Boosting for Online Convex Optimization

Elad Hazan Karan Singh

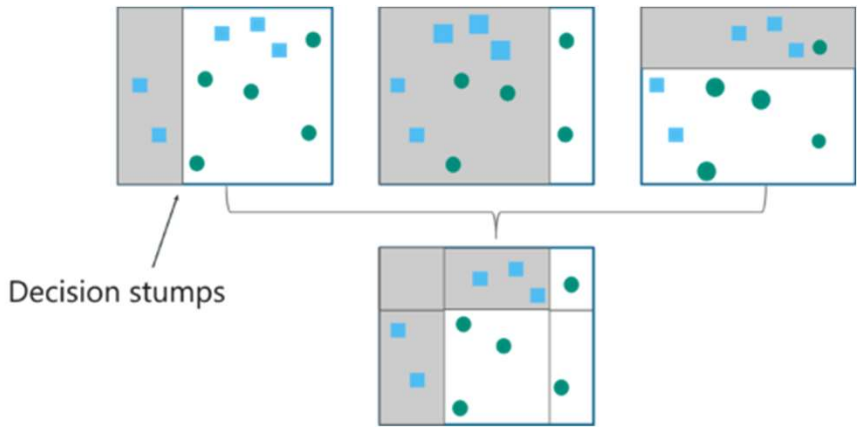
ICML | 2021

PRINCETON UNIVERSITY

Microsoft Research

Boosting

+ Computational Model of Compositional Learning
+ Combine (inaccurate, simple, comp. cheap) **weak** learners
+ To produce (accurate, expressive) **strong** learners



From Edureka

See context c_t
Choose decision $x_t \in K$.
Suffer loss $l_t(x_t)$

Gradient Boosting

Weak Learner
As good as the best
$h \in \mathbb{H}$

Strong Learner
As good as the best
$h \in \boldsymbol{Conv}\,(\mathbb{H})$.

Aim: Enhance Expressivity
(Convex Nonlinear Loss)

IID [MBBF99,ZY05]
Online [BHKL 15,BH20]

Doesn't deal with **approximate** weak learners.

Mostly for special cases of **linear loss**. (specific decision sets)

This Work
Enhance accuracy and expressivity;
Convex losses;
General decision sets.

Classical Boosting

Weak Learner
Slightly better than random guess

Strong Learner
As good as the best
$h \in \mathbb{H}$.

Aim: Increase Accuracy

IID Binary [S90,FS97], Multiclass [FS97,MS11], Multilabel [ADS07]
Online Binary [CLL12,BKL15, BCHM20], Multiclass [JGT 17], Ranking [JT18]

Online Weak Learner

For t = 1..T
- Observe context c_t
- Choose decision $x_t \in K$.
- Suffer loss $l_t(x_t)$

For any sequence of context and linear loss functions, an (*Agnostic*) Weak Learner (WL) ensures

Loss of WL $\leq \gamma \min_{h \in \mathbb{H}}$ Loss of Hypothesis h $+ (1-\gamma)$ Loss of Random Predictor $+ R_W(T)$

$\mathbb{H} \rightarrow \boldsymbol{Conv}(\mathbb{H})$
$\gamma \rightarrow 1$
Linear → Convex

For any sequence of contexts and convex losses,

Loss of Alg $\leq \min_{h \in Conv(\mathbb{H})}$ Loss of Hypothesis h $+ \mathrm{Regret}(T)$

Boosting Guarantee
(Wanted)

Main Result (for Boosting OCO)

An efficient boosting algorithm

Makes N calls to the weak learning algorithm

$$\texttt{Regret}(T) \leq \frac{T}{\gamma\sqrt{N}} + \frac{R_W(T)}{\gamma}$$

Sublinear as long as WL generalizes and $N \geq o(1)$

Running time scales linearly in N

Preview: *What feedback to serve the WL?*

Gradient Boosting says "Take the gradient of the residual loss."

Central Realization: *This is insufficient.*

Instead