This work:
- A formal setting of online boosting for regression.
- An online version of gradient boosting, showing that online boosting is missing.
- Exponentially faster convergence for convex hull of $F$.

1. Motivation and Main Results

Boosting:
- well-known ensemble learning method that combines rules of thumb using a weak learning algorithm.
- extensively studied in the batch setting.

However, online learning setting is gaining more and more attention.

3. Batch Boosting vs Online Boosting

Batch Boosting
Given a batch of $T$ examples, $(x_i, y_i)$ for $t = 1, \ldots, T$.
Learner predicts $\hat{y}_i$ for example $x_i$.

Weak learner (a.k.a. ERM):
\[
\sum_{i=1}^{T} (y_i - \hat{y}_i)^2 \leq \int_{F} \sum_{i=1}^{T} (y_i - f(x_i))^2
\]


Strong learner: For any $f \in \mathcal{F}$,
\[
\sum_{i=1}^{T} (y_i - \hat{y}_i)^2 \leq \int_{F} \sum_{i=1}^{T} (y_i - f(x_i))^2 + \Delta_t
\]
\[\Delta_t \rightarrow 0 \text{ as } N \rightarrow \infty.\]

Online Boosting
Given a sequence of $T$ examples, $(x_i, y_i)$ for $t = 1, \ldots, T$.
Learner predicts $\hat{y}_i$ for example $x_i$ before observing $y_i$.

Weak online learner (a.k.a. online ERM):
\[
\sum_{i=1}^{T} (y_i - \hat{y}_i)^2 \leq \int_{F} \sum_{i=1}^{T} (y_i - f(x_i))^2 + R(T)
\]
\[R(T) \rightarrow 0 \text{ as } N \rightarrow \infty \text{ and } T \rightarrow \infty.\]

4. Gradient Boosting vs Online Gradient Boosting

Gradient Boosting
Input: batch of examples $(x_i, y_i)$ for $t = 1, \ldots, T$.
step-size $\eta$, number of weak learners $N$.
Set pseudo-labels $\tilde{y}_i = y_i$ for all $i$.

For $i = 1, 2, \ldots, N$:
- Pass training examples $(x_i, \tilde{y}_i, \ldots, x_T, \tilde{y}_T)$ to weak learner $f$ with step-size $\eta$.
- Obtain prediction $\hat{y}_i$ for example $x_i$.
- Compute new pseudo-label $\tilde{y}_i \leftarrow \tilde{y}_i - \eta f(x_i)$.
- Obtain final prediction $\hat{y}_i = \sum_{j=1}^{N} \tilde{y}_j^j$ for all $i$.

Online Gradient Boosting
Input: sequence of examples $(x_i, y_i)$ for $t = 1, \ldots, T$.
step-size $\eta$, number of weak learners $N$.
Initialize $\sigma^t = 0$ for all $t$.

For $t = 1, 2, \ldots, T$:
- Obtain predictions of weak learners $\hat{y}_i$ for all $i$.
- Set $\tilde{y}_i = 0$.
- For $i = 1, 2, \ldots, N$:
  - $\hat{y}_i \leftarrow (1 - \sigma^t)\hat{y}_i + \sigma^t y_i$.
  - Predict $\hat{y}_i$ and see true label $y_i$.
- Set pseudo-label $\tilde{y}_i = \hat{y}_i$.
- For $i = 1, 2, \ldots, N$:
  - Pass $(x_i, \tilde{y}_i)$ to weak learner $f$ with step-size $\eta$.
  - Compute new pseudo-label $\tilde{y}_i \leftarrow (1 - \sigma^t)\tilde{y}_i + \sigma^t y_i - y_i$.
- Update $\sigma^t \in [0, \eta]$ using online gradient descent.

5. Theoretical Guarantees

Regret: For any $f \in \mathcal{F}$,
\[
R(T) \leq \left(1 - \frac{\eta}{N}\right)\Delta_0 + O(\|\mathcal{F}_t \cdot (\eta T + R(T) + \sqrt{T})\|)
\]
where $\Delta_0$ is the initial error.
Choosing $\eta = \frac{\Delta_0}{N^2}$, we get $R(T) \rightarrow 0$ as $N \rightarrow \infty$ and $T \rightarrow \infty$.

Lower bound: for any online boosting alg, $R(T) \geq \Omega(\sqrt{T})$ for some $T$ in convex hull of $F$.

For batch setting: our algorithmic technique gives exponentially faster convergence compared to [Zhang and Yu, 2005].

6. Experiments

Setup:
- Implemented within Vowpal Wabbit.
- 14 publicly available data sets.
- Parameters $\eta$ and $N$ tuned via progressive validation.
- Base learners: Default (of VW), Neural Networks, Regression stumps.

7. References