Appendix

A Additional Experiment Details

Datasets: Table 3 summaries the basic statistics of the LIBSVM datasets that were used.

Dataset	task	# samples	# features
a1a	classification	1605	123
w1a	classification	2477	300
diabetes	classification	768	8
german	classification	1000	24
housing	regression	506	13
sonar	classification	208	60

Table 3: Basic statistics of the (real) datasets used.

All fine-tuning experiments: We now look at the testing performance of GBM and AGBM on six datasets with hyperparameter tuning.

For each dataset, we randomly choose 80% as the training and the remaining as the testing dataset. We repeat this splitting 5 times and report mean train and test errors along with standard errors.

We consider depth 3 trees as weak-learners and fix the number of trees to 30, 50 and 100 (notice, that for AGBM that means that the number of boosting iterations is 15, 25 and 50 respectively). We fix learning rate (η) to 0.1 and tune (using 5 fold cross-validation on training dataset with *RandomizedSearchCV* in scikit-learn) the following parameters:

- *min_split_gain* - [10, 5, 2, 1, 0.5, 0.1, 0.01, 0.001, 1e-4, 1e-5]
- l2 regularizer on leaves - [0.01, 0.1, 0.5, 1,2,4, 8, 16, 32, 64]
- momentum parameter γ (only for AGBM): uniform(0.1,1)

We use early stopping for final training on full training dataset (using 5 early stop rounds)

As AGBM has more parameters (namely γ), we did proportionally more iterations of random search for AGBM.

As we can see from Table 2, the accelerated method in general is beneficial for underfiting scenarios (30 and 50 trees). However, for such small datasets, 100 weak learners start overfiting, and accelerated method overfits faster, as expected.

B Extensions and Variants

In this section we study two more practical variants of AGBM. First we see how to restart the algorithm to take advantage of strong convexity of the loss function. Then we will study a straight-forward approach to accelerated GBM, which we call vanilla accelerated gradient boosting machine (VAGBM), a variant of the recently proposed algorithm in Biau et al. (2018), however without any theoretical guarantees.

B.1 A Vanilla Accelerated Gradient Boosting Method

A natural question to ask is whether, instead of adding *two* learners at each iteration, we can get away with adding only *one*? Below we show how such an algorithm would look like and argue that it may not always converge.

Following the updates in Equation (3), we can get a direct acceleration of GBM by using the weak learner fitting the gradient. This leads to an Algorithm 4.

Algorithm 4 Vanilla Accelerated Gradient Boosting Machine (VAGBM)

```
**Input.** Starting function $f^0(x) = 0$, step-size $\eta$, momentum parameter $\gamma \in (0, 1]$.
**Initialization.** $h^0(x) = f^0(x)$, and sequence $\theta_m = \frac{2}{m+2}$.
For $m = 0, \dots, M-1$ do:
**Perform Updates:**
(1) Compute a linear combination of $f$ and $h$: $g^m(x) = (1-\theta_m)f^m(x) + \theta_m h^m(x)$.
(2) Compute pseudo residual: $r^m = -\left[\frac{\partial \ell(y_i, g^m(x_i))}{\partial g^m(x_i)}\right]_{i=1,\dots,n}$.
(3) Find the best weak-learner for pseudo residual: $\tau_m = \arg\min_{\tau \in \mathcal{T}_m} \sum_{i=1}^n (r_i^m - b_\tau(x_i))^2$.
(4) Update the model: $f^{m+1}(x) = g^m(x) + \eta b_{\tau_m}(x)$.
(5) Update the momentum model: $h^{m+1}(x) = h^m(x) + \eta/\theta_m b_{\tau_m}(x)$.
**Output.** $f^M(x)$.
```

Algorithm 4 is equivalent to the recently developed accelerated gradient boosting machines algorithm (Biau et al., 2018; Fouillen et al., 2018). Unfortunately, it **may not always converge** to an optimum or may even **diverge**. This is because b_{τ_m} from Step (2) is only an approximate-fit to r^m, meaning that we only take an *approximate* gradient descent step. While this is not an issue in the non-accelerated version, in Step (2) of Algorithm 4, the momentum term pushes the h sequence to take a large step along the approximate gradient direction. This exacerbates the effect of the approximate direction and can lead to an additive accumulation of error as shown in Devolder et al. (2014). In Section 5.1, we see that this is not just a theoretical concern, but that Algorithm 4 also diverges in practice in some situations.

Remark B.1. *Our* corrected *residual c^m in Algorithm 2 was crucial to the theoretical proof of converge in Theorem 4.1. One extension could be to introduce $\gamma \in (0, 1)$ in step (5) of Algorithm 4 just as in Algorithm 2.*

Remark B.2. *It is worth noting that Vanilla AGBM may bring good empirical performance on small datasets. We hypothesize that the accumulated error in gradient may serve as an additional regularization that slows down overfitting*

C Proof of Theorem 4.1

This section proves our major theoretical result in the paper:

Theorem 4.1 Consider Accelerated Gradient Boosting Machine (Algorithm 2). Suppose ℓ is σ-smooth, the step-size $\eta \le \frac{1}{\sigma}$ and the momentum parameter $\gamma \le \Theta^4/(4+\Theta^2)$. Then for all $M \ge 0$, we have:

$$L(f^M) - L(f^*) \le \frac{1}{2\eta\gamma(M+1)^2}\|f^*(X)\|_2^2 \,.$$

□

Let's start with some new notations. Define scalar constants $s = \gamma/\Theta^2$ and $t := (1-s)/2 \in (0, 1)$. We mostly only need $s + t \le 1$—the specific values of γ and t are needed only in Lemma C.6. Then define

$$\alpha_m := \frac{\eta\gamma}{\theta_m} = \frac{\eta s \Theta^2}{\theta_m} \,,$$

then the definitions of the sequences $\{r^m\}$, $\{c^m\}$, $\hat{h}^m(X)$ and $\{\theta_m\}$ from Algorithm 3 can be simplified as:

$$\begin{aligned}
\theta_m &= \frac{2}{m+2} \\
r^m &= -\left[\frac{\partial l(y_i, g^m(x_i))}{\partial g^m(x_i)}\right]_{i=1,\dots n} \\
c^m &= r^m + (\alpha_{m-1}/\alpha_m)\left[c^{m-1} - b_{\tau^2_{m-1}}(X)\right] \\
\hat{h}^{m+1}(X) &= \hat{h}^m(X) + \alpha_m r^m \,.
\end{aligned}$$

The sequence $\hat{h}^m(X)$ is in fact closely tied to the sequence $h^m(X)$ as we show in the next lemma. For notational convenience, we define $c^{-1} = b_{\tau^2_{-1}}(X) = 0$ and similarly $\frac{\alpha_{-1}}{\theta_{-1}} = 0$ throughout the proof.

Lemma C.1.

$$\hat{h}^{m+1}(X) = h^{m+1}(X) + \alpha_m(c_m - b_{\tau_{m,2}}(X))\,.$$

Proof. Observe that

$$\hat{h}^{m+1}(X) = \sum_{j=0}^{m} \alpha_j r^j \quad \text{and that} \quad h^{m+1}(X) = \sum_{j=0}^{m} \alpha_j b_{\tau_{j,2}}(X)\,.$$

Then we have

$$\begin{aligned}
\hat{h}^{m+1}(X) - h^{m+1}(X) &= \sum_{j=0}^{m} \alpha_j (r^j - b_{\tau_{j,2}}(X)) \\
&= \sum_{j=0}^{m} \alpha_j (r^j - \frac{\alpha_{j-1}}{\alpha_j} b_{\tau^2_{j-1}}(X)) - \alpha_m b_{\tau_{m,2}}(X) \\
&= \sum_{j=0}^{m} \alpha_j (c^j - \frac{\alpha_{j-1}}{\alpha_j} c^{j-1}) - \alpha_m b_{\tau_{m,2}}(X) \\
&= \sum_{j=0}^{m} (\alpha_j c^j - \alpha_{j-1} c^{j-1}) - \alpha_m b_{\tau_{m,2}}(X) \\
&= \alpha_m (c_m - b_{\tau_{m,2}}(X))\,,
\end{aligned}$$

where the third equality is due to the definition of c^m. □

Lemma C.2 presents the fact that there is sufficient decay of the loss function:

Lemma C.2.

$$L(f^{m+1}) \leq L(g^m) - \frac{\eta \Theta^2}{2} \|r^m\|^2\,.$$

Proof. Recall that $\tau_{m,1}$ is chosen such that

$$\tau_{m,1} = \arg\min_{\tau \in \mathcal{T}} \|b_\tau(X) - r^m\|^2\,.$$

Since the class of learners $\mathcal{T}$ is *scalable* (Assumption 2.1), we have

$$\begin{aligned}
\|b_{\tau_{m,1}}(X) - r^m\|^2 &= \min_{\tau \in \mathcal{T}_m} \min_{\sigma \in \mathbb{R}} \|\sigma b_\tau(X) - r^m\|^2 \\
&= \|r^m\|^2 \left[1 - \arg\max_{\tau \in \mathcal{T}} cos(r^m, b_\tau(X))^2\right] \\
&\leq \|r^m\|^2 \left[1 - \Theta^2\right]\,, \qquad (6)
\end{aligned}$$

where the last inequality is because of the definition of Θ, and the second equality is due to the simple fact that for any two vectors a and b,

$$\min_{\sigma \in \mathbb{R}} \|\sigma a - b\|^2 = \|a\|^2 - \max_{\sigma \in \mathbb{R}} \left[\sigma \langle a, b \rangle - \frac{\sigma^2}{2}\|b\|^2\right] = \|a\|^2 - \|a\|^2 \frac{\langle a, b \rangle}{\|a\|^2 \|b\|^2}\,.$$

Now recall that $L(f^{m+1}) = \sum_{i=1}^{n} l(y_i, f^{m+1}(x_i))$ and that $f^{m+1}(x) = g^m(x) + \eta b_{\tau_{m,1}}(x)$. Since the loss function

$l(y_i, x)$ is σ-smooth and step-size $\eta \leq \frac{1}{\sigma}$, it holds that

$$\begin{aligned}
L(f^{m+1}) &= \sum_{i=1}^{n} l(y_i, f^{m+1}(x_i)) \\
&\leq \sum_{i=1}^{n} l(y_i, g^m(x_i) + \eta b_{\tau_{m,1}}(x_i)) \\
&\leq \sum_{i=1}^{n} [*]\, l(y_i, g^m(x_i)) + \frac{\partial l(y_i, g^m(x_i))}{\partial g^m(x_i)} (\eta b_{\tau_{m,1}}(x_i)) + \frac{\sigma}{2} (\eta b_{\tau_{m,1}}(x_i))^2 \\
&\leq \sum_{i=1}^{n} [*]\, l(y_i, g^m(x_i)) + \frac{\partial l(y_i, g^m(x_i))}{\partial g^m(x_i)} (\eta b_{\tau_{m,1}}(x_i)) + \frac{\eta}{2} (b_{\tau_{m,1}}(x_i))^2 \\
&= \sum_{i=1}^{n} [*]\, l(y_i, g^m(x_i)) - r_i^m (\eta b_{\tau_{m,1}}(x_i)) + \frac{1}{2\eta} (b_{\tau_{m,1}}(x_i))^2 \\
&= L(g^m) - \eta \langle r^m, b_{\tau_{m,1}}(X) \rangle + \frac{\eta}{2} \| b_{\tau_{m,1}}(X) \|^2 \\
&= L(g^m) + \frac{\eta}{2} \| b_{\tau_{m,1}}(X) - r^m \|^2 - \frac{\eta}{2} \| r^m \|^2 \\
&\leq L(g^m) - \frac{\Theta^2 \eta}{2} \| r^m \|^2 \,,
\end{aligned}$$

where the final inequality follows from (6). This furnishes the proof of the lemma. □

Lemma C.3 is a basic fact of convex function, and it is commonly used in the convergence analysis in accelerated method.

Lemma C.3. *For any function f and $m \geq 0$,*

$$L(g^m) + \theta_m \langle r^m, h^m(X) - f(X) \rangle \leq \theta_m L(f) + (1 - \theta_m) L(f^m) \,.$$

Proof. For any function f, it follows from the convexity of the loss function l that

$$\begin{aligned}
L(g^m) + \langle r^m, g^m(X) - f(X) \rangle &= \sum_{i=1}^{n} l(y_i, g^m(x_i)) + \frac{\partial l(y_i, g^m(x_i))}{\partial g^m(x_i)} (f(x_i) - g^m(x_i)) \\
&\leq \sum_{i=1}^{n} l(y_i, f(x_i)) = L(f) \,. \qquad (7)
\end{aligned}$$

Substituting $f = f^m$ in (7), we get

$$L(g^m) + \langle r^m, g^m(X) - f^m(X) \rangle \leq L(f^m) \,. \qquad (8)$$

Also recall that $g^m(X) = (1 - \theta_m) f^m(X) + \theta_m h^m(X)$. This can be rewritten as

$$\theta_m (g^m(X) - h^m(X)) = (1 - \theta_m)(f^m(X) - g^m(X)) \,. \qquad (9)$$

Putting (7), (8), and (9) together:

$$\begin{aligned}
&L(g^m) + \theta_m \langle r^m, h^m(X) - f(X) \rangle \\
=& L(g^m) + \theta_m \langle r^m, g^m(X) - f(X) \rangle + \theta_m \langle r^m, h^m(X) - g^m(X) \rangle \\
=& \theta_m [L(g^m) + \langle r^m, g^m(X) - f(X) \rangle] + (1 - \theta_m)[L(g^m) + \langle r^m, g^m(X) - f^m(X) \rangle] \\
\leq& \theta_m L(f) + (1 - \theta_m) L(f^m) \,,
\end{aligned}$$

which finishes the proof. □

We are ready to prove the key lemma which gives us the accelerated rate of convergence.

Lemma C.4. *Define the following potential function $V(f)$ for any given output function f:*

$$V^m(f) = \frac{\alpha_{m-1}}{\theta_{m-1}}\left(L(f^m) - L(f)\right) + \frac{1}{2}\left\|f(X) - \hat{h}^m(X)\right\|^2 . \tag{10}$$

At every step, the potential decreases at least by δ_m:

$$V^{m+1}(f) \le V^m(f) + \delta_m ,$$

where δ_m is defined as:

$$\delta_m := \frac{s\alpha_{m-1}^2}{2t}\|c^{m-1} - b_{\tau^2_{m-1}}(X)\|^2 - (1-s-t)\frac{\alpha_m^2}{2s}\|r^m\|^2 . \tag{11}$$

Proof. Recall that $c^{-1} = b_{\tau^2_{-1}}(X)) = 0$ and $\frac{\alpha_{-1}}{\theta_{-1}} = 0$. It follows from Lemma C.2 that:

$$\begin{aligned}
&L(f^{m+1}) - L(g^m) + \frac{(1-s)\eta\Theta^2}{2}\|r^m\|^2 \\
\le& -\frac{s\eta\Theta^2}{2}\|r^m\|^2 \\
=& -\alpha_m\theta_m\|r^m\|^2 + \frac{\alpha_m\theta_m}{2}\|r^m\|^2 \\
=&\theta_m\left\langle r^m, \hat{h}^m(X) - \hat{h}^{m+1}(X)\right\rangle + \frac{\theta_m}{2\alpha_m}\|\hat{h}^m(X) - \hat{h}^{m+1}(X)\|^2 \\
=&\theta_m\left\langle r^m, \hat{h}^m(X) - f(X)\right\rangle + \frac{\theta_m}{2\alpha_m}\left(\|f(X) - \hat{h}^m(X)\|^2 - \|f(X) - \hat{h}^{m+1}(X)\|^2\right) ,
\end{aligned}$$

where the second equality is by the definition of $\hat{h}^m(x)$ and the third is just mathematical manipulation of the equation (it is also called three-point property). By rearranging the above inequality, we have

$$\begin{aligned}
&L(f^{m+1}) + \frac{(1-s)\eta\Theta^2}{2}\|r^m\|^2 \\
\le&L(g^m) + \left\langle r^m, \hat{h}^m(X) - f(X)\right\rangle + \frac{\theta_m}{2\alpha_m}\left(\|f(X) - \hat{h}^m(X)\|^2 - \|f(X) - \hat{h}^{m+1}(X)\|^2\right) \\
=&L(g^m) + \theta_m\langle r^m, h^m(X) - f(X)\rangle + \frac{\theta_m}{2\alpha_m}\left(\|f(X) - \hat{h}^m(X)\|^2 - \|f(X) - \hat{h}^{m+1}(X)\|^2\right) \\
&\quad + \theta_m\left\langle r^m, \hat{h}^m(X) - h^m(X)\right\rangle \\
\le&\theta_m L(f) + (1-\theta_m)L(f^m) + \frac{\theta_m}{2\alpha_m}\left(\|f(X) - \hat{h}^m(X)\|^2 - \|f(X) - \hat{h}^{m+1}(X)\|^2\right) \\
&\quad + \theta_m\alpha_{m-1}\left\langle r^m, c^{m-1} - b_{\tau^2_{m-1}}(X)\right\rangle ,
\end{aligned}$$

where the first inequality uses Lemma C.3 and the last inequality is due to the fact that $\hat{h}^m(X) - h^m(X) = \alpha_{m-1}(c^{m-1} - b_{\tau^2_{m-1}}(X))$ from Lemma C.1. Rearranging the terms and multiplying by (α_m/θ_m) leads to

$$\begin{aligned}
&\frac{\alpha_m}{\theta_m}(L(f^{m+1}) - L(f)) + \frac{1}{2}\|f(X) - \hat{h}^{m+1}(X)\|^2 \\
\le&\underbrace{\frac{\alpha_m(1-\theta_m)}{\theta_m}}_{:=\mathcal{A}}(L(f^m) - L(f)) + \frac{1}{2}\|f(X) - \hat{h}^m(X)\|^2 + \underbrace{\alpha_m\alpha_{m-1}\left\langle r^m, (c^{m-1} - b_{\tau^2_{m-1}}(X))\right\rangle - \frac{(1-s)\eta\Theta^2\alpha_m}{2\theta_m}\|r^m\|^2}_{:=\mathcal{B}} .
\end{aligned}$$

Let us examine first the term $\mathcal{A}$:

$$\frac{\alpha_m(1-\theta_m)}{\theta_m} = (\eta\Theta^2 s)\frac{1-\theta_m}{\theta_m^2} \le (\eta\Theta^2 s)\frac{1}{\theta_{m-1}^2} = \frac{\alpha_{m-1}}{\theta_{m-1}} .$$

We have thus far shown that

$$V^{m+1}(f) \leq V^m(f) + \mathcal{B}\,,$$

and we now need to show that $\mathcal{B} \leq \delta_m$. Using Mean-Value inequality, the first term in $\mathcal{B}$ can be bounded as

$$\alpha_m \alpha_{m-1} \Big\langle r^m, (c^{m-1} - b_{\tau^2_{m-1}}(X)) \Big\rangle \leq \frac{\alpha_m^2 t}{2s} \|r^m\|^2 + \frac{\alpha_{m-1}^2 s}{2t} \|c^{m-1} - b_{\tau^2_{m-1}}(X)\|^2\,.$$

Substituting it in $\mathcal{B}$ shows:

$$\begin{aligned}
\mathcal{B} &= \alpha_m \alpha_{m-1} \Big\langle r^m, (c^{m-1} - b_{\tau^2_{m-1}}(X)) \Big\rangle - \frac{(1-s)\eta\Theta^2\alpha_m}{2\theta_m} \|r^m\|^2 \\
&\leq \frac{\alpha_m^2 t}{2s} \|r^m\|^2 + \frac{\alpha_{m-1}^2 s}{2t} \|c^{m-1} - b_{\tau^2_{m-1}}(X)\|^2 - \frac{(1-s)\alpha_m^2}{2s} \|r^m\|^2 \\
&= \frac{\alpha_{m-1}^2 s}{2t} \|c^{m-1} - b_{\tau^2_{m-1}}(X)\|^2 - (1-s-t)\frac{\alpha_m^2}{2s} \|r^m\|^2 \\
&= \delta_m\,,
\end{aligned}$$

which finishes the proof. □

Unlike the typical proofs of accelerated algorithms, which usually shows that the potential $V^m(f)$ is a decreasing sequence, there is no guarantee that the potential $V^m(f)$ is decreasing in the boosting setting due to the use of weak learners. Instead, we are able to prove that:

Lemma C.5. *For any given m, it holds that $\sum_{j=0}^m \delta_j \leq 0$.*

Proof. We can rewrite the statement of the lemma as:

$$\sum_{j=0}^{m-1} \alpha_j^2 \|c^j - b_{\tau_{j,2}}(X)\|^2 \leq \frac{t(1-s-t)}{s^2} \sum_{j=0}^{m} \alpha_j^2 \|r^j\|^2\,. \tag{12}$$

Here, let us focus on the term $\|c^{j+1} - b_{\tau^2_{j+1}}(X)\|^2$ for a given j. We have that

$$\begin{aligned}
\left\|c^{j+1} - b_{\tau^2_{j+1}}(X)\right\|^2 &\leq (1-\Theta^2)\left\|c^{j+1}\right\|^2 \\
&= (1-\Theta^2)\left\|r^{j+1} + \frac{\theta_{j+1}}{\theta_j}\left[c^j - b_{\tau_{j,2}}(X)\right]\right\|^2 \\
&\leq (1-\Theta^2)(1+\rho)\left\|r^{j+1}\right\|^2 + (1-\Theta^2)(1+1/\rho)\left\|\frac{\theta_{j+1}}{\theta_j}\left[c^j - b_{\tau_{j,2}}(X)\right]\right\|^2 \\
&\leq (1+\rho)(1-\Theta^2)\left\|r^{j+1}\right\|^2 + (1-\Theta^2)(1+1/\rho)\left\|\left[c^j - b_{\tau_{j,2}}(X)\right]\right\|^2,
\end{aligned}$$

where the first inequality follows from our assumption about the density of the weak-learner class $\mathcal{B}$ (the same of the argument in (6)), the second inequality holds for any $\rho \geq 0$ due to Mean-Value inequality, and the last inequality is from $\theta_{j+1} \leq \theta_j$. We now derives a recursive bound on the left side of (12). From this, (12) follows from an elementary fact of recursive sequence as stated in Lemma C.6 with $a_j = \alpha_j^2\left\|c^j - b_{\tau_{j,2}}(X)\right\|^2$ and $c_j = \alpha_j^2\left\|r^j\right\|^2$. □

Remark C.1. *If $c^m = b_{\tau_{m,2}}(X)$ (i.e. our class of learners $\mathcal{B}$ is* strong*), then $\delta_m = -(1-s-t)\frac{\alpha_m^2}{2s^2}\|r^m\|^2 \leq 0$.*

Lemma C.6 is an elementary fact of recursive sequence used in the proof of Lemma C.5.

Lemma C.6. *Given two sequences $\{a_j \geq 0\}$ and $\{c_j \geq 0\}$ such that the following holds for any $\rho \geq 0$,*

$$a_{j+1} \leq (1-\Theta^2)[(1+1/\rho)a_j + (1+\rho)c_{j+1}]\,,$$

then the sum of the terms a_j can be bounded as

$$\sum_{j=0}^m a_j \leq \frac{t(1-s-t)}{s^2} \sum_{j=0}^m c_j\,.$$

Proof. The recursive bound on a_j implies that

$$\begin{aligned} a_j &\le (1-\Theta^2)[(1+1/\rho)a_{j-1} + (1+\rho)c_j] \\ &\le \sum_{k=0}^{j} [(1+1/\rho)(1-\Theta^2)]^{j-k}(1+\rho)(1-\Theta^2)c_k\,. \end{aligned}$$

Summing both the terms gives

$$\begin{aligned} \sum_{j=0}^{m} a_j &\le \sum_{j=0}^{m}\sum_{k=0}^{j}[(1+1/\rho)(1-\Theta^2)]^{j-k}(1+\rho)(1-\Theta^2)c_k \\ &= \sum_{k=0}^{m}\sum_{j=k}^{m}[(1+1/\rho)(1-\Theta^2)]^{j-k}(1+\rho)(1-\Theta^2)c_k \\ &\le \sum_{k=0}^{m}\left(\sum_{j=0}^{\infty}[(1+1/\rho)(1-\Theta^2)]^{j}\right)(1+\rho)(1-\Theta^2)c_k \\ &= \frac{(1+\rho)(1-\Theta^2)}{1-(1+1/\rho)(1-\Theta^2)}\sum_{k=0}^{m} c_k \\ &= \frac{(1+\rho)(1-\Theta^2)}{\Theta^2-(1-\Theta^2)/\rho}\sum_{k=0}^{m} c_k \\ &= \frac{2(1+\rho)(1-\Theta^2)}{\Theta^2}\sum_{k=0}^{m} c_k \\ &= \frac{2(2-\Theta^2)(1-\Theta^2)}{\Theta^4}\sum_{k=0}^{m} c_k\,, \end{aligned}$$

where in the last two equalities we chose $\rho = \frac{2(1-\Theta^2)}{\Theta^2}$. Now recall that $s \le \frac{\Theta^2}{4+\Theta^2} \in (0,1)$ and that $t = (1-s)/2$:

$$\begin{aligned} \sum_{j=0}^{m} a_j &\le \frac{2(2-\Theta^2)(1-\Theta^2)}{\Theta^4}\sum_{k=0}^{m} c_k \\ &\le \frac{4}{\Theta^4}\sum_{k=0}^{m} c_k \\ &= \left(\frac{4+\Theta^2}{\Theta^2}-1\right)^2 \frac{1}{4}\sum_{k=0}^{m} c_k \\ &\le \left(\frac{1}{s}-1\right)^2 \frac{1}{4}\sum_{k=0}^{m} c_k \\ &= \frac{(1-s)^2}{4s^2}\sum_{k=0}^{m} c_k \\ &= \frac{t(1-s-t)}{s^2}\sum_{k=0}^{m} c_k\,. \end{aligned}$$

□

Lemma C.4 and Lemma C.5 directly result in our major theorem:

Proof of Theorem 4.1 It follows from Lemma C.4 and Lemma C.5 that

$$V^M(f^\star) \le V^{M-1}(f^\star) + \delta_m \le V^0(f^\star) + \sum_{j=0}^{M-1}\delta_j \le \frac{1}{2}\|f^0(X) - f^\star(X)\|^2\,.$$

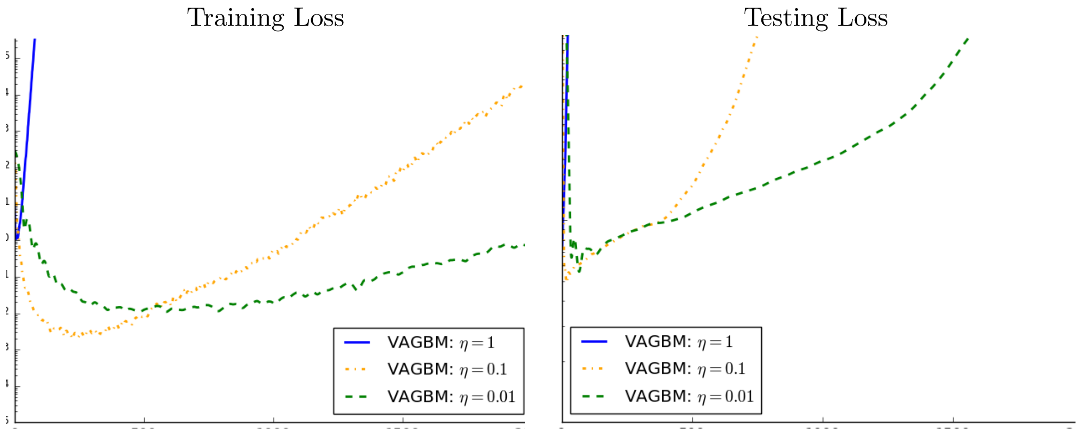

Figure 3: Training and testing loss versus number of trees for AGBM with different η.

Figure 4 presents the performance of different algorithms on tree stumps (namely smaller Θ). They are consistent with Figure 1.

Notice $V^M(f^\star) \geq \frac{\alpha_{m-1}}{\theta_{m-1}}(L(f^M) - L(f^\star))$ as the term $\frac{1}{2}\|f^M(X) - f^\star(X)\|^2 \geq 0$, which induces that

$$L(f^M) - L(f^\star) \leq \frac{\theta_{M-1}}{2\alpha_{M-1}}\|f^0(X) - f^\star(X)\|^2 = \frac{1}{2\gamma\eta} \cdot \frac{\|f^0(X) - f^\star(X)\|^2}{M^2} .$$

□

D Additional Numerical Experiments

D.1 VAGBM may diverge with small η

Figure 3 shows that for smaller η, VAGBM may still diverge. Of course, the smaller the η, the longer VAGBM stay stable.

D.2 Performance of different algorithms with tree stumps

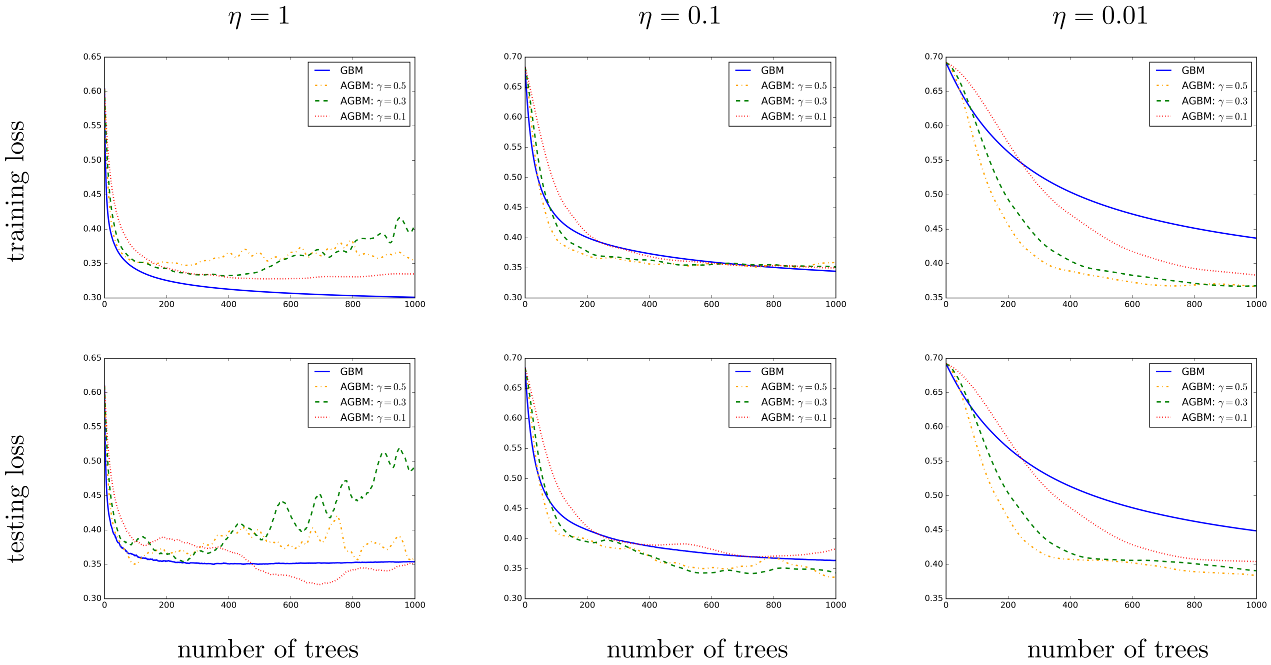

Figure 4: Training and testing loss versus number of trees for logistic regression on a1a with tree stumps (one layer decision trees).