performance only when different parts of the partitioned dataset have similar statistical distributions, which can require expensive random shuffling over the whole large dataset. Moreover, quantized training can be applied along with PV-Tree. This is because PV-Tree reduces communication cost by reducing the number of features to synchronize the statistics in the histograms. Since quantized training can reduce the size of histograms, it brings a general benefit for both PV-Tree and ordinary data-distributed training. DimBoost[14] keeps the number of communicated histograms unchanged, but reduces the size of each histogram by compressing the histograms instead. Histograms represented by floating-point numbers in DimBoost are compressed into low-precision values (8-bit integers) before sending, and decoding into high-precision values. Note that in DimBoost, only the communication message for distributed training is in low-precision, and the compression is lossy. In addition, reduction of the histograms still requires floating-point additions. In comparison, we quantized most parts of the training process of GBDT with low-precision integers, and only need to maintain histograms with integer values.

In this paper, we compare our algorithm with SOTA efficient implementations of GBDT, including XGBoost[5], LightGBM[15] and CatBoost[20]. They support training on both CPUs and GPUs[29], and distributed training. However, none of these support low-precision training. We show that with quantized gradients, GBDT can be trained much faster compared with these SOTA tools in both single process and distributed settings, on both CPUs and GPUs. We implement quantized training based on LightGBM. But our method is general and can be adopted by all these tools.

3 Preliminaries of GBDT Algorithms

Gradient Boosting Decision Trees (GBDT) is an ensemble learning algorithm that trains decision trees as base components iteratively. In each iteration, for each sample, the gradients and hessians of the loss function w.r.t. current prediction values are calculated. Then a decision tree is trained to fit the negative gradients. Formally, in iteration $k+1$, let $\hat{y}_i^k$ be the current prediction value of instance i. And g_i and h_i are the first-order and second-order derivatives of loss function l:

$$g_i = \frac{\partial l(\hat{y}_i^k, y_i)}{\partial \hat{y}_i^k}, h_i = \frac{\partial^2 l(\hat{y}_i^k, y_i)}{\left(\partial \hat{y}_i^k\right)^2} \tag{1}$$

We call g_i's as gradients and h_i's as hessians. For a leaf s, denote I_s as the set of data indices in the leaf. Let $G_s = \sum_{i \in I_s} g_i$ and $H_s = \sum_{i \in I_s} h_i$ be the summations of g_i's and h_i's over instances on leaf s. With the tree structure being fixed in iteration $k+1$, the training loss can be approximated by second-order Taylor series:

$$\mathcal{L}_{k+1} \approx \mathcal{C} + \sum_s \left(\frac{1}{2} H_s w_s^2 + G_s w_s \right) \tag{2}$$

where $\mathcal{C}$ is a constant and w_s is the prediction value of leaf s. By minimizing the approximated loss, we obtain the optimal value for leaf s and the minimal loss contributed by data in I_s:

$$w_s^* = -\frac{G_s}{H_s}, \quad \mathcal{L}_s^* = -\frac{1}{2} \cdot \frac{G_s^2}{H_s}. \tag{3}$$

Finding the optimal tree structure is difficult. Thus the tree is trained by greedily and iteratively splitting a leaf into two child leaves, starting from a single root leaf. When splitting leaf s into two children s_1 and s_2 in tree T_{k+1}, we can calculate the reduction in the approximated loss

$$\Delta\mathcal{L}_{s\to s_1,s_2} = \mathcal{L}_s^* - \mathcal{L}_{s_1}^* - \mathcal{L}_{s_2}^* = \frac{G_{s_1}^2}{2H_{s_1}} + \frac{G_{s_2}^2}{2H_{s_2}} - \frac{G_s^2}{2H_s}. \tag{4}$$

To find the best split condition for leaf s, all candidate splits of all features should be enumerated and the one with the largest loss reduction is chosen.

To accelerate the best split finding process, an algorithm based on histograms is used by most state-of-the-art GBDT toolkits. The basic idea of histogram-based GBDT is to divide the values of a feature into bins, each bin corresponds to a range of the feature values. Then we construct histograms with the bins. Each bin will record the summation of gradients and hessians of data in that bin. Only the boundaries of the bins (ranges) will be considered as candidate split thresholds. Algorithm 1 describes the process of histogram construction. Bin data matrix $\mathbf{data}$ records for each feature and instance the index of the bin in the histogram, i.e. which range the feature value falls in. The g_i's and h_i's are accumulated in the corresponding bins in histograms of the features. Since the split

Algorithm 1 Histogram Construction for Leaf s

Input: Gradients $\{g_1, ..., g_N\}$, Hessians $\{h_1, ..., h_N\}$
Input: Bin data $\mathbf{data}[N][J]$, Data indices in leaf s I_s
Output: Histogram $\mathbf{hist}_s$
for $i \in I_s, j \in \{1...J\}$ **do**
  $bin \leftarrow \mathbf{data}[i][j]$
  $\mathbf{hist}_s[j][bin].g \leftarrow \mathbf{hist}_s[j][bin].g + g_i$
  $\mathbf{hist}_s[j][bin].h \leftarrow \mathbf{hist}_s[j][bin].h + h_i$
end for

criterion (4) only depends on the summation of gradient statistics, we can easily obtain the optimal split threshold by iterating over bins in the histogram. Traditionally 32-bit floating point numbers are used for g_i's and h_i's. And the accumulations in the histograms usually require 32-bit or 64-bit FP numbers. In next section, we show how to quantizes g_i's and h_i's into low-bitwidth integers so that most of the arithmetic operations can be replaced with integer operations with fewer number of bits, which substantially saves computational cost.

4 Quantized Training of GBDT

We describe our quantized training algorithm for GBDT. First we show the overall framework. Then we propose the two critical techniques to preserve the accuracy of quantized traninig, including stochastic gradient discretization and leaf value refitting.

4.1 Framework for Quantized Training

We first quantize g_i and h_i into low-bitwidth integers $\widetilde{g}_i$ and $\widetilde{h}_i$. We divide the ranges of g_i's and h_i's of all instances into intervals of equal length. To use B-bit ($B \geq 2$) integer gradients, we use $2^B - 2$ intervals. Each end of the intervals corresponds to a integer value, resulting in $2^B - 1$ integer values in total. Since the first-order derivatives g_i's can take both positive and negative values, half of the intervals will be allocated for the negatives values, and the other half for the positive values. For the second-order derivatives h_i's, almost all commonly used loss functions of GBDT have non-negative values. We assume that $h_i \geq 0$ in subsequent discussions. Thus the interval lengths are

$$\delta_g = \frac{\max_{i\in[N]} |g_i|}{2^{B-1} - 1}, \quad \delta_h = \frac{\max_{i\in[N]} h_i}{2^B - 2} \tag{5}$$

for g_i's and h_i's, respectively. Then we can calculate the low-bitwidth gradients

$$\widetilde{g}_i = Round\left(\frac{g_i}{\delta_g}\right), \quad \widetilde{h}_i = Round\left(\frac{h_i}{\delta_h}\right). \tag{6}$$

The function $Round$ rounds a floating point number into an integer number. Note that in the case where h_i's are constant, there's no need to quantize h_i. We left the detailed rounding strategy in Section 4.2. We replace g_i's and h_i's in Algorithm 1 with $\widetilde{g}_i$'s and $\widetilde{h}_i$'s. Then addition operations

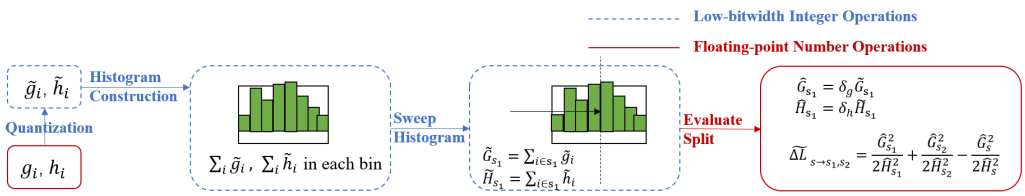


Figure 1: Workflow of the quantized GBDT training.

of the original gradients can be directly replaced by integer additions. Concretely, the statistics g and h in the histogram bins will become integers. Thus accumulating gradients into histogram bins in Algorithm 1 only requires integer additions. In equation (4), G_{s_1}, H_{s_1}, G_{s_2} and H_{s_2} will replaced by their integer counterparts $\widetilde{G}_{s_1}$, $\widetilde{H}_{s_1}$, $\widetilde{G}_{s_2}$, and $\widetilde{H}_{s_2}$. We found that 2 to 4 bits for quantized gradients is enough for good accuracy. As we will discuss in Section 6.1, in most cases, 8-bit or 16-bit integers will be enough to accumulate such low-bitwidth gradients in histograms. Thus most of the operations are done with low-bitwidth integers. We only need floating point operations when calculating the split gain. Specifically, the split gain is estimated as

$$\Delta\widetilde{\mathcal{L}}_{s \to s_1, s_2} = \frac{\left(\widetilde{G}_{s_1}\delta_g\right)^2}{2\widetilde{H}_{s_1}\delta_h} + \frac{\left(\widetilde{G}_{s_2}\delta_g\right)^2}{2\widetilde{H}_{s_2}\delta_h} - \frac{\left(\widetilde{G}_{s}\delta_g\right)^2}{2\widetilde{H}_{s}\delta_h}, \tag{7}$$

where the scale of the gradient statistics are recovered by multiplying δ_g and δ_h. Figure 1 summarizes the workflow for quantized GBDT training.

4.2 Rounding Strategies and Leaf-Value Refitting

We find that the *Round* strategy in equation (6) has a significant impact on the accuracy of quantized training algorithm. Rounding to the nearest integer number seems to be a reasonable choice. However, we found the accuracy drops severely with this strategy, especially with small number of gradient bits. Instead, we adopt a stochastic rounding strategy. $\widetilde{g}_i$ randomly takes values $\lfloor g_i/\delta_g \rfloor$ or $\lceil g_i/\delta_g \rceil$ such that $\mathbb{E}[\widetilde{g}_i] = g_i/\delta_g$. The formal definitions of the two rounding strategies are

$$\mathrm{RN}\left(x\right) = \begin{cases} \lfloor x \rfloor, & x < \lfloor x \rfloor + \frac{1}{2} \\ \lceil x \rceil, & x \geq \lfloor x \rfloor + \frac{1}{2} \end{cases}, \quad \mathrm{SR}\left(x\right) = \begin{cases} \lfloor x \rfloor, & \text{w.p.} \quad \lceil x \rceil - x \\ \lceil x \rceil, & \text{w.p.} \quad x - \lfloor x \rfloor \end{cases} \tag{8}$$

where we use RN for round-to-nearest and SR for stochastic rounding. The key insight is that the split gain is calculated with the summation of gradients. And stochastic rounding would provide an unbiased estimation for the summations, i.e. $\mathbb{E}[\widetilde{G}\delta_g] = G$, and $\mathbb{E}[\widetilde{H}\delta_h] = H$ in equation (7). The importance of stochastic rounding is also recognized in other scenarios for quantization, including NN quantization training[11] and histogram compression in DimBoost[14]. With stochastic rounding, we provide a theorem in Section 5 which ensures that the error in split gain estimation caused by quantized gradients is bounded by a small value with high probability.

With quantized gradients, the optimal leaf value in equation (3) becomes $\widetilde{w}_s^* = -\frac{\widetilde{G}_s\delta_g}{\widetilde{H}_s\delta_h}$. In most cases, $\widetilde{w}_s^*$ is enough for good results. But we found that for some loss functions, especially ranking objectives, refitting the leaf values with the original gradient values after the tree has stopped growing is useful to improve the accuracy of quantized training. BitBoost[8] also recaculates leaf values with real gradients after tree growing. It is cheap to calculate w_s^* in equation (3) given a fixed tree structure, since only a single pass over g_i's and h_i's is needed to sum them up into different leaves.

5 Theoretical Analysis

We prove that the difference in the gain of a split due to gradient quantization is bounded by a small value, with enough training data. We consider loss functions with constant second-order derivatives (i.e., the h_i's are constant). The square error loss function $l(\hat{y}, y) = \frac{1}{2}\left(\hat{y} - y\right)^2$ for regression tasks falls in this category. The analysis is also fit for GBDT without second-order Taylor approximation. We leave the case for loss functions with non-constant second-order derivatives in Appendix B. We focus on the discussion of the conclusion in this section and leave details of the proof in Appendix B.

Our theoretical analysis is based on a more specific form of weak-learnability assumption in boosting. Specifically, the assumption considers only two-leaf decision trees (a.k.a. stumps) as weak learners.

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 $c(\mathbf{x}_i) \in \{-1, 1\}$ and weights $\{w_i\}_{i=1}^N$, where $w_i \geq 0$ 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 weak-learnable by stumps w.r.t. c and $\{w_i\}_{i=1}^N$.

The weak-learnability of stump states that given a binary-class concept c and weights for dataset $\mathcal{D}$, there exists a stump which can produce slightly better classification accuracy than a random guess. Our analysis is based on the following assumption over the data in a single leaf s.

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 weak-learnable by stumps, w.r.t. concept $c(\mathbf{x}_i) = \text{sign}(g_i)$ and weights $w_i = |g_i|$, where $i \in I_s$.

An equivalence of the assumption is: $\mathcal{D}$ is γ_s-empirically weak-learnable by stumps, w.r.t. concept $c(\mathbf{x}_i) = \text{sign}(g_i)$, and weights $w_i = |(g_i)|$ for $i \in I_s$ and $w_i = 0$ for $i \in [N]/I_s$. Similar assumption has been adopted in previous theoretical analysis for gradient boosting[10]. The difference is that here we restrict the weak learner to be a stump. In Appendix B, we show that for any leaf s with positive split gain, there exists such a $\gamma_s > 0$. As we will see in the experiments, for most leaves during the training of GBDT, γ_s won't be too small.

Let $\mathcal{G}_{s \to s_1, s_2} = -\left(\mathcal{L}^*_{s_1} + \mathcal{L}^*_{s_2}\right) \geq 0$, where $\mathcal{L}^*_{s_1}$ and $\mathcal{L}^*_{s_2}$ are defined in equation (3). And let $\widetilde{\mathcal{G}}_{s \to s_1, s_2} = -\left(\widetilde{\mathcal{L}}^*_{s_1} + \widetilde{\mathcal{L}}^*_{s_2}\right) \geq 0$ be the estimated version of $\mathcal{G}_{s \to s_1, s_2}$ after gradient quantization. And let $\mathcal{G}^*_s = \mathcal{G}_{s \to s^*_1, s^*_2}$ be the value for the optimal split $s \to s^*_1, s^*_2$ in leaf s. Note that on leaf s, the optimal split is chosen only according to $\mathcal{G}_{s \to s_1, s_2}$, since $\mathcal{L}^*_s$ is a constant for different splits of leaf s. Then based on Assumption 5.2, we have the following theorem.

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 on leaf s and its descendents, the resultant average change 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 $s \to s_1, s_2$ on leaf s, with probability at least $1 - \delta$,

$$\frac{\left|\widetilde{\mathcal{G}}_{s \to s_1, s_2} - \mathcal{G}_{s \to s_1, s_2}\right|}{\mathcal{G}^*_s} \leq \frac{\max_{i \in [N]} |g_i| \sqrt{2 \ln \frac{4}{\delta}}}{\gamma_s^2 \epsilon \cdot 2^{B-1}} \left(\sqrt{\frac{1}{n_{s_1}}} + \sqrt{\frac{1}{n_{s_2}}}\right) + \frac{\left(\max_{i \in [N]} |g_i|\right)^2 \ln \frac{4}{\delta}}{\gamma_s^2 \epsilon^2 n_s \cdot 4^{B-2}}. \quad (9)$$

In equation (15), the bound of gradients $\max_{i \in [N]} |g_i|$ is determined by the labels and the loss function. For regression tasks, we can normalize the labels to a fixed range, e.g., $[0, 1]$, thus the gradients are bounded. For classification tasks, the gradients are bounded (in the range $[-1, 1]$) by nature with cross-entropy loss. With enough data in leaf s, if the split is not too imbalanced and partitions enough data into both children s_1 and s_2, then we can expect the error caused by quantization is small. Theorem 5.3 ensures that with quantized training, either a split on a leaf has limited impact on the prediction values, or quantization does not change the gains of the splits too much.

In equation (15), we also notice that with more bits (larger B) to represent gradients in quantization, the error will be smaller. This is consistent with our intuition that with more quantized values, we can approximate the original gradients better. But in practice, we found that a smaller number of bits (e.g., 2 to 4 bits) is enough for a good accuracy. With $n_s = 10^7$ is parent node s, and $n_{s_1} = n_{s_2} = 5 \times 10^6$. Let $\epsilon = \max_{i \in [N]} |g_i| /10$, assume that in leaf s we have $\gamma_s = 0.2$ and let $\delta = 0.1$. With 3 bits for gradients, the upper bound in (9) is approximately 0.15. With 4 bits, the value is smaller than 0.08. Thus, with a high probability (greater than 90%), changes in split gain due to quantization is well bounded.

We have a similar analysis for the cases where hessians h_i's are not constant. For non-constant hessian values, the analysis requires additional assumptions. Due to limited space we left the details in Appendix B.

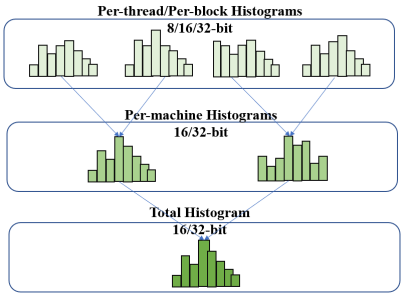


Figure 2: Hierarchical histogram buffers.

Name	#Train	#Test	#Attribute	Task	Metric
Higgs[2]	10,500,000	500,000	28	Binary	AUC↑
Epsilon[2]	400,000	100,000	2,000	Binary	AUC ↑
Kitsune[18]	14,545,195	3,999,993	115	Binary	AUC↑
Criteo[3]	41,256,555	4,584,061	67	Binary	AUC↑
Bosch[4]	1,000,000	183,747	968	Binary	AUC↑
Year [3]	463,715	51,630	90	Regression	RMSE↓
Yahoo LTR [5]	473,134	165,660	700	Ranking	NDCG@10↑
LETOR [21]	2,270,296	753,611	137	Ranking	NDCG@10↑

Table 1: Datasets used in experiments.

6 System Implementation

In this section, we show how to implement quantized GBDT for good efficiency. Current CPU and GPU architectures have limited support for the representation and calculation of low-precision numbers. For example, most CPUs by today only support 8-bit integers as the smallest data type. Under these limitations, the benefits of quantized training cannot be fully exploited. Though the gradients can be compressed into integers with only 2 to 4 bits, we have to use at least 8-bit arithmetics for accumulations in histograms. Even with these limitations, we achieve a considerable speedup on existing CPU and GPU architectures, given the techniques introduced in this section.

6.1 Hierarchical Histogram Buffers

Note that in Algorithm 1 most operations are accumulating gradients. Thus, low-bitwidth gradients will not bring much speedup if high-bitwidth integers are used to store the accumulations. On the contrary, integer overflow may occur with low-bitwidth integers storing the accumulations. To maximally exploit low-precision computation resources while avoiding integer overflow, we partition the training dataset by rows (instances), and assign one partition to a thread on CPU or multiple CUDA blocks on GPU. Thus, each thread or CUDA block constructs a local histogram over the instances in the assigned partition. Since the number of data per partition is much smaller, in most cases 8-bit or 16-bit integers for the accumulations in the histogram is enough. In the end, local histograms are reduced to the total histogram which may use more bits per bin, e.g. with 16-bit or 32-bit integers. Figure 2 displays this divide-and-merge strategy and how the bitwidths differ in local and total histograms.

6.2 Packed Gradient and Hessian

We pack the summations of gradients and hessians in a histogram bin into a single integer in a similar way as SecureBoost+[6]. For example, if we use two 16-bit integers to store the accumulation values, then the packed accumulation value would be a single 32-bit integer. Since gradients can be both negative or positive, the accumulation of gradients must reside in the upper part of the packed integer to exploit the signed bit. The accumulation of hessians resides in the lower part of the packed integer. When accumulating the gradient and hessian of an instance into a bin, the low-bitwidth gradient and hessian are packed in the same way before the addition.

Packed gradient and hessian halves the number of memory accesses in histogram construction. It also halves the number of integer additions. On GPU, the packing contributes significantly to the efficiency. This is partially because histogram construction on GPUs relies on atomic operations to ensure correctness when multiple threads add to the same histogram. Thus packing also reduces the overhead of atomic operations.

7 Experiments

We evaluate our quantized GBDT training system on public datasets, and compare both accuracy and efficiency with other GBDT tools. First, we show that quantized GBDT training preserves the accuracy with low-bitwidth gradients. Then, we assess the efficiency on CPUs and GPU on a single

[2] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/binary.html#epsilon

[3] https://go.criteo.net/criteo-research-kaggle-display-advertising-challenge-dataset.tar.gz

[4] https://www.kaggle.com/c/bosch-production-line-performance

[5] https://webscope.sandbox.yahoo.com/catalog.php?datatype=c

Table 2: Comparison of accuracy, w.r.t. different quantized bits.

Bitwidth	Binary Classification					Regression	Ranking	
	Higgs↑	Epsilon↑	Kitsune↑	Criteo↑	Bosch↑	Year↓	Yahoo LTR↑	LETOR↑
32-bit	0.845694	0.950203	0.950561	0.803791	0.703101	8.956278	**0.793857**	0.524265
2-bit SR_{refit}	0.845587	0.949472	0.952703	0.803293	0.700322	8.953388	0.788579	0.519268
3-bit SR_{refit}	0.845725	0.949884	0.951309	0.803768	0.702756	**8.937374**	0.791077	0.522220
4-bit SR_{refit}	0.845507	0.950049	0.950911	0.803783	0.703315	8.942898	0.792664	0.523796
5-bit SR_{refit}	0.845706	0.950298	0.949229	0.803766	0.702971	8.948542	0.793166	**0.524673**
2-bit $SR_{no\ refit}$	**0.846713**	0.944509	0.952974	0.803750	0.700900	9.112302	0.764862	0.486193
3-bit $SR_{no\ refit}$	0.846040	0.949593	0.951385	**0.803922**	0.702501	8.990034	0.780041	0.507689
4-bit $SR_{no\ refit}$	0.845816	0.950127	0.951197	0.803812	0.703327	8.955256	0.787575	0.515767
5-bit $SR_{no\ refit}$	0.845842	0.950275	0.949794	0.803790	0.703226	8.952768	0.791631	0.520900
2-bit RN_{refit}	0.795991	0.889149	0.962201	0.779906	0.686617	9.429014	0.765103	0.454894
3-bit RN_{refit}	0.830506	0.944329	**0.966606**	0.782732	0.688899	9.062854	0.772364	0.476726
4-bit RN_{refit}	0.840747	0.949946	0.961938	0.795803	0.691469	8.968694	0.777347	0.487256
5-bit RN_{refit}	0.843820	**0.950457**	0.962427	0.802438	0.698954	8.952418	0.784333	0.494951
2-bit $RN_{no\ refit}$	0.836683	0.925220	0.946016	0.768338	0.704445	10.685840	0.632058	0.203732
3-bit $RN_{no\ refit}$	0.843482	0.946850	0.940961	0.791709	**0.708724**	9.377560	0.732487	0.350127
4-bit $RN_{no\ refit}$	0.845788	0.949676	0.949228	0.802689	0.703718	8.969828	0.765432	0.437317
5-bit $RN_{no\ refit}$	0.845765	0.950307	0.952420	0.803645	0.698419	8.965400	0.782608	0.485752

machine, and scalability with distributed training over multiple CPU machines. Ablation study is provided to analyze the techniques for accuracy preserving and system implementation. Table 1 lists the datasets used for the experiments. For Criteo, we encode the categorical features by target and count encoding. We include open-source GBDT tools XGBoost, LightGBM and CatBoost as baselines. We use the leaf-wise tree growing strategy[23] for all baselines. A full description of datasets and hyperparameter settings is provided in Appendix C.

7.1 Accuracy of Quantized Training

We evaluate the accuracy of quantized training. For each dataset and each algorithm, result on the best iteration of test set is reported. We use gradients with 2-bit, 3-bit, 4-bit and 5-bit integers, and compare the results with 32-bit single-precision FP gradients. Table 2 lists the best metric achieved on test set across all boosting iterations. We denote stochastic rounding as SR and round-to-nearest as RN. And use the subscripts to indicate whether to apply leaf-value refitting. Here we focus on the effect of quantization on accuracy only (the SR_{refit} part), and leave the discuss of rounding strategy and leaf-value refitting in 7.4.1. And due to limited space, we list only accuracy on CPUs here. Appendix C provides a similar table for GPU version. With low-bitwidth gradients, the accuracy is not affected much. For some datasets, including Epsilon, Yahoo LTR and LETOR, more gradient bits can improve the performance. Other datasets are less sensitive to the number of bits.

7.2 Speedup on Standalone Machine

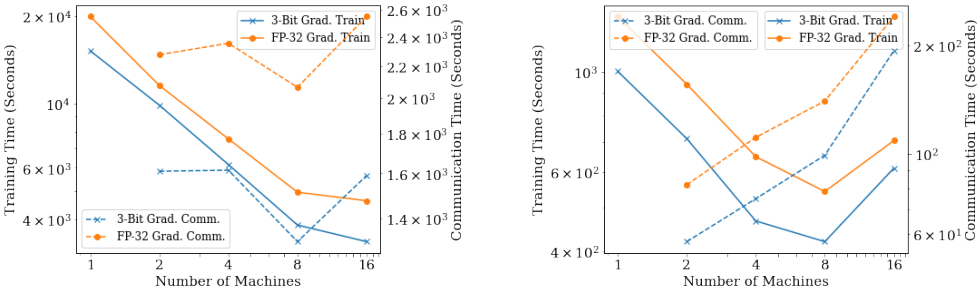
We evaluate the acceleration brought by quantized training on a single machine by comparing it with CPU and CUDA implementations of popular open-source GBDT tools. Our quantized training on CPU is based on LightGBM. We compare our CPU implementation with XGBoost, CatBoost and LightGBM with 32-bit gradients. For the comparison on GPU, we implement a new CUDA version of LightGBM (denoted as LightGBM+, details are in Appendix E) and further implement quantized training based on it. We summarize the benchmark results in the Table 3. Overall, quantized training achieves comparable accuracy with a shorter training time. For the histogram construction time, quantized training speedups histogram construction by up to 3.8 times on GPU, with an overall speedup up to 2.3 times compared with LightGBM+. We also see the acceleration on CPU, but with relatively small gains compared with GPU. In summary, quantized training pushes forward the SOTA efficiency of GBDT algorithms. The number of bits for gradients does not influence the training time much. For histogram construction time, we only list the results with 2-bit gradients due to limited space. A full table can be found in 6 of Appendix C.

7.3 Speedup of Distributed Training

We evaluate quantization in distributed GBDT training on an enlarged version of Epsilon dataset, which duplicates 20 copies of the training set of the original Epsilon dataset, and Criteo. Figure 3 shows how training time and communication time vary with different number of machines. Quan-

Table 3: Detailed time costs for different algorithms in different datasets (seconds).

	Algorithm	Bosch	Criteo	Epsilon	Higgs	Kitsune	Year	Yahoo LTR	LETOR
GPU total time	XGBoost	73	373	125	28	195	**15**	41	48
	CatBoost	21	199	100	59	81	32	58	N/A
	LightGBM+	22	101	86	28	77	24	30	41
	LightGBM+ 2-bit	13	62	**38**	**24**	**37**	18	23	**33**
	LightGBM+ 3-bit	13	60	39	24	39	18	**24**	34
	LightGBM+ 4-bit	13	**59**	39	**24**	40	**17**	26	**33**
	LightGBM+ 5-bit	**12**	**59**	41	**24**	40	**17**	25	34
CPU total time	XGBoost	326	1243	2697	201	606	62	213	155
	CatBoost	2829	11880	1659	1607	2023	130	761	1283
	LightGBM	109	863	846	149	454	23	136	166
	LightGBM 2-bit	**84**	764	808	136	309	24	**95**	125
	LightGBM 3-bit	87	730	**775**	**125**	**289**	23	101	**123**
	LightGBM 4-bit	87	706	**775**	131	291	23	102	129
	LightGBM 5-bit	**84**	**678**	776	127	338	**22**	105	128
GPU Histogram time	LightGBM+	17	70	46	11	54	9	11	17
	LightGBM+ 2-bit	**8**	21	**12**	**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**



(a) Epsilon-8M (b) Criteo

Figure 3: Scaling on distributed systems.

tization can consistently reduce the communication cost and accelerate distributed training. On Criteo dataset, the communication cost dominates with 16 machines, and the training time with 16 machines is even slower than with 4 machines. This shows that there's still a large space to improve the scalability of distributed GBDT training with leaf-wise tree growing strategy.

7.4 Ablation Study

We analyze the effect of proposed techniques that ensure the effectiveness and efficiency of quantized training. We also provide a discussion on the feasibility of the weak-learnability assumption used in Section 5.

7.4.1 Rounding Strategies and Leaf-Value Refitting

Stochastic rounding plays an important role in the accuracy of quantized training. Table 2 also compares the accuracy on test sets with round-to-nearest (denoted as RN) and stochastic rounding (denoted as SR), with different numbers of gradient bits. For both strategies, the results with and without leaf-value refitting are both reported. With smaller numbers of bits, the SR strategy significantly outperforms the RN strategy. Note that leaf value refitting is not able to compensate for the drawback of the RN. Table 2 also demonstrates how leaf value refitting influences accuracy. For binary classification, leaf value refitting is less important as a comparable accuracy could be achieved without it. While for regression and ranking tasks, leaf value refitting is crucial to achieve the comparable accuracy.

7.4.2 Packed Gradient and Hessian

Figure 5 shows the training time and histogram construction time with and without packing gradient and hessian, testing with 3-bit gradients. No-packing version on GPU requires atomic addition for 16-bit integers, which is not natively supported by NVIDIA V100 GPUs. So the no-packing version of GPU is implemented based on atomic compare-and-swap operation, which may incur additional overhead. Overall, packing is important for the speed of histogram construction on both CPU and GPU.

7.4.3 Histogram Bitwidth

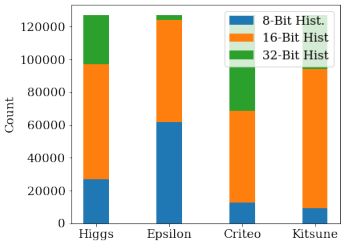


Figure 4: Histogram bitwidth frequency.

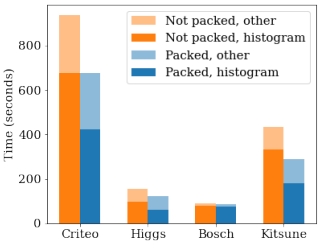


Figure 5: Speedup by packing gradients.

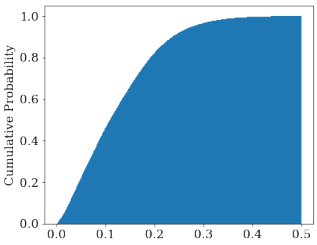


Figure 6: Cumulative distribution of $\hat{\gamma}_s$.

Figure 4 shows the frequency of per-thread/per-block histograms of different bitwidths used during training with 2-bit gradients on CPUs. The counts are the number of leaves in the boosting process using the corresponding bitwidth in per-thread or per-block histograms. On CPUs, our per-thread histograms can use either 8-bit, 16-bit or 32-bit integers. But 8-bit integer histogram is only available with 2-bit gradients. The bitwidth in histograms is determined by the number of data in the leaves, the number of threads/blocks to use, the total number of bins in histograms of all features, and the gradient bitwidth. The minimum bitwidth of histogram integers that won't result in an integer overflow is chosen. The figure shows that quantized training exploits low-precision computations. On GPUs, there's no native support for atomic addition of 8-bit integers. And using 8-bit histograms makes little difference to the training time. So we use 16-bit histograms in CUDA blocks in our current implementation.

7.4.4 Feasibility of Weak-Learnability Assumption

To verify the feasibility of Assumption 5.2, we should evaluate the largest γ_s with which a stump with weighted classification error rate $\frac{1}{2} - \gamma_s$ exists. This requires enumerating all the splits again and calculating the weighted error rate for each split, which is costing. Luckily, we have an easy approach to obtaining a lower bound of the largest γ_s, which can be naturally recorded during training of GBDT, we name the lower bound as $\hat{\gamma}_s$. $\hat{\gamma}_s$ is the weighted error rate of the optimal split in leaf s according to the criterion in (4). Figure 6 shows the cumulative distribution of $\hat{\gamma}_s$ of all leaves during quantized training of Year dataset with 3-bit gradients. For most leaves (over 75%), a $\gamma_s > 0.05$ exists. Thus γ_s won't be too small for most leaves.

8 Discussion and Future Work

In this paper, we try to answer an important but previously neglected question: Can GBDT exploit low-precision training? We propose a quantized training framework with low-bitwidth integer arithmetics which enables the low-precision training of GBDT. We identify the key techniques to preserve accuracy with quantized training, including stochastic rounding and leaf-value refitting. Theoretical analysis shows that quantization has limited impact over the selection of optimal splits in leaves, given enough training data. We implement our algorithm with both CPU and GPU versions. Experiments show that our quantized training GBDT method can achieve comparable accuracy with significant speedup over SOTA GBDT tools, with CPU, GPU and distributed clusters. We believe our method will bring new inspirations in improving GBDT training algorithms, as low-precision computation becomes a trend in machine learning.

References

[1] Firas Abuzaid, Joseph K Bradley, Feynman T Liang, Andrew Feng, Lee Yang, Matei Zaharia, and Ameet S Talwalkar. Yggdrasil: An optimized system for training deep decision trees at scale. In *Advances in Neural Information Processing Systems*, pages 3817–3825, 2016.

[2] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):1–9, 2014.

[3] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. 2011.

[4] Jianfei Chen, Yu Gai, Zhewei Yao, Michael W Mahoney, and Joseph E Gonzalez. A statistical framework for low-bitwidth training of deep neural networks. *Advances in Neural Information Processing Systems*, 33:883–894, 2020.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[6] Weijing Chen, Guoqiang Ma, Tao Fan, Yan Kang, Qian Xu, and Qiang Yang. Secureboost+: A high performance gradient boosting tree framework for large scale vertical federated learning. *arXiv preprint arXiv:2110.10927*, 2021.

[7] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems*, pages 3123–3131, 2015.

[8] Laurens Devos, Wannes Meert, and Jesse Davis. Fast gradient boosting decision trees with bit-level data structures. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 590–606. Springer, 2019.

[9] Fangeheng Fu, Jiawei Jiang, Yingxia Shao, and Bin Cui. An experimental evaluation of large scale gbdt systems. *Proceedings of the VLDB Endowment*, 12(11):1357–1370, 2019.

[10] Alexander Grubb and J Andrew Bagnell. Generalized boosting algorithms for convex optimization. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1209–1216, 2011.

[11] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *International conference on machine learning*, pages 1737–1746. PMLR, 2015.

[12] Ziniu Hu, Yang Wang, Qu Peng, and Hang Li. Unbiased lambdamart: An unbiased pairwise learning-to-rank algorithm. In *The World Wide Web Conference*, pages 2830–2836, 2019.

[13] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. *Advances in neural information processing systems*, 29, 2016.

[14] Jiawei Jiang, Bin Cui, Ce Zhang, and Fangcheng Fu. Dimboost: Boosting gradient boosting decision tree to higher dimensions. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1363–1376, 2018.

[15] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

[16] Qi Meng, Guolin Ke, Taifeng Wang, Wei Chen, Qiwei Ye, Zhi-Ming Ma, and Tie-Yan Liu. A communication-efficient parallel algorithm for decision tree. *Advances in Neural Information Processing Systems*, 29, 2016.

[17] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. In *International Conference on Learning Representations*, 2018.

[18] Yisroel Mirsky, Tomer Doitshman, Yuval Elovici, and Asaf Shabtai. Kitsune: An ensemble of autoencoders for online network intrusion detection. In *Network and Distributed Systems Security (NDSS) Symposium*, 2018.

[19] Richard Nock and Robert Williamson. Lossless or quantized boosting with integer arithmetic. In *International Conference on Machine Learning*, pages 4829–4838. PMLR, 2019.

[20] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6639–6649, 2018.

[21] Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. Letor: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 13(4):346–374, 2010.

[22] Zhen Qin, Le Yan, Honglei Zhuang, Yi Tay, Rama Kumar Pasumarthi, Xuanhui Wang, Michael Bendersky, and Marc Najork. Are neural rankers still outperformed by gradient boosted decision trees? In *International Conference on Learning Representations*, 2020.

[23] Haijian Shi. *Best-first decision tree learning*. PhD thesis, The University of Waikato, 2007.

[24] Meng Wan, Chongxin Deng, Siyu Yu, and JD Vectorlab. Diting: A real-time distributed feature serving system for machine learning. Technical report, EasyChair, 2019.

[25] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: ternary gradients to reduce communication in distributed deep learning. In *31st International Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[26] Zeyi Wen, Qinbin Li, Bingsheng He, and Bin Cui. Challenges and opportunities of building fast gbdt systems. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21) Survey Track*, 2021.

[27] Lei Yang, Hong Wu, Tieying Zhang, Xuntao Cheng, Feifei Li, Lei Zou, Yujie Wang, Rongyao Chen, Jianying Wang, and Gui Huang. Leaper: a learned prefetcher for cache invalidation in lsm-tree based storage engines. *Proceedings of the VLDB Endowment*, 13(12):1976–1989, 2020.

[28] Xiao Yang, Daren Sun, Ruiwei Zhu, Tao Deng, Zhi Guo, Zongyao Ding, Shouke Qin, and Yanfeng Zhu. Aiads: Automated and intelligent advertising system for sponsored search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1881–1890, 2019.

[29] Huan Zhang, Si Si, and Cho-Jui Hsieh. Gpu-acceleration for large-scale tree boosting. In *SysML Conference*, 2018.

[30] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *CoRR*, abs/1606.06160, 2016.

[31] Wenjun Zhou, Taposh Dutta Roy, and Iryna Skrypnyk. The kdd cup 2019 report. *ACM SIGKDD Explorations Newsletter*, 22(1):8–17, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Our implementation is limited by existing processor architecures, which is discussed in section 6.1
   (c) Did you discuss any potential negative societal impacts of your work? [No] We believe our method has no negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] All assumptions are strictly stated.
   (b) Did you include complete proofs of all theoretical results? [Yes] We provide complete and strict proofs in the supplementary.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The data we used are all from public databases.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? All these details are available in our paper and the appendix. Our source code and processed datasets are publicly available. [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Our accuracy evaluation in Table 2 are averaged over 5 different runs. [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Provided in supplementary.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We discuss all the used datasets in the experiment section

(b) Did you mention the license of the assets? [Yes]

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] We use publicly available benchmark datasets.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] We use publicly available benchmark datasets.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not use crowdsourcing or conducted research with human subjects.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not use crowdsourcing or conducted research with human subjects.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not use crowdsourcing or conducted research with human subjects.

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 $c(\mathbf{x}_i) \in \{-1, 1\}$ and weights $\{w_i\}_{i=1}^N$, where $w_i \geq 0$ 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 weak-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 weak-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 on leaf s and its descendents, the resultant average change 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 $s \to s_1, s_2$ on leaf s, with probability at least $1-\delta$,

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

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 change of 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 probability 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 probability 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 probability 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 probability at least $1-\delta$

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

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 on leaf s and its descendants, the resultant weighted average (weighted by h_i) change 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 $s \to s_1, s_2$ on 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

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

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 with 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 probability at least $1 - \frac{\delta}{4}$ by equation (32). Thus with probability 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

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

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

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

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

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

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 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[6]).

[6] 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 Bosch dataset. For 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 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 strategy of XGBoost

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

slows down the training much with max_leaves 255 on Bosch. So 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 has for CatBoost is 35552cf8057447262eedd9671f66fd715af34946. And for XGBoost it is fe4ce920b250d39133a7f6b1128f80da0d4018c6. For LightGBM, we use the modified 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[29, 21, 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 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

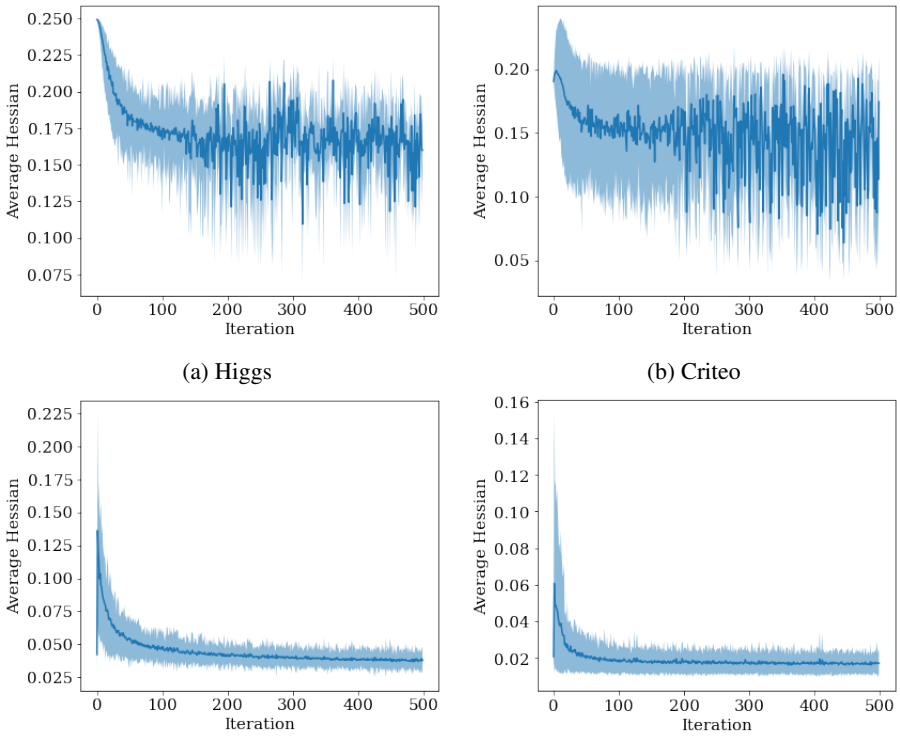
Appendix B.3 provides the theoretical analysis and proof for the error caused by quantization for loss functions with non-constant hessians. The assumption is a little bit stronger than constant hessian loss functions in that we are expecting the average hessian values per leaf won't be too small, so 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 the 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 data point contribute to the approximated loss by second-order Taylor expansion[5]. So 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.