Online Learning of Noisy Data with Kernels

Nicolò Cesa-Bianchi[1], Shai Shalev-Shwartz[2]
and **Ohad Shamir**[2]

[1]Università degli Studi di Milano

[2]The Hebrew University

COLT, June 2010

Online Learning with Partial Information

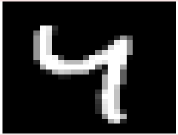
- Standard online learning: After choosing a predictor, Learner sees example chosen by adversary
- Harder setting: Learner only receives partial information on each example

Example (Bandit Learning)

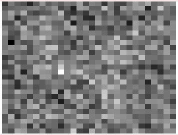
Learner gets to see loss value

This talk

Learner has noisy view of each example



$\Longrightarrow$



Online Learning with Partial Information

Main Results

- Online learning of linear predictors based on noisy views
  - $\mathcal{O}(\sqrt{T})$ regret
  - Noise distribution unknown. Can be chosen adversarially and change for each example
  - Including kernels
  - General technique for unbiased estimators of nonlinear functions

Online Learning with Partial Information

Online Learning of Linear Predictors

On each round t:

- Learner picks predictor $\mathbf{w}_t \in \mathcal{W}$
- Nature picks $(\mathbf{x}_t, y_t)$
- Learner suffers loss $\ell(\langle \mathbf{w}_t, \mathbf{x}_t \rangle, y_t)$
- Learner gets y_t and noisy view of $\mathbf{x}_t$

Learner's goal: minimize **regret**

$$\sum_{t=1}^{T} \ell(\langle \mathbf{w}_t, \mathbf{x}_t \rangle, y_t) - \min_{\mathbf{w} \in \mathcal{W}} \sum_{t=1}^{T} \ell(\langle \mathbf{w}, \mathbf{x}_t \rangle, y_t)$$

First Try

Suppose learner gets $\tilde{\mathbf{x}}_t = \mathbf{x}_t + \mathbf{n}_t$, $\mathbf{n}_t$ random zero-mean noise vector

First Try

Suppose learner gets $\tilde{\mathbf{x}}_t = \mathbf{x}_t + \mathbf{n}_t$, $\mathbf{n}_t$ random zero-mean noise vector

Unfortunately, too hard!

Theorem

If an adversary can choose the noise distribution, and $\ell(\cdot, 1)$ is

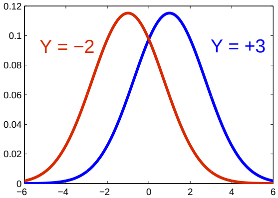
1. *Bounded from below*
2. *Differentiable at 0 with $\ell'(0, 1) < 0$ (a.k.a. classification calibrated)*

then sublinear regret is impossible

Holds even in a stochastic setting

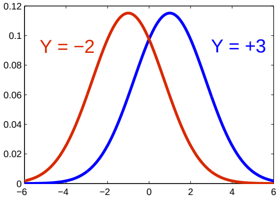
First Try

Suppose data looks like this:

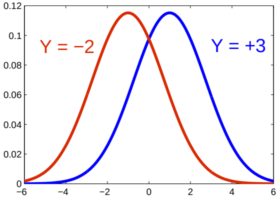


First Try

Suppose data looks like this:

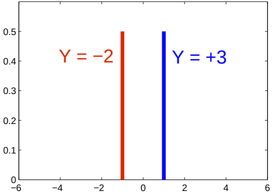


Option A: Data comes from



+ no noise

Option B: Data comes from



+ noise

Information-theoretically impossible to distinguish!

First Try

- Must provide more information to the learner
- Suppose that can get more than one independent copies of $\tilde{\mathbf{x}}_t = \mathbf{x}_t + \mathbf{n}_t$
- Trivial (and unrealistic) setting if unlimited number of copies
- Goal: small number of views, independent of problem scale

Important Technique

Stochastic Online Gradient Descent

- Initialize $\mathbf{w}_1 = \mathbf{0}$
- For $t = 1, \ldots, T$
  - $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \tilde{\nabla}_t$
  - Project $\mathbf{w}_{t+1}$ on ball $\{\mathbf{w} : \|\mathbf{w}\|^2 \leq W\}$

- When $\tilde{\nabla}_t = \nabla \ell(\langle \mathbf{w}_t, \mathbf{x}_t \rangle, y_t)$, this is standard online gradient descent (Zinkevich 2003)

Important Technique

Stochastic Online Gradient Descent

- Initialize $\mathbf{w}_1 = \mathbf{0}$
- For $t = 1, \ldots, T$
  - $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \tilde{\nabla}_t$
  - Project $\mathbf{w}_{t+1}$ on ball $\{\mathbf{w} : \|\mathbf{w}\|^2 \leq W\}$

- When $\tilde{\nabla}_t = \nabla \ell(\langle \mathbf{w}_t, \mathbf{x}_t \rangle, y_t)$, this is standard online gradient descent (Zinkevich 2003)

Theorem

If $\mathbb{E}[\tilde{\nabla}_t] = \nabla \ell(\langle \mathbf{w}_t, \mathbf{x}_t \rangle, y_t), \quad \mathbb{E}[\|\tilde{\nabla}_t\|^2] \leq B$, *expected regret at most*

$$\mathcal{O}\left(\sqrt{BWT}\right)$$

Unknown Noise

Example (Linear predictors, squared loss)

- Gradient is $2(\langle \mathbf{w}_t, \mathbf{x}_t \rangle - y_t)\mathbf{x}_t$
- Unbiased estimate with 2 noisy copies of $\mathbf{x}_t$:

$$2(\langle \mathbf{w}_t, \tilde{\mathbf{x}}_t \rangle - y_t)\tilde{\mathbf{x}}'_t$$

$\Rightarrow$ Can learn in the face of unknown noise

Unknown Noise

Example (Linear predictors, squared loss)

- Gradient is $2(\langle \mathbf{w}_t, \mathbf{x}_t \rangle - y_t)\mathbf{x}_t$
- Unbiased estimate with 2 noisy copies of $\mathbf{x}_t$:

$$2(\langle \mathbf{w}_t, \tilde{\mathbf{x}}_t \rangle - y_t)\tilde{\mathbf{x}}'_t$$

$\Rightarrow$ Can learn in the face of unknown noise

- What if we want other loss functions? Non-linear predictors?

Note:

Technique depended on loss gradient being quadratic in $\mathbf{x}$. Won't work otherwise!

Next: how we can learn with unknown noise using:

- General 'smooth' loss functions
- Non-linear predictors using kernels

Kernels

- Allows to learn highly non-linear predictors
- Idea: instances $\mathbf{x}$ mapped to $\Psi(\mathbf{x})$ in a high dimensional Hilbert space, and a linear predictor learned in that space

Kernels

- Allows to learn highly non-linear predictors
- Idea: instances $\mathbf{x}$ mapped to $\Psi(\mathbf{x})$ in a high dimensional Hilbert space, and a linear predictor learned in that space
- Problematic for our setting: Ψ may be complex and non-linear. In particular, $\mathbb{E}[\Psi(\tilde{\mathbf{x}})] \neq \Psi(\mathbf{x})$

Idea: Unbiased Estimate of Non-Linear Functions

- Suppose X is a real random variable with unknown distribution, and unknown mean μ. Can sample $x_1, x_2, \ldots$
- Want an unbiased estimate of $f(\mu)$

Idea: Unbiased Estimate of Non-Linear Functions

- Suppose X is a real random variable with unknown distribution, and unknown mean μ. Can sample $x_1, x_2, \ldots$
- Want an unbiased estimate of $f(\mu)$
- If f is linear: return $f(x_1)$:

$$\mathbb{E}[f(x_1)] = f(\mathbb{E}[x_1]) = f(\mu)$$

Idea: Unbiased Estimate of Non-Linear Functions

- Suppose X is a real random variable with unknown distribution, and unknown mean μ. Can sample $x_1, x_2, \ldots$

- Want an unbiased estimate of $f(\mu)$

- If f is linear: return $f(x_1)$:

$$\mathbb{E}[f(x_1)] = f(\mathbb{E}[x_1]) = f(\mu)$$

- When f is nonlinear, $\mathbb{E}[f(x_1)] \neq f(\mu)$. In many cases, unbiased estimate of $f(\mu)$ based on $x_1, x_2, \ldots, x_n$ is provably impossible

Idea: Unbiased Estimate of Non-Linear Functions

- Suppose X is a real random variable with unknown distribution, and unknown mean μ. Can sample $x_1, x_2, \ldots$
- Want an unbiased estimate of $f(\mu)$
- If f is linear: return $f(x_1)$:

$$\mathbb{E}[f(x_1)] = f(\mathbb{E}[x_1]) = f(\mu)$$

- When f is nonlinear, $\mathbb{E}[f(x_1)] \neq f(\mu)$. In many cases, unbiased estimate of $f(\mu)$ based on $x_1, x_2, \ldots, x_n$ is provably impossible
- However: What if n is random?

Idea: Unbiased Estimate of Non-Linear Functions

- Suppose f is a continuous function on a bounded interval
- There exist $A_0(\cdot), A_1(\cdot), \ldots$, where $A_n(x) = \sum_{k=0}^{n} a_{n,k} x^k$, such that $A_n(\cdot) \stackrel{n\to\infty}{\longrightarrow} f(\cdot)$
- Let $p_0, p_1, \ldots$ be a distribution over all nonnegative integers

Idea: Unbiased Estimate of Non-Linear Functions

- Suppose f is a continuous function on a bounded interval
- There exist $A_0(\cdot), A_1(\cdot), \ldots$, where $A_n(x) = \sum_{k=0}^{n} a_{n,k} x^k$, such that $A_n(\cdot) \xrightarrow{n\to\infty} f(\cdot)$
- Let $p_0, p_1, \ldots$ be a distribution over all nonnegative integers

Estimator

1. Pick n randomly according to $\Pr(n) = p_n$
2. Sample $x_1, \ldots, x_n$ independently
3. Return

$$\theta = \frac{1}{p_n}\left(\sum_{k=0}^{n} a_{n,k}\left(\prod_{i=0}^{k} x_i\right)\right) - \frac{1}{p_n}\left(\sum_{k=0}^{n} a_{n-1,k}\left(\prod_{i=0}^{k} x_i\right)\right)$$

Theorem

$$\mathbb{E}[\theta] = f(\mu)$$

Idea: Unbiased Estimate of Non-Linear Functions

Proof

$$\theta = \frac{1}{p_n}\underbrace{\left(\sum_{k=0}^{n} a_{n,k}\left(\prod_{i=0}^{k} x_i\right)\right)}_{=A_n(\mu)\ \text{in expectation}} - \frac{1}{p_n}\underbrace{\left(\sum_{k=0}^{n} a_{n-1,k}\left(\prod_{i=0}^{k} x_i\right)\right)}_{=A_{n-1}(\mu)\ \text{in expectation}}$$

Therefore,

$$\begin{aligned}
\mathbb{E}[\theta] &= \mathbb{E}_n\left[\frac{1}{p_n}\left(A_n(\mu) - A_{n-1}(\mu)\right)\right] \\
&= \sum_{n=1}^{\infty}\left(A_n(\mu) - A_{n-1}(\mu)\right) \\
&= f(\mu) - A_0(\mu) \ = \ f(\mu)
\end{aligned}$$

Idea: Unbiased Estimate of Non-Linear Functions

- Technique used in a 1960's paper on sequential estimation (R. Singh, 1964)

Idea: Unbiased Estimate of Non-Linear Functions

- Technique used in a 1960's paper on sequential estimation (R. Singh, 1964)
- Crucial observation: if p_n decays rapidly, then with overwhelming probability, will need just a small number of samples
- When f is analytic, can take $p_n \propto 1/q^n$ for arbitrary q
- We use this technique to learn with noise, using large families of kernels and analytic loss functions

Formal Result - Example

- Consider any *dot product* kernel $k(\mathbf{x}, \mathbf{x}') = G(\langle \mathbf{x}, \mathbf{x}' \rangle)$
  - e.g. $k(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}' \rangle + 1)^n$

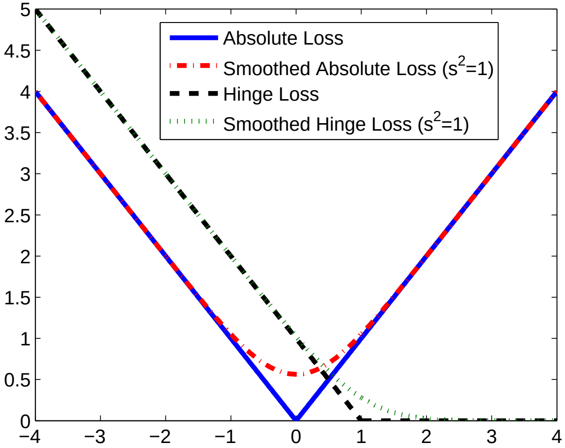
Example (Dot product kernel, squared loss)

Suppose $\mathbb{E}[\|\tilde{\mathbf{x}}\|^2] \leq B$. For any $q > 1.1$, can construct an efficient algorithm which:

- Queries each example $1 + \mathcal{O}_p(1/q)$ times
- Has regret $\mathcal{O}\left(WG(qB)\sqrt{qT}\right)$ w.r.t. $\{\mathbf{w} : \|\mathbf{w}\|^2 \leq W\}$

Tradeoff: Large q implies less queries per example, but larger regret

Smoothed Losses



Summary

- Online Learning with noise
  - Noise distribution may be chosen adversarially
- Quantity makes quality: More examples make up for bad quality of each individual example seen
- General technique to construct unbiased estimators of nonlinear functions
- Can be improved?
  - Upcoming work: yes, if know more about the noise distribution