Learning Drifting Data
Using Selective Sampling

Eli Kravchik

Supervised by Prof. Koby Crammer

Faculty of Electrical Engineering, Technion
Israel Institute of Technology

30.3.2014

Outline

1 Introduction

2 Classification Setting

3 Regression Setting

4 Summary

Online Learning

- Used for many different tasks:

  - Information filtering

  - Market analysis

  - Big data problems

- Data revealed round after round

- Learner has to make prediction online

Online Learning

- At each round t:

  1 Instance x_t observed

  2 Prediction $\hat{y}_t$ issued

  3 Regret R_t suffered

  4 True value y_t revealed

- Regret definition

$$R_t = \mathcal{L}\{\hat{y}_t, y_t\} - \mathcal{L}\{\bar{y}_t, y_t\}$$

$\bar{y}_t$ - optimal prediction, $\mathcal{L}\{,\}$ - loss function

Selective Sampling

- Acquiring true value (or label) y_t can be costly or complicated

- Algorithms can achieve similar results without knowing all true labels y_t

- Only some of the labels are queried. Others remain unknown

- Queries are issued according to confidence of the algorithm

[Cesa-Bianchi et al., 2006, 2009], [Fruend et al., 1997]

Drifting Data

- In some problems data varies over time

- The optimal function from family of functions learned, is not fixed

- Approaches to handle drifting data :

  - Detect the drift

  - Adjust algorithm to drift setting

Drifting Data - Related Work

- Time windows:

    - For drift detection

    - For prediction

    [Klinkenberg, 2004], [Widemer, Kubat 1996]

- Detecting drift using error distribution

    [Gama et al., 2004], [Garcia-Baena et al., 2006]

- Forgetting strategies

    [Vaits and Crammer, 2011]

Drifts vs Switches

Data can change gradually (drift) or suddenly (switch).

Example - market analysis:

- Gradual drift - effect of Israeli real estate bubble
- Abrupt Switch - effect of Russian invasion to Crimea

In this work

- Effect of switch on online learning problems is approached

- Linear classification and linear regression selective sampling settings are examined

- Selective sampling principles suggested to overcome switch effect

Suggested Method

- Implement selective sampling approach to overcome switch in an online learning setting

- Strategy of dealing with switch:

  1 Detect the switch

  2 If switch is undetected - assure that the harm caused by the switch is minor

  3 Small probability for false detections

Suggested Method

- Exploit notion of confidence provided by selective sampling to handle switch

- Avoid unnecessary loss of information while overcoming switch effect

- Time windows used for change detection but not for classification

- Selective sampling is also used in original context - only part of the labels are queried

Problem Setting - Online Classification

Problem setup:

- $x_t \in R^d$

- $y_t \in \{\pm 1\}$

Assumptions on instances:

- $\|x_t\| = 1$

Problem Setting - Online Classification

Assumptions on label distribution:

- $\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1,\, \boldsymbol{u}, \boldsymbol{v} \in R^d$

- For $t \leq \tau$ holds:

  - $\mathrm{E}\left[y_t\right] = \boldsymbol{u}^\top \boldsymbol{x}_t$

  - $\Pr\left[y_t = 1\right] = \frac{1 + \boldsymbol{u}^\top \boldsymbol{x}_t}{2}$

- For $t > \tau$ holds:

  - $\mathrm{E}\left[y_t\right] = \boldsymbol{v}^\top \boldsymbol{x}_t$

  - $\Pr\left[y_t = 1\right] = \frac{1 + \boldsymbol{v}^\top \boldsymbol{x}_t}{2}$

Problem Setting

- At each round t instance $\boldsymbol{x}_t$ observed

- Prediction $\hat{y}_t$ issued

- Regret R_t suffered

- True label y_t can be queried

Linear classification is used to issue prediction:

$$\hat{y}_t = \text{sign} \left\{ \boldsymbol{w}_t^\top \boldsymbol{x}_t \right\}$$

RLS Estimator

RLS estimator (Cesa-Bianchi et al. 2004, 2006, 2009) used:

$$\boldsymbol{w}_t = A_t^{-1} b_t$$

Where:

- $A_t = \left(I + \sum\limits_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top + \boldsymbol{x}_t \boldsymbol{x}_t^\top \right) \in R^{d \times d}$

- $b_t = \sum\limits_{i=1}^{n} y_i \boldsymbol{x}_i \in R^d$

$n = N_t$ - number of queries issued until round $t - 1$

A_t can be viewed as covariance or "confidence " matrix

BBQ Algorithm - Querying Labels

Cesa-Bianchi, Gentile, Orabona 2009:

Selective sampling algorithm -

- Set $\kappa \in (0, 1)$

- Calculate $r_t = \boldsymbol{x}_t^\top A_t^{-1} \boldsymbol{x}_t$

- If $r_t > t^{-\kappa}$ label y_t is queried and A_t, b_t are updated

- If $r_t \leq t^{-\kappa}$ label y_t remains unknown and no update performed

RLS Estimator and BBQ Algorithm Properties

Assuming standard, no switch setting, $(\boldsymbol{u} = \boldsymbol{v})$:

- Logarithmic cumulative regret R_T:

$$R_T \leq O\left(d \ln T\right) + f\left\{\kappa\right\}$$

- Reduced number of queried labels N_T:

$$N_T \sim O\left(dT^\kappa \ln T\right)$$

- Controlled estimator bias B_t:

$$B_t = \boldsymbol{w}_t^\top \boldsymbol{x}_t - \mathrm{E}\left[\boldsymbol{w}_t^\top \boldsymbol{x}_t\right] \leq r_t + \sqrt{r_t}$$

Effect of Switch

- Switch from $\boldsymbol{u}$ to $\boldsymbol{v}$ at round τ increases bias bound:
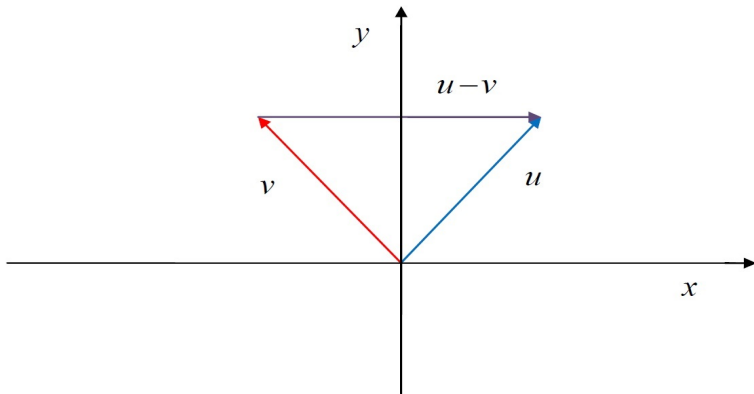
$$B_t \leq r_t + \sqrt{r_t} + N_\tau \|\boldsymbol{v} - \boldsymbol{u}\|\sqrt{r_t}$$

- The bias B_t controls the regret R_t. So switch at round τ increases regret bound:

$$R_T \leq O\left(\left\{\|\boldsymbol{v} - \boldsymbol{u}\|^2 \tau^{2\kappa} (d \ln \tau)^2 + 1\right\} d \ln T\right) + \Gamma\left\{\frac{1}{\kappa}\right\}$$

Effect of Switch

Switch from u to v at round τ:

Using Selective Sampling to Detect Switch

- When switch occurs low cumulative regret can no longer be expected

- Selective sampling approach measures estimator's confidence regarding prediction on given instance x_t

- When confidence on instance x_t is high, prediction should be close to optimal

- Evaluating prediction on "high confidence" instances can be used to detect change

Using Selective Sampling to Detect Switch

Confidence factor $r_t = \boldsymbol{x}_t^\top A_t^{-1} \boldsymbol{x}_t$:

- Small r_t - high confidence regarding instance $\boldsymbol{x}_t$
- Large r_t - high uncertainty (low confidence) regarding instance $\boldsymbol{x}_t$

r_t controls both the bias B_t and the instantaneous regret R_t:

- If r_t is large, low regret R_t can not be assured, switch or no switch.
- If r_t is small, low regret R_T should be expected. Unless a switch had occurred...

Using Selective Sampling to Detect Switch

Main idea - evaluate performance on instances with small r_t to detect switch.

Bad performance will indicate that switch had occurred.

- Performance cannot be evaluated comparing prediction $\hat{y}_t$ to label y_t due to noise

- Even if optimal classifier v is known - error probability will be $\frac{1-|v^\top x_t|}{2}$

- Prediction will be evaluated comparing to optimal classifier $|w_t^\top x_t - v^\top x_t|$

Windowed Demo Classifiers

- Problem - optimal classifier v is unknown.

- Solution - estimate optimal classifier v with demo classifier h_t.

- Demo classifier h_t constructed from a window of last L rounds and should estimate v well enough

- Performance of w_t comparing to h_t will evaluate $\left| w_t^\top x_t - v^\top x_t \right|$ and indicate possible switch

Construction of Windowed Demo Classifier

- Parameter - initial window length $L_0 > 0$

- Calculate window length $L_t = L_0 + \sqrt{t}$

- At round t select a window of last L_t instances

- Calculate $A_{L_t} = \left(I + \sum\limits_{l=t-L_t}^{t-1} \boldsymbol{x}_l \boldsymbol{x}_l^\top \right), b_{L_t} = \sum\limits_{l=t-L_t}^{t-1} y_l \boldsymbol{x}_l$

- Construct demo classifier $h_t = \left(A_{L_t} + \boldsymbol{x}_t \boldsymbol{x}_t^\top \right)^{-1} b_{L_t}$

Resolution of Windowed Demo Classifier

- Demo classifier h_t constructed at round t from a window of last L_t instances

- Demo classifier h_t will be used to evaluate next KL_t instances

- Next demo classifier $h_{t_{next}}$ will be constructed at round $KL_t + 1$

- Only $\frac{T}{K}$ labels will be queried

- Switch detection resolution reduced from L_t to KL_t

Construction and Resolution of Demo Classifier

Demo classifier setup:

$$w_t$$

$$h_{t_{k-2}} \qquad h_{t_{k-1}} \qquad h_{t_k}$$

Algorithm for Detecting Switch

- Calculate estimator $\boldsymbol{w}_t = A_t^{-1} b_t$

  - Where $A_t = \left(I + \sum\limits_{i=1}^{N_t} \boldsymbol{x}_i \boldsymbol{x}_i^\top + \boldsymbol{x}_t \boldsymbol{x}_t^\top\right), b_t = \sum\limits_{i=1}^{N_t} y_i \boldsymbol{x}_i$

- Calculate demo classifier $h_t = \left(A_{L_t} + \boldsymbol{x}_t \boldsymbol{x}_t^\top\right)^{-1} b_{L_t}$

  - $A_{L_t} = \left(I + \sum\limits_{l=m_t-L_t}^{m_t} \boldsymbol{x}_l \boldsymbol{x}_l^\top\right), b_{L_t} = \sum\limits_{l=m_t-L_t}^{m_t} y_l \boldsymbol{x}_l$

- Calculate $r_t = \boldsymbol{x}_t^\top A_t^{-1} \boldsymbol{x}_t$

- Calculate $r_{L_t} = \boldsymbol{x}_t^\top \left(A_{L_t} + \boldsymbol{x}_t \boldsymbol{x}_t^\top\right)^{-1} \boldsymbol{x}_t$

Algorithm for Detecting Switch

- Parameter - $\delta \in (0, 1)$

- Calculate $\delta_t = \frac{\delta}{t(t+1)}$

- Calculate $C_t = \left| \boldsymbol{w}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t \right|$

- Calculate:

$$K_t = \left(\sqrt{2 \ln \frac{2}{\delta_t}} + 1 \right) \sqrt{r_t} + r_t + \left(\sqrt{2 \ln \frac{2}{\delta_t}} + 1 \right) \sqrt{r_{L_t}} + r_{L_t}$$

- If $C_t > K_t$ declare switch and restart classifier w_t from zero. Else continue to next round

Algorithm for Detecting Switch

- If $C_t > K_t$

  - If switch occurred it is detected

  - If no switch occurred $\Pr[C_t > K_t] \leq 2\delta$ - to be proved

- If $C_t \leq K_t$

  - If no switch occurred, no change applied to standard setting

  - If a switch occurred and undetected as $C_t \leq K_t$, additional regret caused would be small - to be proved

Algorithm's Main Result

Main result:

1. If a switch occurs - algorithm detects it, or assures it causes small harm

2. No switch occurs - no false detection

Proving Main Result

Proof structure as follows:

1. Proving undetected switch will cause low regret:

   - Bounding instantaneous regret

   - Summing to bound cumulative regret

2. Proving probability for false positives is small

Proving Main Result - Instantaneous Regret Bound

Instantaneous regret R_t controlled by the term $|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t|$:

$$R_t = \Pr\left[y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t < 0\right] - \Pr\left[y_t \boldsymbol{v}^\top \boldsymbol{x}_t < 0\right] \leq$$

$$\varepsilon I_{\{|\boldsymbol{v}^\top \boldsymbol{x}_t| < \varepsilon\}} + \Pr\left[\left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right| \geq \varepsilon\right]$$

$|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t|$ can be bounded by triangle inequality:

$$\left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right| \leq \left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t\right| + \left|\boldsymbol{v}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t\right|$$

$$= C_t + \left|\boldsymbol{v}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t\right|$$

Proving Main Result - Instantaneous Regret Bound

- C_t is bounded by K_t as a switch was not detected:

$$C_t \leq K_t = \left(\sqrt{2 \ln \frac{2}{\delta_t}} + 1 \right) \sqrt{r_t} + r_t + \left(\sqrt{2 \ln \frac{2}{\delta_t}} + 1 \right) \sqrt{r_{L_t}} + r_{L_t}$$

- From properties of RLS estimator (Cesa-Bianchi et al.) applied to demo classifier h_t:

$$\left| \boldsymbol{v}^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t \right| \leq \left(\sqrt{2 \ln \frac{2}{\delta_t}} + 1 \right) \sqrt{r_{L_t}} + r_{L_t}$$

With probability $1 - \delta_t$.

Proving Main Result - Instantaneous Regret Bound

Combining bounds on C_t and on $\left| \boldsymbol{v}^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t \right|$:

$$\left| \boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t \right| \leq \sqrt{r_t} \left(\sqrt{2 \ln \frac{2}{\delta_t}} + 1 \right) + r_t$$
$$+ 2\sqrt{r_{L_t}} \left(\sqrt{2 \ln \frac{2}{\delta_t}} + 1 \right) + 2r_{L_t}$$

Proving Main Result - Instantaneous Regret Bound

Using identities:

- $\Pr[A] = \mathrm{E}[I_A]$
- $I_{\{x<1\}} \leq e^{1-x}$

final bound instantaneous regret R_t achieved:

$$R_t \leq \varepsilon I_{\{|\boldsymbol{v}^\top \boldsymbol{x}_t| < \varepsilon\}} + \Pr\left[\left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right| \geq \varepsilon\right]$$

$$\leq \varepsilon I_{\{|\boldsymbol{v}^\top \boldsymbol{x}_t| < \varepsilon\}} + 2\exp\left\{1 - \frac{\alpha_{\varepsilon,t}}{r_{L_t}}\right\} + 2\exp\left\{1 - \frac{\beta_\varepsilon}{r_{L_t}}\right\}$$

$$+ \exp\left\{1 - \frac{\alpha_{\varepsilon,t}}{r_t}\right\} + \exp\left\{1 - \frac{\beta_\varepsilon}{r_t}\right\} + \delta_t$$

Proving Main Result - Cumulative Regret Bound

Cumulative regret R_T is given by:

$$R_T = \sum_{t=1}^{T} R_t$$

Cumulative regret R_T will be bounded by summing over bound of instantaneous regret R_t.

Calculation outline:

1. Summation over the r_t terms - separate calculation for:
   - Rounds t for which with $r_t \leq t^{-\kappa}$
   - Rounds t for which with $r_t > t^{-\kappa}$
2. Summation over the r_{L_t} terms.
3. Deriving final bound

Proving Main Result - Cumulative Regret Bound

Summation over r_t terms - for rounds with $r_t > t^{-\kappa}$ -

- Identity $\exp\{-x\} \leq \frac{1}{ex}$ gives:

$$\sum_{t=T_1, r_t > t^{-\kappa}}^{T} \exp\left\{1 - \frac{\alpha_{\varepsilon,t}}{r_t}\right\} \leq \frac{1}{\alpha_{\varepsilon,T}} \sum_{t=T_1, r_t > t^{-\kappa}}^{T} r_t$$

- The result $r_t \leq \left(1 - \frac{\det A_{t-1}}{\det A_t}\right)$ (Cesa-Bianchi et. al 2004) yields:

$$\frac{1}{\alpha_{\varepsilon,T}} \sum_{t=T_1, r_t > t^{-\kappa}}^{T} r_t \leq \frac{1}{\alpha_{\varepsilon,T}} \sum_{t=T_1, r_t > t^{-\kappa}}^{T} \left(1 - \frac{\det A_{t-1}}{\det A_t}\right)$$

Proving Main Result - Cumulative Regret Bound

Summation over r_t terms - for rounds with $r_t > t^{-\kappa}$ -

- Identity $1 - x \leq -\ln x$ (for $x \leq 1$) gives:

$$\frac{1}{\alpha_{\varepsilon,T}} \sum_{t=T_1, r_t > t^{-\kappa}}^{T} \left(1 - \frac{\det A_{t-1}}{\det A_t}\right) \leq \frac{-1}{\alpha_{\varepsilon,T}} \sum_{t=T_1, r_t > t^{-\kappa}}^{T} \ln\left(\frac{\det A_{t-1}}{\det A_t}\right)$$

- Computing the sum will give final expression:

$$-\frac{1}{\alpha_{\varepsilon,T}} \sum_{t=T_1, r_t > t^{-\kappa}}^{T} \ln\left(\frac{\det A_{t-1}}{\det A_t}\right) \leq \frac{1}{\alpha_{\varepsilon,T}} \left\{d \ln T - \ln\left(\det A_{T_1}\right)\right\}$$

Proving Main Result - Cumulative Regret Bound

Summation over r_t terms - for rounds with $r_t \leq t^{-\kappa}$ -

- Substituting $r_t \leq t^{-\kappa}$ and replacing sum with integral yields:

$$\sum_{t=T_1, r_t > t^{-\kappa}}^{T} \exp\left\{1 - \frac{\alpha_{\varepsilon,t}}{r_t}\right\} \leq e \sum_{t=T_1, r_t > t^{-\kappa}}^{T} \exp\left\{-\frac{\alpha_{\varepsilon,t}}{t^{-\kappa}}\right\}$$

$$\leq e \int_{T_1}^{T} \exp\left\{-\alpha_{\varepsilon,T} t^{\kappa}\right\} dt =$$

$$= \frac{e}{\kappa \left(\alpha_{\varepsilon,T}\right)^{\frac{1}{\kappa}}} \left(\Gamma\left\{\frac{1}{\kappa}, \alpha_{\varepsilon,T} T_1^{\kappa}\right\} - \Gamma\left\{\frac{1}{\kappa}, \alpha_{\varepsilon,T} T^{\kappa}\right\}\right)$$

- Last equality follows from the identity:

$$\int \exp\{az^s\} dz = -\frac{z(-az^s)^{-\frac{1}{s}}}{s} \Gamma\left\{\frac{1}{s}, -az^s\right\}$$

Proving Main Result - Cumulative Regret Bound

Summation over r_{L_t} terms -

Matrix Chernoff bound - for a series of random, i.i.d PSD matrices $Z_k \in R^{d \times d}$ holds:

$$\Pr \left[\lambda_{\min} \left\{ \sum_k Z_k \right\} \le (1 - \gamma) \, \mu_{\min} \right] \le d \left(\frac{e^{-\gamma}}{(1 - \gamma)^{(1 - \gamma)}} \right)^{\frac{\mu_{\min}}{\rho}}$$

where:

- $\gamma \in (0, 1)$
- $\mu_{\min} = \lambda_{\min} \left\{ \sum_k \mathrm{E} \left[Z_k \right] \right\}$
- $\lambda_{\max} \left\{ \mathrm{E} \left[Z_k \right] \right\} \le \rho$

Proving Main Result - Cumulative Regret Bound

Summation over r_{L_t} terms -

- Assumption: smallest eigenvalue of covariance matrix grows linearly - $\lambda_{\min} \left\{ \sum\limits_{k=1}^{L} \mathrm{E}\left[\boldsymbol{x}_k \boldsymbol{x}_k^\top \right] \right\} \sim O\left(\frac{L}{d} \right)$

- Using Chernoff matrix bound on $Z_k = \boldsymbol{x}_k \boldsymbol{x}_k^\top$, under the above assumption, yields:

$$\lambda_{\min}\left\{ A_{L_t} \right\} = \lambda_{\min}\left\{ I + \sum_{k=1}^{L_t} \boldsymbol{x}_k \boldsymbol{x}_k^\top \right\} > (1-\gamma)\frac{L_t}{d} + 1$$

Proving Main Result - Cumulative Regret Bound

Summation over r_{L_t} terms -

Using the bound and identities below:

- For unit normed x: $x^\top M x \leq \lambda_{max}\{M\}$
- $\lambda_{max}\{M\} = \frac{1}{\lambda_{min}\{M^{-1}\}}$
- $\lambda_{min}\{A_{L_t}\} > (1 - \gamma)\frac{L_t}{d}$

we get:

$$r_{L_t} = x_t^\top \left(A_{L_t} + x_t x_t^\top \right)^{-1} x_t \leq \frac{d}{(L_t + 2)(1 - \gamma)}$$

Proving Main Result - Cumulative Regret Bound

Summation over r_{L_t} terms -

- Replacing $L_t = L_0 + \sqrt{t}$ into the bound would yield:

$$r_{L_t} \leq \frac{d}{\left(L_0 + \sqrt{t} + 2\right)\left(1 - \gamma\right)}$$

- Substituting bound r_{L_t} into the sum over regret R_T bound:

$$\sum_{t=T_1}^{T} \exp\left\{1 - \frac{\alpha_{\varepsilon,t}}{r_{L_t}}\right\} \leq e \sum_{t=T_1, r_t > t^{-\kappa}}^{T} \exp\left\{-\hat{\alpha}_{\varepsilon,t}\left(L_0 + \sqrt{t}\right)\right\}$$

Proving Main Result - Cumulative Regret Bound

Summation over r_{L_t} terms -

Now replacing sum with integral and solving as before yields:

$$e \sum_{t=T_1}^{T} \exp \left\{ -\hat{\alpha}_{\varepsilon,t} \left(L_0 + \sqrt{t} \right) \right\} \leq$$
$$\frac{2e}{\left(\tilde{\alpha}_{\varepsilon,T} \right)^2} \left(\Gamma \left\{ 2, \tilde{\alpha}_{\varepsilon,T} \left(T_1 - L_0 \right)^{\frac{1}{2}} \right\} - \Gamma \left\{ 2, \tilde{\alpha}_{\varepsilon,T} \left(T - L_0 \right)^{\frac{1}{2}} \right\} \right)$$

Proving Main Result - Cumulative Regret Bound

- Summing all developed bounds yields:

$$R_T \leq O\left(d\left\{\ln T\right\}^2\right)$$

- Cumulative regret controlled and small

- Bound overcomes switch effect - square logarithmic bound in T comparing to more than linear bound in τ

Proving Main Result - No False Positives

- Reminder -switch detection if $C_t > K_t$

- Assuring no false detection - if no switch occurs than $C_t \leq K_t$

- From triangle inequality:

$$C_t = \left| \boldsymbol{w}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t \right| \leq \left| \boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t \right| + \left| \boldsymbol{v}^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t \right|$$

Proving Main Result - No False Positives

- Reminder - from RLS estimator properties, holds with probability $1 - 2\delta_t$:

  - $\left| \boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t \right| \leq \sqrt{r_t} \left(\sqrt{2 \ln \frac{2}{\delta_t}} + 1 \right) + r_t$

  - $\left| \boldsymbol{v}^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t \right| \leq \left(\sqrt{2 \ln \frac{2}{\delta_t}} + 1 \right) \sqrt{r_{L_t}} + r_{L_t}$

- Substituting this into bound:

$$C_t \leq \sqrt{r_t} \left(\sqrt{2 \ln \frac{2}{\delta_t}} + 1 \right) + r_t$$
$$+ \left(\sqrt{2 \ln \frac{2}{\delta_t}} + 1 \right) \sqrt{r_{L_t}} + r_{L_t} \leq K_t$$

Proving Main Result - No False Positives

- Last result assures that if no switch occurred $C_t \leq K_t$ with probability $1 - 2\delta_t$

- Thus the probability for a false detection at round t is $2\delta_t$

- Using union bound - probability for a false detection throughout the algorithm is 2δ
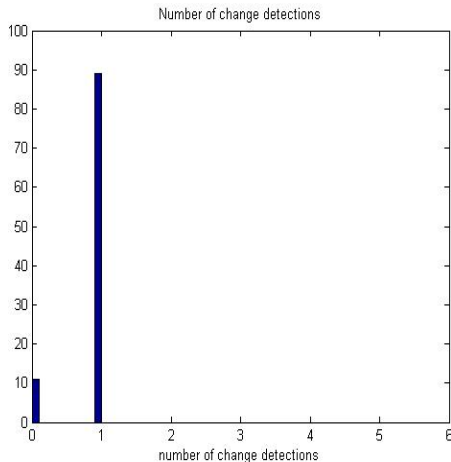
Simulation Results

Synthetic data simulation:

- $T = 10^5$, $\boldsymbol{x}_t \in R^4$, $\kappa = 0.7$, $L_0 = 400$, $K = 6$, $\delta = 0.05$
- Instances $\boldsymbol{x}_t$ drawn randomly from Gaussian distribution and then normalized
- Labels y_t drawn from Bernoulli distribution with $p_t = \frac{1 + \boldsymbol{u}^\top \boldsymbol{x}_t}{2}$

Results, averaged after 100 runs of the algorithm:

- Optimal classifier - 28.81% error (std 0.15%)
- BBQ RLS estimator classifier - 35.76% error (std 4.57%)
- Suggested switch detection algorithm - 29.57% error (std 0.6%)
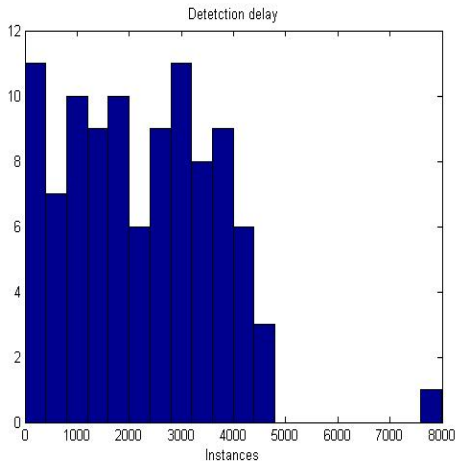- Suggested switch detection algorithm without extra querying - 29.7% error (std 0.48%)

Simulation Results - No false Detections

Number of switch detections, in 100 runs of the algorithm:

Number of change detections

Simulation Results - Switch Detected Relatively Fast

Distribution of switch delay, in 100 runs of the algorithm:
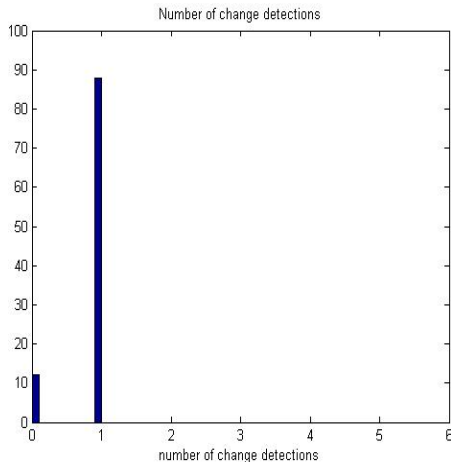

Detection delay

No Extra Querying Version of Algorithm

- Constructing demo classifiers increase number of queried labels

- Simulation results show increase from 12640 to around 28000 (std 873) in number of queried labels

- Version of the algorithm that uses only already queried labels for demo classifier construction was tested

- Algorithm achieves error of 29.7% (std 0.48%), comparing to 28.81% error of optimal classifier, 35.76% of BBQ RLS estimator classifier and 29.57% of previous setting shown

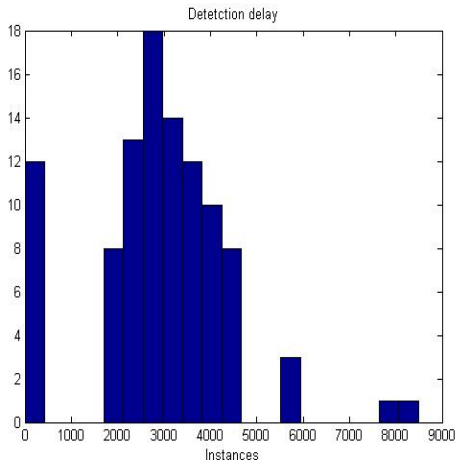- Number of queried labels reduced to 14844 (std 899)

Simulation Results - No false Detections

Number of switch detections, in 100 runs of the algorithm:

Simulation Results - Switch Detected Relatively Fast

Distribution of switch delay, in 100 runs of the algorithm:

Detetction delay

Problem Setting - Online Regression

- Analysis for regression is similar to one presented for classification.

- Differences between the two problems will be discussed.

Problem setup:

- $x_t \in R^d$
- $y_t \in R$

Problem Setting - Online Regression

Assumptions:

- $\|\boldsymbol{x}_t\| = \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1, \boldsymbol{u}, \boldsymbol{v} \in R^d$

- for $t \leq \tau$ holds:

  $y_t = \boldsymbol{u}^\top \boldsymbol{x}_t + \eta_t$ and $\mathrm{E}[y_t] = \boldsymbol{u}^\top \boldsymbol{x}_t$

- for $t > \tau$ holds:

  $y_t = \boldsymbol{v}^\top \boldsymbol{x}_t + \eta_t$ and $\mathrm{E}[y_t] = \boldsymbol{v}^\top \boldsymbol{x}_t$

- η_t i.i.d noise with $\mathrm{E}[\eta_t] = 0, var\{\eta_t\} = \sigma^2$

- $|\eta_t| \leq Z_\eta, Z_\eta$ is known

online Regression - Regret Definition

- Linear regression is to issue prediction:

$$\hat{y}_t = \boldsymbol{w}_t^\top \boldsymbol{x}_t$$

- As in classification $\boldsymbol{w}_t$ is the RLS estimator: $\boldsymbol{w}_t = A_t^{-1} b_t$

- Major difference - instantaneous regret definition:

$$
\begin{aligned}
R_t &= (y_t - \hat{y}_t)^2 - \left(y_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right)^2 \\
&= \left(y_t - \boldsymbol{w}_t^\top \boldsymbol{x}_t\right)^2 - \left(y_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right)^2 \\
&= \left(\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right)^2 - 2\left(y_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right)\left(\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right)
\end{aligned}
$$

Regret Bounds

- RLS properties used to bound $\left| \boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t \right|$

- To bound the cumulative regret R_T:

  1. $\sum\limits_{t=T_1}^{T} \left(\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t \right)^2$ will be bounded using RLS properties

  2. Azuma's inequality will be used to bound
     $\sum\limits_{t=T_1}^{T} \left(y_t - \boldsymbol{v}^\top \boldsymbol{x}_t \right) \left(\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t \right)$

Main Result for Regression

- If switch occurs it is either detected or low regret assured

- In case of non detection - $R_T \leq O\left(\sqrt{d}T^{\left(1-\frac{2\kappa+1}{4}\right)}\ln T\right)$

- Improving expected bound due to effect of switch:
  $R_T \leq O\left(d^2\tau^{2\kappa}\left\{\ln\tau\right\}^2 T^{1-\kappa}\right)$

- Result close to bound in no switch case - $O\left(T^{1-\kappa}\ln T\right)$

- Probability for false detection - 2δ

- Note - in regression problem selective sampling increases regret

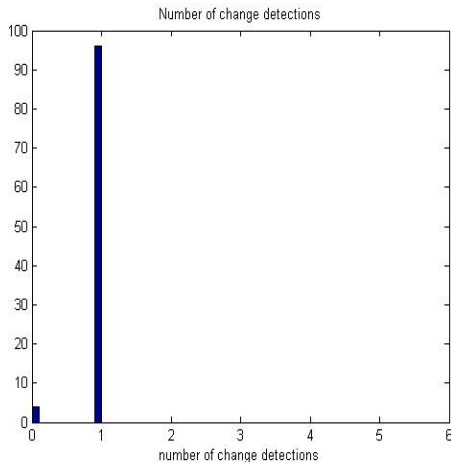## Simulation Results

Synthetic data simulation:

- $T = 10^5$, $x_t \in R^4$, $\kappa = 0.7$, $L_0 = 400$, $K = 6$, $\delta = 0.05$
- Instances x_t drawn randomly from Gaussian distribution and then normalized
- $y_t = u^\top x_t + \eta_t$. η_t Gaussian noise, with $\mathrm{E}[\eta_t] = 0$, $\sigma = 0.4$, $Z_\eta = 2\sigma$

Results, averaged after 100 runs of the algorithm:

- Optimal classifier - 0.1473 mean error (std 0.0005)
- BBQ RLS estimator classifier - 0.2821 mean error (std 0.0629)
- Suggested switch detection algorithm - 0.1602 mean error (std 0.0083)
- Suggested switch detection algorithm without extra querying - 0.1629 mean error (std 0.0061)
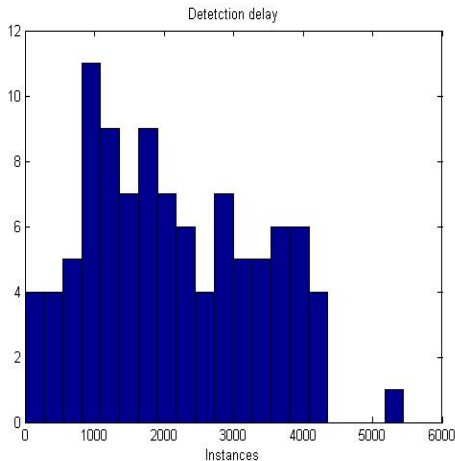
Simulation Results - No false Detections

Number of switch detections, in 100 runs of the algorithm:

Number of change detections

Simulation Results - Switch Detected Relatively Fast

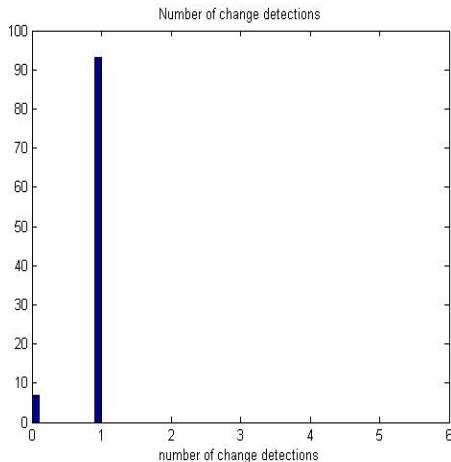Distribution of switch delay, in 100 runs of the algorithm:

Simulation Results - Number of Queries

- Constructing demo classifiers increase number of queried labels

- Simulation results show increase from 12640 to around 28000 (std 616) in number of queried labels

- Version of the algorithm that uses only already quired labels for demo classifier construction was tested

- Algorithm achieves error of 0.1629 (std 0.0061), comparing to 0.1473 error of optimal classifier, 0.2821 of BBQ RLS estimator classifier and 0.1602 of previous setting shown

- Number of queried labels reduced to 14986 (std 748)
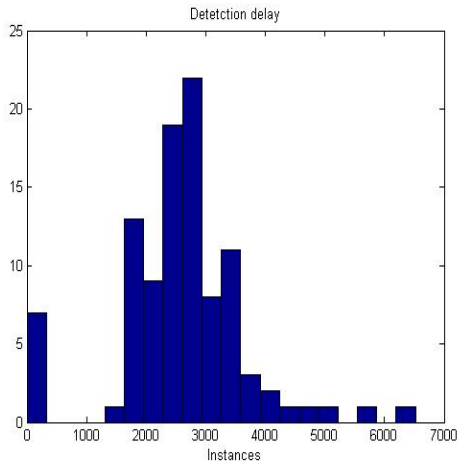
Simulation Results - No false Detections

Number of switch detections, in 100 runs of the algorithm:

Number of change detections

Simulation Results - Switch Detected Relatively Fast

Distribution of switch delay, in 100 runs of the algorithm:

Proposed Method

- Problem approached - switch in data at online linear classification and regression settings

- Proposed solution -

  1 Using confidence notion of selective sampling for switch detection

  2 Constructing demo classifier h_t from recent time window

  3 Difference between estimator and demo classifier $C_t = \left| \boldsymbol{w}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t \right|$ indicates switch

  4 Difference considered with respect to confidence r_t about instance $\boldsymbol{x}_t$

Main Results

- Algorithm either detects switch or assurers low regret in case of non-detection

- Low probability for false detection

- Simulations on synthetic data results:

  - Most switches were detected in relativity short time

  - Error reduced close to optimal result

  - No false detections

Model Limitations

- Strong assumptions on instances x_t distribution

- Strong assumptions on label y_t distribution

- Union bound used on non disjoint events

- In regression setting noise bound Z_η assumed to be known

- Demo classifier construction increases number of queries

- Proposed methods are set for switch but not for drift scenarios

Thank You for Your Time

Questions?

Acknowledgments

- Prof. Koby Crammer for guidance, insights and ideas
- Asaf, Hadas, Daniel, Yonatan, Miri, Yoav, Haim, Itamar, Yehuda, Aviad, Matan, Edward for good times, fun, food and expanding my knowledge in machine learning
- Shahar, Rami, Assaf and Nadav for making the office a great place to be at
- Avinoam, Hagit, Merav, Igal, Misha and Tamir for more than 7 years of friendship in (and more important outside) the Technion
- My mother for everything
- Noa for love and support