Minimax Active Learning via Minimal Model Capacity

Shachar Shayovitz and Meir Feder

IEEE International Workshop on Machine Learning for Signal Processing 2019

TEL AVIV UNIVERSITY אוניברסיטת תל אביב

Outline

- Introduction
  - Passive learning
  - Active learning
  - Existing solutions

Outline

- Introduction
  - Passive learning
  - Active learning
  - Existing solutions
- Minimax active learning

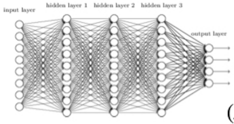
Outline

- Introduction
  - Passive learning
  - Active learning
  - Existing solutions
- Minimax active learning
- Low complexity algorithm for active learning of noisy linear separators

Passive Learning: Random Training

Unlabelled Dataset

Learner

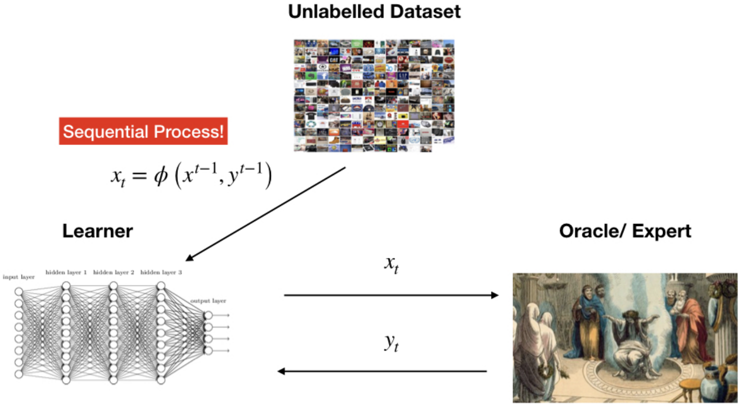


$x_1, x_2, x_3, \ldots, x_N$

Oracle/ Expert

$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$

Active Learning: Interaction with an Expert



Main Objective

How to choose examples interactively to learn faster than passive learning?

Existing Solutions

- Disagreement Region
  - $DIS(H_t) = \{x \in X : \exists f, \tilde{f} \in H_t, f(x) \neq \tilde{f}(x)\}$
  - Querying features in $DIS(H_t)$ will reduce the candidate set H_{t+1}: [CAL1994], A^2 [BBL2006]
  - Label complexity: $exp\left(-\frac{n}{d\theta_c}\right)$, where θ_c is the Disagreement Coefficient.

Existing Solutions

- Disagreement Region
  - $DIS(H_t) = \{x \in X : \exists f, \tilde{f} \in H_t, f(x) \neq \tilde{f}(x)\}$
  - Querying features in $DIS(H_t)$ will reduce the candidate set H_{t+1}: [CAL1994], A^2 [BBL2006]
  - Label complexity: $exp\left(-\frac{n}{d\theta_c}\right)$, where θ_c is the Disagreement Coefficient.
  - **High computational and label complexity (θ_c can be very large)**

Existing Solutions

- Disagreement Region
  - $DIS(H_t) = \{x \in X : \exists f, \tilde{f} \in H_t, f(x) \neq \tilde{f}(x)\}$
  - Querying features in $DIS(H_t)$ will reduce the candidate set H_{t+1}: [CAL1994], A^2 [BBL2006]
  - Label complexity: $exp\left(-\frac{n}{d\theta_c}\right)$, where θ_c is the Disagreement Coefficient.
  - **High computational and label complexity (θ_c can be very large)**
- Information Theoretic measures
  - Maximum Uncertainty (MU): $\max_X H(Y|X, D_{train})$
  - Maximum mutual information [HHGL2011]: $\max_X I(Y, \theta|X, D_{train})$
  - Different methods based on Fisher Information [SALED2017]

Existing Solutions

- Disagreement Region
  - $DIS(H_t) = \{x \in X : \exists f, \tilde{f} \in H_t, f(x) \neq \tilde{f}(x)\}$
  - Querying features in $DIS(H_t)$ will reduce the candidate set H_{t+1}: [CAL1994], A^2 [BBL2006]
  - Label complexity: $exp\left(-\frac{n}{d\theta_c}\right)$, where θ_c is the Disagreement Coefficient.
  - **High computational and label complexity (θ_c can be very large)**
- Information Theoretic measures
  - Maximum Uncertainty (MU): $\max_X H(Y|X, D_{train})$
  - Maximum mutual information [HHGL2011]: $\max_X I(Y, \theta|X, D_{train})$
  - Different methods based on Fisher Information [SALED2017]
  - **Heuristic criteria.**

Information Theoretic Minimax Active Learning

Mathematical Setup

Learning Setting

- Examples (x, y) are drawn from some family of hypotheses $p(y|x, \theta)$ where $\theta \in \Theta$.
- Test feature drawn from $p(x)$ - stochastic setting
- Labeling budget of N queries.
- Probabilistic learners: $q(y|x)$.
- Log-loss cost function: $-\log(q(y|x))$.

Mathematical Setup

Learning Setting

- Examples (x, y) are drawn from some family of hypotheses $p(y|x, \theta)$ where $\theta \in \Theta$.
- Test feature drawn from $p(x)$ - stochastic setting
- Labeling budget of N queries.
- Probabilistic learners: $q\left(y|x\right)$.
- Log-loss cost function: $-\log\left(q\left(y|x\right)\right)$.

Informal Objective

Sequentially select features based on past examples (x^N, y^N) and construct a learner, $q\left(y|x, x^N, y^N\right)$, which will perform well.

Mathematical Setup

Optimal Learner

- Similarly to the statistical learning approach, we would like to find a learner $\hat{q}(y|x)$ which minimizes:

$$\hat{q}(y|x) = \arg\min_{q} E_{p(y|x,\theta)} \left(-\log q(y|x)\right)$$

- Clearly this implies that $\hat{q}(y|x) = p(y|x, \theta)$ in KL sense.

Mathematical Setup

Optimal Learner

- Similarly to the statistical learning approach, we would like to find a learner $\hat{q}(y|x)$ which minimizes:

$$\hat{q}(y|x) = \arg\min_{q} E_{p(y|x,\theta)}\left(-\log q(y|x)\right)$$

- Clearly this implies that $\hat{q}(y|x) = p(y|x,\theta)$ in KL sense.

Problem

- Unfortunately, the learner has no access to the true θ.

Minimax Active Learning Formulation

- Find a sequential selection strategy $\left\{\phi(x_t|x^{t-1}, y^{t-1})\right\}_{t=1}^{N}$ which optimizes the minimax regret to the optimal learner for a random test point (x, y):

$$R = \min_{\{\phi_t\}_{t=1}^{N}} \min_{q} \max_{\theta} E\left\{\log\left(\frac{p\left(y|x,\theta\right)}{q\left(y|x, x^N, y^N\right)}\right)\right\}$$

  where x^N, y^N are the training examples.
- The expectation is performed over the joint probability:

$$p\left(y, x, x^N, y^N|\theta\right) = p\left(y|\theta, x\right)\Pi_{t=1}^{N} p\left(y_t|x_t, \theta\right)\phi\left(x_t|x^{t-1}, y^{t-1}\right)p(x|\theta)$$

Capacity Redundancy Theorem for Active Learning

Theorem [SF19]

The minimax active learning problem is equivalent to the following criterion:

$$R = \min_{\{\phi(x_t|x^{t-1},y^{t-1})\}_{t=1}^{N}} C_{Y;\theta|X,Y^N,X^N}$$

where,

$$C_{Y;\theta|X,Y^N,X^N} = \max_{\pi(\theta)} I\left(Y;\theta|X,Y^N,X^N\right)$$

and the optimal learner is:

$$q^*\left(y|x,x^N,y^N\right) = \sum_{\theta} p\left(\theta|y^N,x^N\right) p\left(y|\theta,x\right)$$

Exploitation - Exploration Trade-Off Interpretation

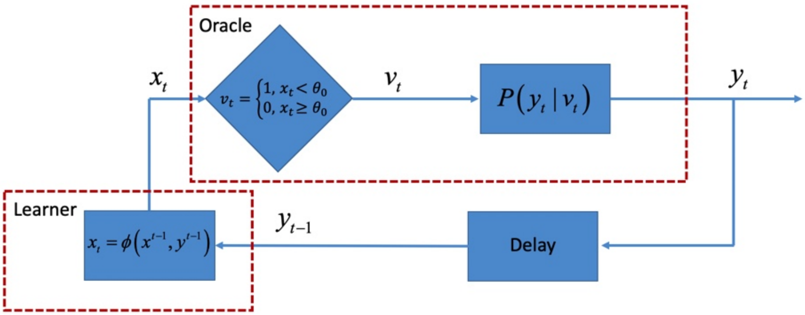
- Our new Active Learning criterion can be upper-bounded by:

$$I\left(Y;\theta|X,Y^N,X^N\right) \leq H\left(Y|X\right)+\sum_{t=1}^{N} H\left(Y_t|X^t,Y^{t-1},X,Y\right)-H\left(Y_t|X^t,Y^{t-1}\right)$$

  - $H\left(Y_t|X^t,Y^{t-1}\right)$ can be viewed as "exploration" and greedy maximization of it is equivalent to MU.
  - $H\left(Y_t|X^t,Y^{t-1},X,Y\right)$ can be viewed as "exploitation".
- Minimizing the difference means that *there is a fundamental trade-off between exploration and exploitation in our criterion.*

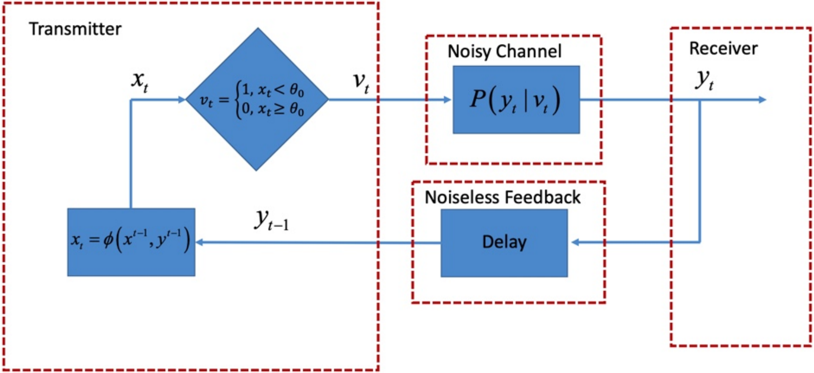
Active Learning of Linear Separators with Label Noise

One Dimensional Linear Separator with Noisy Oracle



Possible Solution for Minimax Active Learning

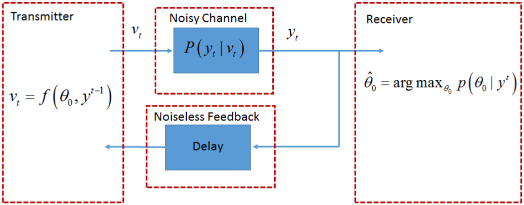
- The Idea is to look at the problem as communicating θ_0 over a noisy channel.
- Pass as much information bits on θ_0 using few channel uses and correctly decode θ_0.



Posterior Matching Scheme

- Capacity achieving scheme proposed by Shayevitz and Feder (2007), suitable for any memory-less channel $P(Y|V)$.
- The estimation error on θ_0 drops exponentially fast.
- Next symbol v_t is computed via:

$$v_t = F_V^{-1}\left(F_{\theta_0|Y^{t-1}}\left(\theta_0|y^{t-1}\right)\right)$$



Posterior Matching Scheme

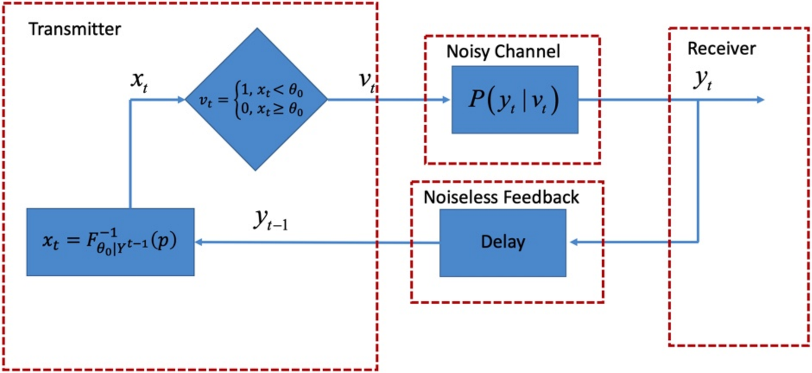
- For a binary valued v_t, with $V \sim Ber(p)$, the PM scheme reduces to:

$$v_t = \begin{cases} 1, & \text{if } \theta_0 > F^{-1}_{\theta|y^{t-1}}(p) \\ 0, & \text{otherwise} \end{cases}$$

- where $Ber(p)$ is the capacity achieving distribution for the noisy channel.

Active Learning with Noisy Labeler - 1d

If we choose $\phi\left(x_t|x^{t-1}, y^{t-1}\right) = F^{-1}_{\theta|y^{t-1}}(p)$, we achieve capacity!



High Dimensional Linear Separators

- Features $\underline{x} \in \mathbb{R}^d$ satisfy $||\underline{x}|| \leq R$ with uniform $p(\underline{x})$.
- The hypotheses class contains all possible hyper-planes with normal vector $\underline{w}$ and threshold b.
- The relation between feature $\underline{x}$ and **clean** label v is defined as,

$$\mathrm{p}(\mathrm{v}|\underline{\mathrm{x}},\underline{\mathrm{w}},\mathrm{b}) = \begin{cases} 1 & \textit{if } \underline{\mathrm{w}}^T \underline{x} > b \\ 0 & \textit{otherwise} \end{cases}$$

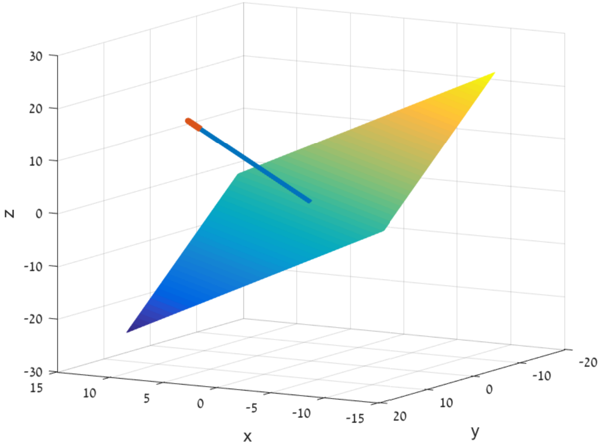
- v passes through a discrete memory-less channel $p(y|v)$ and produces the noisy label - y.

Successive Posterior Matching (SPM)

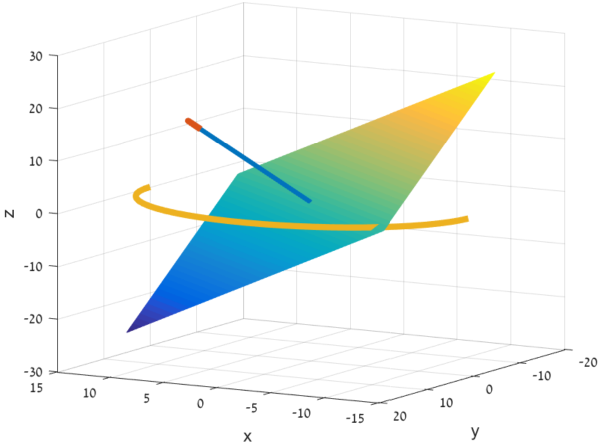
SPM Idea

- True classifier is fully described by its normal vector.
- The idea is to successively localize the spherical coordinates of the normal vector $\underline{w}$ using Posterior Matching.
- Each coordinate lives on the arc: $\theta_i \in [0, \pi]$.
- The intersection of the hyper-plane and the arc is the barrier between classification regions.
- For each spherical coordinate we have a noisy one dimensional barrier problem.

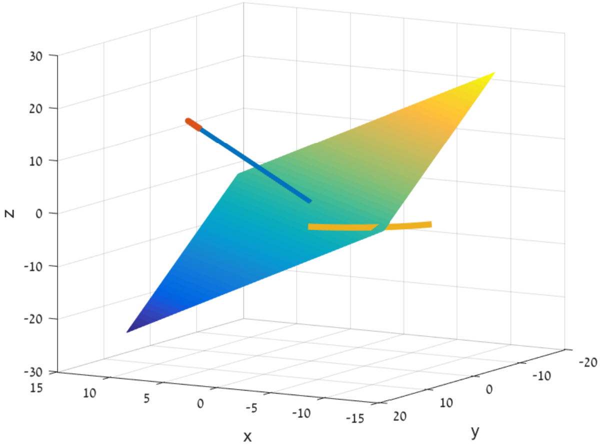
Classifier



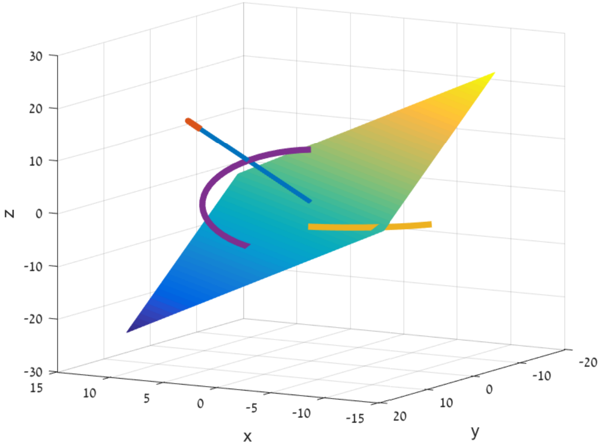
PM on Azimuth



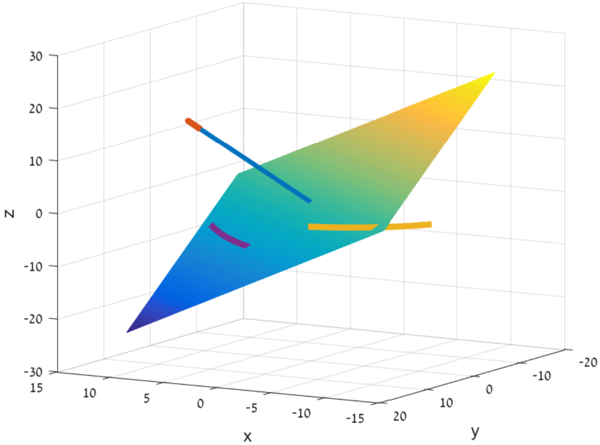
Estimated Barrier between Classification Regions



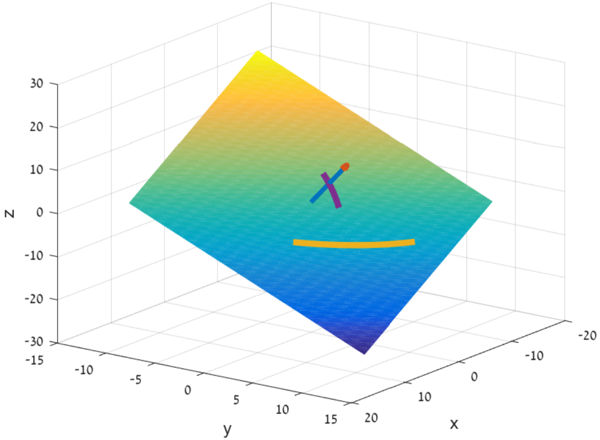
PM on Elevation



Estimated Barrier between Classification Regions



Estimated Normal Vector



Active Learning Criterion with SPM selection

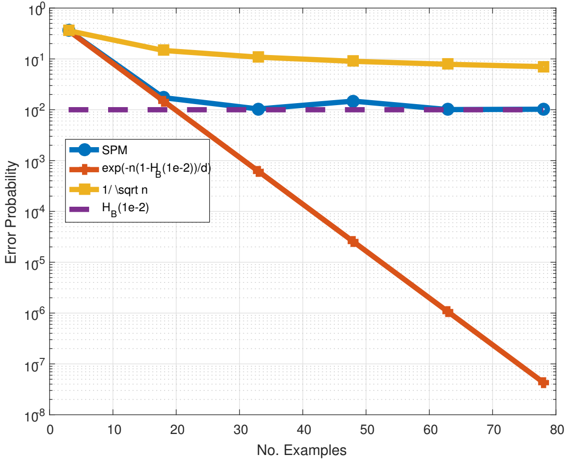
Theorem [SF19]

For the d dimensional binary linear separator hypotheses class with discrete memory-less label noise and uniform $p(\underline{x})$, the SPM algorithm produces a selection policy such that,

$$I\left(\theta; Y|\underline{X}, \underline{X}^N, Y^N\right) \approx O\left(2^{-\frac{N}{d} C_{Channel}}\right)$$

where $C_{Channel}$ is the channel capacity.

SPM: Error probability for BSC label noise



Summary

Minimax Active Learning

- Capacity Redundancy theorem for minimax active learning
- Optimal learner for minimax active learning.

Active Learning of Linear Separators

- Near-optimal, low complexity, algorithm for active learning of Linear Separators with various noise models.
- Explicit expression for the decay factor of the Mutual Information.

Thank You!