Online Prediction: The Role of Convexity and Randomization

Shai Shalev-Shwartz

Toyota Technological Institute at Chicago

Learning Club, The Hebrew University, 2008

Predicting the next element of a binary sequence

Prediction task

For $t = 1, 2, \ldots, T$

- Predict: $\hat{y}_t \in \{\pm 1\}$
- Get: $y_t \in \{\pm 1\}$
- Suffer loss: $\ell_{0-1}(\hat{y}_t, y_t) = \begin{cases} 1 & y_t \neq \hat{y}_t \\ 0 & y_t = \hat{y}_t \end{cases}$

Regret

- Best in hindsight $y^\star = \mathrm{sign}(\sum_t y_t)$
- Regret: $R_T \;=\; \sum_{t=1}^{T} \ell_{0-1}(\hat{y}_t, y_t) - \sum_{t=1}^{T} \ell_{0-1}(\hat{y}_t, y^\star)$

Abstract Prediction Model

- Set of decisions S
- Set of loss functions $\mathcal{L} = \{\ell : S \to \mathbb{R}\}$

Prediction Game

For $t = 1, \ldots, T$

- Learner chooses a decision $\mathbf{w}_t \in S$
- Environment chooses a loss function $\ell_t \in \mathcal{L}$
- Learner suffers loss $\ell_t(\mathbf{w}_t)$

- Goal: Conditions on S, $\mathcal{L}$, and the feedback the learner receives that guarantee low regret

$$R_T \; = \; \sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \sum_{t=1}^{T} \ell_t(\mathbf{w}^\star) \; \stackrel{!}{=} \; o(T)$$

Outline

- Part I: Full Information
  - Motivating example and an abstract online prediction model
  - Cover's impossibility result and randomness
  - A modern view: revealing an underlying convexity
  - Using convex analysis tools for online prediction
  - Sufficient conditions for low regret
  - Tightness
- Part II: Partial Feedback
  (Based on Joint work with S. Kakade and A. Tewari)
  - Motivating application
  - The Banditron
  - Lower regret using inefficient algorithms
  - Open problems

Impossibility Result

- $S = \{\pm 1\}$
- $\mathcal{L} = \{\ell_{0-1}(w_t, 1)\ ,\ \ell_{0-1}(w_t, -1)\}$
- Adversary can make the cumulative loss of the learner to be T by using $\ell_t(\cdot) = \ell_{0-1}(\cdot, -w_t)$
- The constant prediction $w^\star = \mathrm{sign}(\sum_t w_t)$ achieves loss of at most $T/2$
- Regret is at least $T/2$

Conclusion

- In the above example, $|S| = |\mathcal{L}| = 2$.
- Small size does not guarantee low regret

Solution: Randomized Predictions

- Learner predicts $\hat{y}_t = 1$ with probability w_t
- Best in hindsight: $y_t^\star = 1$ with probability $w^\star$ where $w^\star = \frac{|\{t: y_t = 1\}|}{T}$
- Analyze the expected regret:

$$\sum_{t=1}^{T} \mathbb{E}[\hat{y}_t \neq y_t] - \sum_{t=1}^{T} \mathbb{E}[y_t^\star \neq y_t]$$

- There are algorithms that achieve expected regret of $O(\sqrt{T})$

A modern view: revealing an underlying convexity

- Expected zero-one loss can be rewritten as

$$\mathbb{E}[\hat{y}_t \neq y_t] \;=\; \begin{cases} 1 - w_t & \text{if } \; y_t = 1 \\ w_t & \text{if } \; y_t = -1 \end{cases}$$

- Going back to our abstract model, we get that:
  - $S = [0, 1]$
  - $\mathcal{L} = \{\ell(w) = 1 - w \;,\; \ell(w) = w\}$

Properties

- All functions in $\mathcal{L}$ are linear (and thus are convex and Lipschitz)
- S is convex and bounded
- Sufficient conditions for low regret ?

Are we just playing with formalities ?

The convexity assumption is natural in many cases.

Example: Prediction with Expert Advice

- Learner receives a vector $(x_1^t, \ldots, x_d^t) \in [-1, 1]^d$ of experts advice
- Learner needs to predict a target $\hat{y}_t \in \mathbb{R}$
- Environment gives correct target $y_t \in \mathbb{R}$
- Learner suffers loss $|y_t - \hat{y}_t|$
- Goal: Be almost as good as the best experts committee

  $\sum_t |y_t - \hat{y}_t| - \sum_t |y_t - \langle \mathbf{w}^\star, \mathbf{x}^t \rangle| \stackrel{!}{=} o(T)$

Can be modeled as follows:

- S is the d-dimensional probabilistic simplex
- $\mathcal{L} = \{\ell_{\mathbf{x},y}(\mathbf{w}) = |y - \langle \mathbf{w}, \mathbf{x} \rangle| \ : \ \mathbf{x} \in [-1, 1]^d, y \in [-1, 1]\}$

Sufficient Conditions for low regret

The Online Convex Programming (OCP) model

- All functions in $\mathcal{L}$ are convex and L-Lipschitz
- S is convex and $\max\{\|\mathbf{w}\|_2 : \mathbf{w} \in S\} = D$
- Then, there exists an algorithm with regret $O(L\,D\,\sqrt{T})$

Bibliography

- The OCP model was presented by Gordon (1999)
- Zinkevich (2003) introduced the term OCP and proved a regret bound of $O((L^2 + D^2)\sqrt{T})$
- The slightly improved regret bound follows from our analysis below

Tightness

For any prediction algorithm

- Exists S and $\mathcal{L}$ s.t. $L\,D = 1$ and $R_T = \Omega(LD\sqrt{T}) = \Omega(\sqrt{T})$
- (Proof uses probabilistic method)
- Exists S and $\mathcal{L}$ s.t. $L\,D = \sqrt{T}$ and $R_T = \Omega(LD\sqrt{T}) = \Omega(T)$
- (Proof assumes dimension can grow with T)

Dimension independency ?

Yes !

- The regret bound does not depend on the dimensionality of S
- Similarly to Support Vector Machines, we can use Kernel functions

Dimension independency ?

Yes !

- The regret bound does not depend on the dimensionality of S
- Similarly to Support Vector Machines, we can use Kernel functions

Yes ?

- Consider again the prediction with expert advice problem
- d experts, each of which gives an "advice" in $[-1, 1]$
- S is the probabilistic simplex and thus $D = 1$
- Lipschitz constant is $L = \sqrt{d}$
- Regret is $\Omega(\sqrt{d\,T})$.
- Is this the best we can do ?

Low regret algorithmic framework for OCP

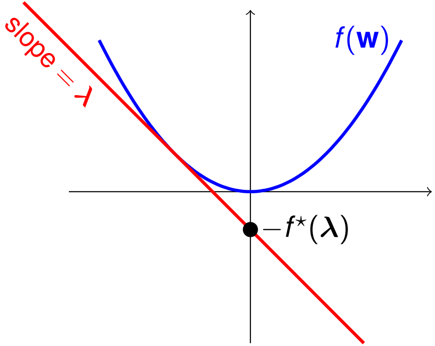
- A low regret algorithmic framework for OCP
- Family of sufficient conditions for low regret
- In particular – Alternatives to the Lipschitz condition
- In the expert committee example – logarithmic dependence on dimension
- Derivation is based on tools from convex analysis

Fenchel Conjugate

The Fenchel conjugate of the function $f : S \to \mathbb{R}$ is $f^\star : \mathbb{R}^d \to \mathbb{R}$

$$f^\star(\boldsymbol{\lambda}) \;=\; \max_{\mathbf{w} \in S} \langle \mathbf{w}, \boldsymbol{\lambda} \rangle - f(\mathbf{w})$$

If f is closed and convex then $f^{\star\star} = f$

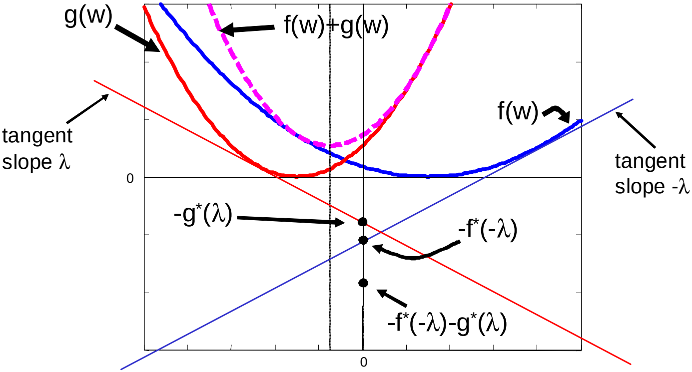


Fenchel Duality

$$\max_{\boldsymbol{\lambda}} \; -f^\star(-\boldsymbol{\lambda}) - g^\star(\boldsymbol{\lambda}) \quad \leq \quad \min_{\mathbf{w}} \; f(\mathbf{w}) + g(\mathbf{w})$$

Fenchel Duality

$$\max_{\boldsymbol{\lambda}} \; -f^\star(-\boldsymbol{\lambda}) - g^\star(\boldsymbol{\lambda}) \;\; \leq \;\; \min_{\mathbf{w}} \; f(\mathbf{w}) + g(\mathbf{w})$$



Regret and Duality

- Recall that our goal is:

$$\forall \mathbf{w}^\star \in S, \quad \sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \sum_{t=1}^{T} \ell_t(\mathbf{w}^\star) \;\leq\; L\,D\sqrt{T}$$

Regret and Duality

- Recall that our goal is:

$$\forall \mathbf{w}^\star \in S, \quad \sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \sum_{t=1}^{T} \ell_t(\mathbf{w}^\star) \;\leq\; L\,D\sqrt{T}$$

- Rewrite it in a 'silly' way

$$\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) \;\leq\; \min_{\mathbf{w}\in S}\; L\,D\sqrt{T} + \sum_{t=1}^{T} \ell_t(\mathbf{w})$$

Regret and Duality

- Recall that our goal is:

$$\forall \mathbf{w}^\star \in S, \quad \sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \sum_{t=1}^{T} \ell_t(\mathbf{w}^\star) \ \le\ L\,D\,\sqrt{T}$$

- Rewrite it in a 'silly' way

$$\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) \ \le\ \min_{\mathbf{w}\in S}\ L\,D\,\sqrt{T} + \sum_{t=1}^{T} \ell_t(\mathbf{w})$$

- Replace $LD\sqrt{T}$ with a function $f : S \to \mathbb{R}$ s.t. $\max_{\mathbf{w}} f(\mathbf{w}) \le LD\sqrt{T}$. E.g. $f(\mathbf{w}) = c\,\|\mathbf{w}\|^2$ for $c = L\sqrt{T}/D$. Obtaining:

$$\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) \ \le\ \min_{\mathbf{w}\in S}\ f(\mathbf{w}) + \sum_{t=1}^{T} \ell_t(\mathbf{w})$$

Regret and Duality

- Recall that our goal is:

$$\forall \mathbf{w}^\star \in S, \quad \sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \sum_{t=1}^{T} \ell_t(\mathbf{w}^\star) \;\leq\; L\,D\sqrt{T}$$

- Rewrite it in a 'silly' way

$$\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) \;\leq\; \min_{\mathbf{w}\in S}\; L\,D\sqrt{T} + \sum_{t=1}^{T} \ell_t(\mathbf{w})$$

- Replace $LD\sqrt{T}$ with a function $f : S \to \mathbb{R}$ s.t. $\max_{\mathbf{w}} f(\mathbf{w}) \leq LD\sqrt{T}$. E.g. $f(\mathbf{w}) = c\,\|\mathbf{w}\|^2$ for $c = L\sqrt{T}/D$. Obtaining:

$$\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) \;\leq\; \min_{\mathbf{w}\in S}\; f(\mathbf{w}) + \sum_{t=1}^{T} \ell_t(\mathbf{w})$$

- Lower bound of a minimization problem. Duality ?

Properties of the dual problem

$$\max_{\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_T} \quad -f^\star(-\sum_t \boldsymbol{\lambda}_t) - \sum_t \ell_t^\star(\boldsymbol{\lambda}_t) \quad \leq \quad \min_{\mathbf{w}\in S} \; f(\mathbf{w}) + \sum_{t=1}^{T} \ell_t(\mathbf{w})$$

Decomposability of the dual

- There's a different dual variable for each online round
- Future loss functions do not affect dual variables of current and past rounds
- Therefore, the dual can be optimized incrementally
- To optimize $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_t$, it is enough to know what the market did until day t

Primal-Dual Online Prediction Strategy

Algorithmic Framework

- Initialize $\boldsymbol{\lambda}_1 = \ldots = \boldsymbol{\lambda}_T = \mathbf{0}$
- For $t = 1, 2, \ldots, T$
  - Construct $\mathbf{w}_t$ from the dual variables
  - Receive ℓ_t
  - Update dual variables $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_t$

Primal-Dual Online Prediction Strategy

Algorithmic Framework

- Initialize $\boldsymbol{\lambda}_1 = \ldots = \boldsymbol{\lambda}_T = \mathbf{0}$
- For $t = 1, 2, \ldots, T$
  - Construct $\mathbf{w}_t$ from the dual variables
  - Receive ℓ_t
  - Update dual variables $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_t$

Lemma

Let $\mathcal{D}_t$ be the dual value at round t and w.l.o.g assume $\mathcal{D}_1 = 0$.

- *Assume that* $\max_{\mathbf{w} \in S} f(\mathbf{w}) \leq a\sqrt{T}$
- *Assume that* $\mathcal{D}_{t+1} - \mathcal{D}_t \geq \ell_t(\mathbf{w}_t) - \frac{a}{\sqrt{T}}$

Then, the regret is bounded by $2a\sqrt{T}$

The proof follows directly from the weak duality theorem

Strong convexity and sufficient dual increase

Strong Convexity w.r.t. norm

A function f is σ-strongly convex over S w.r.t $\|\cdot\|$ if for all $\mathbf{u}, \mathbf{v} \in S$

$$\frac{f(\mathbf{u})+f(\mathbf{v})}{2} \;\geq\; f(\tfrac{\mathbf{u}+\mathbf{v}}{2}) + \tfrac{\sigma}{8}\|\mathbf{u}-\mathbf{v}\|^2$$

Lemma (Sufficient Dual Increase)

Assume:

- *f is σ-strongly convex w.r.t. $\|\cdot\|$*
- *ℓ_t is closed and convex*
- *∇_t is a sub-gradient of ℓ_t at $\mathbf{w}_t$*

Then, there exists a simple dual update rule s.t.

$$\mathcal{D}_{t+1} - \mathcal{D}_t \;\geq\; \ell_t(\mathbf{w}_t) - \frac{\|\nabla_t\|_\star^2}{2\sigma}$$

Generalized Boundedness-Lipschitz condition

Theorem

Assume:

- *Exists* $f : S \to \mathbb{R}$ *which is 1-strongly convex w.r.t.* $\|\cdot\|$
- $D = \max_{\mathbf{w} \in S} \sqrt{f(\mathbf{w})}$
- ℓ_t *is closed and convex*
- $\|\nabla_t\|_\star \leq L$ *(Lipschitz w.r.t. norm* $\|\cdot\|_\star$*)*

Then, there exists an algorithm with regret bound $2\,D\,L\,\sqrt{T}$

Example usage – back to expert problem

- Take f to be the relative entropy
- f is strongly convex w.r.t. $\|\cdot\|_1$ and $D = \sqrt{\log(d)}$
- $\|\nabla_t\|_\star = \|\mathbf{x}^t\|_\infty \leq 1$
- Regret bound becomes $O(\sqrt{\log(d)\,T})$

Self Boundedness instead of Lipschitz

Theorem

Replacing Lipschitz condition with the following self-bounded property:

$$\|\nabla_t\| \leq L\sqrt{\ell_t(\mathbf{w}_t)}$$

Then,

$$R_T \;\leq\; O\left(L\,D\sqrt{\sum_t \ell_t(\mathbf{w}^\star)} + L^2\,D^2\right) \;.$$

Examples

- $\ell(\mathbf{w}) = \frac{1}{2}(\langle \mathbf{w}, \mathbf{x}\rangle - y)^2$ is $(\sqrt{2}\|\mathbf{x}\|)$-self-bounded
- $\ell(\mathbf{w}) = \log(1.26 + \exp(-y\langle \mathbf{w}, \mathbf{x}\rangle))$ is $(\|\mathbf{x}\|)$-self-bounded

Part II
Online learning with partial feedback

Based on joint work with

S. Kakade and A. Tewari

Online learning with partial feedback

Motivating Application – Advertisement on webpages

- k types of ads
- On round t:
  - User submit a query
  - System (the learner) places an ad
  - User either 'clicks' or ignores

A simple formal model – bandit multiclass categorization

- On round t:
  - Environment presents a vector $\mathbf{x}_t$ (encodes user and query)
  - Learner predicts an ad $\hat{y}_t \in \mathcal{Y} = \{1, \ldots, k\}$
  - Environment chooses current user interest $y_t \in \mathcal{Y}$ but only reveals $\mathbf{1}_{[y_t \neq \hat{y}_t]}$

Linear hypotheses and margin

- Linear hypotheses:
  $h : \mathbb{R}^d \to \mathcal{Y}$ s.t. exists a $k \times d$ matrix W s.t.
  $$h(\mathbf{x}) = \operatorname*{argmax}_{r \in \mathcal{Y}} (W\mathbf{x})_r$$
- Separability with margin assumption:
  Exists a matrix $W^\star$ with $\|W^\star\|_F \leq D$ s.t. for all t, $r \neq y_t$,
  $$(W\mathbf{x}_t)_{y_t} \geq (W\mathbf{x}_t)_r$$

The multiclass Perceptron for the full information case

- Initialize $W^1 = \mathbf{0} \in \mathbb{R}^{k \times d}$
- For $t = 1, 2, \ldots, T$
  - Receive $\mathbf{x}_t \in \mathbb{R}^d$
  - Predict $\hat{y}_t = \arg\max_{r \in [k]} (W^t \mathbf{x}_t)_r$
  - Receive feedback y_t
  - Define $U^t \in \mathbb{R}^{k \times d}$ such that: $\quad U^t_{r,j} = x_{t,j} \left(\mathbf{1}_{[r = y_t]} - \mathbf{1}_{[r = \hat{y}_t]} \right)$
  - Update: $W^{t+1} = W^t + U^t$

The Banditron

- Exploration-Exploitation parameter: $\gamma \in (0, 0.5)$
- Initialize $W^1 = \mathbf{0} \in \mathbb{R}^{k \times d}$
- For $t = 1, 2, \ldots, T$
  - Receive $\mathbf{x}_t \in \mathbb{R}^d$
  - Define $\hat{y}_t = \arg\max_{r \in [k]} (W^t \mathbf{x}_t)_r$
  - Exploit: w.p. $1 - \gamma$ predict $\tilde{y}_t = \hat{y}_t$
  - Explore: w.p. γ predict $\tilde{y}_t \in \mathcal{Y}$ uniformly at random
  - Receive partial feedback $\mathbf{1}_{[\tilde{y}_t = y_t]}$
  - Define $\tilde{U}^t \in \mathbb{R}^{k \times d}$ such that: $\tilde{U}^t_{r,j} = x_{t,j} \left(\frac{\mathbf{1}_{[y_t = \tilde{y}_t]} \mathbf{1}_{[\tilde{y}_t = r]}}{P(r)} - \mathbf{1}_{[\hat{y}_t = r]} \right)$
  - Update: $W^{t+1} = W^t + \tilde{U}^t$

The Banditron – Analysis

Theorem (Banditron – separable case)

The expected number of mistakes the Banditron makes on a separable sequence is at most $O(D\sqrt{k\,T})$.

- Proof idea: show that the expected update of the Banditron (i.e. $\tilde{U}^t$) is the Perceptron's update (i.e. U^t)
- We also have bounds for the non-separable case:
  - For 'low noise' the bound is still $O(D\sqrt{k\,T})$.
  - For 'high noise'. The dependence is on $T^{2/3}$.
- Randomness is utilized for obtaining an estimator of the Perceptron's update.
- In the full information case: multiclass Perceptron's bound, $O(D^2)$, does not depend on T

Inefficient algorithms for the separable case

Theorem

- *There exists a deterministic algorithm with mistake bound* $O(k^2 d \log(D\, d))$
- *There exists a randomized algorithm with mistake bound* $O(k^2 D^2 \log(D) \log(T + k)))$

Proof sketch

- Important observation: Halving algorithm works for multiclass problems with partial feedback
- 1st result: Construct a grid that covers matrices with bounded norm
- 2nd result: Use random projections and the JL lemma

Open Problems

- Achievable regret bounds with efficient and inefficient algorithms (lower bounds?)
- When is randomization a must (sometimes it's not necessary; e.g. Halving, Negatron)
- Banditron with multiplicative updates
- More sophisticated exploration vs. exploitation (e.g. self-tuned γ)
- From single label to label ranking