Online Learning with Pairwise Loss Functions

MLSIG Seminar Series, Dept. of CSA, IISc

Joint work with *B. Sriperumbudur, P. Jain, H. Karnick*

Purushottam Kar

MLO Group, Microsoft Research India

Outline

A quick introduction to online learning

Examples of pairwise loss functions

An online learning model+algo for pairwise functions

Outline

- A quick introduction to online learning
  - Notion of regret
  - Generalization error
- Examples of pairwise loss functions
- An online learning model+algo for pairwise functions

Credit Card Fraud Detection

Transaction 1

- Guess ✓
- Truth ✓
- Loss 0

Transaction 2

- Guess ✗
- Truth ✓
- Loss 1

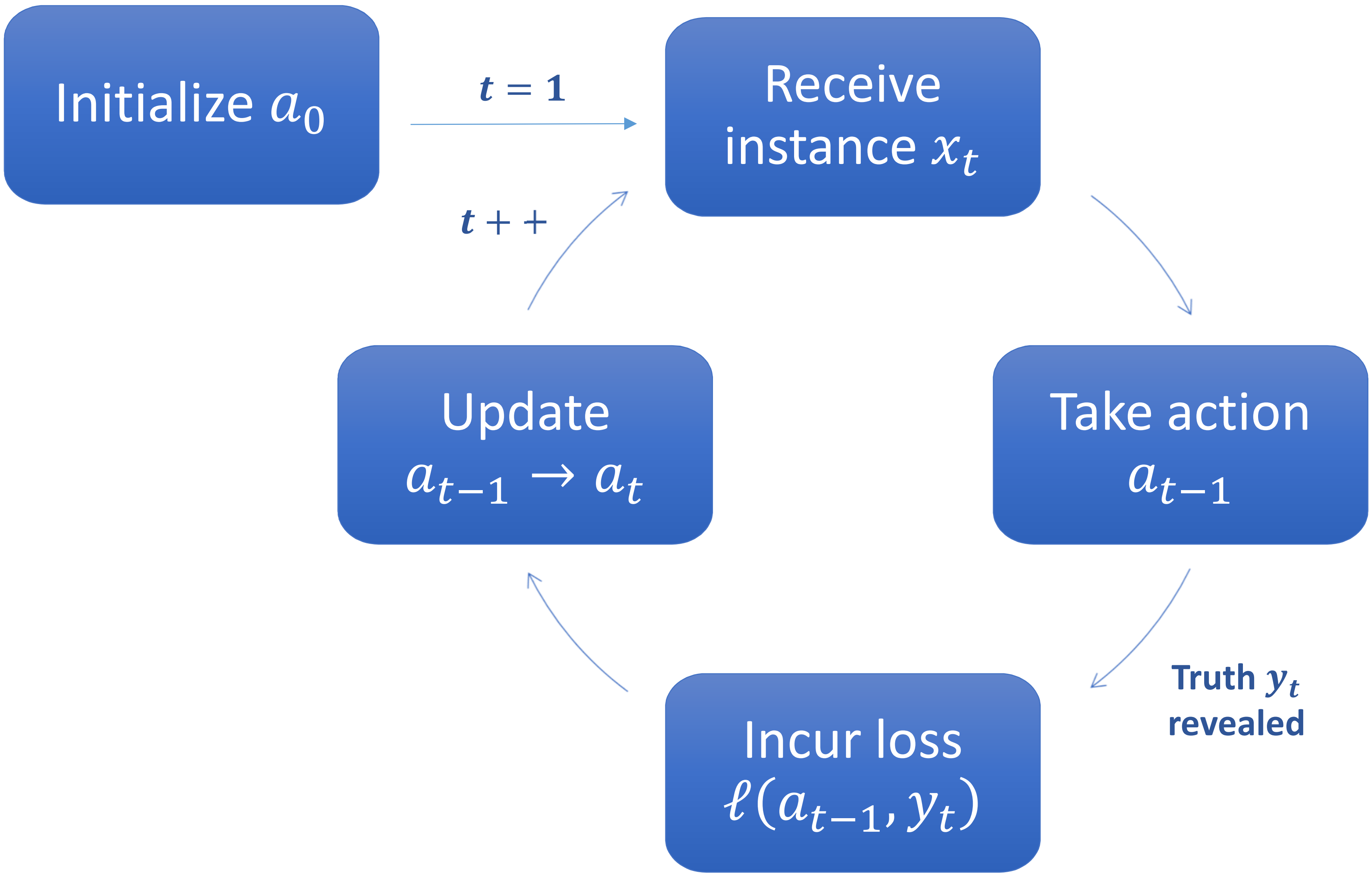
Transaction 3

- Guess ✗
- Truth ✗
- Loss 0

Transaction 4

- Guess ✓
- Truth ✓
- Loss 0

The Online Learning Process



Benefits of Online Learning

- Don't have to wait for all data to arrive
  - Streaming data, Transactional data

- Applications to large scale learning
  - Data too large to fit in memory (or even disk)
  - Solution: stream data into memory from disk or network

- Fast learning
  - Several online learning algorithms have **cheap** updates
$$a_{t-1} \rightarrow a_t$$
  - Online gradient descent, Mirror descent

Example: Online Classification

- Instances are vector-label pairs $z_t = (x_t, y_t)$
  - $x_t \in \mathbb{R}^d, \mathrm{y_t} \in \{-1, +1\}$
- Actions are classifiers e.g. $a_t = \langle w_t, x \rangle, w_t \in \mathcal{W}$
- Loss is the hinge loss function

$$\ell(w_{t-1}, z_t) = [1 - y_t \cdot \langle w_{t-1}, x_t \rangle]_+$$

- Total loss incurred by adaptive classfn $\sum_{t=1}^{T} \ell(w_{t-1}, z_t)$
- Loss of single best classifier $\min_{w \in \mathcal{W}} \sum_{t=1}^{T} \ell(w, z_t)$
  - This is what a "batch" learning algorithm would have given
- The online process suffers
  - Unable to see all data in one go

Regret and Generalization

- Regret: how much the online process suffers

$$\mathfrak{R}_T = \sum_t^T \ell(a_{t-1}, z_t) - \min_{a \in \mathcal{A}} \sum_t^T \ell(a, z_t)$$

- Online learning can compete with batch learning
  - Excess training error $\frac{1}{\mathrm{T}} \mathfrak{R}_T \downarrow 0$ if $\mathfrak{R}_T = o(T)$
  - Performance on unseen points: $\mathcal{L}(a) = \mathop{\mathbb{E}}_{z \sim \mathcal{Z}} \ell(a, z)$
- **Online-to-batch conversion**: For random x_t, convex ℓ

$$\mathcal{L}(\bar{a}) \leq \inf_{a \in \mathcal{A}} \mathcal{L}(a) + \frac{1}{T} \mathfrak{R}_T + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$

where $\bar{a} = \frac{1}{T} \sum a_t$

Outline

- A quick introduction to online learning
  - Notion of regret
  - Generalization error
- Examples of pairwise loss functions
  - Algorithmic challenges
  - Learning theoretic challenges
- An online learning model+algo for pairwise functions

Pointwise Loss Functions

- Loss functions for classification, regression …

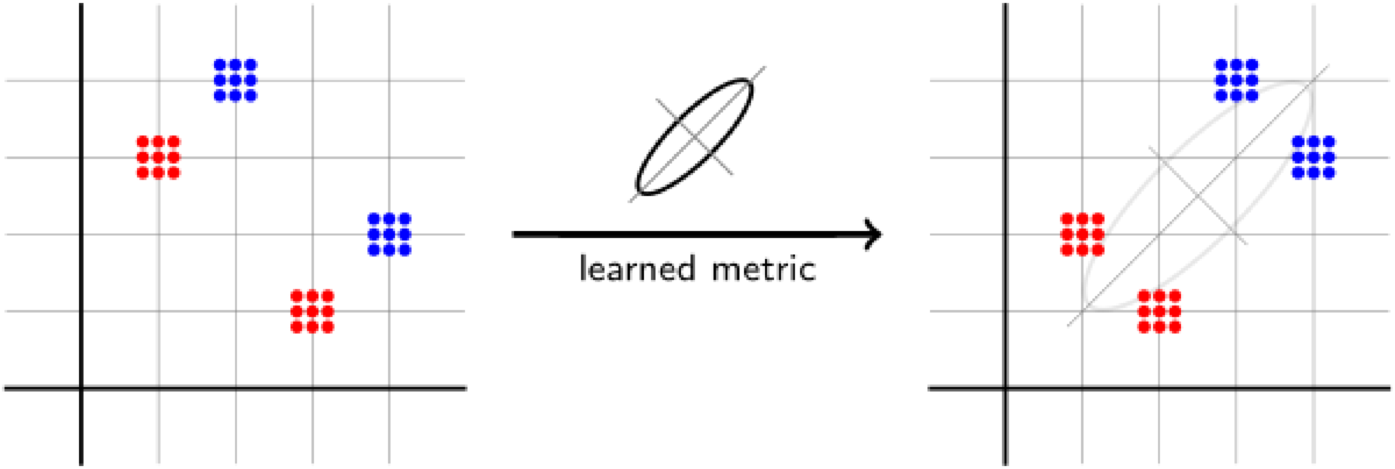
$$\boxed{\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}}$$

- … look at the performance of function at **one** point

Examples

- Hinge loss: $\ell(w, z) = [1 - y \cdot \langle w, x \rangle]_+$
- Logistic loss: $\ell(w, z) = \ln(1 + \exp(y \cdot \langle w, x \rangle))$
- Squared loss: $\ell(w, z) = (y - \langle w, x \rangle)^2$

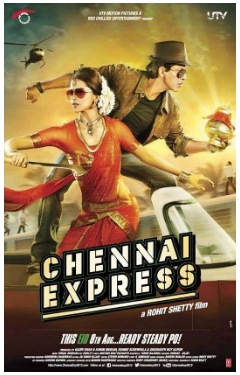
Metric Learning for Classification

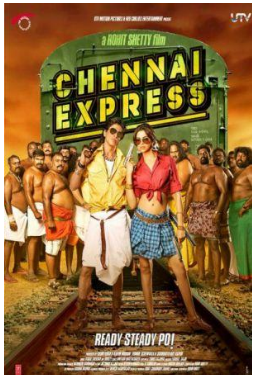


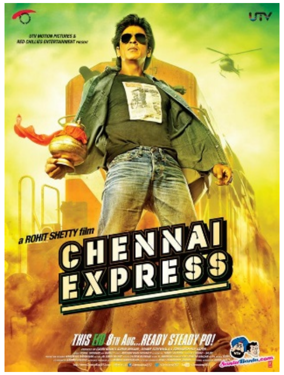
- Penalize metric for bringing **blue** and **red** points close
- Loss function needs to consider two points at a time!
  - … in other words a **pairwise loss function**
- Example: $\ell(M, z_1, z_2) = \begin{cases} 1, y_1 \neq y_2 \text{ and } d_M(x_1, x_2) < \gamma_1 \\ 1, y_1 = y_2 \text{ and } d_M(x_1, x_2) > \gamma_2 \\ 0, \text{otherwise} \end{cases}$

Bipartite Ranking

Chennai Express | Search







- Want relevant results to be ranked above others
- Penalize scoring function $s\colon \mathcal{Z} \to \mathbb{R}$ for each "switch"

s() > s()

- $\ell(s, z_1, z_2) = 1 \text{ iff } r(z_1) > r(z_2) \text{ and } s(z_1) < s(z_2)$

Images taken from cinemahood.com, sify.com, santabanta.com and thehindu.com

Pairwise Loss Functions

$$\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$$

Examples:

- Mahalanobis metric learning
- Bipartite ranking
- Preference learning
- Two-stage multiple kernel learning
- Indefinite kernel learning

Learning with Pairwise Loss Functions

$$\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$$

Algorithmic challenges:

- Training data available as a set $\mathcal{T} = \{z_1, z_2, \ldots, z_T\}$
- Question: **how to create pairs**?
- Solution 1: $\min_{w \in \mathcal{W}} \frac{2}{T(T-1)} \sum_{i<j} \ell(w, z_i, z_j)$
  - Expensive for $T \gg 1$
- Solution 2: Use online techniques for a batch solver
  - Challenge: **Online creation of pairs** from a data stream
  - Desirable: Memory efficiency

Learning with Pairwise Loss Functions

$$\boxed{\ell: \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}}$$

Learning theoretic challenges:

- Batch learning methods: learn from pairs (z_i, z_j)
  - Intersection between pairs: **training data not i.i.d.**
  - Direct application of concentration inequalities not possible
- Online learning methods: let $\{z_i\}$ arrive in a stream
  - Need an appropriate **notion of regret**
  - Classical OTB proofs require i.i.d. data crucially

This talk: mostly algorithmic solutions + hint of theory

Outline

- A quick introduction to online learning
  - Notion of regret
  - Generalization error
- Examples of pairwise loss functions
  - Algorithmic challenges
  - Learning theoretic challenges
- An online learning model+algo for pairwise functions
  - A memory efficient online learning algo
  - Regret and generalization bounds

An Online Learning Model for Pairwise Losses

$$\boxed{\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}}$$

- At each time step t
  - We propose an action a_t (e.g. a scoring function or a metric)
  - We receive **a single point** $z_t = (x_t, y_t)$
- We incur loss ℓ_t on action a_{t-1}
  - **Buffer** B $[z_1, z_2, z_3, \dots]$
  - Pair up z_t with points in buffer (z_t, z_1) (z_t, z_2) … (z_t, z_{t-1})
  - Incur loss

$$\ell_t^\infty(a_{t-1}) = \frac{1}{t-1}\big(\ell(a_{t-1}, z_t, z_1) + \cdots + \ell(a_{t-1}, z_t, z_{t-1})\big)$$

An Online Learning Model for Pairwise Losses

$$\boxed{\ell: \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}}$$

- At each time step t
  - We propose an action a_t (e.g. a scoring function or a metric)
  - We receive **a single point** $z_t = (x_t, y_t)$
- We incur loss ℓ_t on action a_{t-1}
  - **Finite Buffer B** $[\square_1, \square_2, \dots, \square_s]$
  - Pair up z_t with points in buffer (z_t, z_{i_1}) (z_t, z_{i_2}) … (z_t, z_{i_s})
  - Incur loss

$$\ell_t^{\text{buf}}(a_{t-1}) = \frac{1}{s}\left(\ell(a_{t-1}, z_t, z_{i_1}) + \cdots + \ell(a_{t-1}, z_t, z_{i_s})\right)$$

An Online Learning Model for Pairwise Losses

$$\boxed{\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}}$$

Notions of Regret in this Model

- How well are we able to do on **pairs that we have seen**
  - **Finite buffer regret**

$$\mathfrak{R}_T^{\text{buf}} = \sum_{t=1}^{T} \ell_t^{\text{buf}}(a_{t-1}) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell_t^{\text{buf}}(a)$$

- How well are we able to do on **all possible pairs**
  - **All pairs regret**

$$\mathfrak{R}_T^{\infty} = \sum_{t=1}^{T} \ell_t^{\infty}(a_{t-1}) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell_t^{\infty}(a)$$

An Online Learning Algorithm for Pairwise Losses

$$\boxed{\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}}$$

OLP: Online learning with pairwise losses

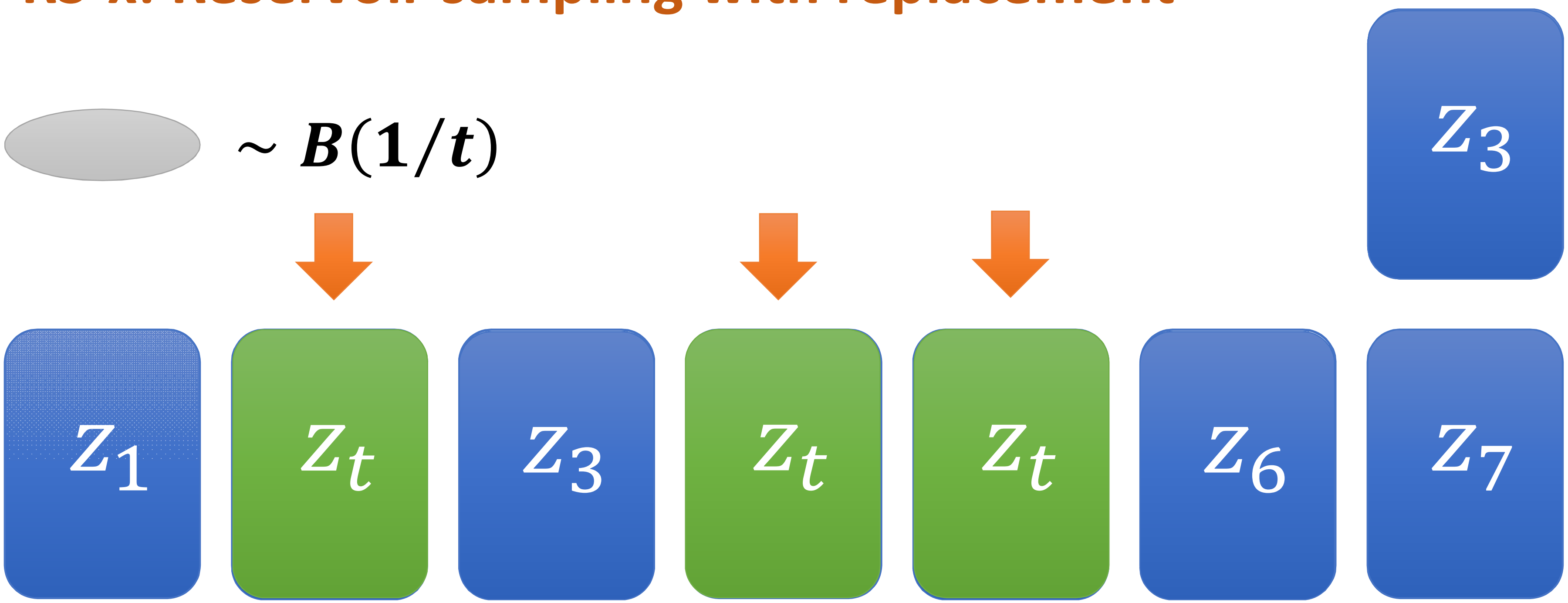
Simple variant of Zinkevich's GIGA

- Start with $w_0 = 0$
- At each $t = 1 \ldots T$
  - Receive a new point z_t
  - Construct appropriate loss function $\ell_t = \ell_t^\infty$ or $\ell_t = \ell_t^{\mathrm{buf}}$
  - $w_t \leftarrow \mathrm{w}_{\mathrm{t}-1} - \frac{\eta}{t} \nabla_w \ell_t(w_{t-1})$
  - If required, update buffer B with z_t

An Online Learning Algorithm for Pairwise Losses

$$\boxed{\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}}$$

RS-x: Reservoir sampling with replacement



An Online Learning Algorithm for Pairwise Losses

$$\boxed{\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}}$$

Guarantees for OLP and RS-x

- **Sampling guarantee**
  At any time $t > s$, the contents of buffer B are s i.i.d. samples from the set $\{z_1, z_2, \ldots, z_{t-1}\}$
- **Regret guarantee**
  OLP guarantees** a finite buffer regret $\frac{1}{T}\mathfrak{R}_T^{\mathrm{buf}} \leq \frac{1}{\sqrt{T}}$

  Finite-to-all-pairs regret conversion

$$\frac{1}{T}\mathfrak{R}_T^{\infty} \leq \frac{1}{T}\mathfrak{R}_T^{\mathrm{buf}} + \sqrt{\frac{\log T}{s}}$$

An Online Learning Algorithm for Pairwise Losses

$$\boxed{\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}}$$

OTB Guarantees for Pairwise loss functions

Define $\mathcal{L}(a) := \underset{z,z' \sim \mathcal{Z}}{\mathbb{E}} \ell(a, z, z')$

- For random x_t, convex ℓ and unbounded buffer

$$\mathcal{L}(\bar{a}) \leq \min_{a \in \mathcal{A}} \mathcal{L}(a) + \frac{1}{T} \mathfrak{R}_T^{\infty} + \mathcal{O}\left(\sqrt{\log T / T}\right)$$

where $\bar{a} = \frac{1}{T} \sum a_t$

An Online Learning Algorithm for Pairwise Losses

$$\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$$

OTB Guarantees for Pairwise loss functions

Define $\mathcal{L}(a) := \underset{z,z' \sim \mathcal{Z}}{\mathbb{E}} \ell(a, z, z')$

- For random x_t, convex ℓ and finite buffer of size s

$$\mathcal{L}(\bar{a}) \leq \min_{a \in \mathcal{A}} \mathcal{L}(a) + \frac{1}{T} \mathfrak{R}_T^{\text{buf}} + \mathcal{O}\left(\sqrt{\log T / s}\right)$$

  where $\bar{a} = \frac{1}{T} \sum a_t$

- **Corollary**: $\mathcal{L}(\bar{a}) \leq \min_{a \in \mathcal{A}} \mathcal{L}(a) + \mathcal{O}\left(\sqrt{\log T / s}\right)$

An Online Learning Algorithm for Pairwise Losses

$$\boxed{\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}}$$

OTB Guarantees for Pairwise loss functions

Define $\mathcal{L}(a) := \underset{z,z' \sim \mathcal{Z}}{\mathbb{E}} \ell(a, z, z')$

- For random x_t, strongly convex ℓ and unbounded buffer

$$\mathcal{L}(\bar{a}) \leq \min_{a \in \mathcal{A}} \mathcal{L}(a) + \frac{1}{T}\mathfrak{R}_T^{\infty} + \mathcal{O}(\log^2 T/T)$$

where $\bar{a} = \frac{1}{T}\sum a_t$

An Online Learning Algorithm for Pairwise Losses

$$\boxed{\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}}$$

OTB Guarantees for Pairwise loss functions

Define $\mathcal{L}(a) := \underset{z,z' \sim \mathcal{Z}}{\mathbb{E}} \ell(a, z, z')$

- For random x_t, strongly convex ℓ and finite buffer

$$\mathcal{L}(\bar{a}) \leq \min_{a \in \mathcal{A}} \mathcal{L}(a) + \frac{1}{T}\mathfrak{R}_T^{\mathrm{buf}} + \mathcal{O}(\log T/s)$$

where $\bar{a} = \frac{1}{T}\sum a_t$

- **Corollary**: $\mathcal{L}(\bar{a}) \leq \min_{a \in \mathcal{A}} \mathcal{L}(a) + \mathcal{O}(\log T/s)$

An Online Learning Algorithm for Pairwise Losses

$$\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$$

Some other details

- Our bounds give dimension independent bounds
  - For Hilbertian norm regularizations: no dependence on d
  - For sparsity inducing regularizations: $\sqrt{\log d}$ dependence
  - Previous work [Wang et al, COLT12]: linear dependence
- Proofs use (modified notions of) Rademacher averages
  - Trickier symmetrization step
  - Previous work: covering number based analysis

Some Open Problems

- Current all-pairs regret bound for finite buffers

$$\mathfrak{R}_T^{\infty} \leq \sqrt{\frac{\log T}{s}}$$

  - Can we get bounds that scale as $1/f(n)$?
- Similar question for OTB conversion bounds

- OTB bounds require *stream-oblivious* buffer updates
  - Update algorithm cannot look at z_t just t
  - Examples: FIFO, RS, RS-x
  - Guarantees for (suitable) stream-aware policies?