Supplementary Material for Gradient Boosting on Stochastic Data Streams

A Proof of Proposition 4.3

Proof. Given that a no-regret online learning algorithm $\mathcal{A}$ running on sequence of loss $\|h(x_t) - y_t\|^2$, we have can easily see that Eqn. 4 holds as:

$$\sum_{t=1}^{T} \|h_t(x_t) - y_t\|^2 \leq \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \|h(x_t) - y_t\|^2 + R_{\mathcal{A}}(T), \tag{11}$$

where $R_{\mathcal{A}}(T)$ is the regret of $\mathcal{A}$ and is $o(T)$. To prove Proposition 4.3, we only need to show that Eqn. 5 holds for some $\gamma \in (0, 1]$. This is equivalent to showing that there exist a hypothesis $\tilde{h} \in \mathcal{H}$ ($\|\tilde{h}\| = 1$), such that $\langle \tilde{h}, f^* \rangle > 0$. To see this equivalence, let us assume that $\langle \tilde{h}, f^*/\|f^*\| \rangle = \epsilon > 0$. Let us set $h^* = \epsilon \|f^*\| \tilde{h}$. Using Pythagorean theorem, we can see that $\|h^* - f^*\|^2 = (1 - \epsilon^2)\|f^*\|^2$. Hence we get γ is at least ϵ^2, which is in $(0, 1]$.

Now since we assume that $f^* \not\perp span(\mathcal{H})$, then there must exist $h' \in \mathcal{H}$, such that $\langle f^*, h' \rangle \neq 0$, otherwise $f^* \perp \mathcal{H}$. Consider the hypothesis $h'/\|h'\|$ and $-h'/\|h'\|$ (we assume $\mathcal{H}$ is closed under scale), we have that either $\langle h', f^* \rangle > 0$ or $\langle -h', f^* \rangle > 0$. Namely, we find at least one hypothesis h such that $\langle h, f^* \rangle > 0$ and $\|h\| = 1$. Hence if we pick $\tilde{h} = \arg\max_{h \in \mathcal{H}, \|h\|=1} \langle h, f^*/\|f^*\| \rangle$, we must have $\langle \tilde{h}, f^*/\|f^*\| \rangle = \epsilon > 0$. Namely we can find a hypothesis $h^* \in \mathcal{H}$, which is $\epsilon \|f^*\| \tilde{h}$, such that there is non-zero $\gamma \in (0, 1]$:

$$\|h^* - f^*\|^2 \leq (1 - \gamma)\|f^*\|^2. \tag{12}$$

To show that we can extend this γ to the finite sample case, we are going to use Hoeffding inequality to relate the norm $\|\cdot\|$ to its finite sample approximation.

Applying Hoeffding inequality, we get with probability at least $1 - \delta/2$,

$$|\frac{1}{T}\sum_{t=1}^{T} \|y_t\|^2 - \langle f^*, f^* \rangle| \leq O\Big(\sqrt{\frac{F^2}{T}\ln(4/\delta)}\Big), \tag{13}$$

where based on assumption that $f^*(\cdot)$ is bounded as $\|f^*(\cdot)\| \leq F$. Similarly, we have with probability at least $1 - \delta/2$:

$$|\frac{1}{T}\sum_{t=1}^{T} \|h^*(x_t) - f^*(x_t)\|^2 - \|h^* - f^*\|^2| \leq O\Big(\sqrt{\frac{F^2}{T}\ln(4/\delta)}\Big), \tag{14}$$

Apply union bound for the above two high probability statements, we get with probability at least $1 - \delta$,

$$\begin{aligned} &|\frac{1}{T}\sum_{t=1}^{T} y_t^2 - \langle f^*, f^* \rangle| \leq O\Big(\sqrt{\frac{F^2}{T}\ln(4/\delta)}\Big), \quad \textit{and,} \\ &|\frac{1}{T}\sum_{t=1}^{T} (h^*(x_t) - f^*(x_t))^2 - \|h^* - f^*\|| \leq O\Big(\sqrt{\frac{F^2}{T}\ln(4/\delta)}\Big). \end{aligned} \tag{15}$$

Now to prove the theorem, we proceed as follows:

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^{T} \|h^*(x_t) - f^*(x_t)\|^2 \\ &\leq \|h^* - f^*\| + O\Big(\sqrt{\frac{F^2}{T}\ln(4/\delta)}\Big) \\ &\leq (1 - \gamma)\|f^*\|^2 + O\Big(\sqrt{\frac{F^2}{T}\ln(4/\delta)}\Big) \\ &\leq (1 - \gamma)\frac{1}{T}\sum_{t=1}^{T} y_t^2 + (1 - \gamma)O\Big(\sqrt{\frac{F^2}{T}\ln(4/\delta)}\Big) + O\Big(\sqrt{\frac{F^2}{T}\ln(4/\delta)}\Big). \end{aligned} \tag{16}$$

Hence we get with probability at least $1-\delta$:

$$\sum_{t=1}^{T} \|h^*(x_t) - f^*(x_t)\|^2 \leq \sum_{t=1}^{T} \|y_t\|^2 + (2-\gamma)O\Big(\sqrt{T\ln(1/\delta)}\Big). \tag{17}$$

Set $R(T) = R_{\mathcal{A}}(T) + (2-\gamma)O\Big(\sqrt{T\ln(1/\delta)}\Big)$, we prove the proposition. □

B Proof of Theorem 5.1

An important property of λ-strong convexity that we will use later in the proof is that for any x and $x^* = \arg\min_x l(x)$, we have:

$$\|\nabla l(x)\|^2 \geq 2\lambda(l(x) - l(x^*)). \tag{18}$$

We prove Eqn. 18 below.

From the λ-strong convexity of $l(x)$, we have:

$$l(y) \geq l(x) + \nabla l(x)(y-x) + \frac{\lambda}{2}\|y-x\|^2. \tag{19}$$

Replace y by x^* in the above equation, we have:

$$\begin{aligned}
& l(x^*) \geq l(x) + \nabla l(x)(x^*-x) + \frac{\lambda}{2}\|x^*-x\|^2 \\
\Rightarrow & 2\lambda l(x^*) \geq 2\lambda l(x) + 2\lambda\nabla l(x)(x^*-x) + \lambda^2\|x^*-x\|^2 \\
\Rightarrow & -2\lambda\nabla l(x)(x^*-x) - \lambda^2\|x^*-x\|^2 \geq 2\lambda(l(x) - l(x^*)) \\
\Rightarrow & \|\nabla l(x)\|^2 - \|\nabla l(x)\|^2 - 2\lambda\nabla l(x)(x^*-x) - \lambda^2\|x^*-x\|^2 \geq 2\lambda(l(x) - l(x^*)) \\
\Rightarrow & \|\nabla l(x)\|^2 - \|\nabla l(x) + \lambda(x^*-x)\|^2 \geq 2\lambda(l(x) - l(x^*)) \\
\Rightarrow & \|\nabla l(x)\|^2 \geq 2\lambda(l(x) - l(x^*)).
\end{aligned} \tag{20}$$

B.1 Proofs for Lemma 4.2

Proof. Complete the square on the left hand side (LHS) of Eqn. 3, we have:

$$\sum \|y_t\|^2 - 2y_t^T h_t(x_t) + \|h_t(x_t)\|^2 \leq (1-\gamma)\sum_t \|y_t\|^2 + R(T). \tag{21}$$

Now let us cancel the $\sum y_t^2$ from both side of the above inequality, we have:

$$\sum -2y_t^T h_t(x_t) \leq \sum -2y_t^T h_t(x_t) + \|h_t(x_t)\|^2 \leq -\gamma\sum \|y_t\|^2 + R(T). \tag{22}$$

Rearrange, we have:

$$\sum 2y_t^T h_t(x_t) \geq \gamma\sum \|y_t\|^2 - R(T). \tag{23}$$

□

B.2 Proof of Theorem 5.1

We need another lemma for proving theorem 5.1:

Lemma B.1. *For each weak learner $\mathcal{A}_i$, we have:*

$$\sum_t \|h_t^i(x_t)\|^2 \leq (4-2\gamma)\sum_t \|\nabla\ell_t(y_t^{i-1})\|^2 + 2R(T). \tag{24}$$

Proof of Lemma B.1. For $\sum_t (h_t^i(x_t))^2$, we have:

$$\begin{aligned}
&\sum_t \|h_t^i(x_t)\|^2 = \sum_t \|h_t^i(x_t) - \nabla\ell_t(y_t^{i-1}) + \nabla\ell_t(y_t^{i-1})\|^2 \\
&\leq \sum_t \|h_t^i(x_t) - \nabla\ell_t(y_t^{i-1})\|^2 + \sum_t \|\nabla\ell_t y_t^{i-1}\|^2 + \sum_t 2(h_t^i(x_t) - \nabla\ell_t(y_t)^{i-1})^T\nabla\ell_t(y_t^{i-1}) \\
&\leq \sum_t 2\|h_t^i(x_t) - \nabla\ell_t(y_t^{i-1})\|^2 + \sum_t 2\|\nabla\ell_t(y_t^{i-1}\|^2 \\
&\leq 2(1-\gamma)\sum_t \|\nabla\ell_t(y_t^{i-1}\|^2 + 2R(T) + 2\sum_t \|\nabla\ell_t(y_t^{i-1}\|^2 \\
&\quad \textit{(By Weak Onling Learning Definition)} \\
&\leq (4-2\gamma)\sum_t \|\nabla\ell_t(y_t^{i-1}\|^2 + 2R(T). \qquad (25)
\end{aligned}$$

□

Proof of Theorem 5.1. For $1 \leq i \leq N$, let us define $\Delta_i = \sum_{t=1}^T (\ell_t(y_t^i) - \ell_t(f^*(x_t)))$. Following similar proof strategy as shown in (Beygelzimer et al., 2015a), we will link Δ_i to Δ_{i-1}. For Δ_i, we have:

$$\begin{aligned}
&\Delta_i = \sum_{t=1}^T (\ell_t(y_t^i) - \ell_t(f^*(x_t))) = \sum_t \ell_t(y_t^{i-1} - \eta h_t^i(x_t)) - \sum_t \ell_t(f^*(x_t)) \\
&\leq \sum_t \big[\ell_t(y_t^{i-1}) - \eta\nabla\ell_t(y_t^{i-1})^T h_t^i(x_t) + \frac{\beta\eta^2}{2}\|h_t^i(x_t)\|^2\big] - \sum_t \ell_t(f^*(x_t)) \\
&\quad \textit{(By } \beta\textit{-smoothness of } \ell_t\textit{)} \\
&\leq \sum_t \big[\ell_t(y_t^{i-1}) - \frac{\eta\gamma}{2}\|\nabla\ell_t(y_t^{i-1})\|^2 + \frac{\eta R(T)}{2} + \frac{\beta\eta^2}{2}\|h_t^i(x_t)\|^2\big] - \sum_t \ell_t(f^*(x_t)) \\
&\quad \textit{(By Lemma 4.2)} \\
&\leq \sum_t \big[\ell_t(y_t^{i-1}) - \frac{\eta\gamma}{2}\|\nabla\ell_t(y_t^{i-1})\|^2 + \frac{\eta R(T)}{2} + \beta\eta^2(2-\gamma)\|\nabla\ell_t(y_t^{i-1})\|^2 + \beta\eta^2 R(T) - \ell_t(f^*(x_t))\big] \\
&\quad \textit{(By Lemma B.1)} \\
&= \Delta_{i-1} - (\frac{\eta\gamma}{2} - \beta\eta^2(2-\gamma))\sum_t \|\nabla\ell_t(y_t^{i-1})\|^2 + (\frac{\eta}{2} + \beta\eta^2)R(T) \\
&\leq \Delta_{i-1} - (\eta\gamma\lambda - \beta\eta^2\lambda(4-2\gamma))\sum_t \big(\ell_t(y_t^{i-1}) - \ell_t(f^*(x_t))\big) + (\frac{\eta}{2} + \beta\eta^2)R(T) \\
&\quad \textit{(By Eqn. 18)} \\
&= \Delta_{i-1}\big[1 - (\eta\gamma\lambda - \beta\eta^2\lambda(4-2\gamma))\big] + (\frac{\eta}{2} + \beta\eta^2)R(T) \qquad (26)
\end{aligned}$$

Due to the setting of η, we know that $0 < (1 - (\eta\gamma\lambda - \beta\eta^2\lambda(4-2\gamma))) < 1$. For notation simplicity, let us first define $C = 1 - (\eta\gamma\lambda - \beta\eta^2\lambda(4-2\gamma))$. Starting from Δ_0, keep applying the relationship between Δ_i and Δ_{i-1} N times, we have:

$$\begin{aligned}
\Delta_N &= C^N\Delta_0 + (\frac{\eta}{2} + \beta\eta^2)R(T)\sum_{i=1}^N C^{i-1} \\
&= C^N\Delta_0 + (\frac{\eta}{2} + \beta\eta^2)R(T)\frac{1-C^N}{1-C} \\
&\leq C^N\Delta_0 + (\frac{\eta}{2} + \beta\eta^2)R(T)\frac{1}{1-C}.
\end{aligned}$$

Now divide both sides by T, and take T to infinity, we have:

$$\frac{1}{T}\Delta_N = C^N\frac{1}{T}\Delta_0 \leq C^N 2B, \qquad (27)$$

where we simply assume that $\ell_t(y) \in [-B, B], B \in \mathcal{R}^+$ for any t and y. Now let us go back to C, to minimize C, we can take the derivative of C with respect to η, set it to zero and solve for η, we will have:

$$\eta = \frac{\gamma}{\beta(8-4\gamma)}. \tag{28}$$

Substitute this η back to C, we have:

$$C = 1 - \frac{\gamma^2\lambda}{\beta(16-8\gamma)} \geq 1 - \frac{\lambda}{8\beta} \geq 1 - \frac{1}{8} = \frac{7}{8}. \tag{29}$$

Hence, we can see that there exist a $\eta = \frac{\gamma}{\beta(8-4\gamma)}$, such that:

$$\frac{1}{T}\Delta_N \leq 2B(1 - \frac{\gamma^2\lambda}{\beta(16-8\gamma)})^N \leq 2B(1 - \frac{\gamma^2\lambda}{16\beta})^N. \tag{30}$$

Hence we prove the first part of the theorem regarding the regret. For the second part of the theorem where ℓ_t and x_t are i.i.d sampled from a fixed distribution, we proceed as follows.

Let us take expectation on both sides of the inequality 30. The left hand side of inequality 30 becomes:

$$\begin{aligned}
&\frac{1}{T}\mathbb{E}\Delta_N = \mathbb{E}\frac{1}{T}\big[\sum_{t=1}^{T}(\ell_t(y_t^N) - \ell_t(f^*(x_t)))\big] = \frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T}\ell_t(-\mu\sum_{i=1}^{N}h_t^i(x_t))\big] - \frac{1}{T}\mathbb{E}_{(\ell_t,x_t)\sim D}[\ell_t(f^*(x_t))] \\
&= \frac{1}{T}\sum_{i=1}^{T}\mathbb{E}_t\big[\ell_t(-\mu\sum_{i=1}^{N}h_t^i(x_t))\big] - \mathbb{E}_{(\ell,x)\sim D}\ell(f^*(x)),
\end{aligned} \tag{31}$$

where the expectation is taken over the randomness of x_t and ℓ_t. Note that h_t^i only depends on $x_1, \ell_1, ..., x_{t-1}, \ell_{t-1}$. We also define $\mathbb{E}_t$ as the expectation over the randomness of x_t and ℓ_t at step t conditioned on $x_1, \ell_1, ..., x_{t-1}, \ell_{t-1}$. Since ℓ_t, x_t are sampled i.i.d from D, we can simply write $\mathbb{E}_t[\ell_t(-\mu\sum_{i=1}^{N}h_t^i(x_t))]$ as $\mathbb{E}_t[\ell(-\mu\sum_{i=1}^{N}h_t^i(x))]$. Now the above inequality can be simplied as:

$$\begin{aligned}
&\frac{1}{T}\mathbb{E}\Delta_N = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_t[\ell(-\mu\sum_{i=1}^{N}h_t^i(x))] - \mathbb{E}_{(\ell,x)\sim D}\ell(f^*(x)) \\
&\geq \mathbb{E}\big[\ell(-\mu\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}h_t^i(x))\big] - \mathbb{E}_{(\ell,x)\sim D}\ell(f^*(x)) \\
&= \mathbb{E}\big[\ell(-\mu\sum_{i=1}^{N}\bar{h}_i(x))\big] - \mathbb{E}_{(\ell,x)\sim D}\ell(f^*(x))
\end{aligned} \tag{32}$$

Now use the fact that $1/T\mathbb{E}\Delta_N \leq 2B(1 - \frac{\gamma^2\lambda}{16\beta})^N$, we prove the theorem. □

C Proof of Theorem 5.2

Lemma C.1. *In Alg. 2, if we assume the 2-norm of gradients of the loss w.r.t. partial sums by G (i.e., $\|\nabla_t^i\| = \|\nabla\ell_t(y_t^{i-1})\| \leq G$), and assume that each weak learner $\mathcal{A}_i$ has regret $R(T) = o(T)$, then we there exists a constant $c = \frac{1-\gamma+\sqrt{1-\gamma(1-\frac{R(T)}{TG^2})}}{\gamma} < \frac{2}{\gamma} - 1$ such that*

$$\sum_{t=1}^{T}\|\Delta_i^t\|^2 \leq c^2G^2T \quad \textit{and} \quad \sum_{t=1}^{T}\|h_i^t(x_t)\|^2 \leq (4-2\gamma)(1+c)^2G^2T + 2R(T) \leq 4c^2G^2T. \tag{33}$$

Proof. We prove the first inequality by induction on the weak learner index i. When $i = 0$, the claim is clearly true since $\Delta_0^t = 0$ for all t. Now we assume the claim is true for some $i \geq 0$, and prove it for $i+1$. We first note that by the inequality $\frac{1}{T}\sum_{t=1}^{T}a_t \leq \sqrt{\frac{\sum_t a_t^2}{T}}$ for all sequence $\{a_t\}_t$, we have

$$\frac{1}{T}(\sum_t\|\Delta_i^t\|)^2 \leq \sum_t\|\Delta_i^t\|^2 \leq c^2G^2T \tag{34}$$

$$\Rightarrow(\sum_t \|\Delta_i^t\|)^2 \leq c^2G^2T^2 \tag{35}$$

$$\Rightarrow \sum_t \|\Delta_i^t\| \leq cGT \tag{36}$$

Then by the assumption that weak learner $\mathcal{A}_i$ has an edge γ with regret $R(T)$, we have from step 14 of Alg. 2:

$$\sum_t \|\Delta_{i+1}^t\|^2 = \sum_t \|\Delta_i^t + \nabla_{i+1}^t - h_{i+1}^t(x_t)\|^2 \leq (1-\gamma)\sum_t \|\Delta_i^t + \nabla_{i+1}^t\|^2 + R(T) \tag{37}$$

$$\leq (1-\gamma)\sum_t \left(\|\Delta_i^t\| + G\right)^2 + R(T) \tag{38}$$

$$\leq (1-\gamma)\left(\sum_t \|\Delta_i^t\|^2 + 2G\sum_t \|\Delta_i^t\| + G^2T\right) + R(T) \tag{39}$$

$$\leq (1-\gamma)(1+c)^2G^2T + R(T) \tag{40}$$

$$= c^2G^2T \tag{41}$$

We have the last equality because c is chosen as the positive root of the quadratic equation: $\gamma c^2 + (2\gamma - 2)c + (\gamma - 1 - \frac{R(T)}{TG^2}) = 0$, which is equivalent to $c^2G^2T = (1-\gamma)(c+1)^2G^2T + R(T)$.

The second inequality of the lemma can be derived from a similar argument of Lemma B.1 by expanding $\| \left(\Delta_{i-1}^t + \nabla_i^t - h_i^t(x_t)\right) - \left(\Delta_{i-1}^t + \nabla_i^t\right) \|^2$ and then applying edge assumption. □

We now use the above lemma to prove the performance guarantee of Alg. 2 as follows.

Proof of Theorem 5.2. We first define the intermediate predictors as: $f_0^t(x) := h_0(x)$, $\hat{f}_i^t(x) := f^{t-1}(x) - \eta_i h_i^t(x)$, and $f_i^t(x) := P(\hat{f}_i^t(x))$. Then for all $i = 1, ..., N$ we have:

$$\|f_i^t(x_t) - f^*(x_t)\|^2 \leq \|\hat{f}_i^t(x_t) - f^*(x_t)\|^2 = \|f_{i-1}^t(x_t) - \eta_i h_i^t(x_t) - f^*(x_t)\|^2 \tag{42}$$

$$= \|f_{i-1}^t(x_t) - f^*(x_t)\|^2 + \eta_i^2\|h_i^t(x_t)\|^2 - 2\eta_i\langle f_{i-1}^t(x_t) - f^*(x_t), h_i^t(x_t) - \Delta_{i-1}^t - \nabla_i^t\rangle$$
$$- 2\eta_i\langle f_{i-1}^t(x_t) - f^*(x_t), \Delta_{i-1}^t + \nabla_i^t\rangle \tag{43}$$

Rearanging terms we have:

$$\langle f^*(x_t) - f_{i-1}^t(x_t), \nabla_i^t\rangle \tag{44}$$

$$\geq \frac{1}{2\eta_i}\|f_i^t(x_t) - f^*(x_t)\|^2 - \frac{1}{2\eta_i}\|f_{i-1}^t(x_t) - f^*(x_t)\|^2 - \frac{\eta_i}{2}\|h_i^t(x_t)\|^2$$
$$- \langle f^*(x_t) - f_{i-1}^t(x_t), h_i^t(x_t) - \Delta_{i-1}^t - \nabla_i^t\rangle - \langle f^*(x_t) - f_{i-1}^t(x_t), \Delta_{i-1}^t\rangle \tag{45}$$

Using λ-strongly convex of ℓ_t and applying the above equality and $\Delta_i^t = \Delta_{i-1}^t + \nabla_i^t - h_i^t(x_t)$, we have:

$$\ell_t(f^*(x_t)) \geq \ell_t(f_{i-1}^t(x_t)) + \langle f^*(x_t) - f_{i-1}^t(x_t), \nabla_i^t\rangle + \frac{\lambda}{2}\|f^*(x_t) - f_{i-1}^t(x_t)\|^2 \tag{46}$$

$$\geq \ell_t(f_{i-1}^t(x_t)) + \frac{1}{2\eta_i}\|f_i^t(x_t) - f^*(x_t)\|^2 - \frac{1}{2\eta_i}\|f_{i-1}^t(x_t) - f^*(x_t)\|^2 - \frac{\eta_i}{2}\|h_i^t(x_t)\|^2$$
$$+ \langle f^*(x_t) - f_{i-1}^t(x_t), \Delta_i^t\rangle - \langle f^*(x_t) - f_{i-1}^t(x_t), \Delta_{i-1}^t\rangle + \frac{\lambda}{2}\|f^*(x_t) - f_{i-1}^t(x_t)\|^2 \tag{47}$$

Summing over $t = 1, ..., T$ and $i = 1, ..., N$ we have:

$$N\sum_{t=1}^T \ell_t(f^*(x_t))$$

$$\geq \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\ell_t(f_{i-1}^t(x_t)) + \left\langle f^*(x_t) - f_{i-1}^t(x_t), \nabla_i^t\right\rangle + \frac{\lambda}{2}\|f^*(x_t) - f_{i-1}^t(x_t)\|^2\right] \tag{48}$$

$$= \sum_{i=1}^{N}\sum_{t=1}^{T}\ell_t(f_{i-1}^t(x_t)) - \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\eta_i}{2}\|h_i^t(x_t)\|^2$$
$$+ \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{2\eta_i}\|f_i^t(x_t) - f^*(x_t)\|^2 - \sum_{i=1}^{N}\sum_{t=1}^{T}(\frac{1}{2\eta_i} - \frac{\lambda}{2})\|f_{i-1}^t(x_t) - f^*(x_t)\|^2$$
$$+ \sum_{i=1}^{N}\sum_{t=1}^{T}\left\langle f^*(x_t) - f_{i-1}^t(x_t), \Delta_i^t\right\rangle - \sum_{i=1}^{N}\sum_{t=1}^{T}\left\langle f^*(x_t) - f_{i-1}^t(x_t), \Delta_{i-1}^t\right\rangle \tag{49}$$

$$= \sum_{i=1}^{N}\sum_{t=1}^{T}\ell_t(f_{i-1}^t(x_t)) - \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\eta_i}{2}\|h_i^t(x_t)\|^2$$
$$+ \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{2\eta_i}\|f_i^t(x_t) - f^*(x_t)\|^2 - \sum_{i=0}^{N-1}\sum_{t=1}^{T}(\frac{1}{2\eta_{i+1}} - \frac{\lambda}{2})\|f_i^t(x_t) - f^*(x_t)\|^2$$
$$+ \sum_{i=1}^{N}\sum_{t=1}^{T}\left\langle f^*(x_t) - f_{i-1}^t(x_t), \Delta_i^t\right\rangle - \sum_{i=1}^{N-1}\sum_{t=1}^{T}\left\langle f^*(x_t) - (f_{i-1}^t(x_t) - \eta_i h_i^t(x_t)), \Delta_i^t\right\rangle$$
$$- \sum_{t=1}^{T}\left\langle f^*(x_t) - f_0^t(x_t), \Delta_0^t\right\rangle \qquad \text{(We switched index and apply } \Delta_0^t = 0 \text{ next.)} \tag{50}$$

$$= \sum_{i=1}^{N}\sum_{t=1}^{T}\ell_t(f_{i-1}^t(x_t)) - \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\eta_i}{2}\|h_i^t(x_t)\|^2 - \sum_{i=1}^{N-1}\sum_{t=1}^{T}\left\langle \eta_i h_i^t(x_t), \Delta_i^t\right\rangle$$
$$+ \sum_{i=1}^{N-1}\sum_{t=1}^{T}\frac{1}{2}\|f_i^t(x_t) - f^*(x_t)\|^2(\frac{1}{\eta_i} - \frac{1}{\eta_{i+1}} + \lambda) - \sum_{t=1}^{T}(\frac{1}{2\eta_1} - \frac{\lambda}{2})\|f_0^t(x_t) - f^*(x_t)\|^2$$
$$+ \sum_{t=1}^{T}\left[\left\langle f^*(x_t) - f_{N-1}^t(x_t), \Delta_N^t\right\rangle + \frac{1}{2\eta_N}\|f_{N-1}^t(x_t) - \eta_N h_N^t(x_t) - f^*(x_t)\|^2\right] \tag{51}$$

(We next apply $\eta_i = \frac{1}{\lambda i}$ and complete the squares for the last sum.)

$$= \sum_{i=1}^{N}\sum_{t=1}^{T}\ell_t(f_{i-1}^t(x_t)) - \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\eta_i}{2}\|h_i^t(x_t)\|^2 - \sum_{i=1}^{N-1}\sum_{t=1}^{T}\left\langle \eta_i h_i^t(x_t), \Delta_i^t\right\rangle$$
$$+ \frac{1}{2\eta_N}\sum_{t=1}^{T}\|\left(f_{N-1}^t(x_t) - f^*(x_t)\right) + \eta_N(\Delta_N^t - h_N^t(x_t))\|^2$$
$$- \frac{\eta_N}{2}\sum_{t=1}^{T}\left(\|\Delta_N^t - h_N^t(x_t)\|^2 - \|h_N^t(x_t)\|^2\right) \tag{52}$$

(We next drop the completed square, and apply Cauchy-Schwarz)

$$\geq \sum_{i=1}^{N}\sum_{t=1}^{T}\ell_t(f_{i-1}^t(x_t)) - \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\eta_i}{2}\|h_i^t(x_t)\|^2 - \sum_{i=1}^{N}\eta_i\sum_{t=1}^{T}\|h_i^t(x_t)\|\|\Delta_i^t\| - \frac{\eta_N}{2}\sum_{t=1}^{T}\|\Delta_N^t\|^2 \tag{53}$$

(We next apply Cauchy-Schwarz again.)

$$\geq \sum_{i=1}^{N}\sum_{t=1}^{T}\ell_t(f_{i-1}^t(x_t)) - \sum_{i=1}^{N}\frac{\eta_i}{2}\sum_{t=1}^{T}\|h_i^t(x_t)\|^2 - \frac{\eta_N}{2}\sum_{t=1}^{T}\|\Delta_N^t\|^2$$
$$- \sum_{i=1}^{N}\eta_i\sqrt{\sum_{t=1}^{T}\|h_i^t(x_t)\|^2\sum_{t=1}^{T}\|\Delta_i^t\|^2} \tag{54}$$

Now we apply Lemma C.1 and replace the remaining $\eta_i = \frac{1}{\lambda i}$. Using $\sum_{i=1}^{N} \frac{1}{i} \leq 1 + \ln N$, we have:

$$
\begin{aligned}
& N \sum_{t=1}^{T} \ell_t(f^*(x_t)) \\
\geq & \sum_{i=1}^{N} \sum_{t=1}^{T} \ell_t(f_{i-1}^t(x_t)) - \sum_{i=1}^{N} \frac{1}{2i\lambda} 4c^2G^2T - \frac{1}{2N\lambda} c^2G^2T - \sum_{i=1}^{N} \frac{1}{i\lambda} 2c^2G^2T \qquad (55) \\
\geq & \sum_{i=1}^{N} \sum_{t=1}^{T} \ell_t(f_{i-1}^t(x_t)) - \frac{4c^2G^2T}{\lambda}(1 + \ln N) - \frac{c^2G^2T}{2N\lambda} \qquad (56)
\end{aligned}
$$

Dividing both sides by NT and rearrange terms, we get:

$$
\frac{1}{TN} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\ell_t(y_t^i) - \ell_t(f^*(x_t)) \right] \leq \frac{4c^2G^2}{N\lambda}(1 + \ln N) + \frac{c^2G^2}{2N^2\lambda}.
$$

Using Jensen's inequality for the LHS of the above inequality, we get:

$$
\frac{1}{T} \sum_{t=1}^{T} \ell_t(\frac{1}{N} \sum_{i=1}^{N} y_t^i) - \ell_t(f^*(x_t)) \leq \frac{4c^2G^2}{N\lambda}(1 + \ln N) + \frac{c^2G^2}{2N^2\lambda},
$$

which proves the first part of the theorem.

For stochastic setting, we can prove it by using similar proof techniques (e.g., take expectation on both sides of Eqn. 57 and use Jensen inequality) that we used for proving theorem 5.1. □

D Counter Example for Alg. 1

In this section, we provide an counter example where we show that Alg. 1 cannot guarantee to work for non-smooth loss. We set $y \in \mathbb{R}^2$, and design a loss function $\ell_t(y) = 2|y_{[1]}| + |y_{[2]}|$, where $y_{[i]}$ stands for the i'th entry of the vector y, for all time step t. The subgradient of this non-smooth loss is $[2, 1]^T$, or $[2, -1]^T$, or $[-2, 1]^T$, or $[-2, -1]^T$, depending on the position of y. We restricted the weak hypothesis class $\mathcal{H}$ to consist of only two types of hypothesis: hypothesis $h(x) = [\alpha, 0]^T$, or hypothesis $h(x) = [0, \alpha]^T$, where $\alpha \in [-2, 2]$. We can show that given a sequence of training examples $\{(x_\tau, g_\tau)\}_{\tau=1}^t$, where g_t is the one of the gradient from the total four possible subgradient of ℓ_t, the hypothesis that minimizes the accumulated square loss $\sum_{\tau=1}^t (h(x_\tau) - g_\tau)^2$ is going to be the type of $h(x) = [\alpha, 0]^T$.

Now we consider using Follow the Leader (FTL) as a no-regret online learning algorithm for each weak learner. Based on the above analysis, we know that no matter what the sequence of training examples each weak learner has received as far, the weak leaners always choose the hypothesis with type $h(x) = [\alpha, 0]^T$ from $\mathcal{H}$. So, for every time step t, if we initialize $y_t^0 = [a, b]^T$, where $a > 0$ and $b > 0$, then the output y_t^N (computed from Line 8 in Alg.1) always have the form of $y_t^N = [\eta, b]$, where $\eta \in \mathbb{R}$. Namely, all weak learners' prediction only moves y_t horizontally and it will never be moved vertically. But note that the optimal solution is located at $[0, 0]^T$. Since for all t, $y_{t_{[2]}}^N$ is also b constant away from 0, the total regret accumulates linearly as bT, regardless of how many weak learners we have.

E Details of Implementation

E.1 Binary Classification

For binary classification, following (Friedman, 2001), let us define feature $x \in \mathbb{R}^n$, label $u \in \{-1, 1\}$. With x_t and u_t, the loss function ℓ_t is defined as:

$$
\ell_t(y) = \ln(1 + \exp(-u_t y)) + \lambda y^2. \qquad (57)
$$

where $y \in \mathbb{R}$. In this setting, we have $\mathcal{H} : \mathbb{R}^n \to \mathbb{R}$. The regularization is to avoid overfitting: we can set $y = \infty * sign(u_t)$ to make the loss close to zero.

The loss function $\ell_t(y)$ is twice differentiable with respect to y, and the second derivative is:

$$\nabla^2 \ell_t(y) = \frac{\exp(u_t y)}{(1+\exp(u_t y))^2} \tag{58}$$

Note that we have:

$$\nabla^2 \ell_t(y) \leq \frac{1}{1/\exp(u_t y) + 2 + \exp(u_t y)} \leq \frac{1}{4}. \tag{59}$$

Hence, $\ell_t(y)$ is $1/4$-smooth.

Under the assumption that the output from hypothesis from $\mathcal{H}$ is bounded as $|y| \leq Y \in \mathbb{R}^+$, we also have:

$$\nabla^2 \ell_t(y) \geq \frac{1}{2+2\exp(Y)} \tag{60}$$

Hence, with boundness assumption, we can see that $\ell_t(y)$ is $1/(2+2\exp(Y))$-strongly convex and $(1/4)$-smooth.

The another loss we tried is the hinge loss:

$$\ell_t(y) = \max(0, 1 - u_t y) + \lambda y^2. \tag{61}$$

With the regularization, the loss $\ell_t(y)$ is still strongly convex, but no longer smooth.

E.2 Multi-class Classification

Follow the settings in (Friedman, 2001), for multi-class classification problem, let us define feature $x \in \mathbb{R}^n$, and label information $u \in \mathbb{R}^k$, as a one-hot representation, where $u[i] = 1$ ($u[i]$ is the i-th element of u), if the example is labelled by i, and $u[i] = 0$ otherwise. The loss function ℓ_t is defined as:

$$\ell_t(y) = -\sum_{i=1}^{k} u_t[i] \ln \frac{\exp(y[i])}{\sum_{j=1}^{k} \exp(y[j])}, \tag{62}$$

where $y \in \mathbb{R}^k$. In this setting, we let weak learner i pick hypothesis h from $\mathcal{H}$ that takes feature x_t as input, and output $\hat{y}_i \in \mathbb{R}^k$. The online boosting algorithm then linearly combines the weak learners' prediction to predict y.